Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 12, Número 1 (2020)

ISSN: 1647-0818

Volume 12, Número 1 – 2020

LinguaMÁTICA

ISSN: 1647–0818

Editores

Alberto Simões
José João Almeida
Xavier Gómez Guinovart

Conteúdo

Artigos de Investigação

Editorial

Caras lectoras e caros lectores,

Chegamos ao duodécimo volume da Linguamática, sumando xa vinte e catro exemplares da revista nos seus doce anos de vida. Nesta longa xeira, tentamos sempre fornecer unha publicación académica de calidade con artigos científicos orixinais que presentasen, en cada momento, o estado da cuestión da investigación no eido do procesamento da linguaxe natural no dominio lingüístico específico das nosas linguas peninsulares: portugués, galego, catalán, castelán, éuscaro, asturiano, mirandés ou aranés. Alén diso, promovemos que os artigos estivesen sempre escritos preferentemente nunha desas nosas linguas, coa aspiración de crear conciencia de comunidade científica e favorecer sinerxías entre os diversos grupos de investigación dentro deste ámbito.

Aínda que non todos os nosos obxectivos foron atinxidos aínda na súa totalidade, o certo é que a Linguamática foi aumentando a súa valoración nos distintos índices mundiais creados para a avaliación da calidade das revistas científicas, o que consideramos un recoñecemento para todos e todas as que colaboramos en levar adiante este proxecto. Neste senso, logo de Scimago/Scopus actualizaren os indicadores SJR das revistas para 2019 e o seu Journal & Country Rank, na nova clasificación a Linguamática ascendeu á clase Q1, isto é, ao primeiro "cuartil" de Scopus no que se catalogan as mellores revistas en cada categoría.

Neste número que xa fai vinte e catro, incluímos oito artigos de investigación elaborados por equipos de investigación en PLN das linguas da Península Ibérica procedentes de Chile, Francia, Brasil, Noruega, España, Galiza, Portugal e o País Vasco. Esperamos que a súa lectura resulte de interese e sexa ben proveitosa.

Finalmente, queremos agradecer de corazón, coma sempre, o labor inmenso desenvolvido por autoras, autores, revisoras e revisores na realización deste número. Sen vós, a Linguamática simplemente non existiría. Obrigados!

Xavier Gómez Guinovart
José João Almeida
Alberto Simões

Comissão Científica

Alberto Álvarez Lugrís,
Universidade de Vigo

Alberto Simões,
Universidade do Minho

Aline Villavicencio,
Universidade Federal do Rio Grande do Sul

Álvaro Iriarte Sanroman,
Universidade do Minho

Ana Frankenberg-Garcia,
University of Surrey

Anselmo Peñas,
Univers. Nac. de Educación a Distancia

Antón Santamarina,
Universidade de Santiago de Compostela

Antoni Oliver González,
Universitat Oberta de Catalunya,

Antonio Moreno Sandoval,
Universidad Autónoma de Madrid

António Teixeira,
Universidade de Aveiro

Arantza Díaz de Ilarraza,
Euskal Herriko Unibertsitatea

Arkaitz Zubiaga,
Dublin Institute of Technology

Belinda Maia,
Universidade do Porto

Bruno Martins,
Instituto Superior Técnico

Carmen García Mateo,
Universidade de Vigo

Diana Santos,
Linguateca/Universidade de Oslo

Ferran Pla,
Universitat Politècnica de València

Gael Harry Dias,
Université de Caen Basse-Normandie

Gerardo Sierra,
Univers. Nacional Autónoma de México

German Rigau,
Euskal Herriko Unibertsitatea

Helena de Medeiros Caseli,
Universidade Federal de São Carlos

Horacio Saggion,
University of Sheffield

Hugo Gonçalo Oliveira,
Universidade de Coimbra

Iñaki Alegria,
Euskal Herriko Unibertsitatea

Irene Castellón Masalles,
Universitat de Barcelona

Iria da Cunha,
Universidad Nacional de Educación a Distancia

Joaquim Llisterri,
Universitat Autònoma de Barcelona

José João Almeida,
Universidade do Minho

José Paulo Leal,
Universidade do Porto

Joseba Abaitua,
Universidad de Deusto

Juan-Manuel Torres-Moreno,
Lab. Informatique d'Avignon - UAPV

Kepa Sarasola,
Euskal Herriko Unibertsitatea

Laura Plaza,
Complutense University of Madrid

Lluís Padró,
Universitat Politècnica de Catalunya

Manex Agirrezabal,
University of Copenhagen (KU), Denmark

Marcos Garcia,
Universidade da Corunha

María Inés Torres,
Euskal Herriko Unibertsitatea

Maria das Graças Volpe Nunes,
Universidade de São Paulo

Mário Rodrigues,
Universidade de Aveiro

Mercè Lorente Casafont,
Universitat Pompeu Fabra

Miguel Solla Portela,
Universidade de Vigo

Mikel Forcada,
Universitat d'Alacant

Pablo Gamallo Otero,
Universidade de Santiago de Compostela

Patrícia Cunha França,
Universidade do Minho

Patricia Martin Rodilla
Universidade de Santiago de Compostela

Ricardo Rodrigues
CISUC / Instituto Politécnico de Coimbra

Rui Pedro Marques,
Universidade de Lisboa

Susana Afonso Cavadas,
University of Exeter

Tony Berber Sardinha,
Pontifícia Univ. Católica de São Paulo

Xavier Gómez Guinovart,
Universidade de Vigo

Artigos de Investigação

Propuesta recibida en mayo 2019 y aceptada para publicación en mayo 2020.

Relación entre calidad de escritura y rasgos lingüístico-discursivos en las introducciones de los trabajos finales de grado de ingeniería civil informática

Relationship between quality of writing and linguistic-discursive features in the introductions of the Undergraduate thesis of Civil Engineering in Computer science

Fernando Lillo-Fuentes
Pontificia Universidad Católica de Valparaíso, Chile
fernando.lillo@pucv.cl

René Venegas
Pontificia Universidad Católica de Valparaíso, Chile
rene.venegas@pucv.cl

Resumen

En este artículo nos proponemos relacionar la calidad de la escritura con un amplio conjunto de rasgos lingüísticos y discursivos presentes en introducciones de los trabajos finales de grado de Ingeniería civil informática.

Para ello se evaluaron 101 introducciones, utilizando una rúbrica diseñada para este efecto. Se realiza un estudio correlacional entre la evaluación de la calidad y 203 variables léxicas, de extensión, sintácticas y discursivas. Luego, se realizó un análisis de regresión lineal con el fin de identificar variables predictoras de la calidad de las introducciones.

Entre nuestros hallazgos destacamos que la extensión (promedio de palabras) presenta una relación negativa fuerte con la calidad de la introducción, a su vez el uso de conectores conclusivos y la complejidad sintáctica se correlacionan de manera positiva con la calidad de las introducciones. Así también, el cumplimiento de propósitos comunicativos, en particular la indicación del vacío, presenta correlaciones positivas medias con la calidad del escrito. Con el análisis de regresión se pudo identificar que las tres primeras variables tienen valores estadísticos significativos para predecir las calidad de la introducción de acuerdo a los valores promedio de evaluación asignada por los evaluadores. Estos resultados son de interés para aplicaciones computacionales de apoyo a la escritura académica.

Palabras clave

calidad escritura, rasgos lingüístico-discursivos, introducción, trabajos finales de grado

Abstract

In this article we propose to relate the quality of writing with a wide range of linguistic and discursive features present in the introductions of the final works of the degree in Civil Engineering in Computer Science.

For this purpose, an evaluation of 101 introductions is carried out, based on a rubric designed for this purpose. A correlation study is carried out between the evaluation of quality and 203 variables identified in the texts. Then, a linear regression analysis was carried out in order to identify predictive variables of the quality of the introductions.

Among our findings we highlight that the length (average of words) presents a strong negative relationship with the quality of the introduction, in turn the use of conclusive connectors and syntactic complexity are positively correlated with the quality of the introductions. Also, the fulfillment of communicative purposes, in particular the indication of the gap, presents mean positive correlations with the quality of the writing. With the regression analysis it was possible to identify that the first three variables have significant statistical values to predict the quality of the introduction according to the average evaluation values assigned by the evaluators. These results are of interest for computational applications to support academic writing.

Keywords

quality of writing, linguistic-discoursive features, introductions, undergraduate graduation project

DOI: 10.21814/lm.12.1.297

1. Introducción

Desde la última década del siglo veinte, ha proliferado la calificación o evaluación automática de ensayos, entendida como una práctica que otorga puntuación y, en algunos casos, retroalimentación a los escritos de los estudiantes (Vajjala, 2018). La masificación de su empleo se ha debido a que los resultados se asemejan a las evaluaciones realizadas por revisores humanos.

A pesar del buen desempeño que han tenido las herramientas desarrolladas con estos fines, los investigadores no han llegado a un consenso acerca de las características o rasgos que permiten evaluar automáticamente la calidad de un texto (Vajjala, 2018). Además, ninguna de ellas se ha enfocado en indagar los rasgos lingüístico-discursivos que tienen una mayor predicción para evaluar los textos escritos, principalmente, porque han sido desarrolladas por investigadores del campo de la informática y se han creado más bien con fines comerciales (Dikli, 2006; Vajjala, 2018). Así, la falta de indagaciones de estos rasgos ha provocado que en cada herramienta se utilicen rasgos diversos, sin una verificación exhaustiva de sus niveles de predictibilidad.

A su vez, gran parte de los estudios en esta área se han desarrollado en torno a las respuestas cortas y ensayos académicos como acreditación del dominio de lenguas extranjeras. Si bien se destaca la importancia de estos géneros en el contexto académico, existen otros que no han recibido la atención necesaria en este campo de investigación. Un ejemplo de lo anterior son las tesis de licenciatura, género discursivo clave para la transición entre la vida académica y la profesional.

Sumado a lo anterior, la mayoría de las investigaciones se han llevado a cabo en inglés, por lo que en español solo existen algunos incipientes acercamientos a evaluaciones automáticas en géneros similares. Sin embargo, estas aproximaciones se han centrado, mayoritariamente, en las técnicas que permitirían evaluar los ensayos, por sobre los rasgos que permitirían predecir la evaluación de la calidad de un género académico en particular. Debido a lo anterior, el objetivo general que guiará este estudio será relacionar la calidad de la escritura con rasgos lingüístico-discursivos presentes en introducciones de los trabajos finales de grado de ingeniería civil informática.

En lo que sigue, se expondrán algunas nociones fundamentales para nuestra investigación, se plantearán los sustentos metodológicos, se expondrán algunos resultados y una breve discusión en torno a los principales hallazgos del estudio. Por último, se darán a conocer las conclusiones y algunas proyecciones, surgidas a propósito de esta investigación.

2. Trabajos relacionados

Project Essay Grade[1], E-Rater[2] e IntelliMetric[3] son algunos ejemplos de los sistemas que se han utilizando para evaluar automáticamente la calidad de los ensayos. Estas herramientas utilizan variadas características o índices para construir sus modelos predictivos y otorgar puntuación al escrito. Así, según Dikli (2006) se entrenan sistemas que predicen el juicio humano de anotación, utilizando índices tales como longitud de palabras y oraciones, ortografía, diversidad léxica y fluidez. A pesar del buen desempeño que han tenido estas herramientas, aún no existe consenso acerca de las características o rasgos lingüísticos que permiten evaluar la calidad un texto (Vajjala, 2018), por lo que se han realizado nuevos acercamientos para mejorar la predicción, a partir de los rasgos lingüísticos implicados en dicha tarea.

En este ámbito, McNamara et al. (2010) utilizando ensayos argumentativos, demostraron que los textos puntuados con alta calidad son los que poseen mayor diversidad léxica, alta frecuencia de palabras y mayor complejidad sintáctica. Complementando lo ya expuesto, en estudios posteriores, Guo et al. (2013) determinaron que la longitud del texto, la sofisticación léxica y la complejidad sintáctica permitían caracterizar textos de alta calidad. También, identificaron que en estos escritos se utilizaba con mayor frecuencia el pasado participio y la voz pasiva. A diferencia de lo propuesto por McNamara et al. (2010), Guo et al. (2013) identificaron una correlación negativa entre la cohesión y la calidad del texto, lo que significa que, a menor cohesión textual mayor calidad del escrito.

Por su parte, Crossley et al. (2014) demostraron que los ensayos de mayor calidad textual son aquellos que contienen mayor diversidad de palabras, palabras más infrecuentes en sus párrafos y no utilizan la segunda persona singular. Posteriormente, Crossley et al. (2016) expusieron que el mejor indicador de la calidad de un escrito es la longitud de la oración, pues permite identificar la calidad del escrito hasta en un 60 %. Además de este indicador, la diversidad léxica y, en una

[1] https://www.measurementinc.com/products-services/automated-essay-scoring

[2] https://www.ets.org/erater/about

[3] http://www.intellimetric.com/direct/

menor medida, la cohesión global, permiten diferenciar entre textos de alta y baja calidad (Crossley et al., 2016). Con respecto a la complejidad sintáctica, los autores concuerdan con McNamara et al. (2010) y Guo et al. (2013) en sostener que esta es un buen indicador, si se entiende como el número medio de palabras que antecede al verbo de la oración principal.

Como se desprende de lo hasta aquí revisado, la cohesión textual fue considerada por la mayoría de los autores como uno de los indicadores de la calidad textual. Sin embargo, en las últimas investigaciones del grupo fundador de CohMetrix, se ha demostrado que este indicador solo predice entre un 6 y 11 % la calidad de un escrito (Crossley et al., 2016; Perin et al., 2017). A diferencia de lo propuesto por el equipo de McNamara, en un estudio posterior, MacArthur et al. (2019) demostraron que la cohesión textual se relaciona con la calidad del texto, pero solo con la cohesión referencial, aunque el valor predictivo no fue considerado dentro de los mejores índices.

3. Marco teórico

3.1. Macrogénero Trabajo final de grado

El trabajo final de grado (TFG) se configura como un macrogénero, pues adquiere diversas formas textuales y distintos nombres (tesis, tesina, artículo, memoria, trabajo final, ensayo, entre otros) según las diferentes comunidades discursivas en las que circule y se desarrolle (Venegas-Velasquez, 2014).

Todos ellos comparten un propósito comunicativo acreditativo-evaluativo, un tipo de audiencia especializada y un registro académico disciplinar por medio del que se instancian (Venegas et al., 2016). Desde esta perspectiva, el TFG es entendido como un trabajo de investigación escrito de carácter evaluativo acreditativo, presentado por los estudiantes universitarios al término de sus estudios, como requisito para la obtención del grado académico de licenciado, de magíster o de doctor. Su finalidad es informar y acreditar los méritos que posee un estudiante como investigador (Venegas-Velasquez, 2014; Zamora & Venegas, 2013; Venegas et al., 2016).

Asimismo, el TFG corresponde a una práctica discursiva clave en el paso de la vida estudiantil universitaria a la académico-científica, ya que se ha transformado en un rito de iniciación para el novato que ingresa a una nueva comunidad (Moyano, 2000; Koutsantoni, 2006; Venegas-Velasquez, 2014).

3.2. Tesis de Licenciatura

Entre los géneros que componen el macrogénero trabajo final de grado, uno de los más destacados es el género tesis. Para Carlino Cantis (2003) "el cumplir con la escritura de la tesis y sus requisitos es trascendental para que el estudiante se inserte en su comunidad y se inicie en el camino de la cultura de la investigación". Asimismo, Meza (2013) expone que la principal característica de una tesis es que los estudiantes de pre y postgrado deben ser capaces de glosar el discurso de otro, transformándolo y apropiándose de él, pues el nuevo integrante debe comunicar el conocimiento y ha de hacerlo de acuerdo a las normas establecidas por su disciplina.

En esta investigación, el género tesis de licenciatura lo entenderemos, siguiendo a Tamola de Spiegel (2005), como un trabajo escrito que cumple con la función de informar acerca del proceso y los resultados de una investigación teórica o empírica con el fin de obtener el grado académico de licenciado. Este texto es dirigido por un docente investigador, quien guiará y revisará el trabajo realizado por el alumno que tiene a su cargo. Para el caso de las tesis de licenciatura, si bien, no se exige un alto nivel de originalidad, sí se requiere que la producción contenga la cantidad de información necesaria para que las conclusiones sean sustentadas. A su vez, el estudiante debe utilizar un método sistemático y un modo de transmisión adecuado, según las expectativas y requerimientos de la comunidad disciplinar.

3.3. Calidad de escritura

Respecto al término calidad de un texto o de escritura, si bien, no existe una definición canónica, algunos autores se han aproximado al concepto desde los parámetros o criterios que permiten evaluarla. Así, Sánchez Ceballos (2014), retomando la investigación de Beaugrande & Dressler (1997), menciona que un texto es de calidad cuando se puede evaluar en todas sus dimensiones formales y cognitivas, a partir de seis de los siete criterios propuestos por los autores, a saber: intencionalidad, situacionalidad, adecuación, informatividad, cohesión y coherencia. Para Sánchez Ceballos (2014), estos criterios no solo permiten evaluar la calidad de un texto, sino que demuestran que está bien construido y lo validan como un evento de comunicación.

Por su parte, Castelló (2002) expone que para producir un texto de calidad se deben tener en consideración siete tipos de dominios, a saber, conocimiento y profundización del tema, conocimiento lingüístico, es decir, cumplimiento

de reglas léxicas, gramaticales y ortográficas, conocimiento de la situación retórica y, finalmente, conocimiento del género textual. Estos dominios han sido considerados por diversos autores a la hora de evaluar la calidad de la producción escrita.

A los dominios ya descritos, se suma la organización o estructura de un escrito, pues Jin (2001) menciona que la estructura y la similitud o cercanía entre las partes consecutivas de un texto permiten evidenciar la calidad del escrito.

A partir de las diversas propuestas que se han revisado para realizar esta investigación, se puede desprender la calidad de la escritura se ha asociado comúnmente con algunos criterios lingüísticos y estructurales que permiten evaluarla (McNamara et al., 2010; Crossley et al., 2014; Vajjala, 2018), pero no se ha encontrado una definición canónica respecto al término. Lo anterior se puede deber a que la calidad de la escritura no se puede evaluar de forma general, pues se debe considerar que cada género discursivo tiene sus propias características y, por ende, sus respectivas exigencias (Figueroa et al., 2019). Además, las convenciones propias de la comunidad discursiva, la situación retórica y la tarea de escritura afectarán lo que en la comunidad se entienda por calidad textual.

Dado lo anterior, para esta investigación entenderemos que la calidad de un texto corresponderá a la adecuación a la tarea de escritura, al género y a todas las condiciones específicas de producción textual en el contexto académico específico consideradas por el escritor en su texto.

4. Metodología

En la presente investigación, de tipo no experimental, alcance correlacional y enfoque cuantitativo, se identifican y relacionan rasgos lingüístico-discursivos con la calidad de las introducciones de los trabajos finales de grado (TFG) de Ingeniería Civil Informática.

4.1. Corpus de estudio

La muestra del presente estudio se compone de 101 introducciones o Macromovidas introducir al lector (MM1) de los trabajos finales de grado de Ingeniería Civil Informática de la PUCV realizadas entre los años 2009 y 2017. Esta muestra fue seleccionada a partir del subcorpus de tesis (TINF) del proyecto *Fondecyt 1140967* y desde el sistema de biblioteca de la universidad. La muestra es representativa con un nivel de confianza del 95 % y un intervalo de confianza de $Z = 1,96$.

A continuación, en el cuadro 1 se ilustran las características de la selección de la muestra y la conformación del corpus empleado en la presente investigación.

Nº introducciones del corpus	101
% introducciones del corpus	96 %
Método selección muestra	Aleatorio
Nº total palabras del corpus	17.946

Cuadro 1: Características del corpus empleado

4.2. Procedimientos generales

Una vez recolectados los TFG mencionados, se procedió a separar las macromovidas 1 (Introducciones) y se convirtieron los archivos a formato txt. Posteriormente, basándonos en la noción de calidad referida en el apartado anterior, se procedió a construir una pauta de evaluación para evaluar manualmente la calidad de los escritos (ver anexo A). El instrumento de evaluación contiene una medición holística y otros indicadores relacionados con la calidad del texto. Para medir cada uno de los criterios, se utilizó una escala con un intervalo de seis puntos (Crossley et al., 2014). En ella se consideró 1 como no cumplido y 6 como cumplido totalmente. Esta pauta fue validada por expertos del área de la lingüística y de la informática, pues se requería una aprobación a partir de sus conocimientos disciplinares, en el caso de los primeros, y de sus conocimientos como profesores en el área, para los últimos.

Una vez validada la pauta, se entrenó a 3 analistas (lingüistas) en un periodo de tres horas cada uno (9 horas total), posteriormente 30 introducciones fueron evaluadas por los mismos. Se calcularon los valores Kappa de Fleiss y se repitió el entrenamiento hasta que se obtuvieran valores superiores a 0.7 (acuerdo considerable). Una vez finalizada la etapa de entrenamiento, se procedió a evaluar las demás introducciones. De esta manera, aquellas que obtenían un porcentaje de logro mayor o igual a 70 % se consideraron de alta calidad, mientras que las con valores menores o iguales a 40 % fueron consideradas de baja calidad. En el análisis estadístico *T-student* se observó diferencia estadística ($p = 5169 \times 10^{-6}$) entre las introducciones de alta calidad ($\overline{x} = 5,03$) y las de baja calidad ($\overline{x} = 3,43$).

Una vez separadas las introducciones en alta y baja calidad, se realizaron las mediciones con los índices que, según la literatura permitían evaluar la calidad de los escritos. De esta manera, se comenzó con los rasgos lingüístico-discursivos. Para efectuar estas mediciones de manera automática,

se empleó SINLP[4] (Crossley et al., 2014) con diccionarios en español y PACTE[5].

Luego, se procedió a evaluar los índices que permiten medir la calidad de los escritos. Para ello, se calculó el grado de asociación o relación entre la calidad del texto, la complejidad sintáctica, la cohesión textual, el vocabulario especializado, la frecuencia de palabras y los índices básicos de información textual. Para ello, se utilizaron los softwares R[6] y Jasp[7]. Posteriormente, se realizó un modelo de regresión lineal con las variables seleccionadas mediante la correlación entre la evaluación holística y las variables independientes. Finalmente, se ajustó el modelo de regresión empleando el método backward,iniciando con un modelo que consideraba todas las variables para luego ir eliminando secuencialmente las que no presentaban significancia estadística (p-value $> 0{,}05$).

4.3. Parámetros empleados para evaluar la calidad de las introducciones

A continuación, se definen, brevemente, las variables que se emplearon en el presente estudio:

- Complejidad sintáctica: Número medio de palabras que anteceden al verbo principal.
- Pronombres personales: Número de pronombres personales empleados por párrafo.
- Cohesión textual: Número de nexos lingüísticos o conectores empleados por párrafo.
- Voz pasiva: Número de casos de voz pasiva por párrafo.
- Vocabulario especializado: frecuencia relativa de palabras propias del discurso académico de Ingeniería Civil Informática.

5. Resultados

5.1. Correlaciones entre rasgos lingüístico-discursivos y calidad de escritura

Las correlaciones efectuadas entre los rasgos lingüístico-discursivos y la evaluación holística se realizaron con el fin de corroborar la existencia de algunos aspectos que se relacionan con la evaluación global que realizan los analistas de las introducciones de los trabajos finales de grado de ingeniería civil en informática. A partir de ellos, se puede afirmar que el rasgo lingüístico-discursivo que más se relaciona con la evaluación holística realizada por los expertos es la complejidad sintáctica ($r = 0{,}779$***)[8]. La correlación con este rasgo resulta un hallazgo importante, puesto que, por un lado, McNamara et al. (2010), Guo et al. (2013) y Crossley et al. (2014) han demostrado que la complejidad suele ser un buen indicador de la calidad del texto cuando se considera como el número medio de palabras que antecede al verbo principal. Pero, por otro lado, investigaciones de MacArthur et al. (2019) consideran que la complejidad posee una correlación negativa en textos escritos en inglés y que no es un buen predictor de la calidad. En nuestra investigación, los resultados respecto a este rasgo nos permiten afirmar que en introducciones escritas en español sí existe una relación positiva entre la calidad del escrito y la complejidad sintáctica del texto.

El número de palabras del texto, a su vez, se correlaciona de manera negativa con el escrito con un valor de $r = -0{,}756$***. Estos resultados se diferencian de investigaciones similares como las de McNamara et al. (2010) y Guo et al. (2013), pues en dichos estudios, los autores establecen que los textos de alta calidad suelen tener mayor cantidad de palabras y, por ende, una mayor extensión. La diferencia con nuestros resultados se explica, debido al género discursivo en cuestión y la disciplina en la que se realiza el estudio, puesto que en la comunidad de ingeniería en informática son poco frecuentes las introducciones extensas y las que lo son, suelen reiterar información y no cumplir con los propósitos comunicativos asociados a los segmentos funcionales del apartado.

Otro rasgo que se vincula con la evaluación holística es el uso de voz pasiva en el escrito ($r = 0{,}613$***). Si bien, este aspecto es más recurrente en inglés que en español, en la disciplina de ingeniería en informática, se suelen escribir las introducciones empleando oraciones impersonales y pasivas en la mayoría de los apartados. Así, solo se utiliza tercera persona singular en la interpretación de los resultados y las conclusiones, por lo que en las introducciones se suelen emplear estos rasgos de forma frecuente. Los resultados a partir de la correlación respecto al empleo de la voz pasiva concuerdan con los de Guo et al. (2013), pues los autores mencionan que los textos de alta calidad suelen emplear la voz pasiva como rasgo prototípico.

[4] http://linguisticanalysistools.org/sinlp.html

[5] http://www.redilegra.com/sistema — herramienta desarrollada por el grupo Redilegra que permite realizar cálculos estadísticos básicos, identificación de N-grams de lemas y POS e identificaciones de modalizadores discursivos, entre otras funciones

[6] https://www.r-project.org/

[7] https://jasp-stats.org/

[8] *** Correlación de pearson, significativa a $p < 0{,}001$

Respecto al uso de pronombres (r = 0,528***), estos se consideran buenos indicadores de la calidad de las introducciones de los TFG de ingeniería civil en informática. Este resultado es relevante pues los pronombres son mecanismos cohesivos que se emplean para mantener los referentes en el escrito y así evitar reiterar las mismas palabras.

Continuando con aspectos ligados a los índices básicos de información textual, el número de oraciones que posee un escrito y la cantidad de párrafos también se correlacionan con la evaluación holística. De esta manera, ambos aspectos poseen una correlación media (oraciones r = 0,475 y párrafos r = 0,470 respectivamente; p = 0,002). Estos resultados concuerdan con los obtenidos por Crossley et al. (2016), pues los autores concluyen de su investigación que la cantidad de oraciones y su longitud es uno de los mejores rasgos para predecir la calidad, ya que permite identificarla hasta en un 60 %.

El vocabulario especializado es otro de los rasgos lingüístico-discursivos que se relaciona de manera media con la evaluación holística de las introducciones (r = 0,468; p = 0,002), lo que indica que, aunque no se considera dentro de los primeros rasgos que se relacionan con la calidad, su empleo junto a otros aspectos podría permitir identificar la calidad del escrito. Nuestros resultados respecto al vocabulario de especialidad concuerdan con los de Guo et al. (2013) y Crossley et al. (2014) debido a que los autores exponen que los textos con mayor calidad, en su caso los ensayos, son aquellos que poseen un mayor número de vocabulario de especialidad y diversidad de palabras.

El siguiente aspecto ligado a la evaluación holística es uno que aún no genera acuerdo entre los especialistas en el tema, pues la cohesión es considerada por algunos investigadores como un rasgo que se liga con la calidad del texto, mientras que para otros no. Nuestros resultados indican que los conectores y mecanismos de cohesión se relacionan con la evaluación holística (r = 0,491***). Si bien, en la presente investigación no se midió la cohesión referencial, por lo que nuestros resultados no pueden ser comparados con los de MacArthur et al. (2019), sí podemos mencionar que nuestros hallazgos se condicen con los de Crossley et al. (2016), pues hemos identificado que existe una relación media entre el uso de mecanismos de cohesión y la calidad del escrito.[9]

[9]Para más detalle de las correlaciones entre los rasgos lingüístico-discursivos y calidad de escritura ver `https://www.dropbox.com/s/oc2qsnc18kdx3bf/correlaciones.docx?dl=0`

5.2. Correlaciones entre calidad de escritura y segmentos funcionales del texto

Este procedimiento se llevó a cabo, porque a partir de los resultados de las relaciones entre los criterios de la pauta de evaluación y la evaluación holística, uno de los aspectos que más se vinculaba con la calidad era el cumplimiento de los propósitos comunicativos. Como forma de operacionalizarlos, se emplearon los trigramas de lemas, pues en investigaciones anteriores (Cotos & Pendar, 2016; Lillo, 2016) han demostrado ser buenas representaciones de las diferentes macromovidas y de las movidas del apartado Introducir al lector en la investigación. Con respecto a la noción de movida, debemos mencionar que en esta investigación la entenderemos como unidad retórica-funcional que realiza un propósito comunicativo de un texto y que, a su vez, se realiza en una unidad textual (Swales, 1990; Askehave & Swales, 2001).

De los resultados, se puede desprender que todos los trigramas de lemas más representativos de las movidas de la introducción se correlacionan con la evaluación holística realizada por los analistas. A su vez, se puede identificar que los trigramas asociados a la movida 2, cuyo propósito es establecer el nicho indicando limitaciones de investigaciones anteriores o mostrando áreas de interés novedosas (Venegas et al., 2016), presentan un valor de r = 0,498***.

Respecto a las relaciones de los trigramas de lema asociados a la movida 3[10] con la evaluación holística, estos poseen un r = 0,358; p = 0,031, lo que relaciona a estos rasgos como significativos con valor de magnitud bajo respecto a la calidad del escrito. Por su parte, los trigramas asociados a la movida 1[11] obtiene valores de correlación de r = 0,338; p = 0,031.

Una posible explicación a la relación existente entre la movida 2 y la evaluación holística es que es uno de los aspectos que menos se presenta, pero que más se valora en las introducciones es la indicación del vacío. Por lo anterior, se podría considerar que en el momento de realizar una evaluación se valora como un aspecto diferenciador de alta y baja calidad el cumplimiento del propósito comunicativo de esta movida (representado, en este caso, por sus trigramas de lemas). Con respecto a los valores de correlación

[10]Ocupar el nicho: movida con el propósito de introducir los aspectos teórico-metodológicos adoptados en la investigación.

[11]Establecer el territorio: movida con el propósito de situar temáticamente la investigación, justificando su relevancia y dando cuenta de investigaciones en el área.

para la movida 1, estos se podrían atribuir a que tanto introducciones de alta como de baja calidad suelen presentar esta movida, pues la Generalización del tópico es el aspecto más prototípico de las introducciones y el que se presentaba en la mayoría de los trabajos que formaron parte de este corpus.

De los resultados expuestos hasta aquí, destaca que las tres movidas de la macromovida Introducir al lector en la investigación se correlacionan con la calidad de las introducciones, por lo que el componente funcional debe ser incluido como un aspecto al momento de evaluar de forma automáticamente la calidad de un escrito. Si bien, sabemos que la representación que hemos realizado de las movidas no dan cuenta totalmente de su presencia o ausencia, consiste en un acercamiento para incluir este aspecto ligado a los segmentos funcionales que ha sido desatendido en investigaciones similares.

5.3. Comparación entre todos los rasgos lingüístico-discursivos en las introducciones de los TFG de alta y baja calidad

Los aspectos lingüístico-discursivos que más se correlacionan entre sí corresponden al uso de pronombres y el número de palabras del texto ($r = 0{,}984^{***}$). A su vez, el número de palabras total del texto también se vincula con el empleo de conectores ($r = 0{,}922^{***}$) y el número de párrafos ($r = 0{,}789^{***}$) de este, por lo que a mayor número de palabras empleadas en el escrito, es decir, extensión, mayor será el uso de estos aspectos. La vinculación de los rasgos anteriormente destacados tiene bastante lógica, pues si el apartado es extenso, en una tesis de alta calidad, se deberán emplear más párrafos, pues se requerirá separar las ideas principales en diferentes párrafos y luego, para unirlas se deberán emplear múltiples mecanismos de cohesión.

Ligado al número de palabras, el número de oraciones se vincula con ella con un $r = 0{,}908^{***}$, lo que significa que en textos de alta calidad, a mayor cantidad de palabras más serán las oraciones y párrafos en el escrito. Lo anterior tiene directa relación con lo planteado por McNamara et al. (2010), pues desde su perspectiva, los textos de alta calidad se caracterizan por tener un gran número de oraciones. A su vez, nuestros resultados también se vinculan a los de Crossley et al. (2016), pues en su investigación, estos autores concluyen que las oraciones se vinculan directamente con la cantidad de párrafos que posee un escrito y que estos dos rasgos unidos son los mejores predictores de la calidad de la escritura.

A su vez, la voz pasiva también se correlaciona con el número de palabras ($r = 0{,}661^{***}$). Estos datos se condicen con lo propuesto por Cubo de Severino (2007), quien expone que, en el discurso académico, entre los procedimientos que se suelen emplear al momento de redactar, predomina la voz pasiva. La complejidad sintáctica se relaciona con la voz pasiva ($r = 0{,}602^{***}$). A partir de esta relación surge otra que se vincula directamente con los resultados obtenidos por el grupo liderado por McNamara, pues a partir de nuestras correlaciones, a mayor número de verbos en las oraciones, mayor será la complejidad sintáctica del escrito, lo que se condice directamente con los hallazgos de Guo et al. (2013) y Crossley et al. (2014).

El vocabulario especializado también se correlaciona con el número de palabras y de oraciones del escrito con valores que varían entre $r = 0{,}905^{***}$ y $r = 0{,}850^{***}$, respectivamente. En cuanto a los conectores, estos de igual manera establecen una correlación positiva de $r = 0{,}799^{***}$ con el vocabulario especializado, por lo que cuando se emplea este último rasgo, las introducciones de alta calidad suelen tener mayor extensión, mayor cantidad de oraciones y mecanismos de cohesión. Respecto al uso de pronombres, estos poseen una correlación de $r = 0{,}881^{***}$ con el empleo de vocabulario especializado. Una posible explicación para estas relaciones podría deberse a que cuando se utiliza vocabulario especializado, resulta complejo sustituir la palabra, pues en muchos casos el término no posee una equivalencia precisa, por lo que se podría emplear pronombres con el fin de mantener el referente a través de una pronominalización.

Con respecto al empleo de pronombres en los textos de alta calidad, estos se correlacionan ($r = 0{,}866^{***}$) con el número de párrafos y la extensión del escrito ($r = 0{,}884^{***}$). A su vez, estas palabras empleadas para sustituir a los nombres se correlacionan ($r = 0{,}882^{***}$) con el empleo de conectores. Respecto a esto último, cabe destacar que una de las formas de mantener la cohesión de un escrito es mediante la pronominalización, por lo que estos dos rasgos estarían directamente vinculados.

Finalmente, en el Cuadro 2 se presentan, de mayor a menor magnitud, los 11 rasgos que se asocian más fuertemente a la calidad de las introducciones de los trabajos finales de grado de ingeniería civil informática.

Orden	Rasgo	Correlación
1	Complejidad sintáctica	0.78***
2	Promedio de palabras	-0.76***
3	Voz pasiva	0.61***
4	Pronombres	0.53***
5	Conclusivos	0.51***
6	Propósito comunicativo	0.50***
7	Conectores	0.48**
8	Ordenadores temporales	0.48**
9	Número de oraciones	0.48**
10	Número de párrafos	0.47**
11	Vocabulario especializado	0.47**

Cuadro 2: Organización de los rasgos lingüístico-discursivos relacionados con la calidad textual en términos de evaluación holística.

5.4. Modelo regresión lineal

Tal como se ha mencionado en apartados anteriores, para el ajuste del modelo se utilizaron las variables seleccionadas (ver Cuadro 2) mediante correlación de Pearson entre los valores promedio de la rúbrica, utilizada para la evaluación de la calidad y las demás variables independientes. Cabe señalar que para la ejecución del modelo de regresión se ha optado por el promedio, pues el valor holístico es binomial lo que no permite cumplir con los supuestos cuantitativos para la posible identificación de una relación lineal. Por otra parte, la correlación existente entre la evaluación holística y el promedio obtenido de los criterios de la rúbrica es de $r = 0{,}95$***. Por lo mismo, la diferencia que se produce en las correlaciones no es muy diferente a las obtenidas con la evaluación holística.

En lo Cuadro 3 se presentan los valores de correlación entre los rasgos y el promedio de la evaluación de la rúbrica.

Orden	Rasgo	Correlación
1	Promedio de palabras	-0.80***
2	Complejidad sintáctica	0.73***
3	Voz pasiva	0.59***
4	Conclusivos	0.53***
5	Propósito comunicativo	0.52***
6	Pronombres	0.47**
7	Ordenadores temporales	0.47**
8	Número de oraciones	0.47**
9	Vocabulario especializado	0.47**
10	Conectores	0.44**
11	Número de párrafos	0.43**

Cuadro 3: Organización de los rasgos lingüístico-discursivos relacionados con la calidad textual en términos del promedio de evaluación con la rúbrica

Como se observa, el orden de los rasgos varía, aunque ello no produce cambios significativos en los valores. Para el análisis de regresión se optó por el método backward, el cual se inicia con un modelo completo para luego ir eliminando secuencialmente las variables que no presenten significancia estadística (p-value $> 0{,}05$). A partir de la tercera regresión se obtuvieron los gráficos post-regresión para determinar normalidad de los residuos, homogeneidad de la varianza, observaciones influyentes y outliers. De acuerdo con este análisis se fueron eliminando una a una las variables contenidas en lo Cuadro 3 que no presentaron significancia estadística. De tal modo, las 3 variables independientes con significancia estadística (p-value $> 0{,}01$) que dan cuenta de un relación lineal con el promedio de calidad son: la complejidad sintáctica, el promedio de palabras y los conectores conclusivos. Cabe destacar que respecto a la bondad de ajuste, el R2 ajustado indica que el modelo representa el 85 % de la variabilidad total del sistema. En el Cuadro 4 se presentan los resultados del último modelo de regresión lineal efectuado y en el Cuadro 5 los valores de validación del Modelo de regresión lineal final.

Variable dependiente	Correlación calidad
Complejidad sintáctica	0.3948416
Promedio de palabras	-0.3673670
Conectores conclusivos	0.3710280

Cuadro 4: Modelo de regresión lineal final

R2	0.868
R2 ajustado	0.849
Akaike	20.595
Error estándar Residual	0.297 (df = 29)
F	47.473*** (df = 4; 29)

Cuadro 5: Valores de validación del Modelo de regresión lineal final

A partir de las tablas anteriores se puede interpretar que si la complejidad sintáctica se incrementa en una unidad de desviación estándar, en promedio, la evaluación de la calidad aumenta en 0,39 unidades de desviación estándar. A su vez, si el promedio de palabras se incrementa en una unidad de desviación estándar la evaluación de la calidad disminuye en 0,37 unidades. Finalmente, si los conclusivos incrementan en una unidad de desviación estándar, en promedio, la calidad aumenta en 0,37 unidades de desviación estándar.

6. Comentarios de cierre

Destacamos en nuestros hallazgos que, a diferencia de investigaciones similares, existe una correlación negativa entre el número de palabras del escrito y la calidad de la escritura, por lo que las introducciones que poseen mayor extensión, tienden a catalogarse de baja calidad. A su vez, identificamos que rasgos como el empleo de conectores conclusivos y la complejidad sintáctica se correlacionan de manera positiva con la calidad de las introducciones de los TFG de ingeniería civil en informática. Este subconjunto de rasgos que representan al plano sintáctico, a la extensión y a la organización argumentativa de la introducción permitirían, de acuerdo con nuestro modelo, predecir la calidad de la introducción en esta disciplina.

Además, como hemos mencionado, en el análisis incluimos los trigramas de lemas para evaluar la relación entre la instanciación de las movidas del apartado y la calidad del escrito, dando como resultado correlaciones positivas y significativas entre los rasgos medidos, aunque no recogidas en el modelo de regresión final.

Debido a lo anterior, como proyecciones de este estudio, se espera poder considerar los aspectos funcionales como parte de los rasgos que permiten evaluar la calidad del escrito e incluir otros índices para medir el cumplimiento de los propósitos comunicativos, identificar otros rasgos e índices de complejidad sintáctica y ampliar el corpus de estudio.

Agradecimientos

Esta investigación fue financiada parcialmente por el proyecto Fondecyt 1190639, titulado "Modelamiento de la práctica discursiva de acreditación del conocimiento por medio de géneros académicos en ingeniería."

Referencias

Askehave, Inger & John M. Swales. 2001. Genre identification and communicative purpose: A problem and a possible solution. *Applied Linguistics* 22(2). 195–212. 10.1093/applin/22.2.195.

Beaugrande, Robert-Alain & Wolfgang Ultich Dressler. 1997. *Introducción a la lingüística del texto*. Grupo Planeta.

Carlino Cantis, Paula. 2003. Alfabetización académica: Un cambio necesario, algunas alternativas posibles. *Educere: Revista Venezolana de Educación* 20. 409–420.

Castelló, Montserrat. 2002. De la investigación sobre el proceso de composición a la enseñanza de la escritura. *Revista Signos: Estudios de Lingüística* 35(51-52). 149–162. 10.4067/S0718-09342002005100011.

Cotos, Elena & Nick Pendar. 2016. Discourse classification into rhetorical functions for AWE feedback. *CALICO Journal* 33(1). 1–22. 10.1558/cj.v33i1.27047.

Crossley, Scott, Kristopher Kyle, Laura Allen, Liang Guo & Danielle McNamara. 2014. Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *The Journal of Writing Assessment* 7(1). 10–16.

Crossley, Scott A., Kristopher Kyle & Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods* 48(4). 1227–1237. 10.3758/s13428-015-0651-7.

Dikli, Semire. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment* 5(1). 4–35.

Figueroa, Javiera, Alejandra Meneses & Eugenio Chandía. 2019. Desempeños en la calidad de explicaciones y argumentaciones en estudiantes chilenos de 8ª básico. *Revista Signos: Estudios de Lingüística* 52(99). 31–54. 10.4067/S0718-09342019000100031.

Guo, Liang, Scott A. Crossley & Danielle S. McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing* 18(3). 218–238. 10.1016/j.asw.2013.05.002.

Jin, Wenjun. 2001. A quantitative study of cohesion in chinese graduate students' writing: variations across genres and proficiency levels. Informe técnico. Northern Illinois University, USA.

Koutsantoni, Dimitra. 2006. Rhetorical strategies in engineering research articles and research theses: Advanced academic literacy and relations of power. *Journal of English for Academic Purposes* 5(1). 19–36. 10.1016/j.jeap.2005.11.002.

Lillo, Fernando. 2016. *Clasificación automática de movidas retóricas en trabajos finales de grado a partir de lemas*: Pontificia Universidad

Católica de Valparaíso, Valparaíso, Chile. Tesis de grado.

MacArthur, Charles A., Amanda Jennings & Zoi A. Philippakos. 2019. Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing* 32. 10.1007/s11145-018-9853-6.

McNamara, Danielle S., Scott A. Crossley & Philip M. McCarthy. 2010. Linguistic features of writing quality. *Written Communication* 27(1). 57–86. 10.1177/0741088309351547.

Meza, Paulina. 2013. *La comunicación del conocimiento en las secciones de tesis de lingüística: Determinación de la variación entre grados académicos*: Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile. Tesis de doctorado.

Moyano, Estela Inés. 2000. *Comunicar ciencia: El artículo científico y las comunicaciones a congresos*. Universidad Nacional de Lomas de Zamora.

Perin, Dolores, Mark Lauterbach, Julia Raufman & Hoori Santikian Kalamkarian. 2017. Text-based writing of low-skilled postsecondary students: relation to comprehension, self-efficacy and teacher judgments. *Reading and Writing* 30. 887–915. 10.1007/s11145-016-9706-0.

Cubo de Severino, Liliana (ed.). 2007. *Los textos de la ciencia*. Cordova: Comunicarte.

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

Sánchez Ceballos, Lina Maria. 2014. La escritura de calidad: base para la transformación de la instrucción en efectiva mediación didáctica. *Revista Reflexión y Saberes* 1(1). 33–37.

Tamola de Spiegel, Diana. 2005. La tesina de licenciatura. En *Los textos de la ciencia. Principales clases del discurso académico-científico*, 235–263. Córdoba, Argentina: Comunicarte.

Vajjala, Sowmya. 2018. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education* 28. 79–105. 10.1007/s40593-017-0142-3.

Venegas, René, Sofía Zamora & Amparo Galdames. 2016. Hacia un modelo retórico discursivo del macrogénero trabajo final de grado en licenciatura. *Revista Signos: Estudios de Lingüística* 49(S1). 247–279. 10.4067/S0718-09342016000400012.

Venegas-Velasquez, Rene. 2014. *Proyecto fondecyt — caracterización de géneros evaluativos como trabajos finales de grado en licenciatura y magíster: Desde los patrones léxico-gramaticales y retórico-estructurales al andamiaje de la escritura académico disciplinar*. ANID: Agencia Nacional de Investigación y Desarrollo.

Zamora, Sofía & René Venegas. 2013. Estructura y propósitos comunicativos en tesis de magíster y licenciatura. *Literatura y Lingüística* 201–218. 10.4067/S0716-58112013000100011.

A. Pauta de evaluación para evaluar manualmente la calidad de los escritos

Descriptor	Nvl 1	Nvl 2	Nvl 3	Nvl 4	Nvl 5	Nvl 6
En la introducción se cumple el propósito de orientar al lector en la temática y presentar el objetivo de investigación.						
En la introducción se destaca la importancia del tema abordado en el Trabajo de título, generalizando y/o mostrando la necesidad de realizar la investigación.						
En la introducción se establece el área específica del Trabajo de título, mostrando las limitaciones de investigaciones previas, aspectos que no se han abordado o áreas de interés novedosas.						
En la introducción, la información se presenta organizada desde la más general a la más específica.						
En la introducción se utilizan diversos conectores discursivos y mecanismos de cohesión para unir la información presente en la introducción del Trabajo de título.						
La información incluida en la introducción se relaciona con el tema abordado en el Trabajo de título.						
El tema se mantiene a lo largo de toda la introducción del trabajo de título.						
La extensión de la introducción es entre una plana y media y dos planas.						
El vocabulario utilizado en la Introducción es acorde a la disciplina y la audiencia del Trabajo de título.						
En la introducción se presenta una correcta ortografía puntual, acentual y literal.						
Las oraciones incluidas en la introducción del Trabajo de título son, en su mayoría, pasivas e impersonales.						
La persona discursiva es consistente a lo largo de la introducción del Trabajo de título.						
Las oraciones de la introducción están construidas con una sintaxis acorde al español y presentan concordancia entre sujeto y predicado en número y persona.						
La introducción del Trabajo de título no presenta vicios idiomáticos tales como: redundancia, discordancia, dequeísmo, queísmo, marcas de oralidad, entre otros.						
La introducción es de alta calidad, pues cumple con el propósito de introducir al lector en la investigación. Además, posee buena redacción, cohesión, coherencia, las ideas son claras, el vocabulario es acorde a la audiencia y al género discursivo. No presenta errores ortográficos (acentual, literal y literal).						

Propuesta recibida en octubre 2019 y aceptada para publicación en mayo 2020.

Generación automática de frases literarias

Automatic Generation of Literary Sentences

Luis-Gil Moreno-Jiménez
Université d'Avignon/LIA
Universidad Tecnológica de la Selva
luis-gil.moreno-jimenez@alumni.univ-avignon.fr

Juan-Manuel Torres-Moreno
Université d'Avignon/LIA
Polytechnique Montréal
juan-manuel.torres@univ-avignon.fr

Roseli S. Wedemann
Universidade do Estado do Rio de Janeiro
roseli@ime.uerj.br

Eric SanJuan
Université d'Avignon/LIA
eric.sanjuan@univ-avignon.fr

Resumen

En este artículo abordamos el tema de la generación automática de frases literarias, que es una parte importante de los estudios relacionados al área de la Creatividad Computacional (CC). Proponemos tres modelos de generación textual guiados por un contexto, basados principalmente en algoritmos estadísticos y análisis sintáctico superficial. Los textos generados fueron evaluados por siete personas a partir de 4 criterios: gramaticalidad, coherencia, relación con el contexto y una adaptación del test de Turing, en donde se pidio a los evaluadores clasificar los textos en: textos generados automáticamente y textos generados por humanos. Los resultados obtenidos son bastante alentadores.

Palabras clave

corpora literarios, generación automática de frases, cadenas de Markov, Word2vec

Abstract

In this article, we regard the task of automatic generation of literary sentences, which is an important topic in the area of Computational Creativity. We propose three generative models mainly based on statistical algorithms and shallow parsing. The generated texts were evaluated by seven persons according to four criteria: grammar, coherence, context related, and an adaptation of the Turing test. We present preliminary results of their implementations that are quite encouraging.

Keywords

literary corpora, automatic sentence generation, Markov chains, Word2Vec

1. Introducción

Los investigadores en Procesamiento de Lenguaje Natural (PLN) durante mucho tiempo han utilizado diversos corpora constituidos por documentos enciclopédicos (principalmente Wikipedia), periodísticos (periódicos o revistas) o especializados (documentos legales, científicos o técnicos) para el desarrollo y pruebas de sus modelos (Torres-Moreno, 2014; Iria et al., 2011; Martínez, 2018).

El uso y análisis de los corpora literarios sistemáticamente han sido dejados a un lado por varias razones. En primer lugar, el nivel de discurso literario es más complejo que los otros géneros. En segundo lugar, a menudo, los documentos literarios hacen referencia a mundos o situaciones imaginarias o alegóricas, a diferencia de los otros géneros que describen sobre todo situaciones o hechos factuales. Estas y otras características presentes en los textos literarios, vuelven sumamente compleja la tarea de análisis automático de este tipo de textos. En este trabajo nos proponemos utilizar corpora literarios, a fin de generar realizaciones literarias (frases nuevas) no presentes en dichos corpora.

La producción de textos literarios es el resultado de un proceso donde una persona hace uso de aptitudes creativas. Este proceso, denominado "proceso creativo", ha sido analizado por Boden (2004), quien propone tres tipos básicos de creatividad: la primera, Creatividad Combinatoria (CCO), donde se fusionan elementos conocidos para la generación de nuevos elementos. La segunda, Creatividad Exploratoria (CE), donde la generación ocurre a partir de la observación o exploración. La tercera, Creatividad Transformacional (CT), donde los elementos generados son producto de alteraciones o experimentaciones aplicadas al dominio de la CE.

DOI: 10.21814/lm.12.1.308

Sin embargo, cuando se pretende automatizar el proceso creativo, la tarea debe ser adaptada a métodos formales que puedan ser realizados en un algoritmo. Este proceso automatizado da lugar a un nuevo campo de investigación, denominado Creatividad Computacional (CC) (Pérez y Pérez, 2015), en donde se retoman los conceptos: CT y la CE propuestos por Boden (2004). Es en este campo donde nosotros hemos trabajado para la generación de frases literarias.

Por otro lado, la definición de literatura no tiene un consenso universal, y muchas variantes de la definición pueden ser encontradas. En este trabajo optaremos por introducir una definición pragmática de frase literaria, que servirá para nuestros modelos y experimentos.

Definición 1 *Una frase literaria es una frase que se diferencia de las frases en lengua general, porque contiene elementos (nombres, verbos, adjetivos, adverbios) que son percibidos como elegantes o menos coloquiales que sus equivalentes en lengua general.*

Por ejemplo, la frase en lengua general:

- *Me paré a ver unos libros viejos en la librería que está en la esquina de mi casa.*

Puede ser ligeramente re-escrita para generar tres frases literarias según nuestra definición:

- *Me detuve a mirar libros antiguos en la librería próxima a mi casa.*
- *Miré durante unos momentos algunos libros antiguos en la librería cercana a mi casa.*
- *Hojeé durante algunos instantes libros viejos en la librería cercana a mi hogar.*

Por supuesto, un autor puede decidir escribir un texto literario basado exclusivamente en frases de lengua general. Por ejemplo, José Agustín en "De perfil" donde el fragmento: *"...me quedé dormido en el Jardín. Supongo que el sol y lo fresco del aire crearon el término exacto para adormecerme."*[1], usa frases literarias dentro de un texto desbordante de lengua general. Sin embargo, nosotros no intentaremos mezclar ambas lenguas y nos restringiremos a analizar y generar frases literarias.

En particular, proponemos crear artificialmente frases literarias, utilizando modelos generativos y aproximaciones semánticas basados en corpora de lengua literaria. La combinación de esos modelos da lugar a una homosintaxis, es decir, la producción de texto nuevo a partir de formas de discurso de diversos autores. La homosintaxis no tiene el mismo contenido semántico, ni siquiera las mismas palabras, aunque guarda la misma estructura sintáctica.

En este trabajo proponemos estudiar el problema de la generación de texto literario original en forma de frases aisladas, no a nivel de párrafos. La generación de párrafos puede ser objeto de trabajos futuros. Una evaluación de la calidad de las frases generadas por nuestro sistema será presentada.

Este artículo está estructurado como sigue. En la Sección 2 presentamos un estado del arte de la generación automática de textos. En la Sección 3 describimos los corpora utilizados. Nuestros modelos son descritos en la Sección 4. Los resultados y su interpretación se encuentran en la Sección 5. Finalmente, la Sección 6 presenta algunas ideas de trabajos futuros antes de concluir.

2. Estado del arte

A continuación, se presenta un estudio del estado del arte en donde se mencionan trabajos para la generaciòn de texto con enfoques variados y objetivos bastante interesantes. En principio, mostramos algunos trabajos que no estàn relacionados a la CC literaria. Posteriormente, analizamos investigaciones dedicadas a la generaciòn textual dentro del marco de la CC, como generaciòn de poemas, poesìas y otras formas literarias.

Durante nuestra investigaciòn, hemos percibido que la CC no busca solucionar los problemas de la sociedad en sus variados aspectos, sino encontrar nuevos paradigmas para la creación de obras con un importante valor cultural y artístico (Colton & Wiggins, 2012). Sin embargo, una gran variedad de modelos de IA han sido adaptados e incluso mejorados para lograr simular el proceso creativo a través de modelos computacionales (Colton, 2012).

2.1. Generación textual no literaria

La generación de texto es una tarea relativamente clásica, que ha sido estudiada en diversos trabajos. Por ejemplo, Szymanski & Ciota (2002) presentam un modelo basado en cadenas de Markov para la generación de texto en idioma polaco. Los autores definen un conjunto de estados actuales y calculan la probabilidad de pasar al estado siguiente. La ecuación (1) calcula la probabilidad de pasar al estado X_i a partir de X_j,

$$P_{ij}(X_i|X_j) = P(X_i \cap X_j)|P(X_j)\,. \qquad (1)$$

[1] J. Agustín. *De perfil*, Joaquín Mortiz, México, 1993.

Para ello, se utiliza una matriz de transición, la cual contiene las probabilidades de transición de un estado actual X_i a los posibles estados futuros X_{i+1}. Cada estado puede estar definido por n-gramas de letras o de palabras.

La tarea inicia en un estado X_i dado por el usuario. Posteriormente, usando la matriz de transición, se calcula la probabilidad de pasar al estado siguiente X_{i+1}. En ese momento el estado predicho X_{i+1} se convierte en el estado actual X_i, repitiendo este proceso hasta satisfacer una condición. Este método tiene un buen comportamiento al generar palabras de 4 o 5 letras. En polaco esta longitud corresponde a la longitud media de la mayor parte de las palabras (Torres-Moreno, 2012).

También hay trabajos que realizan análisis más profundos para generar no solamente palabras, sino párrafos completos. Sridhara et al. (2010) presentan un algoritmo que genera automáticamente comentarios descriptivos para bloques de código (métodos) en Java. Para ello, se toma el nombre del método y se usa como la acción o idea central de la descripción a generar. Posteriormente se usan un conjunto de heurísticas, para seleccionar las líneas de código del método que puedan aportar mayor información, y se procesan para generar la descripción.

La tarea consiste en construir sintagmas, a partir de la idea central dada por el nombre del método, y enriquecerlos con la información de los elementos extraídos. Por ejemplo, si hay un método `removeWall(Wall x)` y se encuentra la llamada al método `removeWall(oldWall)`, la descripción generada podría ser: "Remove old Wall". Obteniéndose la acción (verbo) y el objeto (sustantivo) directamente del nombre del método y el adjetivo a partir de la llamada. Estas ideas permiten a los autores la generación de comentarios extensos sin perder la coherencia y la gramaticalidad.

También existen trabajos con un alcance más limitado pero de mayor precisión. Huang et al. (2012) proponen la evaluación de un conjunto de datos con un modelo basado en redes neuronales para la generación de subconjuntos de multipalabras. Este mismo análisis se considera por Fu et al. (2014), en donde se busca establecer o detectar la relación hiperónimo-hipónimo con la ayuda del modelo Word2vec, también basado en redes neuronales (Mikolov et al., 2013b). Esta propuesta reporta una precisión de 0.70, al ser evaluado sobre un corpus manualmente etiquetado.

2.2. Generación de poesía

También se encuentran trabajos de generación textual que se proponen como meta resultados con un valor literario. La generación de texto literario es un proceso distinto a la generación de texto general (Lebret et al., 2016; Welleck et al., 2019), y ha sido abordado desde los años 60's por investigadores del campo de humanidades, siendo hasta principios del año 2000 abordada fuertemente por la Ciencias Computacionales (Gonçalo Oliveira, 2017), Entre estos, se tienen trabajos para la generación de poesía o poemas.

Zhang & Lapata (2014) proponen un modelo para la generación de poemas que se basa en dos premisas básicas: *¿qué decir?* y *¿cómo decirlo?* La propuesta parte de la selección de un conjunto de frases, tomando como guía una lista de palabras dadas por el usuario. Las frases son procesadas por un modelo de red neuronal (Mikolov & Zweig, 2012), para construir combinaciones coherentes y formular un contexto. Este contexto es analizado para identificar sus principales elementos y generar las líneas del poema, que también pasarán a formar parte del contexto. El modelo fue evaluado manualmente por 30 expertos en una escala de 1 a 5, analizando legibilidad, coherencia y significatividad en frases de 5 palabras, obteniendo una precisión de 0.75. Sin embargo, la coherencia entre frases resultó ser muy pobre.

Gonçalo Oliveira (2012); Gonçalo Oliveira & Cardoso (2015) proponen un modelo de generación de poemas basado en el uso de plantillas. El algoritmo inicia con un conjunto de frases relacionadas a partir de palabras clave. Las palabras clave sirven para generar un contexto. Las frases son procesadas usando el sistema PEN[2] para obtener su información gramatical. Esta información es empleada para la generación de nuevas plantillas gramaticales y finalmente la construcción de las líneas del poema, tratando de mantener la coherencia y la gramaticalidad.

Otros trabajos han sido propuestos para la generación de poesía, como en (Agirrezabal et al., 2013), donde se presenta un método bastante interesante, en donde a partir del análisis de diversos corpora, se extraen las secuencias de etiquetas POS con sus respectivas inflexiones para calcular la probabilidad de aparición de cada una de ellas. Este método estocástico sirve para la generación de nuevas secuencias y posteriormente se procede a la sustitución de las etiquetas POS. Se realizaron tres experimentos para la sustitución. En el primero se sustituyen todas las etiquetas de las

[2]Disponible en: `http://code.google.com/p/pen`

secuencias POS por palabras que respeten la gramaticalidad de la etiqueta. En el segundo se sustituyen únicamente adjetivos y sustantivos bajo la misma condición, y finalmente en el tercer experimento sólo se reemplazan sustantivos con palabras con una relación semántica determinada.

2.3. Generación de narrativas

La literatura es una actividad artística que exige capacidades creativas importantes y que ha llamado la atención de científicos desde hace cierto tiempo. Diversos investigadores han trabajado en proyectos que permiten la generación de texto literario cruzando la frontera de textos cortos como poemas o poesía, para dar lugar a la generación de textos más extensos.

Riedl & Young (2006) presentan un conjunto de algoritmos para la generación de una guía narrativa basada en la idea de Creatividad Exploratoria (Boden, 2004). El modelo establece *i)* un conjunto universal U de conceptos relevantes relacionados a un dominio; *ii)* un modelo generador de texto; *iii)* un subconjunto de conceptos S que pertenecen al conjunto universal U; y *iv)* algoritmos encargados de establecer las relaciones entre U y S para generar nuevos conceptos. Estos nuevos conceptos serán posteriormente comparados con los conceptos ya existentes en U, para verificar la coherencia y relación con la idea principal. Si los resultados son adecuados, estos nuevos conceptos se utilizan para dar continuación a la narrativa.

Son diversos los trabajos que están orientados a la generación de una narrativa ficticia, como cuentos o historias. Clark et al. (2018) proponen un modelo de generación de texto narrativo a partir del análisis de *entidades*. Dichas *entidades* son verbos, sustantivos o adjetivos dentro de un texto, que serán usados para generar la frase siguiente. El modelo recupera las *entidades* obtenidas de tres fuentes principales: la frase actual, la frase previa y el documento completo (contexto), y las procesa con una red neuronal para seleccionar las mejores de acuerdo a diversos criterios. A partir de un conjunto de heurísticas, se analizaron las frases generadas para separar aquellas que expresaran una misma idea (paráfrasis), de aquellas que tuvieran una relación entre sus *entidades* pero con ideas diferentes.

El modelo sentiGAN (Ke & Xiaojun, 2018) pretende generar texto con un contexto emocional. Se trata de una actualización del modelo GAN (*Generative Adversarial Net*) (Goodfellow et al., 2014) que ha producido resultados alentadores en la generación textual, aunque con ciertos problemas de calidad y coherencia. Se utiliza el análisis semántico de una entrada proporcionada por el usuario que sirve para la creación del contexto. La propuesta principal de SentiGAN sugiere establecer un número definido de generadores textuales que deberán producir texto relacionado a una emoción definida. Los generadores son entrenados bajo dos esquemas: *i)* una serie de elementos lingüísticos que deben ser evitados para la generación del texto; y *ii)* un conjunto de elementos relacionados con la emoción ligada al generador. A través de cálculos de distancia, heurísticas y modelos probabilísticos, el generador crea un texto lo más alejado del primer esquema y lo más cercano al segundo.

Pérez y Pérez (2015) presentam una revisión interesante del estado del arte en este tema, donde se mencionan algunos de los primeros intentos de generación automática de textos literarios. Por ejemplo, el modelo "Through the park" (Montfort, 2008b), es capaz de generar narraciones históricas empleando la elipsis. Esta técnica es empleada para manipular, entre otras cosas, el ritmo de la narración. En los trabajos "About So Many Things" (Montfort, 2008c) y "Taroko Gorge" (Montfort, 2009) se muestran textos generados automáticamente. El primero de ellos genera estrofas de 4 líneas estrechamente relacionadas entre ellas. Eso se logra a través de un análisis gramatical que establece conexiones entre entidades de distintas líneas. El segundo trabajo muestra algunos poemas cortos generados automáticamente, con una estructura más compleja que la de las estrofas. El inconveniente de ambos enfoques es el uso de una estructura inflexible, lo que genera textos repetitivos con una gramaticalidad limitada.

El proyecto MEXICA modela la generación colaborativa de narraciones Pérez y Pérez (2015). El propósito es la generación de narraciones completas utilizando obras de la época Precolombina. MEXICA genera narraciones simulando el proceso creativo de E-R *(Engaged and Reflexive)* (Sharples, 1996). Este proceso se describe como la acción, donde el autor trae a su mente un conjunto de ideas y contextos y establece una conexión coherente entre estas (E). Posteriormente se reflexiona sobre las conexiones establecidas y se evalúa el resultado final para considerar si este realmente satisface lo esperado (R). El proceso itera hasta que el autor lo considera concluido.

Otro trabajo que contempla la generacion de narrativas es el que se presenta en (Gervás et al., 2015), en donde se expone el método bajo el cual, algunos escritores reutilizan las experiencias recabadas en textos leídos o escritos, estas son apli-

cadas para la creación de nuevos textos. En este trabajo, estas experiencias son traducidas como escenarios, estructuras, acciones, etc., que sirven como datos de entrenamiento para la generación de nuevas narrativas.

Nuestro modelo de generación de frases no fue concebido para la generación de poesía o narrativa. Mas bien está dentro de un cuadro general de generación automática, teniendo como objetivo la construcción de un generador artificial de texto literario. Desde este punto de vista es sólo un módulo de un sistema mas complejo en perspectiva, que contempla la semántica, el manejo de figuras literarias y las emociones.

3. Corpora utilizados

En esta sección describimos los corpora utilizados en nuestros modelos para los experimentos. Se trata del corpus 5KL y del corpus 8KF, ambos creados en idioma español.

3.1. Corpus 5KL

Este corpus fue constituido con aproximadamente 5 000 documentos (en su mayor parte libros) en español. Los textos, en su mayoria, correspondem a los géneros literarios: narrativa, poesía, teatro, ensayos, etc[3]. Los documentos originales, en formatos muy heterogéneos[4], fueron procesados para crear un único documento codificado en *UTF-8*. Dada su heterogeneidad, este corpus presenta una gran cantidad de errores (palabras cortadas o pegadas, símbolos extraños y disposición no convencional de párrafos).

Las herramientas clásicas como FreeLing (Padró, 2012) tienen mucha dificultad en tratar estes tipos de documentos. Por ello, decidimos construir un segmentador de frases ad hoc para este tipo de corpus ruidoso. Las frases fueron segmentadas automáticamente, usando un programa en Perl y expresiones regulares, para obtener una frase por línea.

Las características del corpus 5KL se encuentran en el Cuadro 1[5]. Este corpus es empleado para el entrenamiento del modelo Word2vec (ver Sección 4).

El corpus literario 5KL posee la ventaja de ser muy extenso y adecuado para el aprendizaje automático. Tiene sin embargo, la desventaja de que no todas las frases son *necesariamente* "frases literarias". Muchas de ellas son frases de lengua general: estas frases a menudo otorgan una fluidez a la lectura y proporcionan los enlaces necesarios a las ideas expresadas en las frases literarias.

	Frases	**Tokens**	**Caracteres**
5KL	9 M	149 M	893 M
Media por documento	2.4 K	37.3 K	223 K

Cuadro 1: Corpus 5KL compuesto de 4 839 obras literarias.

Otra desventaja de este corpus es el ruido que contiene. Por lo que, el proceso de segmentación puede producir errores en la detección de fronteras de frases. También los números de página, capítulos, secciones o índices producen errores. No se realizó ningún proceso manual de verificación, por lo que a veces se introducen informaciones indeseables: *copyrights*, datos de la edición u otros. Estas son, sin embargo, las condiciones que presenta un corpus literario real.

3.2. Corpus 8KF

Decidimos crear un pequeño corpus controlado, exclusivamente compuesto de "frases literarias", que será utilizado en la fase generativa de los modelos propuestos. Un corpus heterogéneo de casi 8 000 frases literarias fue constituido manualmente, a partir de poemas, discursos, citas, cuentos y otras obras.

Se evitaron cuidadosamente las frases de lengua general, y también aquellas demasiado cortas ($N \leq 3$ palabras) o demasiado largas ($N \geq 30$ palabras). Algunos elementos que sirvieron para seleccionar manualmente las frases "literarias" fueron: un vocabulario complejo y estético, el cual rara vez es empleado en el lenguaje común, además de la identificación de ciertas figuras literarias como la rima, la anáfora, la metáfora y otras. Algunos ejemplos de frases literarias son los siguientes:

- *La mentira y la verdad no pueden vivir en paz.*
- *El amor, como la tos, no puede ocultarse.*
- *Si tu belleza fuera enfermedad, vida mía, no habría remedio.*
- *Grabad esto en vuestro corazón: cada día es el mejor del año.*

Las características del corpus 8KF se muestran en el Cuadro 2. Este corpus fue utiliza-

[3]Dada la dimensión de este corpus, no nos fue posible cuantificar los géneros manualmente. Una aproximación automática podrá realizarse a futuro.

[4]pdf, txt, html, doc, docx, odt, etc.

[5]M representa un valor de 10^6 y K de 10^3.

do principalmente en los dos modelos generativos: modelo basado en cadenas de Markov (Sección 4.1.1) y modelo basado en la generación de Texto enlatado (*Canned Text*, Sección 4.1.2).

	Frases	Tokens	Caracteres
8KF	7 679	114 K	652 K
Media por frase	—	15	85

Cuadro 2: Corpus 8KF compuesto de 7 679 frases literarias.

4. Modelos propuestos

En este trabajo proponemos tres modelos híbridos (combinaciones de modelos generativos clásicos y aproximaciones semánticas) para la producción de frases literarias. Hemos adaptado dos modelos generativos, usando análisis sintáctico superficial (*shallow parsing*), combinados con tres modelos de aproximación semántica usando *Word2vec*.

En una primera fase, los modelos generativos recuperan la información gramatical de cada palabra del corpus 8KF (ver Sección 3), en forma de etiquetas POS (*Part of Speech*), a través de un análisis morfosintáctico. Utilizamos FreeLing (Padró, 2012) que permite análisis lingüísticos en varios idiomas[6] y que además de devolvernos las etiquetas POS, que nos permiten saber si la palabra analizada es un verbo, sustantivo, adjetivo, etc., también nos da información acerca de las inflexiones en ella, es decir, conjugaciones, género, número, etc. Por ejemplo, para la palabra "Profesor" FreeLing genera la etiqueta POS `[NCMS000]`. La primera letra indica un sustantivo (**N**oun), la segunda un sustantivo común (**C**ommon); la tercera indica el género masculino (**M**ale) y la cuarta da información de número (**S**ingular). Los 3 últimos caracteres dan información detallada del campo semántico, entidades nombradas, etc.[7] En nuestro caso, usaremos solamente los 4 primeros niveles de las etiquetas.

Con los resultados del análisis morfosintáctico, se genera una salida que llamaremos *Estructura gramatical vacía* (EGV), compuesta exclusivamente de una secuencia de etiquetas POS, o una *Estructura gramatical parcialmente vacía* (EGP), compuesta de etiquetas POS y de palabras funcionales (artículos, pronombres, conjunciones, etc.).

En la segunda fase, las etiquetas POS (en la EGV y la EGP) serán reemplazadas por un vocabulario adecuado usando ciertas aproximaciones semánticas. La producción de una frase $f(Q, N)$ es guiada por dos parámetros: un contexto representado por un término Q (*query*) y una longitud $3 \leq N \leq 15$, dados por el usuario. Los corpora 5KL y 8KF son utilizados en varias fases de la producción de las frases f.

- El Modelo 1 está compuesto por: *i)* un modelo generativo estocástico basado en cadenas de Markov, para la selección de la próxima etiqueta POS usando el algoritmo de Viterbi; y *ii)* un modelo Word2vec, para recuperar el vocabulario que reemplazará la secuencia de etiquetas POS.
- El Modelo 2 es una combinación de: *i)* el modelo generativo de *texto enlatado*; y *ii)* un modelo Word2vec, con un cálculo de distancias entre diversos vocabularios que han sido constituidos manualmente.
- El Modelo 3 utiliza: *i)* la generación de *texto enlatado*; y *ii)* una interpretación geométrica, utilizando redes neuronales con Word2vec. Esta interpretación está basada en una búsqueda de información iterativa (*Information Retrieval*, IR), que realiza simultáneamente un alejamiento de la semántica original y un acercamiento al *query* Q del usuario.

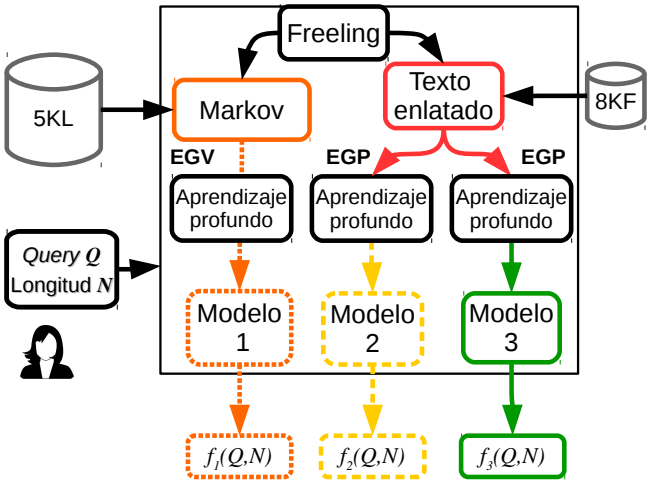


Figura 1: Arquitectura general de los modelos.

La Figura 1 muestra la arquitectura general de nuestro sistema. En la Sección 4.1 se describen los dos modelos generativos, y enseguida los tres modelos de aproximación semántica.

[6] FreeLing ha sido desarrollado en el centro TALP (Universidad Politécnica de Cataluña). Puede ser obtenido en la dirección: `http://nlp.lsi.upc.edu/freeling`

[7] Más detalles de las etiquetas FreeLing en `http://blade10.cs.upc.edu/freeling-old/doc/tagsets/tagset-es.html`

4.1. Modelos generativos

A continuación, se presentan dos modelos generativos de estructuras gramaticales en sus dos variantes, Estructuras Gramaticales Vacías (EGV) y Estructuras Gramaticiales Parcialmente Vacías (EGP), que sirven a los modelos descritos en las secciones 4.2, 4.3 y 4.4 para la generación de frases.

4.1.1. Modelo generativo estocástico usando cadenas de Markov

Este modelo generativo, que llamaremos *Modelo de Markov*, está basado en el algoritmo de Viterbi y las cadenas de Markov (Manning & Schütze, 1999), donde se selecciona una etiqueta POS con la máxima probabilidad de ocurrencia, para ser agregada al final de la secuencia actual.

Utilizamos el corpus de frases literarias 8KF (ver Sección 3.2), que fue convenientemente filtrado para eliminar *tokens* indeseables: números, siglas, horas y fechas. El corpus filtrado se analizó usando FreeLing, que recibe en entrada una cadena de texto y entrega el texto con una etiqueta POS para cada palabra. El corpus es analizado frase a frase, reemplazando cada palabra por su respectiva etiqueta POS. Al final del análisis, se obtiene un nuevo corpus 8KPOS con $s = 7\,679$ secuencias de etiquetas POS, correspondientes al mismo número de frases del corpus 8KF. Las secuencias del corpus 8KPOS sirven como conjunto de entrenamiento para el algoritmo de Viterbi, que calcula las probabilidades de transición, que serán usadas para generar cadenas de Markov.

Las s estructuras del corpus 8KPOS procesadas con el algoritmo de Viterbi son representadas en una matriz de transición $P_{[s \times s]}$. P será utilizada para crear nuevas secuencias de etiquetas POS no existentes en el corpus 8KPOS, simulando un proceso creativo. Nosotros hemos propuesto el algoritmo *Creativo-Markov* que describe este procedimiento.

En este algoritmo, X_i representa el estado de una etapa de la creación de una frase, en el instante i, que corresponde a una secuencia de etiquetas POS. Siguiendo un procedimiento de Markov, en un instante i se selecciona la próxima etiqueta POS_{i+1}, con máxima probabilidad de ocurrencia, dada la última etiqueta POS_i de la secuencia X_i. La etiqueta POS_{i+1} será agregada al final de X_i para generar el estado X_{i+1}. $P(X_{i+1} = Y | X_i = Z)$ es la probabilidad de transición de un estado a otro, obtenido con el algoritmo de Viterbi. Se repiten las transiciones, hasta alcanzar una longitud deseada.

El resultado es una EGV, donde cada cuadro vacío representa una etiqueta POS que será remplazada por una palabra, en la etapa final de generación de la nueva frase. El remplazo se realiza usando el modelo descrito en la Sección 4.2. La arquitectura general de este modelo se muestra en la Figura 2.

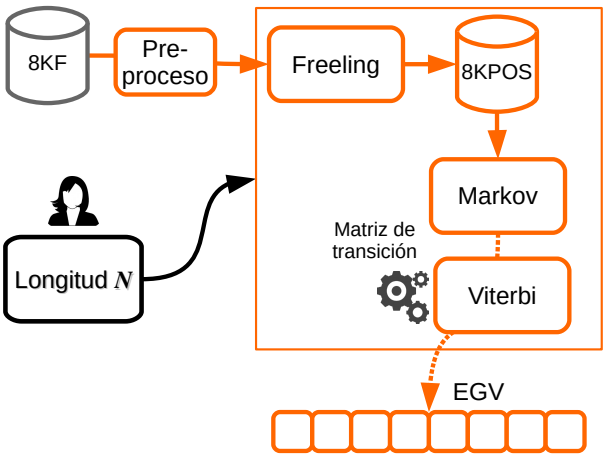


Figura 2: Modelo generativo estocástico (Markov) que produce una estructura gramatical vacía EGV.

4.1.2. Modelo generativo basado en texto enlatado

El algoritmo *Creativo-Markov* del *Modelo de Markov* logra reproducir patrones lingüísticos (secuencias POS) detectados en el corpus 8KPOS, pero de corta longitud. Cuando se intentó extender la longitud de las frases a $N > 6$ palabras, no fue posible mantener la coherencia y legibilidad (como se verá en la Sección 4.2). Decidimos entonces utilizar métodos de generación textual guiados por estructuras morfosintácticas fijas: el *texto enlatado*. Molins & Lapalme (2015) argumentan que el uso de estas estructuras ahorran tiempo de análisis sintáctico y permite concentrarse directamente en el vocabulario.

La técnica de *texto enlatado* ha sido empleada también en varios trabajos, con objetivos específicos. McRoy et al. (2003); van Deemter et al. (2005) desarrollaron modelos para la generación de diálogos y frases simples. Esta técnica es llamada "Generación basada en plantillas" (*Template-based Generation*) o de manera intuitiva, *texto enlatado*[8].

Decidimos emplear *texto enlatado* para la generación textual, usando un corpus de plantillas (*templates*) construido a partir del corpus 8KF (Sección 3). Este corpus contiene estructuras gra-

[8] http://projects.ict.usc.edu/nld/cs599s13/LectureNotes/cs599s13dialogue2-13-13.pdf

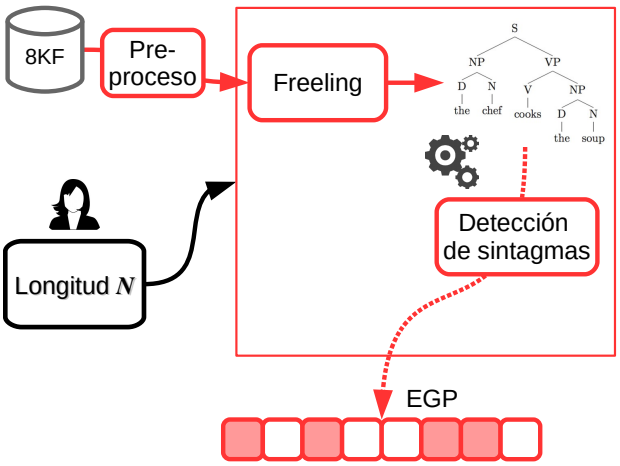


Figura 3: Modelo generativo de Texto enlatado que produce una estructura parcialmente vacía.

maticales flexibles que pueden ser manipuladas para crear nuevas frases. Estas plantillas pueden ser seleccionadas aleatoriamente o a través de heurísticas, según un objetivo predefinido.

Una plantilla es construida a partir de las palabras de una frase f, donde se reemplazan únicamente las palabras llenas de las clases verbo, sustantivo o adjetivo $\{V, S, A\}$, por sus respectivas etiquetas POS. Las otras palabras, en particular las palabras funcionales, son conservadas. Esto producirá una *estructura gramatical parcialmente vacía, EGP*. Posteriormente las etiquetas podrán ser reemplazadas por palabras (términos), relacionadas con el contexto definido por el *query* Q del usuario.

El proceso inicia con la selección aleatoria de una frase original $f_o \in$ corpus 8KF de longitud $|f_o| = N$. f_o será analizada con FreeLing para identificar los sintagmas. Los elementos $\{V, S, A\}$ de los sintagmas de f_o serán reemplazados por sus respectivas etiquetas POS. Estos elementos son los que mayor información aportan en cualquier texto, independientemente de su longitud o género (Bracewell et al., 2005). Nuestra hipótesis es que al cambiar solamente estos elementos, simulamos la generación de frases por homosintaxis: semántica diferente, misma estructura[9].

La salida de este proceso es una estructura híbrida parcialmente vacía (EGP), con palabras funcionales que dan un soporte gramatical y las etiquetas POS. La arquitectura general de este modelo se ilustra en la Figura 3. Los cuadros llenos representan palabras funcionales y los cuadros vacíos etiquetas POS a ser reemplazadas.

[9] Al contrario de la paráfrasis que busca conservar completamente la semántica, alterando completamente la estructura sintáctica.

4.2. Modelo 1: Markov y Word2vec

En este modelo se retoman las estructuras gramaticales vacías (EGV), descritas en la Sección 4.1.1, que pueden ser manipuladas para generar nuevas frases $f(Q, N)$. La idea es que las frases f sean generadas por homosintaxis. En esta sección, proponemos un modelo que combina el modelo generativo de Markov (Sección 4.1.1), con un algoritmo de aproximación semántica que utiliza un modelo de redes neuronales, el Word2vec. La labor de Word2vec es obtener la representación de una palabra en un espacio vectorial (*embeddings*), a través de un análisis contextual[10]. El proceso se describe a continuación.

El corpus 5KL es pre-procesado para uniformizar el formato del texto, eliminando caracteres que no son importantes para el análisis semántico (puntuación, números, etc.). Esta etapa prepara los datos de entrenamiento del Word2vec que utiliza una representación vectorial del corpus 5KL. Para este, utilizamos la biblioteca Gensim[11], una implementación en Python de Word2vec[12]. Con este algoritmo, se obtiene un conjunto de palabras, o *embeddings*, asociadas a un contexto definido por un *query* Q. Word2vec recibe un término Q y devuelve un léxico $L(Q) = (w_1, w_2, \ldots, w_m)$, que representa un conjunto de $m = 10$ palabras semánticamente próximas a Q. El valor de m fue definido de esta manera ya que se se percibió que, mientras más se extiende el número de palabras proximas a Q, estas pierden más su relación con respecto a Q. Formalmente, representamos Word2vec: $Q \rightarrow L(Q)$.

Para el entrenamiento de Word2vec se consideran palabras con más de 5 ocurrencias en el corpus. La ventana contextual definida tiene una dimensión de 10. Para las dimensiones de las representaciones vectoriales se hicieron pruebas dentro de un rango de 50 a 100, siendo 60 la dimensión con la que se obtuvieron embeddings mejores relacionados. El modelo de entrenamiento fue *continuous skip-gram model (Skip-gram)*, el cual funciona mejor con copora de tamaños significativos (Mikolov et al., 2013a).

El próximo paso consiste en procesar la EGV producida por Markov. Las etiquetas POS serán identificadas y clasificadas como POS_Φ funcionales (correspondientes a puntuación y palabras funcionales) y POS_λ llenas $\in \{V, S, A\}$ (Verbos, Sustantivos, Adjetivos).

[10] Word2vec pertenece a un amplio campo de investigación dentro de PLN, conocido como *Representation Learning* (Bengio et al., 2013).

[11] Disponible en: https://pypi.org/project/gensim/

[12] https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

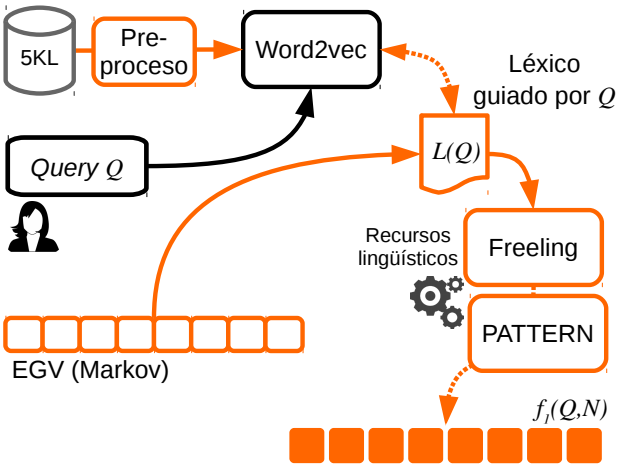


Figura 4: Modelo 1 de aproximación semántica, usando Markov y Word2vec.

Las etiquetas POS_Φ serán reemplazadas por palabras obtenidas de recursos lingüísticos (diccionarios) construídos con la ayuda de FreeLing. Los diccionarios consisten en entradas de pares: POS_Φ y una lista de palabras y signos asociados, formalmente $\text{POS}_\Phi \rightarrow l(\text{POS}_\Phi) = (l_1, l_2, \ldots, l_j)$. Se reemplaza aleatoriamente cada POS_Φ por una palabra de l que corresponda a la misma clase gramatical.

Las etiquetas POS_λ serán reemplazadas por las palabras, $L(Q)$, producidas por Word2vec. Si ninguna de las palabras de $L(Q)$ tiene la forma sintáctica exigida por POS_λ, empleamos la biblioteca PATTERN[13], para realizar conjugaciones o conversiones de género y/o número y reemplazar correctamente POS_λ.

Si el conjunto de palabras $L(Q)$ no contiene ningún tipo de palabra llena, que sea adecuada o que pueda manipularse con la biblioteca PATTERN, para reemplazar las etiquetas POS_λ, se toma otra palabra, $w_i \in L(Q)$, lo más cercana a Q (en función de la distancia producida por Word2vec). Se define un nuevo $Q^* = w_i$ que será utilizado para generar un nuevo conjunto de palabras $L(Q^*)$. Este procedimiento se repite, hasta que $L(Q^*)$ contenga una palabra que pueda reemplazar la POS_λ en cuestión. El resultado de este procedimiento es una nueva frase f que no existe en los corpora 5KL y 8KF. La Figura 4 muestra el proceso descrito.

[13] https://www.clips.uantwerpen.be/pattern

4.3. Modelo 2: Texto enlatado, Word2vec y análisis morfosintáctico

En este modelo proponemos una combinación entre el modelo de *texto enlatado* (Sección 4.1.2) y el algoritmo Word2vec entrenado sobre el corpus 5KL. El objetivo es eliminar las iteraciones del Modelo 1, que son necesarias cuando las etiquetas POS[14] no pueden ser reemplazadas con el léxico $L(Q)$.

Se efectúa un análisis morfosintáctico del corpus 5KL usando FreeLing y se usan las etiquetas POS para crear conjuntos de palabras que posean la misma información gramatical (etiquetas POS idénticas). Una Tabla Asociativa (TA) es generada como resultado de este proceso. La TA consiste en entradas de pares POS_k y una lista de palabras asociadas. Formalmente, se reprenta $\text{POS}_k \rightarrow V_k = \{v_{k,1}, v_{k,2}, \ldots, v_{k,i}\}$. El Modelo 2 es ejecutado una sola vez para cada etiqueta POS_k. La EGP no será reemplazada completamente: las palabras funcionales y los signos de puntuación son conservados.

Para generar una nueva frase se reemplaza cada etiqueta $\text{POS}_k \in \text{EGP}$, $k = 1, 2, \ldots$, por una palabra adecuada. Para cada etiqueta POS_k, se recupera el léxico V_k a partir de TA.

El vocabulario es procesado por el algoritmo Word2vec, que calcula el valor de proximidad (distancia), $dist(Q, v_{k,i})$, entre cada palabra del vocabulario, $v_{k,i}$, y el *query* Q del usuario. Después se ordena el vocabulario V_k en forma descendente según los valores de proximidad $dist(Q, v_{k,i})$ y se escoge aleatoriamente uno de los primeros tres elementos para reemplazar la etiqueta POS_k de la EGP.

[14] Por motivos de claridad de la notación, en esta sección y en la siguiente una etiqueta POS_λ será designada solamente por POS.

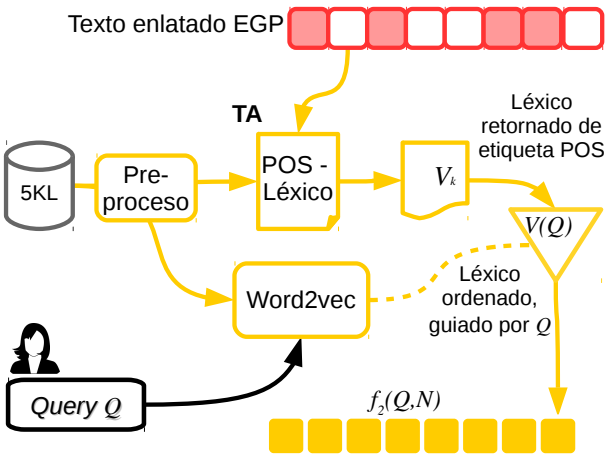


Figura 5: Modelo 2 de aproximación semántica basada en *Word2vec* y análisis morfosintáctico.

El resultado es una nueva frase $f_2(Q, N)$ que no existe en los corpora 5KL y 8KF. El proceso se ilustra en la figura 5.

4.4. Modelo 3: Texto enlatado, Word2vec e interpretación geométrica

El Modelo 3 reutiliza varios de los recursos anteriores: el algoritmo Word2vec, la Tabla Asociativa TA y la estructura gramatical parcialmente vacía (EGP) obtenida del modelo de *texto enlatado*. El modelo utiliza distancias vectoriales para determinar las palabras más adecuadas que sustituirán las etiquetas POS de una EGP y así generar una nueva frase. Para cada etiqueta POS_k, $k = 1, 2, \ldots \in$ EGP, que se desea sustituir, usamos el algoritmo descrito a continuación.

Se construye un vector para cada una de las tres palabras siguientes.

- o: es la palabra k de la frase f_o (Sección 4.1.2), correspondiente a la etiqueta POS_k. Esta palabra permite recrear un contexto del cual la nueva frase debe alejarse, evitando producir una paráfrasis.

- Q: es la palabra que define al *query* proporcionado por el usuario.

- w: la palabra candidata que podría reemplazar POS_k, $w \in V_k$. El vocabulario posee un tamaño $|V_k| = m$ palabras y es recuperado de la TA correspondiente a la POS_k.

Las 10 palabras o_i más próximas a o, las 10 palabras Q_i más próximas a Q y las 10 palabras w_i más próximas a w (en este orden y obtenidas con Word2vec), son concatenadas y representadas en un vector simbólico $\vec{U}$ de 30 dimensiones. El número de dimensiones fue fijado a 30 de manera empírica, como un compromiso razonable entre diversidad léxica y tiempo de procesamiento. El vector $\vec{U}$ puede ser escrito como

$$\vec{U} = (u_1, \ldots, u_{10}, u_{11}, \ldots, u_{20}, u_{21}, \ldots, u_{30}), \quad (2)$$

donde cada elemento $u_j, j = 1, \ldots, 10$, representa una palabra próxima a o; $u_j, j = 11, \ldots, 20$, representa una palabra próxima a Q; y $u_j, j = 21, \ldots, 30$, es una palabra próxima a w. $\vec{U}$ puede ser re-escrito de la siguiente manera,

$$\vec{U} = (o_1, \ldots, o_{10}, Q_{11}, \ldots, Q_{20}, w_{21}, \ldots, w_{30}). \quad (3)$$

o, Q y w generan respectivamente tres vectores numéricos de 30 dimensiones:

$$o : \vec{X} = (x_1, \ldots, x_{10}, x_{11}, \ldots, x_{20}, x_{21}, \ldots, x_{30}),$$
$$Q : \vec{Q} = (q_1, \ldots, q_{10}, q_{11}, \ldots, q_{20}, q_{21}, \ldots, q_{30}),$$
$$w : \vec{W} = (w_1, \ldots, w_{10}, w_{11}, \ldots, w_{20}, w_{21}, \ldots, w_{30}),$$

donde los valores de $\vec{X}$ son obtenidos tomando la distancia entre la palabra o y cada palabra $u_j \in \vec{U}, j = 1, \ldots, 30$. La distancia, $x_j = dist(o, u_j)$ es proporcionada por Word2vec y además $x_j \in [0, 1]$. Evidentemente la palabra o estará más próxima a las 10 primeras palabras u_j que a las restantes.

Un proceso similar permite obtener los valores de $\vec{Q}$ y $\vec{W}$ a partir de Q y w, respectivamente. En estos casos, el *query* Q estará más próximo a las palabras u_j en las posiciones $j = 11, \ldots, 20$ y la palabra candidata w estará más próxima a las palabras u_j en las posiciones $j = 21, \ldots 30$.

Enseguida, se calculan las similitudes coseno entre $\vec{Q}$ y $\vec{W}$ (4) y entre $\vec{X}$ y $\vec{W}$ (5),

$$\theta = \cos(\vec{Q}, \vec{W}) = \frac{\vec{Q} \cdot \vec{W}}{|\vec{Q}||\vec{W}|}, \quad (4)$$

$$\beta = \cos(\vec{X}, \vec{W}) = \frac{\vec{X} \cdot \vec{W}}{|\vec{X}||\vec{W}|}. \quad (5)$$

Estos valores de θ y β están normalizados en [0,1]. El proceso se repite para todas las palabras w del léxico V_k. Esto genera otro conjunto de vectores $\vec{X}, \vec{Q}$ y $\vec{W}$ para los cuales se deberán calcular nuevamente las similitudes. Al final se obtienen m valores de similitudes θ_i y β_i, $i = 1, \ldots, m$, y se calculan los promedios $\langle\theta\rangle$ y $\langle\beta\rangle$.

El cociente normalizado

$$\left(\frac{\langle\theta\rangle}{\theta_i}\right)$$

indica qué tan grande es la similitud de θ_i con respecto al promedio $\langle\theta\rangle$ (interpretación de tipo maximización); es decir, que tan próxima se encuentra la palabra candidata w al *query* Q.

El cociente normalizado

$$\left(\frac{\beta_i}{\langle\beta\rangle}\right)$$

indica qué tan reducida es la similitud de β_i con respecto a $\langle\beta\rangle$ (interpretación de tipo minimización); es decir, qué tan lejos se encuentra la palabra candidata w de la palabra o de f_o.

Estas fracciones se obtienen en cada par (θ_i, β_i) y se combinan (minimización-maximización) para calcular un score S_i, según la ecuación

$$S_i = \left(\frac{\langle\theta\rangle}{\theta_i}\right) \cdot \left(\frac{\beta_i}{\langle\beta\rangle}\right) . \qquad (6)$$

Mientras más elevado sea el valor S_i, mejor obedece a nuestros objetivos: acercarse al *query* y alejarse de la semántica original.

Finalmente, ordenamos en forma decreciente la lista de valores de S_i y se escoge, de manera aleatoria, entre los 3 primeros, la palabra candidata w que reemplazará la etiqueta POS_k en cuestión. El resultado es una nueva frase $f_3(Q, N)$ que no existe en los corpora utilizados para construir el modelo.

En la Figura 6 se muestra una representación del modelo descrito.

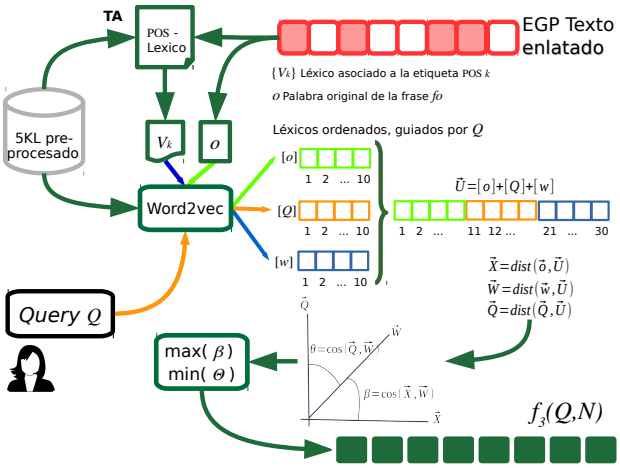


Figura 6: Modelo 3 de aproximación semántica, basada en interpretación geométrica min-max.

5. Experimentos y resultados

Se han diseñado tres experimentos para validar los tres modelos semánticos desarrollados en este trabajo. A partir de un *query* del usuario y de una longitud de palabras, el sistema realiza los procesos siguientes.

- **Modelo 1:** El modelo generativo de cadenas de Markov produce una EGV que se envía al modelo de aproximación semántica para generar las frases f_1.
- **Modelo 2**: El modelo generativo de *texto enlatado* produce una EGP y se envía al modelo de aproximación semántica para generar las frases f_2.
- **Modelo 3**: El modelo de *texto enlatado* produce una EGP y se utiliza el modelo de interpretación geométrica para la generación de las frases f_3.

Dado la especificidad de nuestros experimentos (idioma, corpora disponibles, homosintaxis), no es posible compararse directamente con otros métodos. Tampoco consideramos la utilización de un *baseline* de tipo aleatorio, porque los resultados carecerían de la homosintaxis y sería sumamente fácil obtener mejores resultados. Dicho lo anterior, el Modelo 1 podría ser considerado como nuestro propio *baseline*.

5.1. Resultados

Enseguida presentamos unos ejemplos de las frases obtenidas en función de los experimentos propuestos. Para el *query*, Q, y la longitud en número de palabras, N, los resultados se muestran en el formato,

$$f(Q, N) = \text{frase generada} . \qquad (7)$$

Los resultados han sido generados empleando los *queries* Q=`{GUERRA,SOL}` en todos los casos.

Modelo 1

1. $f(\texttt{GUERRA}, 12)$ = *El ejército conquista mediante el enemigo. La batalla es la guerra desde.*
2. $f(\texttt{GUERRA}, 13)$ = *Toda batalla en rebelión es la guerra contra el ejército en el combate.*
3. $f(\texttt{SOL}, 12)$ = *La luna salvo la lluvia sobre el ocaso hacia el cielo brilla.*
4. $f(\texttt{SOL}, 13)$ = *Cuántos naveguen salvo iluminar para el cielo hacia la aurora es la luna.*

Modelo 2

1. $f(\texttt{GUERRA}, 9)$ = *El incivil comportamiento para la magnificencia es la dicha.*
2. $f(\texttt{GUERRA}, 10)$ = *La cultura es la religión de dogmatizar los bienes caducos.*
3. $f(\texttt{SOL}, 11)$ = *Brilla que contener siempre. Nunca se es dominado de el todo.*
4. $f(\texttt{SOL}, 10)$ = *El rocío exhala el bosque después de haberlo fatigado.*

Modelo 3

1. $f(\texttt{GUERRA}, 9)$ = *Existe demasiada innovacion en torno a muy pocos sucesos.*
2. $f(\texttt{GUERRA}, 9)$ = *En la pelea todo debe motivo, menos la retirada.*

3. $f(\texttt{SOL}, 11)$ = *Con rapidez, los monógamos impedimentos buscan para iluminar nos la luz.*
4. $f(\texttt{SOL}, 10)$ = *Incluso los luceros ingratos son comilones, y por tanto antiguos.*

5.2. Protocolo de evaluación e interpretación de resultados

A continuación presentamos un protocolo de evaluación manual de los resultados obtenidos. El experimento consistió en la generación de 15 frases por cada uno de los tres modelos propuestos. Para cada modelo, se consideraron tres *queries*, $Q = \{\texttt{AMOR, GUERRA, SOL}\}$, generando 5 frases con cada uno. Las 15 frases fueron mezcladas entre sí y reagrupadas por *queries*, antes de presentarlas a los evaluadores.

Para la evaluación, se pidió a 7 personas leer cuidadosamente las 45 frases (15 frases por *query*). Todos los evaluadores poseen estudios universitarios y son hispanohablantes nativos. Se les pidió anotar en una escala de [0,1,2] (donde 0=mal, 1=aceptable y 2=correcto) los criterios siguientes:

- **Gramaticalidad:** ortografía, conjugaciones correctas, concordancia en género y número.
- **Coherencia:** legibilidad, percepción de una idea general.
- **Contexto:** relación de la frase con respecto al *query*.

En el Anexo A se muestran las frases generadas para la evaluación manual.

El mini-test de Turing fue evaluado con una nota de 0 o 1. A los evaluadores se les hizo creer que había algunas frases escritas por personas y otras escritas por los algoritmos. Se les pidió indicar cuáles frases pensaban que habían sido generadas por personas (0) y cuáles por algoritmos (1).

Los resultados de la evaluación se presentan en la Figura 7, en la forma de gráfica de barras, donde cada barra representa un criterio evaluado. Los valores representados corresponden a la moda para cada uno de los criterios.

La primera sección de barras ilustra la evaluación de las frases generadas por el Modelo 1. En ella se puede apreciar como los evaluadores percibieron una estrecha relación con el contexto *query* (barra roja) y una gramática aceptable (barra gris). Sin embargo para este modelo, la barra de coherencia (barra azul) es nula, lo que indica que los evaluadores no perciben las frases como coherentes.

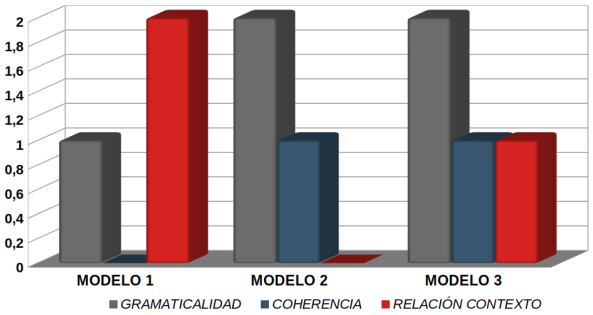


Figura 7: Evaluación de la coherencia, gramaticalida y contexto.

La estrecha relación con el contexto se debe al alto grado de libertad que caracteriza a la EGV generada por el modelo de Markov. Esta EGV permite que todos los elementos de la estructura puedan ser sustituidos por un léxico guiado por los resultados del algoritmo Word2vec.

Para los resultados del Modelo 2, los evaluadores perciben frases razonablemente coherentes y gramaticalmente correctas. Sin embargo, los evaluadores no percibieron una relación evidente entre el contexto de la frase generada y el *query*. Esto se debe a que las frases generadas reportan, en su mayoría, el mismo contexto o idea de la frase original, pudiendo ser interpretado como una paráfrasis elemental, que no es lo que deseamos.

Finalmente, el Modelo 3 genera frases coherentes, gramaticalmente correctas y más bien relacionadas al *query* que el Modelo 2, siendo el único modelo donde los tres criterios evaluados se hacen presentes. Esto se logra siguiendo una intuición opuesta a la paráfrasis: buscamos conservar la estructura sintáctica de la frase original, generando una semántica completamente diferente.

En general se puede apreciar que a diferencia de los modelos 1 y 2, en donde sólo 2 de cada 3 criterios fueron percibidos con claridad, el Modelo 3 es el único que obtuvo resultados positivos en los tres criterios.

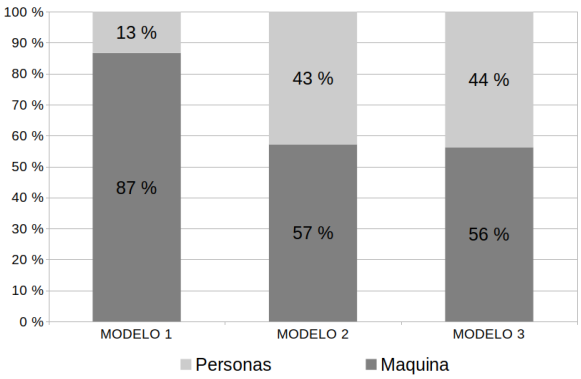


Figura 8: Evaluación del mini-test de Turing.

En la Figura 8, de acuerdo a los resultados del mini-test de Turing, se muestra en tonalidad oscura la moda con la que los evaluadores percibieron las frases como escritas por personas, mientras que en una tonalidad clara (gris) cuando las frases se percibieron como generadas por máquinas. En general, los evaluadores perciben con el 44 % de las frases generadas por el Modelo 3 como generadas por una persona. Esta es la mejor percepción de los tres modelos.

6. Conclusión y trabajo futuro

En este artículo hemos presentado tres modelos de producción de frases literarias. La generación de este género textual necesita sistemas específicos que deben considerar el estilo, la sintaxis y una semántica, que no necesariamente respeta la lógica de los documentos de géneros factuales, como el periodístico, enciclopédico o científico. Los resultados obtenidos son alentadores para el Modelo 3, utilizando *texto enlatado*, Word2vec con redes neuronales y una interpretación del tipo IR.

Uno de los principales problemas detectados en los tres modelos es la pérdida de coherencia en las frases generadas. Esto puede deberse a la ambigüedad gramatical que FreeLing es incapaz de resolver. Por ejemplo, en la frase: *"Es preciso que uno de los tres muera."*, FreeLing detecta el verbo en subjuntivo —muera— como un sustantivo y lo etiqueta como N. Los sustantivos son candidatos a ser subsituidos en nuestros modelos. A partir de ahí, al construir una nueva frase con nuestros sistemas a partir de esta representación y usando el *query*="mundo", obtemos por ejemplo: *"Es necesario que uno de los tres* ***tierra***", que es incoherente.

El trabajo a futuro necesita la implementación de módulos para procesar los *queries* multitérmino del usuario. También se tiene contemplada la generación de frases retóricas en el dominio de discursos políticos, utilizando los modelos aquí propuestos u otros con un enfoque probabilístico (Charton & Torres-Moreno, 2010). Tenemos la hipótesis que una cierta atractividad del discurso político reside no tanto en el contenido mismo, sino en la estructura y en la manera de producir dichas frases (Cossu et al., 2014; Abascal-Mena et al., 2015).

Los modelos aquí presentados pueden ser enriquecidos a través de la integración de otros componentes, como características de una personalidad y/o las emociones (Wedemann & Carvalho, 2012; Wedemann & Plastino, 2016; Edalat, 2017; Siddiqui et al., 2018).

La introducción de la rima puede ser sumamente interesante cuando se produzcan varias frases para constituir un párrafo o una estrofa. El acoplamiento con un generador de rimas asonantes y consonantes (Medina-Urrea & Torres-Moreno, 2019) está previsto.

Finalmente, un protocolo de evaluación semiautomático (y a gran escala) está igualmente previsto.

Agradecimientos

Este trabajo está financiado por el Consejo Nacional de Ciencia y Tecnología (Conacyt, México), beca núm. 661101 y parcialmente por la Université d'Avignon, Laboratoire Informatique d'Avignon (LIA), programa de becas Agricol Perdiguier (France).

Los autores agradecen profundamente a los siete evaluadores anónimos que participaron en este trabajo; y a Carlos González por sus observaciones y sugerencias, asi como a los árbitros de la revista por sus comentarios pertinentes.

Referencias

Abascal-Mena, Rocío, Jean-Valère Cossu, Alejandro Molina-Villegas & Juan-Manuel Torres-Moreno. 2015. Anotación automática de datos acerca de la reputación de los políticos en redes sociales. *Research in Computing Science* 97. 81–99.

Agirrezabal, Manex, Bertol Arrieta, Aitzol Astigarraga & Mans Hulden. 2013. POS-tag based poetry generation with WordNet. En *14th European Workshop on Natural Language Generation*, 162–166.

Bengio, Yoshua, Aaron Courville & Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8). 1798–1828. doi 10.1109/TPAMI.2013.50.

Boden, Margaret A. 2004. *The creative mind: Myths and mechanisms*. Abingdon: Routledge 2ª ed.

Bracewell, David B, Fuji Ren & Shingo Kuriowa. 2005. Multilingual single document keyword extraction for information retrieval. En *International Conference on Natural Language Processing and Knowledge Engineering*, 517–522. doi 10.1109/NLPKE.2005.1598792.

Charton, Eric & Juan-Manuel Torres-Moreno. 2010. Modélisation automatique de connecteurs logiques par analyse statistique du contexte. *Canadian Journal of Information and Library Science* 35(3). 287–306. 10.1353/ils.2011.0017.

Clark, Elizabeth, Yangfeng Ji & Noah A. Smith. 2018. Neural text generation in stories using entity representations as context. En *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2250–2260. 10.18653/v1/N18-1204.

Colton, Simon. 2012. *Automated theory formation in pure mathematics*. London: Springer. 10.1007/978-1-4471-0147-5.

Colton, Simon & Geraint A Wiggins. 2012. Computational creativity: The final frontier? En *20th European Conference on Artificial Intelligence*, 21–26.

Cossu, Jean-Valère, Rocío Abascal-Mena, Alejandro Molina-Villegas, Juan-Manuel Torres-Moreno & Eric SanJuan. 2014. Bilingual and cross domain politics analysis. *Research in Computing Science* 85. 9–19.

van Deemter, Kees, Mariët Theune & Emiel Krahmer. 2005. Real versus template-based natural language generation: A false opposition? *Computational Linguistics* 31(1). 15–24. 10.1162/0891201053630291.

Edalat, Abbas. 2017. Self-attachment: A holistic approach to computational psychiatry. En Péter Érdi, Basabdatta. Sen Bhattacharya & Amy L Cochran (eds.), *Computational Neurology and Psychiatry*, vol. 6, 273–314. Cham: Springer. 10.1007/978-3-319-49959-8_10.

Fu, Ruiji, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang & Ting Liu. 2014. Learning semantic hierarchies via word embeddings. En *52nd Annual Meeting of the Association for Computational Linguistics*, 1199–1209. 10.3115/v1/P14-1113.

Gervás, Pablo, Raquel Hervás & Carlos León. 2015. Generating plots for a given query using a case-base of narrative schemas. En *23rd International Conference on Case-Based Reasoning*, 103–112.

Gonçalo Oliveira, Hugo. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. En *10th International Conference on Natural Language Generation*, 11–20. 10.18653/v1/W17-3502.

Gonçalo Oliveira, Hugo. 2012. PoeTryMe: a versatile platform for poetry generation. En *1st International Workshop on Computational Creativity, Concept Invention and General Intelligence*, s.p.

Gonçalo Oliveira, Hugo & Amílcar Cardoso. 2015. Poetry generation with PoeTryMe. En *Computational Creativity Research: Towards Creative Machines*, 243–266. 10.2991/978-94-6239-085-0_12.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville & Yoshua Bengio. 2014. Generative adversarial nets. En Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence & K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, vol. 27, 2672–2680.

Huang, Eric H., Richard Socher, Christopher D. Manning & Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. En *50th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 873–882.

Iria, da Cunha, M. Teresa Cabré, Eric SanJuan, Gerardo Sierra, Juan-Manuel Torres-Moreno & Jorge Vivaldi. 2011. Automatic specialized vs. non-specialized sentence differentiation. En *12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, 266–276. 10.1007/978-3-642-19437-5_22.

Ke, Wang & Wan Xiaojun. 2018. SentiGAN: Generating sentimental texts via mixture adversarial networks. En *27th International Joint Conference on Artificial Intelligence (IJCAI)*, 4446–4452. 10.24963/ijcai.2018/618.

Lebret, Rémi, David Grangier & Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. En *Conference on Empirical Methods in Natural Language Processing*, 1203–1213. 10.18653/v1/D16-1128.

Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge: The MIT Press.

Martínez, Gerardo Sierra. 2018. *Introducción a los corpus lingüísticos*. Mexico: Instituto de Ingeniería, UNAM.

McRoy, Susan, Songsak Channarukul & Syed Ali. 2003. An augmented template-based approach to text realization. *Natural Language Engineering* 9(4). 381–420. 10.1017/S1351324903003188.

Medina-Urrea, Alfonso & Juan-Manuel Torres-Moreno. 2019. RIMAX: ranking semantic rhymes by calculating definition similarity. *arXiv* CoRR abs/1912.09558.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado & Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. En *26th International Conference on Neural Information Processing Systems*, 3111–3119.

Mikolov, Tomas, Wen-tau Yih & Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. En *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Mikolov, Tomas & Geoffrey Zweig. 2012. Context dependent recurrent neural network language model. En *IEEE Spoken Language Technology Workshop (SLT)*, 234–239. 10.1109/SLT.2012.6424228.

Molins, Paul & Guy Lapalme. 2015. JSrealB: A bilingual text realizer for web programming. En *15th European Workshop on Natural Language Generation (ENLG)*, 109–111. 10.18653/v1/W15-4719.

Montfort, Nick. 2008b. Through the park. https://nickm.com/poems/through_the_park.html.

Montfort, Nick. 2008c. The two. https://nickm.com/poems.

Montfort, Nick. 2009. Taroko gorge. https://nickm.com/poems/taroko_gorge.html.

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

Pérez y Pérez, Rafael. 2015. *Creatividad computacional*. México: Larousse, Grupo Editorial Patria.

Riedl, Mark O. & R. Michael Young. 2006. Story planning as exploratory creativity: Techniques for expanding the narrative search space. *New Generation Computing* 24(3). 303–323. 10.1007/BF03037337.

Sharples, Mike. 1996. *How we write: Writing as creative design*. London: Routledge.

Siddiqui, Maheen, Roseli S. Wedemann & Henrik Jeldtoft Jensen. 2018. Avalanches and generalized memory associativity in a network model for conscious and unconscious mental functioning. *Physica A: Statistical Mechanics and its Applications* 490. 127–138. 10.1016/j.physa.2017.08.011.

Sridhara, Giriprasad, Emily Hill, Divya Muppaneni, Lori Pollock & K. Vijay-Shanker. 2010. Towards automatically generating summary comments for java methods. En *IEEE/ACM International Conference on Automated Software Engineering*, 43–52. 10.1145/1858996.1859006.

Szymanski, Grzegorz & Zygmunt Ciota. 2002. Hidden markov models suitable for text generation. En *International Conference on Signal, Speech and Image Processing*, 3081–3084.

Torres-Moreno, Juan-Manuel. 2012. Beyond stemming and lemmatization: Ultra-stemming to improve automatic text summarization. *arXiv* abs/1209.3126.

Torres-Moreno, Juan-Manuel (ed.). 2014. *Automatic text summarization*. London: ISTE, Wiley.

Wedemann, Roseli S. & Luiz Alfredo Vidal de Carvalho. 2012. Some things psychopathologies can tell us about consciousness. En *International Conference on Artificial Neural Networks (ICANN)*, vol. 7552, 379–386. 10.1007/978-3-642-33269-2_48.

Wedemann, Roseli S. & Angel Ricardo Plastino. 2016. Física estadística, redes neuronales y freud. *Revista Núcleos* 3. 4–10.

Welleck, Sean, Kiante Brantley, Hal Daumé & Kyunghyun Cho. 2019. Non-monotonic sequential text generation. En *36th International Conference on Machine Learning*, 11656–11676.

Zhang, Xingxing & Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 670–680. 10.3115/v1/D14-1074.

A. Frases evaluadas

En este anexo presentamos las 45 (15 frases × 3 sistemas) frases generadas por nuestros modelos, que fueron evaluadas manualmente (ver sección 5).

Modelo 1

1. $f(\texttt{GUERRA}, 10)$ = *La simpatía es el aprecio que ha olvidado la ternura.*
2. $f(\texttt{GUERRA}, 13)$ = *Toda batalla en rebelión es la guerra contra el ejército en el combate.*
3. $f(\texttt{GUERRA}, 12)$ = *El ejército conquista mediante el enemigo. La batalla es la guerra desde.*
4. $f(\texttt{GUERRA}, 13)$ = *El enemigo salvo la batalla en el terrorismo mediante el ejército conquista contra.*
5. $f(\texttt{GUERRA}, 11)$ = *Su atómica lucha. la guerra desde el combate. la derrota es.*
6. $f(\texttt{SOL}, 12)$ = *La luna salvo la lluvia sobre el ocaso hacia el cielo brilla.*
7. $f(\texttt{SOL}, 13)$ = *Cuántos naveguen salvo iluminar para el cielo hacia la aurora es la luna.*
8. $f(\texttt{SOL}, 13)$ = *Nuestro cielo es verdaderamente el que luna es la lluvia bajo la aurora.*
9. $f(\texttt{SOL}, 9)$ = *Cuántos perezcas durante amanecer el ocaso bajo la luna.*
10. $f(\texttt{SOL}, 6)$ = *El resplandor deshoja bajo la aurora.*
11. $f(\texttt{AMOR}, 13)$ = *Toda amistad contra compasión por la ternura es la pasión en el afecto.*
12. $f(\texttt{AMOR}, 13)$ = *Todos durante la ternura con el afecto hacia la compasión por la virtud.*
13. $f(\texttt{AMOR}, 12)$ = *El cariño con el afecto hacia el amado aborrece salvo en sentimiento.*
14. $f(\texttt{AMOR}, 11)$ = *La ternura envidia entre el cariño. El amado es demasiada compasión.*
15. $f(\texttt{AMOR}, 11)$ = *Cuánto bien deseo sin la amistad. El afecto es otra ternura.*

Modelo 2

1. $f(\texttt{GUERRA}, 9)$ = *El incivil comportamiento para la magnificencia es la dicha.*
2. $f(\texttt{GUERRA}, 9)$ = *Mi anciana: tú felicidad no la alumbra ninguna autoridad.*
3. $f(\texttt{GUERRA}, 9)$ = *No hay hipocresía más impopular que la historia simulada.*
4. $f(\texttt{GUERRA}, 10)$ = *La cultura es la religión de dogmatizar los bienes caducos.*
5. $f(\texttt{GUERRA}, 10)$ = *De el temperamento a la entereza hay una velocidad terrible.*
6. $f(\texttt{SOL}, 9)$ = *El color que reanima más es una picardía suprema.*
7. $f(\texttt{SOL}, 10)$ = *La paz es el suelo artificial de la luz moderna.*
8. $f(\texttt{SOL}, 10)$ = *En el vocabulario está el bosque mixto de una política.*
9. $f(\texttt{SOL}, 10)$ = *El rocío exhala el bosque después de haber lo fatigado.*
10. $f(\texttt{SOL}, 11)$ = *Brilla que contener siempre. Nunca se es dominado de el todo.*
11. $f(\texttt{AMOR}, 9)$ = *Estorbas una amada calle de un amor de fantasías.*
12. $f(\texttt{AMOR}, 9)$ = *Jamás hubo una conquista nueva o una amistad extraña.*
13. $f(\texttt{AMOR}, 11)$ = *Dios dejó la desesperacion para trabajar la y no para desilusionarla.*
14. $f(\texttt{AMOR}, 11)$ = *Abultar se en cualquier mentira, es conveniente que no porfiar nada.*
15. $f(\texttt{AMOR}, 10)$ = *Por culpa, el anhelo no suprime siempre con el deseo.*

Modelo 3

1. $f(\texttt{GUERRA}, 9)$ = *Existe demasiada innovacion en torno a muy pocos sucesos.*
2. $f(\texttt{GUERRA}, 9)$ = *En la pelea todo debe motivo, menos la retirada.*
3. $f(\texttt{GUERRA}, 10)$ = *La nueva pelea se combate cuando se abandona la civilización.*
4. $f(\texttt{GUERRA}, 10)$ = *La codicia, siempre adversa, es terrible engendrada contra un desgraciado.*
5. $f(\texttt{GUERRA}, 10)$ = *La retirada es el vapor remediable de el lucha ilimitada.*
6. $f(\texttt{SOL}, 9)$ = *Si tus dulces fueran amanecer, mis ojos marchitas fueran.*
7. $f(\texttt{SOL}, 11)$ = *Con rapidez, los monógamos impedimentos buscan para iluminar nos la luz.*
8. $f(\texttt{SOL}, 10)$ = *Incluso los luceros ingratos son comilones, y por tanto antiguos.*
9. $f(\texttt{SOL}, 9)$ = *El ocaso es una extraña frente de la inmortalidad.*
10. $f(\texttt{SOL}, 9)$ = *La aurora es el amanecer que ha olvidado la calma.*
11. $f(\texttt{AMOR}, 10)$ = *En el aprecio está el cariño forzoso de una simpatía.*
12. $f(\texttt{AMOR}, 10)$ = *Los cariños no conocen de nada a un respeto loco.*
13. $f(\texttt{AMOR}, 10)$ = *No está la simpatía en las bondades de la envidia.*
14. $f(\texttt{AMOR}, 9)$ = *Si el respeto es felicidad, que oculten los cariños.*
15. $f(\texttt{AMOR}, 10)$ = *Acostumbramos de lamentar aquello que se ha enseñado a comprender.*

Proposta recebida em September 2019 e aceite para publicação em April 2020.

Análise da Lei de Menzerath no Português Brasileiro

Menzerath's law analysis in Brazilian Portuguese

Leonardo Araujo
Universidade Federal de São João del Rei
leolca@ufsj.edu.br

Aline Benevides
Universidade de São Paulo
benevides.aline12@gmail.com

Marcos Pereira
Universidade Federal de São João del Rei
marcos.vinicius@ufsj.edu.br

Resumo

Sob a ótica da Linguística Quantitativa, este trabalho revisita a Lei de Menzerath, aplicando-a aos dados do português brasileiro, a partir das seguintes unidades de análise: palavras, sílabas e fonemas. Os dados foram extraídos do Corpus ABG. Análises estatísticas foram realizadas nos modelos propostos, as quais demonstraram uma relação de decrescimento entre o comprimento médio das palavras (em sílabas) e o comprimento médio das sílabas (em fonemas); resultados esses que corroboram a Lei de Menzerath. Além disso, constatou-se, de maneira geral, que melhores medições ou a existência de variáveis não consideradas no modelo poderão ser utilizadas para melhorá-lo.

Palavras chave

lei de Menzerath, linguística quantitativa, análise estatística

Abstract

Under the perspective of Quantitative Linguistics, this paper revisits the Menzerath's Law, applying it to data from Brazilian Portuguese, using the following unities of analysis: words, syllables and phonemes. The data was extracted from the ABG Corpus. Statistical analyses are performed on the proposed models, corroborating the existence of a decay relationship between the mean length of words (in syllables) and the average length of syllables (in phonemes); what corroborates the Menzerath Law. It is noticed that better measures or variables not considered in the model might be used to improve it.

Keywords

Menzerath's law, quantitative linguistics, statistical analysis

1. Introdução

A linguística é o campo das ciências que estuda a linguagem, analisando a sua forma, utilização, significado e contexto. Ela se estabeleceu, de forma consistente, a partir do século XX, através da publicação do Curso de Linguística Geral, de Ferdinand de Saussure, em 1916, dando início a Linguística Moderna. Neste período, a primeira orientação teórica, denominada de Linguística Estruturalista, emerge com o objetivo de estabelecer e de descrever os sistemas linguísticos, valendo-se, para tanto, da noção de valor, a partir de distinções teóricas como *língua* vs. *fala*, *forma* vs. *substância*. Os estruturalistas, por serem avessos ao estudo do sentido, de caráter mental e, portanto, pertencente à psicologia individual, dedicam-se ao estudo da forma. Nesse sentido, a língua passa a ser analisada do ponto de vista sistêmico por meio de relações de oposição (noção de valor). Sob orientação de Leonard Bloomfield, o estruturalismo americano estabelece que a análise das estruturas e das categorias gramaticais devem ser realizadas a partir de dados de sentenças ou de textos, e não mais extraídos de experiências prévias (Ilari, 2003).

A Linguística Quantitativa ganha abrangência, nesse sentido, ao conjugar métodos estatísticos e computacionais, a partir da análise de corpora linguísticos, a fim de caracterizar a linguagem, sua evolução e estrutura. O seu principal propósito é estabelecer leis que modelem a linguagem ou a comunicação e produzir formulações para uma teoria geral da linguagem (Altmann & Schwibbe, 1989; Köhler, 2005). Atribui-se o status de leis científicas àquelas formulações que podem ser derivadas a partir de axiomas, criando uma estrutura firme de rede nomológica.

Sob o auspício da tradição das ciências teoréticas, a Linguística Quantitativa busca estabelecer leis e formulações capazes de inter-

DOI: 10.21814/lm.12.1.300

relacioná-las, através de proposições e derivações lógicas. É sob essa perspectiva que o presente trabalho se baseia. Propomo-nos, neste artigo, a testar a predição da Lei de Menzerath a partir da análise de um corpus linguístico do português brasileiro, o Corpus ABG (Benevides & Guide, 2017). A Lei de Menzerath é conhecida por predizer, em termos gerais, que "um som é mais curto quão maior o todo em que ele ocorre (lei da quantidade)" e que "quantos mais sons possuir uma sílaba, menor serão seus comprimentos relativos"(Menzerath, 1954, p. 100).

Este artigo estrutura-se da seguinte maneira: na Seção 1, apresenta-se uma contextualização da área de Linguística Quantitativa (Seção 1.1), da Lei de Menzerath-Altmann (Seção 1.2) e de sua formulação matemática (Seção 1.3); a Seção 2 apresenta a abordagem deste trabalho, destacando as unidades linguísticas e o corpus com os dados do português brasileiro que foram utilizados para realizar as análises desta pesquisa; a Seção 2.4 apresenta os resultados gerados a partir da análise dos dados do corpus e os resultados de ajustes dos modelos matemáticos; na Seção 2.5, busca-se explorar os resultados e contrastá-los com as teorias linguísticas; e, por fim, conclui-se o trabalho na Seção 3.

1.1. Linguística Quantitativa

Os trabalhos em Linguística Quantitativa buscam explicações provenientes de leis estocástico-linguísticas que venham a estabelecer uma teoria geral da linguagem. Uma rápida revisão a respeito das principais leis estatísticas na linguística pode ser encontrada em Altmann & Gerlach (2016) ou também na enciclopédia *on-line* destinada à Linguística Quantitativa, Glottopedia (2019). A lei mais conhecida é a Lei de Zipf (Zipf, 1935, 1949; Ferrer-i-Cancho & Solé, 2002; Mitzenmacher, 2004; Ferrer-i-Cancho, 2006). Zipf observou, por meio da quantidade de ocorrências de palavras em um corpus, a existência de uma proporção inversa entre a frequência da palavra e o seu ranque, quando ordenadas por frequência. O mesmo tipo de relação também é verificada em outros fenômenos da natureza, por exemplo: na magnitude de terremotos (Abe & Suzuki, 2005), na população de cidades (Gabaix, 1999) e no número de requisições de páginas na internet (Adamic & Huberman, 2002). Há, ainda, trabalhos que buscam relacionar diferentes leis da Linguística Quantitativa, como o trabalho de Lü et al. (2010) que buscou relacionar a Lei de Zipf com a Lei de Heaps (Grzybek, 2007; Lü et al., 2010; Heaps, 1978; Herdan, 1960), a qual descreve o crescimento sublinear do vocabulário com o comprimento do corpus.

Tendo em vista que as bases ontológicas são essenciais para a construção de uma verdadeira teoria da linguagem e da comunicação sob o prisma da Linguística Quantitativa, transcreve-se abaixo um trecho do capítulo inicial de Altmann & Schwibbe (1989):

> *Leis são hipóteses bem fundamentadas e confirmadas. Uma generalização empírica nunca poderá se tornar uma lei, a menos que sejamos capazes de derivar a teoria de uma hipótese correspondente a ela. Nas ciências empíricas, esta é a forma mais comum de pesquisa: observações são feitas sob o pano de fundo de uma 'teoria' ainda embrionária, vaga e não formalizada, levando a generalizações empíricas, para a qual uma teoria correspondente é construída. Sem o estabelecimento de leis, um conjunto de afirmações dificilmente poderá ser chamado de teoria. Por esta razão, hoje não podemos falar na existência de uma teoria da linguagem, teoria gramatical, e assim por diante. A maioria dos conceitos linguísticos, embora bem complicados, consiste em uma gama de generalizações empíricas.* (Altmann & Schwibbe, 1989, p. 1)

1.2. O todo sem a parte não é todo

É notório que todas as línguas, apesar de terem uma representação fonológica das unidades linguísticas, se manifestam de forma que há grande variação em suas realizações, seja entre grupos ou indivíduos, ou mesmo quando se analisa diferentes realizações de um mesmo indivíduo. Os falantes não são meros usuários passivos, mas são parte integrante e ativa na dinâmica de uma língua. O uso acarreta mudanças, sendo que diversos fatores contribuem para tal. Fatores cognitivos, culturais, sociais e históricos, por exemplo, levam a mudanças constantes que, ao longo do tempo, podem provocar o surgimento de novos dialetos e idiomas (Bybee, 2015). Além dos fatores extralinguísticos, fatores internos à língua, como fonêmicos, morfológicos, sintáticos e semânticos, também contribuem para as constantes mudanças. Estas podem ser visualizadas de forma sincrônica ou diacrônica e sempre evidenciam a existência da ordem imanente. Quer a análise de uma língua seja feita no nível de sentenças, palavras, morfemas, sílabas ou fonemas, ordem e desordem são forças inerentes ao

processo linguístico.[1] Embora muito tenha sido investigado sobre a estrutura e a forma da língua e sua variação no tempo, ainda não existem certezas rígidas a respeito da aquisição e do processamento linguístico. Uma das hipóteses seria a existência de processos primários que guiariam a estruturação e o uso da língua, processos esses que atuariam em níveis mais altos e que poderiam ser compreendidos como generalizações (ou abstrações) de vários processos linguísticos, como leis fonéticas e gramaticais.

Em consonância com essa hipótese, vários estudos buscam encontrar e analisar quais seriam os motores atuantes em níveis hierárquicos mais altos. A Lei de Menzerath, por exemplo, é uma das leis mais conhecidas e corroboradas pela Linguística Quantitativa. Ela foi inicialmente elaborada por Menzerath (1928), sendo matematicamente formulada por Altmann (1980) e, posteriormente, confirmada pelos trabalhos de Hřebíček (1995); Andres (2010), dentre outros. Menzerath (1928, p. 104) propõe que "um som é mais curto quão maior o todo em que ele ocorre (lei da quantidade)" e "quantos mais sons possuir uma sílaba, menor serão seus comprimentos relativos". Tal postulação foi realizada a partir de uma análise do léxico da língua alemã, em que Menzerath (1954, p. 101) concluiu que "o número relativo de sons em uma sílaba decresce quando o número de sílabas em uma palavra aumenta", cunhando a frase "quão maior o todo, menores as partes!".

Antes de Menzerath, outros pesquisadores já demonstraram algumas observações que vão ao encontro da formulação de Menzerath. Jespersen (1904), por exemplo, analisou o comprimento das sílabas do francês, em especial a duração da vogal *a* em palavras como *pâtisserie*, *pâte* e *pâté*, e verificou que a vogal era sistematicamente mais curta em palavras mais longas. Outros autores também observaram semelhante tendência de redução da duração das vogais (Meyer, 1904; Roudet, 1910). Essas observações foram sistematizadas e estruturadas por Menzerath (1954), em uma análise sobre a estrutura morfológica do alemão. De forma semelhante ao *Princípio do Esforço Mínimo* formulado por Zipf (1935, 1949), Menzerath (1954) utilizou essa mesma proposição filosófica, chamando-a de *Princípio da Economia Cognitiva*, o que se manifestaria como um "fluxo constante de informação linguística" (Fenk & Fenk-Oczlon, 2013).

Conforme salientado, a proposta de Menzerath (1954) ganhou formulação matemática com o trabalho de Altmann (1980), que buscou descrever a relação entre os constituintes e os construtos. A sua validade foi observada, inicialmente, no indonésio e no inglês com Altmann (1980), seguida pelos trabalhos de Gerlach (1982), Hřebíček (1995), Polikarpov (2000a) para as línguas alemã, turca e russa, respectivamente. Além desses trabalhos, verificou-se, mais recentemente, a aplicação da Lei de Menzerath em uma análise paralela de um mesmo texto em 50 línguas diferentes (Coloma, 2015). Mais tarde, a mesma relação foi mostrada válida em música (Boroda & Altmann, 1991), cromossomos e genes (Wilde & Schwibbe, 1989; Ferrer-I-Cancho & Forns, 2009; Li, 2011; Nikolaou, 2014), proteínas (Shahzad et al., 2015) e, ainda, na compressão de dados, sob a ótica da teoria da informação (Gustison et al., 2016; Ferrer-i-Cancho, 2017).

1.3. Formulação Matemática da Lei de Menzerath

A formulação matemática, apresentada por Altmann (1980), propõe uma taxa de decrescimento constante do comprimento do componente, ou seja, $(1/y)\mathrm{d}y/\mathrm{d}x = -c$, onde y é o comprimento do constituinte e x o comprimento do construto. A formulação usual se restringe aos componentes de unidades imediatamente vizinhas, como oração e sentença ou palavras e sílabas. Entretanto, é possível estabelecermos relações compostas entre unidades que não sejam imediatamente vizinhas hierárquicas, por exemplo, palavras e sentenças, fonemas e palavras. Essa formulação pode ser necessária em línguas que apresentam palavras não silábicas,[2] como o russo (Grzybek & Altmann, 2002). Um refinamento no modelo é feito, considerando, além do decrescimento constante, um membro adicional inversamente proporcional ao comprimento do construto, como apresentado na Equação (1).

$$\frac{\mathrm{d}y/\mathrm{d}x}{y} = -c + \frac{b}{x} \tag{1}$$

[1] A linguagem pode ser vista como um sistema complexo e adaptativo, consistindo de múltiplos agentes que interagem entre si. O comportamento de cada agente depende de suas experiências passadas e também do ambiente e do contexto em que está inserido. O comportamento de um agente é fruto de diversos fatores, desde restrições perceptuais até motivações sociais. Como resultado da desordem criada nesse processo complexo, há a emergência de regularidades e de padrões (Beckner et al., 2010).

[2] Algumas línguas possuem palavras não silábicas, usualmente constituídas por uma ou duas consoantes (não silábicas) e sem a presença de vogais. O russo, por exemplo, possui as preposições *k*, *v* e *s* que funcionam como proclíticos para as palavras seguintes, contribuindo, assim, para o seu comprimento.

A taxa relativa de mudança no comprimento do constituinte ($^{(\mathrm{d}y/\mathrm{d}x)}/_{y}$) é uma soma de duas parcelas: a primeira, inversamente proporcional ao comprimento do construto ($^{b}/_{x}$), e a segunda, um fator constante ($-c$). Essa equação diferencial em (1) pode ser solucionada pela integração direta,[3] resultando em

$$y = Ax^b e^{-cx}, \tag{2}$$

onde o termo e^{-cx} e a constante A são sempre maiores do que zero. A curva dada pela eq. (2) é convexa crescente quando $b > 1$, uma curva côncava crescente quando $0 < b < 1$ e uma curva convexa decrescente quando $b < 0$ (Altmann, 1980). A eq. (2) pode assumir diferentes formas para $b = 0$, $b \neq 0$, $c = 0$ e $c \neq 0$, conforme a Tabela 1.

$b = 0$		$y = Ae^{-cx}$	modelo I	(3)
$b \neq 0$	$c = 0$	$y = Ax^b$	modelo II	(4)
	$c \neq 0$	$y = Ax^b e^{-cx}$	modelo III	(5)

Tabela 1: Soluções para diferentes possibilidades das constantes b e c (Altmann, 1980).

Cada uma das soluções apresentadas na Tabela 1 pode ser linearizada aplicando o logaritmo natural a ambos os lados, como nas eqs. (6) a (8):

$$\log y = \log A - cx \qquad \text{I} \tag{6}$$

$$\log y = \log A + b \log x \qquad \text{II} \tag{7}$$

$$\log y = \log A + b \log x - cx \qquad \text{III} \tag{8}$$

Note que, em cada uma das eqs. (6) a (8), tem-se uma relação linear entre as formas transformadas dessas variáveis:

$$\begin{aligned} y' = \log y &= \log A - cx \\ &= \beta_0 + \beta_2 x \end{aligned} \tag{9}$$

$$\begin{aligned} y' = \log y &= \log A + b \log x \\ &= \eta_0 + \beta_1 x' \end{aligned} \tag{10}$$

$$\begin{aligned} y' = \log y &= \log A + b \log x - cx \\ &= \beta_0 + \beta_1 x' + \beta_2 x \end{aligned} \tag{11}$$

onde $\beta_0 \triangleq \log A$, $\beta_1 \triangleq b$, $x' \triangleq \log x$ e $\beta_2 \triangleq -c$.

[3] Verifica-se a seguir que a Equação (2) é de fato solução:

$$\begin{aligned} {}^{\mathrm{d}y}/_{\mathrm{d}x} &= Abx^{b-1}e^{-cx} - cAx^b e^{-cx} \\ &= Ax^b e^{-cx}\left({}^{b}/_{x} - c\right) \\ &= y\left({}^{b}/_{x} - c\right), \end{aligned}$$

ou seja, obtém-se a Equação (1).

É possível, então, utilizar um modelo linear para relacionar a versão transformada da variável independente (variável explicativa), chamada de x', com a versão transformada da variável dependente (variável de resposta), chamada de y'. A essência desse modelo pode ser expressa, de forma geral, por

$$E[Y'|x] = \beta_0 + \beta_1 x' + \beta_2 x, \tag{12}$$

onde $E[\cdot]$ representa o valor esperado e $Y'|x$ indica a busca de possíveis valores de Y' (em que $Y' = \log Y$), o que restringe x a um único valor (consequentemente, x' também estará restrito). O parâmetro β_0 é o intercepto, β_1 e β_2 são as constantes de proporcionalidade em relação a cada um dos fatores (x' e x, respectivamente). Dado um conjunto de observações, é possível ajustar o modelo,[4] ou seja, encontrar os parâmetros β_0, β_1 e β_2 que melhor (sob algum critério a ser definido) explicam os dados. Para realizar o ajuste do modelo, deve-se ter um conjunto de dados, relacionando y e x.

Para uma regressão linear simples, a principal hipótese nula é $H_0 : \beta_1 = 0$ e $\beta_2 = 0$, ou seja, a média populacional de Y' é β_0 para todo valor de x, implicando que x não possui efeito em Y. A hipótese alternativa, dessa maneira, será H_1: $\beta_1 \neq 0$ e/ou $\beta_2 \neq 0$, implicando que mudanças em x acarretam mudanças em Y. Em alguns casos, é razoável considerar uma hipótese nula diferente, por exemplo, quando comparado com um padrão de referência. Nesse caso, a hipótese nula usualmente considerada é $H_0 : \beta_1 = 1$, com a hipótese alternativa $H_1 : \beta_1 \neq 1$. Neste trabalho, assume-se a seguinte hipótese nula: o comprimento médio das sílabas (em termos do número de fonemas que as constituem) é constante, para qualquer comprimento de palavra (em número de sílabas), ou seja, H_0: $\beta_1 = 0$ e $\beta_2 = 0$.

Para que o modelo I, dado pela Equação (3), descreva uma relação de decrescimento entre componentes e construtos, devemos ter $c > 0$. No caso do modelo II, dado pela Equação (4), devemos ter $b < 0$. Já o modelo III, dado pela Equação (5), terá derivada dada por

$$\begin{aligned} {}^{\mathrm{d}y}/_{\mathrm{d}x} &= Abx^{b-1}e^{-cx} - Acx^b e^{-cx} \\ &= (b - cx)Ae^{-cx}x^{b-1}. \end{aligned} \tag{13}$$

[4] Se o ajuste for realizado através do método dos mínimos quadrados, o critério escolhido será minimizar o erro quadrático médio; se o modelo for ajustado usando o método da máxima verossimilhança, busca-se maximar a probabilidade de que os dados observados sejam provenientes do modelo encontrado, ou seja, maximizar a verossimilhança.

Considerando que o comprimento do construto x não pode assumir valores negativos, para que a Equação (13) seja negativa, e assim exista uma relação de decrescimento, será necessário ter $(b - cx) < 0$, que poderá ocorrer quando $b < c$ e $c > 0$, considerando $x \geq 1$ ($x_{\min} = 1$), ou quando $b < cx_{\sup}$ e $c < 0$, onde $x_{\sup}$ é o limite superior para os valores que o comprimento do construto pode assumir.

Em algumas contextos, podemos analisar uma condição em que há um elemento hierarquicamente intermediário. Nesses casos, podemos obter uma relação decrescente entre construto e constituinte ou observar uma relação crescente entre eles (Altmann & Schwibbe, 1989; Prün, 1994; Grzybek & Stadlober, 2007). Suponha, então, que z seja nosso elemento intermediário entre y e x. Vamos analisar cada um dos modelos a seguir.

Para o modelo I, teremos:

$$y = Ae^{-cz}, \tag{14}$$

$$z = A'e^{-c'x}, \tag{15}$$

e, assim, podemos escrever y em função de x,

$$y = Ae^{-c(A'e^{-c'x})}. \tag{16}$$

Para que exista uma relação de decrescimento entre y e z, devemos ter $c > 0$. Da mesma forma, analisando z e x, devemos ter $c' > 0$. Para que também exista uma relação de decrescimento entre y e x, devemos analisar a derivada de Equação (16).

$$\mathrm{d}y/\mathrm{d}x = AA'cc'e^{-cA'e^{-c'x}-c'x}. \tag{17}$$

A Equação (17) será positiva para $c > 0$ e $c' > 0$ e, portanto, haverá uma relação crescente entre os vizinhos indiretos. Caso exista uma relação crescente apenas entre um dos vizinhos diretos, $c < 0$ ou $c' < 0$, haverá uma relação decrescente entre os vizinhos indiretos. Se existir uma relação crescente para ambos os vizinhos diretos, a relação também será crescente para os vizinhos indiretos.

Analisando agora o modelo II, teremos:

$$y = Az^b, \tag{18}$$

$$z = A'x^{b'}, \tag{19}$$

dessa maneira, a relação entre y e z será da forma

$$y = A(A')^b x^{b'b} = A''x^{b''}, \tag{20}$$

onde definimos $A'' = A(A')^b$ e $b'' = b'b$.

Para este modelo composto, poderemos ter uma relação crescente ou decrescente entre os vizinhos indiretos, dependendo do valor das constantes b e b'. Se os vizinhos diretos possuírem o mesmo tipo de relação entre eles, crescente ou decrescente, ou seja, $b > 0$ e $b' > 0$ ou $b < 0$ e $b' < 0$, respectivamente, teremos $b'' > 0$ e, por conseguinte, existirá uma relação crescente entre y e x, vizinhos indiretos. Se a relação entre os vizinhos diretos não for a mesma, teremos um expoente positivo e outro negativo, de forma que obteremos $b'' < 0$, uma relação decrescente entre vizinhos indiretos.

Para o caso mais geral, dado pelo modelo III (Equação (5)), teremos as seguintes relações entre vizinhos hierárquicos:

$$y = Az^b e^{-cz}, \tag{21}$$

$$z = A'x^{b'}e^{-c'x}. \tag{22}$$

A partir das eqs. (21) e (22), podemos estabelecer uma relação entre vizinhos indiretos y e x, nos mesmos moldes da Equação (5):

$$\begin{aligned} y &= A(A'x^{b'}e^{-c'x})^b e^{-c(A'x^{b'}e^{-c'x})} \\ &= A(A')^b x^{b'b} e^{-(c'bx+cA'x^{b'}e^{-c'x})} \\ &= A''x^{b''}e^{-(c''x+c'''x^{b'}e^{-c'x})}, \end{aligned} \tag{23}$$

onde definimos $A'' = A(A')^b$, $b'' = b'b$, $c'' = c'b$, $c''' = cA'$. O uso do modelo mais geral acaba levando a uma relação intrincada entre as variáveis. A derivada de y será dada por

$$\begin{aligned} \mathrm{d}y/\mathrm{d}x = -(&A(c'x - b')e^{A'cx^{b'}(-e^{-c'x})-c'x} \\ &(A'x^{b'}e^{-c'x})^b(be^{c'x} - A'cx^{b'}))/x. \end{aligned} \tag{24}$$

A relação entre x e y será decrescente se,

- $c'x - b' > 0$ e $be^{c'x} - A'cx^{b'} > 0$, ou
- $c'x - b' < 0$ e $be^{c'x} - A'cx^{b'} < 0$.

Em geral, os trabalhos que analisam a Lei de Menzerath utilizam os modelos simplificados e limitam suas análises a um mesmo sistema hierárquico, ainda que possa haver alguma sobreposição e interação entre sistemas hierárquicos distintos (Pike, 1967). Os modelos analisados buscam descrever a relação entre construtos e constituintes, sendo regidos por determinadas formulações matemáticas que utilizam constantes a serem determinadas ao ajustar o modelo aos dados observados. Entretanto, ainda não é claro qual é a relação entre as constantes e a interpretação delas sob a ótica da teoria da linguagem (Prün, 1994; Köhler, 1989; Altmann &

Schwibbe, 1989). Existem, entretanto, regiões de valores que aparentemente possuem relação com o nível linguístico de análise, havendo a formação de grupos quando observamos os parâmetros dos modelos (Cramer, 2005).

É importante ressaltar que, do ponto de vista linguístico, quando analisamos tais construtos, constituintes e suas relações, estamos diante de conjuntos e relações, de certa forma, difusos, sobretudo quando hierarquias lexicais e fonológicas são analisadas concomitantemente. As unidades linguísticas e a forma como seus sistemas hierárquicos se relacionam variam de uma língua para outra. Os modelos aqui propostos podem se adequar melhor ou pior a cada caso, não podendo ser considerados como uma forma de abarcar todas as nuances de uma língua.

2. Abordagem deste trabalho

Para realizar a análise da Lei de Menzerath em uma língua, deve-se, inicialmente, definir sob qual nível será realizada a análise. Este trabalho atém-se às unidades: palavras, sílabas e fonemas. Em uma análise fonológica, cada uma dessas unidades encontra-se em um nível hierárquico distinto da língua. Ainda que seja possível realizar análises no nível morfológico (Altmann, 1980; Gerlach, 1982; Polikarpov, 2000b; Krott, 1996), ou mesmo considerando a taxa de elocução (Menzerath, 1954), iremos nos ater a palavras, sílabas e fonemas, unidades essas decorrentes do nível fonológico, que estão acessíveis no Corpus ABG e que são escopo deste trabalho.

Embora uma análise simplista conceba a palavra como uma unidade situada entre dois espaços em branco, do ponto de vista linguístico, o conceito de palavra ainda não é claramente definido. Assumimos, aqui, a definição de palavra empregada por Cristófaro-Silva (2011, p. 169), apresentada no *Dicionário de Fonética e Fonologia*, a qual consiste em uma "unidade linguística que agrega som e significado em uma unidade. Pode ser compreendido [*sic*] como a menor unidade de significado em uma língua". Para Coulmas (2002, p. 38), "palavras são unidades na fronteira entre morfologia e sintaxe, possuindo importante função ao levar concomitantemente informação semântica e sintática, estando assim sujeitas à variação. Em algumas línguas, palavras parecem ser mais bem definidas e estáveis que em outras. A estrutura que constitui as palavras depende das características tipológicas das línguas".

Segundo Martinet (1978), as palavras podem ser segmentadas, do ponto de vista morfológico, em morfemas, que consistem na menor unidade linguística portadora de significado, e, do ponto de vista fonológico, em sílabas e em fonemas, sendo estes as menores unidades distintivas de significado. Os fonemas, dependendo do contexto em que ocorrem, podem ser expressos de diferentes formas, sendo estas chamadas de fones. Uma elocução qualquer pode ser descrita por uma sequência de fones que convencionalmente são representados graficamente através de um alfabeto fonético.

De modo semelhante à palavra, a definição de sílaba ainda é alvo de inúmeros debates na literatura linguística. Steriade (2002, p. 1) define-a como "uma sequência de segmentos agrupados em torno de uma vogal obrigatória ou de um elemento vocálico (silábico)", mas tal definição não é consensual. De forma que, para alguns linguistas, a sílaba é elemento central na teoria fonológica (Selkirk, 1982), enquanto para outros é um construto teórico desnecessário (Köhler, 1966). Não adentraremos aqui nesse embate teórico, tendo em vista que o corpus utilizado neste estudo já apresenta silabificação proveniente de um silabificador automático pautado na escala de sonoridade (Selkirk, 1984). Independente do método de silabificação empregado, seja por meio do dicionário, seja pelo silabificador automático, não se espera grandes discrepâncias que venham a prejudicar a análise realizada (Marchand et al., 2009).

No âmbito da escrita, as unidades da língua (fonemas, sílabas e palavras) usualmente são representadas por meio do seu sistema ortográfico, o qual requer a dissecação do fluxo da fala em partes distinguíveis. Para Coulmas (2002, p. 151), "todo sistema de escrita mapeia um sistema linguístico, incorpora e exibe visivelmente a dissecação das unidades da língua e, portanto, realiza uma análise linguística".

Embora os corpora linguísticos, em geral, apresentem apenas a representação ortográfica da palavra, o Corpus ABG, base de dados desta pesquisa, foi selecionado por já dispor da transcrição fonêmica das palavras, bem como de outras informações fonológicas. Deve-se salientar que, por questões metodológicas, foram utilizados caracteres do teclado para representar os fonemas, o que não significa que eles sejam grafemas em si. São, na verdade, apenas símbolos diferentes do próprio IPA (*International Phonetic Alphabet*), mas que têm a mesma função: representar e expressar os sons das línguas.

2.1. Tokenização

O primeiro problema com o qual se deve lidar ao trabalhar com um corpus escrito é o da *tokenização*. *Tokens* são realizações de uma determinada unidade, isto é, toda ocorrência de uma sequência idêntica de caracteres (unidades) representa uma realização de um tipo (entidade abstrata) na forma de um *token* (instância concreta). Estabelecer, a partir de uma sequência de caracteres, a sua divisão, eliminando os caracteres irrelevantes, como os de pontuação, não é uma tarefa trivial. A simples remoção da pontuação e a utilização dos espaços em branco para delimitar os *tokens* geram alguns resultados indesejados, como em palavras com apóstrofe e/ou com hífen. Por exemplo, como devemos lidar com as sequências: "*ex-presidente*", "*dona-de-casa*", "*arqui-inimigo*" e "*copo-d'água*"? Deveríamos separá-los e gerar os *tokens*: "*ex*", "*presidente*", "*dona*", "*de*", "*casa*", "*arqui*", "*inimigo*", "*copo*", "*d*" e "*água*"? Ou devemos tratar a sequência como um único *token*: "*ex-presidente*", "*dona-de-casa*", "*arqui-inimigo*" e "*copo-d'água*"? Observe que, ao utilizar o hífen e o apóstrofe como caracteres que delimitam a fronteira de palavra, geram-se palavras malformadas, como *arqui* e *d*. Para os casos de hífen, de forma geral, uma estratégia aparentemente mais apropriada seria assumi-las como palavras compostas e, conforme a análise linguística a ser realizada, considerá-las ou não no corpus em investigação - abordagem adotada neste estudo, uma vez que essa foi a abordagem utilizada para criar o Corpus ABG (Benevides & Guide, 2017).

2.2. Frequência de Ocorrência

Vários estudos mostram que o efeito de frequência é ubíquo nas áreas ligadas à cognição e ao comportamento humano (Sikström, 2002; Nosofsky, 1988; Ellis, 2015), sobretudo, na aquisição de linguagem (Ambridge et al., 2015; Ellis, 2002), no processamento de linguagem (Ellis, 2002) e nas mudanças linguísticas (Bybee, 2010). Especificamente, a frequência de ocorrência pode ser utilizada como um instrumento de análise quantitativa de vários aspectos linguísticos.[5] Segundo Bybee (2001), frequência de ocorrência consiste na quantidade de vezes que uma unidade, em geral uma palavra, ocorre em determinado corpus ou texto. Para calculá-la, deve-se contabilizar quantas vezes cada uma delas aparece em uma dada amostra. Ela vem sendo incorporada na descrição e na análise de diversos estudos linguísticos, tanto no estudo de língua adulta, como de aquisição de linguagem (Bybee, 1995, 2001; Pierrehumbert, 2003; Jarosz et al., 2016).

Na Linguística Quantitativa, diversas leis examinam o comportamento da frequência de ocorrência de tipos e sua relação com propriedades linguísticas. A mais conhecida é a Lei de Zipf, já mencionada na Seção 1. Zipf (1935, 1949) propôs, no âmago de sua teoria, o *Princípio do Esforço Mínimo*, sendo este responsável por explicar as observações que obteve sobre a frequência de ocorrência de palavras.

As unidades linguísticas, como fonemas, letras, sílabas, morfemas e palavras, podem ser estudadas através de suas frequências de ocorrência. Tal análise é possível até mesmo em estruturas que ocupam nível hierárquico mais alto. A frequência de ocorrência pode ser utilizada para inferir, por exemplo, sobre a distribuição subjacente da fonte que produz uma sequência de símbolos observada. Busca-se, assim, deduzir e quantificar a informação produzida por uma fonte. Esta também é usada em psicolinguística para explicar o fenômeno de recuperação lexical, considerado um dos processos centrais do processamento da linguagem, o qual consiste na transformação de um conceito abstrato em uma realização concreta, a elocução da palavra (Gleason & Ratner, 1998; Marantz, 2015). A frequência de ocorrência também é utilizada em outras áreas como no estudo do aprendizado, da organização e do desenvolvimento de uma língua (Phillips, 2006; Lieberman et al., 2007; Bybee, 2007, 2015).

2.3. Base de dados para o presente trabalho

Este trabalho verificou a aplicação da Lei de Menzerath em um corpus do português, analisando a relação entre o número de sílabas das palavras e o comprimento médio das sílabas, em fonemas. Para tanto, fez-se necessária uma base de dados que possuísse o número de sílabas e o número de fonemas para cada palavra, além de sua frequência de ocorrência no corpus. Utilizou-se como base de dados o Corpus ABG (Benevides & Guide, 2017).[6] Este é um corpus linguístico do português brasileiro (PB) que pode ser utilizado como fonte de extração de dados fonológicos, contendo, para cada palavra, sua frequência de ocorrência (oral e escrita), transcrição fonológica,

[5] Aprendizado, memorização, percepção, recuperação lexical, regularização, redução fonética, dentro muitos outros. Estes são apenas alguns exemplos da relevância e da atuação da frequência de ocorrência.

[6] O Corpus ABG está disponível no sítio `https://github.com/SauronGuide/corpusABG`.

codificação da estrutura da sílaba,[7] além de categoria morfológica, lema e acentuação. Mais informações sobre a criação, a estruturação e a utilização do corpus podem ser obtidas em Benevides & Guide (2017).

2.4. Lei de Menzerath no português brasileiro

São poucos os trabalhos que analisam a Lei de Menzerath no português. Por exemplo, Coloma (2015) faz um paralelo com 50 línguas distintas, dentre elas o português. O foco principal deste trabalho é comparar o ajuste de dois modelos distintos: o tradicional, de Menzerath-Altmann (Altmann, 1980), e o modelo hiperbólico proposto por Milička (2014). Utilizou, para isso, o conto "O Vento Norte e o Sol" de Esopo, em suas diversas traduções contidas no *Handbook of the International Phonetic Association*. Este livro é um guia sobre como utilizar o alfabeto fonético IPA e assim apresenta, além dos textos traduzidos, suas transcrições fonéticas. A versão traduzida para o português possui apenas 8 orações, 98 palavras e 380 fonemas. Ao analisar a relação entre fonemas por palavra e palavras por oração nas 50 línguas, Coloma (2015) concluiu que ambos os modelos apresentam igualmente bem a correlação negativa visualizada entre número de fonemas por palavra e palavras por oração.

Rothe-Neves et al. (2018), por sua vez, analisou a relação entre a duração média das sílabas em elocuções e o número de sílabas nas sentenças, a partir da elocução de 20 sujeitos para 40 sentenças, que variam de 2 a 29 sílabas. O trabalho buscou investigar a tendência geral de compressão da duração dos sons da fala em função do número de sons em uma sentença, independente do seu conteúdo linguístico. Para tanto, Rothe-Neves et al. (2018) ajustaram o modelo II (Equação (4)) para os dados de cada um dos sujeitos e também para o conjunto de todos os dados. Suas observações corroboraram a hipótese de independência entre os parâmetros A e b (Milička, 2014; Rothe-Neves et al., 2018), em oposição à argumentação de que tais parâmetros seriam dependentes da língua (Cramer, 2005; Kelih, 2010; Kułacka, 2010); e, sobretudo, corroboraram a Lei de Menzerath ao observar uma tendência de encurtamento na duração das sílabas conforme o alongamento dos enunciados.

Embora ambos os trabalhos sugiram que a Lei de Menzerath descreve adequadamente o comportamento de encurtamento do constituinte em relação ao construto, eles analisam dados em níveis hierárquicos linguísticos distintos, os quais também são diferentes daqueles que são abordados neste trabalho, e, principalmente, possuem uma base de dados de pequeno volume, o que compromete o estabelecimento de conclusões robustas. Diante dessa carência, este trabalho apresenta os resultados da aplicação da Lei de Menzerath em dados do português brasileiro, a partir do Corpus ABG, que "contabiliza 3.616.625 ocorrências de palavras e 92.602 tipos de palavras, sendo que 1.938.805 ocorrências são provenientes dos corpora de fala e 1.676.820 ocorrências dos corpora escritos" (Benevides & Guide, 2017, p. 1).

O corpus foi submetido ao *script* de tratamento e de criação do banco de dados desta pesquisa. Como o corpus já apresenta dados de transcrição fonológica e de estrutura silábica, o número de fonemas e o número de sílabas das palavras foram extraídos diretamente do corpus. Ao final, foi criada uma base de dados com informações de frequência de ocorrência, número de fonemas e número de sílabas para cada uma das palavras do Corpus ABG.

Os dados obtidos, a partir da relação entre o número de fonemas e o número de sílabas das palavras do PB, são expostos nas Figuras 1 e 2. Como o número de sílabas e o número de palavras são valores inteiros, o comprimento médio das sílabas só poderá assumir alguns valores fracionários possíveis. Têm-se, assim, apenas alguns níveis discretos de comprimento médio, o que resultou em uma grande sobreposição dos dados. Os gráficos de densidade expostos na Figura 2 apresentam, portanto, uma interpolação dos valores. Para representar graficamente a sobreposição de dados, nas Figuras 1 e 3, utilizou-se o tamanho dos círculos para representar o número de palavras encontradas com determinado número de sílabas e de fonemas. A Figura 1 apresenta também a distribuição marginal do número de fonemas e do número de sílabas no Corpus ABG.

O ajuste dos mínimos quadrados[8] aponta para uma relação decrescente entre o número de sílabas das palavras e o tamanho médio das sílabas, em número de fonemas. Os resultados estatísticos exibidos na Tabela 4 evidenciam que

[7] A codificação utilizada consistiu em: C para consoante, V para vogal, G para glide e S para as fricativas alveolares em posição de coda em final de palavra. Esta codificação foi empregada em decorrência da invisibilidade de tal segmento às regras acentuais (Bisol, 1994; Lee, 1995).

[8] O método dos mínimos quadrados aproxima um dado modelo (isto é, encontra o conjunto de parâmetros), buscando a solução que minimize a soma dos quadrados do resíduo.

palavra	frequência	transcrição fonética*	número de fones	tipo silábico	número de sílabas
de	125749	de	2	CV	1
que	116882	ke	2	CV	1
a	102779	a	1	V	1
o	91246	o	1	V	1
e	87868	e	1	V	1
é	61550	3	1	V	1
eu	46558	eW	2	VG†	1
do	46538	do	2	CV	1
não	43919	nAW	3	CVG	1
da	40205	da	2	CV	1
em	37053	EJ	2	VG	1
um	35188	U	1	V	1
você	29544	vo-se	4	CV-CV	2
na	29447	na	2	CV	1
com	29013	kO	2	CV	1
uma	28659	u-ma	3	V-CV	2
no	28427	no	2	CV	1
né	25291	n5	2	CV	1
assim	24666	a-sI	3	V-CV	2

* transcrição fonológica adotada no Corpus ABG Benevides & Guide (2017)
† Glide (ou semivogal).

Tabela 2: Exemplificação de alguns dados contidos no Corpus ABG.

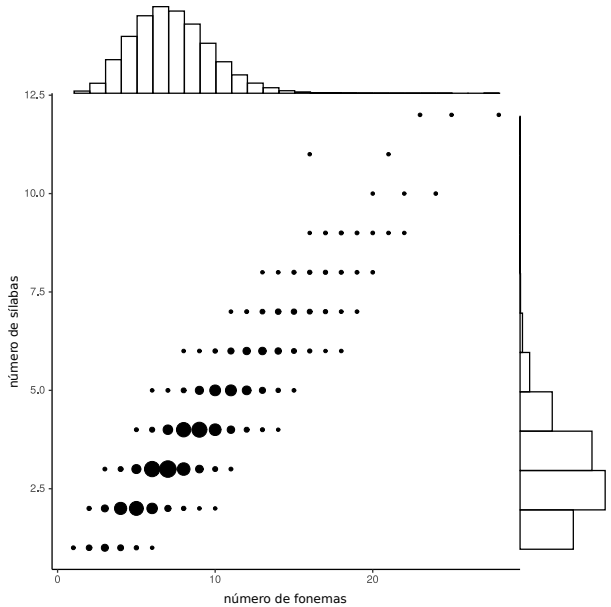
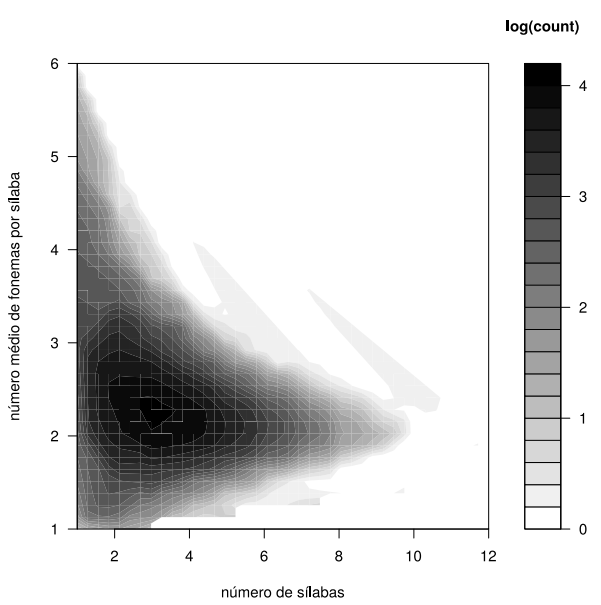


Figura 1: Relação entre o número de sílabas e o número de fonemas nas palavras do PB. A figura apresenta os círculos com tamanho proporcional ao número de palavras com cada uma das relações encontradas entre número de fonemas e sílabas. Nas laterais são apresentados os histogramas do número de sílabas e do número de fonemas.

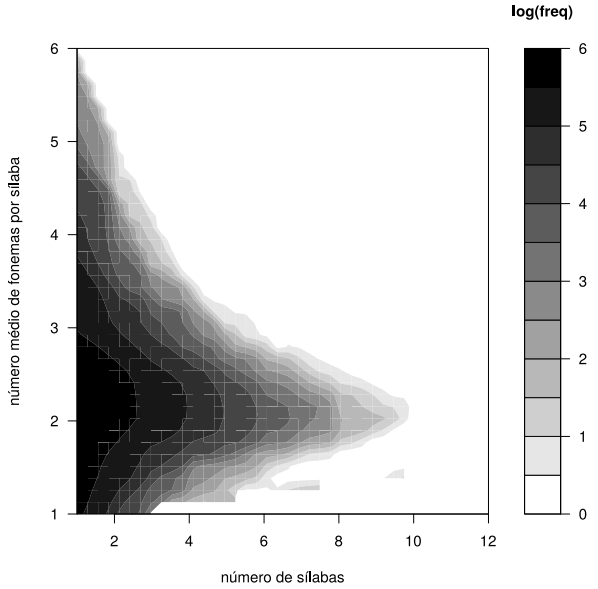
é possível descartar com segurança a hipótese nula: a hipótese de que não existe relação entre o número de sílaba e o tamanho médio das sílabas.[9] Em outros termos, observa-se que existe uma tendência a se utilizar sílabas menores em palavras com mais sílabas. Os modelos obtidos aqui também foram comparados utilizando ANOVA, sendo que os resultados apresentados na Tabela 3 evidenciam que o modelo III é mais explicativo para os dados observados.

Pelo gráfico dos resíduos, exibido nas Figuras 4a, 4c e 4e (veja Apêndice A), constata-se que, em média, o modelo é adequado, visto que os resíduos encontram-se centrados em zero, o que pode ser observado pela linha média dos resíduos ao longo do eixo da variável independente (número de sílabas). Note também que o resíduo não é sistematicamente grande ou pequeno em diferentes regiões, o resíduo está simetricamente distribuído em torno do zero, não sendo assim possível estabelecer alguma predição sobre os resíduos a partir da variável independente. Conclui-se, dessa maneira, que não há in-

[9] O coeficiente de determinação aparentemente é baixo, quando comparado aos valores observados na literatura. Entretanto, devemos observar que utilizamos aqui um volume de dados muito maior que o usual e, ainda mais, o modelo é determinado para todos os dados da amostra, enquanto na literatura obtém-se valores altos de R^2 tão somente por ajustarem o modelo às médias.



(a) Considerando a frequência de tipo das palavras.



(b) Considerando a frequência de ocorrência de cada palavra.

Figura 2: Relação entre o número de sílabas e o número médio de fonemas por sílabas observada nas palavras do Corpus ABG.[10]

formação explanatória do modelo sendo perdida e, consequentemente, sendo observada através dos resíduos.

Observa-se, porém, que a variância do resíduo cresce com a variável independente, indicando a presença de heterocedasticidade.[11] Isto pode invalidar os testes estatísticos de significância, pois estes pressupõem que os erros na modelagem são descorrelacionados e uniformes. A ausência de homoscedasticidade pode também indicar a existência de não-linearidade nos dados. Heteroscedasticidade usualmente ocorre quando há uma grande diferença no tamanho das observações. No caso em questão, quanto maior o número de sílabas de uma palavra, maior a variabilidade das sílabas usadas para construir essa palavra. Quando a palavra é pequena, a variabilidade é menor. A existência de heteroscedasticidade implica que o teorema de Gauss-Markov[12] não é válido para o caso em questão, de forma que é possível que o estimador linear de mínimos quadrados ordinário[13] utilizado não seja a melhor escolha, em termos de prover a menor variância dentre todos estimadores não polarizados. Com isso, observa-se que não há correlação entre o valor ajustado pelo modelo e o resíduo, como era esperado, e, ainda, que há heteroscedasticidade nos modelos propostos. Essencialmente, qualquer modelo é impreciso, desta forma, dentre as inúmeras opções de modelos que poderiam ser propostas, busca-se sempre o mais simples, que seja capaz de explicar os dados observados (princípio da Navalha de Occam). Talvez os modelos propostos possam ser melhorados se for acrescentada uma nova variável, por eles não abarcada, ou ainda, uma melhor estimativa das variáveis envolvidas poderia ser suficiente e, com isso, o efeito da heteroscedasticidade poderia ser diminuído, mantendo as demais qualidades dos modelos utilizados. Trabalhos como o de van Heuven et al. (2014) mostram a importância de

[10] Conforme exposto na Figura 2, é possível encontrar no Corpus ABG, e no português, palavras simples com 6, 7 e até mesmo 8 sílabas, como *pro.gres.si.va.men.te*, *pro.ble.ma.ti.za.re.mos* e *tra.di.ci.o.na.lís.si.mo*, ainda que elas sejam poucas. Há, ainda, palavras com extensão igual ou superior a elas, que são, em geral, palavras compostas, como *ex-pro.cu.ra.dor-ge.ral*, *la.ti.no-a.me.ri.ca.na* e *pre.si.dên.cia-e.xe.cu.ti.va*.

[11] Diz-se que há heterocedasticidade em um conjunto de variáveis aleatórias quando existe subpopulações com diferentes variabilidades. Se a variabilidade de uma variável não se mantém igual ao longo da extensão de uma segunda variável que a prediz, então, diz-se que há heterocedasticidade. Algumas possíveis causas da heterocedasticidade são: a própria natureza de algumas variáveis que apresentam tendência à heterocedasticidade, a existência de valores extremos e as falhas na especificação do modelo. A existência de heterocedasticidade pode comprometer alguns testes de significância em uma análise de regressão.

[12] O teorema de Gauss-Markov estipula que, em um modelo de regressão linear, sob certas condições, o estimador linear não polarizado com menor variância é dado pelo estimador dos mínimos quadrados ordinário, se existir.

[13] O estimador de mínimos quadrados ordinário é aquele que utiliza a soma dos quadrados das diferenças entre a variável dependente observada e o valor predito pelo modelo linear, ou seja, $\sum_x (y(x) - \hat{y}(x))^2$, onde x é a variável independente, y a variável dependente e $\hat{y}$ o valor predito pelo modelo.

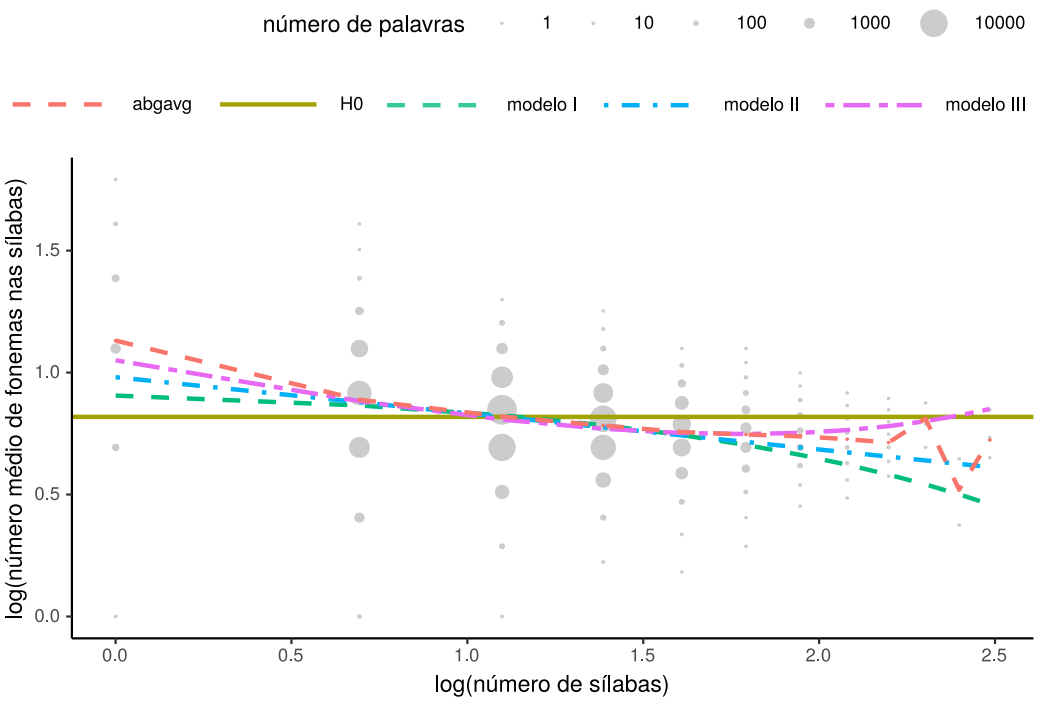


Figura 3: Relação entre o número de sílabas e o número médio de fonemas nas palavras do PB. São traçados os modelos I, II e III obtidos pelo ajuste de modelo linear a partir da base de dados ABG e a hipótese nula como referência.

	Res.Df	RSS	Df	Soma Qd.	F	Pr(>F)
I	92602	2096,3				
III	92601	2020,5	1	75,829	3475,3	$< 2{,}2 \times 10^{-16}$ *
II	92602	2043,7				
III	92601	2020,5	1	23,207	1063,6	$< 2{,}2 \times 10^{-16}$ *
Nota:						$^{*}p < 0{,}001$

Tabela 3: Análise de Variância (ANOVA) comparando modelo I com modelo III e comparando modelo II com modelo III para os dados do Corpus ABG.

uma boa estimação de variáveis como frequência de ocorrência, comprimento e similaridade de palavras, uma vez que a melhor estimação dessas variáveis é capaz de agregar maior explicação sobre a variância dos dados do que a inclusão de novas variáveis.

O gráfico Q-Q normal (*normal quantile-quantile plot*), apresentado nas Figuras 4b, 4d e 4f (veja Apêndice A), revela um desvio em direção a uma distribuição com cauda longa, o que pode ser constatado pela observação de valores mais extremos do que o esperado, caso os dados fossem provenientes de uma distribuição normal. Avaliar a normalidade dos dados é importante, pois muitos procedimentos de inferência estatística presumem que as amostras, resultantes de um conjunto fixo de valores de uma variável dependente, provenham de uma distribuição normal. Violações severas dessa suposição podem levar a resultados errados em valor de p e de intervalo de confiança. Embora, no caso em questão, constate-se a não normalidade dos resíduos, ainda assim a normalidade é uma suposição razoável.

2.5. A Lei de Menzerath frente aos dados do português

No presente trabalho, conforme apresentado na seção 2.4, buscou-se verificar a veracidade da Lei de Menzerath frente aos dados de um corpus linguístico do português brasileiro. A Lei prediz que o número de fonemas nas sílabas diminui à medida que o número de sílabas da palavra aumenta. A fim de testar essa predição, utilizou-se como fonte de dados o Corpus ABG.

Observa-se, de maneira geral, que o número de fonemas diminui à medida que o número de sílabas aumenta, o que corrobora a Lei de Menzerath. Tal proporcionalidade é claramente vi-

	Variável dependente:		
	log(comprimento médio em fonemas)		
	(I)	(II)	(III)
número de sílabas	0,041*		-0,054*
	(0,0004)		(0,002)
log(número de sílabas)		−0,148*	-0,319*
		(0,001)	(0,005)
Constante	0,946*	0,981*	0,996*
	(0,002)	(0,002)	(0,002)
Observações	92.604	92.604	92.604
R^2	0,091	0,114	0,124
R^2 Ajustado	0,091	0,114	0,124
Erro padrão residual	0,150 (df = 92.602)	0,149 (df = 92.602)	0,148 (df = 92.601)
Estatística F	9.281,210* (df = 1; 92.602)	11.904,500* (df = 1; 92.602)	6.552,366* (df = 2; 92.601)
Nota:			*$p < 0,01$

Tabela 4: Resultados do ajuste dos modelos lineares para os dados do Corpus ABG.

sualizada na Figura 2a, a qual expressa uma maior densidade de tipos de palavras de até 4 sílabas com até 3 fonemas (por sílaba). É interessante notar que a redução do número de fonemas por sílaba conduz, consequentemente, a predileção por sílabas mais simples, em geral, CV, V e CVC. Essa preferência é esperada, tendo em vista que, segundo Crystal (1988), as duas primeiras sílabas são universais nas línguas naturais. No português, as três figuram entre as mais frequentes, com índices, respectivamente, de 192.532, 26.907 e 24.055 em termos de frequência de tipo,[14] segundo dados do Corpus ABG (Benevides & Guide, 2017).

Além de nos mostrar que, quanto maior a palavra, menor a quantidade de fonemas por sílaba, as distribuições expostas nas Figuras 2 e 3, a partir da Lei de Menzerath, demonstram que, quanto maior a estrutura da palavra, menor a quantidade de tipos de palavra. Isto é, as estruturas de palavras mais frequentes da língua tendem a ter estruturas silábicas mais simples (CV). Tal afirmativa é visualizada no Corpus ABG, o qual apresenta as seguintes estruturas mais frequentes: CV-CV-CV (5.243 tipos), CV-CV (3.553), CV-CV-CV-CV (3.229), CV-CV-CVS (1.671) e V-CV-CV-CV (1.625). Note que a única estrutura com ramificação de rima ocupa a quarta posição e ainda assim é preenchida por *s*, marcador, em geral, de plural, o qual é tido por diversas propostas como invisível a alguns fenômenos fonológicos, como o acento (Massini-Cagliari, 1992; Bisol, 1994). Tal correlação poderia ser estendida, no caso do português, a frequência das palavras, tendo em vista que, segundo Araújo et al. (2007), os pentassílabos figuram entre as palavras com índice de frequência mais raro da língua, com 41,5%, em comparação a 21% dos trissílabos, por exemplo. A raridade de frequência das palavras tende a aumentar conforme aumenta as suas extensões, como demonstra a Figura 2b.

Diante disso, destaca-se, no presente artigo, a aplicabilidade da Lei de Menzerath para o corpus do português, o que permite tecer análises quantitativas relevantes com relação à estrutura da palavra e das sílabas. A Linguística Quantitativa, dessa maneira, permite realizar descrições essenciais às análises fonológicas em geral.

[14] Assume-se, aqui, a concepção de frequência de tipo de (Bybee, 2001, p. 10), segundo a qual "frequência de tipo refere-se à frequência de dicionário de um padrão específico"; neste caso, a quantidade de tipos de palavras que possui semelhante estrutura silábica.

3. Conclusão

O presente estudo não fornece resposta a todas as questões em aberto que permeiam a Lei de Menzerath-Altmann na linguagem. Buscou-se aqui analisar os modelos frente aos dados do português brasileiro; a partir dos quais verificamos a existência de uma tendência à diminuição do número de constituintes na composição de construtos maiores, corroborando assim a Lei de Menzerath. No contexto de uma teoria geral da comunicação, argumenta-se que mecanismos de percepção e de cognição devam ser considerados para explicar essa observação. Zipf (1935, 1949) usa o *Princípio do Esforço Mínimo* como fundamento para suas observações. Mais tarde, Köhler (1989) utiliza um modelo de processamento de linguagem para fundamentar as observações feitas através da Lei de Menzerath. O processamento de linguagem é sequencial, ao menos no nível mais baixo, uma vez que a fala se realiza sequencialmente ao longo do tempo, através da sucessão linear de perturbações acústicas com características que se modificam no decorrer do tempo. Além disso, Köhler (1989) utiliza como argumento a capacidade finita de processamento e memória em tarefas cognitivas. Esse raciocínio leva à conclusão de que construtos mais complexos e mais longos necessitam utilizar-se de constituintes menores e mais simples. Desta forma, não apenas se observa a tendência à utilização de constituintes mais simples na composição de construtos mais complexos, como também uma menor variabilidade. Já na construção de construtos mais simples, a variabilidade dos constituintes é maior. Tal argumentação entra em consonância com o limite da memória de curta duração (Miller, 1956), segundo o qual o número de objetos que em média uma pessoa é capaz de manter na memória de trabalho é de 7 ± 2, a chamada Lei de Miller. As tarefas analisadas por Miller (1956) mostram que há queda na performance à medida que o número de estímulos diferentes (variando apenas um dos atributos) aumenta para além de cinco ou seis. A capacidade de memória imediata, o presente psicológico, é dito ter uma duração de 1,5 a 3 segundos, operando principalmente no nível de processamento de sentenças. Além disso, para palavras em que a relação entre o número de sílaba e o número de fonemas se mantém constante, a duração temporal é fator determinante para a recuperação de palavras, apresentando melhor desempenho aquelas de curta duração (Baddeley et al., 1975). Essa análise sugere a necessidade de se analisar também a duração de palavras e de sentenças em elocuções para verificar se a Lei de Menzerath também se faz presente.

Nos modelos aqui contemplados, através da análise dos resíduos, é possível verificar que não são satisfeitas as considerações de heteroscedasticidade e de normalidade dos resíduos. Isto indica que possivelmente existem fatores que não foram abarcados pelo modelo proposto, ou que é necessário obter medidas mais acuradas das variáveis envolvidas, ou ainda que é necessário modificar o modelo, utilizando algum termo de outra sorte para conseguir um melhor ajuste. A frequência de ocorrência de palavras pode ser utilizada para ponderar a relação entre o número de sílabas em palavras e o comprimento médio das sílabas em fonemas. A influência dessa nova variável pode alterar a relação observada entre comprimento médio das sílabas e o comprimento das palavras. Essa variável, ainda não considerada, pode inclusive ser fator causador da heterocedasticidade observada. Outros fatores importantes também poderiam ser considerados, como a frequência de ocorrência de sequências de palavras, a complexidade articulatória, a duração e a distintividade linguística de fonemas, sílabas e palavras. Afinal, fatores relacionados à produção e à percepção também possuem papel importante no uso da linguagem.

Deve-se ter em mente a separação entre os três processos centrais que foram abarcados no presente trabalho: a seleção do modelo, a estimação dos parâmetros e a predição de resultados a partir do modelo e dos parâmetros dos dados. O modelo correto nunca será conhecido, mas, como ponderou Box (1976), "alguns são úteis". A análise de um problema não deve consistir na aplicação dos três passos descritos uma única vez, mas deve retornar ao passo anterior sempre que se verificar falsas suposições que foram rejeitadas nas etapas seguintes. Durante a construção e a aplicação de um modelo, muitas vezes adotam-se generalizações; esse processo pode ser perigoso, pois incertezas estão inseridas em distintas etapas.

A partir do estudo aqui depreendido sobre a relação entre construto e constituinte na comunicação restritos ao âmbito fonológico, constatase que as observações quantitativas e o tratamento matemático tornam-se necessários para a descrição de fenômenos que não podem ser representados por um arquétipo de uma determinada categoria, nem explicado por regras simbólicas estruturais, ou seja, é uma abordagem necessária para estudar a variabilidade e a imprecisão nas línguas naturais. Prefere-se lidar aqui com tendências e com preferências a lidar com relações estáveis de estruturas bem definidas. As relações dinâmicas e a variabilidade revelam mais sobre o fenômeno do que o funcionamento rígido de um sistema estrutural bem definido, uma vez que os modelos tesos e austeros culminam em uma descrição imprecisa e inconsistente. As expressões quantitativas permitem uma melhor adequação à realidade que aquelas qualitativas, permitem ainda uma análise mais fina em diferentes resoluções, uma gradação contínua de formatos representacionais que, sob outra ótica, seriam discretos e muitas vezes impossibilitariam o estabelecimento de inter-relações entre diferentes níveis de análise. A análise quantitativa da linguagem busca explorar os fundamentos, estabelecer explicações e construir uma teoria consistente com hipóteses refutáveis.

Referências

Abe, Sumiyoshi & Norikazu Suzuki. 2005. Scale-free statistics of time interval between successive earthquakes. *Physica A: Statistical Mechanics and its Applications* 350(2–4). 588–596. 10.1016/j.physa.2004.10.040.

Adamic, Lada A. & Bernardo A. Huberman. 2002. Zipf's law and the internet. *Glottometrics* 3. 143–150.

Altmann, Eduardo G. & Martin Gerlach. 2016. Statistical laws in linguistics. Em M. Degli Esposti, E.G. Altmann & F. Pachet (eds.), *Creativity and Universality in Language*, 7–26. Springer. 10.1007/978-3-319-24403-7_2.

Altmann, Gabriel. 1980. Prolegomena to Menzerath's law. *Glottometrika* 2(2). 1–10.

Altmann, Gabriel & Michael H. Schwibbe. 1989. *Das Menzerathsche Gesetz in informationsverarbeitenden systemen*. Hildesheim: Olms.

Ambridge, Ben, Evan Kidd, Caroline F. Rowland & Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language* 42(2). 239–273. 10.1017/s030500091400049x.

Andres, Jan. 2010. On a conjecture about the fractal structure of language. *Journal of Quantitative Linguistics* 17(2). 101–122. 10.1080/09296171003643189.

Araújo, Gabriel A. de, Zwinglio O. Guimarães Filho, Leonardo Oliveira & Viaro M. Eduardo. 2007. As proparoxítonas e o sistema acentual do português. Em *O acento em português: abordagens fonológicas*, 37–60. Parábola.

Baddeley, Alan D., Neil Thomson & Mary Buchanan. 1975. Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior* 14(6). 575–589. 10.1016/s0022-5371(75)80045-4.

Beckner, Clay, Nick C. Ellis, Richard Blythe, John Holland, Joan Bybee, Jinyun Ke, Morten H. Christiansen, Diane Larsen-Freeman, William Croft & Tom Schoenemann. 2010. Language is a complex adaptive system: Position paper. Em Nick C. Ellis & Diane Larsen-Freeman (eds.), *Language as a Complex Adaptive System*, 1–26. University of Michigan: Wiley-Blackwell.

Benevides, Aline De Lima & Bruno Ferrari Guide. 2017. Corpus ABG. *Texto Livre: Linguagem e Tecnologia* 10(1). 139–163. 10.17851/1983-3652.10.1.139-163.

Bisol, Leda. 1994. O acento e o pé métrico. *Letras de Hoje* 29(4). 25–36.

Boroda, Mojsej G. & Gabriel Altmann. 1991. Menzerath's law in musical texts. *Musikometrica* 3. 1–13.

Box, George E. P. 1976. Science and statistics. *Journal of the American Statistical Association* 71(356). 791–799. 10.1080/01621459.1976.10480949.

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10(5). 425–455. 10.1080/01690969508407111.

Bybee, Joan. 2001. *Phonology and language use* Cambridge Studies in Linguistics. Cambridge University Press. 10.1017/CBO9780511612886.

Bybee, Joan. 2007. *Frequency of use and the organization of language*. USA: Oxford University Press. 10.1093/acprof:oso/9780195301571.001.0001.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge University Press. 10.1017/cbo9780511750526.

Bybee, Joan. 2015. *Language change* Cambridge Textbooks in Linguistics. Cambridge University Press. 10.1017/CBO9781139096768.

Coloma, Germán. 2015. The Menzerath-Altmann law in a cross-linguistic context. *SKY Journal of Linguistics* 28. 139–159.

Coulmas, Florian. 2002. *Writing systems: An introduction to their linguistic analysis*. Cambridge University Press. 10.1017/CBO9781139164597.

Cramer, Irene. 2005. The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics* 12(1). 41–52. 10.1080/09296170500055301.

Cristófaro-Silva, Thaís. 2011. *Dicionário de fonética e fonologia*. Contexto.

Crystal, David. 1988. *Dicionário de linguística e fonética*. J. Zahar Editor.

Ellis, Nick C. 2002. Frequency effects in language processing. *Studies in Second Language Acquisition* 24(2). 143–188. 10.1017/s0272263102002024.

Ellis, Nick C. 2015. Cognitive and social aspects of learning from usage. Em *Usage-Based Perspectives on Second Language Learning*, 49–74. De Gruyter. 10.1515/9783110378528-005.

Fenk, August & Gertraud Fenk-Oczlon. 2013. Menzerath's Law and the constant flow of linguistic information. Em Reinhard Köhler & Burghard B. Rieger (eds.), *Contributions to Quantitative Linguistics: Proceedings of the First International Conference on Quantitative Linguistics*, Springer. 10.1007/978-94-011-1769-2_2.

Ferrer-i-Cancho, Ramon. 2006. On the universality of Zipf's law for word frequencies. Em P. Grzybek & R. Köhler (eds.), *Exact methods in the study of language and text. In honor of Gabriel Altmann*, 131–140. Gruyter.

Ferrer-i-Cancho, Ramon. 2017. The placement of the head that maximizes predictability. An information theoretic approach. *Glottometrics* 39. 38–71.

Ferrer-I-Cancho, Ramon & Núria Forns. 2009. The self-organization of genomes. *Complexity* 15. 34–36. 10.1002/cplx.20296.

Ferrer-i-Cancho, Ramon & Ricard V. Solé. 2002. Zipf's law and random texts. *Advances in Complex Systems* 5(1). 1–6. 10.1142/S0219525902000468.

Gabaix, Xavier. 1999. Zipf's law for cities: An explanation. *Quarterly Journal of Economics* 114(3). 739–767. 10.1162/003355399556133.

Gerlach, Rainer. 1982. Zur Überprüfung des menzerath'schen gesetzes im bereich der morphologie. *Glottometrika* 4. 95–102.

Gleason, Jean Berko & Nan Bernstein Ratner. 1998. *Psycholinguistics*. Fort Worth: Harcourt Brace College Publishers.

Glottopedia. 2019. The free encyclopedia of linguistics. http://www.glottopedia.org/.

Grzybek, Peter. 2007. *Contributions to the science of text and language: Word*

length studies and related issues Text, Speech and Language Technology. Springer. 10.1007/978-1-4020-4068-9.

Grzybek, Peter & Gabriel Altmann. 2002. Oscillation in the frequency-length relationship. *Glottometrics* 2. 97–107.

Grzybek, Peter & Ernst Stadlober. 2007. Do we have problems with arens' law? a new look at the sentence-word relation. Em Peter Grzybek & Reinhard Köhler (eds.), *Exact Methods in the Study of Language and Text*, 205–218. Walter de Gruyter. 10.1515/9783110894219.205.

Gustison, Morgan L., Stuart Semple, Ramon Ferrer i Cancho & Thore J. Bergman. 2016. Gelada vocal sequences follow menzerath's linguistic law. *Proceedings of the National Academy of Sciences of the USA* 113(19). E2750–E2758. 10.1073/pnas.1522072113.

Heaps, Harold Stanley. 1978. *Information retrieval, computational and theoretical aspects* Library and information science. USA: Academic Press.

Herdan, Gustav. 1960. *Type-token mathematics*, vol. 4 Janua linguaarum, studia memoriae Nicolai van Wijk dedicata. Series maior. Mouton & Cie.

van Heuven, Walter J. B., Pawel Mandera, Emmanuel Keuleers & Marc Brysbaert. 2014. Subtlex-UK: A new and improved word frequency database for british english. *Quarterly Journal of Experimental Psychology* 67(6). 1176–1190. 10.1080/17470218.2013.850521.

Hřebíček, Luděk. 1995. *Text levels: Language constructs, constituents and the menzerath-altmann law* Quantitative linguistics. Wissenschaftlicher Verlag Trier.

Ilari, Rodolfo. 2003. *A lingüística e o ensino da língua portuguesa*. São Paulo: Martins Fontes.

Jarosz, Gaja, Shira Calamaro & Jason Zentz. 2016. Input frequency and the acquisition of syllable structure in polish. *Language Acquisition* 24(4). 361–399. 10.1080/10489223.2016.1179743.

Jespersen, Otto. 1904. *Lehrbuch der phonetik*. Leipzig: Teubner.

Kelih, Emmerich. 2010. Parameter interpretation of the menzerath law: evidence from serbian. Em Peter Grzybek, Emmerich Kelih & Ján Mačutek (eds.), *Text and Language: Structures, Functions, Interrelations: Quantitative Perspectives*, 71–79. Praesens.

Köhler, Konrad. J. 1966. Is the syllable a phonological universal? *Journal of Linguistics* 2(2). 207–208. 10.1017/S0022226700001493.

Köhler, Reinhard. 1989. Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. Em Gabriel Altmann & Michael H. Schwibbe (eds.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, Hildesheim: Olms.

Köhler, Reinhard. 2005. Gegenstand und Arbeitsweise der Quantitativen Linguistik. Em Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative Linguistik. Ein internationales Handbuch*, 1–16. de Gruyter.

Krott, Andrea. 1996. Some remarks on the relation between word length and morpheme length. *Journal of Quantitative Linguistics* 3(1). 29–37. 10.1080/09296179608590061.

Kułacka, Agnieszka. 2010. The coefficients in the formula for the menzerath-altmann law. *Journal of Quantitative Linguistics* 17(4). 257–268. 10.1080/09296174.2010.512160.

Lee, Seung Hwa. 1995. *Morfologia e fonologia lexical do português do Brasil*. Campinas: Universidade Estadual de Campinas. Tese de Doutoramento.

Li, Wentian. 2011. Menzerath's law at the geneexon level in the human genome. *Complexity* 17(4). 49–53. 10.1002/cplx.20398.

Lieberman, Erez, Jean-Baptiste Michel, Joe Jackson, Tina Tang & Martin A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* 449(7163). 713–716. 10.1038/nature06137.

Lü, Linyuan, Zi-Ke Zhang & Tao Zhou. 2010. Zipf's law leads to heaps' law: Analyzing their relation in finite-size systems. *PLoS ONE* 5(12). e14139. 10.1371/journal.pone.0014139.

Marantz, Alec. 2015. Morphology. Em Gregory Hickok & Steven L. Small (eds.), *Neurobiology of Language*, chap. 13, 153–163. Academic Press.

Marchand, Yannick, Connie R. Adsett & Robert I. Damper. 2009. Automatic syllabification in english: A comparison of different algorithms. *Language and Speech* 52(1). 1–27. 10.1177/0023830908099881.

Martinet, André. 1978. *Elementos de lingüística geral*. Martins Fontes.

Massini-Cagliari, Gladis. 1992. *Acento e ritmo*. São Paulo: Contexto.

Menzerath, Paul. 1928. Über einige phonetische Probleme. Em *Actes du premier Congres International de Linguistes*, 104–105.

Menzerath, Paul. 1954. *Die architektonik des deutschen Wortschatzes* Phonetische Studien. F. Dümmler.

Meyer, Ernst Alfred. 1904. Zur Vokaldauer im Deutschen. Em Adolf Gotthard (ed.), *Nordiska studier tillegnade Adolf Noreen på hans 50-årsdag den 13 Mars 1904*, Wentworth Press.

Milička, Jiří. 2014. Menzerath's law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics* 21(2). 85–99. 10.1080/09296174.2014.882187.

Miller, George A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63(2). 81–97. 10.1037/h0043158.

Mitzenmacher, Michael. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1(2). 226–251. 10.1080/15427951.2004.10129088.

Nikolaou, Christoforos. 2014. Menzerath-altmann law in mammalian exons reflects the dynamics of gene structure evolution. *Computational Biology and Chemistry* 53, Part A. 134–143. 10.1016/j.compbiolchem.2014.08.018.

Nosofsky, Robert M. 1988. Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(1). 54–65. 10.1037/0278-7393.14.1.54.

Phillips, Betty. 2006. *Word frequency and lexical diffusion*. Basingstoke England New York: Palgrave Macmillan.

Pierrehumbert, Janet. 2003. Probabilistic phonology: Discrimination and robustness. Em Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probability Theory in Linguistics*, 177–228. Cambridge MA: The MIT Press.

Pike, Kenneth Lee. 1967. *Language in relation to a unified theory of the structure of human behavior* Janua Linguarum. Series Maior. De Gruyter.

Polikarpov, Anatoliy A. 2000a. Menzerath's law for morphemic structures of words: A hypothesis for the evolutionary mechanism of its arising and its testing. Em *Qualico: Quantitative Linguistics Conference*, .

Polikarpov, Anatoliy Anatolyevich. 2000b. Menzerath's law for morphemic structures of words: A hypothesis for the evolutionary mechanism of its arising and its testing. Em *Qualico: Quantitative Linguistics Conference*, .

Prün, Claudia. 1994. Validity of menzerath-altmann's law: Graphic representation of language, information processing systems and synergetic linguistics. *Journal of Quantitative Linguistics* 1(2). 148–155. 10.1080/09296179408590009.

Rothe-Neves, Rui, Bárbara Marques Bernardo & Robert Espesser. 2018. Shortening tendency for syllable duration in brazilian portuguese utterances. *Journal of Quantitative Linguistics* 25(2). 156–167. 10.1080/09296174.2017.1360172.

Roudet, Léonce. 1910. *Éléments de phonétique générale*. Welter.

Selkirk, Elisabeth. 1982. The syllable. Em Harry van der Hulst & Norval Smith (eds.), *The structure of phonological representations*, Foris.

Selkirk, Elisabeth. 1984. On the major class features and syllable theory. Em Mark Aronoff & Richard T. Oehrle (eds.), *Language Sound and Structure*, The MIT Press.

Shahzad, Khuram, Jay E. Mittenthal & Gustavo Caetano-Anollés. 2015. The organization of domains in proteins obeys Menzerath-Altmann's law of language. *BMC Systems Biology* 9. 44. 10.1186/s12918-015-0192-9.

Sikström, Sverker. 2002. Habituation during encoding of episodic memory. Em *Connectionist Models of Cognition and Perception*, 107–117. 10.1142/9789812777256_0009.

Steriade, Donca. 2002. The syllable. Em William Frawley (ed.), *International Encyclopedia of Linguistics*, Oxford University Press.

Wilde, Joachim & Michael H. Schwibbe. 1989. Organizationsformen von Erbinformation Im Hinblick auf die Menzerathsche Regel. Em Gabriel Altmann & Michael H. Schwibbe (eds.), *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, 92–107. Olms.

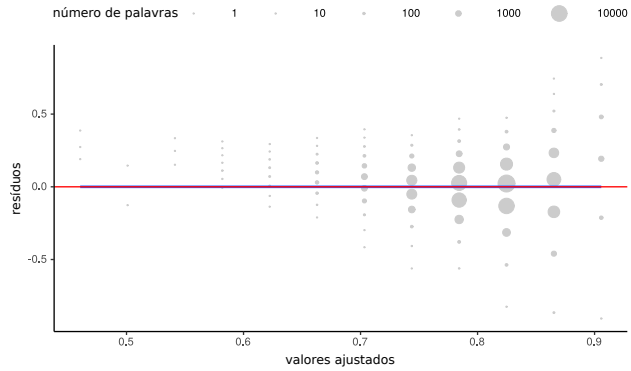
Zipf, George Kingsley. 1935. *The psycho-biology of language: an introduction to dynamic philology*. The MIT Press.

Zipf, George Kingsley. 1949. *Human behaviour and the principle of least effort: An introduction to human ecology*. Hafner Pub. Co.

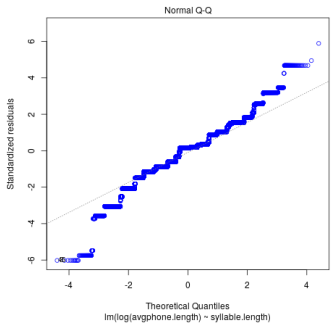
A. Dados, scripts e outros resultados

Os dados, os *scripts* desenvolvidos para a realização deste trabalhos e todos os resultados estão disponíveis em repositórios no GitHub.

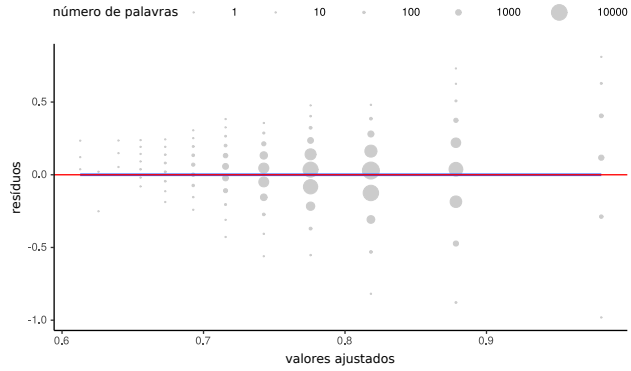
1. Corpus ABG: https://github.com/SauronGuide/corpusABG.
2. *Scripts* voltados para a área de Linguística Quantitativa e Computacional: https://github.com/leolca/clscripts.
3. *Notebook* com anotações, códigos e resultados gerados para este trabalho: https://github.com/leolca/clscripts/blob/master/menzerath_abg.ipynb.



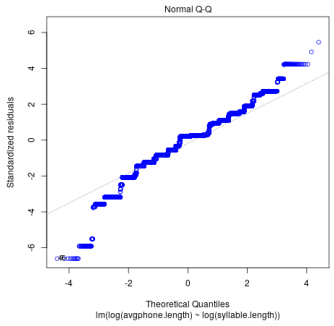
(a) Gráfico dos resíduos para o modelo I.



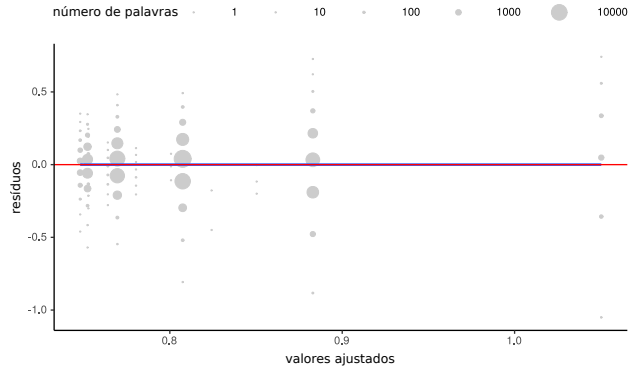
(b) Gráfico QQ para o modelo I.



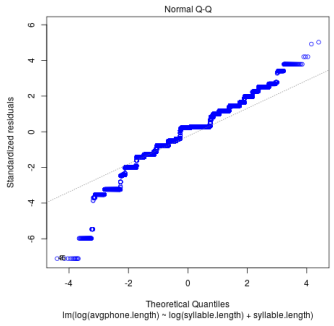
(c) Gráfico dos resíduos para o modelo II.



(d) Gráfico QQ para o modelo II.



(e) Gráfico dos resíduos para o modelo III.



(f) Gráfico QQ para o modelo III.

Figura 4: Análise dos resíduos para os modelos utilizando os dados do Corpus ABG.

Proposta recebida em Dezembro 2019 e aceite para publicação em Abril 2020.

Reescrita sentencial baseada em traços de personalidade

Personality-dependent sentence rewriting

Georges Basile Stavracas Neto
Universidade de São Paulo (USP)
georges.stavracas@gmail.com

Ivandré Paraboni
Universidade de São Paulo (USP)
ivandre@usp.br

Resumo

Sistemas de Geração de Língua Natural (GLN) são centrais para o desenvolvimento de comunicação humano-computador realista e psicologicamente plausível que não recorra ao uso de texto fixo ou predefinido, fazendo uso de uma ampla gama de estratégias para modelar alguma forma de variação estilística. Entre estas estratégias, o uso de modelos computacionais da personalidade humana emergiu como uma alternativa popular na área. Neste contexto, o presente trabalho apresenta um modelo de GLN do tipo texto-para-texto (ou reescrita sentencial) para o português que leva em conta, além da sentença a ser reescrita, informações sobre a personalidade de um locutor-alvo de interesse. Mais especificamente, o modelo transforma a sentença de entrada em outra na qual certas formas lexicais são substituídas por termos mais adequados ao tipo de personalidade-alvo fornecido. Resultados sugerem que as sentenças geradas com base em personalidade são mais próximas das que seriam produzidas por um locutor humano com as características de personalidade fornecidas do que seria possível sem acesso a essa informação, e abrem assim caminho para futuros estudos de geração de língua natural personalizada em português.

Palavras chave

geração de língua natural, personalidade

Abstract

Natural Language Generation (NLG) systems are central to the development of psychologically plausible human-computer communication that does not rely on canned text, and which makes use of a wide range of strategies to model some stylistic variation. Among these, the use of computational models of human personality has emerged as a popular alternative in the field. In this context, the present work presents a text-to-text (or sentential rewriting) GLN model for Portuguese that takes into account, in addition to the sentence to be rewritten, information about the personality of a target speaker of interest. More specifically, the model transforms the input sentence into another one in which certain lexical forms are replaced by terms more suited to a certain personality type. Results suggest that personality-based generation produces sentences that are closer to those produced by a human speaker with those personality traits than what would be possible without access to this information, thus paving the way for future studies of speaker-dependent natural language generation in Portuguese.

Keywords

natural language generation, personality

1. Introdução

Sistemas de geração de língua natural (GLN) produzem texto a partir de uma representação linguística ou não-linguística fornecida como entrada, e são centrais para o desenvolvimento de comunicação humano-computador realista e psicologicamente plausível que não recorra ao uso de texto fixo ou predefinido. Aplicações de GLN incluem, por exemplo, a geração de resumos textuais a partir de dados obtidos de unidades de terapia neonatal intensiva (Portet et al., 2009), históricos de pacientes e relatórios de enfermagem (Hunter et al., 2012; di Eugenio et al., 2014; Jordan et al., 2014), cartas personalizadas de cessação do tabagismo (Reiter et al., 2003), boletins de previsões meteorológicas (Reiter et al., 2005), diálogos e textos narrativos (Walker et al., 2011a,b), poesia (Zhang & Lapata, 2014), legendas de imagens (Karpathy & Fei-Fei, 2015; Xu et al., 2015) e muitas outras.

Podemos em princípio distinguir dois tipos de sistemas de GLN (Reiter & Dale, 2000): sistemas do tipo dados-para-texto, que recebem uma representação não linguística como entrada (e.g., dados numéricos provenientes de sensores) e produzem texto como saída (e.g., relatórios do mercado financeiro), e sistemas do tipo texto-para-texto, que recebem como entrada um texto já redigido em língua natural (possivelmente por um autor humano) e produzem uma versão modifi-

DOI: 10.21814/lm.12.1.312

cada deste mesmo texto como saída (e.g., uma versão simplificada, resumida, traduzida, adaptada estilisticamente etc.) No presente trabalho será abordado exclusivamente o segundo tipo de sistema, tratando da questão da reescrita de sentenças preexistentes com base em critérios a serem discutidos a seguir.

Sistemas de GLN podem, em princípio, produzir sempre o mesmo texto de saída para uma determinada entrada. No entanto, sistemas que visam gerar texto de maneira mais natural ou semelhante ao humano geralmente implementam uma ampla gama de estratégias para modelar variação estilística, incluindo o controle de parâmetros que afetam o tamanho ou complexidade das sentenças, uso de pausas, ênfases, pontuação, escolha lexical e muitos outros. Uma visão detalhada destas estratégias é apresentada por Mairesse (2008). Entre estas possibilidades, o uso de modelos computacionais da *personalidade humana* emergiu como uma alternativa popular na área (Mairesse & Walker, 2010).

De especial interesse para o presente trabalho, consideraremos o uso do modelo dos Cinco Grandes Fatores (CGF) da personalidade humana, ou *Big Five* (Goldberg, 1990). O modelo CGF baseia-se no pressuposto de que as diferenças de personalidade são reveladas pela maneira como os indivíduos se expressam em língua natural e, em especial, nas escolhas lexicais realizadas. Dada a sua motivação linguística, o modelo CGF tem sido aplicado a uma ampla gama de estudos tanto na interpretação (Plank & Hovy, 2015; Álvarez-Carmona et al., 2015; González-Gallardo et al., 2015) como na geração de língua natural (Mairesse & Walker, 2011; Herzig et al., 2017) e, embora haja várias outras teorias de personalidade em discussão (em especial, o modelo MBTI de Briggs Myers & Myers (2010)), a quase totalidade dos estudos em Ciência da Computação tende a ser baseado no modelo CGF.

O modelo CGF compreende cinco dimensões fundamentais da personalidade: *Extroversão*, *Agradabilidade*, *Conscienciosidade*, *Neuroticismo* e *Abertura à experiência*. Cada dimensão CGF é associada a um escore que representa a intensidade com que um indivíduo manifesta cada aspecto da personalidade. Assim, por exemplo, um escore abaixo da média (de acordo com uma população de interesse) para Extroversão indica um indivíduo introvertido, enquanto que um escore acima da média indica um indivíduo verdadeiramente extrovertido. Escores de personalidade podem ser obtidos por diversos métodos, incluindo a aplicação de questionários específicos (John et al., 2008).

Para melhor apreciar o papel das dimensões de personalidade CGF na produção humana de língua natural, considere-se a tarefa de produzir uma descrição textual simples de uma determinada cena como na figura 1, extraída da base GAPED de Dan-Glauser & Scherer (2011).

Figura 1: Um exemplo de imagem extraída da base GAPED.

Em uma situação desse tipo, diferentes locutores poderiam produzir um grande número de descrições alternativas do mesmo contexto. Por exemplo, um indivíduo com maior escore para a dimensão *Agradabilidade* pode descrever o personagem da cena como "uma linda criança", enquanto que um indivíduo de menor escore pode descrevê-lo como "um pirralho sujo", dentre muitas outras possibilidades. De forma análoga, podemos também conceber aplicações computacionais em que um sistema não apenas descreve uma imagem seguindo um padrão fixo ou pré-definido, mas o faz impondo um estilo específico ditado por uma personalidade-alvo de interesse.

A variação humana em GLN é um tópico recorrente na pesquisa na área (Viethen & Dale, 2010; Mairesse & Walker, 2011; Ferreira & Paraboni, 2014) e será também o tema do presente trabalho. De forma mais específica, propõe-se desenvolver um modelo de reescrita sentencial para o português que leva em conta, além da sentença a ser reescrita, informações sobre a personalidade de um locutor-alvo de interesse. Modelos deste tipo são potencialmente úteis, por exemplo, quando procura-se obter maior engajamento do leitor (ou usuário), como no caso de aplicações em educação, no relacionamento com clientes ou consumidores e outras.

O modelo a ser apresentado objetiva transformar uma sentença fornecida como entrada em outra na qual certas formas lexicais são substituídas por termos mais adequados ao tipo de

personalidade-alvo fornecido[1]. Neste contexto, a hipótese a ser investigada é a de que as sentenças geradas com base em personalidade sejam mais próximas das que seriam produzidas por um locutor humano com as características de personalidade fornecidas do que as que seria possível obter utilizando-se um modelo de geração de texto sem acesso à informação de personalidade.

O restante deste artigo está organizado da seguinte forma. Na Seção 2 são discutidos trabalhos relacionados à área de GLN baseada em personalidade. Na Seção 3.2 é descrito o córpus a ser empregado neste trabalho, a anotação semântica realizada e os modelos de lexicalização e reescrita sentencial propostos. Na Seção 4 estes modelos são avaliados tanto sob uma perspectiva de aprendizado de máquina como orientada à tarefa de reescrita propriamente dita. Os resultados destas duas formas de avaliação são apresentados nas Seções 4.1 e 4.2. Finalmente, a Seção 5 apresenta algumas conclusões e trabalhos futuros.

2. Trabalhos relacionados

Estudos de GLN baseado em personalidade são de modo geral escassos, e não foi possível identificar na literatura iniciativas deste tipo (ou sistemas de GLN minimamente completos) voltados para o idioma português. Nesta seção são discutidos assim alguns estudos da área de GLN baseada em personalidade dedicados ao idioma inglês.

Estudos de GLN são fortemente relacionados ao trabalho seminal de Mairesse (2008) e suas extensões. Todos estes estudos abordam a geração de texto em língua inglesa, e nenhum deles trata da tarefa de geração texto-para-texto (como no caso da reescrita sentencial tratada no presente trabalho), mas apenas da geração de dados-para-texto. Estes estudos são brevemente discutidos a seguir.

O estudo de Mairesse & Walker (2010) abordou uma ampla gama de decisões de geração que podem ser influenciadas por um perfil de personalidade-alvo. O trabalho se concentra na geração de língua natural do tipo ponta-a-ponta, ou dados para texto, apresentando um sistema de GLN configurável que produz recomendações textuais de restaurantes. O sistema resultante — chamado PERSONAGE— é treinado com base em dados anotados com informação de personalidade proveniente do modelo CGF, e os textos assim produzidos foram reconhecidos por juízes humanos como refletindo certos traços de personalidade específicos. A lexicalização no sistema PERSONAGE é realizada para cada palavra de conteúdo no texto com base em três parâmetros: a frequência do léxico, o tamanho da palavra e a força verbal (por exemplo, "sugerir" seria mais fraco do que "recomendar"). Esses parâmetros fazem uso do conhecimento obtido a partir de vários recursos lexicais on-line (por exemplo, WordNet e VERBOCEAN) e de contagens de frequência de corpus.

Diversos estudos subsequentes foram desenvolvidos como extensões do sistema PERSONAGE. Estes estudos incluem uma série de melhorias na arquitetura do sistema original, e acrescentam suporte aos outros quatro traços de personalidade do modelo CGF (além de Extroversão) (Mairesse & Walker, 2011). Além disso, a arquitetura PERSONAGE original foi também aplicada a outros domínios, incluindo a geração personalizada de fofocas (Khosmood & Walker, 2010), diálogos sobre jogos de computador (McCormick, 2012), escrita criativa (Lukin et al., 2014), narração de histórias (Bowden et al., 2016), geração de gestos (Aly & Tapus, 2016) e geração de *feedback* a clientes (Herzig et al., 2017).

No caso do idioma português, não foram identificados estudos de GLN baseados em personalidade exceto dois trabalhos prévios relacionados ao projeto atual para tratamento de outras questões de pesquisa, e ambos utilizando o próprio córpus *b*5 a ser discutido na próxima seção. Estes dois estudos, detalhados por Paraboni et al. (2017) e Lan & Paraboni (2018), que abordam a questão da relação entre personalidade e a tarefa de seleção de conteúdo na geração de expressões de referência (GER). Esta tarefa, que é um subcomponente da arquitetura de sistemas de GLN, utiliza algoritmos próprios para decidir, nos estágios iniciais da geração de texto, qual o conteúdo semântico a ser expresso na forma de descrições definidas (e.g., "a mão esquerda do menino"). Uma visão geral da área, que está fora do escopo do presente trabalho, é apresentada por van Deemter (2016).

3. Método

Nesta seção são descritos o córpus empregado neste trabalho (Seção 3.1), a preparação de dados realizada (Seção 3.2) e os modelos de lexicalização e reescrita sentencial propostos (Seção 3.3).

[1] O presente foco na questão da escolha lexical é motivado pela natureza lexical do próprio modelo CGF, mas outras direções possíveis de pesquisa são discutidas na seção 5.

3.1. O córpus b5

Neste trabalho será utilizado o córpus *b5* (Ramos et al., 2018) de textos em português brasileiro rotulados com escores de personalidade do modelo *CGF* relativos aos seus autores. O córpus foi utilizado em estudos prévios de geração de texto baseada em personalidade (Paraboni et al., 2017) e inferência de traços de personalidade a partir de textos e caracterização autoral (dos Santos et al., 2017; Hsieh et al., 2018; Silva & Paraboni, 2018b,a). Detalhes desta organização são discutidos por Ramos et al. (2018).

Os textos a serem utilizados são provenientes de duas bases (ou subcórpus) do córpus *b*5 denominados *b5-text* e *b5-caption*, em ambos os casos coletados em uma série de experimentos presenciais envolvendo participantes humanos engajados na tarefa de descrição de imagens. Para este fim, foram empregados 10 estímulos visuais extraídos da base GAPED (Dan-Glauser & Scherer, 2011) de imagens classificadas por valência e significância normativa, e designadas de modo a despertar diferentes graus e tipos de emoção. Um exemplo de imagem deste tipo foi apresentado na Figura 1 da Seção 1.

O córpus contempla duas subtarefas de descrição de imagens: em versão detalhada (em texto multi-sentencial) e em versão resumida (na forma de uma sentença única, ao estilo de uma legenda que resume o conteúdo da imagem). Um exemplo de descrição textual resumida para a imagem anterior, tal qual observado em *b5-caption*, poderia ser simplesmente "*Trabalho infantil*". A versão detalhada desta mesma descrição, tal qual observada em *b5-text*, é apresentada a seguir.

> *Criança de cerca de 5 anos, trabalhando injustamente, despejando terra num balde por algum motivo para ajudar alguém. Cenário de pobreza no qual ao seu fundo existe uma casa construída de sacos plásticos e alguns tijolos.*

O córpus *b*5 contém descrições textuais de 10 contextos visuais como o da figura anterior. Todas as imagens foram descritas por um grupo de 151 participantes, havendo portanto 151 versões longas e outras 151 versões curtas de cada uma —todas potencialmente influenciadas pelas diferentes personalidades de seus autores. Todos os 1510 textos longos e 1510 legendas encontram-se rotulados com escores de personalidade CGF obtidos por meio de inventários de personalidade dos participantes que os escreveram.

A porção *b5-text* dos dados contém 84463 *tokens* e a porção *b5-caption* contém 4896. Os textos completos da porção *b5-text* são usados para fins de treinamento dos modelos a serem discutidos. A porção *b5-caption* foi utilizada para fins de teste tal qual descrito na Seção 4.

3.2. Preparação dos dados

A geração de texto com variação lexical decorrente de personalidade (ou outros fatores) requer a identificação de conceitos de nível semântico a serem lexicalizados. Assim, o uso do córpus *b*5 descrito na seção anterior demanda uma tarefa de anotação semântica com o objetivo de estabelecer um mapeamento entre conceitos e suas lexicalizações possíveis. Esta atividade é descrita a seguir.

Os modelos de reescrita a serem propostos tratam da lexicalização de três tipos de conceitos semânticos representados nas cenas do córpus *b5-text* (Ramos et al., 2018): conceitos a serem realizados como substantivos (como a maior parte das entidades das cenas do córpus, pessoas, objetos etc.), adjetivos (e.g., características concretas e abstratas das entidades da cena, como cores, emoções etc.) e verbos (e.g., ações que as entidades presentes na cena executam). Estes conceitos são aqui denominados CRS, CRA e CRV, respectivamente, em referência à forma de realização superficial (substantivo, adjetivo ou verbo) que cada um assume no texto, e foram escolhidos por estarem entre os mais facilmente observáveis nas imagens do córpus.

Dado que um determinado conceito pode ser lexicalizado de muitas formas, o primeiro passo desse estudo consistiu em computar todos os conceitos referenciados no córpus *b*5 e então associar cada conceito a uma lista de formas lexicais possíveis. Para este fim, foi atribuído um rótulo único a cada conceito, aqui designado na forma $nome, onde 'nome' é um termo representado em inglês como forma de distingui-lo de suas formas lexicais em português. A seguir apresentamos um exemplo deste tipo de notação, em que o CRS $bucket é associado a três formas lexicais possíveis.

$bucket→ {'balde', 'recipiente', 'pote'}

O procedimento para cômputo dos mapeamentos de conceitos para lexicalizações difere para cada tipo de conceito (CRS, CRA ou CRV) por motivos discutidos a seguir. Em todos os casos, a anotação dos dados foi realizada por dois avaliadores, e nos poucos casos de ambiguidade observados, foram decididos por um terceiro.

No caso dos CRS, uma vez que os textos do córpus *b*5 são fortemente apoiados nos elementos visuais de cada cena, optou-se por modelar apenas a lexicalização de conceitos que representam elementos referenciáveis das cenas, incluindo por exemplo personagens e objetos físicos, entidades representando indivíduos ou grupos, e entidades representando um todo ou suas partes, mas excluindo-se conceitos mais abstratos ou de caráter retórico presentes no discurso (e.g., problemas, questões, diferenças etc.) Assim, criou-se inicialmente uma lista de todos elementos visuais que poderiam ser realizadas na forma de substantivos, atribuindo-se um rótulo identificador a cada um. Por exemplo, para a Figura 1 anterior poderiam ser enumerados, dentre outros, os CRS $bucket, $child, $tent para os objetos balde, criança e barraca, respectivamente.

A seguir, o córpus foi analisado sintaticamente com uso da ferramenta PALAVRAS (Bick, 2000), escolhida pela conveniência de uso de sua versão online, e por de modo geral apresentar bom desempenho para o Português brasileiro. Foram selecionados todos os substantivos com um mínimo de cinco ocorrências nos textos, um limite mínimo foi adotado para evitar casos muito infrequentes, e que não seriam passíveis de aprendizado automático. Estes substantivos foram então manualmente associados a cada um dos CRS enumerados no passo anterior examinando-se as imagens quando necessário para fins de desambiguação, objetivando agrupar todos os sinônimos referentes a cada um dos conceitos. Um exemplo de mapeamento de um CRS para uma lista de substantivos foi ilustrado no exemplo acima, para o caso do CRS $bucket.

Neste processo, alguns novos CRS que não tinham sido originalmente listados foram descobertos a partir dos próprios substantivos, e a lista de CRS foi atualizada de acordo quando pertinente. O resultado desta atividade foi a produção de uma série de mapeamentos de CRS para suas formas lexicais possíveis, contendo 151 CRS associados a 191 substantivos únicos com cinco ou mais ocorrências no córpus.

Uma vez definidos os CRS possíveis, o cômputo dos CRA foi realizado de forma semi-automática utilizando-se a informação sintática disponível. De forma mais específica, os adjetivos associados a cada substantivo ou pronome[2] de interesse no córpus foram manualmente associados a cada um dos CRS definidos no passo anterior. Assim, $black-bucket é o CRA que representa a cor escura de um balde, e $black-skin é o CRA que representa a cor escura da pele de um indivíduo. A definição de conceitos distintos associados a cada tipo de objeto a que se referem é motivada pelas observações de que seus conjuntos de lexicalizações possíveis podem não coincidir totalmente, e de que mesmo sendo idênticas estas lexicalizações podem ser empregadas de forma diferente para cada conceito por locutores com determinados tipos de personalidade.

Cada grupo de adjetivos assim identificado recebeu um rótulo único representando seu significado e contendo indicação do CRS ao qual se refere. Um exemplo de mapeamento de um CRA para uma lista de adjetivos é ilustrado a seguir, representando a qualidade do objeto balde de possuir uma cor escura.

$$\$\text{black-bucket} \rightarrow \{\text{‘escuro’,‘preto’,‘negro’}\}$$

Seguindo-se este procedimento, 114 CRA foram associados a 191 adjetivos únicos com cinco ou mais ocorrências no córpus.

No caso dos CRV, que representam ações que podem estar relacionadas a múltiplas entidades, optou-se simplesmente por computar todos os verbos do córpus e agrupá-los em conjuntos de sinônimos (ou *synsets*) obtidos a partir do dicionário TeP 2.0 (Maziero et al., 2008). A cada *synset* foi então atribuído um rótulo arbitrário único. Um exemplo de mapeamento de um CRV para uma lista de verbos é ilustrado a seguir.

$$\$\text{v1} \rightarrow \{\text{‘cansar’, ’debilitar’}\}$$

Como resultado, foram identificados 198 CRV associados a 352 verbos únicos com cinco ou mais ocorrências no córpus.

Finalmente, dado o objetivo de explorar lexicalizações alternativas (i.e., de acordo com uma personalidade-alvo) dos CRS, CRA e CRV identificados, apenas conceitos com mais de uma lexicalização possível foram mantidos. A Tabela 1 sumariza os três conceitos mais frequentes de cada tipo, o número de vezes que suas lexicalizações são utilizadas no córpus, e suas possíveis lexicalizações.

Os mapeamentos entre conceitos (CRA, CRS ou CRV) e suas possíveis lexicalizações serão tomados por base para o modelo de reescrita sentencial baseada em personalidade discutido na próxima seção.

[2] A identificação de antecedentes anafóricos foi feita de forma manual. Para abordagens computacionais, ver por exemplo Paraboni & de Lima (1998); Cuevas & Paraboni (2008).

Tipo	ID	Instâncias	Lexicalizações
CRS	$child	1399	criança, menino, garoto, bebê
	$scene	761	imagem, cena, cenário, foto
	$soil	437	terra, barro, areia
CRA	$white-human	989	claro, branco
	$black-human	823	escuro, preto, negro
	$pale-object	674	pálido, desmaiado, claro
CRV	$v1	8024	haver, ser, estar, existir
	$v2	4592	querer, relevar, estar, assentar
	$v3	3144	ser, servir

Tabela 1: Três conceitos mais frequentes de cada classe (CRS, CRA e CRV).

3.3. Modelo proposto

Nesta seção é apresentado o modelo de reescrita sentencial que é o foco principal deste trabalho. O modelo recebe como entrada uma sentença s em Português e os escores das cinco dimensões de personalidade P de um locutor-alvo de interesse, e produz como saída uma nova sentença s' que é uma variação de s em que certas palavras foram substituídas por sinônimos que seriam tipicamente empregados por um locutor de personalidade P. Neste modelo, a tarefa de escolha lexical é implementada com uso de métodos de aprendizado de máquina supervisionado (ou classificadores). Estes dois componentes —classificadores e o reescrevedor sentencial propriamente dito— são ilustrados na Figura 2, e serão detalhados separadamente nas próximas seções.

3.3.1. Classificadores de escolha lexical

O componente principal do modelo de reescrita é formado por um conjunto de 10 classificadores (um para cada contexto do córpus *b5-text*) que recebe como entrada um conceito semântico representado por *id* e escores de uma personalidade-alvo P, e retorna a lexicalização w de *id* que seria tipicamente empregada por locutores de personalidade P para descrever o conceito *id*. Por exemplo, dado *id* = $child e um perfil P de maior agradabilidade, a classificação pode sugerir a forma lexical w = 'criança'. Com menor agradabilidade, por outro lado, o classificador pode sugerir, e.g., w = 'pirralho'.

O uso de classificadores individuais para cada contexto do córpus é motivado pela observação de que cada imagem é efetivamente um domínio distinto, e conceitos superficialmente semelhantes (como 'mulher') possuem significados e lexicalizações distintas em cada um (i.e., cada imagem representa uma mulher diferente, que em alguns casos pode ser descrita como 'moça' e em outras não). Assim, cada subconjunto do córpus *b5-text* precisa ser tratado com um córpus distinto.

Para cada palavra w representando um adjetivo, substantivo ou verbo com 50 ou mais ocorrências no córpus, foram criadas instâncias de aprendizado supervisionado como segue. Primeiramente, cada palavra w é pesquisada nos mapeamentos de conceitos para palavras (cf. seção anterior) e o conceito *id* correspondente é recuperado. Em paralelo a isso, o autor do texto é pesquisado na base de participantes do córpus *b*5 e seus escores de personalidade P são também recuperados. Assim, a instância de aprendizado é formada pelos atributos *id* e P, e a classe a ser apreendida é a palavra w. Havendo mais de um mapeamento para a palavra em questão, múltiplas instâncias de aprendizado são criadas. Finalmente, todas as instâncias relativas a conceitos com menos de cinco ocorrências foram descartadas.

Este procedimento resultou em um conjunto de dados contendo 35k instâncias de aprendizado de 734 lexicalizações possíveis para todos os CRA, CRS e CRV com frequência mínima no córpus e que apresentavam alguma variação lexical[3]. A distribuição geral de classes por tipo de conceito é ilustrada na Tabela 2.

Tipo de Conceito	Instâncias de aprendizado	Lexicalizações possíveis
CRS	10428	191
CRA	5741	191
CRV	18905	352

Tabela 2: Número de instâncias e lexicalizações por tipo de conceito.

[3] O número de possíveis escolhas lexicais para cada conceito variou entre 2 e 12 alternativas.

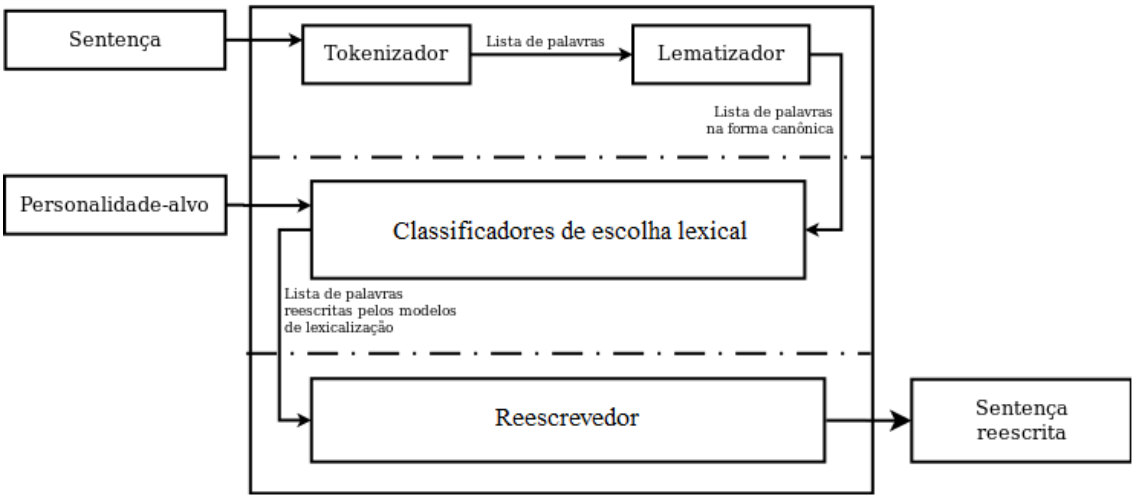


Figura 2: Fases da reescrita sentencial.

Em um experimento-piloto foi avaliado o uso de máquinas de vetor de suporte (SVM), indução de árvore de decisão e classificação Naive Bayes na tarefa de lexicalização. Como resultado, optou-se por construir todos os classificadores utilizando-se SVMs com núcleo de base radial. Os parâmetros ótimos de cada modelo foram obtidos por *grid search*, e os classificadores resultantes foram então incorporados ao modelo de reescrita sentencial propriamente dito.

3.3.2. Reescrevedor sentencial

Os classificadores descritos na seção anterior são combinados em um modelo reescrevedor que recebe como entrada uma sentença em língua natural e efetua substituições lexicais de acordo com um determinado perfil de personalidade fornecido. Assim, é possível converter uma dada sentença em versões alternativas que refletem um maior ou menor grau de extroversão, agradabilidade, conscienciosidade, neuroticismo ou abertura à experiência, tal qual definido no modelo dos Cinco Grandes Fatores (Goldberg, 1990).

O processo de reescrita segue um *pipeline* de três fases onde a sentença fornecida é analisada e modificada conforme o diagrama da Figura 2 anterior. Primeiramente, a sentença de entrada passa por um processo de *tokenização* e lematização dos termos constituintes, gerando assim uma lista de palavras na sua forma canônica. A seguir, os mapeamentos de palavras para conceitos discutidos na Seção 3.2 são utilizados para obter os conceitos que as palavras de entrada representam. Finalmente, os classificadores de escolha lexical são invocados para predizer a forma modificada (baseada em personalidade) de cada palavra original e, caso a predição resulte em uma forma válida para o conceito fornecido[4], a substituição lexical é realizada. Nota-se também que, caso a palavra de entrada já esteja na forma adequada para o tipo de personalidade em questão, a lexicalização sugerida pelo classificador pode ser idêntica à palavra de entrada, e neste caso nenhuma substituição é realizada.

3.3.3. Exemplo de funcionamento

Esta seção descreve um exemplo simplificado de funcionamento do modelo proposto. Para este fim, considere-se a tarefa de reescrever a frase de entrada a seguir com base em um conjunto de traços de personalidade-alvo P.

Homem maduro demonstrando preocupação.

Inicialmente, é obtido o conceito (CRS, CRV ou CRA) referente a cada substantivo, verbo ou adjetivo da frase fornecida. Neste exemplo seriam obtidos os conceitos $man (com possíveis lexicalizações como 'homem', 'senhor' etc.) $old ('maduro', 'idoso', 'velho' etc.), $show ('demonstrar', 'mostrar', 'revelar' etc.) e $negemotion ('preocupação', 'sofrimento', 'desespero' etc.)

Cada conceito, associado a um conjunto de traços P fornecido, forma uma instância de teste a ser submetida ao respectivo classificador de escolha lexical (de CRS, CRA ou CRV, conforme o caso) de modo a obter a lexicalização mais adequada (i.e., de acordo com a personalidade-alvo

[4]Ou seja, dado que os classificadores não possuem desempenho perfeito, é possível que façam predições inconsistentes a serem desconsideradas.

fornecida). Supondo-se, respectivamente, escores de Extroversão, Agradabilidade, Conscienciosidade, Neuroticismo e Abertura à experiência como $P = \{3.0, 3.0, 2.4, 2.0, 3{,}5\}$, um exemplo de instância de aprendizado a ser submetidas ao classificador de lexicalização de CRA seria como segue:

[$old, 3.0, 3.0, 2.4, 2.0, 3.5]

A seguir, os classificadores de CRS, CRA e CRV são invocados, retornando a lexicalização de cada conceito fornecido. No exemplo, seriam obtidas as lexicalizações 'homem', 'velho', 'mostrar' e 'desespero', respectivamente. Finalmente, os termos originais da frase de entrada são substituídos pelas formas lexicais sugeridas pelos classificadores com a devida flexão, resultando na seguinte sentença reescrita:

Homem velho mostrando desespero.

Variando-se os escores de personalidade em P, lexicalizações alternativas podem ser sugeridas pelos classificadores, resultando em sentenças distintas porém de significado aproximado. Por exemplo, um maior grau de Extroversão modificaria a lexicalização do verbo $show, e um maior grau de Neuroticismo modificaria simultaneamente as lexicalizações do adjetivo $old e do substantivo $negemotion. Assim, dado um perfil de personalidade $P = \{4.0, 3.0., 2.4, 4.1, 3, 5\}$, a seguinte sentença seria obtida.

Homem idoso revelando sofrimento.

Cabe destacar, entretanto, que a interação entre os cinco traços de personalidade é de modo geral complexa, e que o efeito de um traço pode ser atenuado ou amplificado por outro. Por exemplo, um escore mais baixo de Conscienciosidade neste exemplo faria com que a lexicalização do adjetivo $old fosse definida como 'idoso' independentemente do grau de Neuroticismo.

4. Avaliação e resultados

A avaliação dos modelos propostos foi realizada de duas formas: considerando-se o desempenho individual dos classificadores de escolha lexical, e considerando-se o uso destes classificadores como parte de um modelo completo de reescrita sentencial. As duas formas de avaliação são detalhadas nas seções a seguir.

Dado que os classificadores, se tomados isoladamente, não possuem função prática, sua avaliação objetiva unicamente ilustrar seu desempenho na tarefa de lexicalização de conceitos individuais sob a perspectiva de aprendizado de máquina. Para este fim, foi realizada a comparação entre os resultados dos modelos baseados em personalidade com uma alternativa de *baseline* na qual os cinco atributos de aprendizado representando os escores de personalidade do autor-alvo são omitidos. Assim, a escolha lexical neste caso passa a ser feita apenas com base no conceito fornecido como entrada. Todos os classificadores com e sem acesso à informação de personalidade foram avaliados utilizando-se o conjunto de dados completo do córpus *b5-text* com validação cruzada de 10 partições, e computando-se sua medida F_1 média.

No caso da reescrita sentencial, por outro lado, objetiva-se verificar se o uso de informação de personalidade permite gerar sentenças mais próximas das produzidas por humanos (tal qual observadas no córpus) do que uma alternativa que não tenha acesso a este tipo de informação (ou seja, construída utilizando-se os classificadores de *baseline* acima.) Para este fim, todos os modelos com e sem acesso à informação de personalidade foram treinados com o conjunto completo de textos do córpus *b5-text*, e aplicados à geração de sentenças do córpus *b5-caption*, que contém legendas das mesmas imagens e escritas pelos mesmos autores dos textos de treinamento. A avaliação neste caso foi realizada seguindo-se prática comum na avaliação de sistemas de GLN, em especial no que diz respeito à tarefa de realização superficial (Reiter & Belz, 2009), comparando-se cada sentença original do córpus e sua versão modificada e medindo-se a distância de edição (Damerau, 1964) entre ambas. É importante observar também que os reflexos de diferentes personalidades na produção textual tende a ser sutis (Walker et al., 2011b), o que dificulta o uso de métodos de avaliação mais sofisticados (e possivelmente de maior custo) como o uso de julgamento humano.

4.1. Resultados dos modelos de classificação de escolha lexical

A Tabela 3 apresenta os resultados médios de lexicalização dos CRS, CRA e CRV do córpus *b5-text*.

Para os três tipos de conceitos, observa-se que o uso de informação de personalidade (à direita na tabela) apresenta resultados superiores ao modelo de *baseline* que não tem acesso a esse tipo de informação, e que os resultados da lexicalização na forma de adjetivos são consideravelmente superiores aos demais. Este comportamento era em grande parte esperado, já que o modelo CGF é mais fortemente relacionado ao uso

	sem personalidade			com personalidade		
Tipo	P	R	F_1	P	R	F_1
CRS	0,45	0,64	0,52	0,73	0,75	**0,70**
CRA	0,78	0,79	0,76	0,85	0,82	**0,82**
CRV	0,76	0,77	0,74	0,78	0,78	**0,76**

Tabela 3: Médias ponderadas de precisão (P), revocação (R) e medida F_1 do modelo de lexicalização completo. O maior valor de medida F_1 de cada tipo é destacado.

de adjetivos (Goldberg, 1990), e estes seriam assim os elementos que melhor refletem a distinção entre tipos de personalidade.

Dado que estes resultados incluem diversos conceitos com pequeno número de ocorrências, é possível obter uma visão mais clara do desempenho efetivo dos classificadores analisando-se apenas o resultado médio dos 10 conceitos mais frequentes de cada tipo. Os resultados desta análise são apresentados na Tabela 4.

	sem personalidade			com personalidade		
Tipo	P	R	F_1	P	R	F_1
CRS	0,53	0,58	0,52	0,6	0,58	**0,56**
CRA	0,48	0,52	0,48	0,74	0,67	**0,68**
CRV	0,55	0,52	0,51	0,72	0,56	**0,58**

Tabela 4: Médias ponderadas de precisão (P), revocação (R) e medida F_1 do modelo de lexicalização para os dez conceitos mais frequentes de cada tipo. O maior valor de medida F_1 de cada tipo é destacado.

Novamente, observa-se que os classificadores com acesso à informação de personalidade apresentam resultados mais próximos do desempenho humano do que os classificadores de *baseline*.

Por fim, como forma de comparar o desempenho dos classificadores com e sem acesso à informação de personalidade por meio de exemplos, as Tabelas 5, 6 e 7 apresentam resultados das alternativas de lexicalização dos conceitos mais frequentes de cada tipo (CRA, CRS e CRV).

No caso dos CRS mais frequentes listados na Tabela 5, observa-se que o uso de personalidade foi superior em seis das onze lexicalizações mais frequentes, havendo também um empate (no grupo criança/jovem) e quatro casos em que é ligeiramente inferior, embora todos concentrados nas classes menos frequentes. Estes resultado era de certa forma esperado dado que substantivos tendem a apresentar menor variação de sinônimos do que adjetivos e verbos.

Substantivos	sem personalidade	com personalidade
foto	0,45	**0,53**
imagem	0,56	**0,62**
cena	0,00	**0,33**
criança	0,64	**0,64**
jovem	0,00	0,00
casaco	**0,69**	0,67
jaqueta	0,00	**0,33**
senhora	**0,69**	0,68
mulher	**0,64**	0,62
calça	**0,86**	0,82
jeans	0,00	**0,35**

Tabela 5: Medida F_1 média obtida para os cinco CRS mais frequentes, por forma lexical.

Adjetivos	sem personalidade	com personalidade
alto	0,83	**0,85**
forte	0,65	**0,71**
bravo	0,69	**0,82**
simples	**0,89**	0,89
singelo	0,00	**0,44**
rural	**0,71**	0,63
rústico	0,00	**0,30**
velho	0,63	**0,71**
idoso	0,74	**0,75**
preocupado	0,76	**0,92**
apreensivo	0,00	**0,80**

Tabela 6: Medida F_1 média obtida para os cinco CRA mais frequentes, por forma lexical.

Verbox	sem personalidade	com personalidade
mostrar	0,73	**0,74**
dar	0,50	**0,60**
revelar	0,00	**0,67**
dividir	0,33	**0,75**
destacar	0,80	**0,81**
separar	0,48	**0,56**
carregar	**0,58**	0,56
passar	**0,61**	0,60
tirar	**0,93**	0,92
separar	0,48	**0,56**
afastar	0,00	**0,20**
espalhar	0,56	**0,58**
dar	0,50	**0,60**

Tabela 7: Medida F_1 média obtida para os cinco CRV mais frequentes, por forma lexical.

No caso dos CRA a vantagem dos modelos baseados em personalidade é mais expressiva, sendo superiores em 9 dos 11 casos de lexicalização considerados. Finalmente, no caso dos CRV, o uso de informação de personalidade é superior em 10 de 13 casos.

4.2. Resultados de reescrita sentencial

Embora os resultados da avaliação do classificador apontem uma certa vantagem no uso de informação de personalidade na tarefa de lexicalização, resta a questão de quão efetiva seria a tarefa de reescrita sentencial baseada nestes classificadores. A avaliação global do modelo de reescrita é descrita a seguir.

A Tabela 8 apresenta a distância de edição média obtida pelas duas variantes do modelo proposto —com e sem acesso à informação de personalidade— para cada contexto do córpus.

	sem personalidade		**com personalidade**	
Contexto	Média	DP	Média	DP
1	3,28	3,01	**2,81**	3,29
2	3,52	1,93	**2,23**	2,41
3	3,02	3,69	**2,91**	3,73
4	**0,45**	3,28	0,56	3,51
5	**1,11**	3,39	1,26	3,25
6	**1,07**	2,65	1,30	3,06
7	2,89	2,70	**2,28**	3,14
8	**0,32**	1,47	0,36	1,72
9	3,34	4,22	**2,70**	3,74
10	2,27	3,61	**1,57**	2,92
Média	3,37	3,36	**2,53**	3,36

Tabela 8: Médias e desvios padrão das distâncias de edição por contexto. O melhor resultado para cada contexto é destacado.

Com base nesses resultados, observa-se que a distância de edição média dos modelos com acesso à personalidade é inferior à dos modelos sem acesso a esta informação. Além disso, os modelos baseados em personalidade possuem, em média, menor distância de edição em 6 dos 10 contextos visuais do córpus, e observa-se ainda que os 4 contextos em que o uso de informação de personalidade foi prejudicial são justamente aqueles em que ocorreu menor volume de substituição lexical (indicado pelo menor grau de distância de edição).

A diferença entre a distância de edição média dos modelos de reescrita sentencial com e sem personalidade é significativa de acordo com um teste ANOVA de fator único ($F(1, 1624) = 13.23$, $MSE = 11.39$, $p < 0.001$). Esse resultado oferece suporte à hipótese de pesquisa deste estudo, ou seja, de que o uso de informação de personalidade permite a geração de sentenças mais próximas daquelas que seriam geradas por um locutor humano com estas características.

5. Conclusão

Este trabalho apresentou um modelo de reescrita sentencial que leva em conta, além da sentença a ser reescrita, informações sobre a personalidade de um locutor-alvo de interesse. O modelo proposto efetua substituições lexicais com base no modelo de personalidade CGF e é, até onde temos conhecimento, o primeiro deste tipo para o português (pelo menos em sua variante brasileira). Resultados sugerem que levar em conta a personalidade do autor-alvo de interesse produz sentenças mais próximas das que seriam produzidas por um locutor humano com as mesmas características do que um modelo de geração sem acesso a este tipo de informação, e sugerem assim a possibilidade de geração de texto mais natural e/ou psicologicamente plausível.

Uma das possíveis limitações do presente trabalho é o fato de ser baseado em um conjunto de textos rotulados manualmente com informações de conceitos semânticos de interesse para fins de aprendizado de máquina supervisionado. Assim, o uso dos métodos aqui discutidos em outro domínio em princípio exigiria um novo trabalho de anotação de córpus, possivelmente com avaliação do grau de concordância entre juízes (no presente caso omitido dada a simplicidade do domínio e do pequeno tamanho do córpus considerado). Alternativas mais generalizáveis, como as baseadas em métodos semi ou não supervisionados de aprendizado, são deixadas como trabalho futuro.

Finalmente, observa-se que o foco do presente trabalho foi a questão da escolha lexical feita por indivíduos com diferentes personalidades, e foi motivado pela própria natureza lexical do modelo CGF. No entanto, para além deste modelo específico, é possível considerar diversas outras formas de variação humana na produção de língua natural. Estas variações incluem o uso de certas construções sintáticas preferenciais, tempos verbais e até mesmo a determinação do conteúdo semântico a ser representado em forma textual. Estas variações, que seriam uma extensão natural do presente trabalho, também são deixadas como trabalho futuro.

Agradecimentos

Este trabalho contou com apoio FAPESP nro. 2016/14223 0.

Referências

Álvarez-Carmona, Miguel A., A. Pastor López-Monroy, Manuel Montes y Gómez, Luis Villaseñor-Pineda & Hugo Jair Escalante. 2015. INAOE's participation at PAN'15: Author profiling task. Em *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF)*, s.p.

Aly, Almir & Adriana Tapus. 2016. Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction. *Autonomous Robots* 40(2). 193–209. 10.1007/s10514-015-9444-1.

Bick, Eckhard. 2000. *The parsing system PALAVRAS: Automatic grammatical analysis of portuguese in a constraint grammar famework*: Aarhus University. Tese de Doutoramento.

Bowden, Kevin K., Grace Lin, Lena Reed, Jean E. Fox Tree & Marilyn A. Walker. 2016. M2D: Monolog to dialog generation for conversational story telling. Em *9th International Conference on Interactive Digital Storytelling*, 12–24. 10.1007/978-3-319-48279-8_2.

Briggs Myers, Isabel & Peter B. Myers. 2010. *Gifts differing: Understanding personality type*. Davies-Black 2nd edn.

Cuevas, Ramon Re Moya & Ivandré Paraboni. 2008. A machine learning approach to Portuguese pronoun resolution. Em *Advances in Artificial Intelligence (IBERAMIA)*, 262–271. 10.1007/978-3-540-88309-8_27.

Damerau, Fred J. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM* 7(3). 171–176. 10.1145/363958.363994.

Dan-Glauser, Elise S. & Klaus R. Scherer. 2011. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods* 43(2). 468–477. 10.3758/s13428-011-0064-1.

van Deemter, Kees. 2016. *Computational models of referring. a study in cognitive science*. Cambridge: MIT Press.

di Eugenio, Barbara, Andrew Boyd, Camillo Lugaresi, Abhinaya Balasubramanian, Gail Keenan, Mike Burton, Tamara Goncalves Rezende Macieira, Jianrong Li & Yves Lussier. 2014. PatientNarr: Towards generating patient-centric summaries of hospital stays. Em *8th International Natural Language Generation Conference (INLG)*, 6–10. 10.3115/v1/W14-4402.

Ferreira, Thiago Castro & Ivandré Paraboni. 2014. Classification-based referring expression generation. Em *Computational Linguistics and Intelligent Text Processing (CICLing)*, 481–491. 10.1007/978-3-642-54906-9_39.

Goldberg, Lewis R. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology* 59. 1216–1229. 10.1037/0022-3514.59.6.1216.

González-Gallardo, Carlos E., Azucena Montes, Gerardo Sierra, J. Antonio Núñez-Juaréz, Adolfo Jonathan Salinas-López & Juan Ek. 2015. Tweets classification using corpus dependent tags, character and POS n-grams. Em *Working notes of the Conference and Labs of the Evaluation Forum (CLEF)*, s.p.

Herzig, Jonathan, Michal Shmueli-Scheuer, Tommy Sandbank & David Konopnicki. 2017. Neural response generation for customer service based on personality traits. Em *10th International Conference on Natural Language Generation*, 252–256. 10.18653/v1/W17-3541.

Hsieh, Fernando Chiu, Rafael Felipe Sandroni Dias & Ivandré Paraboni. 2018. Author profiling from Facebook corpora. Em *11th International Conference on Language Resources and Evaluation (LREC)*, 2566–2570.

Hunter, James, Yvonne Freer, Albert Gatt, Ehud Reiter, Somayajulu Sripada & Cindy Sykes. 2012. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial intelligence in medicine* 56(3). 157–172. 10.1016/j.artmed.2012.09.002.

John, Oliver P., Laura P. Naumann & Christopher J. Soto. 2008. Paradigm shift to the integrative Big-Five trait taxonomy: History, measurement, and conceptual issues. Em *Handbook of personality: Theory and research*, 114–158. New York: Guilford Press.

Jordan, Pamela, Nancy Green, Chistopher Thomas & Susan Holm. 2014. TBI-Doc: Generating patient & clinician reports from brain

imaging data. Em *8th International Natural Language Generation Conference (INLG)*, 143–146. 10.3115/v1/W14-4423.

Karpathy, Andrej & Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. Em *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3128–3137. 10.1109/CVPR.2015.7298932.

Khosmood, Foaad & Marilyn Walker. 2010. Grapevine: a gossip generation system. Em *5th International Conference on the Foundations of Digital Games*, 92–99.

Lan, Alex Gwo Jen & Ivandré Paraboni. 2018. Definite description lexical choice: taking speaker's personality into account. Em *11th International Conference on Language Resources and Evaluation (LREC)*, 2999–3004.

Lukin, Stephanie M., James O. Ryan & Marilyn Walker. 2014. Automating direct speech variations in stories and games. Em *Games and Natural Language Processing Workshop (GAMNLP)*, s.p.

Mairesse, François & Marilyn A. Walker. 2010. Towards personality-based user adaptation: psychologically informed stylistic language generation. *User Model. User-Adapt. Interaction* 20(3). 227–278. 10.1007/s11257-010-9076-2.

Mairesse, François & Marilyn A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics* 37(3). 455–488. 10.1162/COLI_a_00063.

Mairesse, François. 2008. *Learning to adapt in dialogue systems: Data-driven models for personality recognition and generation*: University of Sheffield. Tese de Doutoramento.

Maziero, Erick G., Thiago A. S. Pardo, Ariani di Felippo & Bento C. Dias da Silva. 2008. A base de dados lexical e a interface web do TeP 2.0: Thesaurus eletrônico para o português do Brasil. Em *14th Brazilian Symposium on Multimedia and the Web*, 390–392. 10.1145/1809980.1810076.

McCormick, Christopher. 2012. *Evaluating the perception of personality and naturalness in computer generated utterances*: University of Dublin, Trinity College. Tese de Mestrado.

Paraboni, Ivandré & Vera Lucia Strube de Lima. 1998. Possessive pronominal anaphor resolution in Portuguese written texts. Em *17th International Conference on Computational Linguistics*, 1010–1014. 10.3115/980691.980735.

Paraboni, Ivandré, Danielle Sampaio Monteiro & Alex Gwo Jen Lan. 2017. Personality-dependent referring expression generation. Em *Text, Speech and Dialogue (TSD)*, 20–28. 10.1007/978-3-319-64206-2_3.

Plank, Barbara & Dirk Hovy. 2015. Personality traits on Twitter - or - how to get 1,500 personality tests in a week. Em *6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 92–98. 10.18653/v1/W15-2913.

Portet, François, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer & Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence* 173(7–8). 789–816. 10.1016/j.artint.2008.12.002.

Ramos, Ricelli Moreira Silva, Georges Basile Stavracas Neto, Barbara Barbosa Claudino Silva, Danielle Sampaio Monteiro, Ivandré Paraboni & Rafael Felipe Sandroni Dias. 2018. Building a corpus for personality-dependent natural language understanding and generation. Em *11th International Conference on Language Resources and Evaluation (LREC)*, 1138–1145.

Reiter, Ehud & Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics* 35(4). 529–558. 10.1162/coli.2009.35.4.35405.

Reiter, Ehud & Robert Dale. 2000. *Building natural language generation systems*. New York: Cambridge University Press.

Reiter, Ehud, Roma Robertson & Liesl Osman. 2003. Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence* 144(1–2). 41–58. 10.1016/S0004-3702(02)00370-3.

Reiter, Ehud, Somayajulu Sripada, Jim Hunter & Jin Yu. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence* 167(1–2). 137–169. 10.1016/j.artint.2005.06.006.

dos Santos, Vitor Garcia, Ivandré Paraboni & Bárbara Barbosa Claudino Silva. 2017. Big five personality recognition from multiple text genres. Em *Text, Speech and Dialogue (TSD)*, 29–37. 10.1007/978-3-319-64206-2_4.

Silva, Bárbara Barbosa Claudino & Ivandré Paraboni. 2018a. Learning personality traits from Facebook text. *IEEE Latin America Transactions* 16(4). 1256–1262. 10.1109/TLA.2018.8362165.

Silva, Bárbara Barbosa Claudino & Ivandré Paraboni. 2018b. Personality recognition from Facebook text. Em *13th International Conference on the Computational Processing of Portuguese (PROPOR)*, 107–114. 10.1007/978-3-319-99722-3_11.

Viethen, Jette & Robert Dale. 2010. Speaker-dependent variation in content selection for referring expression generation. Em *Australasian Language Technology Association Workshop*, 81–89.

Walker, Marilyn, Grace Lin, Jennifer Sawyer, Ricky Grant, Michael Buell & Noah Wardrip-Fruin. 2011a. Murder in the arboretum: Comparing character models to personality models. Em *Intelligent Narrative Technologies*, Santa Cruz: AAAI.

Walker, Marilyn A., Ricky Grant, Jennifer Sawyer, Grace I. Lin, Noah Wardrip-Fruin & Michael Buell. 2011b. Perceived or not perceived: Film character models for expressive NLG. Em *International Conference on Interactive Digital Storytelling: Interactive Storytelling (ICIDS)*, 109–121. 10.1007/978-3-642-25289-1_12.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel & Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. Em *32nd International Conference on Machine Learning, ICML*, 2048–2057.

Zhang, Xingxing & Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 670–680. 10.3115/v1/D14-1074.

Proposta recebida em Dezembro 2019 e aceite para publicação em Março 2020.

Subjetividade em correção de redações: detecção automática através de léxico de operadores de viés linguístico

Subjectivity in essay grading: automatic detection through language bias operator lexicon

Márcia Cançado
FALE/Univ.Federal de Minas Gerais
mcancado@ufmg.br

Luana Amaral
FALE/Univ. Federal de Minas Gerais
luanalopes@ufmg.br

Evelin Amorim
DCC/Univ. Federal de Minas Gerais
evelin.amorim@gmail.com

Adriano Veloso
DCC/Univ. Federal de Minas Gerais
adrianov@dcc.ufmg.br

Heliana Mello
FALE/Univ. Federal de Minas Gerais
heliana.mello@gmail.com

Resumo

As redações são instrumentos avaliativos muito importantes para os estudantes brasileiros. Mesmo que seja assumido que a subjetividade esteja presente em todo e qualquer texto, espera-se que as correções dessas redações sejam feitas com o mínimo de subjetividade possível. Entretanto, a partir da análise de uma amostra de correções de redação, percebemos um alto grau de subjetividade nesses textos. Baseados nessa pré-análise, feita de forma manual, levantamos a hipótese de que o gênero "correção de redação" é mais subjetivo do que se esperaria. Para corroborar essa hipótese, elaboramos uma lista de operadores linguísticos, marcadores de viés, dividida em quatro categorias: operadores argumentativos, operadores de pressuposição, operadores de modalidade e operadores de opinião e valoração. Essa lista foi aplicada, através de uma metodologia de detecção automática de linguagem enviesada, a um *corpus* de correções de redação. A partir disso, quantificamos os operadores de viés presentes nesses textos. Foram também analisados esses operadores de viés em dois outros *corpora*: de resumos acadêmicos e de resenhas de produtos publicadas em sites de vendas na *internet*. A ideia dessa análise foi compararmos a distribuição dessas marcas de viés nas correções de redação e em gêneros reconhecidamente menos subjetivos (resumos acadêmicos) e reconhecidamente mais subjetivos (resenhas). Para tal comparação, lançamos mão de uma ferramenta estatística muito utilizada na análise de comparação de dados, os *boxplots*. Os nossos resultados mostraram que a distribuição de operadores de viés linguístico nas correções de redação se aproxima mais da distribuição desses itens em resenhas do que em resumos acadêmicos. Isso corrobora nossa hipótese e indica que o grau de subjetividade das correções é alto, estando mais próximo do grau de subjetividade de um texto como as resenhas. Concluímos, portanto, que essas correções refletem pontos de vista do corretor, que se afastam dos critérios de correção, o que coloca dúvidas sobre a consideração desse gênero como um instrumento avaliador isento e justo.

Palavras chave

correção de redação, subjetividade, léxico de operador de viés, detecção automática

Abstract

Essays are very important assessment tools for Brazilian students. Therefore, it is expected that the grading of these texts will be made with as little subjectivity as possible. However, in an analysis of a sample of grading sheet comments by evaluators, we have noticed a high degree of subjectivity in these texts. From this first analysis, carried manually, we proposed the hypothesis that this genre is more subjective than one would expect. In order to corroborate this hypothesis, we have drawn up a list of linguistic bias markers, divided into four categories: argumentative operators, presupposition operators, modalization operators, and opinion and value operators. This list was applied to a *corpus* of essay grading sheet comments by evaluators, using an automatic language bias detection methodology. From this, we were able to quantify the linguistic bias markers present in these texts. These bias markers were also analyzed in two other *corpora*: abstracts and product reviews published on internet sales sites. We have compared the percentage of these markers in

DOI: 10.21814/lm.12.1.313

evaluators' comments with the percentage numbers of these markers in genres admittedly less subjective (abstracts) and admittedly more subjective (reviews). For such comparison, we have used *boxplots*, a statistical tool widely used in data comparison analysis. Our results indicated that the grading sheets, as for the number of bias markers, are closer to more subjective texts than to less subjective texts. This corroborates our hypothesis and indicates that these grading sheets present a high degree of subjectivity, closer to the degree of a more subjective text. Thus, we conclude that these grading sheets reflect the personal views of the evaluator, deviating from the correction criteria, which raises doubts about considering this genre an exempt and fair assessment instrument.

Keywords

essay grading, subjectivity, bias operator lexicon, automatic detection

1. Introdução

Seguindo a linha teórica da Semântica Argumentativa, Anscombre & Ducrot (1976) e Ducrot (1987) (e trabalhos subsequentes) afirmam que há pistas na língua que indicam uma orientação, ou um viés, do locutor. Essas pistas podem ser detectadas por operadores argumentativos, tipos específicos de verbos, modalizadores e outras expressões linguísticas. A distribuição numérica dessas marcas no texto pode indicar graus de subjetividade, o que chamamos aqui de "viés linguístico". Mesmo que seja assumido que a subjetividade esteja presente em todo e qualquer texto, ainda assim, é possível fazer distinções de grau e, ainda mais, é possível detectar a orientação de textos muito marcados nesse sentido.

Tendo em vista essa afirmação, espera-se que um texto como a correção de uma redação não apresente um alto índice de viés linguístico. Essa expectativa deve-se ao fato de que, primeiramente, a redação é usada como um instrumento avaliativo durante toda a trajetória de formação escolar de um estudante do ensino básico; e, um segundo fato, de grande relevância, é que redações são amplamente utilizadas como instrumento de classificação em concursos importantes, realçando, no Brasil, a utilização desse procedimento no ENEM (Exame Nacional do Ensino Médio). Neste último caso, a nota da redação é a responsável por uma grande parte da classificação de um candidato, fechando ou abrindo as portas de entrada em nossas universidades públicas. Por isso, espera-se que esse instrumento de avaliação seja o mais isento possível de qualquer tipo de viés do corretor, não apresentando, assim, desigualdades nos resultados.

Neste artigo, propomos que o grau de viés linguístico apresentado pelo corretor de redações pode ser detectado por pistas linguísticas nos textos dessas correções. Analisamos, preliminarmente, uma amostra de 50 exemplares do gênero. Identificamos, manualmente, a partir do nosso conhecimento teórico e da nossa intuição de falantes do português, marcas linguísticas de subjetividade nesses textos. Os resultados dessa primeira análise indicaram o oposto do esperado: encontramos correções com um número alto de operadores linguísticos marcadores de viés. Além disso, ao fazermos nós mesmos a correção das redações, percebemos que as notas que se distanciavam muito das notas atribuídas por nós, seja para cima ou para baixo, eram acompanhadas de justificativas muito subjetivas, segundo critérios preliminares que estabelecemos. Dessa forma, chegamos à hipótese de que o gênero correção de redação não é tão isento de subjetividade quanto se esperaria e esses textos, sendo permeados de marcas de viés linguístico, indicam também um viés do corretor na correção de redações.

Na mesma linha da nossa hipótese, Mendes (2013), em seu artigo sobre a subjetividade na avaliação de redações, demonstra a inviabilidade de se manter tal tipo de prova em processos seletivos, dada a impossibilidade de se fazer justiça aos candidatos. A pesquisadora aponta que nos próprios critérios usados na correção do ENEM "encontram-se evidências do problema da subjetividade na descrição dos níveis das competências. Como se poderá verificar, esses descritores são eivados de modalizadores, apresentam vagueza conceitual e termos obscuros" (Mendes (2013), p. 439). A autora apresenta, ainda, uma pesquisa empírica sobre o sistema de avaliações de redações no ENEM, em que ela e mais um grupo de docentes e alunos de pós-graduação corrigiram 700 redações feitas por alunos de um curso pré-vestibular. Essa correção foi feita seguindo os mesmos critérios e sistemática de correção do ENEM. A partir dessa pesquisa empírica, o artigo apresenta interessantes resultados, como a discrepância entre as notas atribuídas pelos avaliadores, a discrepância entre as notas dependendo do horário da correção e até mesmo diferenças em notas atribuídas a uma mesma redação por um mesmo corretor, em momentos diferentes. A autora aponta que, além da subjetividade intrínseca ao próprio avaliador, outros fatores também subjetivos interferem na confiabilidade da avaliação, como o perfil do avaliador, a sua tolerância ao cansaço, a interação entre avaliadores e erros na avaliação.

Para provarmos essa hipótese de forma mais rigorosa, neste trabalho, nos valemos de uma metodologia de análise usada na Linguística Computacional, chamada de "detecção automática de linguagem enviesada" (Recasens et al., 2013). A proposta dessa metodologia é que, a partir de um trabalho manual de análise linguística feita por falantes nativos para o entendimento da realização dos vieses linguísticos, pode-se propor um detector automático de viés composto de uma lista de palavras indutoras de vieses linguísticos. Assim, elaboramos uma lista de operadores, marcadores de viés linguístico, dividida em quatro categorias: operadores argumentativos, operadores de pressuposição, operadores de modalização e operadores de opinião e valoração. Para a elaboração dessa lista, apoiamo-nos nos pressupostos teóricos da Semântica Argumentativa, principalmente por Anscombre & Ducrot (1976); Ducrot (1987); Koch (2011, 2015). Apresentamos também, como objetivo do artigo, a elaboração dessa lista de operadores de viés linguístico para o português.

Dando continuidade à nossa pesquisa, extraímos, de forma automática, um *corpus* de textos do gênero correção de redação, disponíveis no banco de redações do UOL, para procedermos, também de forma automática, à identificação e à contagem dos operadores de viés listados. Mas somente a contagem desses operadores de viés não comprovaria nossa hipótese, já que é consenso na literatura em Linguística Textual que marcas de subjetividade são encontradas em qualquer texto. Então, assumimos uma perspectiva comparativa, partindo do pressuposto de que, entre os vários gêneros textuais, é possível detectar distinções do grau de subjetividade. Assim, comparamos a distribuição dos operadores nas correções de redações com a distribuição dessas marcas em um *corpus* de resumos acadêmicos (textos com menor grau de subjetividade) e em um *corpus* de resenhas (textos com maior grau de subjetividade). Nosso propósito, ao fazer tal comparação, é verificar de qual polo de subjetividade as correções mais se aproximam.

Para fazer a comparação entre os três gêneros, utilizamos uma ferramenta gráfica muito usada em estatística na análise de comparação de dados: o *boxplot*. O resultado que obtivemos é que textos do gênero correção de redação, em relação ao grau de subjetividade, se aproximam mais do comportamento de textos do gênero resenha do que de textos do gênero resumo acadêmico. Consideramos que tal resultado não é o desejado para um gênero textual que deveria apresentar um baixo grau de subjetividade, devendo, a princípio, se aproximar mais dos textos de gêneros acadêmicos. Esse resultado confirma nossa hipótese e os resultados de Mendes (2013) e coloca em dúvida a forma de correção de redações utilizada atualmente, como um instrumento avaliador isento e justo.

O artigo está organizado da seguinte forma: a próxima seção apresenta os procedimentos metodológicos da pesquisa; a Seção 3 traz a construção da lista de operadores de viés; a Seção 4 apresenta os resultados e análises e a Seção 5 inclui uma breve apreciação da possível extensão desses resultados para as correções de redação do ENEM. A Seção 6 apresenta as considerações finais do trabalho.

2. Metodologia

2.1. Metodologia linguística

Tomamos como objeto principal de investigação os textos de correção de redação publicados no banco de redações do UOL. Esse banco é um serviço *online* que tem como objetivo estimular estudantes a treinar produção de textos do tipo argumentativo. Professores associados ao banco corrigem os textos enviados, que são publicados mensalmente no *site*, juntamente com as respectivas correções. As avaliações são baseadas nos critérios adotados para a correção da redação do ENEM, que são os seguintes: domínio da norma culta do português, compreensão e desenvolvimento do tema baseados em conhecimentos gerais, capacidade de argumentação, conhecimento de mecanismos linguísticos de coesão e elaboração de uma proposta de intervenção, respeitando os valores humanos e a diversidade cultural.

Preocupados em manter um maior rigor na quantificação dos dados, compilamos um *corpus* de correções composto de 610.543 palavras. Para chegar a esse valor, seguimos as afirmações de Aluísio & de Barcellos Almeida (2006) e Sardinha (2002). Conforme Aluísio & de Barcellos Almeida (2006), para estudos de processos gramaticais, é necessário um *corpus* de 500 mil a 1 milhão de palavras. Conforme Sardinha (2002), um *corpus* médio para trabalhos na Linguística de *Corpus* é composto de 500 mil palavras.

Como já mencionamos, além desse *corpus*, utilizamos dois outros *corpora*: de resumos acadêmicos e de resenhas. Tivemos o cuidado de comparar textos do mesmo tipo textual da correção de redação, o argumentativo *stricto sensu*. Conforme Marcuschi (2002), os tipos textuais podem ser classificados em categorias

como a narração, a exposição, a descrição, a injunção (ou prescrição) e a argumentação *stricto sensu*. Esses tipos são definidos por estruturas linguísticas. O tipo argumentativo *stricto sensu* se caracteriza pela sequência de relações entre argumentos e conclusões, sendo marcado pela presença de "operadores argumentativos". Todos os textos analisados aqui se caracterizam por apresentarem esse esquema básico de estrutura linguística, em que argumentos levam a uma conclusão e são marcados pela presença de operadores argumentativos, que podem ser vistos nos exemplos de correção (1), de resumo (2) e de resenha (3):

(1) Na realidade, o texto não argumenta a respeito de aspectos positivos ou negativos da questão, *mas* apenas arrola elementos do senso comum para discorrer sobre o tema.[1]

(2) *Além disso*, os professores e os alunos estão submetidos a uma alta infraestrutura e alto conhecimento em informática...[2]

(3) E ótima facil de usar tem boa imagem e *até* na grvação,exelente. (sic)[3]

Esses textos argumentativos, apesar de compartilharem características tipológicas, se distinguem em gênero, por terem funções sociocomunicativas distintas. O gênero correção de redação tem como função apontar problemas de uma redação, de acordo com uma lista de critérios. Sua finalidade é justificar a nota atribuída à redação. Já o resumo acadêmico se caracteriza pela função de levar o alocutário a compreender o conteúdo de um texto base, sem ter tido acesso a ele. O resumo deve ser sucinto e fiel ao conteúdo do texto original. Resumos acadêmicos obedecem a rígidas regras de formulação, que incluem restrições sobre formatação, tamanho e conteúdo (Motta-Roth & Hendges, 2010; Japiassú, 2013). E a resenha é um gênero que compreende textos em que o locutor decididamente explicita sua opinião sobre algo. Sua função é avaliar um determinado objeto, com base em experiências prévias, e recomendá-lo ou não ao alocutário. Assim, apesar de serem do mesmo tipo, por possuírem funções sociocomunicativas diferentes, esses gêneros também possuem diferentes marcas de subjetividade. Além dos operadores argumentativos, outros tipos de operadores linguísticos também contribuem para a argumentação, tornando os textos mais ou menos subjetivos.

Tendo isso como premissa, e também tendo como norte parâmetros de uniformidade entre os três *corpora*, no que diz respeito ao tamanho de cada texto e ao número total de palavras, foram feitas buscas na *internet* para compor os grupos de comparação. O *corpus* de resumos acadêmicos foi construído através de consultas e extrações automáticas do Google Scholar[4]. As áreas de pesquisa selecionadas foram as mais variadas, desde Ciências da Computação até Artes e Educação Física. Adotamos essa diversidade para mantermos um equilíbrio entre o caráter mais ou menos subjetivo da linguagem adotada nas diversas áreas do conhecimento. O número de palavras total desse corpus é 460.011. E coletamos também textos do gênero resenha, extraídos do *corpus* já compilado "Buscapé" (Hartmann et al., 2014)[5], sobre os mais variados temas, desde câmeras fotográficas, passando por balança e até apontador. Esses textos são comentários feitos por consumidores sobre produtos em *sites* de vendas na *internet*. Esse corpus possui 717.096 palavras.

Após a compilação dos *corpora*, elaboramos, manualmente, a lista de operadores de viés linguístico para ser aplicada automaticamente aos três grupos analisados. Para elaborarmos essa lista, nos valemos da metodologia de análise de "detecção de linguagem enviesada" (Recasens et al., 2013), previamente utilizada para o inglês. Uma primeira ideia foi que poderíamos usar a tradução dessas palavras já analisadas em inglês para a nossa análise de textos no português. Entretanto, como é assumido na área de Semântica Lexical (Levin & Rappaport Hovav, 1995), a tradução de palavras de uma língua para outra não é algo trivial; sempre existem nuances de sentido e muita interferência de contexto sentencial. Com isso, optamos por nos pautar somente na metodologia de desenvolvimento da lista das palavras indutoras de viés e elaborar uma lista específica para o português.

Recasens et al. (2013) propõem analisar exemplos reais de edição de textos da Wikipedia para remover os vieses linguísticos de artigos, que deveriam ser textos isentos de opinião. A ideia é que a partir do entendimento da realização desses vieses, feita por falantes nativos do inglês, pode-se propor um detector de viés automático composto de palavras indutoras de vieses linguísticos. Os autores mostram que o modelo de informação linguística de viés desenvolvido teve uma per-

[1] https://educacao.uol.com.br/bancoderedacoes/redacao/ult4657u82.jhtm. Acesso: 27/08/2019.

[2] Damasceno et al. (2016)

[3] https://sites.google.com/icmc.usp.br/opinando/. Acesso: 05/12/2019.

[4] https://scholar.google.com/

[5] https://sites.google.com/icmc.usp.br/opinando

formance muito próxima da análise de linguagem enviesada feita pelos falantes nativos. Recasens et al. (2013) se valem de trabalhos propostos na literatura que abordam o viés linguístico associado a pistas lexicais e gramaticais sobre subjetividade (Wiebe et al., 2004), sobre sentimento (Liu et al., 2005; Lin et al., 2011; Turney, 2002) e, especialmente, sobre atitudes (Lin et al., 2006; Somasundaran & Wiebe, 2010; Yano et al., 2010; Conrad et al., 2012). Os autores dividem os tipos de vieses em dois grandes grupos: os vieses epistemológicos e os vieses estruturais. Os vieses epistemológicos podem ser detectados pela presença de verbos factivos (Kiparsky & Kiparsky, 1968), verbos implicativos (Karttunen, 1971), verbos assertivos (Hooper, 1974) e operadores de modalização '*hedges*' (Yule, 1996); os vieses estruturais podem ser detectados por intensificadores subjetivos e termos tendenciosos (Lin et al., 2006).

Ao iniciarmos a análise dos nossos textos, porém, percebemos que os dois grandes grupos propostos pelos autores não captavam uma questão fundamental para a compreensão e detecção dos vieses de forma automática, que seria o problema do sentido dependente de contexto. Dividimos, então, as palavras e expressões que comporiam a nossa lista em dois grandes grupos distintos dos propostos por Recasens et al. (2013): expressões carregadas de sentido que variavam a significação segundo o contexto, ou seja, expressões dependentes de contexto (geralmente, as classes gramaticais abertas, como nomes, verbos e adjetivos) e expressões cujos significados não dependem de contexto, consideradas mais "leves" semanticamente (geralmente, as classes gramaticais fechadas, como advérbios, preposições e conjunções). Pode-se associar os tipos de vieses propostos por Recasens et al. (2013) a esses dois grupos da seguinte forma: expressões dependentes de contexto são os termos tendenciosos, os verbos implicativos e os verbos assertivos; expressões não-dependentes de contexto são os operadores de modalização, os verbos factivos e os intensificadores subjetivos.

Com esses pressupostos teóricos e a nossa intuição de falantes do português, partimos para o trabalho minucioso da análise dos textos coletados. Fizemos primeiramente a análise de 50 correções, aleatoriamente selecionadas, apontando, segundo os critérios linguísticos acima, todas as palavras e expressões que nos pareciam indicar algum tipo de subjetividade do corretor.

Entretanto, ao examinarmos expressões dependentes de contexto, deparamo-nos com outro problema. Por exemplo, detectamos a seguinte sentença no nosso *corpus* de correções de redação: *a frase final está abstrata demais*[6]. A palavra *abstrata* seria um operador de viés. O que se espera de um corretor de redação são avaliações sobre a coesão, domínio da norma culta etc. e não a sua opinião sobre "o que é ser abstrato". Portanto, para se detectar um viés linguístico em correções de redação, essa é uma palavra adequada. Por outro lado, se pensamos em um resumo acadêmico na área de Artes, em que o autor descreve uma *pintura abstrata*, certamente não teremos viés nessa palavra. Como essa, percebemos outras tantas expressões que se comportavam de acordo com o tipo de conteúdo do texto.

Considerando o nosso objetivo de estabelecer uma lista de operadores de viés linguístico que gere um modelo de detecção de viés para qualquer tipo de texto, na nossa proposta, precisamos que essa lista detecte as palavras e expressões enviesadas de correções de redação, de resumos acadêmicos e de textos opinativos de forma indistinta. A partir daí, pretendemos que essa lista seja suficientemente abrangente para a utilização em modelos computacionais para qualquer gênero textual. Contudo, estabelecer os vários contextos para que o significado de expressões com sentidos dependentes de contexto seja detectado em um modelo computacional não é uma tarefa trivial. Percebemos mesmo que a lista de expressões que estávamos gerando apresentava, ela própria, uma natureza enviesada, baseada na nossa intuição de falantes. Em vista desses resultados preliminares, optamos por seguir um caminho distinto do seguido em Recasens et al. (2013) e resolvemos descartar as expressões dependentes de contexto e criar uma lista somente com palavras não-dependentes de contexto.

Isso não significa, porém, que elementos de viés dependentes de contexto não serão também detectados nesse tipo de lista. É importante realçar que essas palavras e expressões funcionam como operadores sobre outros itens lexicais, apontando assim para uma porção maior do grau de subjetividade do texto. Fazendo uma analogia, *grosso modo*, com a Matemática, as palavras não-dependentes de contexto funcionam como os operadores da adição, subtração, multiplicação etc., que apontam para a relação pretendida entre os números, gerando um produto final. Operadores linguísticos funcionam da mesma forma: estabelecem uma relação entre parcelas do texto, gerando uma interpretação subjetiva como produto. Assim, detectar operadores de viés não-

[6] https://educacao.uol.com.br/bancoderedacoes/redacao/ilusao.jhtm. Acesso: 06/09/2019.

dependentes de contexto em um determinado texto, mesmo que em pouca quantidade, indicará também a presença de viés em outras partes de texto, sobre as quais esses operadores têm escopo.

Dessa forma, voltamos nossa atenção para as palavras e expressões não-dependentes de contexto, que apresentam o mesmo tipo de significação em qualquer situação. À primeira vista, essas palavras e expressões parecem apresentar menos conteúdo semântico, tendo uma função essencialmente gramatical. Entretanto, para os estudiosos da Semântica Argumentativa, essas expressões são alvo de muita atenção e pesquisa. O que se assume nessa linha teórica é que essas palavras ou expressões "despretensiosas" são as responsáveis por grande parte da força argumentativa dos textos, ou seja, a direção ou sentido para o qual apontam. Segundo Anscombre & Ducrot (1976); Ducrot (1987), toda língua possui em sua gramática mecanismos que captam essa subjetividade através de marcas linguísticas da argumentação; entre essas marcas, encontram-se as palavras e expressões que denominamos "não-dependentes de contexto".

Com esse procedimento, elaboramos uma lista de 578 itens. Baseados na classificação encontrada em Koch (2011, 2015), dividimos esses operadores em quatro tipos: operadores argumentativos, operadores de pressuposição, operadores de modalidade e operadores de opinião e valoração[7].

A partir desse trabalho metodológico, a lista construída foi aplicada a cada um dos *corpora* mencionados acima e os operadores de viés de cada texto foram quantificados e comparados. A seguir, indicamos com mais detalhes como foram feitas as compilações dos *corpora* e como fizemos a quantificação dos operadores de viés nas correções de redação e a comparação com os demais gêneros analisados.

2.2. Metodologia computacional

A compilação dos *corpora* utilizados nesta pesquisa foi feita de forma automática, a partir de buscas em bancos de textos dos gêneros analisados. Nosso *corpus* principal se compõe de textos do gênero correção de redação. Esse *corpus* foi construído a partir da extração automática de textos do banco de redações do UOL, considerando-se um período de cerca de 10 anos. Relembrando, o número total de palavras é de 610.543.

Adicionalmente, construímos dois outros *corpora* para comparação: um de resumos acadêmicos, com 460.011 palavras e outro de resenhas, com 717.096 palavras[8]. Para a composição do *corpus* de resumos, os textos foram extraídos de consultas automáticas ao Google Scholar, considerando-se diversas áreas do conhecimento, retiradas do site da Universidade Federal de Minas Gerais. Para se determinar os tipos de consulta, lançamos mão do repositório de teses e dissertações dessa universidade[9], selecionando as palavras-chave em teses ou dissertações. Usamos um programa que fazia consultas e *downloads* de artigos no Google Scholar e foram extraídos os resumos desses textos. Se a quantidade de palavras era pequena (menos de 50) ou muito grande (mais de 1000), os textos eram descartados. Foi feita uma verificação manual de alguns desses textos e constatamos que a metodologia de extração de dados era confiável.

O *corpus* de resenhas consiste num fragmento, também extraído de forma automática, do *Corpus* Buscapé. A utilização de apenas um fragmento do *Corpus* Buscapé foi necessária para manter a uniformidade com os demais *corpora*, tanto no número total de palavras quanto no tamanho dos textos. Utilizamos apenas as resenhas do *Corpus* Buscapé com mais de 200 palavras, pois muitas delas eram curtas e não serviam à nossa comparação, e tentamos manter um número próximo ao total de palavras dos demais *corpora*.

No processamento computacional que realizamos desses dados, cada texto, seja uma correção de redação, uma resenha ou um resumo, foi considerado um "documento", representado por *d*. Cada documento foi processado a fim de se computar a proporção de ocorrências de operadores de viés, em relação ao número total de palavras do texto. A metodologia utilizada para fazer esse cálculo seguiu os seguintes passos:

1. Divisão dos documentos em *tokens*.

Dividimos os documentos em unidades linguísticas menores: as palavras e expressões fixas comparáveis a palavras, que chamamos de *tokens*. Os documentos foram, assim, *tokenizados* e essas unidades, ou *tokens*, foram identificadas a partir de sua separação gráfica

[7] Esse léxico de operadores de viés está disponível em https://bit.ly/2sJzt21

[8] Apesar de haver uma diferença no número de palavras total dos *corpora* (de 25% entre os resumos e as correções e de 35% entre as resenhas e as correções), os cálculos da frequência de ocorrência dos operadores de viés em cada *corpus* foram feitos proporcionalmente. Assim, essa diferença não teve impacto em nossos resultados.

[9] https://repositorio.ufmg.br/custom/presentation.jsp

por símbolos, como espaço e hífen. Utilizamos para esse fim a função *word_tokenize*, da biblioteca de código NLTK[10]. Essa biblioteca possui diversas funcionalidades prontas para uso em processamento de linguagem natural. A partir desse processo, é possível identificar e quantificar operadores de viés em um texto, já que ocorrências desses operadores fazem parte do conjunto de *tokens* de um documento.

2. Classificação de partes do discurso.

O próximo passo da nossa metodologia foi determinar a categoria gramatical dos *tokens*, em termos de classes de palavras. Para isso, utilizamos um classificador de partes do discurso, o *POS Tagging*, também da biblioteca NLTK. Para construir o modelo do classificador de partes do discurso, a base de dados utilizada foi o Mac Morpho (Aluísio et al., 2003) e o código para sua construção pode ser facilmente reproduzido.

3. Lematização de operadores verbais.

Categorizar os *tokens* nos documentos analisados foi essencial para esta terceira etapa metodológica, a lematização. Lematização é uma técnica utilizada para buscar palavras, abrangendo um paradigma de opções relacionado a elas. No caso de verbos que se encontram na nossa lista de operadores de viés, é feita a lematização, de forma que seja possível a busca por todo o paradigma de flexões desse item. Um exemplo é o verbo *começar*, para o qual a utilização do lematizador possibilita a busca por todo o paradigma flexional: *comecei, começamos, começando* etc. Para palavras invariáveis e expressões fixas, mesmo que sejam verbais, não é feita a lematização. Por exemplo, as expressões *desde que* ou *é bom que* não passam pelo lematizador, sendo buscadas sempre na mesma forma fixa. Esse processo permite que sejam identificados, dentre os *tokens* de um documento, todas as formas possíveis de um operador de viés verbal. Para esta etapa, utilizamos a linguagem de programação Java, pois a biblioteca de código, chamada de LemPORT, é baseada em Java. O processo utilizado por esta biblioteca de lematização está descrito em Rodrigues et al. (2014).

4. Uniformização dos *tokens* através da utilização de caixa baixa.

A terceira etapa da metodologia computacional consistiu em colocar todos os caracteres dos *tokens* de cada documento em letras minúsculas, para que esses *tokens* se tornem "visíveis" a programas computacionais utilizados no processamento dos documentos.

5. Identificação dos operadores de viés em cada documento.

A fim de contabilizarmos o número de ocorrências dos operadores de viés nos documentos dos *corpora*, primeiramente, identificamos em cada documento os *tokens* correspondentes a operadores da nossa lista. Para isso, comparamos os *tokens* dos documentos de cada *corpus* com cada operador de viés listado. Sendo o *token* uma ocorrência de um desses operadores, ele seria, então, computado em um acumulador que chamamos de i_c. O resultado final desse cálculo i_c é, portanto, o número de ocorrências de operadores de viés de um determinado tipo em um documento. Cada documento apresentou, assim, quatro números desse tipo, um para cada tipo de operador de viés.

6. Quantificação de ocorrências de operadores de viés em relação ao número de palavras.

A partir do número de i_c foi possível, então, contabilizar o percentual de *tokens* de um determinado documento que se caracterizavam como operadores de viés. Esse cálculo foi feito da seguinte forma. Para cada documento e para cada tipo de operador de viés, dividimos o acumulador i_c pelo número de *tokens* do documento, representado por $|d|$. Assim, a proporção de ocorrências de operadores de viés, de cada tipo, representada por P_c, é dada pela fórmula

$$P_c = \frac{i_c}{|d|}.$$

Os passos descritos acima foram implementados na linguagem de programação Python em sua versão 3[11]. E com essas etapas metodológicas chegamos ao número de proporção de ocorrências de operadores de viés por documento. Após esse procedimento, analisamos estatisticamente os resultados, a fim de fazer uma comparação entre os três gêneros. Nosso objetivo não foi somente computar os operadores do gênero em análise, as correções, mas observar se esses textos se aproximam mais dos resumos acadêmicos, o polo menos subjetivo, ou das resenhas, o polo mais subjetivo. Para isso usamos uma ferramenta estatística, que será apresentada a seguir.

[10] https://www.nltk.org

[11] https://www.python.org

2.2.1. Ferramenta de análise estatística

Como ferramenta de análise estatística, utilizamos o diagrama de caixa, ou *boxplot*. O *boxplot* é uma ferramenta não-paramétrica, auto evidente, e não deixa dúvidas quanto à distribuição empírica dos dados: esse gráfico permite visualizar a distribuição de uma variável em termos da sua locação (mediana/quartis), dispersão (variabilidade), grau de assimetria, presença de valores extremos/discrepantes (*outliers*), entre outros. Para gerar os *boxplot*, utilizamos a linguagem R[12]. A Figura 1 mostra um *boxplot* padrão, que está marcado com seus principais elementos.

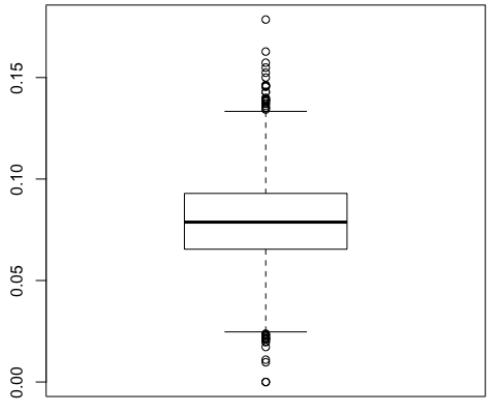


Figura 1: Exemplo de *boxplot*

Essa figura contém uma "caixa", o retângulo central da figura, com uma linha horizontal mais grossa, duas linhas horizontais menores, uma linha vertical e pontos acima e abaixo das linhas horizontais menores, que aparecem fora da caixa. Cada um desses elementos gráficos possui um significado estatístico, em relação à variação dos dados observados de uma variável numérica. Assim, o gráfico permite a visualização de valores que dividem um conjunto de dados ordenados em quatro partes iguais, que são chamados de "quartis".

A primeira parte, ou primeiro quartil, é o valor que divide 25% dos menores valores do conjunto de dados dos 75% restantes. O primeiro quartil é representado pela linha inferior da caixa (abaixo da caixa estão os 25% dos dados que possuem menor valor). O segundo quartil, ou a mediana, é o valor que divide os dados ordenados ao meio. A mediana é definida como o valor que está no meio de um conjunto de valores (Lane et al., 2017). Outra forma de pensar a mediana é que metade do conjunto de valores, 50%, será menor que o valor da mediana e a outra metade do conjunto, 50%, será maior. Por esta razão, a mediana também é chamada de percentil 50. A linha horizontal mais grossa representa a mediana. O terceiro quartil representa o valor que separa dos demais os 25% dos dados que possuem o maior valor. No *boxplot*, o terceiro quartil é graficamente representado pela linha horizontal mais acima na caixa. As linhas horizontais menores, que aparecem fora da caixa, indicam os limites dos dados até 1,5 vezes a distância entre os quartis (distribuição interquartílica). Os pontos acima e abaixo dessas linhas representam valores discrepantes, ou *outliers*, do conjunto. Esses pontos mostram valores que estão no extremo da distribuição. Por fim, o tamanho da caixa está relacionado com a variabilidade do conjunto de dados. Assim, se a distância entre a mediana e os *outliers* for maior de um lado do que do outro, os dados possuem uma distribuição assimétrica; caso contrário, simétrica. Na Figura 1, o tamanho da caixa nos dois quartis é similar, portanto, a distribuição dos dados é simétrica.

O *boxplot* é muito útil para a visualização de diferentes conjuntos de dados com relação a uma variável. Este é o caso deste trabalho, já que temos três conjuntos de dados (três gêneros textuais) e quatro variáveis (quatro tipos de operadores de viés). Portanto, para cada tipo de operador, geramos um *boxplot* e analisamos os resultados. Na seção seguinte, antes de partirmos para a explicitação desses resultados, explicamos com detalhes cada um desses tipos de operadores.

3. Léxico de operadores de viés linguístico

Conforme Anscombre & Ducrot (1976) e Ducrot (1987) (e trabalhos subsequentes), a argumentação, em um sentido mais amplo, está inscrita na língua, não depende exclusivamente de um tipo textual e pode ser guiada por itens específicos da gramática. Como mostramos na metodologia (2.1), o que nos interessa na nossa análise são exatamente esses itens, que apontam para a direcionalidade da argumentação, deixando à mostra o grau de subjetividade do texto. Portanto, são essas marcas linguísticas, que chamamos de "operadores de viés linguístico", que fazem parte da nossa lista, que pode ser também entendida como um "léxico". Lembramos que nosso foco é em expressões não-dependentes de contexto e que essas expressões são operadores que agem sobre porções maiores de texto, resultando em sentenças com implicações subjetivas.

Seguimos as propostas de Koch (2011, 2015), para quem as marcas de argumentação em um texto podem ser expressas através das seguintes categorias: operadores argumentativos, ope-

[12]Disponível em https://www.rdocumentation.org/packages/graphics/versions/3.6.1/topics/boxplot.

radores de pressuposição, operadores de modalidade, operadores de opinião e valoração, tempos verbais e índices de polifonia. Usamos as quatro primeiras categorias, marcas possíveis de serem utilizadas em um detector automático de viés linguístico. Também ampliamos a lista dada pela autora, usando expressões encontradas em nossos dados. A lista completa é composta por 578 itens, sendo 118 operadores argumentativos, 212 operadores de pressuposição, 93 modalizadores e 155 operadores de opinião e valoração. A seguir explicitamos essas quatro categorias[13].

3.1. Operadores argumentativos

Os operadores argumentativos têm a função de estabelecer uma ligação entre as orações, períodos ou parágrafos, deixando transparecer a intenção do locutor. São as estruturas que caracterizam os textos de tipo argumentativo *stricto sensu*. Segundo Koch (2011, 2015), esses operadores encontram-se divididos em alguns tipos.

Um primeiro tipo é o dos operadores que assinalam o argumento mais ou menos proeminente de uma escala orientada no sentido de determinada conclusão, os chamados "operadores de escala argumentativa". Operadores como esses mostram que o locutor constrói uma escala de argumentos, sendo o mais forte ou o mais fraco introduzido pelo operador. Por exemplo:

(4) A escolha das palavras é, em geral, confusa e ambígua, comprometendo a compreensão das ideias e *até mesmo* mantendo num nível superficial a análise que faz do tema.[14]

A sentença em (4), retirada do *corpus* de correções, mostra que o operador *até mesmo* indica que há uma escala construída pelo locutor: confusa → ambígua → superficial, em que o valor alto da escala, e talvez não-esperado, "ser superficial", reflete a opinião de quem usa o operador.

Um segundo tipo de operador argumentativo é o que soma um argumento a favor da afirmação feita anteriormente, de uma forma positiva ou mesmo negativa:

(5) A objetividade, *aliás*, é o ponto mais alto dessa redação.[15]

Na sentença (5), pode-se perceber que há um desfecho glorioso em relação às propriedades da redação, na perspectiva de quem elabora o enunciado.

Um terceiro tipo de operador argumentativo são os que introduzem a conclusão à qual o locutor quer que seu alocutário chegue, a partir dos argumentos apresentados anteriormente:

(6) Não desenvolve, *portanto*, uma dissertação argumentativa.[16]

Em (6), *portanto* indica que o fato de que o aluno não desenvolve uma dissertação argumentativa é uma conclusão, que pode ser tirada a partir de argumentos elencados anteriormente. O operador é utilizado pelo produtor do texto como um "guia" para que o alocutário compartilhe com ele essa mesma posição.

Um quarto tipo de operador introduz argumentos alternativos que marcam uma conclusão oposta à afirmação anterior, indicando a dúvida do locutor:

(7) Seria melhor substituir a expressão "o casal e os amigos" por "amigos" somente *ou então* por "par".[17]

Um quinto tipo é o que estabelece relações de comparação entre elementos, indicando uma dada conclusão por parte do locutor:

(8) Mal o aluno acaba de falar de uma coisa, ele pula para outra *como se* continuasse a falar da mesma.[18]

Um sexto tipo é o que estabelece relações de condição entre elementos, indicando uma imposição por parte do locutor:

(9) Isso se chama tautologia, e deve ser evitado; pode-se repetir, *desde que* avançando nas ideias.[19]

[13] O léxico completo com todos os operadores encontra-se disponível em https://bit.ly/2sJzt21.

[14] https://educacao.uol.com.br/bancoderedacoes/redacao/a-importancia-de-biografias-autorizadas.jhtm. Acesso: 06/09/2019.

[15] https://educacao.uol.com.br/bancoderedacoes/redacao/educacao-e-mais-forte-que-violencia.jhtm. Acesso: 06/09/2019.

[16] https://educacao.uol.com.br/bancoderedacoes/redacao/bandido-bom-e-bandido-recuperado.jhtm?action=print. Acesso: 11/09/2019.

[17] https://educacao.uol.com.br/bancoderedacoes/redacao/ult4657u564.jhtm. Acesso: 06/09/2019.

[18] https://educacao.uol.com.br/bancoderedacoes/redacoes/policia-e-bandido-violencia-no-seculo-xxi.htm. Acesso: 06/09/2019.

[19] https://educacao.uol.com.br/bancoderedacoes/redacao/ult4657u488.jhtm. Acesso: 06/09/2019.

Um sétimo operador introduz uma justificativa ou explicação do locutor, em relação ao enunciado anterior:

(10) . . . *desse modo*, fica impossível atribuir alguma nota às competências 2, 3 e 4.[20]

Um oitavo tipo contrapõe argumentos, indicando conclusões contrárias:

(11) . . . a análise das causas é simplificadora e superficial, *apesar de* plausível[21]

O locutor faz uma afirmação, mas introduz uma contraposição ao seu argumento, criando uma expectativa contrária no alocutário. Esses operadores ligam dois argumentos opostos, mas sempre colocam mais relevância para um dos argumentos, aquele que corrobora a conclusão pretendida.

3.2. Operadores de pressuposição

Um segundo grupo de operadores de viés linguístico são expressões que têm a função de introduzir no texto um conteúdo pressuposto, aos quais chama-se na literatura de "desencadeadores de pressuposição". Quando usamos itens linguísticos que trazem em si uma verdade assumida anteriormente, estamos impondo aos alocutários essa verdade, o que nem sempre pode ser um fato, mas pode corresponder à opinião de quem a enuncia, ou seja, o uso de operadores de pressuposição é uma "manobra argumentativa"(Fiorin, 2007).

Um primeiro tipo desses operadores (*agora, ainda, atualmente* etc.) é exemplificado a seguir, com trechos retirados das próprias correções que compõem o nosso *corpus*.

(12) Acrescenta-se uma informação *já* esboçada anteriormente....[22] (pressuposto: a informação não devia aparecer novamente)

Além dessas marcas temporais, existem alguns conectores circunstanciais, sobretudo quando a sentença introduzida por eles vem anteposta, que são desencadeadores de pressuposição; toma-se como fato o conteúdo da sentença introduzida pelo conector:

(13) *Depois de* um caótico e redundante parágrafo introdutório, o autor começa a dizer contrassensos ou a usar conceitos vagos. . .[23] (pressuposto: o parágrafo introdutório é caótico e redundante)

Outros exemplos de operadores de pressuposição são expressões verbais que denotam uma mudança ou uma manutenção do estado; essas expressões pressupõem o estado anterior à mudança ou a ocorrência prévia do estado que se mantém; por exemplo:

(14) O que o autor parece não perceber é que o trote vem *se tornando* mais violento hoje em dia. (pressuposto: o trote antes era menos violento)

Um último tipo de operador de pressuposição são os verbos chamados "factivos". Quando empregados, esses verbos desencadeiam a pressuposição da verdade do seu complemento sentencial. Um exemplo desse tipo de verbo é *notar*:

(15) *Note-se* ainda que o texto não justifica o título.[24]

Na sentença acima, toma-se como verdade que "o texto não justifica o título". Quando expressões contendo esses verbos são usadas, há uma indução do alocutário a aceitar uma opinião do locutor como verdade.

3.3. Operadores de modalidade

Os operadores de modalidade, ou modalizadores, são elementos lexicais (e gramaticais, que não abordaremos aqui) pelos quais o locutor manifesta uma determinada atitude em relação ao conteúdo de seu próprio enunciado. Pode-se dividir, de uma maneira mais ampla, os modalizadores em dois tipos, segundo Pires de Oliveira (2001): os de possibilidade e os de necessidade. Esses dois tipos de modalizadores podem expressar a possibilidade e a necessidade de acordo com normas morais ou legais, os modalizadores deônticos, ou podem expressar a necessidade e a possibilidade em relação ao conhecimento e à crença do locutor, os modalizadores epistêmicos.

[20] https://educacao.uol.com.br/bancoderedacoes/redacoes/uma-perda-historica-para-o-pais.htm. Acesso: 06/09/2019.

[21] https://educacao.uol.com.br/bancoderedacoes/redacao/violencia-comeca-em-casa.jhtm. Acesso: 05/09/2019

[22] https://educacao.uol.com.br/bancoderedacoes/redacoes/violencia-gera-consequencia.htm. Acesso: 06/09/2019.

[23] https://educacao.uol.com.br/bancoderedacoes/redacoes/brasil-por-uma-migracao-segura-ordenada-e-regular.htm. Acesso: 06/09/2019.

[24] https://educacao.uol.com.br/bancoderedacoes/redacao/os-desafios-da-interacao.jhtm. Acesso: 06/09/2019.

Esses operadores linguísticos podem vir lexicalizados na forma "verbo *ser* + adjetivo"):

(16) *É preciso* evitar esse tipo de redundância, resultante da falta de planejamento na criação do texto.[25] (necessidade)

Os modalizadores também podem aparecer na forma de advérbios ou locuções adverbiais (*certamente, sem dúvida, de fato*):

(17) Texto fraco, marcado pelo uso inadequado do vocabulário, em que as palavras são usadas com significados que, *de fato*, não têm.[26] (necessidade)

Ainda, modalizadores podem ser verbos modais:

(18) É o máximo que *se pode* extrair de um texto cuja linguagem obscura parece usada para iludir e impressionar o alocutário...[27] (possibilidade)

A modalização aparece também em orações modalizadoras (*tenho a certeza de que, existe a possibilidade de* etc.):

(19) *Não há necessidade de* usar essa locução no trecho assinalado.[28] (possibilidade)

E, por fim, aparece em certos tipos de verbos que denotam crença ou sentimento (*achar, acreditar, desejar* etc.):

(20) O autor *crê que* o leitor deva subentender o que ele está dizendo com essas expressões incorretas, mas não é assim que funciona a comunicação escrita.[29] (possibilidade)

3.4. Operadores de opinião e valoração

Finalmente, apresentamos os operadores de opinião e valoração. Operadores de opinião apontam para um determinado estado psicológico, um certo tipo de sentimento, de opinião que o locutor assume em relação ao conteúdo de seu enunciado. Podem ser indicadores desse tipo expressões adverbiais, expressões verbais e alguns adjetivos que não têm alteração do sentido segundo o contexto. Veja um exemplo:

(21) Além disso, em vermelho, é *ruim* usar duas vezes seguidas o "porém, atualmente"[30]

Ainda, dentro dessa categoria, podem-se incluir os operadores de valoração. Segundo Koch (2015), a atitude subjetiva do locutor pode ser medida também pela avaliação e valoração dos fatos, estados ou qualidades atribuídas a um referente. Em geral, essa medição se dá por expressões adjetivas e intensificadores. Veja um exemplo de valoração extraído das correções:

(22) A sugestão que conclui a redação também é retórica e panfletária, propondo manifestações contrárias ao que está aí, que foi analizado (sic) de modo *excessivamente* genérico...[31]

Apresentamos, em seguida, os resultados da análise do viés linguístico nas correções, feita a partir da aplicação automática do léxico dos operadores de viés. Apresentamos também o resultado comparativo da aplicação dessa lista nos textos dos nossos outros dois *corpora*, o de resumos acadêmicos e o de resenhas.

4. Resultados e análise dos dados

Antes de procedermos à análise, é importante observar que os números de porcentagens obtidos podem não ser expressivos numericamente, mas indicam com clareza como se dá a distribuição dos operadores nos três *corpora* comparados. Como já observado, o número de operadores em relação ao número de outras categorias linguísticas que compõem um texto é muito pequeno, assim como são poucos os operadores matemáticos em relação aos números. Então, é de se esperar que a porcentagem dessas ocorrências nos textos analisados seja um número baixo. Mas, como mostramos na Seção 3, esses operadores agem sobre expressões linguísticas e até mesmo sobre sentenças inteiras, resultando em porções maiores de texto com implicações subjetivas. Analisando dessa forma, o grau de subjetividade não se encontra somente no número absoluto de operadores, mas, sim, em todo o produto

[25] https://noticias.uol.com.br/educacao/bancoderedacoes/redacao/ult4657u67.jhtm Acesso: 06/09/2019

[26] https://educacao.uol.com.br/bancoderedacoes/redacao/horario-apolitico.jhtm. Acesso: 05/09/2019.

[27] https://educacao.uol.com.br/bancoderedacoes/redacoes/menos-influencia-estatal-nao-significa-menos-ordem.htm. Acesso: 05/09/2019.

[28] https://educacao.uol.com.br/bancoderedacoes/redacao/ult4657u140.jhtm. Acesso: 05/09/2019

[29] https://educacao.uol.com.br/bancoderedacoes/redacoes/justica-pra-alguns-stf-pra-outros.jhtm. Acesso: 18/12/2019

[30] https://noticias.uol.com.br/educacao/bancoderedacoes/redacao/ult4657u27.jhtm. Acesso: 05/09/2019.

[31] https://educacao.uol.com.br/bancoderedacoes/redacoes/educacao-brasileira-conhecimento-ja.htm. Acesso: 05/09/2019.

final da interpretação daquela porção linguística gerada pelo operador. Por isso, para uma análise automática, assumimos que a presença dos operadores nos textos nos dá evidências quantitativas de que uma porção maior do texto apresenta um grau de viés linguístico.

Para iniciar a análise, apresentamos, na Tabela 1, o número absoluto de ocorrências de cada tipo de operador de viés em cada *corpus* analisado.

Tipo de operador de viés	Gênero		
	Correção	Resenha	Resumo
Argumentativo	48912	39282	18949
Opinião/valoração	12742	22604	3903
Pressuposição	6081	10750	947
Modalidade	5446	4539	1074

Tabela 1: Ocorrências de operadores de viés por gênero

Os números da Tabela 1 mostram que os operadores argumentativos e de opinião e valoração ocorrem em maior quantidade que os demais operadores, nos três gêneros analisados. Ainda, as correções apresentam números mais próximos da resenha que do resumo.

Porém, uma análise por valores absolutos não é suficiente para mostrar e comparar como essas ocorrências são distribuídas pelos textos específicos que compõem cada *corpus*. Por isso, analisamos os números de ocorrências de cada tipo de operador distribuídos por cada texto de cada gênero, o que nos dá a proporção desses operadores em relação ao número de *tokens* de cada documento, a partir da fórmula $P_c = \frac{i_c}{|d|}$, como explicitado na metodologia. Partindo-se, então, dessa proporção, foram gerados *boxplots*, a fim de facilitar a visualização dessa distribuição. Os diagramas das Figuras 2, 3, 4 e 5, ilustrativos para cada tipo de operador de viés, são mostrados a seguir.

Primeiramente, temos o diagrama que mostra o comportamento dos operadores argumentativos nos três gêneros analisados na Figura 2.

Na Figura 2, a linha no meio dos retângulos aponta para a mediana das ocorrências por *corpus*. Comparando os três grupos, a mediana nas correções de redação é de 0,08, nas resenhas é de 0,05 e nos resumos é de 0,04. Esses números indicam que a porcentagem de ocorrência desses operadores nas correções é maior do que nas resenhas e nos resumos, o que pode ser verificado visualmente no diagrama. Essa constatação também pode ser evidenciada pelos dados da Tabela 1, que apresenta o valor absoluto das

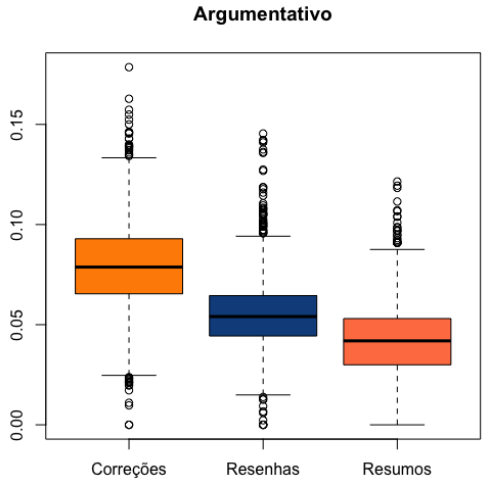


Figura 2: *Boxplot* para operadores argumentativos

ocorrências dos operadores argumentativos nos três gêneros. As correções também apresentam o maior número desses operadores em valor absoluto.

Outra observação a respeito dos dados da Tabela 1 é que os operadores argumentativos ocorrem em um número bem maior do que os outros operadores. Isso se deve ao fato de serem os três gêneros do tipo textual argumentativo e, por isso, apresentam mais operadores desse tipo. Esse é um resultado esperado. Entretanto, a diferença grande entre as medianas e a diferença numérica entre os gêneros é um resultado inesperado: sendo esses textos do mesmo tipo, argumentativo, era de se esperar uma distribuição e quantidade de operadores mais equilibrada entre eles. Como operadores argumentativos são marcas de subjetividade, isso nos leva a uma primeira análise de que, em relação a essas marcas, o gênero correção de redação já apresenta um maior grau de viés linguístico, na comparação dos três gêneros.

Os pontos de discrepância, ou seja, os *outliers*, que representam 1% da população para cima ou 1% da população para baixo, também estão presentes nos três gêneros, mas isso indica apenas que alguns textos contêm pouquíssimos operadores ou muito mais operadores do que os 50% distribuídos dentro do retângulo, não trazendo significação para a análise de subjetividade.

O segundo diagrama a ser apresentado é o dos operadores de opinião e valoração que está na Figura 3.

No diagrama da Figura 3, as medianas dividem a distribuição dos operadores de opinião e valoração de forma assimétrica nos três gêneros. A mediana nas correções de redação é de 0,02,

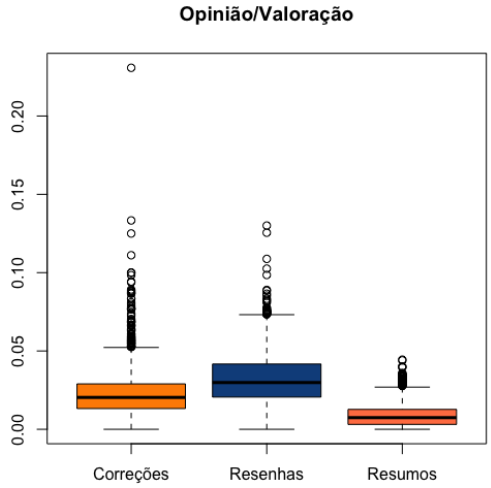


Figura 3: *Boxplot* para operadores de opinião e valoração

nas resenhas é de 0,03 e nos resumos é muito próxima de 0. Esses números indicam que a proporção desses operadores nas resenhas é maior do que nas correções e nos resumos, o que pode ser verificado visualmente no diagrama. Esse é um resultado esperado, já que resenhas certamente vão apresentar mais palavras que expressem opinião, visto que a função do gênero é precisamente avaliar. Resumos já apresentam um grau baixíssimo de operadores de opinião e valoração, já que sua função é relatar resultados de uma pesquisa acadêmica da forma menos subjetiva possível. Porém, pensando nas correções e em sua função de ser um instrumento avaliativo isento, o número de ocorrências dos operadores de opinião e valoração deveria estar mais próximo dos resumos, não das resenhas. E, como podemos observar, não é isso que ocorre, já que as correções se aproximam mais da resenha do que do resumo acadêmico, em relação a esses marcadores e, consequentemente, ao grau de subjetividade. Essa é mais uma evidência em direção à nossa hipótese inicial, de que as correções são mais subjetivas do que se esperaria.

Também os dados da Tabela 1 evidenciam que, em termos de valor absoluto das ocorrências dos operadores de opinião e valoração, as correções se aproximam mais das resenhas do que dos resumos. É uma grande diferença em termos numéricos: as resenhas apresentam 22.604 ocorrências desses operadores, as correções 12.742 e os resumos apenas 3.903.

Os *outliers* do diagrama da Figura 3 ainda nos dão um outro tipo de evidência: quando existe algum desvio nas correções, esses desvios são somente para cima, ou seja, documentos que se desviam do comportamento da maioria apresentam sempre um número maior de operadores de opinião e valoração e chegam mesmo a ultrapassar os desvios dos textos de resenhas, em amplitude e em quantidade. Não seria de se esperar que pudessem aparecer no *corpus* correções com um grau de viés maior do que o das resenhas.

Passemos, agora, para os diagramas com os operadores de pressuposição e de modalidade, que estão respectivamente nas Figuras 4 e 5.

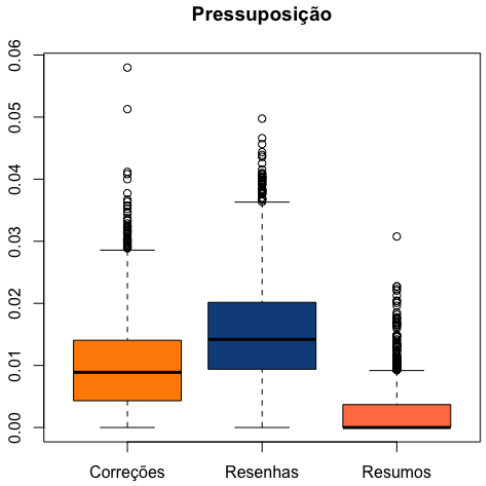


Figura 4: *Boxplot* para operadores de pressuposição

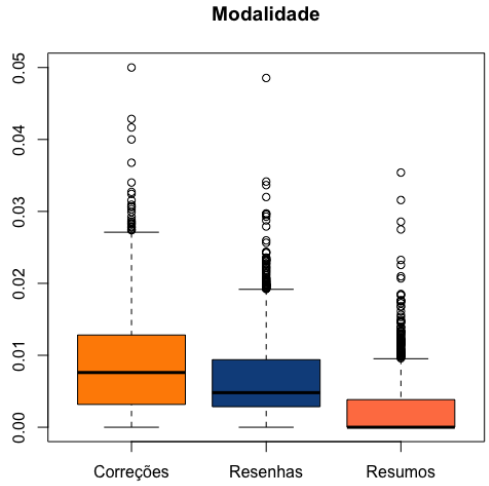


Figura 5: *Boxplot* para operadores de Modalidade

Esses dois tipos de operadores apresentam uma menor relevância no grau do viés, pois, como pode ser visto nos diagramas, as medianas dos três gêneros não ultrapassam 0,01. Entretanto, ainda assim, é possível notar visualmente nos *boxplots* que a mediana das correções, tanto na pressuposição como na modalidade, se aproxima mais das resenhas. E, no número dos operadores de modalidade, as correções chegam a ultrapassar as resenhas. Mostramos na Tabela 2 as medianas desses dois operadores.

Tipo de operador de viés	Gênero		
	Correção	Resenha	Resumo
Pressuposição	0,009	0,014	0
Modalidade	0,008	0,005	0

Tabela 2: Medianas para operadores de pressuposição e modalidade

Em relação aos *outliers*, um fato interessante a se notar é que os desvios ocorrem sempre em uma direção maior de viés e em uma maior amplitude nas correções, assim como ocorre para os operadores de opinião e valoração.

Também nos valores da Tabela 1 as correções se aproximam mais das resenhas do que dos resumos, apresentando um valor maior para os modalizadores. O que se pode concluir desse padrão é que também no comportamento desses dois operadores, ainda que em menor relevância, as correções se aproximam mais das resenhas, inclusive ultrapassando-as na categoria de modalidade. Esse comportamento não seria esperado, dada a função do gênero correção de redação. Esses resultados corroboram, novamente, a hipótese levantada.

Como última evidência para a nossa hipótese, fizemos uma comparação em termos de percentual de operadores por tipo de gênero, tomando cada *corpus* como um único extrato do gênero e como sendo uma população homogênea. Obtivemos o resultado descrito na Tabela 3.

Tipo de operador de viés	Gênero		
	Correção	Resenha	Resumo
Argumentativo	0.080	0.055	0.041
Opinião/valoração	0.020	0.032	0.008
Pressuposição	0.010	0.015	0.002
Modalidade	0.009	0.006	0.002

Tabela 3: Percentual dos operadores por gênero

Os dados da Tabela 3 corroboram todas as análises anteriores, feitas para a distribuição desses operadores por documentos de cada gênero. Primeiramente, pode-se notar que o percentual de ocorrências de operadores argumentativos, nos três gêneros, também é maior do que dos outros operadores. Entre os gêneros, as correções apresentam o maior número dessas marcas, indicando o maior grau de viés. Em relação aos operadores de opinião e valoração, também temos um maior número de ocorrências para o gênero resenha, o esperado para um gênero avaliativo e opinativo. O número de ocorrências para as correções também é alto e se aproxima muito mais das resenhas, o que não é esperado para um texto que deveria ter um caráter menos subjetivo. Finalmente, como nos *boxplots*, a pressuposição e a modalidade apresentam um número menos relevante para o diagnóstico de viés nos três gêneros. Mas, mesmo assim, a correção também se aproxima mais das resenhas do que dos resumos e, ainda, apresenta valores que chegam a superar aqueles encontrados para as resenhas.

Com a Tabela 3, que mostra os percentuais de ocorrência dos tipos de marcadores no conjunto de cada *corpus*, tomado como uma unidade, evidenciamos que as características de subjetividade encontradas para as correções pertencem ao gênero, e não provêm de peculiaridades individuais de corretores. As análises por documento e a análise geral do *corpus* levam à mesma conclusão e nossos resultados indicam uma uniformidade de comportamento desses textos, enquanto parte de um mesmo gênero textual.

Concluindo a análise, os resultados apresentados nos levam a confirmar nossa hipótese: as correções de redação apresentam um maior grau de viés linguístico do que o esperado para esse gênero textual. Em termos de distribuição de operadores de viés, as correções se aproximam mais de resenhas (o polo mais subjetivo) do que de resumos acadêmicos (o polo menos subjetivo), até mesmo superando as resenhas para alguns tipos de operadores.

5. Uma breve reflexão sobre a correção das redações do ENEM

Com a análise apresentada aqui, vimos que os textos de correção de redação se assemelham a resenhas. Isso evidenciou o caráter mais subjetivo, não esperado, para o gênero. Tal resultado coloca em dúvida também a isenção do processo de correção das redações do ENEM. Como as correções de redação do nosso *corpus*, retiradas do banco do UOL, se assemelham muito às correções das redações do ENEM, tanto nos critérios adotados, quanto no perfil dos corretores, o resultado encontrado nesta pesquisa é uma evidência do possível grau de viés linguístico existente também na correção de redações feitas nesse exame. Entendemos que, para confirmarmos tal hipótese, o ideal seria termos acesso aos textos das próprias correções de redações do ENEM. Contudo, essa avaliação é feita por uma grade de cinco competências, já estabelecidas pelos elaboradores da prova, em que o corretor deve apontar uma nota específica de 0 a 200 para cada habilidade. Assim, o processo individual de como cada corretor chegou a uma determinada pontuação fica oculto.

Acrescentamos ao nosso resultado, ainda, três outras evidências do alto nível de subjetividade das correções de redação do ENEM, já apontadas de forma detalhada em Mendes (2013). A primeira evidência diz respeito ao processo, propriamente, de correção das redações. As redações do ENEM são submetidas a dois avaliadores distintos, que não se comunicam. Eles corrigem pelo computador em torno de 50 a 70 redações por dia. O avaliador é remunerado por produtividade; quanto mais redações corrigir, maior será a remuneração. Um novo pacote de textos só é enviado ao corretor quando ele já tiver concluído o anterior. Nota-se nesse processo uma clara sobrecarga dos avaliadores. Nesse ponto, vale a afirmação feita em Mendes (2013) (p. 450) a respeito da interferência do cansaço na atribuição de notas na correção: "Segundo Wolcott & Legg (1998), em situação de cansaço, a atenção do avaliador começa a fugir, levando-o a sobrepontuar ou penalizar um texto. A tendência, nesse caso, é sacrificar a precisão pela rapidez."

Outro aspecto a se notar é a disparidade das avaliações. Cada avaliador atribui uma nota entre 0 e 200 pontos para cada uma das cinco competências, e a soma desses pontos compõe a nota final de cada avaliador, que pode chegar a 1.000 pontos. As notas atribuídas por esses avaliadores são somadas e é feita a média aritmética. Quando há uma divergência (mais de 100 pontos de diferença entre os dois corretores ou diferença superior a 80 pontos em qualquer uma das competências), essa redação é submetida a uma terceira correção independente. A nota final será a média aritmética das duas notas totais que mais se aproximarem. E, ainda havendo divergência, essa redação é submetida a uma quarta correção, feita por uma banca presencial composta por três professores, que atribuirá a nota final do candidato[32]. Segundo dados do Inep[33], o número de divergência entre notas atribuídas por diferentes corretores é bastante alto. Temos o seguinte quadro de correções feitas nos anos de 2014, 2015 e 2016, descrito pela Tabela 4.

A partir desses dados, vemos que uma terceira correção é feita em metade das redações e, ainda, em torno de 200.000 redações são submetidas a uma quarta correção. Esses números corroboram a nossa hipótese de que também no ENEM a correção é altamente subjetiva, já que não há uniformidade nas notas atribuídas em muitos casos.

[32]Ver como se dá esse processo em: http://download.inep.gov.br/educacao_basica/enem/guia_participante/2017/manual_de_redacao_do_enem_2017.pdf.

[33]Fonte: Pedido de informação SIC - Inep nº 23480.004970/2017-81

	2014	2015	2016
Redações corrigidas	4.338.259	5.598.545	5.637.484
Submetidas a 3ªcorreção	2.823.128	2.160.115	2.162.044
Submetidas a 4ªcorreção	295.161	163.620	164.200

Tabela 4: Redações do ENEM corrigidas em 2014, 2015 e 2016

Uma última evidência vem do próprio texto dos critérios de correção[34]. Esse texto, a princípio, deveria apresentar uma sequência tipológica injuntiva (prescritiva), apresentando instruções ao alocutário em relação a um procedimento. A linguagem utilizada nessa sequência é menos subjetiva e não se espera a presença de operadores do tipo que utilizamos em nossa análise. Entretanto, essa expectativa não se cumpre. Há várias marcas de subjetividade no texto dos critérios de correção de redação do ENEM. Mendes (2013) mostra que a definição desses critérios envolve valores subjetivos, expressos por modalizadores, vagueza conceitual e termos obscuros. Aplicando a nossa lista de operadores de viés linguístico no texto, que é composto por 640 palavras, observamos 51 ocorrências de operadores argumentativos, o que aponta aproximadamente para 8%, e 12 ocorrências de operadores de opinião e valoração, o que representa aproximadamente 2% do total de palavras do texto. Não aparecem operadores de pressuposição e de modalidade. Se os próprios critérios que norteiam a correção apresentam alto grau de viés linguístico, não podemos esperar que as correções sejam isentas de viés.

Os nossos resultados, acrescidos das três evidências aqui apresentadas, nos levam a concluir, em concordância com Mendes (2013), que a redação não é um bom instrumento avaliativo quando é imperiosa a isenção.

6. Considerações finais

Este artigo teve como principal objetivo demonstrar a subjetividade na correção de redações através da detecção automática de linguagem enviesada. Fizemos uma análise quantitativa de operadores de viés linguístico em textos de correção de redação, coletados no banco de redações do UOL. Assumimos, a partir da Semântica Argumentativa, que há certas palavras e expressões que explicitam pontos de vista e a distribuição numérica desses itens no texto pode indicar graus de subjetividade, o viés

[34]Disponíveis em: http://download.inep.gov.br/educacao_basica/enem/guia_participante/2017/manual_de_redacao_do_enem_2017.pdf.

linguístico. A partir disso, elaboramos uma lista desses operadores, baseados na classificação encontrada em Koch (2011, 2015). Essa lista foi, assim, aplicada ao *corpus* de correções de redação. Foram também analisados esses operadores de viés linguístico em resumos acadêmicos e em resenhas de produtos publicadas em *sites* de vendas na *internet*. Esses resultados numéricos foram comparados aos resultados encontrados para as correções, utilizando como ferramenta estatística o gráfico *boxplot*. Para essa comparação, partimos do pressuposto de que resumos acadêmicos são menos subjetivos e resenhas são mais subjetivas. Nosso propósito era, assim, verificar de qual polo de subjetividade as correções mais se aproximavam, maior subjetividade ou menor subjetividade.

Considerando a importância da redação em processos seletivos, espera-se que um texto como a correção de uma redação não apresente um alto nível de viés linguístico, aproximando-se do grau de subjetividade encontrado em resumos acadêmicos. Porém, nossa análise apontou para a direção oposta a essa expectativa. A análise quantitativa revela que os textos de correção de redação apresentam um número alto de operadores de viés, quando comparados aos demais gêneros, se aproximando do polo mais subjetivo e, até mesmo, superando os textos desse polo, as resenhas, para alguns operadores. Assim, apontamos como conclusão em nossa pesquisa que há um alto grau de subjetividade na correção de redações. Consideramos que tal resultado não é o desejado para um gênero textual que deveria apresentar um baixo grau de subjetividade, devendo, a princípio, se voltar para os critérios de correção, e não para a opinião do corretor. Ainda, estando nossos resultados de acordo com o que já apontou Mendes (2013), concluímos que os resultados obtidos através da análise do nosso *corpus* indicam a subjetividade do processo avaliativo como um todo.

Agradecimentos

Agradecemos ao Prof. Frederico R.B. Cruz (DEST/UFMG) as úteis orientações em relação à análise estatística. Os professores Márcia Cançado, Adriano Veloso e Heliana Mello agradecem o apoio financeiro do CNPq.

Referências

Aluísio, Sandra, Jorge Pelizzoni, Ana Raquel Marchi, Lucélia de Oliveira, Regiana Manenti & Vanessa Marquiafável. 2003. An account of the challenge of tagging a reference corpus for brazilian portuguese. Em *6th international conference on Computational processing of the Portuguese language (PROPOR)*, 110–117.

Aluísio, Sandra Maria & Gladis Maria de Barcellos Almeida. 2006. O que é e como se constrói um corpus? lições aprendidas na compilação de vários corpora para pesquisa linguística. *Calidoscópio* 4(3). 156–178.

Anscombre, Jean-Claude & Oswald Ducrot. 1976. L'argumentation dans la langue. *Langages* 42. 5–27.

Conrad, Alexander, Janyce Wiebe & Rebecca Hwa. 2012. Recognizing arguing subjectivity and argument tags. Em *Workshop on extra-propositional aspects of meaning in computational linguistics*, 80–88.

Damasceno, Adriana Carla, Rafael Andrade, Israel Almeida, Mayrlla Lopes & Silvana Nóbrega. 2016. Descrevendo o uso dos computadores nas escolas públicas da paraíba. *Revista Brasileira de Informática na Educação* 24(3). 10.5753/rbie.2016.24.3.47.

Ducrot, Oswald. 1987. *O dizer e o dito*. Pontes Editores.

Fiorin, José Luiz. 2007. A linguagem em uso. Em Ana Scher, Antonio V. Pietroforte, Diana P. Barros, Esmeralda V. Negrão, Evani Viotti, Luiz Tatit, Margarida Petter, Paulo Chagas, Raquel Santos & Ronald Beline (eds.), *Introdução à Linguística*, 165–185. São Paulo: Editora Contexto.

Hartmann, Nathan, Lucas Avanço, Pedro Paulo Balage Filho, Magali Sanches Duran, Maria Das Graças Volpe Nunes, Thiago Alexandre Salgueiro Pardo & Sandra M Aluísio. 2014. A large corpus of product reviews in portuguese: Tackling out-of-vocabulary word. Em *International Conference on Language Resources and Evaluation (LREC)*, 3865–3871.

Hooper, Joan B. 1974. *On assertive predicates*. Indiana University Linguistics Club.

Japiassú, André Miguel. 2013. Como elaborar e submeter resumos de trabalhos científicos para congressos. *Revista Brasileira de Terapia Intensiva* 25(2). 77–80. 10.5935/0103-507X.20130016.

Karttunen, Lauri. 1971. Implicative verbs. *Language* 47(2). 340–358. 10.2307/412084.

Kiparsky, Paul & Carol Kiparsky. 1968. *Fact*. Linguistics Club, Indiana University.

Koch, Ingedore Grunfeld Villaça. 2011. *Argumentação e linguagem.* Cortez Editora 13th edn.

Koch, Ingedore Grunfeld Villaça. 2015. *A interação pela linguagem.* Editora Contexto 11th edn.

Lane, David M, David Scott, Mikki Hebl, Rudy Guerra, Dan Osherson & Heidi Zimmer. 2017. *Introduction to statistics, online edition.* Rice University, University of Houston Clear Lake, and Tufts University.

Levin, Beth & Malka Rappaport Hovav. 1995. *Unaccusativity: At the syntax-lexical semantics interface.* MIT press.

Lin, Chenghua, Yulan He & Richard Everson. 2011. Sentence subjectivity detection with weakly-supervised learning. Em *5th International Joint Conference on Natural Language Processing*, 1153–1161.

Lin, Wei-Hao, Theresa Wilson, Janyce Wiebe & Alexander Hauptmann. 2006. Which side are you on?: identifying perspectives at the document and sentence levels. Em *10th conference on computational natural language learning*, 109–116.

Liu, Bing, Minqing Hu & Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. Em *14th international conference on World Wide Web*, 342–351. 10.1145/1060745.1060797.

Marcuschi, Luiz Antônio. 2002. Gêneros textuais: definição e funcionalidade. Em Angela Paiva Dionisio; Anna Rachel Machado; Maria Auxiliadora Bezerra (ed.), *Gêneros textuais e ensino*, 19–36. Lucerna.

Mendes, Eliana Amarante de Mendonça. 2013. A avaliação da produção textual nos vestibulares e outros concursos: a questão da subjetividade. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)* 18(2). 435–458. 10.1590/S1414-40772013000200011.

Motta-Roth, Désirée & Graciela Rabuske Hendges. 2010. *Produção textual na universidade.* Parábola Editorial.

Pires de Oliveira, Roberta. 2001. *Semântica formal: uma breve introdução.* Mercado de Letras.

Recasens, Marta, Cristian Danescu-Niculescu-Mizil & Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. Em *51st Annual Meeting of the Association for Computational Linguistics*, 1650–1659.

Rodrigues, Ricardo, Hugo Gonçalo Oliveira & Paulo Gomes. 2014. LemPORT: a high-accuracy cross-platform lemmatizer for portuguese. Em *3rd Symposium on Languages, Applications and Technologies*, 267–274. 10.4230/OASIcs.SLATE.2014.267.

Sardinha, Tony Berber. 2002. Tamanho de corpus. *The Especialist* 23(2). 103–122.

Somasundaran, Swapna & Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. Em *Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 116–124.

Turney, Peter D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Em *40th Annual Meeting of the Association for Computational Linguistics*, 417–424. 10.3115/1073083.1073153.

Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell & Melanie Martin. 2004. Learning subjective language. *Computational linguistics* 30(3). 277–308. 10.1162/0891201041850885.

Wolcott, Willa & Sue M Legg. 1998. *An overview of writing assessment: Theory, research, and practice.* ERIC.

Yano, Tae, Philip Resnik & Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. Em *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 152–158.

Yule, George. 1996. *Pragmatics.* Oxford University Press.

Proposta recebida em Janeiro 2020 e aceite para publicação em Maio 2020.

Periodização automática: Estudos linguístico-estatísticos de literatura lusófona

Automatic literary school assignment: Linguistic-statistical studies of lusophone literature

Diana Santos
Linguateca & Universidade de Oslo
d.s.m.santos@ilos.uio.no

Emanoel Pires
Universidade Estadual do Maranhão
emanoel.uema@gmail.com

Cláudia Freitas
Linguateca & PUC-Rio
maclaudia.freitas@gmail.com

Rebeca Schumacher Fuão
Linguateca
rebischu@gmail.com

João Marques Lopes
Linguateca
marqueslopes1928@hotmail.com

Resumo

Neste artigo usamos um conjunto de características sintático-semânticas da língua portuguesa para classificar em períodos literários dois conjuntos de obras. Em que medida tais características são capazes de refletir distinções relevantes no âmbito dos estudos literários é uma das questões que pretendemos investigar.

O primeiro grupo de obras corresponde à replicação do trabalho relatado em 2009 por Barufaldi et al., que usaram métodos de compressão de dados sobre uma série de obras brasileiras classificadas em quatro períodos literários: barroco, arcadismo, romantismo e realismo, desde o Padre António Vieira até Raul Pompéia, contabilizando 15 autores diferentes e totalizando 37 obras.

O segundo grupo inclui muito mais obras (192), tanto portuguesas como brasileiras, mas apenas integra romances ou novelas publicadas no período de 1840 a 1919. As escolas literárias escolhidas foram o realismo, o romantismo, o simbolismo, o naturalismo, o decadentismo e o modernismo, mas, ao contrário da classificação anterior, permitimos que uma mesma obra pertença a várias escolas.

Usamos técnicas de classificação em R para a primeira tarefa, e análise de correspondências para a segunda. Também aplicamos técnicas de modelos de tópicos à segunda coleção para ver se é possível obter tópicos representativos de escolas literárias diferentes.

Palavras chave

leitura distante, linguística com corpos, literatura lusófona, escola literária, português, literatura brasileira, literatura portuguesa

Abstract

In this paper we use a set of syntactic and semantic features of Portuguese to automatically classify literary works in literary periods and/or schools, and address the issue of their appropriateness, for two different literary collections.

The first task attempts to replicate the work by Barufaldi and colleagues, who applied compression methods on 37 Brazilian works by 15 different authors and classified the works in 4 different literary schools.

The second collection, of 192 novels published in Portugal and Brazil in the period 1840 to 1919, features many works who cannot be singly accomodated in one literary school only, and which have been (not mutually exclusively) classified as romantic, realist, naturalist, symbolist, decadent and modernist.

We use classification techniques in R, such as discriminant analysis and support vector models for the first task, and correspondence analysis for the second collection. We also apply topic modeling to (distinct subsets of) the second collection in order to investigate whether this technique can provide us with recurrent topics for different literary schools.

Keywords

distant reading, corpus linguistics, literary school, Portuguese, Brazilian literature, Portuguese literature, lusophone literature

1. Introdução

O objetivo do presente artigo é avaliar se a informação linguística que temos vindo a associar, em estudos linguísticos da língua portuguesa, a várias obras literárias pode também ser usada para responder a questões do foro dos estudos literários.

DOI: 10.21814/lm.12.1.314

Para tal, compilámos uma lista de características sintáticas e semânticas a que temos acesso na Literateca (Santos, 2019b; Santos & Simões, 2019) e usámo-las em dois problemas, que passamos a descrever sucintamente:

- atribuir 37 obras de 15 autores brasileiros diferentes a quatro períodos literários, replicando um trabalho anterior feito com métodos de compressão
- organizar 192 romances ou novelas de autores portugueses e brasileiros publicadas entre 1840 e 1919, tentando apreciar semelhanças entre autores e escolas literárias, seguindo a proposta de Santos et al. (2018b) inspirada por Moretti (2000)

A nossa posição é a de explorar a informação que temos e não a de demonstrar que estes métodos resolvem os problemas literários. Na medida em que for possível encontrar formas de identificar semelhanças e grupos que concordem com a autoridade literária, ou que levem a perguntas pertinentes, estes métodos de leitura distante poderão contribuir para os estudos literários. Se, pelo contrário, indicarem outros agrupamentos, tal não deve ser considerado como uma teoria alternativa, mas apenas como demonstrando que as características escolhidas não eram relevantes para o problema em questão.

Em relação às obras usadas como material para a nossa pesquisa, a lista exata encontra-se em apêndice. Além de material compilado pela própria Linguateca, usamos textos gentilmente cedidos pelos seguintes projetos irmãos *Corpus Histórico do Português Tycho Brahe* (Galves & Faria, 2010) e *Colonia - Corpus of Historical Portuguese* (Zampieri & Becker, 2013). Exceto no caso dos textos provenientes do corpo PANTERA (Santos, 2019c), trabalhamos com textos completos.

2. Características usadas

Na Literateca, além do acesso ao texto em formato eletrônico, temos a vantagem de ter (e disponibilizar para consulta) todo o material anotado gramatical e semanticamente. A anotação morfossintática é feita pelo analisador PALAVRAS (Bick, 2000).

Para ambas as tarefas calculámos um conjunto extenso de características de cada texto (128) que nos pareceram de interesse para uma possível descrição do estilo, que passamos a elencar.

A partir da anotação morfossintática do PALAVRAS, levamos especificamente em conta em nossa análise a presença de adjetivos (índices de qualificação) e nomes próprios (instâncias de classes genéricas como pessoas/personagens e locais, entre outros); a presença e a distribuição de construções como voz passiva ou a forma progressiva; orações relativas e completivas (índices de complexidade estrutural); a presença de coordenações e conjunções coordenativas, bem como de vírgulas e outros sinais de pontuação (índices de ritmo). As indicações de tempo, modo e aspeto verbal também foram consideradas potenciais elementos caracterizadores (especificamente o modo conjuntivo, o pretérito perfeito composto, o pretérito imperfeito, o perfeito, o mais que perfeito e os aspetualizadores), bem como a presença de verbos na primeira pessoa, e de palavras no género morfológico feminino. Pontos de exclamação, de interrogação e travessões, e elementos de negação também foram utilizados como índices potencialmente caracterizadores de autores, obras e/ou estilos.

De um ponto de vista estilístico, o número de palavras por frase é um elemento que costuma ser utilizado na diferenciação de autores e obras —veja-se, por exemplo, a matéria de Almeida & Mariani (2019), que utiliza este traço para produzir gráficos relativos a obras da literatura brasileira —, e por isso usamos o número de frases por obra.

Além da anotação morfossintática, o material da Literateca conta também com a anotação de diversos campos semânticos[1]. Neste trabalho, levamos em conta os campos dos verbos de fala, da saúde/doença, cores, corpo humano, família, roupas e emoções.

Com relação ao campo do dizer Freitas et al. (2016), partimos de um léxico de verbos específicos e de regras que indicam se os verbos estão sendo utilizados para introduzir a fala de alguém (relato direto ou indireto) ou se apenas se trata da menção a algum evento comunicativo ("...e não falou mais no assunto"). Como características, usamos três: verbos de relato direto, verbos de relato indireto, e verbos de fala somente.

O campo semântico das emoções conta com variadas palavras, de diferentes classes gramaticais, distribuídas em 24 grandes grupos, como amor, coragem, desejo, desespero, felicidade,

[1]Por campo semântico denotamos uma área de conhecimento refletida na língua, como a cor, ou a família. Infelizmente este é um uso completamente distinto daquele que é definido pelo Dicionário Terminológico, conforme nos chamou a atenção Àlvaro Iriarte Sanroman.

fúria, admiração, inveja e medo, entre outros (veja-se o Emocionário[2] para sua documentação cabal). O número de palavras em cada um destes grupos é uma característica, assim como o total de palavras de emoção.

No campo da saúde (Santos, 2019a), usamos os seguintes indicadores: o número de palavras desse campo, a presença do lema dor, e a presença de palavras dos subcampos progressão (da doença), causa (da doença), palavras genéricas sobre saúde ou doença, remédios, acessórios (relacionados com saúde/doença), medicina e saúde psicológica.

No campo da cor (Silva & Santos, 2012), usamos o total de palavras de cor, o total de palavras de cor com sentido de cor, assim como o número de palavras pertencente a cada grupo de cor (Laranja, Vermelho, Dourado, etc.) e o total de palavras de cor com sentido não cor, ou seja, presentes em expressões fixas como *buraco negro*, *luz verde*, *lista negra*.

Para o corpo humano (Freitas et al., 2015), usamos o total de palavras de corpo, o total de palavras de corpo com sentido literal, e o número de palavras pertencentes a cada parte do corpo (Cabeça, Sexual, Pernas, etc).

Para a roupa (Santos et al., 2011), usamos o total de palavras de roupa, e o número de palavras pertencentes a cada grupo de roupa (Calças, RoupaDormir, Calçado, etc.).

Finalmente, para o campo da família usamos o número de palavras relacionadas com a família, assim como o campo mais específico de parentesco. Veja-se o trabalho de Higuchi et al. (2019) para uma motivação deste campo.

A lista completa, por ordem alfabética, encontra-se no sítio da Linguateca[3]. Convém esclarecer que a marcação destes campos semânticos é feita automaticamente através de regras, e tem alguma margem de erro.

Uma primeira discussão destes indicadores no contexto da literatura está presente em Santos (2019b). Mas desde esse trabalho, que data de 2017 embora apenas tenha sido publicado em 2019, adicionámos várias características e várias obras.

3. Primeira tarefa: repetir o trabalho de Barufaldi et al.

Barufaldi et al. (2010) usaram métodos de compressão de dados sobre uma série de obras brasileiras classificadas em quatro períodos literários: barroco, arcadismo, romantismo e realismo, desde o Padre António Vieira até Raul Pompéia, contabilizando 15 autores diferentes e totalizando 37 obras.

Tentando obter exatamente o mesmo material utilizado por Barufaldi et al. (2010), optamos, sempre que possível, por obras que estão disponibilizadas em sítios como o da Biblioteca Digital de Literaturas de Língua Portuguesa[4] e do Domínio Público[5]. Ainda assim, como é restrita a informação mencionada pelos autores no que diz respeito às edições utilizadas, é possível que haja mudanças, ainda que mínimas em alguns casos, nos textos das obras escolhidas. Ademais, a publicação inicial em folhetim e posterior edição em formato livro também pode ser outro fator que acarrete variações nas edições escolhidas. Sobre alterações nas edições das obras de Machado de Assis, por exemplo, conferir Campos (2018).

De todo modo, parece-nos que a possível distinção mais radical entre os arquivos das obras utilizadas possa estar em *14 de Julho na Roça*, de Raul Pompéia. Como se trata de uma coletânea de contos nomeada de *Contos* na Biblioteca de Literaturas de Língua Portuguesa e de *14 de Julho na Roça* no sítio do Domínio Público, restou a dúvida se os autores utilizaram apenas o conto inicial, intitulado de *14 de Julho na Roça*, ou se a obra completa. Como estamos tratando de textos escritos pelo mesmo autor e em um espaço de tempo muito próximo, optamos por utilizar a obra completa, admitindo que todos os contos têm o mesmo estilo de época.

Quanto ao processo computacional, aplicámos duas técnicas (Baayen, 2008) usando o ambiente R (R Core Team, 2018), empregando as características descritas acima para o mesmo fim:

- análise de discriminantes com base em componentes principais (ver Figuras 1 2 3)
- máquinas de vetores de apoio (support vector machines) (ver Tabela 1)

[2] https://www.linguateca.pt/Gramateca/Emocionario.html

[3] https://www.linguateca.pt/Gramateca/Literateca/lista_caracteristicas.txt

[4] https://www.literaturabrasileira.ufsc.br/

[5] http://www.dominiopublico.gov.br/pesquisa/PesquisaObraForm.jsp

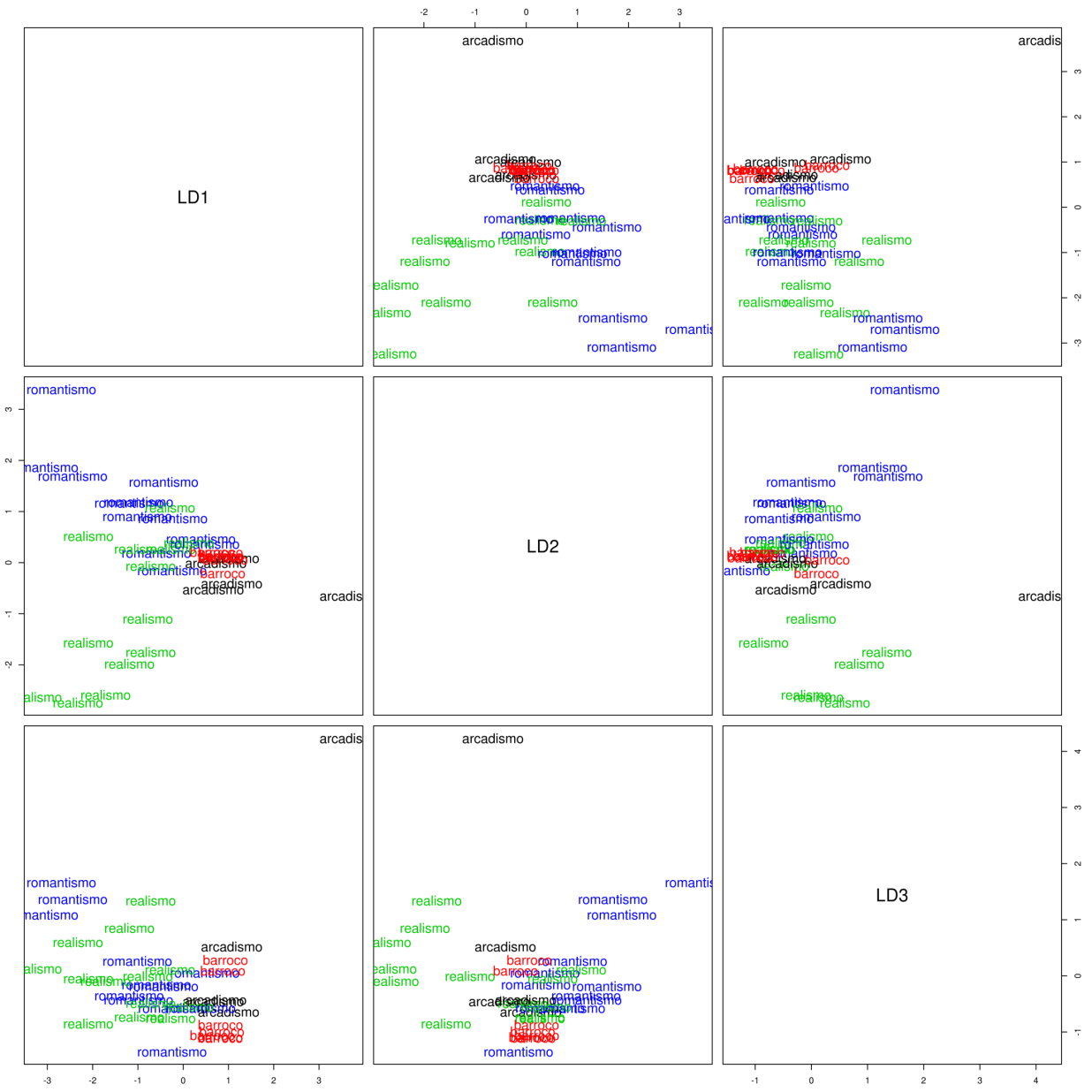


Figura 1: Análise de discriminantes, global

	arcad.	barr.	real.	romant.
arcadismo	3	2	0	0
barroco	0	7	0	0
realismo	0	0	10	3
romantismo	0	1	0	11

Tabela 1: Resultado da classificação com máquinas de vetores de apoio

Ainda não nos está suficientemente claro os motivos por detrás da nossa classificação equivocada das obras *O Uraguai*, de Basílio da Gama, e *Coletânea de obras*, de Alvarenga Peixoto. Em Barufaldi et al. (2010), a *Coletânea de obras líricas*, de Gregório de Matos, também foi classificada erroneamente. As causas podem estar relacionadas ao conjunto de características marcadas nas obras e utilizadas como parâmetro nas análises como, também, à falta de marcações que estejam mais relacionadas com elementos que sinalizem de maneira mais efetiva o estilo na poesia, como os processos de acomodação silábica. Mittmann et al. (2016) utilizam uma ferramenta de escansão automática, o Aoidos[6], que, em estudos futuros, poderá ajudar nos casos de confusão.

Sobre *Ubirajara*, de José de Alencar, ter sido classificado como pertencente ao Barroco, mesmo sendo em prosa, a hipótese inicial com a qual tra-

[6] https://aoidos.ufsc.br/

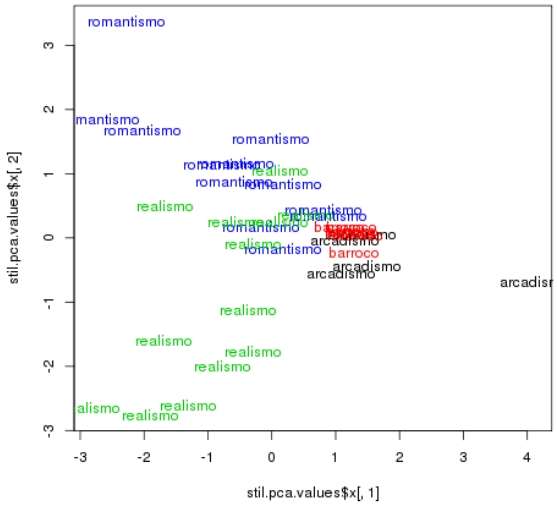


Figura 2: Análise de discriminantes, mostrando o primeiro e o segundo

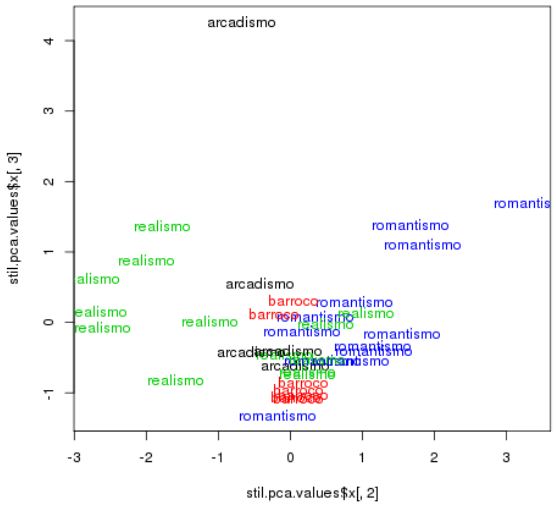


Figura 3: Análise de discriminantes, mostrando o segundo e o terceiro

balhamos diz respeito ao tamanho médio das frases do romance de Alencar, que, por serem muito curtas, se diferenciam em muito do estilo empregado nos demais romances do período romântico que foram incluídos na análise.

Seja como for, os resultados indicam que estas características parecem ser apropriadas para distinguir entre os quatro períodos ou escolas literárias selecionados pelos nossos antecessores, embora com uma leve tendência a privilegiar o barroco. Mas pensamos que a tarefa pode ter sido demasiado simples, dado que os diferentes períodos também implicam diferenças tão abissais como poesia vs. prosa e correspondem a épocas razoavelmente distintas.

4. Segunda tarefa: romances e novelas portugueses e brasileiros do período 1840–1919

O segundo conjunto de obras pode ser mais complexo de classificar, visto que se refere a um período de apenas 80 anos, e a duas formas muito semelhantes: o romance e a novela, ambas correspondentes à *novel* inglesa, daí a razão da amálgama[7]. Contém autores que, devido à sua longevidade e/ou génio, produziram obras que são tradicionalmente consideradas de escolas diferentes, e possui elementos que muitos estudiosos consideram inqualificáveis, por únicos.

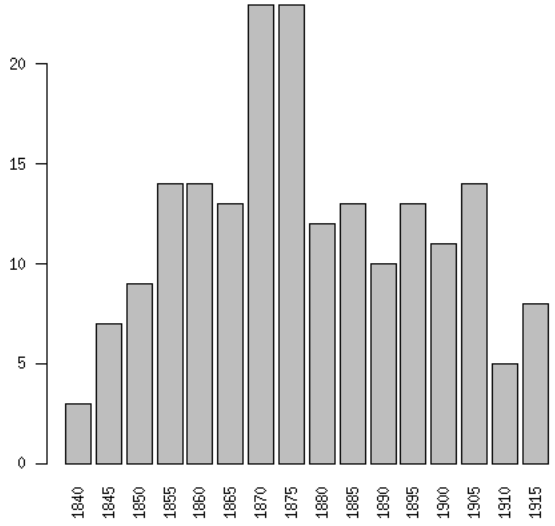


Figura 4: Segundo conjunto de obras por data de publicação

Não cabendo fazer aqui a discussão pormenorizada do assunto, limitamo-nos a sinalizar dois exemplos de complexidade e singularidade, remetendo o leitor para a lista de referências consultadas[8]. O primeiro é *O Ateneu*, do brasileiro Raul Pompeia, que tem uma longa recepção crítica em que uns o consideram naturalista ou realista, outros o tacham de impressionista, outros indicam o predomínio do simbolismo e há ainda quem assinale o expressionismo ou o cruzamento de duas ou mais escolas no seu interior, conforme se pode constatar nos estudos de Araújo (2011) e de Quintale Neto (2007). O segundo é *Os Maias*, do

[7] Convirá a este propósito mencionar que o interesse especial por este período vem da ação COST *Distant Reading for European Literary History*, https://www.distant-reading.net/, em cujo âmbito estamos a produzir duas coleções de obras em português, uma coleção portuguesa Herrmann et al. (2020) e uma coleção lusófona, também com obras brasileiras do mesmo período.

[8] Acessível de https://www.linguateca.pt/OBRAS/siglas_Literateca.pdf

português Eça de Queirós, em cuja classificação Carlos Reis (Reis, 2012), reconhecidamente um dos maiores especialistas neste autor, oscila entre realismo, naturalismo e pós-naturalismo. Na Literateca, e não querendo escolher entre as várias escolas, *O Ateneu* está marcado como `impressionismo-naturalismo-realismo-simbolismo`, e *Os Maias* como `realismo-naturalismo-pós-naturalismo`[9].

Portanto, isso fez com que a tarefa de atribuir um rótulo a cada obra não fosse algo linear, indo de casos cuja taxinomia foi efetivamente unívoca (por exemplo, ninguém duvida de que *Eurico, o presbítero*, de Alexandre Herculano, é um marco do romantismo português) a casos mais complicados como os que acabamos de referir.

Devido à facilidade em adicionar a informação de que nos encontrávamos em presença de um romance histórico, essa classificação foi também adicionada aquando da classificação (e cobriu apenas obras classificadas como romantismo).

Vale ainda mencionar que algumas obras não possuem uma classificação, e muitas vezes nem são citadas, em nenhuma obra de referência sobre a história da literatura de Portugal ou Brasil. Isso ocorre porque não são obras canônicas. Nesses casos, foi preciso desenvolver um método que nos permitisse classificar essas obras de forma coerente com o conjunto que possuímos.

Primeiramente, realizamos um mapeamento de características que nos permitissem identificar a escola à qual uma obra pertence, tais como: tempo, narrador, espaço, personagens, temas, finais felizes ou infelizes, etc. Após esse mapeamento, partimos para a leitura de trechos ou da obra completa para então discutir e determinar em que escola poderíamos enquadrá-la.

O resultado deste trabalho encontra-se sumariado na Tabela 2.

Dito isso, a coleção que usámos é a seguinte: todos os romances e novelas em português em formato eletrónico a que tínhamos acesso à data de 25 de outubro de 2019 —estão em curso diversas iniciativas para aumentar este acervo, mas estes são os que pudemos coligir nessa altura e que foram publicados no período já mencionado (1840-1919).

Isso corresponde a 192 obras (listadas no anexo), das quais 123 portuguesas e 69 brasileiras. O autor com mais obras é Camilo com 37, seguido de Machado de Assis com 13, Aluísio de Azevedo com 11, Eça de Queirós com 11, José de Alencar com 9, Júlio Dinis com 8, e Alexandre Herculano com 6. Os restantes autores têm entre uma a quatro obras nesta coleção. (Cinco obras são traduzidas, duas por Machado de Assis, uma por Eça de Queirós, outra por Camilo Castelo Branco, e uma adaptada por Pedro Supico de Morais. Estas obras são úteis para servirem de teste.)

Escola literária	**Quantos**
decadentismo	1
expressionismo	1
expressionismo-simbolismo	1
ficcaocient	1
histórico	3
historico-romantismo	2
impress.-natural.-realismo-simbol.	1
indianismo-romantismo	2
modernismo	2
naturalismo	12
naturalismo-realismo	2
naturalismo-realismo-romantismo	1
naturalismo-regionalismo	1
picaresco-realismo	1
realismo	20
realismo-naturalismo	8
realismo-pos-naturalismo	1
realismo-posnaturalismo	5
realismo-regionalismo-romantismo	3
realismo-romantismo	1
regionalismo	2
romantismo	78
romantismo-decadentismo	1
romantismo-histórico	16
romantismo-indianismo	1
romantismo-indianismo-histórico	1
romantismo-realismo	15
romantismo-realismo-naturalismo	5
romantismo-regionalismo	3
simbolismo	1

Tabela 2: Escolas literárias atribuídas

Na Figura 4 pode ver-se a distribuição da data de publicação destas obras por períodos de cinco anos.

Uma análise de correspondências mostra-nos como as características que selecionámos colocam os diferentes autores, e as diferentes escolas, no plano definido por estas.

Na Figura 5 vê-se cada obra com uma cor diferente por autor, além de apresentar as características mais discriminadoras neste conjunto de obras a vermelho, nomeadamente o número de completivas, de interrogativas, a menção de humildade e a referência a medicina ou progressão de uma doença.

[9] De notar que a escolha das classes foi feita com base nos especialistas que sobre os autores e obras se pronunciaram, o que resultou em que por exemplo apenas um autor, Eça de Queirós, tem (em algumas obras) a classificação `pós-naturalismo`, e uma autora, Ana de Castro Osório, `pós-romantismo`.

Vemos que há autores, como Aluísio de Azevedo e Júlio Dinis, que são bem fiéis a si próprios, definindo portanto áreas bem claras no plano, enquanto que outros, como Machado de Assis ou Eça de Queirós, têm obras espalhadas por vários quadrantes.

Olhando apenas para as obras destes quatro autores, é interessante reparar que, enquanto as obras mais extremas de Machado são as traduções (uma em cada extremo da Figura 6), no caso de Eça de Queirós (Figura 7) a tradução não se demarca de forma alguma das suas outras obras. Não cabe aqui a análise do perfil destes autores como tradutores[10], mas notamos que esse será um tema interessante para os estudos de tradução, assim como é um argumento para analisar a obra de escritores incluindo também as traduções que escreveram.

Nas Figuras 8 e 9, relativas a Aluísio de Azevedo e a Júlio Dinis, as obras encontram-se mais perto no plano.

Repare-se que o canto superior esquerdo, na Figura 5 é quase só Camilo, que também tem obras no canto inferior esquerdo. Lembramos que o nosso conjunto não é balanceado entre autores, nem entre escolas literárias, como o demonstra a Figura 10.

Convém explicar que para esta figura "traduzimos" a Tabela 2 para uma classificação muito mais simples, que mostramos na Tabela 3. Basicamente, usámos as seguintes regras para "traduzir" a pertença para apenas uma escola: qualquer menção a romantismo, indianismo ou histórico ganhava a classificação de romantismo. Depois, qualquer menção a realismo ou regionalismo ficavam realismo puro. Em seguida, se naturalismo era mencionado, ficava naturalismo, enquanto simbolismo e decadentismo eram amalgamados em simbolismo. Obviamente outras formas de reclassifiar seriam possíveis, por exemplo usando a primeira classificação em vez de ordenar a decisão da forma que fizemos.

Seja como for, temos claramente regiões em que as escolas se localizam, mesmo que não haja regiões sem sobreposição. A Figura 11 mostra a situação sem simplificações, ou seja, cada obra aparece com o conjunto de escolas que lhe foram atribuídas (veja-se de novo a Tabela 2).

Nova escola literária	**Quantos**
expressionismo	1
ficcaocient	1
modernismo	2
naturalismo	12
realismo	42
romantismo	131
simbolismo	3

Tabela 3: Escola literária simplificada

num.	tópico
13	porta rua janela parede luz
16	homem arma inimigo soldado guerra
52	padre santo igreja missa religião
59	cavalo caminho estrada homem cavaleiro
90	sala baile festa salão sociedade
93	livro poeta romance verso obra

Tabela 4: Tópicos obtidos sobre as obras classificadas como românticas

5. Análise de tópicos

Desviando-nos um pouco da análise linguística, resolvemos também usar o método estatístico mais comum dos estudos literários: a análise de tópicos (*topic modelling*), ver Jockers (2013), que apenas usa as palavras e calcula os tópicos sem acesso a outras classificações (exceto que as palavras usadas são exclusivamente as das classes gramaticais substantivo, adjetivo e advérbio, obtidas pela análise do PALAVRAS).

Usando blocos de 500 dessas palavras consecutivas, e pedindo 100 tópicos, o sistema mallet (McCallum, 2002) produziu uma lista[11] de XXXX entradas para a coleção completa.

Apresentamos alguns tópicos que nos parecem esclarecedores, pela consistência e facilidade de interpretação, na Tabela 4.

Alguns destes apresentamos também em nuvem de palavras, nas Figuras 12 e 13.

Outros há que não são facilmente interpretáveis, enquanto outros ainda são mais específicos, como 4 (romano lusitano povo exército cidade) ou 80 (gaúcho sertanejo vez animal fazenda).

[10] Embora seja necessário mencionar que uma das traduções de Machado de Assis foi continuada por outro autor.

[11] https://www.linguateca.pt/Gramateca/Literateca/artigoEscolas/topicos_todas_as_obras/topicosNA_todos_tam500.txt

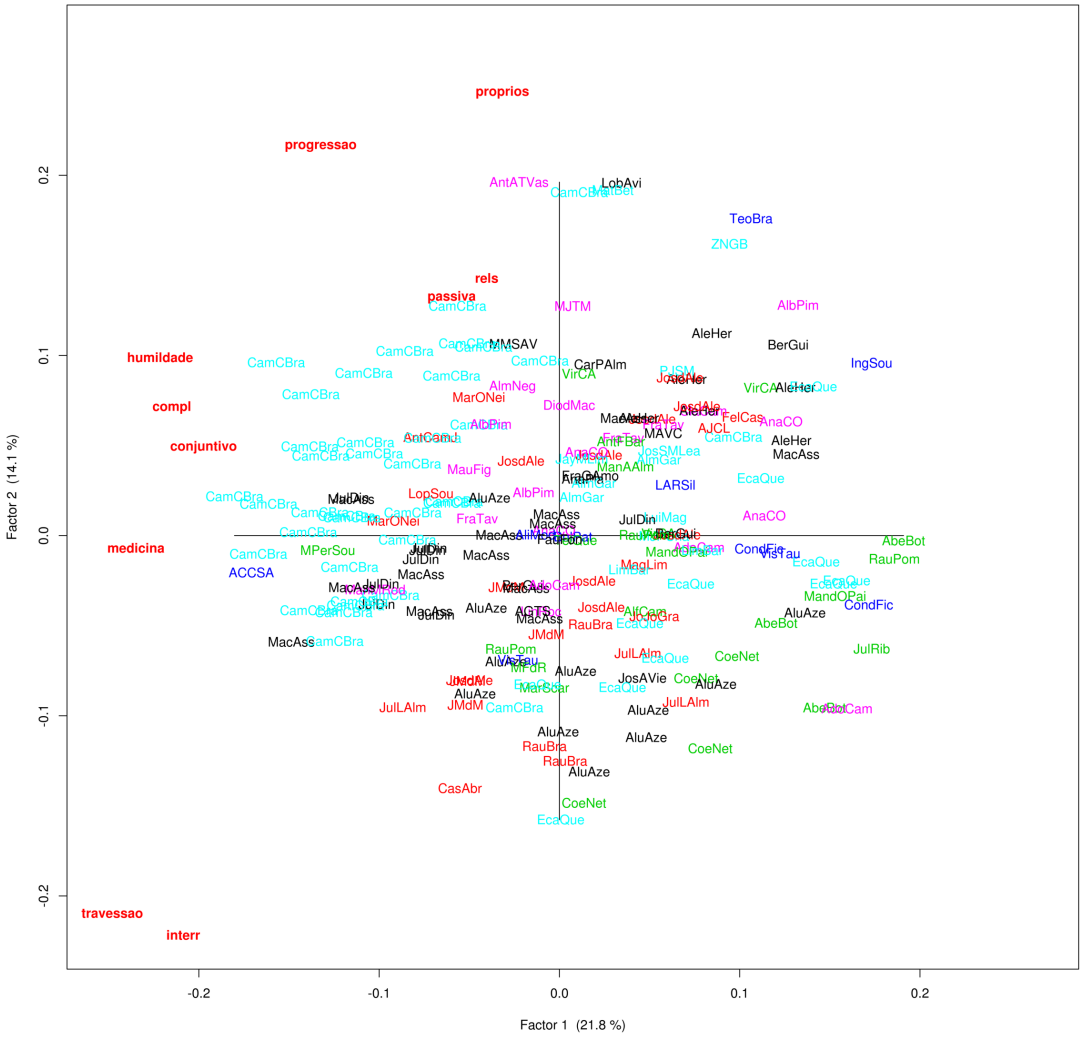


Figura 5: Análise de correspondências

Selecionando apenas as obras marcadas (não necessariamente exclusivamente) com a classificação de românticas (124 obras), por um lado, e realistas e/ou naturalistas, por outro (68 obras), obtemos duas novas listas (romantismo[12] e realismo[13]). Na Tabela 5 apresentamos tópicos realistas/naturalistas, e na Tabela 6 românticos.

num.	tópico
25	médico dia doente febre saúde
44	dinheiro conto negócio real carta
55	mulher amor paixão vida beijo
60	estudante colégio professor diretor livro
73	casa porta noite sala quarto

Tabela 5: Tópicos obtidos sobre as obras classificadas como realistas ou naturalistas

num.	tópico
1	mar vento praia onda tempestade
14	guerreiro chefe virgem cabana taba
19	cavaleiro homem rei batalha namorado
68	flor sombra sol doce jardim
89	navio homem marinheiro bordo capitão
90	leito doente quarto corpo morte

Tabela 6: Tópicos obtidos sobre as obras classificadas como românticas

[12] https://www.linguateca.pt/Gramateca/Literateca/artigoEscolas/topicos_romantismo/topicosNA_romantismo_tam500.txt

[13] https://www.linguateca.pt/Gramateca/Literateca/artigoEscolas/topicos_realismo/topicosNA_realismo_tam500.txt

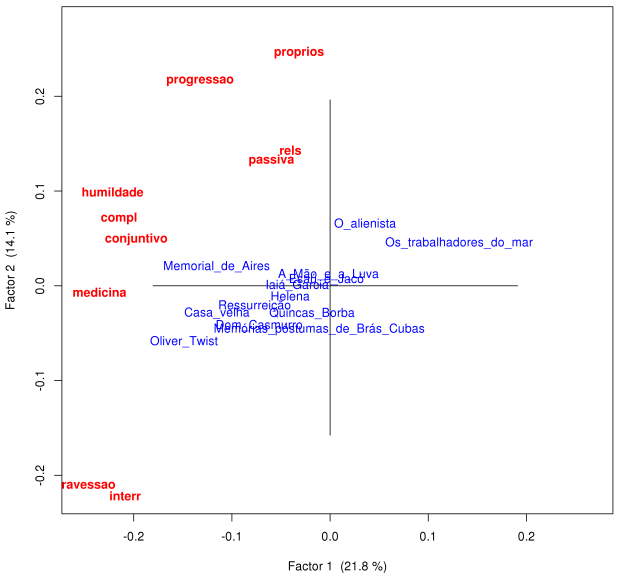


Figura 6: Análise de correspondências mostrando apenas as obras de Machado de Assis

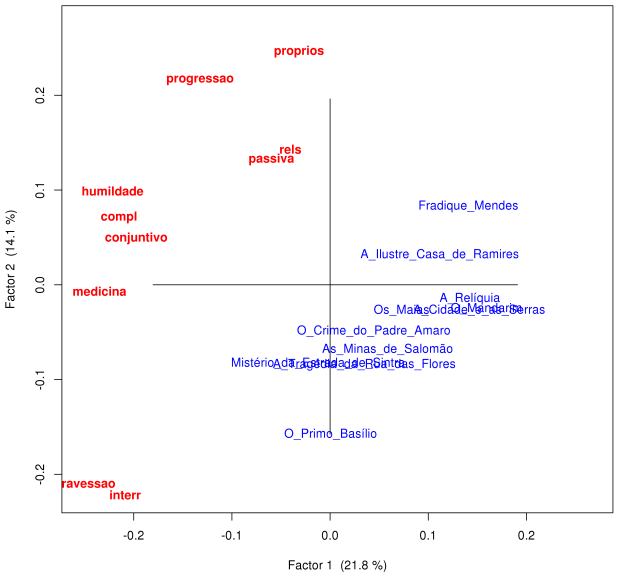


Figura 7: Análise de correspondências mostrando apenas as obras de Eça de Queirós

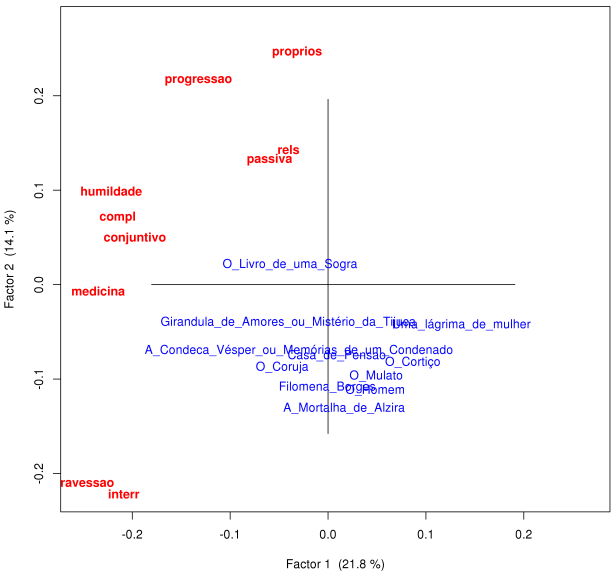


Figura 8: Análise de correspondências mostrando apenas as obras de Aluísio de Azevedo

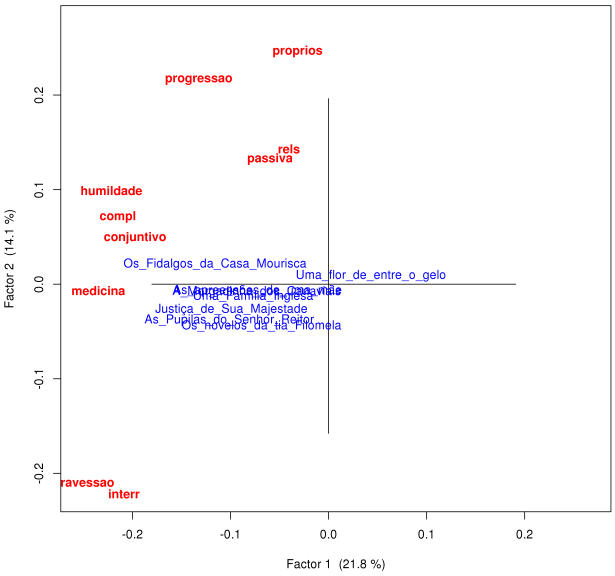


Figura 9: Análise de correspondências mostrando apenas as obras de Júlio Dinis

6. Comentários e Trabalho futuro

Este trabalho marca o início de um programa de colaboração entre estudos literários e linguística computacional para mutuamente enriquecer ambas as disciplinas. Dessa forma, em vez de muitos resultados, temos muitas interrogações, e vias de desenvolvimento futuro.

Se por um lado pensamos ter mostrado que as características linguísticas (nas secções 3 e 4) e o conteúdo lexical (na Secção 5) são úteis para a exploração e estudo da literatura, estamos plenamente conscientes de que muito mais trabalho tem de ser feito em relação à identificação e correta anotação de muitas destas características, e pretendemos efetuar muito brevemente estudos de algumas em particular, como as emoções e o corpo.

Por outro lado, foi evidente que a noção de escola literária não era uma questão simples, e que muitas outras características e interrogações seriam possíveis, desde o género do autor, data de escrita, local de escrita (por exemplo Brasil ou Portugal) ao tipo de obra (romance histórico, romance de costumes, etc.).

Além disso, o facto de termos um número considerável de obras que caíram no esquecimento, e que provavelmente nunca foram colocadas numa escola literária pelos teóricos da literatura, pode também levantar a questão de que as escolas do

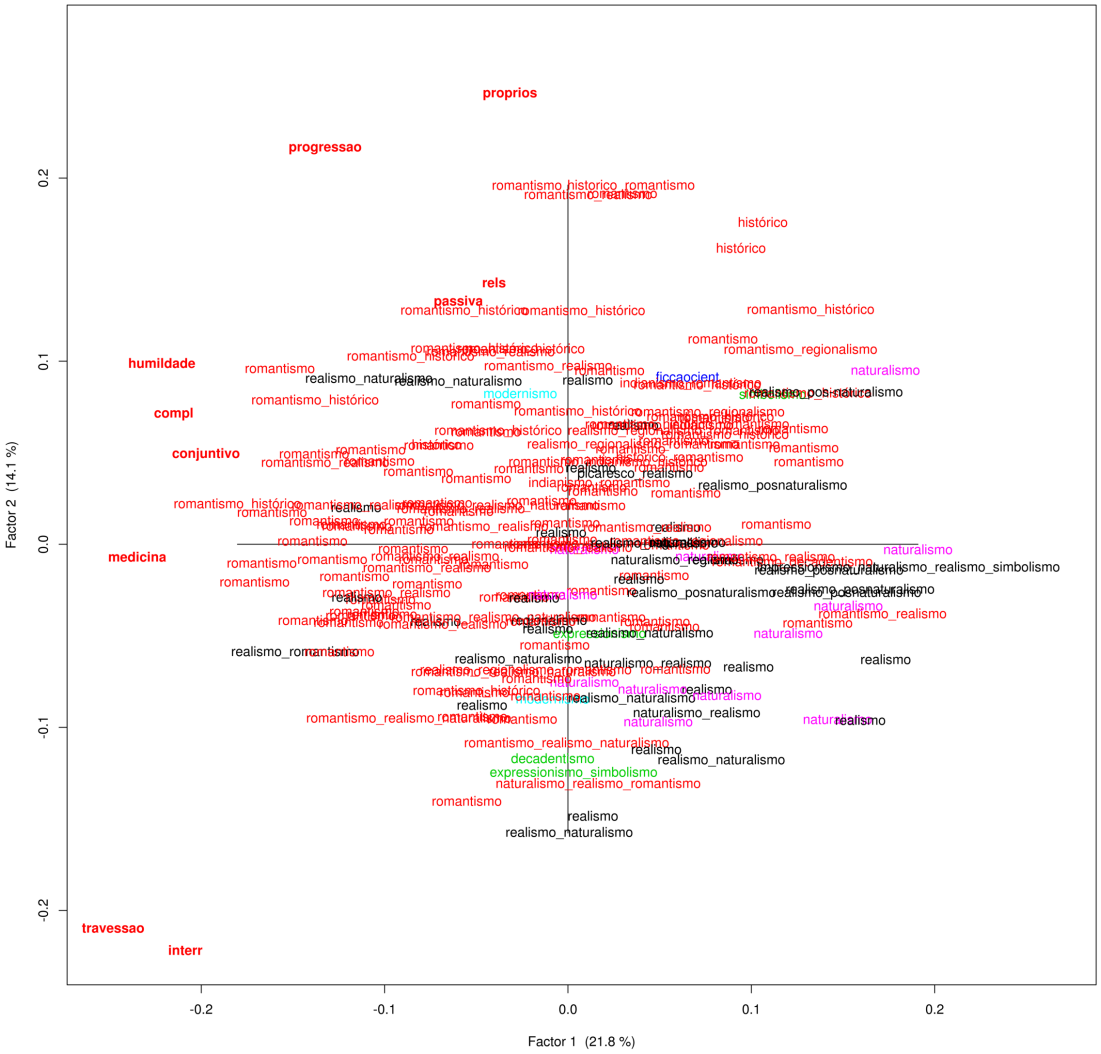


Figura 10: Análise de correspondências mostrando cada escola com uma cor diferente

cânone literário poderão não ser as únicas, e que outras fações ou movimentos podem ser sugeridos através do estudo de mais obras — uma das vantagens da leitura distante.

O peso de determinados autores assim como a possibilidade de discordância em relação à escola ou escolas atribuídas mostra como muito do que por exemplo encontrámos no romantismo poderia mudar se seguíssemos a opinião de João Gaspar Simões (Simões, 1967) e considerássemos Camilo um autor por si só (e não romântico!)[14]

Será pois relevante, de um ponto de vista literário, experimentar fazer outras escolhas (mesmo entre as obras que temos) para criar novas (sub)coleções, para não deixar que um autor, uma época ou um tipo de escrita pese demais no grupo.

Muitas outras técnicas estatísticas, assim como tarefas, podiam ter sido realizadas com este material, que se encontra público[15] para que outros o possam experimentar e avançar no estudo da literatura lusófona[16]. Mencionamos como uma das mais evidentes a tentativa de classificação da segunda coleção através do método das árvores de decisão.

[14] Convém referir que muitos autores, incluindo, aliás, Camilo, tiveram obras classificadas em escolas diferentes... Cada obra recebeu uma classificação distinta.

[15] Em https://www.linguateca.pt/Gramateca/Literateca/artigoEscolas/ podem encontrar-se quer os dados como os comandos utilizados em R.

[16] Vários dos corpos de onde estas obras foram retiradas estão acessíveis da Linguateca, veja-se por exemplo o OBras (Santos et al., 2018a), ou de outros projetos como o Tycho Brahe (Galves & Faria, 2010).

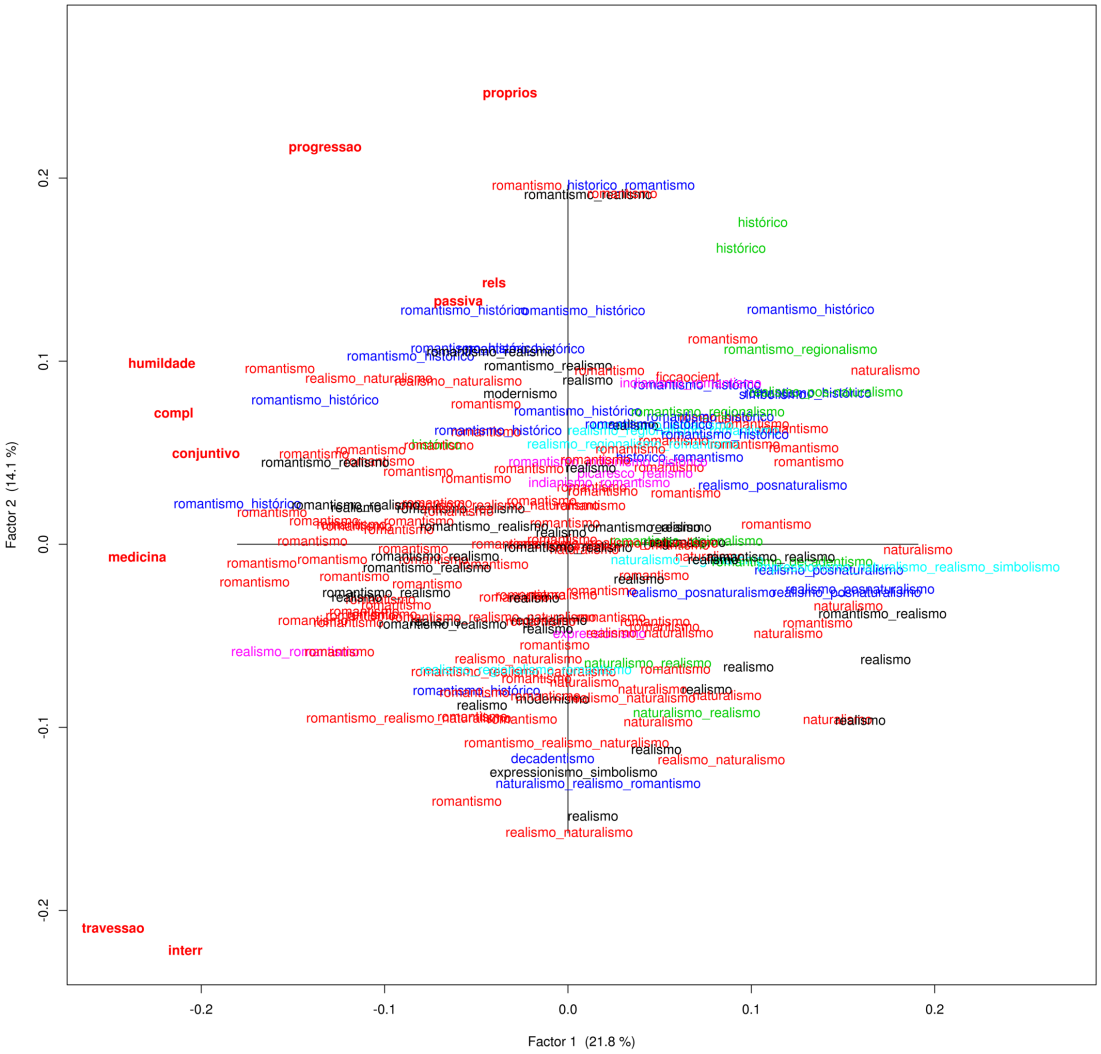


Figura 11: Análise de correspondências mostrando todas as classificações de escola literária

Figura 12: Palavras que constituem o tópico 13

Figura 13: Palavras que constituem o tópico 90

Concluindo, este trabalho é apenas um primeiro passo no uso de métodos estatísticos e linguísticos para reconsiderar a literatura lusófona. Ao tornarmos públicos os documentos e as análises, assim como os primeiros resultados, esperamos que alguns nos sigam no destrinçar de características, influências e semelhanças entre muitos autores que escreveram em português, assim como desejamos que este tipo de explorações nos dê mais conhecimento sobre o estilo e a "alma" linguística da língua portuguesa.

Agradecimentos

Estamos muito gratos a Alckmar Luiz dos Santos pelas sugestões e críticas feitas em Oslo, à audiência da APL em Braga pelas perguntas pertinentes, e aos recensores Miguel Anxo Solla Portela e Álvaro Iriarte Sanroman pela revisão aturada de uma primeira versão deste trabalho.

Agradecemos à FCCN pelo alojamento da Linguateca nos seus servidores, ao grupo de Research Computing da Universidade de Oslo pelo apoio informático, e à UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway pelos recursos computacionais cedidos para o processamento dos corpos e a obtenção de resultados.

Este artigo não existiria se não tivesse sido desencadeado pela ação COST "Distant reading for European literary history", financiada pelo EU Framework Programme da União Europeia, Horizon 2020.

Finalmente, Emanoel Pires agradece à FAPEMA pelo apoio ao projeto "Estudos estatístico-literários em literatura lusófona: junção de esforços entre a Linguateca e o Portal Maranhão".

Referências

Almeida, Rodolfo & Daniel Mariani. 2019. O ritmo e o estilo de diferentes obras literárias brasileiras. https://www.nexojornal.com.br/grafico/2017/01/30/O-ritmo-e-o-estilo-de-diferentes-obras-liter%C3%A1rias-brasileiras.

Araújo, Francisco Magno da Silva de. 2011. *O Ateneu e a nostalgia da forma*. Natal: Centro de Ciências Humanas, Letras e Artes da Universidade Federal do Rio Grande do Norte. Tese de Mestrado.

Baayen, Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Barufaldi, Bruno, Eduardo F. Santana, José Rogério B. B. Filho, Jan Kees van der Poel, Milton Marques Júnior & Leonardo Vidal Batista. 2010. Classificação Automática de Textos por Período Literário Utilizando Compressão de Dados Através do PPM-C. *Linguamática* 2(1). 35–44.

Bick, Eckhard. 2000. *The parsing system "Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework*. Aarhus, Denmark: Aarhus University. Tese de Doutoramento.

Campos, Alex Sander Luiz. 2018. Edições de Machado de Assis: por quê, para quê? *Machadiana Eletrônica* 1(1). 131–150.

Freitas, Cláudia, Bianca Freitas & Diana Santos. 2016. QUEMDISSE?: Reported speech in Portuguese. Em *10th International Conference on Language Resources and Evaluation (LREC)*, 4410–4416.

Freitas, Cláudia, Diana Santos, Cristina Mota, Bruno Carriço & Heidi Jansen. 2015. O léxico do corpo e anotação de sentidos em grandes corpora: o projeto esqueleto. *Revista de Estudos da Linguagem* 23(3). 641–680. doi 10.17851/2237-2083.23.3.641-680.

Galves, Charlotte & Pablo Faria. 2010. Tycho Brahe parsed corpus of historical Portuguese. http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html.

Herrmann, J. Berenike, Carolin Odebrecht, Diana Santos & Pieter Francois. 2020. Towards modeling the european novel. Introducing ELTeC for multilingual and pluricultural distant reading. Em *Digital Humanities Conference*, Abstract Book.

Higuchi, Suemi, Diana Santos, Cláudia Freitas & Alexandre Rademaker. 2019. Distant reading Brazilian history. Em *4th Conference of The Association Digital Humanities in the Nordic Countries*, 190–200.

Jockers, Matthew L. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

McCallum, Andrew Kachites. 2002. MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu.

Mittmann, Adiel, Aldo von Wangenheim & Alckmar Luiz dos Santos. 2016. A system for the automatic scansion of poetry written in Portuguese. Em *17th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, 611–628.

Moretti, Franco. 2000. Conjectures on world literature. *New Left review* 1. 54–68.

Quintale Neto, Flávio. 2007. *Idéias estéticas e filosóficas nos romances O Ateneu, de Raul Pompéia e Die Vervirrungen des Zöglings Törless, de Robert Musil*: Universidade de São Paulo. Tese de Doutoramento.

R Core Team. 2018. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available online at https://www.R-project.org/.

Reis, Carlos. 2012. Trajeto literário. https://queirosiana.wordpress.com/trajeto-literario/.

Santos, Diana. 2019a. Distant reading health: A pilot study on health and disease in lusophone literature. Illness and disability in literary and cultural texts: an international seminar. https://www.linguateca.pt/Diana/download/DRHealth.pdf.

Santos, Diana. 2019b. Literature studies in literateca: between digital humanities and corpus linguistics. Em Martin Doerr, Øyvind Eide & Oddrun Grønvik ans Bjørghild Kjelsvik (eds.), *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore*, 89–109. Novus forlag.

Santos, Diana. 2019c. PANTERA: a parallel corpus to study translation between Portuguese and Norwegian. *Bergen Language and Linguistics Studies* 10(1). doi 10.15845/bells.v10i1.1372.

Santos, Diana, Cláudia Freitas & Eckhard Bick. 2018a. OBras: a fully annotated and partially human-revised corpus of brazilian literary works in the public domain. Em *Latin American and Iberian Languages Open Corpora Forum (OpenCor)*, s.p.

Santos, Diana, Cláudia Freitas & João Marques Lopes. 2018b. Ler e estudar a literatura lusófona como parte da literatura mundial: recursos para leitura distante em português. Em *I Congresso Internacional em Humanidades Digitais no Rio de Janeiro (HdRio)*, 375–383.

Santos, Diana, Augusto Soares da Silva & Cristina Mota. 2011. Guarda-fatos: notas sobre a anotação do campo semântico do vestuário em português. Relatório técnico. Linguateca. http://www.linguateca.pt/acesso/GuardaFatos.pdf.

Santos, Diana & Alberto Simões. 2019. Towards a computational environment for studying literature in portuguese. Apresentação na conferência Digital Humanities.

Silva, Rosário & Diana Santos. 2012. Arco-íris: notas sobre a anotação do campo semântico da cor em português. Relatório técnico. Linguateca. http://www.linguateca.pt/acesso/ArcoIris.pdf.

Simões, João Gaspar. 1967. *História do Romance Português*. Estúdios Cor.

Zampieri, Marcos & Martin Becker. 2013. Colonia: Corpus of historical portuguese. *ZSM Studien* 5. 77–84.

Lista de textos

1843 *O Bobo*, de Alexandre Herculano

1844 *A Moreninha*, de Joaquim Manuel de Macedo

1844 *Eurico o Presbítero*, de Alexandre Herculano

1845 *O Arco de Santana*, de J. B. da Silva L. de Almeida Garrett

1845 *O moço louro*, de Joaquim Manuel de Macedo

1846 *O Galego*, de Alexandre Herculano

1846 *Viagens na Minha Terra*, de J. B. da Silva L. de Almeida Garrett

1848 *O Monge de Cister I*, de Alexandre Herculano

1848 *O Monge de Cister II*, de Alexandre Herculano

1848 *Os Dois Amores*, de Joaquim Manuel de Macedo

1851 *Anátema*, de Camilo Castelo Branco

1851 *O Pároco de Aldeia*, de Alexandre Herculano

1852 *Memórias de um sargento de milícias*, de Manuel de Almeida

1853 *Coisas que só eu sei*, de Camilo Castelo Branco

1854 *A Filha do Arcedíago*, de Camilo Castelo Branco

1854 *Helena*, de J. B. da Silva L. de Almeida Garrett

1854 *Mistérios de Lisboa I*, de Camilo Castelo Branco

1854 *Mistérios de Lisboa II*, de Camilo Castelo Branco

1854 *Mistérios de Lisboa III*, de Camilo Castelo Branco

1855 *Livro Negro de Padre Dinis I*, de Camilo Castelo Branco

1855 *Livro Negro de Padre Dinis II*, de Camilo Castelo Branco

1855 *O Cura de São Lourenço*, de M M S A e Vasconcelos

1856 *Carolina*, de Casimiro de Abreu

1856 *Onde Esta a Felicidade*, de Camilo Castelo Branco

1856 *Um Homem de Brios*, de Camilo Castelo Branco

1857 *A viuvinha*, de José de Alencar

1857 *O Guarani*, de José de Alencar

1857 *O soldado de Aljubarrota*, de Matilde Isabel de Santana e Vasconcelos Moniz Bettencourt

1857 *Os tripeiros: Crónica do século XIV*, de António José Coelho Lousada

1858 *A Vingança*, de Camilo Castelo Branco

1858 *O Que Fazem Mulheres*, de Camilo Castelo Branco

1859 *Maria ou a menina roubada*, de Antônio Gonçalves Teixeira e Souza

1859 *Úrsula*, de Maria Firmina dos Reis

1861 *A chave do enigma*, de António Feliciano de Castilho

1861 *Romance dum Homem Rico*, de Camilo Castelo Branco

1862 *Amor de Perdição*, de Camilo Castelo Branco

1862 *Coisas Espantosas*, de Camilo Castelo Branco

1862 *Coração Cabeça e Estômago*, de Camilo Castelo Branco

1862 *Infaustas Aventuras de Mestre Marçal Estouro: Vítima duma paixão*, de José da Silva Mendes Leal

1863 *Adelina*, de Ana Plácido

1863 *Aventuras de Basílio Fernandes Enxertado*, de Camilo Castelo Branco

1863 *O Bem e o Mal*, de Camilo Castelo Branco

1864 *A Filha do Doutor Negro*, de Camilo Castelo Branco

1864 *A pálida estrela*, de Bulhão Pato

1864 *Amor de Salvação*, de Camilo Castelo Branco

1864 *No Bom Jesus do Monte*, de Camilo Castelo Branco

1864 *Vinte Horas de Liteira*, de Camilo Castelo Branco

1865 *Iracema, lenda do Ceará*, de José de Alencar

1866 *A Queda dum Anjo*, de Camilo Castelo Branco

1866 *A conquista de Lisboa*, de Carlos Pinto de Almeida

1866 *Os trabalhadores do mar*, de Machado de Assis

1867 *A Doida do Candal*, de Camilo Castelo Branco

1867 *As Pupilas do Senhor Reitor*, de Júlio Dinis

1867 *Henriqueta*, de Maria Peregrina de Sousa

1868 *A Morgadinha dos Canaviais*, de Júlio Dinis

1868 *O Retrato de Ricardina*, de Camilo Castelo Branco

1868 *O ermitão do Muquém*, de Bernardo Guimaraes

1868 *Uma Família Inglesa*, de Júlio Dinis

1869 *A luneta mágica*, de Joaquim Manuel de Macedo

1869 *Os Brilhantes do Brasileiro*, de Camilo Castelo Branco

1870 *A Rosa do Adro*, de Manuel Maria Rodrigues

1870 *A ermida de Castromino*, de Antonio Augusto Teixeira de Vasconcellos

1870 *A pata da gazela*, de José de Alencar

1870 *As apreensões de uma mãe*, de Júlio Dinis

1870 *Justiça de Sua Majestade*, de Júlio Dinis

1870 *Mistério da Estrada de Sintra*, de José Maria Eça de Queirós

1870 *O gaúcho*, de José de Alencar

1870 *Oliver Twist*, de Machado de Assis

1870 *Os novelos da tia Filomela*, de Júlio Dinis

1870 *Uma flor de entre o gelo*, de Júlio Dinis

1871 *Herança de lágrimas*, de Lopo de Sousa

1871 *Os Fidalgos da Casa Mourisca*, de Júlio Dinis

1872 *A Infanta Capelista*, de Camilo Castelo Branco

1872 *Inocência*, de Visconde de Taunay

1872 *O Carrasco de Vitor Hugo*, de Camilo Castelo Branco

1872 *O seminarista*, de Bernardo Guimaraes

1872 *Ressurreição*, de Machado de Assis

1873 *A alma de Lázaro*, de José de Alencar

1873 *A filha do Cabinda*, de Alfredo Campos

1873 *O Annel Mysterioso: Scenas da Guerra Peninsular*, de Alberto Pimentel

1873 *Um conto portuguez: episódio da guerra civil: a Maria da Fonte*, de Miguel J T Mascarenhas

1874 *A Mão e a Luva*, de Machado de Assis

1874 *Ubijarara*, de José de Alencar

1875 *A Escrava Isaura*, de Bernardo Guimaraes

1875 *A Filha do Regicida*, de Camilo Castelo Branco

1875 *A Freira no Subterraneo*, de Camilo Castelo Branco

1875 *A senhora viscondessa*, de S de Magalhães Lima

1875 *Novelas do Minho I*, de Camilo Castelo Branco

1875 *O Crime do Padre Amaro*, de José Maria Eça de Queirós

1875 *O sertanejo*, de José de Alencar

1875 *Os selvagens*, de Francisco Gomes de Amorim

1875 *Senhora*, de José de Alencar

1876 *A Caveira da Mártir*, de Camilo Castelo Branco

1876 *Helena*, de Machado de Assis

1876 *O Cabeleira*, de Franklin Távora

1876 *O Christão novo*, de Diogo de Macedo

1877 *Alice*, de Maria Amália Vaz de Carvalho

1877 *Novelas do Minho II*, de Camilo Castelo Branco

1878 *A Tragédia da Rua das Flores*, de José Maria Eça de Queirós

1878 *Iaiá Garcia*, de Machado de Assis

1878 *O Matuto*, de Franklin Távora

1878 *O Primo Basílio*, de José Maria Eça de Queirós

1879 *Eusébio Macário*, de Camilo Castelo Branco

1879 *O Romance da Rainha Mercedes*, de Alberto Pimentel

1879 *O Sacrifício*, de Franklin Távora

1879 *Uma lágrima de mulher*, de Aluisio Azevedo

1880 *A Corja*, de Camilo Castelo Branco

1880 *O Mandarim*, de José Maria Eça de Queirós

1881 *Memórias póstumas de Brás Cubas*, de Machado de Assis

1881 *O Mulato*, de Aluisio Azevedo

1882 *A Brasileira de Prazins*, de Camilo Castelo Branco

1882 *A Condeca Vésper ou Memórias de um Condenado*, de Aluisio Azevedo

1882 *As jóias da Coroa*, de Raul Pompéia

1882 *Girandula de Amores ou Mistério da Tijuca*, de Aluisio Azevedo

1882 *O alienista*, de Machado de Assis

1882 *Uma tragédia no Amazonas*, de Raul Pompéia

1884 *Casa de Pensao*, de Aluisio Azevedo

1884 *Filomena Borges*, de Aluisio Azevedo

1885 *Casa velha*, de Machado de Assis

1886 *O Brasileiro Soares*, de Luís Magalhães

1886 *Quincas Borba*, de Machado de Assis

1886 *Vulcões de Lama*, de Camilo Castelo Branco

1887 *A Relíquia*, de José Maria Eça de Queirós

1887 *O Homem*, de Aluisio Azevedo

1888 *A Carne*, de Júlio Ribeiro

1888 *Mais Uma*, de Conde de Ficalho

1888 *O Ateneu*, de Raul Pompéia

1888 *Os Maias*, de José Maria Eça de Queirós

1888 *Uma Eleição Perdida*, de Conde de Ficalho

1889 *No declínio*, de Visconde de Taunay

1889 *O Coruja*, de Aluisio Azevedo

1890 *O Cortiço*, de Aluisio Azevedo

1891 *As Minas de Salomão*, de José Maria Eça de Queirós

1891 *Dona Guidinha do Poço*, de Manuel de Oliveira Paiva

1891 *O Barão de Lavos*, de Abel Botelho

1891 *O missionário*, de Inglês de Sousa

1891 *O último cartuxo da Scala Caeli de Évora: Romance histórico (1808-1865)*, de António Francisco Barata

1892 *Noites de Cintra*, de Alberto Pimentel

1892 *O Dr. Luiz Sandoval*, de Alice Moderno

1893 *A Normalista*, de Adolfo Caminha

1894 *A Mortalha de Alzira*, de Aluisio Azevedo

1895 *A viúva Simões*, de Júlia Lopes de Almeida

1895 *Miragem*, de Coelho Neto

1895 *O Bom-Crioulo*, de Adolfo Caminha

1895 *O Livro de uma Sogra*, de Aluisio Azevedo

1895 *O mundo no ano 3000*, de Pedro José Supico de Morais

1896 *Tentacao*, de Adolfo Caminha

1897 *Pero da Covilhan: Episódio Romântico do Século XV*, de Zeferino Norberto Gonçalves Brandão

1898 *A descoberta e conquista da Índia pelos portugueses: romance historico*, de Artur Lobo d'Avila

1899 *A afilhada*, de Manuel de Oliveira Paiva

1899 *A conquista*, de Coelho Neto

1899 *Dom Casmurro*, de Machado de Assis

1899 *Elle*, de Claudia de Campos

1899 *Transviado*, de Jayme de Magalhães Lima

1900 *A Ilustre Casa de Ramires*, de José Maria Eça de Queirós

1900 *Fradique Mendes*, de José Maria Eça de Queirós

1900 *O exilado*, de Maurícia C de Figueiredo

1901 *A Cidade e as Serras*, de José Maria Eça de Queirós

1901 *A falência*, de Júlia Lopes de Almeida

1901 *Amanhã*, de Abel Botelho

1903 *A Farsa*, de Raúl Brandão

1904 *Esaú e Jacó*, de Machado de Assis

1904 *Os filhos do padre Anselmo*, de António da Costa Couto Sá de Albergaria

1904 *Turbilhão*, de Coelho Neto

1904 *Viriato*, de Teófilo Braga

1905 *A Ala dos Namorados*, de António Campos Junior

1905 *A Intrusa*, de Júlia Lopes de Almeida

1906 *A Divorciada*, de José Augusto Vieira

1906 *A Lenda da Meia-Noite*, de Manuel Joaquim Pinheiro Chagas

1906 *Os Bravos do Mindelo*, de Faustino da Fonseca

1906 *Os Pobres*, de Raúl Brandão

1908 *A Casa dos Fantasmas*, de Luís Augusto Rebelo da Silva

1908 *A feiticeira*, de Ana de Castro Osório

1908 *A vinha*, de Ana de Castro Osório

1908 *Diário de uma criança*, de Ana de Castro Osório

1908 *Memorial de Aires*, de Machado de Assis

1908 *Sacrificada*, de Ana de Castro Osório

1909 *O Salústio Nogueira*, de Teixeira de Queirós

1909 *Recordações do escrivão Isaías Caminha*, de Lima Barreto

1910 *Maria Dusá*, de Lindolfo Rocha

1911 *Triste Fim de Policarpo Quaresma*, de Lima Barreto

1913 *A Confissão de Lúcio*, de Mário de Sá-Carneiro

1914 *A Marquesa de Vale Negro*, de Maria O'Neill

1914 *Por bom caminho*, de Maria O'Neill

1915 *A capital federal*, de Coelho Neto

1915 *A engomadeira: novela vulgar lisboeta*, de José Sobral de Almada Negreiros

1916 *A morte vence*, de João José Grave

1916 *Decameron*, de Virgínia de Castro e Almeida

1916 *Innocente*, de Virgínia de Castro e Almeida

1916 *O Solar dos Pavões*, de Virgínia de Castro e Almeida

1919 *Amor crioulo*, de Abel Botelho

1919 *Húmus*, de Raúl Brandão

Propuesta recibida en enero 2020 y aceptada para publicación en junio 2020.

Una aplicación tecnológica que ayuda a la ciudadanía a escribir textos a la Administración pública

A technological application that helps citizens write texts to the Public Administration

Iria da Cunha
Universidad Nacional de Educación a Distancia (UNED)
iriad@flog.uned.es

Resumen

En este trabajo se presenta una aplicación tecnológica gratuita y en línea que ayuda a la ciudadanía a escribir textos dirigidos a la Administración pública. Concretamente, ayuda a redactar cinco géneros textuales: alegación, carta de presentación, queja, reclamación y solicitud. La aplicación tiene forma de editor de textos e incluye tres módulos para: I) estructurar y añadir contenidos en el texto, II) corregirlo ortográficamente y darle formato, y III) obtener sugerencias de mejora sobre aspectos léxicos y discursivos. Integra diferentes herramientas de Procesamiento del Lenguaje Natural (PLN), como un analizador morfosintáctico y un segmentador discursivo. Las evaluaciones *data-driven* y *user-driven* realizadas ofrecen resultados positivos.

Palabras clave

administración pública, ciudadanía, géneros textuales, asistente a la redacción en español

Abstract

This article presents a free and online technological application that helps citizens write texts addressed to the Public Administration. Specifically, it helps to draft five textual genres: allegation, cover letter, letter of complaint, claim and application. The technological application is a text editor that includes three modules: I) structure and contents of the text, II) spelling and format correction, and III) suggestions on vocabulary and discourse. It integrates different Natural Language Processing (NLP) tools, such as a morphosyntactic tagger and a discourse segmenter. The data-driven and user-driven evaluations performed show positive results.

Keywords

public administration, citizenship, textual genres, Spanish writing assistant

1. Introducción

Escribir textos de ámbitos especializados no es fácil, ya que han de poseer unas características muy concretas que deben tenerse en cuenta para que sean adecuados (Cabré, 1999; Gotti, 2008). Estas características dependerán del ámbito de los textos y del género textual que se quiera producir (Van Dijk, 1989; Bhatia, 1993; Parodi, 2010). Como indica Swales (1990), un género textual es una estructura convencionalizada a través de la cual se organizan los intercambios comunicativos de una determinada comunidad discursiva. Por lo general, cuando se habla de géneros textuales producidos en ámbitos especializados, suele partirse de la idea de que los autores serán especialistas de dichos ámbitos (Cabré, 1999), como por ejemplo un abogado en el ámbito legal. Sin embargo, hay determinados ámbitos especializados en los que son los ciudadanos quienes deben enfrentarse a la tarea de la redacción de textos. Un ejemplo muy evidente es el ámbito de la Administración pública, a quien los ciudadanos deben dirigirse por escrito en muchas ocasiones, para presentar por ejemplo una reclamación, una alegación o una queja. La ciudadanía no suele estar familiarizada con este tipo de géneros textuales administrativos, y su redacción resulta difícil y frustrante, lo cual se agrava al tratarse de documentos que tienen una repercusión directa en su vida y en su bienestar.

En el contexto de la lengua española, en los últimos años ha habido algunos esfuerzos por listar los géneros textuales prototípicos del ámbito administrativo y caracterizarlos lingüísticamente (Ayala et al., 2000; Castellón, 2001; Sánchez Alonso, 2014). A partir de estos trabajos, en da Cunha & Montané (2019) se seleccionaron los géneros textuales que tienen como emisores a los ciudadanos y, posteriormente, se analizaron empíricamente cuáles de estos géneros deben redactar con más frecuencia y cuáles les generan una mayor dificultad de redacción.

DOI: 10.21814/lm.12.1.316

En las conclusiones de ese estudio se destacaron cinco géneros: alegación, carta de presentación, queja, reclamación y solicitud. Asimismo, en ese trabajo, se analizaron los problemas que plantean los ciudadanos a la hora de escribir textos destinados a la Administración, y se concluyó que son principalmente cuatro: 1) estructurar el texto, 2) utilizar el vocabulario adecuado a la situación comunicativa, 3) decidir el contenido del texto y 4) utilizar el grado de formalidad adecuado. Recientemente, en da Cunha & Montané (2020), se realizó un análisis lingüístico (textual, léxico y discursivo) basado en corpus de los cinco géneros mencionados, para obtener una perspectiva global de las características de cada uno de ellos.

Tomando estos estudios como base, el objetivo del presente trabajo ha sido desarrollar una aplicación tecnológica que ayuda a la ciudadanía a escribir textos dirigidos a la Administración. Esta aplicación se ha denominado sistema arText e integra diferentes herramientas de Procesamiento del Lenguaje Natural (PLN).

En el Apartado 2 se presenta un estado de la cuestión sobre los recursos de ayuda a la redacción en el ámbito de la e-Administración. En el Apartado 3 se detalla la metodología del trabajo, haciendo hincapié en el marco teórico, en las funcionalidades de arText y en su implementación. En el Apartado 4 se exponen los resultados y la evaluación realizada, tanto *data-driven* como *user-driven*. En el Apartado 5 se incluyen las conclusiones y las líneas de trabajo futuro.

2. Estado de la cuestión

En España, la comunicación electrónica entre el ciudadano y la Administración es el procedimiento habitual en la actualidad, ya que se está tendiendo a la e-Administración. El 2 de octubre del 2015 se publicó en el Boletín Oficial del Estado (BOE) la Ley 39/2015, de 1 de octubre, del Procedimiento Administrativo Común de las Administraciones Públicas[1]. Esta Ley establece una regulación completa de las relaciones entre las Administraciones y los administrados, en diversos aspectos. Uno de ellos está relacionado con las Tecnologías de la Información y la Comunicación (TIC), cuyo desarrollo en los últimos años ha repercutido en la forma y contenido de las relaciones entre la Administración, y los ciudadanos y las empresas. En la introducción de la Ley 39/2015 se indica:

> *Si bien la Ley 30/1992, de 26 de noviembre, ya fue consciente del impacto de las nuevas tecnologías en las relaciones administrativas, fue la Ley 11/2007, de 22 de junio, de acceso electrónico de los ciudadanos a los Servicios Públicos, la que les dio carta de naturaleza legal, al establecer el derecho de los ciudadanos a relacionarse electrónicamente con las Administraciones Públicas, así como la obligación de éstas de dotarse de los medios y sistemas necesarios para que ese derecho pudiera ejercerse.* [...] *Porque una Administración sin papel basada en un funcionamiento íntegramente electrónico no sólo sirve mejor a los principios de eficacia y eficiencia, al ahorrar costes a ciudadanos y empresas, sino que también refuerza las garantías de los interesados.*

Sin embargo, hasta ahora son pocos los esfuerzos que se han hecho para desarrollar herramientas TIC que tengan como objetivo mejorar la comunicación escrita entre la Administración y la ciudadanía, a pesar de que ya en 2010 en el Informe de la Comisión de modernización del lenguaje jurídico del Ministerio de Justicia[2] se especificaba que:

> *Las tecnologías de la información pueden prestar soporte para el análisis de la claridad de los textos, facilitando una redacción más comprensible, sin suponer por ello una ralentización en los tiempos de escritura.*
> *En este sentido, son de especial utilidad los analizadores gramaticales y estadísticos de textos (longitud de las frases, número de oraciones, longitud media de las palabras, etc.), así como los programas de escritura semiautomática empleados por los profesionales de la interpretación y traducción. Estos últimos, por ejemplo, almacenan repositorios de frases o párrafos predeterminados que, una vez validados, se proponen a los redactores de los textos.*
> *En definitiva, se recomienda a las instituciones del sector promover el uso de este tipo de programas entre los profesionales del derecho e invertir en el desarrollo de nuevas aplicaciones.*

Es cierto que desde hace años y hasta la actualidad se han escrito diversas publicaciones interesantes que han tratado sobre las características del lenguaje jurídico (Rodríguez-Aguilera, 1969; Duarte & Martínez, 1995; Alcaraz & Hughes,

[1] https://boe.es/boe/dias/2015/10/02/pdfs/BOE-A-2015-10565.pdf

[2] https://lenguajeadministrativo.com/wp-content/uploads/2013/05/cmlj-recomendaciones.pdf

2002; Samaniego, 2005; González Salgado, 2009; Montolío Durán, 2012; Jiménez Yañez, 2016, entre otros). Sin embargo, los trabajos centrados en el lenguaje administrativo son más escasos, y principalmente han abordado el tema desde un punto de vista descriptivo y, por lo general, teniendo en cuenta como potencial emisor de los textos a los profesionales del ámbito y no a la ciudadanía (Ayala et al., 2000; de Miguel Aparicio, 2000; Castellón, 2001; Sánchez Alonso, 2014, entre otros).

También existen algunos sistemas automáticos en el ámbito del PLN que pueden ser de utilidad a la hora de escribir textos, ya que detectan diferentes tipos errores, principalmente ortográficos, gramaticales y de estilo, como por ejemplo:[3]

- Stilus,[4] textos generales en español;
- LanguageTool,[5] textos generales en español e inglés, entre otras lenguas;
- Estilector,[6] textos académicos universitarios en español;
- Grammarly,[7] textos generales y algunos géneros textuales en inglés;
- ACROLINX,[8] textos generales en inglés.
- GrammarChecker,[9] textos generales en inglés.
- SWANfootnote,[10] textos científicos en inglés;
- CALeSE,[11] textos científicos en inglés;
- eWriting Pal,[12] textos académicos en inglés.

Otro caso interesante dentro de la Península Ibérica es el de LinguaKit[13], un corrector que analiza el texto buscando errores ortográficos, léxicos, gramaticales o de estilo. Actualmente únicamente funciona para el gallego, pero, según los autores, su arquitectura se podría adaptar a cualquier lengua, como el español, en caso de tener los recursos necesarios para ello, como diccionarios y analizadores morfosintácticos (Gamallo Otero et al., 2015).

No obstante, ninguno de estos sistemas se centra en textos del ámbito de la Administración y, además, la mayor parte de ellos son sistemas comerciales. Es de destacar también que, en el ámbito del PLN, normalmente se emplea información lingüística de varios tipos, pero en el campo de la redacción asistida es difícil encontrar investigaciones basadas en conocimiento extraído del análisis del discurso especializado para desarrollar herramientas informáticas relacionadas con la lengua, aunque en los últimos años ha habido autores que han remarcado la necesidad de usar información discursiva como entrada para dichas herramientas (Zhou et al., 2014).

3. Metodología

En este apartado se detalla el marco teórico del trabajo, junto con las funcionalidades de arText y los detalles de su implementación.

3.1. Marco teórico

En este trabajo se emplean tres marcos teóricos complementarios. En primer lugar, con respecto al nivel textual, se parte, por un lado, de los trabajos de Van Dijk (1977, 1989), quien afirma que los géneros textuales siguen un patrón claramente codificado y ampliamente aceptado. Por ejemplo, un artículo de investigación suele incluir ciertos apartados prototípicos: Introducción, Estado de la cuestión, Metodología, Resultados y Conclusiones. Este autor, además, define la superestructura como la estructura organizativa textual, que varía dependiendo del tipo de texto y que, en general, se muestra mediante distintos apartados, que suelen incluir títulos y diferentes contenidos. Por otro lado, se utiliza también el concepto *moves* ("movidas retóricas" o "movimientos") de Swales (1990) para caracterizar la superestructura de los cinco géneros textuales que se incluyen en este trabajo, en la línea del análisis basado en corpus propuesto por Biber et al. (2007). Concretamente, en este marco, un *move* representa un fragmento de texto que sirve para una función comunicativa y semántica particular (Upton & Cohen, 2009), y que suele insertarse en alguno de los apartados del documento.

En segundo lugar, en cuanto al nivel léxico, se sigue la Teoría Comunicativa de la Terminología (TCT) de Cabré (1999), que es una teoría de la terminología y del discurso especializado que pone de relieve la dimensión comunicativa de los textos de especialidad. Según la TCT, los géneros textuales producidos en ámbitos especializados presentan algunas características globales, como

[3] Aunque este trabajo se centra en la lengua española, se incluyen aquí también referencias de sistemas automáticos para el inglés para dejar patente la escasa investigación previa en el ámbito incluso en esta lengua, que es la lengua para la cual, por lo general y como es sabido, se suelen desarrollar más aplicaciones y herramientas informáticas.

[4] http://www.mystilus.com/

[5] https://www.languagetool.org/

[6] http://www.estilector.com/

[7] http://www.grammarly.com/

[8] http://www.acrolinx.com/

[9] https://www.e-uned.es/correctme/cm_english/

[10] https://cs.joensuu.fi/swan/

[11] http://www.nilc.icmc.usp.br/calese/

[12] http://www.ewritingpal.com/

[13] https://linguakit.com/es/supercorrector

la precisión, la concisión, la sistematicidad, la impersonalidad y la objetividad. Estas características se hacen evidentes en los textos por medio que diferentes rasgos lingüísticos. Por ejemplo, la concisión se puede lograr a través del uso de siglas y la objetividad se puede alcanzar evitando marcadores de subjetividad, entre otras estrategias (Cabré et al., 2010).

En tercer lugar, en relación con el nivel discursivo, se emplea la *Rhetorical Structure Theory* (RST) de Mann & Thompson (1981), que es una teoría de organización textual que permite describir un documento caracterizando su estructura mediante las relaciones retóricas que mantienen sus segmentos discursivos. Estas relaciones pueden ser núcleo-satélite o multinucleares. En las relaciones núcleo-satélite hay dos elementos: uno de ellos es más relevante de cara a los propósitos del emisor (el núcleo), mientras que el otro (el satélite) aporta una información adicional sobre el núcleo. En las relaciones multinucleares, en cambio, puede haber más de dos elementos, todos ellos núcleos, que se relacionan al mismo nivel, es decir, todos tienen la misma importancia de cara a los propósitos del autor. Estos elementos se llaman segmentos discursivos o *elementary discourse units* (EDUs), y pueden definirse como indican (Tofiloski et al., 2009, p.77):

> *Discourse segmentation is the process of decomposing discourse into elementary discourse units (EDUs), which may be simple sentences or clauses in a complex sentence, and from which discourse trees are constructed.*

Los criterios de segmentación discursiva específicos para el español utilizados en este trabajo se extrajeron de da Cunha & Iruskieta (2010). En el ejemplo (1) se incluye una oración (que podría ser parte de una alegación) donde se han marcado los tres segmentos discursivos que contiene (indicados entre corchetes, junto con el número de segmento):

(1) [Ese día el Ayuntamiento no tuvo en cuenta la seguridad ciudadana,]SEGMENTO1 [es decir, no adoptó las medidas necesarias que garantizasen un correcto funcionamiento de los servicios públicos municipales,]SEGMENTO2 [pero sí es cierto que supo reconocer su error públicamente.]SEGMENTO3

da Cunha & Montané (2020) seleccionaron ocho relaciones discursivas para el análisis junto con sus correspondientes conectores discursivos: siete relaciones núcleo-satélite (antítesis, causa, concesión, condición, propósito, reformulación y resumen) y una relación multinuclear (contraste). Esta selección se realizó a través de la búsqueda de relaciones discursivas que aparecen de manera frecuente en el corpus de referencia anotado con relaciones discursivas en español, el *RST Spanish Treebank*[14], y que a la vez se evidencian habitualmente por medio de conectores. En el marco del presente trabajo, y de cara a integrarlos en arText, se hace una diferencia entre conectores discursivos intraoracionales (conectores que enlazan segmentos dentro de una misma oración) y conectores interoracionales (conectores que enlazan diferentes oraciones). Por ejemplo, la oración mostrada en el ejemplo (1) incluye dos conectores intraoracionales que encabezan los segmentos 2 y 3: "es decir", que refleja una relación de reformulación, y "pero", que marca una relación de antítesis. Si esta oración se dividiese en tres oraciones diferentes, más breves, esos dos conectores intraoracionales deberían sustituirse por conectores interoracionales que expresasen la misma relación discursiva, como por ejemplo "en otras palabras" y "sin embargo", respectivamente, tal como se recoge en el ejemplo (2), donde ambos conectores se han marcado en cursiva:

(2) [Ese día el Ayuntamiento no tuvo en cuenta la seguridad ciudadana.]SEGMENTO1 [*En otras palabras*, no adoptó las medidas necesarias que garantizasen un correcto funcionamiento de los servicios públicos municipales.]SEGMENTO2 [*Sin embargo*, sí es cierto que supo reconocer su error públicamente.]SEGMENTO3

Partiendo de estos tres marcos teóricos complementarios, uno de los resultados del trabajo de da Cunha & Montané (2020) fue una estructura modelo de cada uno de los cinco géneros textuales analizados. La estructura modelo de cada género incluye sus apartados prototípicos, la indicación de si estos suelen incluir un título y cuál es en caso de existir, y los *moves* (Swales, 1990) habituales en cada apartado. Además, se ofrece fraseología habitual que puede utilizarse para expresar dichos *moves*. Por ejemplo, en una alegación los apartados prototípicos detectados son los siguientes: "Cabecera", "Identificación del emisor", "Exposición de hechos", "Presentación de alegaciones", "Petición" y "Cierre". En cuanto a los *moves*, en el caso del apartado "Presentación de alegaciones", por ejemplo, se identificaron cuatro habituales: "Fórmula de introducción de alegaciones", "Título introductorio al listado

[14] http://www.corpus.unam.mx/rst/index_es.html

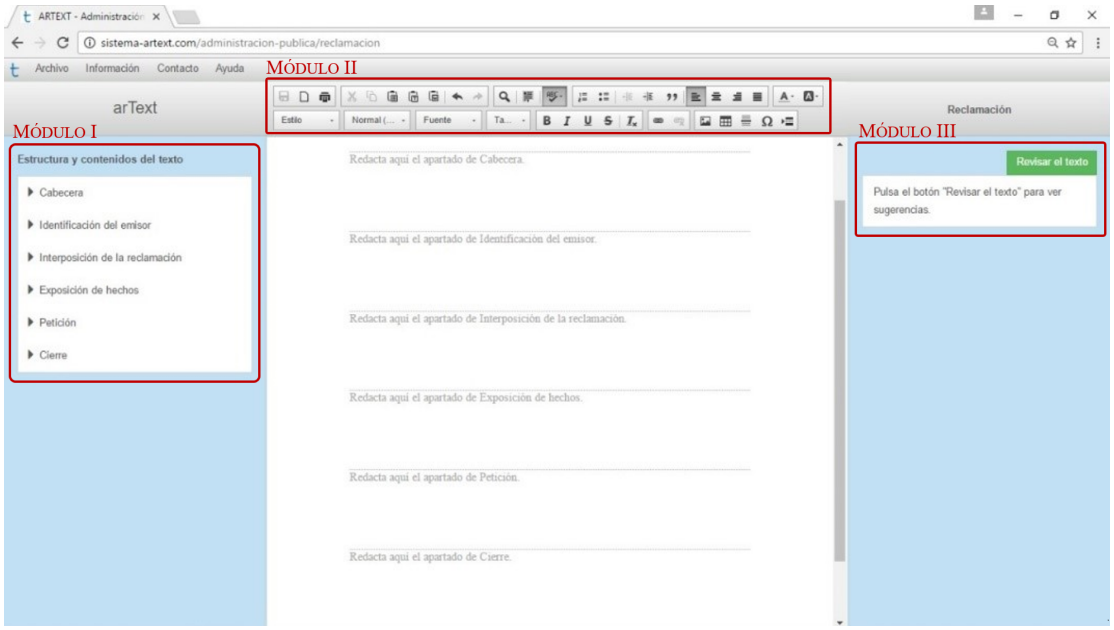


Figura 1: Captura de pantalla del editor de textos en línea de arText, con indicación de sus tres módulos en color rojo.

de alegaciones", "Listado de las alegaciones presentadas por el emisor" y "Mención de los documentos adjuntos". En el caso del primero de estos cuatro *moves*, en la estructura modelo se incluyen algunas unidades fraseológicas (que en este caso se corresponden con oraciones completas, aunque no siempre es así, ya que depende del género textual del que se trate) que pueden ayudar a expresarlo en un texto, como por ejemplo:[15]

> *Que mediante este escrito, y sin perjuicio de lo que pueda manifestar en el trámite de audiencia, formula las siguientes alegaciones.*
>
> *Que no estando conforme con los hechos denunciados, interpone el presente escrito en base a las siguientes alegaciones.*
>
> *Que dentro del plazo que le ha sido concedido y al amparo de lo previsto en [documentación legal aplicable: artículos de leyes, leyes, reglamentos, etc.], formula alegaciones y presenta los documentos y justificaciones necesarios.*

Otro resultado de dicho trabajo fue una caracterización lingüística de cada género textual, que incluye tanto ciertos aspectos léxicos como discursivos. En relación con los aspectos léxicos, se incluye, por ejemplo, información sobre el uso de siglas, de unidades subjetivas, y de la primera persona del singular y del plural. En cuanto a aspectos discursivos, se aporta información sobre la cantidad de oraciones, de segmentos discursivos y de conectores discursivos que suelen contener.

El diseño de arText, cuyo desarrollo es el objetivo del presente artículo, parte, como se verá en el apartado 3.2, por un lado, de las estructuras modelo mencionadas para cada uno de los cinco géneros textuales, y, por otro, de su caracterización lingüística.

3.2. Módulos y funcionalidades de arText

El sistema arText es una aplicación tecnológica que tiene forma de editor de textos en línea, y que puede utilizarse gratuitamente y sin necesidad de registro[16]. Hay disponible un tutorial en línea sobre sus funcionalidades en Canal UNED[17]. Una vez el usuario entra en la página principal del sistema, debe seleccionar el botón "EMPIEZA A USAR ARTEXT" y, a continuación, indicar el género textual que desea redactar. El usuario puede elegir entre los cinco géneros textuales del ámbito de la Administración mencionados en el apartado 1: alegación, carta de presentación, queja, reclamación y solicitud. Una vez seleccionado el género textual, el usuario entrará automáticamente en el editor en línea (véase Figura 1, en donde se ha seleccionado el género reclamación). Desde este editor, puede comenzar a redactar su escrito en la hoja de texto central, utilizando los tres módulos de ayuda con los que cuenta el sistema, que se detallan a continuación:

[15] Los corchetes indican información variable en cada texto.

[16] http://sistema-artext.com/

[17] https://canal.uned.es/mmobj/index/id/54433

I. Módulo de estructura y contenidos del texto

El primer módulo ayuda a estructurar el documento y a comenzar a redactarlo, utilizando las estructuras modelo de cada género. Este módulo se encuentra en la columna izquierda de la pantalla y permite:

- Insertar los apartados prototípicos del documento.
- Añadir contenidos habitualmente presentes en cada apartado (los mencionados *moves* de Swales (1990)).
- Incorporar fraseología relacionada con los contenidos.

Tal como muestra la Figura 1, lo primero que visualiza el usuario en la columna izquierda es la lista de apartados sugeridos por arText para el género textual seleccionado, la reclamación. Asimismo, como se aprecia también en la Figura 1, el usuario verá estos mismos apartados en la hoja de texto, siempre precedidos por la siguiente indicación en color gris: "Redacta aquí el apartado de. . . ". Una vez el usuario comience a escribir el texto en el apartado correspondiente, cada una de estas indicaciones se eliminará automáticamente.

Al hacer clic en alguno de los apartados sugeridos en la columna izquierda, se despliega debajo una lista que incluye los contenidos que se recomienda insertar en él, en el orden propuesto. El título del apartado, si procede incorporarlo, se muestra como el primer contenido de la lista. Al situar el cursor sobre un contenido, este se evidencia mediante un recuadro con fondo azul. En la Figura 2 se ofrece un ejemplo, en que el apartado "Exposición de hechos" incluye dos contenidos: "Explicación del hecho o hechos que motivan la reclamación" y "Mención de los documentos adjuntos".

A su vez, al hacer clic sobre cada contenido, se muestra una lista de frases prototípicas que pueden utilizarse para expresarlo, que no aparecen en un orden determinado. El usuario debe elegir cuáles de estas frases quiere incorporar en su texto, teniendo en cuenta el objetivo del documento que está redactando. Estas frases están escritas en color azul y, al poner el cursor encima de una de ellas, aparecerá subrayada. Para incorporar las frases deseadas en el documento, basta con hacer clic sobre cada una de ellas en la columna de la izquierda y automáticamente se cargarán en el texto, dentro del apartado correspondiente. Las frases propuestas por arText pueden incluir varios componentes:

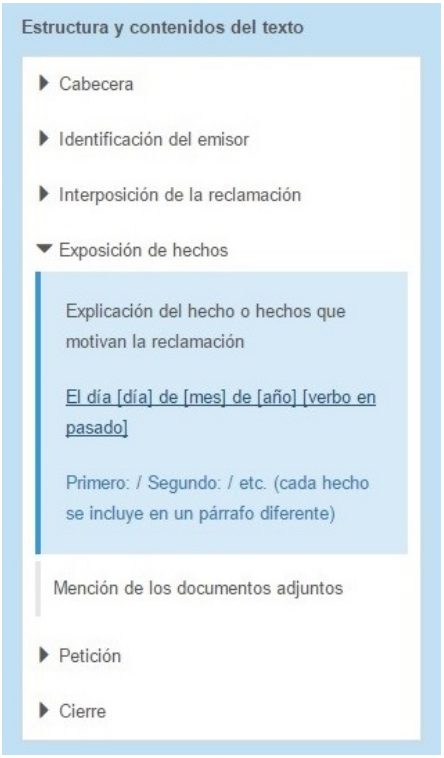


Figura 2: Detalle de una captura de pantalla de arText donde se muestran los apartados y contenidos incluidos en el módulo 1 para el género reclamación.

- Elementos sin marca alguna: son fragmentos literales que el usuario podría incorporar en su texto directamente, como por ejemplo: "Mediante el presente escrito presenta reclamación de responsabilidad patrimonial de la administración en base a los siguientes hechos:".
- Elementos entre corchetes: se trata de componentes variables que el usuario debe modificar para adaptar a su texto. Por ejemplo, si arText ofrece la siguiente secuencia "El día [día] de [mes] de [año] [verbo en pasado]" el usuario debería transformarla sustituyendo la información entre corchetes, como por ejemplo de la siguiente manera: "El día 5 de enero de 2015 solicitó".
- Elementos entre paréntesis: se trata de instrucciones sobre la estructura textual sugeridas al usuario. Ej. "(cada hecho se incluye en un párrafo diferente)" quiere decir que cada uno de los hechos redactados por el usuario debe ocupar un párrafo distinto en el texto.

En la Figura 2 se incluye la fraseología asociada al contenido "Explicación del hecho o hechos que motivan la reclamación", ubicado en el apartado "Exposición de hechos".

II. Módulo de corrección ortográfica y formato

El segundo módulo incorpora una barra superior que incluye diferentes opciones de formato mediante las cuales el usuario tendrá acceso a las funcionalidades más habituales de los editores de texto, como seleccionar el tamaño de letra y la fuente, asignar estilos, insertar tablas, hacer listas, copiar, pegar, buscar, etc. (véase Figura 1). Asimismo, en esta barra de formato se ha integrado un corrector ortográfico.

III. Módulo de sugerencias sobre léxico y discurso

El sistema arText cuenta con un tercer módulo de sugerencias que permite al usuario procesar lingüísticamente su texto, y visualizar recomendaciones sobre cuestiones relacionadas con aspectos lingüísticos (léxicos y discursivos, principalmente) que se podrían mejorar. Estas recomendaciones están basadas en la caracterización lingüística de los cinco géneros textuales analizados por da Cunha & Montané (2020).

Para visualizar las sugerencias, el usuario debe tener escrito su texto en el editor. En el ejemplo (3) se incluye un texto ficticio que se corresponde con el género textual de la reclamación (donde se ha eliminado la información sobre el emisor y el receptor):

(3) Mediante el presente escrito presenta reclamación de responsabilidad patrimonial de la administración en base a los hechos que se describen a continuación.

El día 2 de mayo de 2019 el personal del Ayuntamiento instaló cuatro pivotes de gran tamaño en el número 46 de la calle Galileo. Creemos que esta decisión no se debió de tomar teniendo en cuenta las normas urbanísticas del Plan General de Ordenación Urbana de Madrid (PGOUM). En este Plan General de Ordenación Urbana de Madrid se determinaron las situaciones en las cuales es posible instalar este tipo de elementos urbanos, es decir, se definieron supuestos en los que dicha instalación se debería efectuar. Evidentemente, la instalación de estos cuatro bolardos supone una negligencia, porque imposibilitan la salida de vehículos por la puerta del garaje de la vivienda que se encuentra delante de los mismos, pero considero que es posible solucionar la situación mediante una opción alternativa. Esta opción sería eliminar dos de los cuatro pivotes, es decir, mantener únicamente dos, separados a una distancia de 3 metros. Creo que debemos ser partícipes de las decisiones tomadas por la administración que nos afecten, es decir, debe tenerse en cuenta la opinión de la comunidad de vecinos de la zona porque es su derecho poder decidir sobre estas cuestiones. En otras comunidades, como la AVIT, hace tiempo que negocian sobre cómo alcanzar un pacto para la regeneración de ideas en relación con esta materia. Deberíamos fijarnos también en la CVBM para observar maneras distintas de gestionar nuestros procesos de toma de decisiones, porque no es deseable caer en los errores de siempre.

Se adjunta la siguiente documentación justificativa:

Ley 30/1992, de 26 de noviembre, de Régimen Jurídico de las Administraciones Públicas y del Procedimiento Administrativo Común.

Real Decreto 429/1993, de 26 de marzo, por el que se aprueba el Reglamento de las Administraciones Públicas en materia de responsabilidad patrimonial.

Solicita al Ayuntamiento que admita esta reclamación patrimonial, junto con los documentos que se acompañan.

Una vez escrito el texto, el usuario debe hacer clic sobre el botón "Revisar el texto" de la columna derecha de la pantalla. Entonces, el sistema le ofrecerá en esa misma columna diferentes recomendaciones de mejora del texto, como se recoge en la Figura 3.

En función de la recomendación sobre la que el usuario haga clic, aparecerán marcadas en el texto que ha escrito diferentes cuestiones, como oraciones largas, siglas, palabras repetidas, etc. A continuación se detallan las recomendaciones de mejora ofrecidas al usuario al procesar el texto, junto con un ejemplo representativo de cada una de ellas:

a) "*División de oraciones largas*. Parece que las oraciones marcadas podrían dividirse en otras más cortas. Te recomendamos que lo hagas. Haz clic en cada oración para ver dónde podrías segmentarla."

En este caso, el sistema marca en amarillo en el texto las oraciones consideradas demasiado largas para el género textual seleccionado (véase Figura 4)[18].

[18] Se toman como referencia los umbrales de palabras para cada género detectados por da Cunha & Montané (2020).

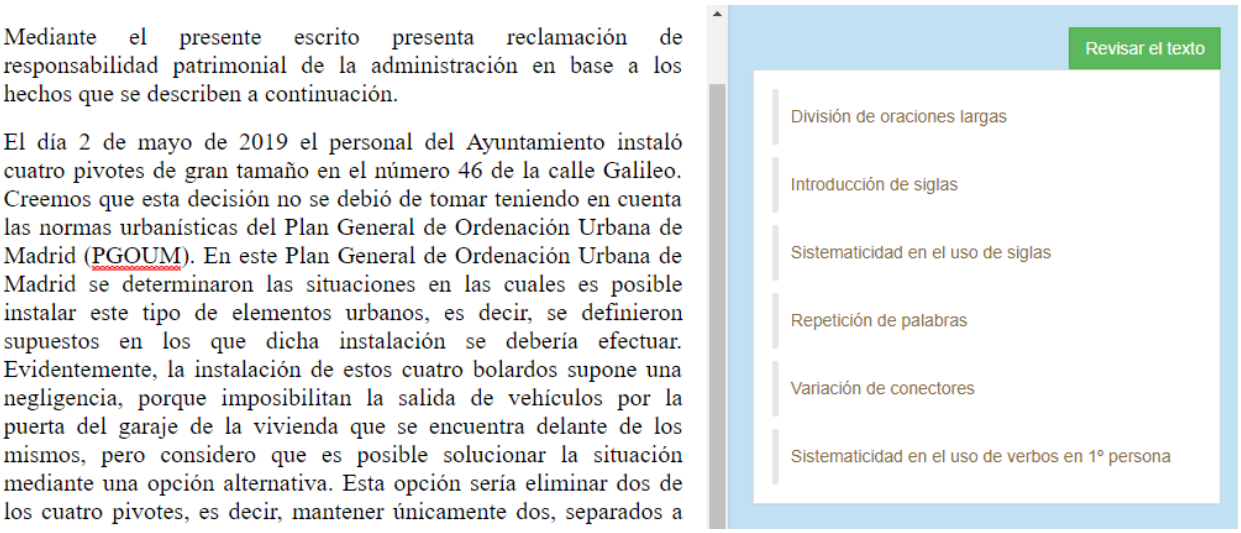


Figura 3: Detalle de una captura de pantalla de arText que muestra las recomendaciones de mejora obtenidas al procesar un texto.

Además, si el sistema logra dividir automáticamente la oración en segmentos discursivos más breves, también ofrece esta información al usuario en la columna derecha. Veamos el ejemplo (4), que incluye una oración extraída del texto del ejemplo (3):

(4) Evidentemente, la instalación de estos cuatro bolardos supone una negligencia, porque imposibilitan la salida de vehículos por la puerta del garaje de la vivienda que se encuentra delante de los mismos, pero considero que es posible solucionar la situación mediante una opción alternativa.

En este caso, arText propone al usuario dividir esta oración larga en tres oraciones más breves, teniendo en cuenta los tres segmentos siguientes (véase Figura 4):

I. *Evidentemente, la instalación de estos cuatro bolardos supone una negligencia,*

II. *porque imposibilitan la salida de vehículos por la puerta del garaje de la vivienda que se encuentra delante de los mismos,*

III. *pero considero que es posible solucionar la situación mediante una opción alternativa.*

Asimismo, en caso de que los segmentos propuestos comiencen por conectores discursivos intraoracionales, el sistema propone al usuario opciones de conectores interoracionales alternativos que expresen la misma relación discursiva. Por ejemplo, siguiendo con la misma oración, en el texto se marcarían en naranja dos conectores: "porque" (que refleja una relación de causa) y "pero" (que refleja una relación de antítesis). Si el usuario hace clic en "pero", verá en la columna de la derecha las siguientes propuestas de conectores interoracionales alternativos, que expresan la misma relación de antítesis: "Con todo", "De todas formas", "De todas maneras", "De todos modos", "No obstante" y "Sin embargo" (véase Figura 4). De esta manera, el usuario obtendrá información útil que le ayudará a decidir si desea conservar en su texto la oración original o si, por el contrario, prefiere dividirla en otras oraciones más breves y mantener explícita la relación discursiva que existe entre ellas a través de conectores. Es el usuario quien decide en todo momento qué recomendaciones desea incorporar en su texto. Para hacerlo, simplemente debe realizar los cambios directamente en la hoja del editor de textos.

b) "*Introducción de siglas.* Las unidades marcadas parecen siglas. Si es así, ten en cuenta que la primera vez que se utiliza una sigla en un texto suele ir acompañada del término desplegado."

En este caso, el sistema marca en amarillo en el texto las siglas que no se acompañan de su correspondiente término desplegado la primera vez que aparecen en el texto, lo cual no es recomendable. El sistema solo detecta siglas propias (Giraldo, 2008), es decir, aquellas que están formadas únicamente por las unidades léxicas incluidas en la estructura sintagmática del término desplegado, como por ejemplo "UNED", que se corresponde con "Universidad Nacional de Educación a Distancia". En la Figura 5 puede verse que el sistema ha detectado dos de estos casos, como son "AVIT" y "CVBM".

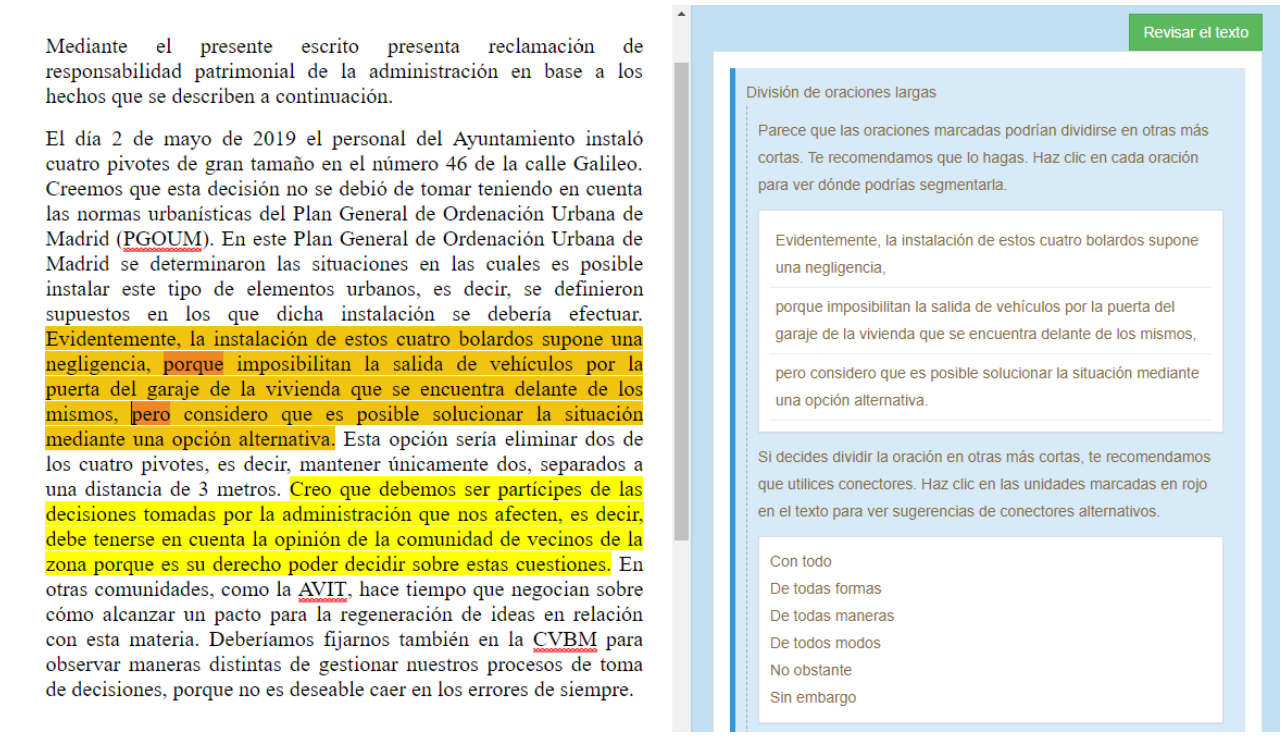


Figura 4: Detalle de una captura de pantalla de arText que refleja la recomendación *División de oraciones largas*.

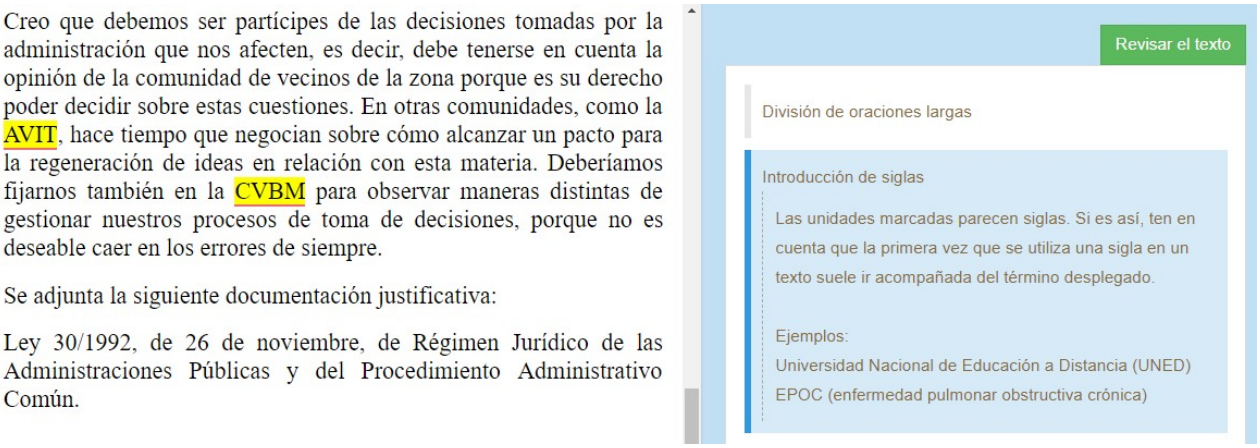


Figura 5: Detalle de una captura de pantalla de arText que refleja la recomendación *Introducción de siglas*.

c) "*Sistematicidad en el uso de siglas*. Las unidades marcadas parecen el término desplegado de siglas que utilizas en el texto. Si es así, ten en cuenta que, una vez se introduce una sigla en un texto, se suele seguir utilizando la sigla y no el término desplegado."

Esta recomendación también está relacionada con las siglas. Sin embargo, en esta ocasión el sistema marca en el texto las ocurrencias de términos desplegados que ya habían sido introducidos en el texto previamente junto con su correspondiente sigla, lo cual es desaconsejable. La Figura 6 muestra un ejemplo, en que la unidad "Plan General de Ordenación Urbana de Madrid" está marcada dos veces. La primera vez aparece en una oración junto con su sigla entre paréntesis, correctamente. Sin embargo, también aparece marcada en la siguiente oración, ya que se quiere hacer notar al usuario que en este caso sería más conveniente sustituir esta unidad por su sigla.

d) "*Repetición de palabras*. Las unidades de la lista siguiente se repiten varias veces en el texto. Ten en cuenta que en este tipo de textos puedes utilizar variantes como sinónimos, explicaciones, paráfrasis, etc. Haz clic en cada unidad para visualizar sus ocurrencias en el texto."

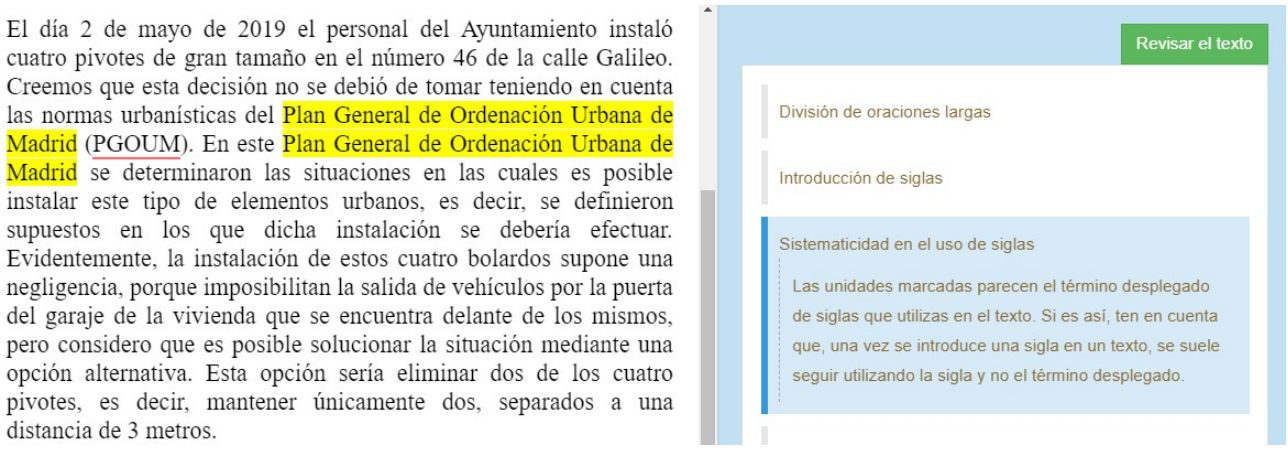


Figura 6: Detalle de una captura de pantalla de arText que refleja la recomendación *Sistematicidad en el uso de siglas*.

Esta recomendación da al usuario una lista de las unidades léxicas repetidas en el texto, concretamente de cuatro categorías gramaticales: nombres, verbos, adjetivos y adverbios. Cuando el usuario hace clic en alguna de estas unidades, en el texto se destacan en amarillo sus diferentes ocurrencias. El sistema lematiza todas las palabras del documento, por lo que se marcarán en el texto todas las formas detectadas de los nombres y de los adjetivos (singular, plural, masculino y femenino) y de los verbos (independientemente de su tiempo, modo, número y persona). Por ejemplo, en la Figura 7, se observa en la columna derecha un fragmento de la lista de palabras repetidas, que incluye "administración", "deber", "decir", "decisión", "patrimonial" y "urbano". En el caso del verbo "deber", se marcan en el texto las formas "debería", "debemos", "debe" y "deberíamos". Es importante destacar que el hecho de que se marquen estas unidades en el texto no quiere decir que sean incorrectas. Lo que se pretende con esta recomendación es ofrecer al usuario información para decidir si desea eliminar o modificar alguna de ellas.

e) "*Variación de conectores*. Los conectores de la lista siguiente se repiten varias veces en el texto. Haz clic en cada conector para ver sugerencias de conectores alternativos."

En este caso, el sistema ofrece en la columna derecha un listado de los conectores discursivos que se repiten en el texto tres o más veces. Si el usuario hace clic en alguno de ellos, verá en el texto marcadas sus ocurrencias y además obtendrá en la columna derecha un listado de conectores alternativos (tanto intraoracionales como interoracionales) que expresan la misma relación discursiva. Por ejemplo, en la Figura 8 se observan marcadas en el texto tres ocurrencias del conector de reformulación "es decir" y en la columna derecha una lista con cuatro propuestas de conectores alternativos, como son "dicho de otro modo", "en otras palabras", "esto es" y "o sea".

f) "*Sistematicidad en el uso de verbos en 1º persona*. Las unidades marcadas en verde parecen verbos en 1º persona del singular y las marcadas en azul parecen verbos en 1ª persona del plural. Te recomendamos que optes por el singular o el plural para que el texto sea sistemático."

Esta recomendación tiene que ver con la falta de sistematicidad en el uso de la primera persona del singular y del plural, especialmente en relación con el autor del escrito. Hay ocasiones en que en un texto es necesario utilizar ambas opciones, pero, en los textos administrativos, es importante que el autor utilice el singular o el plural de manera sistemática, en función de quién o quiénes emitan la alegación, queja, reclamación, etc. Para ello, el sistema marca en el texto las formas verbales en singular en verde y las formas en plural en azul, dando así información al autor para tomar la decisión de hacer modificaciones en este sentido. Por ejemplo, en la Figura 9, se marcan en verde las formas en singular "considero" y "creo", mientras que se destacan en azul las formas "creemos", "debemos" y "deberíamos".

g) "*Uso de indicadores de subjetividad*. Las unidades marcadas podrían ser indicadoras de subjetividad. Ten en cuenta que este tipo de textos suelen ser objetivos. Te recomendamos que revises estas unidades para confirmar que son adecuadas en tu texto."

Las recomendaciones anteriores se aplican a los cinco géneros textuales incluidos en arText.

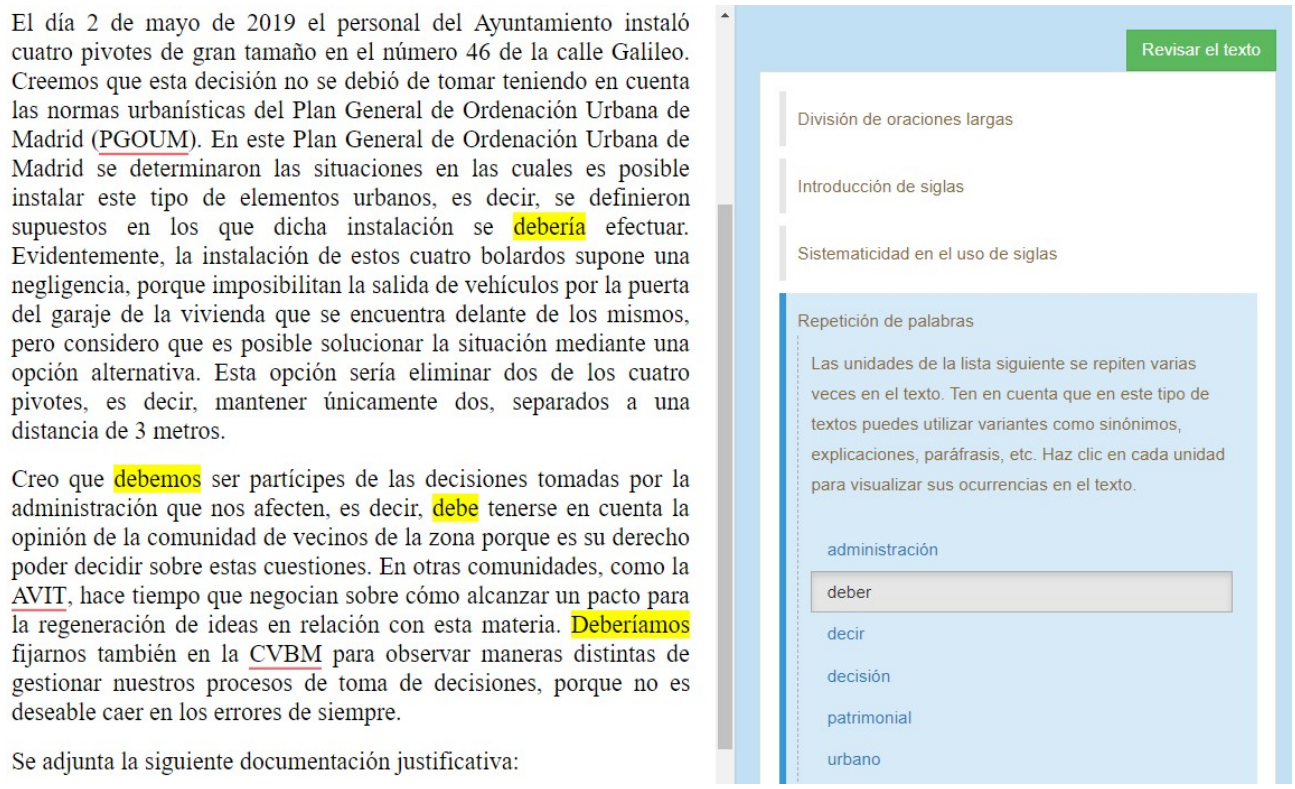


Figura 7: Detalle de una captura de pantalla de arText que refleja la recomendación *Repetición de palabras*.

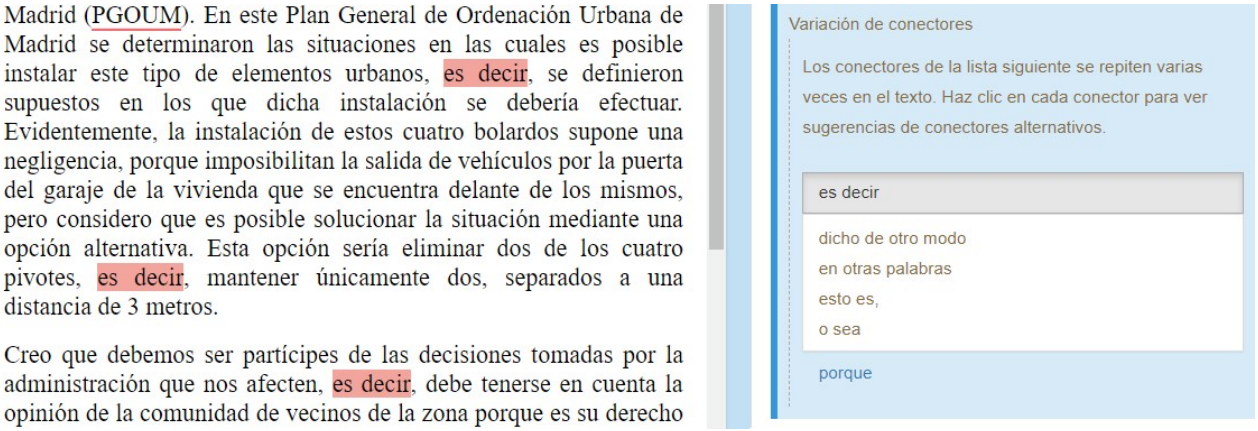


Figura 8: Detalle de una captura de pantalla de arText que refleja la recomendación *Variación de conectores*.

Sin embargo, en este caso, se trata de una recomendación específica para el género textual solicitud, ya que en da Cunha & Montané (2020) se observó que en este tipo de género se tiende a evitar la subjetividad (en contraposición con una queja o una carta de presentación, por ejemplo). Así, como puede observarse en la Figura 10, si se selecciona este género, se marcan en el texto las unidades indicadoras de subjetividad, como "evidentemente", para que el usuario decida si quiere eliminarlas de su escrito.

3.3. Implementación del sistema

El sistema arText se desarrolló en un entorno Linux usando un servidor Apache. Se utilizaron también diferentes recursos tanto en el *back-end* (Bash, Perl y PHP, con un entorno de trabajo Laravel) como en el *front-end* (HTML, CSS, JavaScript, con AJAX y jQuery). Algunos de los recursos principales se detallan en este apartado. El sistema está optimizado para su utilización con el navegador Google Chrome.

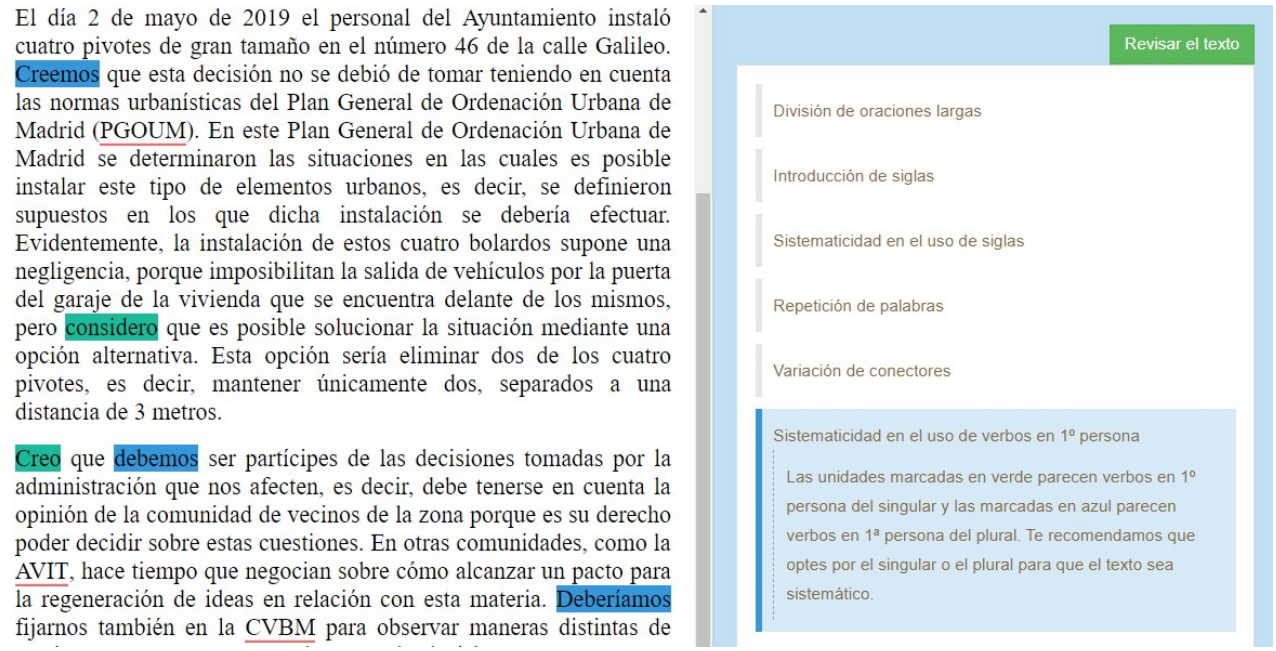


Figura 9: Detalle de una captura de pantalla de arText que refleja la recomendación *Sistematicidad en el uso de verbos en 1º persona*.

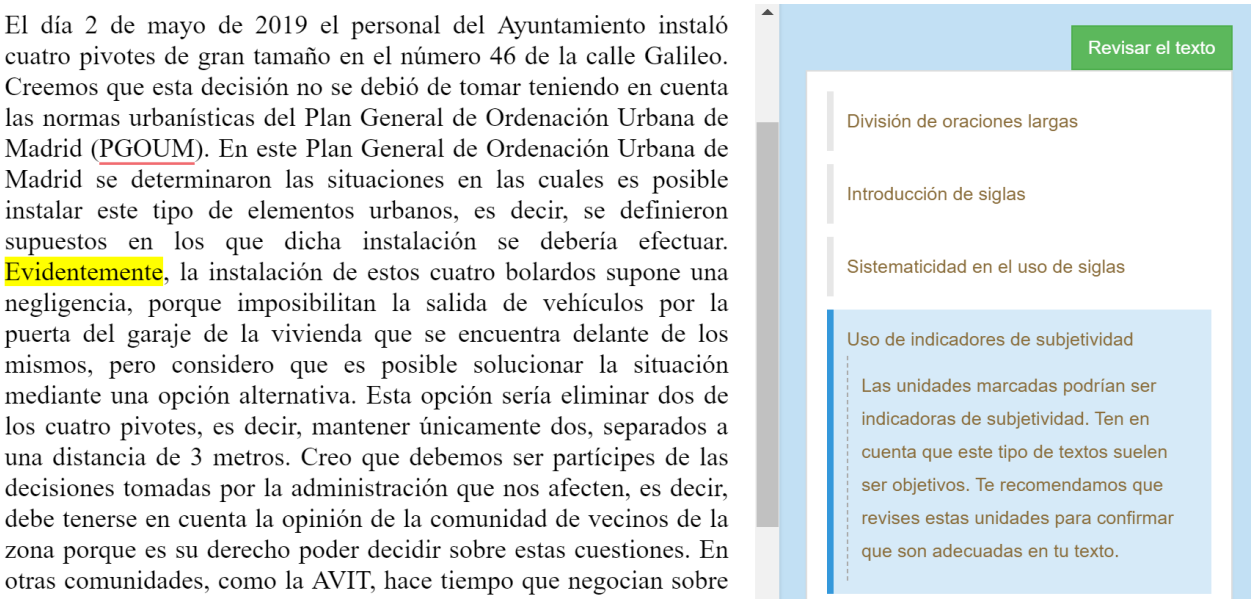


Figura 10: Detalle de una captura de pantalla de arText que refleja la recomendación *Uso de indicadores de subjetividad* en el género textual solicitud.

El principal recurso empleado para la implementación del Módulo I fue una base de datos MySQL, donde se guarda la información relativa a la estructura textual de cada uno de los cinco géneros que incluye arText, es decir, los apartados, los contenidos (incluyendo aquí los títulos) y la fraseología. MySQL es un sistema de gestión de bases de datos relacionales de código abierto con un modelo cliente-servidor que nos pareció adecuado por la facilidad en la gestión de los datos y por los requerimientos del sistema, ya que se necesitaba flexibilidad para crear o editar la estructura (ya que se tiene prevista la creación de nuevos ámbitos y géneros, con diferentes apartados, contenidos y fraseología).

Por su parte, el Módulo II, además de la barra de formato, incorpora un corrector ortográfico *opensource* (WebSpellChecker Ltd). El acceso a WebSpellChecker se hace por medio de una API desde la nube. Se trata de un *plugin* del editor de texto de código abierto CKEditor, utilizado también en la implementación del sistema.

Finalmente, el Módulo III incluye dos herramientas de PLN existentes actualmente para el español que permiten realizar un procesamiento lingüístico del texto escrito por el usuario. En primer lugar, incorpora el analizador morfosintáctico de Freeling (Atserias et al., 2006), mediante el cual se lematizan todas las unidades léxicas del texto y se asigna una categoría gramatical a cada una de ellas. En segundo lugar, incorpora el segmentador discursivo DiSeg (da Cunha et al., 2012), que permite dividir el texto en oraciones y, además, en segmentos discursivos intraoracionales, siguiendo, el concepto de segmento discursivo de Tofiloski et al. (2009) y los criterios de segmentación para el español de da Cunha & Iruskieta (2010), mencionados en el apartado 3.1.

Asimismo, se implementaron en el Módulo III diferentes algoritmos que toman como entrada el texto procesado lingüísticamente por las dos herramientas de PLN mencionadas. Estos algoritmos permiten detectar en el texto escrito por el usuario los elementos lingüísticos necesarios para poder ofrecer las recomendaciones correspondientes asociadas a cada uno de ellos. Estos elementos son:

- Oraciones largas, con un umbral diferente de palabras para cada género textual.
- Conectores discursivos interoracionales e interoracionales que evidencian las ocho relaciones discursivas utilizadas en la investigación, extraídos del trabajo de da Cunha & Montané (2020).
- Conectores discursivos que se repiten tres o más veces en el texto, extraídos del trabajo de da Cunha & Montané (2020).
- Unidades léxicas que indican subjetividad, como marcas de superlativos (ej. "-ísimo"), y ciertos adjetivos (ej. "bueno"), adverbios (ej. "evidentemente") y frases (ej. "sin ninguna duda"), extraídas del trabajo de Otaola Olano (1988).
- Siglas propias y sus correspondientes términos desplegados. En este caso, para hacer la correlación entre la sigla y su término desplegado, se tiene en cuenta que la letra inicial de las unidades léxicas incluidas en el término (excepto las *stopwords*) se correspondan, en el mismo orden, con las mismas letras que incluye la sigla.

En el Módulo III se gestiona el lado cliente por medio de clases/objetos JavaScript que se encuentran en un fichero PHP (*Hypertext Pre-Processor*). Los algoritmos implementados procesan los resultados de las clases lingüísticas que se encuentran en este fichero PHP. La clase principal es "app" y esta procesa varias clases, como oraciones siglas, oraciones, unidades subjetivas, conectores discursivos, etc. El motivo de esta elección es que JavaScript es un lenguaje de programación que permite realizar actividades complejas en una página web, y PHP es un lenguaje de código abierto muy popular especialmente adecuado para el desarrollo web y que puede ser incrustado en HTML. Por tanto, al ser arText un sistema para ser empleado en la web, nos pareció una opción acertada.

La arquitectura general del Módulo III de arText puede verse en la Figura 11.

En cuanto al formato de persistencia de los documentos de arText, se usó HTML5, que es la última versión de HTML, con nuevos elementos, atributos y comportamientos, y que contiene un conjunto más amplio de tecnologías que permite a los sitios web y a las aplicaciones ser más diversas y de gran alcance. La elección de este formato se debe a que su uso es requerido por el editor CKEditor.

En relación con la exportación e importación de documentos, por una cuestión de protección de datos, se decidió que el sistema no guardase en su servidor los textos escritos por los usuarios. Por tanto, si el usuario desea guardar un texto escrito en línea en arText, debe hacerlo en local. Para ello, existen varias opciones de exportación de documentos: .pdf, .txt, .html, y .arText. Para poder importar un texto posteriormente en arText debe utilizare el formato creado específicamente para esta aplicación, el formato .arText.

El sistema permite incluir en el documento imágenes, siempre que estas tengan asignada una URL, es decir, una dirección web. No se permite subir imágenes desde local porque, como actualmente no es necesario registrarse para utilizar el sistema, nuestro equipo no puede hacerse responsable de las imágenes subidas por los usuarios desde sus ordenadores personales. En la opción "Cómo subir imágenes desde Google Drive" (ubicada en la pestaña de Ayuda en la parte superior izquierda del editor; véase Figura 1) se explica cómo asignar una URL a una imagen.[19] Todas estas cuestiones se explican con detalle en el manual de uso del sistema.[20]

[19] http://sistema-artext.com/doc/como_subir_imagenes_desde_Drive.pdf

[20] http://sistema-artext.com/doc/manual.pdf

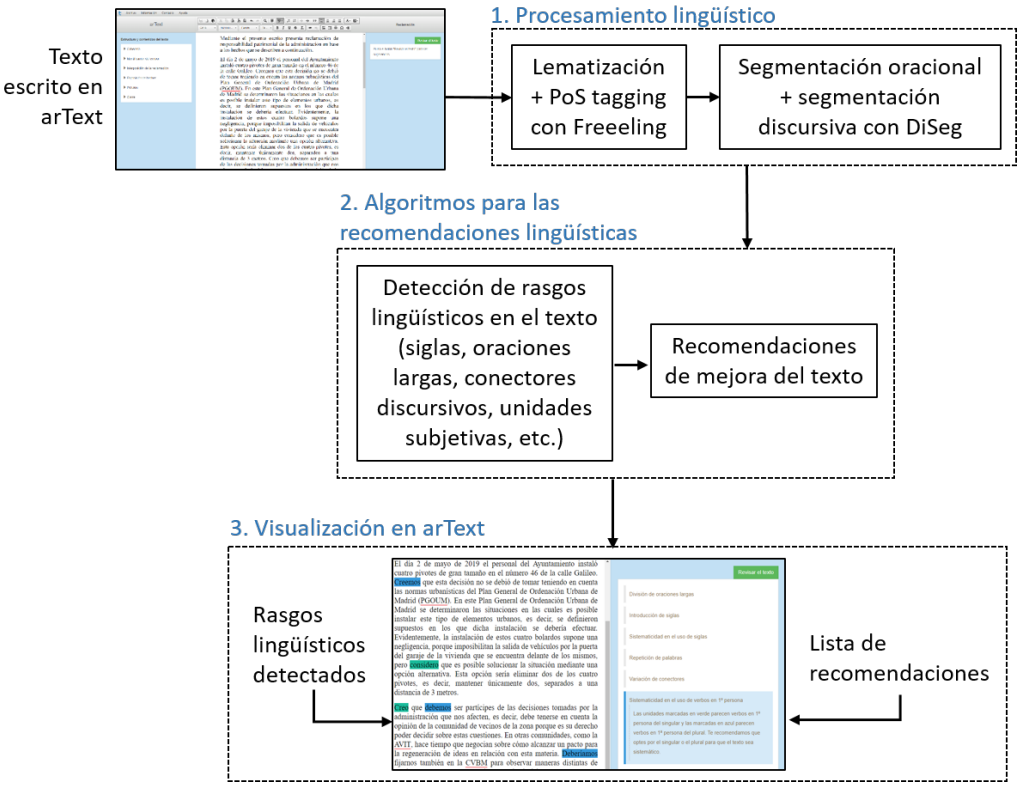


Figura 11: Arquitectura general del Módulo III de arText.

4. Resultados y evaluación

Se realizaron dos tipos de evaluaciones de arText: *data-driven* (basada en datos) y *user-driven* (basada en usuarios). Por un lado, la evaluación *data-driven* se llevó a cabo principalmente para comprobar que los algoritmos desarrollados en el Módulo III funcionaban correctamente. Para ello, en primer lugar, se compiló un corpus específico para la evaluación, que incluyó un total de 24 textos. De estos, 8 fueron del ámbito administrativo (concretamente, solicitudes), pero también se decidió añadir textos de diferentes ámbitos y géneros textuales, para confirmar que arText puede funcionar con otro tipo de textos. Así, se añadieron al corpus de evaluación 8 resúmenes de artículos de investigación del ámbito médico y 8 artículos de divulgación de revistas especializadas en turismo. En segundo lugar, un anotador con formación en lingüística y experiencia en anotación de corpus anotó manualmente los diferentes rasgos lingüísticos que se pretendían evaluar, incluidos en la Tabla 1.

Rasgos anotados
a) División de oraciones largas.
a1) Detección de oraciones largas.
a2) Segmentación discursiva.
a3) Detección de conectores discursivos interoracionales e intraoracionales que expresan la misma relación discursiva.
b) Introducción de siglas.
c) Sistematicidad en el uso de siglas.
e) Variación de conectores.
f) Sistematicidad en el uso de verbos en 1º persona.
f1) Sistematicidad en el uso de verbos en 1ª persona singular.
f2) Sistematicidad en el uso de verbos en 1ª persona singular.

Tabla 1: Rasgos anotados para la evaluación *data-driven* de arText.

En tercer lugar, se revisaron los textos del corpus automáticamente con arText. Finalmente, se compararon los resultados provenientes del análisis manual y de la revisión automática realizada con el sistema. Para ello se calculó la precisión y cobertura de los resultados de arText en contraposición con la anotación manual. Los resultados se muestran en la Tabla 2. Como puede observarse, los resultados obtenidos son en general muy positivos para todos los rasgos lingüísticos analizados.

Destaca especialmente que el sistema detectó correctamente todas las oraciones largas del corpus, todos los conectores discursivos incluidos en la Base de datos 1 y todas las unidades subjetivas incluidas en la Base de datos 2.

Recomendación	Precisión	Cobertura
a1)	1	1
a2)	0,74	0,87
a3)	1	1
b)	0,76	0,68
c)	0,75	0,94
e)	1	1
f1)	0,87	0,97
f2)	1	0,97

Tabla 2: Resultados de la evaluación *data-driven* de arText.

En el caso de la recomendación d) *Repetición de palabras*, la evaluación fue ligeramente diferente: no se anotaron manualmente en el corpus todas las palabras repetidas, sino que se comprobó si las palabras listadas por arText estaban marcadas en el texto y eran realmente diferentes formas del mismo lema. Los resultados indicaron que el 93,2 % de las unidades detectadas sí lo eran. Las principales dificultades se observaron en la detección de abreviaturas como "Ilmo." o "Sr.".

Con respecto a las causas de los errores detectados, pueden dividirse en errores derivados de las dos herramientas de PLN integradas en arText y errores de los algoritmos propios del sistema. Con respecto a los primeros, los errores relacionados con las recomendaciones sobre palabras repetidas y verbos en primera persona vienen derivados del analizador morfosintáctico, mientras que los errores en relación con la recomendación sobre segmentación discursiva son generados por el segmentador discursivo. Al tratarse de herramientas externas, en estos casos será difícil realizar mejoras en el sistema para evitar este tipo de errores. En cambio, sí podremos trabajar en el futuro en la búsqueda de soluciones a los errores producidos por los algoritmos propios de arText. Estos tienen que ver principalmente con la cobertura en la detección de siglas (como se ha mencionado en el apartado 3.2, actualmente únicamente se detectan siglas propias) y sus correspondientes términos desplegados.

Por otro lado, la evaluación *user-driven* permitió conocer la percepción de la utilidad de la aplicación por parte de usuarios reales. Para ello, en primer lugar, se diseñó un cuestionario de valoración con la herramienta Google Forms en que se pedía a los usuarios que probasen arText y, a continuación, se les preguntaba su opinión sobre diferentes aspectos. El cuestionario contiene 3 bloques, que incluyen distintas preguntas, detalladas en el Apéndice A:

1. Breve cabecera de contextualización del proyecto.
2. Pasos a seguir para la utilización de arText.
3. Preguntas de respuesta cerrada sobre:
   - Accesibilidad.
   - Estructuración y contenidos del texto.
   - Corrección ortográfica y formato.
   - Revisión del texto.
   - Valoración final.

En segundo lugar, se distribuyó el cuestionario a 25 ciudadanos, con estudios universitarios, entre 30 y 50 años, y con manejo de internet. En general, los participantes hicieron una valoración muy positiva de la aplicación. La valoración general indica que el 84 % de los usuarios considera que arText es muy útil y el 100 % lo recomendaría a otras personas. Las opciones ofrecidas por el editor resultaron muy claras o bastante claras al 100 % de los usuarios. El aspecto mejor valorado fue el Módulo I de arText, que incluye información sobre la estructura textual, ya que un 80 % de los usuarios considera que le ha parecido de mucha utilidad. El segundo aspecto mejor valorado tiene que ver con las sugerencias de revisión lingüística del texto ofrecidas en el Módulo III, puesto que al 76 % de los usuarios le han parecido muy útiles o bastante útiles, y, además, al 80 % le ha parecido que están redactadas de manera muy clara o bastante clara. La barra de formato del Módulo II parece que también ha sido de bastante utilidad a los usuarios para maquetar el documento.

El manual de uso, la forma de exportar e importar documentos, y la manera de insertar imágenes tuvieron asimismo una valoración positiva, aunque hay un porcentaje de usuarios que reconoce que no usó el manual (24 %), que no exportó o importó ningún documento (12 % y 20 %, respectivamente), o que no insertó ninguna imagen (24 %).

Uno de los aspectos en donde hay menos consenso y que parece ser de los peor valorados es el uso de las frases sugeridas en el Módulo I, ya que, aunque un 36 % de los usuarios dice haber insertado muchas de las frases y un 28 % ha insertado bastantes, un 16 % indica que no ha insertado ninguna y un 20 % ha insertado pocas. Las repuestas sobre la utilidad del corrector ortográfico también van en esta línea.

En el Apéndice A se recogen los resultados del cuestionario de valoración de arText.

5. Conclusiones y trabajo futuro

Como se ha visto, aunque la comunicación electrónica entre el ciudadano y la Administración es el procedimiento habitual en España en el contexto de la e-Administración, son escasas las iniciativas para desarrollar herramientas TIC que tengan como objetivo mejorar la comunicación escrita entre ambos. Este trabajo busca contribuir en este sentido. Así, el objetivo de este artículo ha sido mostrar el diseño e implementación de arText, una aplicación tecnológica que ayuda a la ciudadanía a escribir textos dirigidos a la Administración pública, y que integra diferentes herramientas y recursos de PLN, como un analizador morfosintáctico y un segmentador discursivo.

La aplicación tiene forma de editor de textos en línea e incluye tres módulos, que ayudan al usuario a: I) estructurar y redactar el documento, II) darle formato y corregirlo ortográficamente, y III) revisar la adecuación del texto, mediante recomendaciones lingüísticas relacionadas con oraciones largas, conectores discursivos, siglas, palabras repetidas, verbos en primera persona del singular y plural, y unidades léxicas subjetivas. El Módulo I es especialmente relevante porque no existía hasta la fecha ninguna aplicación tecnológica que ayudase a la ciudadanía a estructurar géneros textuales del ámbito administrativo, incorporando sus apartados prototípicos, y añadiendo en cada uno de ellos los títulos, contenidos y fraseología habituales. El Módulo III es útil principalmente para revisar la adecuación del texto escrito por la ciudadanía, atendiendo a las características globales que debe cumplir este tipo de documentos, como son la adecuación gramatical y estilística, la precisión, la concisión, la objetividad y la sistematicidad.

Las evaluaciones *data-driven* y *user-driven* realizadas a arText ofrecen resultados positivos. La evaluación *data-driven* se realizó para evaluar el funcionamiento de los algoritmos utilizados en el Módulo III. En este caso, se observa que los resultados obtenidos de precisión y cobertura son en general buenos para todos los rasgos lingüísticos analizados. La evaluación *user-driven* permitió conocer la percepción de la utilidad del sistema por parte de usuarios reales, quienes hicieron una valoración muy favorable, destacando la utilidad y claridad de la aplicación. El aspecto mejor valorado fue el Módulo I, que incluye información sobre la estructura textual, y el segundo fue el Módulo III, que ofrece sugerencias de revisión lingüística del texto.

El aspecto en el que hubo menos consenso en la evaluación fue el uso de las frases sugeridas en el Módulo I. Por tanto, la revisión, modificación y ampliación de estas unidades será una de las líneas prioritarias de trabajo futuro. Asimismo, se refinarán las recomendaciones relacionadas con las siglas en el Módulo III, sobre todo para que logren una mayor cobertura, es decir, para que se detecten un mayor número de casos. También se investigará sobre la posibilidad de incluir nuevas recomendaciones no contempladas por el momento. En cuanto a la evaluación, sería interesante ampliar los textos del corpus, realizar la anotación por parte de diferentes anotadores y medir el acuerdo entre ellos (*interannotator agreement*). Asimismo, se podría realizar una evaluación *user-driven* adicional enfocada a otros colectivos, con distintos perfiles (en cuanto a edad y formación) y capacidades (en cuanto a manejo de internet). Finalmente, se investigará sobre la posibilidad de adaptar arText a otros géneros textuales, ámbitos especializados y lenguas. Actualmente el sistema ya está disponible para la redacción de géneros textuales en español de otros dos ámbitos: medicina y turismo. En el caso de la medicina, permite redactar un artículo de investigación, un artículo de revisión, una historia clínica, un resumen de artículo de investigación y un Trabajo de Fin de Grado (TFG). En el caso del turismo, permite redactar un artículo de divulgación, una entrada de blog de viajero, un informe, una normativa y un plan de negocio.

Agradecimientos

Este trabajo se ha llevado a cabo en el marco de un contrato Ramón y Cajal (RYC-2014-16935) financiado por el Ministerio de Economía, Industria y Competitividad, vinculado al Departamento de Filologías Extranjeras y sus Lingüísticas de la Facultad de Filología de la Universidad Nacional de Educación a Distancia (UNED). Los resultados se derivan de dos proyectos de investigación. Por un lado, del proyecto "Un sistema automático de ayuda a la redacción de textos especializados de ámbitos relevantes en la sociedad española actual", financiado en la Convocatoria 2015 de Ayudas Fundación BBVA a Investigadores y Creadores Culturales. Por otro lado, del proyecto "Tecnologías de la Información y la Comunicación para la e-Administración: hacia la mejora de la comunicación entre Administración y ciudadanía a través del lenguaje claro (TIC-eADMIN)", financiado por el Ministerio de Ciencia, Innovación e Universidades en la convocatoria 2018 de Proyectos I+D del Subprograma Es-

tatal de Generación de Conocimiento (PGC2018-099694-A-I00). Ambos proyectos se han desarrollado en el marco del grupo de investigación ACTUALing de la UNED, en colaboración con el grupo de investigación IULATERM del Institut de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF). Quiero agradecer su participación en la investigación a todos los miembros del equipo de trabajo de ambos proyectos, especialmente a M. Amor Montané y a Luis Hysa.

Referencias

Alcaraz, Enrique & Brian Hughes. 2002. *El español jurídico* Ariel Derecho. Grupo Planeta.

Atserias, Jordi, Bernardino Casas, Elisabet Comelles, Meritxell González, Lluís Padró & Muntsa Padró. 2006. Freeling 1.3. syntactic and semantic service in an open-source NLP library. En *5th International Conference on Language Resources and Evaluation (LREC)*, 48–55.

Ayala, Pilar, Elena Domínguez Ortega, Francisca Martel Ruiz, Jacqueline Montelongo Sanchéz, Dolores Morales Sosa, Orlando J. Socorro Lorenzo & Elena Suárez Manrique de Lara. 2000. Manual de normalización de documentos administrativos. Informe técnico. Universidad de Las Palmas de Gran Canaria.

Bhatia, Vijay Kumar. 1993. *Analyzing genre:language use in professional settings*. Longman.

Biber, Douglas, Ulla Connor & Thomas A. Upton. 2007. *Discourse on the move: Using corpus analysis to describe discourse structure*. John Benjamins. 10.1075/scl.28.

Cabré, M. Teresa, Carme Bach, Iria da Cunha, Albert Morales & Jorge Vivaldi. 2010. Comparación de algunas características lingüísticas del discurso especializado frente al discurso general: el caso del discurso económico. *Modos y formas de la comunicación humana, Ways and Modes of Human Communication* 453–460.

Cabré, Maria Teresa. 1999. *La terminología: Representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. IULA, Universitat Pompeu Fabra.

Castellón, Heraclia. 2001. *El lenguaje administrativo. Formas y uso*. Editorial La Vela.

da Cunha, Eric SanJuan, Iria, Juan M. Torres-Moreno, Marina Lloberes & Irene Castellón. 2012. DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Systems with Applications* 39(2). 1671–1678. 10.1016/j.eswa.2011.06.058.

da Cunha, Iria & Mikel Iruskieta. 2010. Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies* 12(5). 563–598. 10.1177/1461445610371054.

da Cunha, Iria & M. Amor Montané. 2019. Textual genres and writing difficulties in specialized domains. *Revista Signos. Estudios de Lingüística* 52(99). 4–30. 10.4067/S0718-09342019000100004.

da Cunha, Iria & M. Amor Montané. 2020. A corpus-based analysis of textual genres in the administration domain. *Discourse Studies* 22(1). 3–31. 10.1177/1461445619887538.

Duarte, Carles & Anna Martínez. 1995. *El lenguaje jurídico*. AZ Editora.

Gamallo Otero, Pablo, Marcos García, Iria del Río & Isaac González López. 2015. Avalingua: Natural language processing for automatic error detection. En *Learner Corpora in Language Testing and Assessment*, 35–58. John Benjamins. 10.1075/scl.70.02gam.

Giraldo, John J. 2008. *Análisis y descripción de las siglas en el discurso especializado de genoma humano y medio ambiente*: IULA, Universitat Pompeu Fabra. Tesis Doctoral.

González Salgado, J. Antonio. 2009. El lenguaje jurídico del siglo XXI. *THEMIS: Revista de Derecho* 57. 235–245.

Gotti, Maurizio. 2008. *Investigating specialized discourse*. Peter Lang.

Jiménez Yañez, Ricardo-Maria. 2016. *Escribir bien es de justicia*. Pamplona, Navarra: Editorial Aranzadi.

Mann, William & Sandra A. Thompson. 1981. Rhetorical structure theory: Toward a functional theory of text organization. *Text & Talk* 8(3). 243–281. 10.1515/text.1.1988.8.3.243.

de Miguel Aparicio, Elena. 2000. El texto jurídico-administrativo: Análisis de una orden ministerial. *Círculo de Lingüística Aplicada a la Comunicación* 4.

Montolío Durán, Estrella. 2012. *Hacia la modernización del discurso jurídico*. Publicacions i Edicions de la Universitat de Barcelona.

Otaola Olano, Concepción. 1988. La modalidad (con especial referencia a la lengua española). *Revista de Filología Española* 68(1/2). 97–117.

Parodi, Giovanni. 2010. *Academic and professional discourse genres in Spanish*. John Benjamins. 10.1075/scl.40.

Rodríguez-Aguilera, Cesáreo. 1969. *El lenguaje jurídico*. Barcelona Bosch.

Samaniego, Eva. 2005. El lenguaje jurídico: Peculiaridades del español jurídico. En *Lengua y Sociedad: Aportaciones recientes en lingüística cognitiva, lenguas de contacto, lenguajes de especialidad y lingüística del corpus*, 273–310. Universidad de Valladolid, Centro Buendia.

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

Sánchez Alonso, Fernando. 2014. *Lenguaje y estilo administrativo. Redacción de documentos*. Escuela de Formación e Innovación. Administración Pública.

Tofiloski, Milan, Julian Brooke & Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. En *47th Annual Meeting of the Association for Computational Linguistics*, 77–80.

Upton, Thomas A. & Mary A. Cohen. 2009. An approach to corpus-based discourse analysis: The move analysis as example. *Discourse Studies* 11(5). 585–605. 10.1177/1461445609341006.

Van Dijk, Teun A. 1977. *Text and context: Explorations in the semantics and pragmatics of discourse*. Longman.

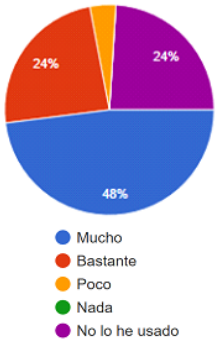
Van Dijk, Teun A. 1989. *La ciencia del texto: un enfoque interdisciplinario*. Paidós Comunicación.

Zhou, Lanjun, Li Binyang, Wei Zhongyu & Wong Kam-Fai. 2014. The CUHK Discourse Treebank for Chinese: Annotating Explicit Discourse Connectives for the Chinese Treebank. *International Conference on Language Resources and Evaluation (LREC)* 942–949.

A. Resultados de las preguntas del cuestionario de valoración de arText

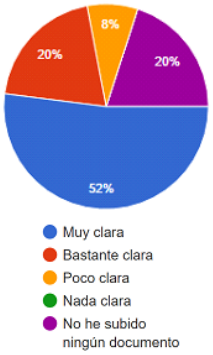
¿Le han parecido claras las opciones ofrecidas por el editor?

¿Le ha parecido útil el manual de uso?

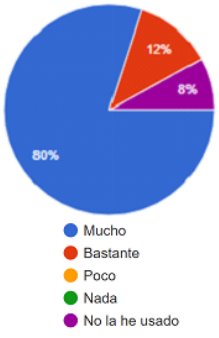


¿Le ha parecido clara la manera de exportar documentos?

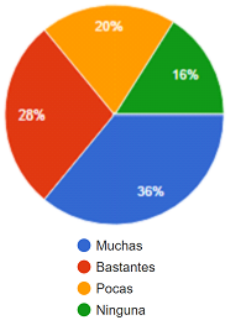
¿Le ha parecido clara la manera de subir al sistema documentos guardados previamente?



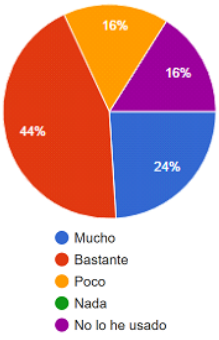
¿Le ha parecido útil la información incluida en la opción "Estructuración y contenidos de arText"?



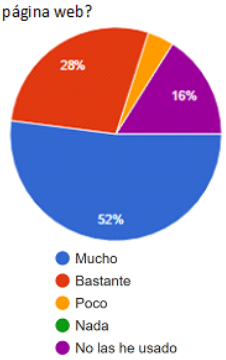
¿Ha insertado en el texto alguna de las frases sugeridas?



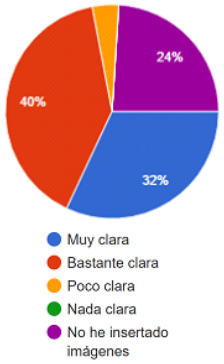
¿Le ha parecido útil el corrector ortográfico, que subraya las palabras desconocidas en rojo?



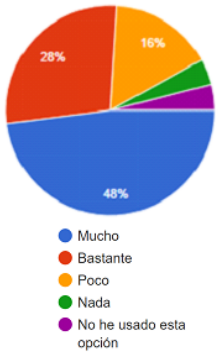
¿Le han parecido útiles las opciones de formato que incluye la barra superior de la página web?



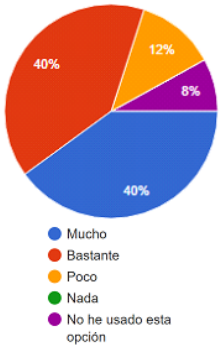
¿Le ha parecido clara la manera de insertar imágenes en el texto?



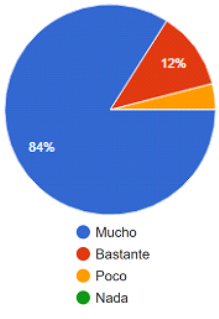
¿Le han parecido útiles las sugerencias ofrecidas al seleccionar la opción "Revisar el texto"?



¿Le ha parecido clara la redacción de las sugerencias ofrecidas?



En general, ¿considera útil este sistema?



Proposta recebida em Janeiro 2017 e aceite para publicação em Maio 2017.

Distância diacrónica automática entre variantes diatópicas do português e do espanhol

Automatic diachronic distance between diatopic variants of Portuguese and Spanish

José Ramom Pichel
imaxin software
jramompichel@imaxin.com

Pablo Gamallo
Universidade de Santiago de Compostela
pablo.gamallo@usc.es

Marco Neves
Universidade Nova de Lisboa
mfneves@fcsh.unl.pt

Iñaki Alegria
Universidade do País Basco (EHU/UPV)
i.alegria@ehu.eus

Resumo

O objetivo deste trabalho é aplicar uma metodologia baseada na perplexidade, para calcular automaticamente a distância interlinguística entre diferentes períodos históricos de variantes diatópicas de idiomas. Esta metodologia aplica-se a um corpus construído *adhoc* em ortografia original, numa base equilibrada de ficção e não-ficção, que mede a distância histórica entre o português europeu e do Brasil, por um lado, e o espanhol europeu e o da Argentina, por outro. Os resultados mostram distâncias muito próximas em ortografia original e transcrita automaticamente, entre as variedades diatópicas do português e do espanhol, com ligeiras convergências/divergências desde meados do século XX até hoje. É de salientar que o método não é supervisionado e pode ser aplicado a outras variedades diatópicas de línguas.

Palavras chave

distância linguística, linguística diacrónica, perplexidade

Abstract

The objective of this work is to apply a perplexity-based methodology to automatically calculate the cross-lingual distance between different historical periods of diatopic language variants. This methodology applies to an *adhoc* constructed corpus in original spelling, on a balanced basis of fiction and non-fiction, which measures the historical distance between European and Brazilian Portuguese on the one hand, and European and Argentinian Spanish on the other. The results show very close distances, both in original spelling and automatically transcribed spelling, between the diatopic varieties of Portuguese and Spanish, with slight convergences/divergences from the middle of the 20th century until today. It should be noted that the method is not supervised and can be applied to other diatopic varieties of languages.

Keywords

language distance, diachronic linguistics, perplexity

1. Introdução

Os idiomas e as suas variedades diatópicas mudam constantemente ao longo da história (Millar & Trask, 2015), pelo que medir esta distância de forma automática é um desafio.

Historicamente houve diferentes abordagens para calcular esta distância, nomeadamente com base nos estudos filogenéticos no âmbito da Linguística Histórica (Petroni & Serva, 2010), da dialectologia (Nerbonne & Heeringa, 1997), do campo da aquisição de segunda língua (Chiswick & Miller, 2004), ou da identificação automática da língua (Malmasi et al., 2016).

Para Gamallo et al. (2016), o conceito de distância linguística está intimamente relacionado com o processo de identificação automática da língua. Na verdade, quanto mais difícil for a identificação das diferenças entre duas línguas ou variedades linguísticas, menos distância existe entre elas.

Com este fim, os melhores sistemas de identificação automática de línguas baseiam-se em modelos de n-gramas de caracteres extraídos de corpora textuais (Malmasi et al., 2016). Os n-gramas de caracteres não só codificam informação léxica e morfológica, mas também características fonológicas, uma vez que os sistemas fonográficos escritos estão relacionados com a forma como as línguas eram pronunciadas no passado.

DOI: 10.21814/lm.12.1.319

Tendo isto em mente, o objectivo principal do presente artigo é aplicar uma metodologia para medir a distância diacrónica entre duas variedades diatópicas do português e duas do espanhol. Para isso utilizaremos modelos de n-gramas de caracteres obtidos a partir de corpus histórico construído *adhoc*, e a métrica chamada *Perplexity Language Distance* (PLD), baseada na perplexidade e definida em Pichel et al. (2019a). A distância automática entre variedades diacrónicas de línguas diferentes será referida de forma abreviada como *CrossDiaDist*.

A nossa metodologia orientada por corpus não é supervisionada e, portanto, só necessitamos de corpora históricos em bruto. Os textos sobre os quais realizamos as experiências de distância linguística preservam a ortografia original; também calculamos essa distância entre esses mesmos textos transliterados para uma ortografia comum às duas línguas. Um trabalho similar de transcrição foi realizado em Simões et al. (2012), com o objectivo de modernizar ortograficamente versões antigas de palavras em português num dicionário.

De agora em diante, usaremos as siglas *OS* para ortografia original e *TS* para ortografia transcrita.

Em resumo, o nosso objetivo é tentar verificar se as duas variedades de línguas têm uma *CrossDiaDist* estável ou se, pelo contrário, têm períodos convergentes e/ou divergentes. Além disso, tentamos também medir até que ponto a ortografia desempenha um papel nesta *CrossDiaDist* entre variedades diatópicas nos períodos históricos estudados.

O artigo está organizado da seguinte forma: em primeiro lugar, descrevemos alguns estudos sobre distância automática entre línguas com diferentes abordagens na Secção 2. Depois, descrevemos o corpus usado e o conceito de perplexidade na Secção 3. Posteriormente, na Secção 4, apresentamos a metodologia, baseada na perplexidade, a aplicar ao corpus diacrónico. Por fim, na Secção 5, apresentamos e discutimos os resultados, comentando as conclusões e o trabalho futuro na Secção 6.

2. Trabalho relacionado

Para medir a proximidade ou distanciamento entre línguas ou variedades diatópicas de línguas, existem diferentes abordagens: identificação automática de línguas, filogenética e cálculo da distância automática entre línguas.

2.1. Identificação automática de línguas

A identificação automática de línguas é um campo da linguística computacional ainda com desafios por resolver, tais como a diferenciação automática de línguas muito próximas (por exemplo, checo e eslovaco, croata e bósnio) ou variedades diatópicas na mesma língua (por exemplo, espanhol argentino e espanhol europeu, português de Angola e português de Portugal).

Para esta identificação de línguas têm sido usadas diferentes abordagens: dicionários baseados em listas de palavras e heurísticas (ortografia, morfologia, características sintácticas) ou abordagens estatísticas baseadas em modelos de língua (nomeadamente, n-gramas de caracteres ou n-gramas de palavras) a partir de corpora.

Estes últimos, especialmente os baseados em n-gramas de caracteres, costumam ser os melhores sistemas de identificação linguística (Malmasi et al., 2016). A razão provável é que os n-gramas de caracteres não só codificam informações lexicais e morfológicas, mas também características fonológicas, uma vez que os sistemas fonográficos escritos estão relacionados com a forma como as línguas eram pronunciadas no passado. Se os n-gramas forem longos (por exemplo, $\geq$ 6–gramas), também codificam relações sintácticas, pois podem representar o fim de uma palavra e o início da próxima numa sequência. Também podemos destacar, no que toca à identificação eficiente de idiomas próximos, o trabalho de Tiedemann & Ljubešić (2012) baseado em n-gramas de palavras utilizando *blacklists*.

Entre os estudos mais relevantes e pioneiros devemos destacar os de Cavnar & Trenkle (1994) e Dunning (1994), que são os primeiros a usar n-gramas para identificação automática de línguas.

Também existem trabalhos para classificar línguas próximas ou variedades diatópicas (Malmasi et al., 2016; Zampieri et al., 2018; Kroon et al., 2018), e também para a detecção de línguas em textos curtos e com muito ruído como tweets (Gamallo et al., 2014; Zubiaga et al., 2016).

Finalmente, existem abordagens relacionadas com a aprendizagem profunda (*deep learning* (Lopez-Moreno et al., 2014; Gonzalez-Dominguez et al., 2014). Na *Evaluation Campaign* mais recente organizada no Workshop on Natural Language Processing for Similar Languages, Varieties and Dialects (VarDial-2019), confirma-se que as abordagens mais sofisticadas baseadas em aprendizagem profunda e vectores contextuais não melhoram os resultados das estratégias mais tradicionais com modelos de n-gramas de caracteres e classificadores de tipo *Naive Bayes* ou *Support Vector Machine* (Zampieri et al., 2019).

2.2. Filogenética

Na filogenética, para calcular a distância ou proximidade entre línguas, a estratégia consiste em classificar as línguas através da construção de uma árvore enraizada que descreve a história evolutiva de um conjunto de línguas ou variedades relacionadas.

Para isso existem diferentes metodologias, como as baseadas em comparar cognatos lexicais, ou seja, palavras que têm uma origem histórica comum (Nakhleh et al., 2005; Holman et al., 2008; Bakker et al., 2009; Petroni & Serva, 2010; Barbançon et al., 2013). Também existem aproximações lexico-estatísticas baseadas em listas de palavras em vários idiomas, por exemplo, Swadesh list (Swadesh, 1952) ou a base de dados ASJP (Brown et al., 2009), que medem automaticamente distâncias usando a percentagem de cognatos compartilhados. Também a distância Levenshtein entre as palavras numa lista cross-lingual (Yujian & Bo, 2007) é uma das métricas mais comuns usadas neste campo (Petroni & Serva, 2010). Finalmente, também usando uma distância baseada na perplexidade, Gamallo et al. (2017a) construíram uma rede que representa o mapa actual de semelhanças e divergências entre as principais línguas da Europa.

2.3. Distância entre idiomas

Inicialmente houve abordagens como as de Nerbonne & Heeringa (1997) e Kondrak (2005) a partir da comparação entre formas fonéticas de idiomas, "mas alguns pesquisadores têm argumentado contra a possibilidade de obter resultados significativos a partir da comparação entre formas fonéticas de idiomas" (Singh & Surana, 2007).

Em tempos recentes o cálculo da distâncias entre línguas baseiam-se sobretudo em modelos de língua construídos a partir de corpora paralelos. Estes modelos são construídos a partir das co-ocorrências de palavras e, portanto, a distância entre línguas é resultado da similaridade interlinguística entre estas co-ocorrências (Liu & Cong, 2013; Gao et al., 2014; Asgari & Mofrad, 2016).

Também existem outras aproximações baseadas na entropia para investigar a mudança diacrónica no inglês científico, como em (Degaetano-Ortlieb et al., 2016) (Rama & Borin, 2015), utilizando a cross-entropy. Finalmente esta distância tem sido calculada utilizando a perplexidade em corpus sincrónicos Gamallo et al. (2017a) e diacrónicos Pichel et al. (2018).

3. Materiais e ferramentas

3.1. Corpora

Para a elaboração das nossas experiências, criámos um corpus diacrónico em OS para o português europeu, português do Brasil, espanhol europeu e espanhol da Argentina.

No que toca ao tamanho deste corpus, seguimos os critérios dos autores do Helsinki Corpus of Historical English (Rissanen et al., 1993), que indicam: "O primeiro problema a ser decidido na compilação de um corpus é o seu tamanho" e "O tamanho do corpus básico é de cerca de 1,5 milhões de palavras".

Em relação aos períodos, como só queremos estudar a distância entre as variantes diatópicas do português e do espanhol em períodos recentes, vamos dividir o nosso corpus exclusivamente em dois períodos históricos: segunda metade do século XX (XX-2) e século XXI até ao presente (XXI-1). Também para tornar este corpus representativo de todas as variantes diatópicas de português e espanhol, tendo em conta a representatividade definida por Biber (1993), incluímos 50% de ficção e 50% de não-ficção para cada período. Além disso, como queremos ver o papel que a ortografia desempenha na distância entre as variedades diatópicas, incluímos sempre textos em OS.

Tendo em conta todas estas características, alargámos o corpus histórico *Carvalho* em OS já desenvolvido em Pichel et al. (2019b) para o português europeu (Carvalho-PT-PT) e espanhol europeu (Carvalho-ES-ES), com o português do Brasil (Carvalho-PT-BR) e o espanhol da Argentina (Carvalho-ES-AR). Temos portanto o português europeu, português do Brasil, espanhol europeu e espanhol da Argentina para os períodos XX-2 e XXI-1. Além disso, os textos incluídos neste corpus estão na ortografia mais próxima possível do original, uma vez que as experiências que iremos realizar serão desenvolvidas tanto em OS como em TS automático. Criado para estas experiências, Carvalho[1] é um corpus histórico em OS disponível gratuitamente para inglês, português europeu, português do Brasil, espanhol europeu e espanhol da Argentina.

Finalmente, Carvalho-PT-PT, Carvalho-PT-BR, Carvalho-ES-ES, Carvalho-ES-AR foram divididos em dois subcorpora (treino e teste) para calcular a distância entre variedades diatópicas baseadas na perplexidade. A Tabela 1 mostra o tamanho dos corpora de treino e de teste nos dois

[1] https://github.com/gamallo/Perplexity/tree/master/resources/Carvalho

Carvalho	Train-pt	Test-pt	Train-br	Test-br	Train-es	Test-es	Train-arg	Test-arg
XX-2	1.688M	363K	1.261M	342K	1.231M	250K	1.280M	256K
XXI-1	1.389M	336K	1.222M	315K	1.270M	285K	1.202M	285K

Tabela 1: Tamanho dos corpora de treino e teste em dois períodos históricos de espanhol-Espanha (es), espanhol-Argentina (arg), português-Portugal (pt) e português-Brasil (br)

períodos de cada variante diatópica de português e espanhol para os períodos XX-2 e XXI-1.

A próxima secção descreve as características do corpus diacrónico de Carvalho para cada uma das variedades diatópicas das línguas. Vamos concentrar-nos nos diferentes repositórios de onde foram extraídos todos os documentos e nas características significativas de cada língua.

3.1.1. Corpus do Português Europeu e do Brasil

Para a elaboração dos corpora Carvalho-PT-PT e Carvalho-PT-BR, seleccionámos textos com a ortografia o mais próxima possível do original (OS). Há que ter em conta que nessa OS estão incluídos textos com e sem o Acordo Ortográfico de 1990 (AO'90). As diferentes versões do português (português europeu, português do Brasil, português europeu AO'90 e português do Brasil AO'90) podem ser vistas na Tabela 2.

O português europeu e o português do Brasil têm variado especialmente no século XX do ponto de vista do padrão e da ortografia. Assim, desde o ano 1779 em Portugal, a Academia das Ciências de Lisboa tem promovido diferentes padrões e normas ortográficas (e.g.: 1885, 1911, 1945, 1973, 1990). Por sua vez, a Academia Brasileira de Letras tem convergido ou divergido com estas propostas (e.g.: 1907, 1915, 1919, 1924, 1929, 1931, 1943, 1971, 1986) até ao Acordo Ortográfico de 1990 (AO'90), que ainda hoje é objeto de grande controvérsia em ambos os países e não está totalmente espalhado.

Para criar os corpora de português Carvalho-PT-PT e Carvalho-PT-BR nos subperíodos XX-2 e XXI-1, identificámos e seleccionámos documentos dos seguintes repositórios: Wiki source[2], OpenLibrary[3], Linguateca[4], *Domínio Público*[5] e *TesesUSP*[6]

3.1.2. Corpus do Espanhol Europeu e da Argentina

No caso do espanhol, as mudanças relevantes na ortografia ocorreram especialmente desde o aparecimento em 1713 da Real Academia Espanhola e mais tarde em 1741, com um padrão ortográfico diferente do resto das línguas românicas. Esta norma foi consolidada ao longo do tempo com pequenas variações na história, embora houvesse gramáticas na Argentina com orientações divergentes em relação ao espanhol europeu, como em Bello (1984) e Bello et al. (1951). Durante o século XX, a ortografia em espanhol europeu e argentino mudou muito pouco (1952, 1959 e 1999), mas houve contribuições para a gramática da Academia Argentina de las Letras fundada em 1931.

Na Tabela 3, mostram-se trechos do espanhol europeu e espanhol argentino. Para a realização dos corpora Carvalho-ES-ES e Carvalho-ES-AR, obtivemos documentos de ficção e não-ficção nos seguintes repositórios: OpenLibrary[7], Wiki source[8], *Repositorio Institucional CONICET Digital*[9] e *TesesUniversidadBuenosAires*[10]

3.2. Perplexidade

Para medir a qualidade dos modelos de linguagem construídos com n-gramas extraídos a partir de corpora (Chen & Goodman, 1996; Sennrich, 2012; Dieguez-Tirado et al., 2005) utilizamos a perplexidade:

$$PP(CH, LM) = \sqrt[n]{\prod_{i}^{n} \frac{1}{P(ch_i|ch_1^{i-1})}} \quad (1)$$

onde as probabilidades de n-grama $P(\cdot)$ são definidas desta forma:

$$P(ch_n|ch_1^{n-1}) = \frac{C(ch_1^{n-1}ch_n)}{C(ch_1^{n-1})} \quad (2)$$

[2] https://en.wikisource.org/wiki/Category:Portuguese_authors
[3] https://openlibrary.org/
[4] https://www.linguateca.pt/
[5] http://www.dominiopublico.gov.br/
[6] https://www.teses.usp.br
[7] https://openlibrary.org/
[8] https://en.wikisource.org/wiki/Category:Spanish_authors
[9] https://ri.conicet.gov.ar/
[10] http://repositoriouba.sisbi.uba.ar/

Portugal (OS)	**Brasil(OS)**	**PT(AO'90) (OS)**	**BR(AO'90) (OS)**
o princípio da *acção* ou, também, a função essencial da vida animal. (. . .)	Ele existe — mas quase só por intermédio da *ação* das pessoas: de bons e maus. (. . .)	em primeiro lugar, porque tais deuses de *facto* não existem, (. . .)	Só o mau *fato* de se topar com eles, dava soloturno sombrio. (. . .)

Tabela 2: *Diferenças* entre variedades diatópicas do português europeu (OS), português do Brasil (OS), e ambos com Acordo Ortográfico (AO'90). Este extratos pertencem a documentos dos corpora Carvalho-PT-PT e Carvalho-PT-BR

Espanhol europeu (OS)	**Espanhol da Argentina (OS)**
-¿*Sabes* lo que te digo? -¡Qué! -Que si *tú* fueses el novio de mi hermana, te hubiera matado. (. . .)	Pero es que *vos* ya lo *sabés*, decía la Maga, resentida. (. . .)

Tabela 3: *Diferenças* entre variedades diatópicas do espanhol europeu (OS) e o espanhol da Argentina (OS) em documentos do corpus Carvalho-ES-ES e Carvalho-ES-AR

Esta métrica está orientada para conferir se um modelo de língua é bom a prever uma amostra de texto. Assim, se a perplexidade é baixa, o modelo de língua é bom a prever a amostra. Pelo contrário, uma perplexidade alta mostra que o modelo de linguagem não é bom a prever a amostra em questão.

A perplexidade tem sido usada também em tarefas muito específicas, tais como medir a dificuldade das tarefas de reconhecimento da fala (Jelinek et al., 1977), para classificar tweets formais e coloquiais (González, 2015), ou para identificar automaticamente línguas estreitamente relacionadas e até variedades diatópicas de línguas (Gamallo et al., 2016).

Tendo em conta isto, definimos recentemente em Pichel et al. (2019b) uma distância baseada na perplexidade chamada Perplexity Language Distance (PLD), para medir a distância diacrónica intralinguística em línguas como o inglês, português e espanhol. A PLD também foi aplicada para medir a *CrossDiaDist* entre duas línguas (Pichel et al., 2019a).

No nosso caso a *CrossDiaDist* será entre duas variedades diatópicas da mesma língua. Esta é definida comparando os n-gramas de um texto numa variedade da língua (português europeu) com o modelo de n-gramas treinado para a outra variedade de língua (português do Brasil). Esta comparação deve ser feita nas duas direcções, dado que PP é uma divergência com valores assimétricos. Além disso esta comparação ao ser diacrónica é por cada período histórico.

Finalmente, para tornar a medida simétrica, a perplexidade do texto do teste CH na variedade diatópica $VL1.2$, dado o modelo da linguagem LM da variedade diatópica $VL1.1$, bem como a perplexidade do texto do teste em $VL1.1$, dado o modelo da linguagem LM de $VL1.2$, são utilizadas para definir *CrossDiaDist* baseada na perplexidade, PLD, entre $VL1.1$ e $VL1.2$, da seguinte forma:

$$PLD(VL1.1, VL1.2) = \frac{PP(A) + PP(B)}{2} \quad (3)$$

$$PP(A) = PP(CH_{VL1.2}, LM_{VL1.1}) \quad (4)$$

$$PP(B) = PP(CH_{VL1.1}, LM_{VL1.2}) \quad (5)$$

No trabalho actual, o nosso objectivo é aplicar a PLD para medir a *CrossDiaDist* entre variedades diatópicas de línguas nos mesmos períodos históricos. Com este fim, utilizámos modelos de linguagem baseados em 7-gramas de caracteres, que incorporam uma técnica de alisamento baseada em interpolação linear. Os corpora de treino/teste contêm aproximadamente 1,25M/250K palavras, respectivamente, para que os nossos resultados possam ser comparados e comentados mais tarde na Secção 5.

Finalmente, para que se possa medir a PLD entre períodos de qualquer outro idioma, outros pares de idiomas ou outros pares de variedades diatópicas de idiomas, desenvolvemos uma arquitetura de pipeline em Perl, disponível em GitHub[11].

4. Métodos e procedimento

O nosso método para calcular a *CrossDiaDist* entre variantes diatópicas de línguas está dividido nas seguintes tarefas sequenciais:

[11] https://github.com/gamallo/Perplexity

1. Definir períodos históricos comuns para todas as línguas ou variedades diatópicas das línguas. No nosso caso teremos dois períodos (XX-2 e XXI-1) para as seguintes línguas: português europeu, português do Brasil, espanhol europeu e espanhol do Brasil.
2. Obter textos suficientes para todas as variedades diatópicas dos idiomas nos períodos históricos previamente definidos. Antes de incorporá-los no corpus é importante verificar se estão em OS. Para isso, temos de olhar para a história das mudanças ortográficas de cada variedade diatópica. Os excertos em qualquer outra língua são eliminados.
3. Dividir o corpus anterior em treino e teste para cada um dos períodos históricos. A tipologia dos textos deve estar equilibrada em 50% aproximadamente entre ficção e não-ficção. O treino contém pelo menos 1,25M palavras por período, enquanto o teste tem pelo menos 20% do tamanho da partição do treino, ou seja, entre 250K e 350K palavras.
4. Realização da *CrossDiaDist* em OS, que será calculada entre cada variedade diatópica de idioma (PLD(VL1.1, VL1.2), PLD(VL2.1, VL2.2)), e para cada período.
5. Realização da *CrossDiaDist* em TS. A TS é o resultado da aplicação de uma normalização ortográfica nos textos com a finalidade de unificar ortograficamente os textos das variedades do português europeu e do Brasil, e também da variedade do espanhol europeu e do da Argentina. Uma vez unificados ortograficamente, é calculada a *CrossDiaDist*, mas em TS. Para isso, foi implementado um transcritor cujo alfabeto consiste em 34 símbolos, representando 10 vogais (incluindo acentos) e 24 consoantes, destinados a cobrir a maioria dos sons mais comuns, incluindo várias palatizações. A codificação é, portanto, próxima da fonológica e, assim, permite simplificar e homogeneizar os casos em que sons semelhantes (geralmente palatalizações) são transcritos de forma diferente em diferentes idiomas. Como as grafias do português europeu e do português do Brasil são muito próximas, a normalização da TS só afecta especialmente a diferenças nas acentuações gráficas. Por exemplo, "acadêmico" no português do Brasil e "académico" no português europeu, ou "assembléia" no português do Brasil e "assembleia" no português europeu são unificados em TS como "academico" e "assembleia". O mesmo acontece com o espanhol europeu e espanhol da Argentina, embora sem diferenças ortográficas salientáveis.
6. Finalmente, avaliação dos resultados finais da *CrossDiaDist* em OS e TS.

5. Avaliação

Após aplicar a metodologia para o cálculo da *CrossDiaDist* baseado em PLD em OS e TS, sobre os corpora Carvalho-PT-PT (português europeu), Carvalho-PT-BR (português do Brasil), Carvalho-ES-ES (espanhol europeu) e Carvalho-ES-AR (espanhol da Argentina), e sobre os dois períodos históricos XX-2 e XXI-1, obtemos os resultados que serão explicados a seguir.

5.1. Resultados

A Tabela 4 mostra os resultados da aplicação da metodologia para os corpora de português europeu e português do Brasil nos dois períodos XX-2 e XXI-1 tanto em OS como em TS. Nela vemos que a distância aumenta ligeiramente desde o período XX-2 até a actualidade, entre o português europeu e o português do Brasil, tanto em OS com em TS. Em OS aumenta de PLD: 4,12 para PLD: 4,36 e em TS aumenta de PLD: 3,65 para 3,83.

A Tabela 5 mostra os resultados para o espanhol espanhol europeu e o espanhol da Argentina em OS e TS. Para as variedades diatópicas do espanhol, vemos que a distância diminui ligeiramente entre espanhol de Espanha e espanhol da Argentina entre os períodos XX-2 e XXI-1 em OS e também em TS. Assim, em OS diminui a PLD: 4,27 para PLD: 4,04 e em TS diminui de PLD: 3,60 para 3,45.

Finalmente as Figuras 1 e 2 retratam a informação da distância entre as variedades diatópicas do português europeu e do português do Brasil, e do espanhol europeu e o espanhol da Argentina.

5.2. Discussão

Em primeiro lugar, observamos que a *CrossDiaDist* entre as variedades diatópicas do português e do espanhol são muito semelhantes em OS e TS sendo a PLD inferior a 5. Assim a distância mais pequena é de 3.45, entre espanhol de Espanha e

PLD(PT/BR)	PLD (OS)	PLD (TS)
XX-2	4.12	3.65
XXI-1	4.36	3.83

Tabela 4: Distância diacrónica (PLD) entre o português europeu e o português do Brasil nos períodos XX-2 e XXI-1 em OS e TS.

PLD(ES/AR)	PLD (OS)	PLD (TS)
XX-2	4.27	3.60
XXI-1	4.04	3.45

Tabela 5: Distância diacrónica (PLD) entre o espanhol europeu e o espanhol da Argentina nos períodos XX-2 e XXI-1 em OS e TS.

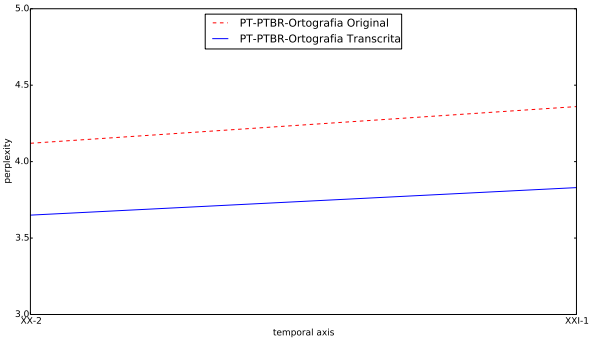


Figura 1: *CrossDiaDist* entre o português europeu e o português da Brasil através do eixo temporal em OS e TS.

espanhol da Argentina em TS, e a máxima é de 4.36 entre o português europeu e português do Brasil em OS. Segundo os resultados reportados em Gamallo et al. (2016), línguas muito próximas como bósnio e croata têm em TS uma distância muito superior, com PLD: 5,90.

Para o caso do português europeu e do português do Brasil, observamos um ligeiro distanciamento no século XXI. Por um lado, talvez este distanciamento se fique a dever a Portugal e o Brasil funcionarem como sistemas culturais diferenciados. O AO'90 foi apresentado como factor de aproximação mas, no entanto, tem tido uma implementação lenta e com muitas resistências, o que talvez seja sintoma das barreiras culturais entre os dois países. De qualquer forma, os valores que apresentamos em TS mostram que a ortografia é um fator pouco relevante no que toca à distância entre o português de Portugal e o português do Brasil. Por outro lado, os valores relativos ao espanhol mostram que é possível registar uma aproximação entre variantes nacionais da mesma língua.

Pelo contrário, no caso do espanhol europeu e do espanhol argentino, vemos que existe uma ligeira aproximação no mesmo período (XXI-1), talvez devido aos esforços de coordenação entre as diferentes academias de língua espanhola e à existência de mais troca de materiais entre os sistemas culturais de Espanha e Argentina.

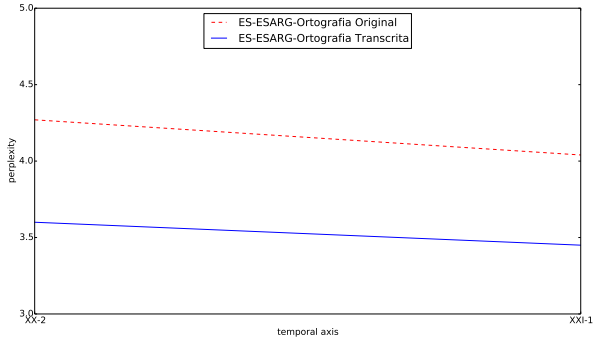


Figura 2: *CrossDiaDist* entre o espanhol europeu e o espanhol da Argentina através do eixo temporal em OS e TS.

Finalmente, observamos que a ortografia entre as duas variantes diatópicas de português e espanhol não desempenha um papel importante nesta distância, pois quando calculamos a *PLD* em TS, ela diminui ligeiramente, mantendo a mesma tendência que em OS.

6. Conclusões e trabalhos futuros

Compilaremos agora as principais conclusões das nossas experiências a partir da aplicação da metodologia de cálculo da distância diacrónica *CrossDiaDist* a variantes diatópicas do português e do espanhol. Também detalharemos na Secção 6.2 próximas investigações em relação à distância automática entre idiomas.

6.1. Conclusões

O cálculo da distância entre idiomas ou variantes diatópicas baseado na perplexidade (*PLD*) identifica automaticamente idiomas e variantes diatópicas de idiomas (Gamallo et al., 2017b), mede a distância síncrona entre idiomas (Gamallo et al., 2017a), a distância diacrónica intralinguística em varias línguas Pichel et al. (2018), a *CrossDiaDist* entre línguas (Pichel et al., 2019a) e agora a *CrossDiaDist* entre variantes diatópicas.

Observamos que esta distância entre as variedades diatópicas de português e espanhol é inferior à distância entre línguas muito próximas. Além disso, vemos que o português europeu e o português do Brasil estão a distanciar-se ligeiramente no século XXI. Pelo contrário, o espanhol europeu e o espanhol da Argentina estão a aproximar-se.

Finalmente, a ortografia nestas variantes diatópicas do português e do espanhol não desempenha um papel relevante, pois estas variantes são escritas com ortografias muito próximas ou indistinguíveis.

6.2. Trabalhos futuros

Queremos alargar esta metodologia ao cálculo de distância entre três línguas. Aplicaremos esta metodologia a três línguas muito próximas, como é o caso do galego em relação ao português e ao castelhano.

Outro objectivo é construir um corpus de redes sociais (p.e.: twitter) e comentários em plataformas digitais (p.e: Tripadvisor, AirBnB, Booking, etc.), para variedades diatópicas de português e espanhol, e observar a distância linguística com um corpus de textos mais afastados da gramática padrão e mais próximo das falas populares.

Finalmente, gostaríamos de investigar a relação entre a distância do idioma usando *PLD* e a estimativa da qualidade da tradução automática (Specia et al., 2018; Han et al., 2013).

Agradecimentos

Estamos muito gratos aos professores Dr. Carlos Quiroga e Dr. José António Souto Cabo da Universidade de Santiago de Compostela, Dr. Fernando Venâncio da Universidade de Amesterdão pelas suas observações sobre a história do português europeu e do Brasil, para além da ajuda na escolha de textos de Portugal e do Brasil. Também à professora Maria Isabel Fernández Domínguez pelo seu conhecimento sobre a história do espanhol europeu e ao Dr. Ernesto Vázquez Souza no que toca à história do espanhol da Argentina. Também a ambos, pela ajuda na escolha de textos de referência de ambas as variedades diatópicas. Finalmente, ao Dr. Marcos Garcia da Universidade da Corunha pelos seus conselhos durante as experiências.

Referências

Asgari, Ehsaneddin & Mohammad R. K. Mofrad. 2016. Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (WELD) as a quantitative measure of language distance. Em *Workshop on Multilingual and Cross-lingual Methods in NLP*, 65–74. 10.18653/v1/W16-1208.

Bakker, Dik, Andre Muller, Viveka Velupillai, Soren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant & Eric W. Holman. 2009. Adding typology to lexicostatistics: A combined approach to language classification. *Linguistic Typology* 13(1). 169–181. 10.1515/LITY.2009.009.

Barbançon, François, Steven N. Evans, Luay Nakhleh, Don Ringe & Tandy Warnow. 2013. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica* 30(2). 143–170. 10.1075/dia.30.2.01bar.

Bello, Andrés. 1984. *Gramática de la lengua castellana*. EDAF.

Bello, Andrés et al. 1951. *Gramática: gramática de la lengua castellana destinada al uso de los americanos*. Caracas: Ministerio de Educación.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and linguistic Computing* 8(4). 243–257. 10.1093/llc/8.4.243.

Brown, Cecil H., Eric W. Holman, Søren Wichmann & Viveka Velupilla. 2009. Automated classification of the world's languages: a description of the method and preliminary results. *Language Typology and Universals* 61(4). 285–308. 10.1524/stuf.2008.0026.

Cavnar, William B & John M Trenkle. 1994. N-gram-based text categorization. Em *3rd annual symposium on document analysis and information retrieval*, 161–175.

Chen, Stanley F. & Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. Em *34th Annual Meeting on Association for Computational Linguistics*, 310–318. 10.3115/981863.981904.

Chiswick, Barry R. & Paul W. Miller. 2004. *Linguistic distance: A quantitative measure of the distance between english and other languages*. Bonn: IZA Discussion Papers.

Degaetano-Ortlieb, Stefania, Hannah Kermes, Ashraf Khamis & Elke Teich. 2016. An information-theoretic approach to modeling diachronic change in scientific english. Em *From Data to Evidence in English Language Research*, 258–281. Brill. 10.1163/9789004390652_012.

Dieguez-Tirado, Javier, Carmen Garcia-Mateo, Laura Docio-Fernandez & Antonio Cardenal-Lopez. 2005. Adaptation strategies for the acoustic and language models in bilingual speech transcription. Em *IEEE International Conference on Acoustics, Speech, and Signal Processing*, I/833–I/836. 10.1109/ICASSP.2005.1415243.

Dunning, Ted. 1994. *Statistical identification of language*. Computing Research Laboratory, New Mexico State University.

Gamallo, Pablo, Inaki Alegria, José Ramom Pichel & Manex Agirrezabal. 2016. Comparing two basic methods for discriminating between similar languages and varieties. Em *3rd Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 170–177.

Gamallo, Pablo, Marcos Garcia, Susana Sotelo & José Ramom Pichel. 2014. Comparing ranking-based and naive bayes approaches to language detection on tweets. Em *Workshop TweetLID: Twitter Language Identification Workshop at SEPLN 2014*, 12–16.

Gamallo, Pablo, José Ramom Pichel & Iñaki Alegria. 2017a. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications* 484. 152–162. 10.1016/j.physa.2017.05.011.

Gamallo, Pablo, Jose Ramom Pichel, Santiago de Compostela & Inaki Alegria. 2017b. A perplexity-based method for similar languages discrimination. *4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)* 109–114. 10.18653/v1/W17-1213.

Gao, Yuyang, Wei Liang, Yuming Shi & Qiuling Huang. 2014. Comparison of directed and weighted co-occurrence networks of six languages. *Physica A: Statistical Mechanics and its Applications* 393. 579–589. 10.1016/j.physa.2013.08.075.

González, Meritxell. 2015. An analysis of Twitter corpora and the differences between formal and colloquial tweets. Em *Tweet Translation Workshop 2015*, 1–7.

Gonzalez-Dominguez, Javier, Ignacio Lopez-Moreno, Haşim Sak, Joaquin Gonzalez-Rodriguez & Pedro J Moreno. 2014. Automatic language identification using long short-term memory recurrent neural networks. Em *15th Annual Conference of the International Speech Communication Association*, .

Han, Aaron Li-Feng, Yi Lu, Derek F Wong, Lidia S Chao, Liangye He & Junwen Xing. 2013. Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling. Em *8th Workshop on Statistical Machine Translation*, 365–372.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Muller & Dik Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica* 42(3–4). 331–354. 10.1515/FLIN.2008.331.

Jelinek, Fred, Robert L Mercer, Lalit R Bahl & James K Baker. 1977. Perplexity: a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* 62(S1). S63. 10.1121/1.2016299.

Kondrak, Grzegorz. 2005. N-gram similarity and distance. Em *International Symposium on String Processing and Information Retrieval (SPIRE)*, 115–126. 10.1007/11575832_13.

Kroon, Martin, Masha Medvedeva & Barbara Plank. 2018. When simple n-gram models outperform syntactic approaches: Discriminating between Dutch and Flemish. Em *5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 244–253.

Liu, HaiTao & Jin Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin* 58. 1139–1144. 10.1007/s11434-013-5711-8.

Lopez-Moreno, Ignacio, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez & Pedro Moreno. 2014. Automatic language identification using deep neural networks. Em *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5337–5341. 10.1109/ICASSP.2014.6854622.

Malmasi, Shervin, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali & Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL Shared Task. Em *3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (VarDial)*, 1–14.

Millar, Robert McColl & Larry Trask. 2015. *Trask's historical linguistics*. Abington, UK: Routledge.

Nakhleh, Luay, Donald A Ringe & Tandy Warnow. 2005. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language* 81(2). 382–420.

Nerbonne, John & Wilbert Heeringa. 1997. Measuring dialect distance phonetically. Em *3rd Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, 11–18.

Petroni, Filippo & Maurizio Serva. 2010. Measures of lexical distance between languages. *Physica A: Statistical Mechanics and its Applications* 389(11). 2280–2283. 10.1016/j.physa.2010.02.004.

Pichel, José Ramom, Pablo Gamallo & Iñaki Alegria. 2018. Measuring language distance among historical varieties using perplexity. application to european portuguese. Em *5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 145–155.

Pichel, José Ramom, Pablo Gamallo & Iñaki Alegria. 2019a. Cross-lingual diachronic distance: Application to portuguese and spanish. *Procesamiento del Lenguaje Natural* 63. 77–84.

Pichel, José Ramom, Pablo Gamallo & Iñaki Alegria. 2019b. Measuring diachronic language distance using perplexity: Application to english, portuguese, and spanish. *Natural Language Engineering* 1–22. 10.1017/S1351324919000378.

Rama, Taraka & Lars Borin. 2015. Comparative evaluation of string similarity measures for automatic language classification. Em *Sequences in Language and Text*, 171–200. De Gruyter Mouton. 10.1515/9783110362879-012.

Rissanen, Matti, Merja Kytö & Minna Palander-Collin. 1993. *Early english in the computer age: Explorations through the helsinki corpus*. Berlin: De Gruyter Mouton.

Sennrich, Rico. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. Em *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 539–549.

Simões, Alberto, Álvaro Iriarte Sanromán & José João Almeida. 2012. Dicionário-aberto: A source of resources for the portuguese language processing. Em *International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 121–127. 10.1007/978-3-642-28885-2_14.

Singh, Anil Kumar & Harshit Surana. 2007. Can corpus based measures be used for comparative study of languages? Em *9th meeting of the ACL special interest group in computational morphology and phonology*, 40–47.

Specia, Lucia, Carolina Scarton & Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies* 11(1). 1–162. 10.2200/S00854ED1V01Y201805HLT039.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to north american indians and eskimos. *American Philosophical Society* 96(4). 452–463.

Tiedemann, Jörg & Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. Em *International Conference on Computational Linguistics (COLING)*, 2619–2634.

Yujian, Li & Liu Bo. 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence* 29(6). 1091–1095. 10.1109/TPAMI.2007.1078.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann et al. 2018. Language identification and morphosyntactic tagging: The second vardial evaluation campaign. Em *5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 1–17.

Zampieri, Marcos, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru & Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. Em *6th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 1–16. 10.18653/v1/W19-1401.

Zubiaga, Arkaitz, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza & Víctor Fresno. 2016. TweetLID: a benchmark for tweet language identification. *Language Resources and Evaluation* 50. 729–766. 10.1007/s10579-015-9317-4.

http://www.linguamatica.com/

linguaMATICA

Artigos de Investigação