Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 11, Número 2 (2019)

ISSN: 1647-0818

linguaMATICA

Volume 11, Número 2 – 2019

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

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

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 octubre de 2018 y aceptada para publicación en noviembre de 2019.

Estrategia multidimensional para la selección de candidatos de traducción automática para posedición

Multidimensional strategy for the selection of machine translation candidates for post-editing

Ona de Gibert
Universidad del País Vasco UPV/EHU
ona.degibert@ehu.eus

Nora Aranberri
Grupo IXA, Universidad del País Vasco UPV/EHU
nora.aranberri@ehu.eus

Resumen

Una integración eficiente de un sistema de traducción automática (TA) en un flujo de traducción conlleva la necesidad de distinguir entre oraciones que se benefician de la TA y las que no antes de que pasen a manos del traductor. En este trabajo, cuestionamos el uso por separado de las dimensiones de esfuerzo de posedición de Krings (2001) para clasificar oraciones en aptas para traducir o poseditar al entrenar modelos de predicción y abogamos por una estrategia multidimensional. A partir de una tarea de posedición en un escenario real, se recogen mediciones de los tres parámetros de esfuerzo, a saber, tiempo, tasa de palabras poseditadas, y percepción del esfuerzo, como representativos de las tres dimensiones (temporal, técnica y cognitiva). Los resultados muestran que, a pesar de que existen correlaciones entre las mediciones, los parámetros difieren en la clasificación de un número elevado de oraciones. Concluimos que la estrategia multidimensional es necesaria para estimar el esfuerzo real de posedición.

Palabras clave

traducción automática, esfuerzo de posedición, estimación de calidad

Abstract

An efficient integration of a machine translation (MT) system within a translation flow entails the need to distinguish between sentences that benefit from MT and those that do not before they are presented to the translator. In this work we question the use of **?** post-editing effort dimensions separately to classify sentences into suitable for translation or for post-editing when training predictions models and propose a multidimensional strategy instead. We collect measurements of three effort parameters, namely, time, number of post-edited words and perception of effort, as representative of the three dimensions (temporal, technical and cognitive) in a real post-editing task. The results show that, although there are correlations between the measurements, the effort parameters differ in the classification of a considerable number of sentences. We conclude that the multidimensional strategy is necessary to estimate the overall post-editing effort.

Keywords

machine translation, post-editing effort, quality estimation

1 Introducción

La integración de un sistema de traducción automática (TA) en un flujo de traducción de un proveedor de servicios lingüísticos conlleva la toma de varias decisiones. Por una parte, se encuentran aquellos aspectos de índole más técnica que abarcan desde la compatibilidad de herramientas y formatos, hasta cuestiones de rendimiento de las máquinas. Por otra parte, se han de abordar asuntos relacionados con la aplicación y uso que se quieran hacer de las propuestas de TA. Existen varias posibilidades a la hora de emplear dichas propuestas, entre otras:

- Presentar las propuestas de TA al traductor para todos los segmentos.
- Presentar las propuestas de TA al traductor siempre que se consideren de calidad adecuada para posedición.
- Presentar las propuestas de TA al traductor únicamente cuando no existan propuestas de una memoria de traducción (MT) a partir de un umbral preestablecido.
- Presentar las propuestas de TA al traductor únicamente cuando no existan propuestas de MT con un umbral preestablecido y se consideren de calidad adecuada para posedición.

La elección de una u otra opción dependerá de la calidad global del sistema de TA, así como de las características y los requisitos de los

DOI: 10.21814/lm.11.2.277

encargos de traducción. Claramente, la primera opción sería la más sencilla de implementar. Bastaría con traducir el texto completo con el sistema de TA y entregárselo al traductor para su posedición. Para ser eficiente, este escenario requeriría de propuestas de TA de calidad constante por encima de un umbral determinado, debido a que al no considerar la ayuda de segmentos de MT el traductor deberá trabajar continuamente con las propuestas de TA independientemente de su calidad.

Con respecto a las opciones de implementación descritas, aquellas que consideran la calidad de los segmentos que se proponen al traductor, independientemente de si proceden de una MT o de un sistema de TA, plantean un escenario óptimo para el uso eficiente de las tecnologías disponibles (Parra Escartín et al., 2017). Por una parte, se facilita la reutilización de segmentos ya traducidos y validados anteriormente, y por otra, se optimiza la tarea de realizar nuevas traducciones, ya sea con la ayuda de propuestas de TA en caso de que resulten adecuadas para una posedición eficiente, ya sea sin ellas, en cuyo caso el traductor formulará su propuesta sin necesidad de trabajar con propuestas de TA de mala calidad.

Sin embargo, este escenario requiere una implementación más compleja. En primer lugar, se debe establecer el umbral de coincidencia parcial para filtrar las propuestas de la MT. Aun siendo ésta una práctica habitual, para establecer el umbral idóneo en el escenario que nos ocupa, se deberá comparar la aportación de un segmento de MT con la de una propuesta de TA además de con la traducción manual (Forcada & Sánchez-Martínez, 2015). Por lo tanto, este umbral no será necesariamente el utilizado en un flujo sin TA. En segundo lugar, se debe establecer una manera de filtrar aquellas propuestas de TA que se presentarán al traductor y aquellas que se descartarán a favor de la traducción manual. Estos filtrados, y el segundo en particular, conllevan una tarea compleja, ya que requieren entrenar modelos automáticos de predicción que se han de desarrollar en una etapa previa al trabajo de traducción con datos reales del esfuerzo de posedición, e integrarlos en el flujo de traducción para que decidan automáticamente si una oración deberá presentarse al traductor para traducir o, junto con su traducción automática, para poseditar.

En este trabajo, nos centramos en esa etapa previa de entrenamiento de modelos automáticos de predicción, en concreto, en estudiar el tipo de información que se debería utilizar para entrenar dichos predictores. Tomando como base las tres dimensiones de esfuerzo de posedición identificadas por Krings (2001), partimos de la hipótesis de que el esfuerzo real de posedición no estará reflejado en su totalidad en dichos modelos si no se consideran las tres dimensiones, ya que por separado, las dimensiones pueden no alcanzar a medir parte del esfuerzo total incluido en la tarea. Por ello, proponemos el uso de una estrategia multidimensional que combine información referente a las tres dimensiones. Los datos de esfuerzo temporal, técnico y cognitivo se podrían integrar en el proceso de aprendizaje de los modelos de diversas maneras. Por una parte, se podrían incluir como características para entrenar un clasificador que prediga si una oración se debería poseditar o traducir. Por otra, se podrían utilizar como objetivos a predecir para (1) crear tres modelos que predigan cada dimensión de manera independiente, que se combinarían posteriormente para decidir si traducir o poseditar, (2) crear un único modelo que combine las tres dimensiones de esfuerzo como objetivos, o (3) una versión mixta en la que algunas dimensiones sean características y otras objetivos a predecir.

Dado que es posible recopilar información sobre las tres dimensiones durante un trabajo de posedición rutinario, realizamos un estudio preliminar que nos permite una primera aproximación al estudio de la relación entre las dimensiones citadas. Un primer análisis muestra que, tras clasificar manualmente las oraciones de un corpus, los conjuntos de decisiones (poseditar o traducir) obtenidos para cada dimensión varían. Este hecho parece indicar la necesidad de considerar las tres dimensiones para entrenar modelos de predicción, de lo contrario, podríamos no estar considerando el esfuerzo real de posedición.

2 Trabajos relacionados

Son diversos los trabajos que se han centrado en el desarrollo de modelos automáticos de estimación que tratan de predecir la forma más eficiente de editar una oración, bien traduciéndola bien poseditándola (He et al., 2010; Callison-Burch et al., 2012; Parra Escartín & Arcedillo, 2015; Bojar et al., 2017). En general, estos modelos se crean a partir de un proceso de aprendizaje automático en cuyo entrenamiento se utiliza información sobre esfuerzo de posedición real. Es precisamente este esfuerzo de posedición lo que determina si es más eficiente traducir o poseditar una oración.

Una de las mayores dificultades de este proceso es precisamente la medición del esfuerzo de posedición. En el estudio de referencia sobre posedición, Krings (2001) afirma que el esfuerzo de po-

sedición está compuesto por tres dimensiones, a saber, la dimensión temporal, la dimensión técnica y la dimensión cognitiva. Atendiendo en mayor o menor medida a esta teoría, estudios previos han utilizado diferentes parámetros para representar el esfuerzo de posedición (véase la Figura 1). Algunos trabajos se han basado en recopilar la percepción del esfuerzo como parámetro de la dimensión cognitiva, es decir, se ha recopilado la percepción de un evaluador sobre lo difícil que resultaría poseditar una propuesta de TA en una escala del 1 al 5 al analizar la propuesta de traducción y una versión de posedición realizada por un traductor, sin que tuviera que completar la posedición él mismo previamente (Specia et al., 2010; Felice & Specia, 2012; Shah et al., 2015). Moorkens et al. (2015), en cambio, han estudiado la posibilidad de utilizar el tiempo de fijación de la mirada como parámetro para esta dimensión.

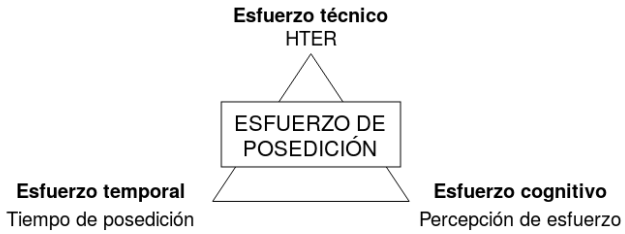

Figura 1: Las tres dimensiones del esfuerzo de posedición según Krings (2001) junto con posibles parámetros de medición

Otros trabajos consideran que el parámetro que se debería utilizar para medir el esfuerzo y clasificar las oraciones es el tiempo de ejecución (Parra Escartın & Arcedillo, 2015). En estos trabajos se llevan a cabo tests de productividad que comparan el ratio de segundos por palabra requeridos para traducir y poseditar una serie de oraciones (Plitt & Masselot, 2010). Esta información les permite concluir cuál de las dos aproximaciones, traducir o poseditar, es más rápida y clasificar las oraciones en la clase correspondiente. Si bien es cierto que en las tareas compartidas de estimación de calidad de 2013 y 2014 varios trabajos se centraron en crear modelos de estimación de tiempo, no se conoce ningún trabajo aplicado que recojan estos modelos y los utilice para desarrollar modelos de estimación binarios.

A pesar de que existen trabajos que consideran la percepción de esfuerzo y el tiempo de ejecución como posibles indicadores del esfuerzo de posedición, el parámetro más extendido y utilizado en las tareas compartidas de estimación de calidad (2012–2018) (Callison-Burch et al., 2012; Bojar et al., 2017) es la métrica HTER (Snover et al., 2009). Esta métrica es una variación de la tasa de edición de la traducción (TER, del inglés translation error rate) (Snover et al., 2006) que calcula de manera automática el número de ediciones (inserciones, eliminaciones, sustituciones y reordenaciones) que un traductor realiza en la propuesta de TA para que coincida con una traducción de referencia. En HTER, los traductores crean una nueva traducción de referencia o versión de posedición propia que sea fiel al original en términos de fluidez y significado realizando el número mínimo de ediciones. Después, se calcula el número de ediciones necesarias para transformar la propuesta de TA en la versión de posedición final.

Una rápida revisión de la literatura muestra que se ha experimentado con diferentes parámetros para medir el esfuerzo de posedición, si bien la métrica HTER es el estándar en el área, promovida por su supuesta correlación con la percepción de esfuerzo humana (Snover et al., 2006; Specia & Farzindar, 2010) y, sobre todo, por tratarse del parámetro más sencillo de obtener. Otra de las conclusiones que sacamos es que los estudios que tratan de medir o representar el esfuerzo de posedición se limitan al uso de un solo parámetro de esfuerzo, dando por sentado que la información de una única dimensión es suficiente para representar el esfuerzo real, si bien esto no está demostrado.

Si, como parece apuntar este trabajo, el resultado de la medición del esfuerzo de posedición varía dependiendo de la dimensión utilizada, aquellos flujos de traducción que únicamente consideren una de ellas podrían llevar a una toma de decisiones subóptimas. Por ejemplo, si una empresa utiliza información temporal para entrenar un modelo de clasificación binario, es muy probable que el modelo siga presentando a la traductora oraciones para poseditar que supongan gran esfuerzo técnico o cognitivo. Asimismo, si una empresa utiliza la dimensión técnica como criterio para clasificar oraciones el aptas para traducir o poseditar, podría estar agrupando oraciones que requieran un esfuerzo cognitivo o temporal diverso. Creemos que el uso combinado de la información representativa de las tres dimensiones de esfuerzo podría llevar a una estimación más precisa del esfuerzo real de posedición, ofreciendo la oportunidad de crear entornos de trabajo óptimos y tarifas más justas, entre otros.

3 Propuesta multidimensional

Tal y como refleja el análisis previo, el parámetro más común utilizado para clasificar una oración como adecuada para poseditar o idónea pa-

ra traducir es la medida HTER. Podemos argumentar que esta medida viene a reflejar, si bien de manera incompleta, la dimensión de esfuerzo técnico, ya que considera las ediciones necesarias para transformar la propuesta de TA en un segmento de calidad requerida. Decimos que recoge el esfuerzo técnico de manera parcial porque el recuento de ediciones llevado a cabo por la métrica no considera todas las ediciones realizadas durante el proceso de posedición de la oración sino que calcula las ediciones partiendo de la versión final de posedición, es decir, no tiene en cuenta las rectificaciones ni los cambios realizados antes de fijar la versión final de posedición.

Parece innegable que esta manera de medir el esfuerzo es considerablemente limitada, ya que excluye la dimensión temporal así como la cognitiva. Sin embargo, en línea con las argumentaciones de algunos trabajos que tratan de estudiar las relaciones entre las dimensiones de esfuerzo, así como de los distintos parámetros que se utilizan para representarlos, se podría inferir que, de manera indirecta, la métrica también reconoce parte del esfuerzo temporal, puesto que cuanto mayor sea el número de cambios, más tiempo requiere la posedición. Sin embargo, este razonamiento sólo es cierto si todas las ediciones requieren el mismo tiempo y, tal y como apunta Temnikova (2010), ciertas ediciones suponen un esfuerzo cognitivo mayor, lo que hace suponer que requerirán mayor tiempo. De hecho, Koponen (2012) estudia la relación entre la dimensión técnica y cognitiva, y concluye que HTER y la percepción de esfuerzo pueden dar resultados diferentes. A partir de oraciones traducidas automáticamente del inglés al español, la autora obtiene puntuaciones sobre la percepción del esfuerzo de posedición para representar la dimensión de esfuerzo cognitivo, y calcula la métrica TER para atender a la dimensión de esfuerzo técnico. Los resultados no siempre son equiparables. Lo mismo ocurre en el estudio de Moorkens et al. (2015) en el cual se concluye que la percepción de esfuerzo no es equiparable al esfuerzo real determinado por el tiempo de posedición, el tiempo de fijación de mirada y TER. También podría ocurrir, asimismo, que el esfuerzo de diferentes tipos de ediciones se compense entre ellos en cada segmento y siga habiendo buena correlación.

Podemos extraer dos conclusiones de los trabajos que estudian la relación entre las dimensiones de esfuerzo. Primero, que los parámetros que utilizamos para abordar las diferentes dimensiones de esfuerzo, si bien están enfocadas a una de ellas de manera más explícita, podrían atender a las demás en cierto grado. Segundo, que los resultados de los distintos parámetros no siempre son equiparables. Por tanto, al utilizar únicamente un parámetro para representar el esfuerzo de posedición y al centrarse éste especialmente en una de las tres dimensiones, es posible que estemos descuidando parte del esfuerzo real. Es evidente que carecemos de resultados concluyentes y que es necesario seguir investigando en este campo. Curiosamente, los trabajos que se centran en desarrollar modelos de predicción para el esfuerzo de posedición no han abordado este aspecto, a excepción de Aranberri & Pascual (2018), quienes proponen la inclusión de varios parámetros de esfuerzo como características de entrenamiento para modelos binarios.

En este trabajo, exploramos y comparamos las mediciones de esfuerzo de las tres dimensiones por separado. En esta primera aproximación, recogemos información sobre el tiempo de posedición, la percepción de esfuerzo según el poseditor y HTER durante un proceso de posedición rutinaria, y tras estudiar los resultados, proponemos utilizar los datos derivados de las tres mediciones de manera unificada para entrenar modelos de clasificación automáticos que, una vez implementados en el flujo de traducción, separen las oraciones bien para poseditar o traducir.

4 Ejemplo de aplicación

En esta sección exponemos la metodología seguida para la obtención de una clasificación binaria (poseditar, traducir) manual para estudiar la relación entre los resultados obtenidos para las distintas dimensiones. En primer lugar, se describe el corpus, atendiendo a los textos incluidos. En segundo lugar, se describe el proceso de recopilación de los datos correspondientes a los parámetros de las tres dimensiones de esfuerzo, es decir, el tiempo de posedición, la percepción de esfuerzo y los valores de HTER. Finalmente se establecen los umbrales para crear los subconjuntos de oraciones óptimas para traducir o para poseditar para cada parámetro, y se analiza la relación entre dichos subconjuntos antes de presentar la selección final.

4.1 Compilación del corpus

Describimos en primer lugar los textos incluidos en el corpus, atendiendo a los tres factores que, según Bernth & Gdaniec (2001), determinan la calidad de una traducción automática y es, por lo tanto, imprescindible exponerlos para tener una visión completa del corpus: el sistema de TA, el par de lenguas y el dominio de los tex-

tos. Estos factores fueron predeterminados por el contexto real para el cual se realizó este estudio.

Los documentos recopilados son textos técnicos que pertenecen al área de la construcción. Específicamente, se trata de documentos de instalación y mantenimiento de equipamiento. Debido a la tipología textual, los textos incluyen un alto número de repeticiones, listas y enumeraciones, entre otros. El subcorpus anotado para este estudio consta de 7 textos con un total de 509 oraciones y 6.542 palabras (veáse el Cuadro 1).

Texto	# palabras	# frases	palabras/frase
1	360	25	14,40
2	663	74	8,72
3	444	53	8,38
4	726	70	10,37
5	1.198	95	12,61
6	246	17	14,47
7	2.905	174	16,70
Total	6.542	509	12,83

Cuadro 1: Descripción del corpus

Los textos están redactados originalmente en español y la empresa recibe el encargo de traducirlos al inglés (entre otras lenguas). Por lo tanto, a diferencia de estudios de investigación anteriores que se centran en el inglés como lengua de origen (Felice & Specia, 2012; Hardmeier, 2011), este trabajo se centra en el español como lengua de origen y el inglés como lengua de destino.

La traducción automática de los textos del español al inglés se obtuvo con el sistema de TA utilizado por la empresa dentro de su flujo de producción. Se trata de un sistema de TA neuronal (arquitectura de codificador–decodificador con mecanismos de atención (Bahdanau et al., 2014)) desarrollado específicamente para la empresa y personalizado para el respectivo cliente.

4.2 Compilación de parámetros de esfuerzo: percepción, tiempo y HTER

Los datos para los tres parámetros de esfuerzo de posedición se recogieron durante una tarea de posedición en la cual se extrajo información real de los parámetros que nos atañen. Es precisamente durante el trabajo de posedición cuando es posible medir el tiempo de ejecución. A su vez, la fiabilidad sobre la percepción del esfuerzo de posedición aumenta si la traductora realiza dicho trabajo, y no simplemente imagina el esfuerzo que conllevaría. Es necesario subrayar que esta técnica es dependiente de la capacidad de comunicación del evaluador, por lo tanto, en este trabajo hablaremos de percepción *comunicada* del esfuerzo. Finalmente, para calcular los valores de HTER es imprescindible contar con una versión poseditada de la propuesta de TA, por lo que se calcularon una vez terminada la tarea. En los siguientes párrafos se describe el proceso seguido para la obtención de las mediciones de cada uno de los parámetros de esfuerzo.

Es evidente que sería necesario contar con múltiples traductoras que completaran varias tareas extensas de forma paralela para poder recopilar datos suficientes con los que obtener resultados concluyentes. En este trabajo en concreto, sin embargo, contamos con recursos limitados. Tuvimos a nuestra disposición tres traductoras profesionales internas durante un tiempo limitado. Según la encuesta realizada, las tres traductoras son mujeres, nacieron entre 1987 y 1991 y han estado trabajando en la industria de la traducción de 2 a 4 años. Todas cuentan con estudios universitarios de traducción, con especialidad científica, técnica y literaria, y trabajan con inglés y español como lenguas de trabajo. Con respecto a la posedición, todas tenían experiencia en este campo. Sin embargo, su actitud hacia el uso de TA para posedición es algo negativa y afirman que la traducción manual es más efectiva, ya que resulta más fácil y más rápida.

Debido a la disponibilidad limitada de las traductoras, el trabajo de posedición se dividió en tres partes comparables teniendo en cuenta los límites y la longitud de los textos, y cada traductora poseditó una de ellas. Las tres partes contaban aproximadamente con el mismo número de palabras (∼2.180) (véase el Cuadro 3).

El trabajo de posedición se realizó en la plataforma en línea Matecat (Federico et al., 2014), la cual nos permitió recopilar información para las tres dimensiones de manera sencilla y relativamente precisa, específicamente, el tiempo total de edición para cada oración, la percepción comunicada del esfuerzo y HTER.

Se prepararon las tareas para cada traductora por separado de manera que cada una accedía al texto original y a la traducción automática obtenida por el sistema de TA mencionado anteriormente para su parte del trabajo. Su trabajo, por lo tanto, consistió en poseditar la TA para lograr una traducción de buena calidad. Con el objetivo de garantizar un trabajo coherente y lo más homogéneo posible, se les proporcionó una guía de posedición con pautas específicas (ver Anexo I). No se restringió ni el tipo ni el número de ediciones posibles, pero se les pidió que modificaran la propuesta de TA lo mínimo posible.

Valor	Descripción	
1	Sin sentido	La traducción era totalmente incomprensible.
2	No utilizable	La traducción contenía tantos errores que claramente hubiera sido más rápido traducir.
3	Neutral	La traducción contenía bastantes errores, no está claro qué hubiera sido más rápido.
4	Utilizable	La traducción contenía algunos errores, pero sería más útil poseditar.
5	Muy buena	La traducción era correcta o casi correcta.

Cuadro 2: Escala de valoración de la percepción de esfuerzo de posedición

Además de poseditar su parte correspondiente, las traductoras calificaron el esfuerzo que les había supuesto el trabajo realizado con cada oración. Seguimos la metodología utilizada por Lacruz et al. (2014), quienes piden al evaluador que califique la idoneidad de un segmento para posedición después de haberlo editado, según una escala del 1 al 5 (véase el Cuadro 5). Ambas tareas se realizaron a la par para atenuar su dificultad y baja fiabilidad descrita en estudios anteriores (Callison-Burch et al., 2012). Se pidió a las traductoras que tras realizar la posedición añadieran al final de cada segmento la puntuación que considerasen oportuna. Esta opción ofrecía una manera sencilla de combinar las tareas y extraer dichas puntuaciones en un paso posterior. De esta manera conseguimos que la percepción de la dificultad se capturase inmediatamente después de editar cada segmento, lo más inalterada posible. Sin embargo, esta opción puede llegar a desvirtuar en cierto grado la medición del tiempo en caso de que la toma de decisión de las traductoras varíe de manera significativa para los distintos segmentos. Para minimizar este riesgo se pidió a las traductoras que limitaran el tiempo de decisión lo máximo posible. Es importante mencionar que el 1 % de las oraciones (5 oraciones en total) carecían de puntuación de esfuerzo. Estos valores fueron reemplazados por la mediana de los valores totales. Así, se les asignó un 5, ya que más de la mitad de las traducciones recibieron un 5.

La medición del tiempo fue sencilla desde el punto de vista técnico, ya que Matecat calcula el tiempo que la traductora emplea en cada segmento. Este cálculo se realiza acumulando el tiempo que cada segmento está activo, es decir, que el cursor se encuentra en la celda que corresponde a la traducción de un segmento concreto. Esta opción permite a la traductora retomar un segmento cuantas veces sea necesario y acumular el tiempo correspondiente. Es cierto que una traductora puede leer y considerar el trabajo realizado en un segmento mientras otro está activado, lo cual muestra una limitación de la herramienta para este tipo de mediciones. Sin embargo, las traductoras fueron informadas de la distorsión que este hecho podía introducir en los datos a fin de minimizarla. Por lo tanto, el tiempo dispuesto para la posedición de cada oración se obtuvo de manera automática tras la finalización de los trabajos.

A pesar de que existen herramientas que pueden capturar todas las ediciones realizadas durante el proceso de posedición (Aziz et al., 2012), y por consiguiente, recopilar el esfuerzo técnico de manera más precisa, decidimos centrar nuestro estudio en la métrica HTER, ya que es la más extendida tanto en trabajo de investigación como en la práctica industrial. Tal y como se ha indicado en la sección anterior, HTER es una métrica que calcula, de manera automática, el número de ediciones necesarias para transformar una propuesta de TA en una oración de calidad adecuada. Para ello, la métrica requiere el texto producido por el sistema de TA, así como una versión final de éste. En nuestro caso, calculamos el HTER tras haber completado el trabajo de posedición con la versión de TA presentada a las traductoras y las posediciones creadas por ellas. Este cálculo nos informa del número de ediciones necesario en cada una de las oraciones de la tarea.

Examinemos el trabajo de las traductoras individualmente. Podemos ver un resumen de su labor en el Cuadro 3. La traductora 1 trabajó con el conjunto que ofrecía mayor variación de oraciones, ya que incluía oraciones de cuatro textos diferentes. El tiempo medio empleado por oración fue de 31 segundos (0,57 min). Resultó ser la más rápida. Asimismo, observamos que es precisamente esta traductora quien modifica la traducción automática en menor proporción, ya que su promedio de HTER global es el más bajo con un valor de 6,83. Para ella, la valoración media de esfuerzo de posedición es de 4,66.

La traductora 2 abordó oraciones de dos textos diferentes. El tiempo medio empleado por oración es de 1 minuto. Su promedio de HTER es de 14,90 ediciones por segmento, lo cual la sitúa como la traductora que introduce el mayor número de cambios durante la posedición. La valoración de esfuerzo se sitúa en 4,56, aunque mínimamente, por debajo de la valoración asignada por sus compañeras a sus respectivos trabajos.

Traductora	# textos	# frases	# palabras	tiempo	seg/palabra	seg/frase	HTER	percepción comunicada
Traductora 1	4	222	2.168	02h:07m	3,5	34	6,83	4,66
Traductora 2	2	150	2.173	02h:34m	4,3	62	14,90	4,56
Traductora 3	1	137	2.198	04h:59m	8,2	131	10,14	4,66

Cuadro 3: Descripción de la labor de las traductoras en la tarea de posedición

La traductora 3 contó con oraciones de un solo texto. El tiempo medio empleado por oración es de 2,2 minutos. Se trata de la traductora más lenta. Su promedio de HTER es de 10,14. Casualmente, la valoración de esfuerzo es exactamente la misma que la de la Traductora 1: 4,66.

Los datos reportados dejan en evidencia que a pesar de las medidas tomadas para dividir la tarea de forma uniforme, la tarea y proceso de posición de cada traductora difiere, sobre todo en lo referente a la longitud de las oraciones poseditadas y la velocidad de trabajo. Curiosamente, podemos observar cómo, independientemente del tiempo empleado o del número de textos abordados, la percepción comunicada de esfuerzo es consistente entre las tres traductoras. Su valoración indica que el sistema de TA proporciona, en general, un número considerable de oraciones aptas para posedición. Estos resultados están en yuxtaposición directa con los resultados obtenidos en la encuesta. Al preguntarles respecto a la tarea específica de PE que llevaron a cabo, las traductoras consideraron que la TA no era útil, si bien una de ellas afirmó que era bastante precisa. De acuerdo con esto, podemos decir que la opinión de las traductoras sobre el uso de TA para posedición es claramente negativa, ya que aún habiendo valorado el esfuerzo de posedición como bajo (4,6 de media), siguen respondiendo que la tecnología no es provechosa.

Tras completar la tarea de posedición y extraer la información necesaria, los datos recopilados para cada oración incluyen la oración original en español, la traducción automática al inglés, su versión poseditada en inglés, una valoración del esfuerzo de posedición, el tiempo de posedición y el valor de HTER. Es conveniente apuntar que debido a que el estudio se realizó en una empresa privada, no ha sido posible hacer público el conjunto de datos.

4.3 Correlación de los parámetros de esfuerzo

En esta sección se inspecciona la relación entre los resultados de los tres parámetros propuestos como representativos de las dimensiones del esfuerzo de posedición (percepción comunicada de esfuerzo, tiempo de posedición y valor de HTER) para cada segmento. Para ello, se ha calculado la correlación de Spearman. Es importante apuntar que, pese a las diferencias que emergieron respecto al trabajo de las traductoras, el análisis de correlación de parámetros de esfuerzo se realizó con la combinación de todos los datos recopilados. Es sabido que el proceso de traducción/posedición varía dependiendo del profesional que lo realiza aun tratándose del mismo conjunto de segmentos, y por consiguiente, un análisis siempre conlleva esta variabilidad en mayor o menor grado. Además, la unificación de los datos nos permite examinar un conjunto más representativo. No obstante, se realizó el mismo análisis para los datos de cada traductora por separado, lo cual confirmó que, sin bien los valores absolutos no coincidían, las tendencias eran las mismas, llevándonos a las mismas conclusiones. Los resultados se muestran en el Cuadro 4.

Parámetro de esfuerzo	(1)	(2)	(3)
(1) Percepción comunicada	1,000		
(2) Tiempo	-0,386*	1,000	
(3) HTER	-0,712*	0,443*	1,000

Cuadro 4: Correlación de Spearman de los parámetros de esfuerzo de posedición (*p<0.001)

Como podemos observar, la percepción comunicada de esfuerzo y HTER obtienen la correlación más alta, logrando una fuerte correlación. Esto podría indicar que, en cierta medida, el número de ediciones necesarias está relacionado con una percepción de mayor o menor de esfuerzo, es decir, a menor número de ediciones, más sencillo parece el trabajo. Sin embargo, la débil y moderada correlación entre el tiempo transcurrido durante la posedición, y la percepción de esfuerzo y HTER, respectivamente, no parece respaldar esta hipótesis.

En la Figura 2 podemos observar más claramente la correlación negativa entre la percepción de esfuerzo y HTER, es decir, cuanto más óptima parece una oración para posedición, menor es el valor HTER de dicha oración, menos ediciones necesita la propuesta de TA. La mayoría de las oraciones etiquetadas como 5, que indica que la oración es impecable y requiere un esfuerzo mínimo de posedición, tiene un HTER de 0, es decir, que no requiere ningún cambio.

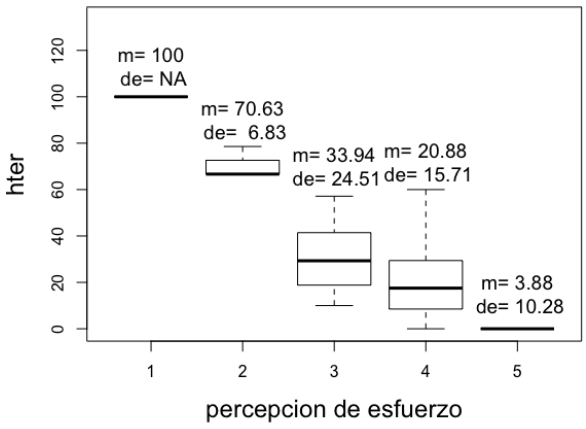

Figura 2: Correlación entre HTER y percepción comunicada de esfuerzo, donde *m* indica la media y *de* la desviación estándar

Se aprecia un gran salto entre los niveles 3 y 2. Para el nivel 2, el HTER es el segundo más alto. Esto significa que estas oraciones se han modificado en gran medida.

En la Figura 3 podemos observar la correlación negativa entre tiempo y percepción. Intuitivamente podríamos pensar que cuanto menor parece el esfuerzo requerido por una propuesta de TA, menor es el tiempo invertido en su posedición. Es interesante ver que esta correlación es débil. Por una parte, observamos que el tiempo medio necesario para poseditar las oraciones clasificadas en los niveles 2–5 no varía excesivamente. Si bien es cierto que la caja para el nivel 5 es comparativamente estrecha, lo cual sugiere que estas oraciones se poseditaron en un intervalo de tiempo similar, y que, por el contrario, la caja para el nivel 3 es comparativamente ancha, lo que indica que el tiempo de posedición varía más entre las traductoras en este caso, las medias se centran en torno a 3–8 segundos. Sin embargo, el tiempo medio necesario para poseditar las oraciones clasificadas como 1 se eleva a 36 segundos.

Por último, en la Figura 4, podemos ver cómo la correlación entre HTER y tiempo es positiva. Aunque la mayoría de los casos cuentan con un valor de HTER y un tiempo bajo, hay una tendencia que indica que cuanto más alto es el HTER, más alto es el tiempo de posedición. Esto parece indicar que cuantas más ediciones (inserciones, eliminaciones, sustituciones y reordenaciones, representadas por HTER) se hagan, más tiempo se necesita. Aun así, observamos que esta correlación es moderada, lo cual podría señalar la existencia de un tercer factor que distorsiona la relación directa entre HTER y el tiempo.

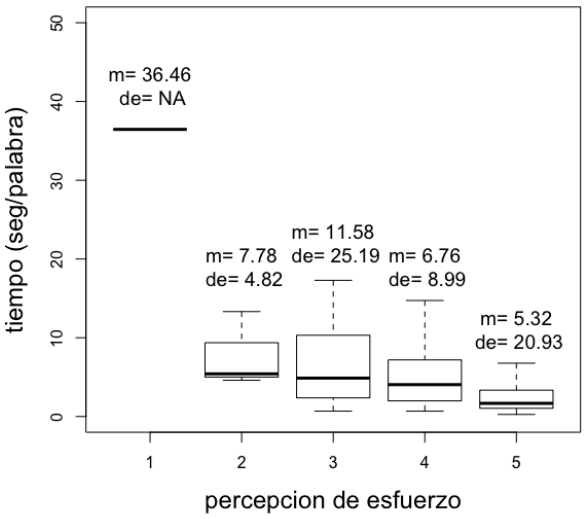

Figura 3: Correlación entre tiempo de posedición y percepción comunicada de esfuerzo, donde *m* indica la media y *de* la desviación estándar

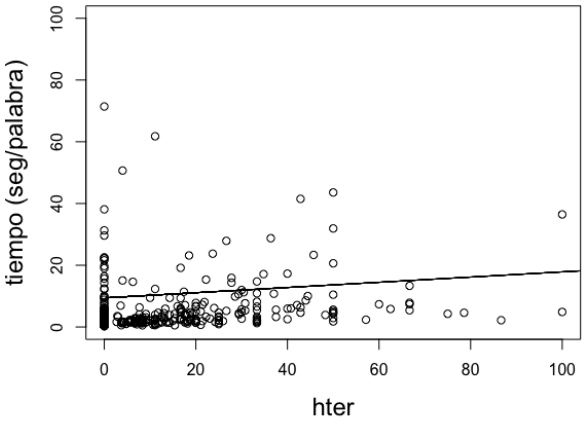

Figura 4: Correlación entre HTER y tiempo de posedición

4.4 Umbrales para los parámetros de esfuerzo

Como hemos mencionado anteriormente, el objetivo que perseguimos es estudiar la relación entre los resultados de los parámetros de las tres dimensiones para identificar qué parámetros de esfuerzo se deberían incluir en el entrenamiento de un modelo de predicción que sea capaz de recomendar si una oración nueva debería traducirse manualmente o si se debería poseditar. Sin embargo, los datos recopilados no responden a una división binaria; los parámetros de esfuerzo utilizados asignan a cada oración un rango de valores continuos (en el caso del tiempo y HTER) o dis-

cretos (en el caso de la percepción). Por lo tanto, una vez recopilados los datos, debemos establecer un umbral para clasificar las oraciones bien como óptimas para poseditar (PE) o para traducir (T) para cada uno de los parámetros de esfuerzo.

La literatura recoge varias aproximaciones a este paso. Una de las estrategias consiste en recopilar tareas de traducción y posedición para el mismo conjunto de oraciones y realizar una comparación directa de los resultados obtenidos para cada tarea (Aranberri & Pascual, 2018). Otros enfoques se basan en establecer valores de umbral específicos para cada parámetro de esfuerzo, como puede ser el de HTER (Parra Escartín & Arcedillo, 2015). Al carecer de datos de comparación directa (recordemos que las traductoras realizaron un trabajo de posedición pero no de traducción), proponemos umbrales para cada uno de los parámetros atendiendo al entorno experimental. Pese a que sería posible seleccionar umbrales distintos, que probablemente resultarían en una clasificación final distinta, nuestro objetivo principal no es obtener la selección final de las oraciones a traducir o poseditar sino mostrar la diferencia existente entre cada una de las dimensiones.

Percepción comunicada de esfuerzo > 3

El Cuadro 5 muestra la escala de percepción comunicada de esfuerzo de 1 a 5 según la cual las traductoras tuvieron que calificar cada oración de TA. Dada la definición asignada a cada valor, las valoraciones de 4 y 5 suponen que la calidad de las propuestas de TA es adecuada para posedición, mientras que en la valoración de 3 es dudosa, y para los valores 2 y 1 claramente no es adecuada. Teniendo en cuenta la opinión negativa de las traductoras hacia la traducción automática y el trabajo de posedición, decidimos excluir de la tarea de posedición todas aquellas oraciones que no facilitasen notablemente la posedición. Partiendo de esta decisión, establecemos el umbral para la percepción de esfuerzo en 3. De esta manera, las oraciones con una valoración de 4 y 5 se clasificarán para posedición y las oraciones con valoraciones de 1, 2 y 3 para traducción.

Tiempo < 11,5 seg/palabra

Durante la tarea de posedición, la plataforma Matecat registró el tiempo invertido por las traductoras en cada oración. Asumimos que cuanto menor es el tiempo empleado para poseditar una propuesta de TA, más sencilla resulta la tarea. El tiempo medio de posedición para las oraciones de nuestro conjunto de datos es de 6,1 segundos por palabra. El recuento del tiempo nos ofrece la posibilidad de ordenar las oraciones según el tiempo de trabajo requerido, normalizado por el número de palabras, pero aún así necesitamos establecer un umbral para poder obtener la clasificación binaria que buscamos. Para ello nos basamos en el tiempo medio asignado por la empresa para las tareas de traducción, que asciende a 313 palabras por hora y, por lo tanto, a 11,5 segundos por palabra. Así, aquellas oraciones que estén por debajo de ese tiempo se clasificarán como óptimas para poseditar y viceversa, las oraciones que superen los 11,5 segundos por palabra se clasificarán para traducir.

HTER < 33

HTER representa el número de ediciones necesarias para obtener una versión de traducción adecuada. Por lo tanto, cuanto más bajo sea el valor de HTER, mejor será la calidad de la propuesta de TA. Los pocos estudios que se han centrado en establecer el valor óptimo de HTER a partir del cual las oraciones se deberían poseditar o traducir, sugieren que éste se encuentra entre los 30–35 puntos (Parra Escartín & Arcedillo, 2015), por lo que nos centraremos en esa franja. Si observamos la Figura 5, vemos que para el nivel de percepción comunicada de esfuerzo 3, el HTER promedio de nuestros datos es del 33 %. Viendo que este valor se encuentra dentro del rango sugerido por los estudios previos, establecemos el umbral en el 33 % y asumimos que las oraciones con un HTER inferior al 33 % son de una calidad lo suficientemente buena como para ser seleccionadas para posedición.

4.5 Clasificación de las oraciones para los parámetros de esfuerzo

En resumen, atendiendo a los umbrales definidos, clasificamos las oraciones en óptimas para posedición (PE) o traducción (T), creando tres conjuntos, uno para cada dimensión tratada. Para dividir el conjunto según la percepción de esfuerzo, unimos las oraciones de los grupos 1, 2 y 3 asignándoles la categoría de T y las oraciones de los grupos 4 y 5 en PE. Para el conjunto de tiempo, aquellas oraciones que cuenten con un total de tiempo que suponga una media superior a 11 segundos por palabra formarán el subconjunto T, mientras que el formarán el subconjunto de PE. Finalmente, la división del conjunto según HTER se realizará asignando aquellas oraciones con un valor superior a 33 al subconjunto T y las oraciones con un valor inferior a 33 al subconjunto PE.

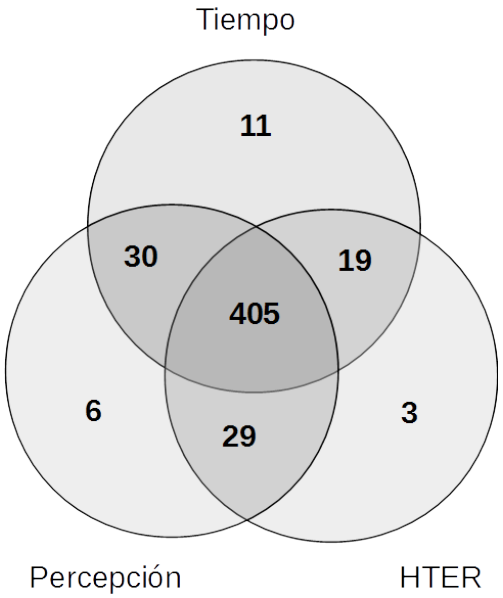

Figura 5: Número de elementos clasificados para poseditar comunes a las distintas dimensiones de esfuerzo de posedición.

El Cuadro 5 muestra el número de oraciones asignadas a cada clase para cada uno de los parámetros. A pesar de que el tamaño de ambas clases es muy similar independientemente del parámetro de esfuerzo utilizado, no se puede concluir que el comportamiento de los tres parámetros de esfuerzo es semejante. Esto se debe a que estos porcentajes no consideran que las oraciones incluidas en cada clase sean las mismas. Con el objetivo de estudiar la clasificación exacta de cada oración, calculamos cuáles de las asignadas para posedición son comunes a los tres parámetros de esfuerzo, cuáles a dos de ellos y cuáles únicamente a uno (ver Figura 5).

Clase	Percepción	Tiempo	HTER
PE	470 (92.34 %)	465 (91.36 %)	456 (89.59 %)
T	39 (7.66 %)	44 (8.64 %)	53 (10.41 %)

Cuadro 5: Número de oraciones asignadas a las clases PE y T para cada parámetro

Los resultados muestran que un total de 405 oraciones pertenecen a la intersección de los tres parámetros de esfuerzo, es decir, que los tres parámetros las clasifican para posedición. De la misma manera, en la Figura 5 también se puede observar el número de oraciones que son comunes únicamente a dos de los parámetros (por ejemplo, 30 en el caso del tiempo y la percepción). Finalmente, también se observa el número de oraciones que son comunes a un único parámetro de esfuerzo (por ejemplo, 3 en el caso de HTER).

Estos resultados muestran que si bien los tres parámetros de esfuerzo han asignado un porcentaje elevado de oraciones a la misma clase, la clasificación de alrededor de un 15 % difiere. Estos resultados parecen indicar que los parámetros de esfuerzo evaluados no son capaces de representar el esfuerzo global de posedición por separado, y por lo tanto, sugieren que se deberían utilizar de manera conjunta para una medición más exacta del esfuerzo de posedición. Es posible que al limitarnos al uso de los parámetros por separado, como es común en el desarrollo de modelos automáticos de estimación, no estemos proporcionando a dichos modelos datos completos para un entrenamiento óptimo.

A la vista de la discordancia en la clasificación de oraciones dependiendo del parámetro de esfuerzo, en este trabajo adoptamos una visión conservadora y proponemos clasificar como oraciones óptimas para posedición únicamente aquellas que son comunes a las tres dimensiones, que en este caso supondría el 79,5 % de las oraciones totales. Esta decisión se basa en la visión general negativa sobre la TA por parte de las traductoras. Con los datos recopilados y la clasificación realizada, se proseguiría al entrenamiento de un modelo automático de estimación que clasificase las nuevas oraciones en óptimas para posedición o traducción. Éste propondría para poseditar aquellas oraciones que fueran óptimas desde el punto de vista de las tres dimensiones. De esta manera, pretendemos que las traductoras tengan una experiencia lo más positiva posible durante las tareas de posedición y comiencen a vincular la TA con unos hábitos de trabajo satisfactorios.

5 Conclusiones

Este trabajo cuestiona el uso por separado de información referente a las dimensiones de esfuerzo propuestas por Krings (2001) (temporal, cognitiva y técnica) a la hora de medir el esfuerzo del trabajo de posedición y aboga por la inclusión de información de las tres dimensiones de manera conjunta. Se propone una estrategia multidimensional para la selección de información incluida en los modelos automáticos de selección de candidatos de TA para poseditar. Proponemos combinar parámetros que representan las tres dimensiones para establecer un umbral más preciso para clasificar oraciones en óptimas para traducir o para poseditar. Específicamente, en este trabajo se estudia la posibilidad de utilizar el tiempo como medida de la dimensión temporal, la percepción comunicada de esfuerzo de posedición como medida del esfuerzo cognitivo, y HTER como medida del esfuerzo técnico.

En el trabajo se presenta la recopilación de datos de un entorno real, los cuales se analizan siguiendo la propuesta multidimensional en preparación a utilizarlos posteriormente para el entrenamiento de modelos automáticos de clasificación que, una vez optimizados, podrían incluirse en el flujo de traducción para seleccionar las propuestas de traducción automática que se deberían presentar a las traductoras y excluir aquellas que perjudicarían la producción.

El análisis de las clasificaciones de este pequeño conjunto de datos reales utilizado a modo de ejemplo parece indicar que distintos parámetros de esfuerzo de posedición (percepción, tiempo y HTER) que atienden a las tres dimensiones en grados diferentes no valoran de igual manera si es más eficiente traducir o poseditar una oración. Los resultados obtenidos parecen sugerir que las dimensiones por separado no son capaces de describir el esfuerzo real de posedición y que, por lo tanto, el uso de un único parámetro de esfuerzo que se centre particularmente en una de las dimensiones no es adecuado para calcular el esfuerzo de posedición real.

Debido a la naturaleza preliminar de este trabajo, recordamos que los resultados deben tomarse con cautela. El conjunto de datos utilizado a modo de ejemplo cuenta con limitaciones tanto de extensión como de participantes, de modo que la unificación de los datos podría no ser completamente representativa del trabajo de posedición para el tipo de texto y sistema de traducción automática estudiados. Así, apuntamos como trabajo futuro replicar este análisis con un mayor número de traductoras, que realicen tareas paralelas, e incluyan conjuntos más extensos, para poder obtener datos más sólidos respecto a las correlaciones entre los parámetros de medición de las distintas dimensiones. Además, convendría extender el estudio de distintas combinaciones de umbrales.

Si bien parece conveniente combinar las tres dimensiones de esfuerzo, sería necesario continuar investigando qué parámetros representan las distintas dimensiones de manera más precisa y completa. Este trabajo ha explorado el uso de HTER como parámetro de esfuerzo técnico. Sin embargo, como ya se ha mencionado anteriormente, sería interesante utilizar el número total de ediciones realizadas por las traductoras durante el proceso de edición. Durante el proceso de traducción y posedición, es práctica común reformular una y otra vez distintas partes de una oración hasta conseguir una versión final, lo cual probablemente vaya ligado a la dificultad que entraña el segmento en cuestión. El uso de HTER podría ocultar el esfuerzo técnico real llevado a cabo en segmentos complejos.

Asimismo, cabría explorar otras mediciones más objetivas para el esfuerzo cognitivo, es decir, que no se basen en la opinión de las traductoras, quizá en la línea de Koponen et al. (2012) y Temnikova (2010), o incluso Moorkens et al. (2015). En este punto es necesario subrayar la importancia de la metodología seguida y las herramientas seleccionadas para recopilar los datos. Por ejemplo, en este trabajo, se ha querido primar la obtención del esfuerzo cognitivo de manera inmediata y para cada uno de los segmentos. Como el contexto del experimento exigía realizar la recopilación en línea, las aplicaciones disponibles no eran lo suficientemente flexibles como para realizar la medición temporal y de percepción completamente por separado. Si bien, como en este trabajo, se pueden tomar medidas adicionales para evitar que los datos se desvirtúen, consideramos que estrategias más rigurosas aportarán una mayor solidez de las conclusiones.

Finalmente, sería de gran interés analizar la relación existente entre las tres dimensiones de posedición para estudiar la aportación específica de cada dimensión al esfuerzo de posedición. Este estudio podría aportar mayor información a la hora de seleccionar o definir parámetros de esfuerzo para cada dimensión.

Agradecimientos

Las autoras quieren agradecer a los revisores, cuyos comentarios han contribuido a mejorar la versión original del trabajo. Este trabajo ha sido financiado parcialmente por el proyecto Modena (KK-2018/00087) del Departamento de Desarrollo Económico e Infraestructuras del Gobierno Vasco, el proyecto UnsupNMT (TIN2017-91692-EX) del Ministerio de Economía, Industria y Competitividad de España y el proyecto DOMINO (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE).

Referencias

Aranberri, Nora & Jose A Pascual. 2018. Towards a post-editing recommendation system for Spanish–Basque machine translation. En *21st Annual Conference of the European Association for Machine Translation (EAMT)*, 21–30.

Aziz, Wilker, Sheila Castilho Monteiro de Sousa & Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. En

8th Language Resources and Evaluation Conference (LREC), 3982–3987.

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint,* arXiv:1409.0473.

Bernth, Arendse & Claudia Gdaniec. 2001. MTranslatability. *Machine Translation* 16(3). 175–218. 10.1023/A:1019867030786.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia & Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT). En *2nd Conference on Machine Translation, Volume 2: Shared Task Papers*, 169–214. 10.18653/v1/W17-4717.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut & Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. En *7th Workshop on Statistical Machine Translation*, 10–51.

Federico, Marcello, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines et al. 2014. The MateCat tool. En *25th International Conference on Computational Linguistics: System Demonstrations (COLING)*, 129–132.

Felice, Mariano & Lucia Specia. 2012. Linguistic features for quality estimation. En *7th Workshop on Statistical Machine Translation*, 96–103.

Forcada, Mikel L. & Felipe Sánchez-Martínez. 2015. A general framework for minimizing translation effort: towards a principled combination of translation technologies in computer-aided translation. En *18th Annual Conference of the European Association for Machine Translation (EAMT)*, 27–34.

Hardmeier, Christian. 2011. Improving machine translation quality prediction with syntactic tree kernels. En *15th Annual Conference of the European Association for Machine Translation (EAMT)*, 233–240.

He, Yifan, Yanjun Ma, Josef van Genabith & Andy Way. 2010. Bridging SMT and TM with translation recommendation. En *48th Annual Meeting of the Association for Computational Linguistics (ACL)*, 622–630.

Koponen, Maarit. 2012. Comparing human perceptions of post-editing effort with post-editing operations. En *7th Workshop on Statistical Machine Translation*, 181–190.

Koponen, Maarit, Wilker Aziz, Luciana Ramos & Lucia Specia. 2012. Post-editing time as a measure of cognitive effort. *Workshop on Post-Editing Technology and Practice* 11–20.

Krings, Hans P. 2001. *Repairing texts: empirical investigations of machine translation post-editing processes*, vol. 5. Kent State University Press.

Lacruz, Isabel, Michael Denkowski & Alon Lavie. 2014. Cognitive demand and cognitive effort in post-editing. En *3rd Workshop on Post-Editing Technology and Practice*, 73–84.

Moorkens, Joss, Sharon O'Brien, Igor da Silva, Norma de Lima Fonseca & Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation* 29(3–4). 267–284. 10.1007/s10590-015-9175-2.

Parra Escartín, Carla & Manuel Arcedillo. 2015. Living on the edge: productivity gain thresholds in machine translation evaluation metrics. En *4th Workshop on Post-Editing Technology and Practice (WPTP)*, 46–56.

Parra Escartın, Carla & Manuel Arcedillo. 2015. Machine translation evaluation made fuzzier: A study on post-editing productivity and evaluation metrics in commercial settings, 131–144.

Parra Escartín, Carla, Hanna Béchara & Constantin Orăsan. 2017. Questing for quality estimation: A user study. *The Prague Bulletin of Mathematical Linguistics* 108(1). 343–354. 10.1515/pralin-2017-0032.

Plitt, Mirko & François Masselot. 2010. A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics* 93. 7–16. 10.2478/v10108-010-0010-x.

Shah, Kashif, Trevor Cohn & Lucia Specia. 2015. A bayesian non-linear method for feature selection in machine translation quality estimation. *Machine Translation* 29(2). 101–125. 10.1007/s10590-014-9164-x.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. En *Association*

for Machine Translation in the Americas, 223–231.

Snover, Matthew, Nitin Madnani, Bonnie J. Dorr & Richard Schwartz. 2009. Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. En *4th Workshop on Statistical Machine Translation*, 259–268.

Specia, Lucia & Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with HTER. En *Workshop Bringing MT to the User: MT Research and the Translation Industry (AMTA)*, 33–41.

Specia, Lucia, Dhwaj Raj & Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation* 24(1). 39–50. doi 10.1007/s10590-010-9077-2.

Temnikova, Irina P. 2010. Cognitive evaluation approach for a controlled language post–editing experiment. En *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, 3485–3490.

ANEXO I: Instrucciones de posedición

You will be presented machine translated (MT) sentences in English, together with their source sentence in Spanish. You need to modify (postedit) the machine translated sentence as little as possible so that it bares the same meaning as its source. This postedition may include substitutions, deletions, insertions or reorderings. For each postedition, three things will be recorded:

- Your postedited version of the sentence
- A number provided by you indicating the quality of the machine translated sentence
- The time you take to postedit (this will be recorded automatically)

Here you can find an example of the task you'll need to perform:

	Ejemplo
Source	Es necesario tapar esos agujeros ya que las patas traseras de la armadura apoyan sobre ellos.
MT	These holes have to be covered as the frame rear feet support them.
Postedited version	These holes need to be covered as the frame back legs support on them.
Quality	1– 2 – 3 – **4** –5

The tool: Matecat

The platform we will be using is called Matecat. Matecat is an open source online CAT tool. If you want more information, check this link: https://www.matecat.com/about/.

To perform this task, you will be given two links, one for each task (practice task and real task). In one, you will have 5 sentences to familiarize yourself with the environment. In the other, you will have between 130 and 230 sentences to postedit, around 2180 words. When opening the link, you will be directed to the environment in which you will work. This is the following:

The platform shows the source sentence on the left side and its machine translation on the right side. This machine translated proposal is taken from a translation memory that contains all segments translated automatically. You need to post-edit the Machine Translate sentence in the right window taking into account the source sentence shown on the left window.

Quality rating

To rate the quality of the Machine Translation, please follow the scale proposed by Lacruz, Denkowski & Lavie (2014):

Valor	Descripción	
1	Sin sentido	La traducción era totalmente incomprensible.
2	No utilizable	La traducción contenía tantos errores que claramente hubiera sido más rápido traducir.
3	Neutral	La traducción contenía bastantes errores, no está claro qué hubiera sido más rápido.
4	Utilizable	La traducción contenía algunos errores, pero sería más útil poseditar.
5	Muy buena	La traducción era correcta o casi correcta.

After post-editing each sentence, you need to add a numberat the end of each sentence indicating the quality of the machine translated sentence following the aforementioned scale (1-5). Then, hit the button TRANSLATED to finish.

Please keep in mind that you need to write this number for each sentence.

If you feel that a certain machine translated sentence is perfectand needs no post-editing, just add a numberfrom 1 to 5 at the end and hit the TRANSLATED button directly.

Propuesta recibida en noviembre de 2018 y aceptada para publicación en octubre de 2019.

Formalización de reglas para la detección del plural en castellano en el caso de unidades no diccionarizadas

Formalization of rules for the detection of plurals in Spanish in the case of out-of-vocabulary units

Rogelio Nazar
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl

Amparo Galdames
Instituto de Literatura y Ciencias del Lenguaje
Pontificia Universidad Católica de Valparaíso
amparo.galdames@pucv.cl

Resumen

En este artículo ofrecemos una formalización de reglas de pluralización en castellano para ser utilizada concretamente en el procesamiento de términos especializados, ya que con frecuencia estos no se encuentran registrados en los diccionarios de lengua general y, por tanto, no son reconocidos su categoría y lema. Esto tiene consecuencias negativas en tareas como la extracción de terminología, especialmente en el caso de lenguas con riqueza morfológica. Enfrentamos el problema con un diseño en forma de cascada de reglas de sustitución, expresiones regulares y adquisición léxica a partir de corpus de grandes dimensiones. Los resultados experimentales muestran una reducción significativa de la tasa de error de dos etiquetadores ampliamente utilizados: TreeTagger y UDPipe. Ofrecemos una implementación en código abierto que funciona como posproceso del etiquetado.

Palabras clave

etiquetado morfosintáctico, lematización, reglas de pluralización del castellano, unidades no diccionarizadas

Abstract

This paper presents a formalization of rules on plural formation in Spanish to be used in the processing of specialized terminology, as it is frequently the case that terms are not found in dictionaries of general language and therefore they cannot be lemmatized or POS-tagged. The absence of terms in general dictionaries has negative effects in tasks such as terminology extraction, particularly in the case of morphologically rich languages. We attack the problem by cascading through multiple trasnfser rules, regular expressions and lexical aquisition from large corpora. Results show significant reduction of the error rate of two POS-taggers: TreeTagger and UDPipe. We offer an open-source implementation which works as a post-process, cleaning up after the tagger.

Keywords

part-of-speech tagging, lemmatization, rules for plural in Spanish, out-of-vocabulary units

1 Introducción

Una de las tareas más elementales del área del procesamiento del lenguaje natural (PLN) es el etiquetado morfosintáctico o POS-tagging. Esto incluye la lematización de las formas que aparecen en el texto y la asignación de una categoría gramatical, con o sin análisis morfológico específico, que indique persona, género, número, etc. (Manning & Schütze, 1999). En otras palabras, se trata de reconocer la asociación entre una palabra flexionada y, por un lado, su lema o lexema —la forma elegida como entrada en los diccionarios— y, por otro lado, la categoría gramatical.

Los etiquetadores morfosintácticos fueron diseñados inicialmente como sistemas basados en reglas (Greene & Rubin, 1971), pero fueron gradualmente reemplazados por familias de algoritmos probabilísticos (Church, 1989; Schmid, 1994) y, más recientemente, neuronales (Ling et al., 2015; Straka & Straková, 2017; Qi et al., 2018). Estos últimos han mejorado sustancialmente la calidad del resultado y se han aplicado también al análisis sintáctico. El problema general de los métodos cuantitativos, sin embargo, es que si bien reducen el esfuerzo de la creación de reglas, exigen el de la producción de material de entrenamiento en forma de un corpus ya etiquetado o –al menos– revisado manualmente. El inconveniente radica en que entre el tamaño de este corpus de entrenamiento y la calidad del resultado existe una relación que parece describir un modelo logarítmico: se requiere cada vez más corpus de entrenamiento para obtener cada vez menos mejora en la calidad del resultado. Esta circunstancia

DOI: 10.21814/lm.11.2.285

representa una motivación para explorar formas híbridas de etiquetadores probabilísticos con reglas desarrolladas de manera manual o, al menos, un proceso posterior al etiquetado que corrija algunos de los errores. En el pasado se exploraron alternativas en este sentido, con algoritmos que complementaban sistemas de reglas con análisis probabilístico (Brill, 1992, 1995), pero no representan una tendencia actual, al menos no en este campo.

Está claro, en cualquier caso, que la tasa de error de los etiquetadores se ha ido reduciendo progresivamente. Sin embargo, y a pesar de estas mejoras, al día de hoy los problemas de los etiquetadores persisten. La desambiguación sigue siendo un desafío, ya que en muchos casos depende del sentido de la oración o incluso de la intención del emisor, con todos los matices y sutilezas que puedan darse (la ironía, el humor, etc.). Otro desafío en el análisis morfosintáctico es el de las unidades no diccionarizadas (out-of-vocabulary units; en adelante, UND), es decir, formas cuyo lema no es posible reconocer porque no está en el diccionario o corpus de entrenamiento del etiquetador. Esta es una dificultad especialmente frecuente cuando se trabaja en terminología, en que existe gran innovación léxica. Todo ello, además, tiene como consecuencia que no es posible proporcionar ni la etiqueta ni el lema de la forma analizada. Aplicar al procesamiento automático de textos especializados analizadores morfosintácticos que no cuentan con información morfológica específica de los términos genera una mayor tasa de error, lo que afecta tareas posteriores como la extracción de terminología. El problema se agudiza en particular en el caso de lenguas morfológicamente ricas como las romances.

Los Cuadros 1 y 2 ilustran la problemática. Se trata de fragmentos del resultado del análisis morfosintáctico de un corpus especializado (cf. Sección 3.4) después de haber sido sometido al analizador morfosintáctico TreeTagger (Schmid, 1994), que produce lema y categoría gramatical de las palabras ingresadas, y al analizador UDPipe (Straka & Straková, 2017), que además de esto ofrece un análisis sintáctico completo (nivel de análisis que para los fines del presente estudio no hemos tenido en cuenta). En general, se aprecia una mejora en la calidad de la lematización del segundo respecto al primero, pero en ambos casos es fácil advertir también la persistencia de errores cuando se trabaja en corpus especializados. En ambos casos se observa la lematización errónea de unidades léxicas desconocidas para el etiquetador, como *dopaminérgicos* en un caso y *micro-diálisis* y *demuestra* en el otro.

Forma	**POS-tag**	**Lema**
Existen	VLfin	existir
grupos	NC	grupo
de	PREP	de
fármacos	NC	fármaco
que	CQUE	que
tienen	VLfin	tener
afinidad	NC	afinidad
hacia	PREP	hacia
los	ART	el
diversos	QU	diversos
receptores	NC	receptor
dopaminérgicos	ADJ	UNKNOWN
.	FS	.

Cuadro 1: Resultado con TreeTagger. Se advierte error en la lematización de la forma *dopaminérgicos*.

Forma	**POS-tag**	**Lema**
Es	AUX	ser
así	ADV	así
como	SCONJ	como
Gingrich	PROPN	Gingrich
,	PUNCT	,
in	NOUN	in
vivo	ADJ	vivo
con	ADP	con
micro-diálisis	NOUN	micro-diálisi
demuestra	VERB	demuestro
liberación	NOUN	liberación
de	ADP	de
DA	PROPN	Da
en	ADP	en
el	DET	el
N	PROPN	N
Acc	PROPN	Acc

Cuadro 2: Resultado con el etiquetador UDPipe. Se advierte error en la lematización de las formas *micro-diálisis* y *demuestra*

En este contexto, el presente trabajo tiene por objetivo proponer la formalización de las reglas específicamente para la pluralización en castellano. La propuesta no pretende resolver todos los problemas; sin embargo, solo reconocer la flexión de número en sustantivos y adjetivos, como un proceso posterior al etiquetado, ya implicaría una mejora sustancial de la calidad del resultado. El método tiene la forma de una cascada de reglas de sustitución y expresiones regulares para restituir el lema de formas del plural en sustantivos y adjetivos. La propuesta se centra en el caso específico de los términos especializados porque es

el campo en el que el problema de las UND tiene mayor incidencia. En una serie de ensayos en un corpus especializado compuesto por artículos de investigación de una revista de neuropsiquiatría, obtuvimos mejoras significativas en la calidad del etiquetado de TreeTagger y de UDPipe para el caso específico del reconocimiento del plural.

La siguiente sección proporciona los antecedentes teóricos asociados al problema de las UND en el PLN en general, el POS-tagging en castellano en particular y las normas de pluralización en esta lengua, foco de interés de la presente investigación. Posteriormente, la Sección 3 presenta la propuesta metodológica que incluye la descripción del algoritmo diseñado y las reglas construidas para la lematización de las UND. La Sección **??** expone los resultados de una evaluación en el corpus de neuropsiquiatría. Finalmente, la Sección **??** presenta nuestras conclusiones y sugerencias para trabajo futuro. Hemos implementado este algoritmo en un prototipo en lenguaje Perl para que funcione como posproceso del etiquetado. Un demostrador del prototipo y su código fuente están disponibles en el sitio web del proyecto[1].

2 Marco Teórico

2.1 El problema de las unidades no diccionarizadas

El problema de las UND ha sido explorado en distintas áreas del PLN, como la adquisición léxica y el reconocimiento de habla (automatic speech recognition), donde las UND se presentan normalmente en forma de nombres propios de persona o lugar, pero también en unidades del vocabulario general, ya que este se encuentra en permanente cambio (Manning & Schütze, 1999; Bazzi & Glass, 2002; Bazzi, 2002; Parada et al., 2011; Qin, 2013). En el caso del reconocimiento de habla, el problema de las UND es particularmente agudo porque multiplica los errores en el reconocimiento de palabras vecinas: "When encountering an OOV word, the recognizer will incorrectly recognize the OOV word with one or more similar sounding in-vocabulary (IV) words. In addition, OOV words also affect the recognition performance of their surrounding IV words" (Qin, 2013, p. 3).

En el caso del POS-tagging, las UND han sido una de las primeras dificultades a superar: "Unknown words are a major problem for taggers, and in practice, the differing accuracy of different taggers over different corpora is often mainly determined by the proportion of unknown words" (Manning & Schütze, 1999, p. 351). Algunas propuestas para resolver el problema incluyen recurrir a pistas morfológicas calculando las probabilidades de combinación de raíces, sufijos y categorías morfológicas: "If we have never seen the word 'rakishly', then knowledge that 'ly' typically ends an adverb will improve our accuracy on this word –similarly, for 'randomizing' [etc.]" (Charniak et al., 1993, p. 787).

Manning (2011) calculó en 4.5 % el porcentaje de error que las UND representan respecto de la totalidad de errores cometidos en inglés, pero hay que suponer que la tasa de error siempre será más alta en el caso de los corpus especializados y en lenguas morfológicamente ricas. En castellano, no conocemos datos del porcentaje que representan las UND frente al total de los errores cometidos por los etiquetadores, pero esto variará de la misma manera en función de la naturaleza del texto analizado.

El mismo Manning (2011) y otros autores (*cf.* Biemann, 2006) propusieron nuevas formas de pensar el problema, entre las que se reconoce la aplicación de medidas de similitud distribucional para comparar las UND con las unidades que sí están en el vocabulario. Asimismo, la idea del modelamiento morfológico en la línea de Charniak et al. (1993) también ha sido explorada en distintas investigaciones posteriores para el tratamiento de las UND (Adams et al., 1994; Creutz et al., 2007). Una de las propuestas que se reconoce como la más reciente es la aplicación de las representaciones vectoriales de las secuencias de caracteres al interior de las palabras, secuencias transportadoras de información morfosintáctica (Santos & Zadrozny, 2014; Ling et al., 2015).

2.2 El POS-tagging en castellano

Al igual que en el caso del inglés y de otras lenguas, los primeros etiquetadores morfosintácticos para el castellano eran sistemas basados en reglas codificadas de manera manual, como sería el caso de Moreno & Goni (1995). Pero eventualmente fueron también los algoritmos probabilísticos los que fueron ganando popularidad. Entre estos destaca el ya mencionado TreeTagger (Schmid, 1994), basado en aprendizaje automático que está entre los más utilizados actualmente, al menos en el caso de las lenguas romances (Allauzen & Bonneau-Maynard, 2008; Parra Escartín & Martínez Alonso, 2015). Otras propuestas incluyen sistemas híbridos que combinan análisis probabilístico con sistemas de reglas manualmente codificadas, como Freeling (Carreras et al., 2004).

[1] http://www.tecling.com/pullpos

La lematización y el etiquetado morfosintáctico han cobrado nuevo impulso en el último tiempo, tanto en castellano como en otras lenguas. La aparición de nuevos productos parece sugerir que se producirán cambios en el escenario de los etiquetadores y que el análisis de dependencias sintácticas puede tener un rol en la desambiguación de categorías gramaticales. El POS-tagger que propuso Manning (2011) está basado en un clasificador que funciona con modelos de máxima entropía (Maximum Entropy models) y está disponible para el castellano entre otras lenguas, aunque no lematiza. El proyecto de las 'IXA pipes' (Agerri et al., 2014) contiene un POS-tagger basado en el mismo tipo de modelo, pero a diferencia del anterior sí incluye lematización, para lo cual utiliza un diccionario de 600.000 entradas que a su vez es expandido mediante la herramienta de análisis morfológico Morfologik (Miłkowski, 2010). Han aparecido recientemente distintas propuestas de etiquetadores que pueden aplicarse al castellano, como MateTools (Bohnet & Nivre, 2012) o spaCy (Honnibal, 2016; Honnibal & Johnson, 2015). En particular destacamos el sistema Lemming (Müller et al., 2015) porque incorpora algunas de las ideas que se presentan también en este artículo, tales como observar la frecuencia de aparición en corpus (Wikipedia en el caso de Lemming) para ponderar la validez de un candidato a lema. Sin embargo, su enfoque es distinto al nuestro, entre otras razones porque utilizan recursos externos, tales como el diccionario ASPELL.

Al margen de lo anterior, la razón fundamental de la renovación del interés por los etiquetadores viene dada por los nuevos avances en el campo de las redes neuronales, ya que dada su naturaleza también pueden aplicarse al castellano. En particular, el modelo Bi-LSTM (*Bidirectional long short-term memory*, un subtipo de redes neuronales recurrentes) es el que actualmente está dando mejores resultados (Ling et al., 2015; Plank et al., 2016) al aplicarse a la representación vectorial no ya de palabras (*word embeddings*), sino de fragmentos inferiores a la palabra (*subword* o *subtoken representations*). Trabajar al nivel inferior a la palabra resulta ser la clave para la lematización de las UND porque se puede generalizar a partir de la información morfológica de la palabras que sí son conocidas.

Dentro de la familia de algoritmos a los que se ha hecho referencia, destacamos el ya mencionado UDPipe (Straka et al., 2016; Straka & Straková, 2017), uno de los sistemas más recientes y con mejor desempeño tanto en etiquetado morfológico como en análisis sintáctico, que resultó el mejor entre 26 participantes de la *CoNLL 2018 UD Shared Task* (Straka, 2018). Cabe destacar, además, que ya es posible aplicarlo a una amplia variedad de lenguas.

Una evaluación extensiva de todos los sistemas existentes escapa a los límites del presente artículo. Sin embargo, ya un examen superficial de los distintos etiquetadores revela que aún existe amplio margen de mejora. Cuando los autores informan un 99 % de precisión, es necesario tener en cuenta, como ya se advirtió, que en corpus especializado el desempeño puede ser muy inferior. Además, por regla general, cuando se informan estos porcentajes de precisión, esto se hace teniendo en cuenta la precisión por token ("token ratio"), es decir, que se incluyen en el conteo aquellos tokens que solo tienen una etiqueta y una lematización. Incluso si se confirmaran estos diagnósticos optimistas sobre el desempeño de los etiquetadores, dos o tres errores cada cien palabras resultarán en una proporción mucho mayor de oraciones mal analizadas ("per-sentence ratio").

2.3 La pluralización en castellano

El plural en castellano es relativamente sencillo en comparación con otras lenguas que cuentan con una variación numérica como el dual, para dos elementos, o el paucal, para un grupo reducido de elementos (Dixon, 2009). Sin embargo, y a pesar de la aparente simplicidad del sistema castellano, la flexión de número ha sido un fenómeno que, en conjunto con la flexión de género, ha suscitado la atención de muchos gramáticos (Sánchez Corrales, 1994).

En castellano, los adjetivos tienen flexión de género y número y los sustantivos solo aparecen en su forma singular o plural, más allá de casos específicos como los nombres de cargos o profesiones, que también presentan flexión de género. En el plural, la flexión responde a una serie de restricciones que considera, entre otros aspectos, la estructura fonológica de la forma singular (Cedeño et al., 2014). De esta manera, el sonido, la acentuación y la estructura silábica han sido aspectos comunes en gran parte de las propuestas que explicitan una variedad de especificaciones en las normas para la formación de los plurales.

Dentro de las primeras formalizaciones que podemos reconocer para la flexión de número se encuentra la *Gramática de la lengua castellana* (De Nebrija, 1492). En esta publicación se utilizan ya los términos de singular y plural y se proponen las normas que permiten formar los plurales, básicamente el uso de las terminaciones *-s* y *-es*. Posteriormente, la *Ortografía española* (Real Academia Española, 1741), la *Gramática*

de Andrés Bello (1847), la *Gramática de la lengua castellana* (Real Academia Española, 1920) y, más recientemente, la Nueva gramática de la lengua española (Real Academia Española, 2009) han mantenido estas normas en lo esencial.

Alemany (1920) también mantiene en esencia las mismas normas, pero las complejiza al advertir que las marcas de la pluralización dependen de ciertas características del lema principal. Estas suelen asociarse a la combinatoria de letras en el término de la palabra, su sílaba átona o tónica y su dependencia gramatical en la sintaxis. Sin embargo, este autor aborda también el tema de la pluralización de los compuestos, a los que distingue entre perfectos e imperfectos. El primero corresponde a aquellos sustantivos que combinan dos elementos nominales (adjetivos+sustantivos, sustantivos+sustantivos, adjetivo+adjetivo, etc.) conformando una nueva forma nominal (*ferrocarril, portafusil*). Esta nueva construcción aceptaría el plural en su segundo término. Por su parte, los compuestos imperfectos son las formas que se componen por más de una palabra ortográfica, conformando así unidades poliléxicas. Dentro de estos casos, la pluralización interna se encarga de marcar solo el primer componente, que además es el núcleo del término (*hombres rana, cartas suicida, palabras clave*). El tema ha sido desarrollado por autores contemporáneos como Moyna (2011) o de León (2015). En el presente artículo, sin embargo, nos restringimos a la construcción de los plurales de unidades monoléxicas y fuera de contexto ("types").

Otros fenómenos dignos de atención respecto a la pluralización en español son 1) el caso del morfema ∅ (morfema cero) en sustantivos como *lunes, martes* o *paréntesis*, ya que en su forma singular acaban en *-s* y que no son agudos y reconocen el mismo uso tanto en su forma singular como plural (Real Academia Española, 1920); y 2) el de los *pluralia tantum*, que remiten generalmente a un solo objeto pero compuesto (*tijeras, pinzas, pantalones*).

La revisión de la bibliografía permite identificar que existen numerosos trabajos enfocados en la descripción de la pluralización español (Stockwell et al., 1965; Saporta, 1965; Foley, 1967; Alcina & Blecua, 1975; Hernández Alonso, 1984; Ambadiang, 1999, entre otros), y todos coinciden en que las marcas que afectan a esta flexión se identifican con *-s* y *-es*, dependiendo de los aspectos estructurales del lema y de otras características fonológicas de la palabra. Como tendencia general, se puede identificar que para la forma del singular terminada en vocal, suele añadirse *-s* a su forma plural; mientras que para la forma singular que acaba con consonante, es más recurrente añadir *-es* para su forma plural.

Entre las obras consultadas, nos hemos decantado por la sistematización que ofrece el *Diccionario panhispánico de dudas* (Asociación de Academias de la Lengua Española, 2005) para la implementación de las reglas. La obra no es del todo reciente y algunas de las reglas son de naturaleza normativa, lo que nos obligaría en el futuro a hacer algunos ajustes. Sin embargo, nos pareció una buena síntesis al menos para esta etapa inicial del proyecto. A continuación se presentan los 10 casos principales entre los 17 que regularizan la flexión del plural, ya que los restantes corresponden a particularidades asociadas a latinismos, notas musicales, plural de nombres de las letras del abecedario, abreviaturas y símbolos, entre otros.

(a) Sustantivos y adjetivos terminados en vocal átona o en *-e* tónica. Forman el plural con *-s* (*casas, estudiantes, taxis, planos, tribus, comités*).

(b) Sustantivos y adjetivos terminados en *-a* o en *-o* tónicas. forman el plural únicamente con *-s*. (*sofás, rococós, dominós*).

(c) Sustantivos y adjetivos terminados en *-i* o en *-u* tónicas. Admiten generalmente dos formas de plural, una con *-es* y otra con *-s*, aunque en la lengua culta suele preferirse *-es*. (*bisturíes/bisturís, carmesíes/carmesís, tabúes/tabús*).

(d) Sustantivos y adjetivos terminados en *-y* precedida de vocal. Forman tradicionalmente su plural con *-es*. (*rey/reyes; ley/leyes; buey/bueyes*). Los sustantivos y adjetivos con esta misma terminación que se han incorporado provienen por lo general de otras lenguas. Aplica para este caso que la *y* del singular pasar a escribirse *i* (*espray/espráis; yóquey/yoqueis*).

(e) Voces extranjeras terminadas en *-y* precedida de consonante. Deben adaptarse gráficamente al español sustituyendo la *-y* por *-i*. Su plural se forma añadiendo una *-s*. (*dandis, pantis, ferris*).

(f) Sustantivos y adjetivos terminados en *-s* o en *-x*. Si son monosílabos o polisílabos agudos, forman el plural añadiendo *-es* (*tos/toses; vals/valses, fax/faxes; compás/compases; francés/franceses*). En el resto de los casos, permanecen invariables (*crisis/crisis; tórax/tórax; fórceps/fórceps*).

(g) Sustantivos y adjetivos terminados en *-l, -r, -n, -d, -z, -j*. Si no van precedidas de otra consonante, forman el plural con *-es*. Los extranjerismos deben aplicar la misma regla (*dócil/dóciles; color/colores; pan/panes; césped/céspedes; cáliz/cálices; reloj/relojes*).

(h) Sustantivos y adjetivos terminados en consonantes distintas de *-l, -r, -n, -d, -z, -j, -s, -x, -ch* pluralizan en *-s*. Esta norma incluye onomatopeyas o voces procedentes de otros idiomas (*crac/cracs; zigzag/zigzags; esnob/esnobs; chip/chips; mamut/mamuts; cómic/cómics*).

(i) Sustantivos y adjetivos terminados en *-ch*. Procedentes todos ellos de otras lenguas, o bien se mantienen invariables en plural (*(los) crómlech, (los) zarévich, (los) pech*), o bien hacen el plural en *-es* (*sándwich/sándwiches; maquech/maqueches*).

(j) Sustantivos y adjetivos terminados en grupo consonántico. Procedentes todos ellos de otras lenguas, forman el plural con *-s* salvo aquellos que terminan ya en *-s*, y que siguen la regla (f) (*gong/gongs; iceberg/icebergs; récord/récords*). Se exceptúan de esta norma las voces *compost, karst, test, trust y kibutz*, que permanecen invariables en plural, porque añadir *-s* en estos casos daría lugar a una secuencia de difícil articulación en castellano.

Creemos que una sistematización e implementación computacional de estas reglas puede representar una ayuda para identificar el plural de lemas no incluidos en los diccionarios, mejorando así la calidad del trabajo en el procesamiento de terminología especializada.

3 Metodología

La propuesta metodológica está basada en un sistema que se puede aplicar con posterioridad al POS-tagger para detectar los casos de error con los plurales y corregirlos asignando el lema correspondiente del plural encontrado, si es que se trata realmente de una forma plural, porque si la unidad encontrada no es un plural, se asignará la misma forma como lema. Describimos a continuación los pasos del algoritmo que hemos diseñado para esta tarea. En primer lugar, la Sección 3.1 describe la adquisición léxica de un corpus de referencia de gran tamaño. La Sección 3.2 describe la implementación de las reglas de pluralización en castellano descritas ya en la Sección 2.3, pero utilizando la frecuencia de aparición de los elementos observada en el corpus de referencia como fundamento para la toma de decisión. La Sección 3.3 describe el proceso de extracción de parejas singular-plural por medio de la aplicación de las reglas de pluralización al corpus de referencia.

El proceso de adquisición léxica debe realizarse solamente una vez, ya que al completarse esta primera acción, el algoritmo pasa a la segunda fase, que es la que puede realizarse n veces. Esta segunda fase corresponde al análisis de un corpus especializado en particular (Sección 3.4), de ahora en adelante designado como el corpus-objetivo, que ha sido previamente etiquetado con un POS-tagger. Sobre este corpus se procederá con la detección y corrección de errores.

3.1 Adquisición de un formario amplio

El punto de partida es la generación automática de un amplio formario del castellano, entendiendo por ello un listado de formas léxicas distintas. Para nuestros experimentos utilizamos un fragmento del corpus de castellano EsTenTen (Kilgarriff & Renau, 2013), compuesto por páginas web descargadas de manera aleatoria. El tamaño de la muestra utilizada es de aproximadamente dos mil millones de palabras sobre un total de 10^{10} que tiene el corpus.

El EsTenTen se ofrece ya etiquetado, pero nuestra metodología no utiliza esas etiquetas y por eso hemos procedido a eliminarlas dejando solo el texto plano (*cf.* comentarios al respecto en la Sección 5 al proyectar las posibilidades de trabajo futuro). En esta etapa del proceso, el algoritmo produce una tabla como la que se muestra en el Cuadro 3, con todo el vocabulario del corpus asociado a su frecuencia de aparición, reteniendo solamente aquellos elementos con frecuencia mínima $>= u$ ($u = 5$). El umbral es arbitrario pero obedece a cuestiones prácticas: un umbral muy bajo haría que el tamaño del formario sea demasiado grande como para manejarlo con facilidad, mientras que uno muy alto reduciría la cobertura del sistema. El tamaño del formario obtenido mediante este procedimiento es de 1.054.411 registros.

3.2 Reglas de pluralización

A partir de lo revisado sobre las normas de pluralización en castellano derivamos una serie de reglas para reconocer la relación entre un singular y un plural. En los casos de las formas más frecuentes, utilizamos simples reglas de sustitución porque consideramos que son un apuesta segura.

Forma	**Frecuencia**
...	...
zigzagueaba	13
zigzagueaban	7
zigzagueado	8
zigzagueamos	5
zigzaguean	42
zigzagueando	140
zigzagueante	258
zigzagueantes	108
zigzaguear	63
zigzagueo	54
zigzagueos	40
zigzagues	5
zigzagueó	6
...	...

Cuadro 3: Fragmento del formario del EsTenTen

Si se trata de terminaciones típicamente adjetivales o de nombres de profesiones, la forma singular remite directamente al masculino. En el resto de los casos, conservamos la forma terminada en *a*, ya que la decisión final se tomará en un momento posterior del análisis (Sección 3.4.3). En total, desarrollamos 55 reglas de este tipo. Los siguientes son algunos ejemplos:

```
-áceos     =>   -áceo
-ares      =>   -ar
-ari[ao]s  =>   -ario
-ciones    =>   -ción
-eces      =>   -ez
-eones     =>   -eón
-ices      =>   -iz
-idades    =>   -idad
-ines      =>   -ín
-siones    =>   -sión
```

Para los casos de palabras que no cumplen con este tipo de morfología, utilizamos cascadas de expresiones regulares que nos permiten establecer generalizaciones más amplias según la normativa de la pluralización del español. Para ello, tomamos como base la sistematización del *Diccionario panhispánico de dudas* expuesta en la Sección 2.3.

3.3 Derivación de parejas singular-plural a partir del formario extraído del corpus

A partir del formario obtenido en la Sección 3.1 y aplicando las reglas de la Sección 3.2, generamos un recurso léxico en forma de parejas singular-plural. Esta tabla utilizará luego el algoritmo para el análisis de cada corpus-objetivo.

Solo se admite una pareja en este diccionario si la proporción entre la frecuencia del singular y el plural se encuentran dentro de unos límites que se asignaron de manera empírica. Fundamentamos esta decisión en la expectativa de encontrar determinado equilibrio entre las formas singular y plural. Por ejemplo, si una palabra no aparece casi nunca en plural, se asume que en realidad no pluraliza. Definimos en (1) la razón r entre la frecuencia observada de una forma plural $c(p)$ sobre la frecuencia de la forma singular $c(s)$, y en (2) una función binaria $a(p,s)$ que decide si el algoritmo admite o no la pareja (p,s).

$$r(p,s) = \frac{c(p)}{c(s)+1} \tag{1}$$

$$a(p,s) = \begin{cases} 1 & x > r(p,s) < z \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

De esta forma, la función $a(p,s)$ controla que la razón $r(p,s)$ se encuentre dentro de la banda de tolerancia definida por los parámetros x y z ($x = 0{,}001 \wedge z = 120$). Esto impide que se produzca una disparidad excesiva entre la frecuencia del singular y del plural. En el primer caso se evitaría, por ejemplo, la unión de *algos*, con frecuencia 132, vs. *algo*, con frecuencia 1.186.957. En el caso opuesto, se evita por ejemplo la unión de *vívere*, con frecuencia 8, y *víveres*, con frecuencia 2.427.

A modo de orientación para la estimación de los parámetros x y z, incluimos la Figura 1, en la que se puede apreciar el porcentaje de error que se obtiene tomando muestras aleatorias de 50 parejas en distintos intervalos. Tal como se puede apreciar en esa Figura, los valores de precisión disminuyen hacia los extremos mayor y menor.

Mediante este procedimiento se generó un total de 143.159 parejas (un fragmento se muestra en el Cuadro 4). Las parejas se componen principalmente de sustantivos y adjetivos, pero no porque el algoritmo utilice la información proporcionada por el etiquetador en el corpus de referencia. La razón, en cambio, es que son estas categorías gramaticales las que cumplen con las reglas de pluralización, y para una forma singular se ha encontrado efectivamente una coincidencia en el formario con una forma plural. Esto no es posible en el caso de otras categorías gramaticales, aunque sí se introducen errores que corresponden a formas en otras lenguas o nombre propios, pero la mayor parte de estos errores se corrigen en la etapas del análisis de un corpus-objetivo.

Plural	**Singular**	**Frec 1**	**Frec 2**	**r(p,s)**	**a(p,s)**
...	...	...	...	...	
luís	luí	12880	54	238.5185185	0
extremis	extremi	2124	9	236	0
holmes	holme	7073	30	235.7666667	0
escalopines	escalopín	244	11	22.1818182	1
fotomecánicas	fotomecánico	24	16	1.5	1
claroscuristas	claroscurista	11	13	0.8461538	1
antieconómicas	antieconómico	68	161	0.4223602	1
linfocíticas	linfocítico	8	20	0.4	1
fototérmicos	fototérmico	9	24	0.375	1
autoproclamaciones	autoproclamación	7	69	0.1014493	1
esquizofrénicas	esquizofrénico	92	1067	0.0862231	1
jurisprudencias	jurisprudencia	62	18092	0.0034269	1
moderaciones	moderación	36	10547	0.0034133	1
nazismos	nazismo	11	4523	0.002432	1
comos	como	651	10574252	0.0000616	0
madrids	madrid	17	1239084	0.0000137	0
relacionares	relacionar	5	425566	0.0000117	0
...	...	...	...	...	

Cuadro 4: Fragmento del listado de parejas singular-plural. El valor de la columna *a(p,s)* define si el algoritmo aceptará o no la pareja propuesta

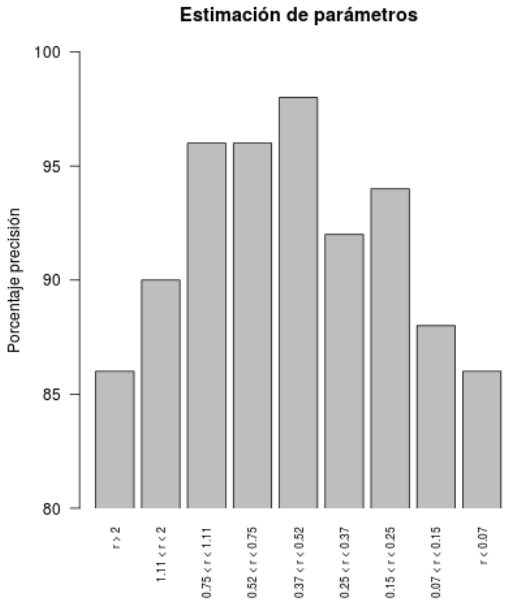

Figura 1: Diagrama de barras con valores orientativos para estimar los parámetros x y z

3.4 Análisis de un corpus-objetivo

Ya con los recursos generados en la etapa anterior (tabla de parejas singular-plural + reglas de pluralización), pasamos ahora a describir cómo es la metodología de análisis de un corpus especializado en particular, el que denominamos corpus-objetivo.

3.4.1 Etiquetado y posprocesamiento del corpus-objetivo

Tal como hemos descrito en las secciones anteriores, proponemos una secuencia de pasos en los que, en primer lugar, se somete el corpus-objetivo a un etiquetado con la herramienta habitual. A continuación, se reenvía el resultado de esta operación al script que implementa nuestro algoritmo de corrección del plural.

Por cada forma del corpus, este sistema puede detectar casos en los que la palabra termina en *-s*. Llamaremos W_p a una palabra que cumpla con esta condición, como por ejemplo *anticoagulantes*. W_p será sometida a una función binaria $f(W_p)$ (3) que determinará si W_p tiene o no lematización.

$$f(W_p) = \begin{cases} 1 & W_p \rightarrow W_s \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Para cada W_p existen solo dos resultados posibles: la lematización correspondiente (1) o su rechazo como forma plural (0). El valor 1 se devuelve en el caso de cumplirse $W_p \rightarrow W_s$, es decir, que para la supuesta forma plural W_p se ha encontrado una forma singular satisfactoria W_s (por ejemplo, *anticoagulante*). Por el contrario, si $f(W_p) = 0$, se considerará un caso de no plural y no presentarán flexión de número (sería el caso, por ejemplo, de una forma como *apoptosis*).

A continuación se explica en detalle este proceso de decisión.

3.4.2 Primeras operaciones de descarte

Antes de someter a una determinada forma a la batería de análisis para encontrar su forma singular, aplicamos primero una serie de filtros que reducen considerablemente la tasa de error de entrada del algoritmo. Estos filtros se aplican primero por eficiencia computacional.

En primer lugar, aplicamos una expresión regular para detectar y descartar formas verbales. En castellano encontramos terminaciones en *-s* en formas verbales de distintos tiempos y modos, como en segunda persona del singular (*vienes, cantes, salgas*, etc.) y en primera persona del plural (*venimos, cantamos, salgamos*, etc.) así como en los pronombres enclíticos (*cantarles, llamábales, arreglándoselas*, etc.).

Optamos por detectar estas terminaciones por medio de expresiones regulares. Si de esta forma se consigue una coincidencia, entonces el script cambia la etiqueta NC (o ADJ) del etiquetador reemplazándola por la categoría verbo. En esta ocasión no hemos resuelto recuperar el infinitivo del verbo en cuestión, ya que consideramos que esa es otra tarea que dejamos para un trabajo futuro.

El siguiente filtro es el de los elementos que típicamente no son plurales en castellano, y que aparecen en el caso de los corpus especializados con mayor o menor frecuencia dependiendo del dominio:

```
(us|[oipeas][lrxtfsd]is|[íó][dtf][aei][dnl]esus|[oipeas]
[lrxtfsd]is|[íó][dtf][aei][dnl]es)$
```

Al igual que en el caso de los verbos, las unidades que coincidan con este modelo morfológico serán clasificadas como elementos no plurales, y se asignará como lema la misma forma encontrada.

Un tercer filtro es el de las formas en inglés. Ocurre con frecuencia, particularmente cuando se trabaja con corpus especializados, que fragmentos del corpus están escritos en otra lengua, usualmente en inglés, debido a los resúmenes, las palabras clave, las citas y los títulos bibliográficos que aparecen en los artículos especializados. Idealmente, deberíamos implementar un detector de fragmentos en inglés, pero en lugar de eso nos pareció más sencillo aplicar una nueva regla de filtrado para detectar la morfología característica del plural en inglés, y que m

```
(ys[ei]s|nces|tr?ics|ions|[nl]ess|ties|tions|en[td]s|[ae]
cts|sters|oids|ishes|ous|ers|[csr]ies|ants|[aoe]ss)$
```

Si una unidad léxica terminada en *-s* W_p coincide con estas terminaciones se le asigna la etiqueta "Palabra en inglés", y queda así también fuera del análisis. Si, en cambio, W_p supera estos filtros, se registra su frecuencia en el corpus-objetivo y es sometida al resto del proceso.

3.4.3 Consulta a la tabla de parejas singular-plural y procedimientos auxiliares

La tabla de parejas plural-singular, cuyo fragmento se mostró en el Cuadro 4, es leída y cargada en memoria como *hash table* al principio del proceso y ofrece dos informaciones: por un lado, la forma singular del elemento en cuestión y, por el otro, la frecuencia de aparición de cada elemento en el corpus de referencia.

Definimos en (4) una función $m(W_f)$ exclusivamente para los casos en los que encontramos que la forma singular termina con la letra *a*, lo que podría ser indicativo de que se trata de una forma en femenino tal como sucede normalmente en el caso de los adjetivos. La manera de determinar esto es comprobar si existe en el corpus de referencia la forma masculina correspondiente.

$$m(W_f) = \begin{cases} 1 & \text{máx}\,(c(W_o), c(W_s)) > u \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

La función $m(W_f)$ toma en consideración entonces la frecuencia de aparición en el corpus de referencia de las formas masculina y femenina del término en cuestión. Así, si W_f es una palabra terminada en *a*, definimos un candidato a forma masculina que puede ser con o sin *o* (W_o y W_s, respectivamente) en función de la frecuencia de aparición en el corpus de referencia ($c(W_o)$ y $c(W_s)$). Para ello se exploran dos posibilidades:

1. Reemplazar la última letra *a* por *o*.
2. Eliminar directamente la letra *a*.

Si alguna de las dos formas resultantes tiene una frecuencia igual o superior al umbral de frecuencia u ya definido anteriormente, entonces $m(W_f)$ decide "masculinizar" la forma terminada en *-a*. Si ambos cumplen esta condición, se elige el más frecuente. En cambio, si $m(W_f) = 0$, se asume que no se trata de un adjetivo y por tanto debe lematizarse con la forma terminada en *-a*. Naturalmente, este será el caso de sustantivos femeninos, como *rosa*, que debe ser lematizado de esa manera, pero también casos como el de *pediatra*, que termina en *-a* pero no es un sustantivo femenino y su lema también permanece inalterado.

3.4.4 Estrategias de repliegue

A pesar del gran tamaño del corpus de referencia utilizado, es normal observar que el corpus-objetivo contiene formas léxicas que no aparecen en el de referencia. La primera estrategia de repliegue en este caso es reintentar el proceso, pero esta vez reemplazando el corpus de referencia por el mismo corpus-objetivo. Particularmente, esto permite resolver casos de terminología propia del dominio.

En los casos en que una determinada palabra no se ha observado en ninguno de los dos corpus y no ha podido ser clasificada con ninguna de las estrategias anteriores, aplicamos entonces la segunda estrategia de repliegue, que consiste en el análisis de la posible prefijación de la palabra para los casos de un elemento léxico que sí es conocido, pero que ha sufrido un proceso de derivación mediante la adición de prefijos. Aplicamos en estos casos una regla de determinación de prefijos, para lo cual utilizamos la siguiente lista de prefijos en castellano (Real Academia Española, 2006):

```
^(ad|ana|anti|auto|cata|co|cuasi|de|di|em|en
|entre|ex|extra|hetero|hiper|hipo|in|infra
|inter|macro|meta|micro|mono|neuro|para|per
|poli|pos|post|pre|pro|p?seudo|psico|radio
|re|sin|son|sub|super|tele|tran?s|ultra)
```

No es una lista exhaustiva, pero reúne los prefijos de alta productividad en dominios de especialidad. Si con la ayuda de esta lista podemos separar la palabra en dos partes y descubrir la unidad léxica conocida, sometemos esa unidad al mismo proceso descrito en las secciones anteriores, con la única diferencia de que al lema resultante volvemos a añadir el prefijo encontrado. Así, por ejemplo, encontramos que la forma *antidopaminérgicos* no aparece con la frecuencia mínima en los corpus, pero luego de la eliminación del prefijo *anti-* se descubre una forma que sí es conocida (*dopaminérgicos*) y que sí admite un singular con nuestra metodología (*dopaminérgico*). De esta forma, se restituye el prefijo elidido anteriormente y se recupera de esta forma el singular *antidopaminérgico*.

Si esta segunda estrategia de repliegue también falla, es decir, si no estamos ante un caso de derivación por prefijos, entonces esta circunstancia nos obliga a dar cuenta de estos elementos e intentar también ofrecer un lema aunque se trate de una entidad puramente teórica. Cabe esperar que entren en esta categoría los errores de tipeo encontrados en el corpus. No hemos pretendido dar solución aquí el problema del error de tipeo porque consideramos que es una investigación diferente (*cf.* comentarios al respecto en la Sección 5, cuando mencionamos posibilidades de trabajo futuro). En lugar de esto, optamos por proporcionar un lema teórico debido que entre los errores de tipeo cabe esperar también la aparición de unidades terminológicas genuinas que aun no han sido registradas (neologismos).

Claramente se trata de una estrategia arriesgada en este último caso por lo que, cuando esto ocurre, añadimos a la lematización la etiqueta UNKNOWN para que el usuario sepa que el sistema aquí está "inventando" un lema cuya existencia no ha sido documentada. Esto se consigue aplicando la misma batería de reglas descrita en la Sección 3.2, pero ahora ya sin el apoyo empírico que fundamenta la decisión de clasificarlo como un lema genuino.

4 Resultados

El primer paso para la evaluación de los resultados fue constituir un corpus especializado para ser utilizado como corpus-objetivo. Con la autorización de la *Revista Chilena de Neuropsiquiatría*, conformamos un corpus-objetivo a partir una muestra de artículos aleatoria (800.000 palabras) desde su sitio web[2]. Transformamos el material a texto plano y, por medio de expresiones regulares, eliminamos resúmenes y palabras clave en inglés, y así quedó así conformado el corpus-objetivo[3]. Para nuestros experimentos, etiquetamos este corpus con TreeTagger, que arrojó un tokenizado de 805.624 líneas, y luego con UDPipe, que verticalizó el corpus en 797.812 tokens. Las diferencias entre ambos resultados se deben a las distintas estrategias de tokenización de cada programa. En particular, difieren en la forma de encapsular, por ejemplo, locuciones o marcadores discursivos como una sola unidad léxica (*en general, por ejemplo, sin embargo*, etc.).

Utilizamos estos dos etiquetadores por las razones expuestas en la Sección 2.2: principalmente, lo extendido que se encuentra su uso en la actualidad y el hecho de que representen estrategias de lematización distintas (probabilístico vs. neuronal). Sin embargo, como ya advertimos, no es el propósito de esta investigación hacer una evaluación exhaustiva de todos los etiquetadores actualmente existentes, ya que el hecho de que un etiquetador tenga mejor desempeño que otro no tiene relación directa con el método que propone-

[2] La revista se puede consultar en línea y los artículos son *open access*: http://www.sonepsyn.cl

[3] Las pruebas de desarrollo del algoritmo se hicieron en un proyecto de extracción de terminología en un corpus de biología en castellano. Las dificultades allí encontradas con el plural motivaron esta investigación.

mos. Dicho de forma más general, dado un corpus etiquetado con una calidad X, nuestro método producirá una cantidad Y de reducción del error ($1 - X$ en un intervalo comprendido entre 0 y 1).

El procedimiento para evaluación consistió entonces en aplicar nuestro script al resultado obtenido con cada etiquetador. Cada vez que nuestro script encuentra una forma terminada en *-s* se considera que esto es indicio de que podría tratarse de una forma plural. A partir de aquí, los siguientes resultados son posibles:

1. El algoritmo decide que la forma corresponde efectivamente a un plural. En este caso se propone el lema correspondiente en singular.
2. Decide que la forma no es un plural, por lo tanto deja como lema la misma forma.
3. Decide que la palabra está en inglés, por tanto es un elemento no analizable (el lema queda igual a la forma y se identifica con la etiqueta "Palabra en inglés").
4. Decide que la palabra es una forma verbal (no corresponde por tanto aplicar las reglas para la detección de plural y se identifica con la etiqueta "verbo").

Después de la aplicación de los etiquetadores y de nuestro script encontramos, en el caso de TreeTagger, un total de 9.763 formas distintas (types) terminadas en *-s*. En más de la mitad de estos casos (5.081) nuestro script modificó el lema originalmente asignado por el etiquetador. En el caso de UDPipe, en tanto, encontramos 9.576 types, de los cuales 2.364 tuvieron una modificación del lema por parte de nuestro script. Este primer dato ya arroja una estimación de la diferencia en el desempeño de los etiquetadores, ya que en el segundo caso fueron requeridas menos modificaciones. Es necesario tener en cuenta, sin embargo, que la discrepancia no significa necesariamente un error en la lematización. Es posible asignar lemas distintos y que ambos sean aceptables. El caso más frecuente sería el de los participios, que pueden lematizarse con un infinitivo (en la lectura de forma verbal) o bien como un adjetivo (dejando el participio en la forma masculino singular).

Para llevar a cabo la evaluación, tomamos muestras aleatorias de 600 casos de cada uno de los dos resultados, únicamente cuando el lema propuesto por nuestro script es distinto al lema producido por los etiquetadores, ya que suponemos que si el lema es el mismo entonces el resultado debe ser correcto. Hicimos un muestreo dividido por categorías: 300 casos de formas reconocidas como plural (que es el fenómeno que más nos interesa estudiar en esta investigación); 100 casos de formas no plurales; 100 casos de formas clasificadas por nuestro script como palabras en inglés, y finalmente 100 casos de formas verbales.

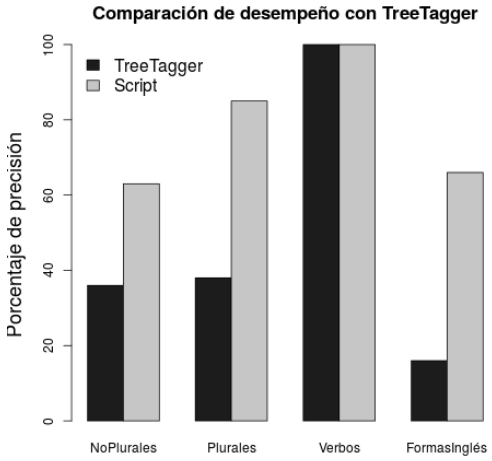

Figura 2: Diagrama de barras con la comparación del desempeño entre TreeTagger y el script

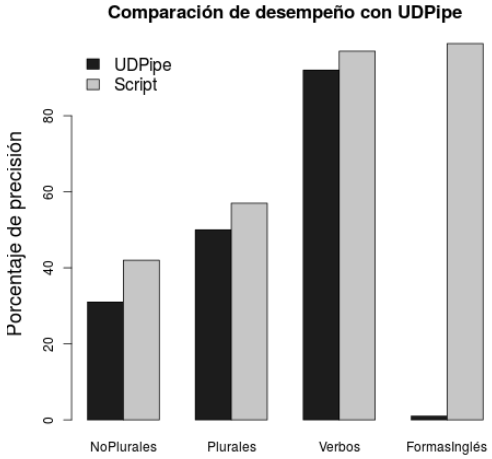

Figura 3: Diagrama de barras con la comparación del desempeño entre UDPipe y el script

Las figuras 2 y 3 muestran diagramas de barras que comparan el desempeño de nuestro prototipo con el de TreeTagger y UDPipe, respectivamente. Es necesario recalcar aquí que los porcentajes de precisión corresponden a las muestras de palabras terminadas en *-s* en las que existe una discrepancia entre el lema asignado por el etiquetador y el asignado por nuestro script. En ambas figuras se puede observar que cuando hay disparidad en la lematización, en general nuestro script tiende a ser más preciso, tendencia que se mantiene en las cuatro categorías de error analizadas.

En el caso de las formas que fueron clasificadas como plurales y fueron lematizadas, encontramos que, en la comparación con TreeTagger, un 85 % de las veces en que hubo disparidad con este etiquetador, el lema propuesto por el script fue correcto, frente a un 38 % de este etiquetador (los porcentajes no suman 100 porque, como ya se indicó, hay casos en que lematizaciones distintas son aceptables). Esta diferencia se reduce considerablemente en el caso de UDPipe. En parte, y como ya vimos, el desacuerdo con nuestro script es mucho menor. Pero cuando existe desacuerdo, el desempeño de nuestro script es mejor, aunque el margen es más reducido: 57 % de nuestro script frente a 50 % de UDPipe.

La mayoría de los errores cometidos por el script en esta categoría corresponden a palabras en inglés, con un 35 % (*caregivers, remarks, substances*, etc.). A veces los lemas que propone el script son igualmente correctos aunque estén en inglés, pero esto es irrelevante porque es un resultado casual, no parte del diseño, y debemos considerarlos por tanto como errores, ya que el sistema debería haberlos clasificado como formas en inglés. El porcentaje restante de errores lo conforman apellidos con el 11 % (como *Hodges, Hopkins, Duprés*, etc.), siglas y abreviaturas con el 5 % (*irss, mgrs, tlrs*, etc.), entre otros fenómenos. En cuanto a los resultados de la detección de no-plurales, el porcentaje de precisión en los casos en los que hay discrepancia con el etiquetador, la precisión fue de 63 % contra 36 % en el caso de TreeTagger, y 42 % frente a 31 % en el caso de UDPipe.

A modo de ilustración, el Cuadro 5 ofrece un fragmento de los resultados en la categoría "plurales". Allí, el único error registrado es de tipeo *cuaiïdades* [*sic*], que recibió el "lema" *cuaiïdad*, y que tenemos que marcar como error del procedimiento, ya que es consecuencia de apostar por una forma cuya existencia no está documentada en corpus. Se trata de un problema que podría tener solución con un algoritmo de corrección ortográfica, pero esto es, como ya hemos señalado, un problema distinto. En ninguno de los casos etiquetados como elementos desconocidos se encontraron errores de otro tipo. El Cuadro 6, por su parte, muestra resultados aleatorios en la categoría "NoPlurales". Allí, la columna "Error" muestra los errores que marcamos durante la revisión manual de los resultados. Tal como se puede apreciar, los errores cometidos son principalmente formas en inglés o nombres propios.

Los datos también permiten hacer observaciones sobre la comparación del desempeño entre los dos etiquetadores. En general se observa que UD-

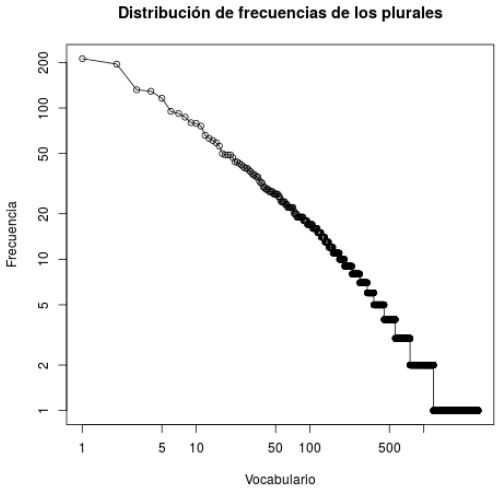

Figura 4: Distribución de frecuencias de los plurales encontrados (escala logarítmica en ambos ejes)

Pipe tiene mejor desempeño porque el porcentaje de mejora o de corrección de error que introduce nuestro script es menor. Sin embargo, hay ciertas categorías en las que tiende a cometer más errores que TreeTagger, que no intenta lematizar formas que no conoce (asigna el lema "unknown"). UDPipe sí lo hace y comete más errores, lo que se ve en particular en el caso de las formas en inglés, en donde el desempeño de UDPipe se deteriora notablemente.

Finalmente, la Figura 4 muestra la distribución de frecuencias de los elementos corregidos en el caso de TreeTagger. Como cabía esperar, la figura muestra que se trata de una distribución típicamente zipfeana.

5 Conclusiones

En esta propuesta metodológica hemos presentado una formalización de las reglas para la pluralización de sustantivos y adjetivos en el español y un algoritmo que ofrece una solución de alta efectividad para el problema de las UND. Consideramos que el sistema será de utilidad para el trabajo de los terminólogos en corpus especializado en castellano.

En este trabajo hemos optado por acotar el problema al caso del plural debido a su alta incidencia y su efecto perjudicial en tareas más avanzadas como la extracción de terminología o la revisión ortográfica en procesadores de texto. En ese sentido, nuestra investigación es de interés desde un punto de vista eminentemente práctico. Sin embargo, más allá de este aspecto práctico, hay que destacar la simplicidad del algoritmo,

Forma	Lema	Desconocido	Error
...	...		
hipersensibles	[hipersensible]	*	
escretoras	[escretor]	*	
anorectics	[PALABRA EN INGLÉS]		
fuentes	[fuente]	*	
cuaiidades	[cuaiidad]	*	!
perdonamos	[FORMA VERBAL]		
anatómicos	[anatómico]		
inexplicadas	[inexplicado]		
intangibles	[intangible]	*	
contraindicadas	[contraindicado]		
teriovenosas	[teriovenoso]	*	
existenciarios	[existenciario]		
fetales	[fetal]		
...	...		

Cuadro 5: Fragmento de salida del algoritmo para la categoría "plurales", con indicación de elementos desconocidos y eliminación de formas verbales y palabras en inglés

Forma	Error
...	...
sarcoidosis	
angus	
dermis	
epistaxis	
linfocitosis	
crisis	
periartritis	
rawlins	!
polyhidramnios	!
neurogénesis	
meningitis	
losephs	!
enuresis	
alcalosis	
...	...

Cuadro 6: Fragmento de salida del script para la categoría de "NoPlurales"

sobre todo en comparación con la complejidad de los modelos basados en redes neuronales. Esto tiene importancia práctica, la que se traduce en facilidad de uso y rapidez de ejecución, pero creemos que también tiene un atractivo metodológico y conceptual: el seguir el principio de parsimonia (*keep it small and simple*). Se requieren pocos conocimientos de programación para adaptar esta implementación a otra lengua, suponiendo que es posible sistematizar también reglas de formación de plural.

El presente trabajo sugiere, además, varias líneas de trabajo futuro, puesto que representa una invitación a enfrentar con un razonamiento similar otros tipos de error cometidos por los etiquetadores. Por ejemplo, una línea de investigación que consideramos complementaria y de gran interés sería la determinación del género de los sustantivos, que en este trabajo apenas hemos gestionado. Otra posibilidad en la misma línea sería corregir las etiquetas gramaticales que asignan los etiquetadores, tema que tampoco dejamos resuelto en este artículo, aunque ya damos algunas orientaciones sobre cómo podría hacerse al buscar las desinencias verbales, por ejemplo. Otro caso de esto último sería comprobar si una etiqueta de NC asignada por el etiquetador no debería ser en realidad ADJ cuando se observa que la palabra en cuestión tiene flexión de género además de la de número. Esto, naturalmente, sin que se observe la presencia de determinantes en la posición inmediatamente anterior a la palabra analizada con una frecuencia significativa en el corpus de referencia, lo que acusaría su uso como sustantivo.

En este artículo hemos mencionado también, pero no resuelto, el tema de la corrección ortográfica, ya que los errores de tipeo o de ortografía son frecuentes incluso en las publicaciones especializadas. Pensamos que se podría explorar, por ejemplo, una medida de similitud ortográfica entre palabras para descubrir lo que el autor quiso decir (ej. *cualidades* por *cuaiidades*). Sin embargo, en un caso así, una medida de similitud ortográfica no sería suficiente, ya que se debería complementar con una medida de similitud distribucional para controlar que no se ofrezcan palabras ortográficamente similares pero semánticamente distintas.

Finalmente, consideramos también la siguiente línea de trabajo futuro, inspirada en parte en el artículo de Brill (1992): investigar hasta qué punto se puede utilizar el etiquetado que ya trae el corpus de referencia para crear el modelo de la pluralización a partir de las etiquetas dejadas por este. El corpus etiquetado funcionaría así como un material de entrenamiento para un algoritmo de clasificación. De esta manera, es posible aprender a reconocer la morfología del plural (o la del género, etc.) a través de las palabras que sí son reconocidas por el etiquetador, para extrapolar esa morfología aprendida hacia las palabras que no son reconocidas. Esta línea de investigación es interesante sobre todo por las posibilidades de generalización hacia otras lenguas.

Agradecimientos

Este trabajo ha sido posible gracias a la financiación del Proyecto Fondecyt Regular 1191481: "Inducción automática de taxonomías de marcadores discursivos a partir de corpus multilingües", dirigido por el primer autor (Fondo Nacional de Desarrollo Científico y Tecnológico, Gobierno de Chile). Queremos agradecer también a los revisores Gerardo Sierra y Marcos García, ya que con sus comentarios mejoraron sustancialmente el artículo.

Referencias

Adams, Greg, Beth Millar, Eric Neufeld & Tim Philip. 1994. Ending-based strategies for part-of-speech tagging. En *Uncertainty in artificial Intelligence*, 1–7. Elsevier. 10.1016/B978-1-55860-332-5.50005-5.

Agerri, Rodrigo, Josu Bermudez & German Rigau. 2014. IXA pipeline: Efficient and ready to use multilingual NLP tools. En *Language Resources and Evaluation Conference (LREC)*, 3823–3828.

Alcina, Juan & José Manuel Blecua. 1975. *Gramática española*, vol. 1991. Barcelona: Ariel.

Alemany, José. 1920. *Tratado de la formación de palabras en la lengua castellana*. Madrid: Librería general de Victoriano Suárez.

Allauzen, Alexandre & Hélene Bonneau-Maynard. 2008. Training and evaluation of POS taggers on the French MULTITAG corpus. En *Language Resources and Evaluation Conference (LREC)*, s/p.

Ambadiang, Théophile. 1999. La flexión nominal: género y número. En *Gramática descriptiva de la lengua española*, 4843–4914. Espasa Calpe.

Asociación de Academias de la Lengua Española. 2005. *Diccionario panhispánico de dudas*. Real Academia Española.

Bazzi, Issam. 2002. *Modelling out-of-vocabulary words for robust speech recognition*: Massachusetts Institute of Technology. Tesis Doctoral.

Bazzi, Issam & James Glass. 2002. A multiclass approach for modelling out-of-vocabulary words. En *7th International Conference on Spoken Language Processing*, 1613–1616.

Bello, Andrés. 1847. *Gramática de la lengua castellana destinada al uso de los americanos*. Chile: Imprenta del Progreso.

Biemann, Chris. 2006. Unsupervised part-of-speech tagging employing efficient graph clustering. En *21st Conference on Computational Linguistics and 44th Meeting of the Association for Computational Linguistics: student research workshop*, 7–12.

Bohnet, Bernd & Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. En *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1455–1465.

Brill, Eric. 1992. A simple rule-based part of speech tagger. En *3rd Conference on Applied Natural Language Processing*, 152–155. 10.3115/974499.974526.

Brill, Eric. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics* 21(4). 543–565.

Carreras, Xavier, Isaac Chao, Lluis Padró & Muntsa Padró. 2004. FreeLing: An open-source suite of language analyzers. En *Language Resources and Evaluation Conference*, 239–242.

Cedeño, Rafael A. Núñez, Sonia Colina & Travis G. Bradley. 2014. *Fonología generativa contemporánea de la lengua española*. Washington, DC: Georgetown University Press.

Charniak, Eugene, Curtis Hendrickson, Neil Jacobson & Mike Perkowitz. 1993. Equations for part-of-speech tagging. En *11th Conference on Artificial Intelligence (AIII)*, vol. 93, 784–789.

Church, Kenneth Ward. 1989. A stochastic parts program and noun phrase parser for unrestricted text. En *International Conference on Acoustics, Speech, and Signal Processing*, 695–698. 10.3115/974235.974260.

Creutz, Mathias, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar & Andreas Stolcke. 2007. Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages. En *Proceedings of Human Language Technologies Conference*, 380–387.

De Nebrija, Antonio. 1492. *Gramática castellana*. Salamanca.

Dixon, Robert. 2009. *Basic linguistic theory volume 1: Methodology*. Oxford: Oxford University Press.

Foley, James. 1967. Spanish plural formation. *Language* 43(2). 486–493. 10.2307/411548.

Greene, Barbara B. & Gerald M. Rubin. 1971. *Automatic grammatical tagging of English*. Providence, Rhode Island: Department of Linguistics, Brown University.

Hernández Alonso, César. 1984. *Gramática funcional del español*. Madrid: Gredos.

Honnibal, Mathew. 2016. Spacy. https://spacy.io/. Accessed: 2018-10-30.

Honnibal, Matthew & Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1373–1378. 10.18653/v1/D15-1162.

Kilgarriff, Adam & Irene Renau. 2013. esTenTen, a vast web corpus of Peninsular and American Spanish. *Procedia - Social and Behavioral Sciences* 95. 12 – 19. 10.1016/j.sbspro.2013.10.617.

de León, Ramón Zacarías Ponce. 2015. Flexión de número en la composición nominal del español: estructura morfológica y rutinización. *Anuario de Letras. Lingüística y Filología* 2(2). 101–131.

Ling, Wang, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fermandez, Silvio Amir, Luís Marujo & Tiago Luís. 2015. Finding function in form: Compositional character models for open vocabulary word representation. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1520–1530. 10.18653/v1/D15-1176.

Manning, Christopher D. 2011. Part-of-speech tagging from 97 % to 100 %: is it time for some linguistics? En *International Conference on Intelligent Text Processing and Computational Linguistics*, 171–189. 10.1007/978-3-642-19400-9_14.

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

Miłkowski, Marcin. 2010. Developing an open-source, rule-based proofreading tool. *Software: Practice and Experience* 40(7). 543–566.

Moreno, Antonio & José M Goni. 1995. GRAMPAL: a morphological processor for spanish implemented in prolog. *arXiv preprint cmp-lg/9507004* .

Moyna, María Irene. 2011. *Compound words in spanish: theory and history*, vol. 316. John Benjamins Publishing.

Müller, Thomas, Ryan Cotterell, Alexander Fraser & Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2268–2274. 10.18653/v1/D15-1272.

Parada, Carolina, Mark Dredze & Frederick Jelinek. 2011. OOV sensitive named-entity recognition in speech. En *12th Conference of the International Speech Communication Association (INTERSPEECH)*, s/p.

Parra Escartín, Carla & Héctor Martínez Alonso. 2015. Choosing a spanish part-of-speech tagger for a lexically sensitive task. *Procesamiento del Lenguaje Natural* 54. 29–36.

Plank, Barbara, Anders Søgaard & Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. En *54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 412–418. 10.18653/v1/P16-2067.

Qi, Peng, Timothy Dozat, Yuhao Zhang & Christopher D. Manning. 2018. Universal dependency parsing from scratch. En *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 160–170.

Qin, Long. 2013. *Learning out-of-vocabulary words in automatic speech recognition*: School of Computer Science, Carnegie Mellon University. Tesis Doctoral.

Real Academia Española. 1741. *Orthographia española, compuesta, y ordenada por la real academia española*. Madrid: Imprenta de la Real Academia Española.

Real Academia Española. 1920. *Gramática de la lengua castellana*. Madrid: Perlado Páez y compañía, Impresores y Libreros de la Real Academia Española.

Real Academia Española. 2006. *Diccionario esencial de la lengua española*. Espasa Calpe.

Real Academia Española. 2009. *Nueva gramática de la lengua española*. Espasa Libros.

Sánchez Corrales, Víctor. 1994. La categoría morfosintáctica número en el sustantivo español. *Revista de filología y lingüística de la Universidad de Costa Rica* 20(1). 155–168.

Santos, Cicero D & Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. En *31st International Conference on Machine Learning (ICML)*, 1818–1826.

Saporta, Sol. 1965. Ordered rules, dialect differences, and historical processes. *Language* 41(2). 218–224. 10.2307/411875.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. En *International Conference on New Methods in LanguageProcessing*, 25–36.

Stockwell, Robert P, J Donald Bowen & John W Martin. 1965. *The grammatical structures of english and spanish*. University of Chicago Press.

Straka, Milan. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. En *CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 197–207. 10.18653/v1/K18-2020.

Straka, Milan, Jan Hajič & Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-u files performing tokenization, morphological analysis, POS tagging and parsing. En *10th Language Resources and Evaluation Conference (LREC)*, 4290–4297.

Straka, Milan & Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. En *CoNLL Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99. 10.18653/v1/K17-3009.

Proposta recebida em Março 2019 e aceite para publicação em Novembro 2019.

O uso da análise de clusters na identificação de padrões de transitividade linguística

The use of cluster analysis for identification of linguistic transitivity patterns

Marcus Lepesqueur
Universidade Federal de Minas Gerais
marcus.le@gmail.com

Ilka Afonso Reis
Universidade Federal de Minas Gerais
ilka@est.ufmg.br

Resumo

Este trabalho visa a apresentar a técnica de agrupamento hierárquica para análise de padrões semânticos e sintáticos da transitividade no nível oracional. Partindo de uma perspectiva empírica e baseando-se em dados reais da língua em uso, esse tipo de metodologia se mostrou útil na investigação dos padrões linguísticos a que os falantes são expostos, chegando a resultados semelhantes a categorias teoricamente conhecidas. Em um processo de amostragem simples sem reposição, foram selecionadas 690 unidades oracionais de um *corpus* de 23 entrevistas orais. Essas unidades oracionais foram analisadas em termos de nove parâmetros de transitividade e de sua respectiva sintaxe oracional. O objetivo foi identificar grupos de orações que compartilham semelhanças em termos de um conjunto de traços semânticos e morfossintáticos. Os grupos encontrados revelam um tipo de significado protoconceptual das orações, que inclui traços aspectuais e actanciais que se correlacionam. Os resultados evidenciam três cenários micro-narrativos básicos sobre os quais se desenrola o evento expresso na oração.

Palavras chave

transitividade, análise de *cluster*, semântica, sintaxe oracional

Abstract

This paper aims to present a hierarchical clustering technique for the analysis of semantic and syntactic patterns of transitivity at clausal level. From an empirical and usage-based approach, this type of methodology has proved useful for the investigation of linguistic patterns to which speakers are exposed, reaching similar results found in theoretically categories. In a simple sampling procedure without replacement, 690 oral units were selected from a corpus of 23 oral interviews. These sentence units were analyzed in terms of nine transitivity parameters and their clausal syntax. The goal was to identify groups of sentences that share similarities in terms of this set of traits. The groups found reveal a kind of protoconceptual meaning of the sentences, which includes correlated aspectual and actantial traits. The results show three basic micro-narrative scenarios on which the event expressed in clausal unfolds.

Keywords

transitivity, cluster analysis, semantics, clausal syntax

1 Introdução

A transitividade tem um papel central em grande parte das teorias linguísticas, principalmente pelo fato de que um número muito significativo de línguas apresenta uma estrutura formal —a morfossintaxe transitiva— cuja principal função é expressar um conjunto específico de propriedades semânticas (Næss, 2007). A questão principal em torno do fenômeno da transitividade é que as características semântica e sintática da transitividade tendem a covariar, sendo um fenômeno universal, ou ao menos *quasi* universal das línguas humanas. Givón (2001) aponta que, apesar das características transitivas de uma oração parecerem independentes, é um fato, na maioria das línguas, que as estruturas sintática e semântica da transitividade se sobrepõem, de forma que grande parte das orações semanticamente transitivas são também sintaticamente transitivas. De forma semelhante, Næss (2007) tenta demonstrar que, em muitas línguas, uma oração que é formalmente distinta da oração transitiva também se desvia dessa oração transitiva em termos das suas propriedades semânticas, ou seja, a escolha por uma estrutura linguística diferente reflete o desejo do falante de exprimir uma semântica diferente do protótipo da transitividade.

Mas apesar dos esforços das ultimas décadas, a linguística contemporânea não encontrou uma explicação completa capaz de abraçar a complexidade desse fenômeno. Primeiro, porque os verbos variam as características da sua regência

DOI: 10.21814/lm.11.2.292

em diferentes contextos de uso; segundo, porque as definições tradicionais da transitividade tratam, comumente, da mesma forma, elementos sintáticos e semânticos, que não apenas são distintos, mas que também interagem de forma complexa (Lucena & Cunha, 2012). Influenciada por fatores diacrônicos e sincrônicos, a morfossintaxe transitiva pode acomodar uma gama de valores semânticos não transitivos e, inversamente, a semântica transitiva pode ser expressa por mais de um padrão formal.

Por exemplo, no português do Brasil, como em muitas outras línguas, algumas orações sintaticamente transitivas não apresentam a semântica transitiva. Enquanto orações transitivas prototípicas como em (1) expressam ações, alguns verbos psicológicos transitivos, como em (2), têm significado estático, sendo considerados verbos de estado mental.

(1) eu raspei a barba[1]

(2) ela sabe tudo

Na oração transitiva, normalmente temos, como sujeito sintático, o causador do evento e, como objeto, o participante afetado. No entanto, encontramos casos em que essas funções não são claras. Næss (2007) usa o termo "inversão do vetor semântico" para descrever uma relação semântica aparentemente invertida. Em (3), por exemplo, o sujeito sintático "Eu" não é exatamente o causador do evento, e o objeto sintático "ela" não é afetado pelo evento.

(3) Eu vi ela

Tradicionalmente, a literatura linguística tenta associar a sintaxe transitiva à semântica de um sistema causal físico; no entanto, um número significativo de estruturas transitivas exibe relações não-físicas ou sem uma conexão direta do tipo causa-efeito. Por exemplo, "convidar" em (4a) e "chamar" em (4b), mesmo em estruturas transitivas, não expressam, necessariamente, um efeito direto no participante que se encontra na posição do objeto direto.

(4) a. Dilma convidou eu
b. Chamei ela

Com o objetivo de abordar essa complexa interface da sintaxe e semântica transitiva, este artigo investigou, em uma perspectiva empírica, alguns aspectos gramaticais das unidades oracionais no português do Brasil. Mais precisamente, esse trabalho visa a apresentar uma técnica hierárquica de agrupamento para analisar grupos de orações que compartilham aspectos semânticos e sintáticos da transitividade. Partindo de uma perspectiva empírica e baseando-se em dados reais da língua em uso, esse tipo de metodologia se mostrou útil na investigação dos padrões linguísticos a que os falantes são expostos, chegando a resultados semelhantes a categorias teoricamente conhecidas.

2 Parâmetros de Transitividade

Em um artigo seminal sobre o tema, Hopper & Thompson (1980) isolaram alguns dos componentes da noção de transitividade e estudaram a forma como eles são tipicamente codificados na gramática. A partir de um conjunto de evidências translinguísticas os autores propuseram um conjunto de parâmetros semânticos e morfossintáticos relacionados ao fenômeno da transitividade.

Hopper & Thompson (1980) argumentam que as gramáticas das línguas agrupam estes parâmetros em função de uma escala de transitividade, i.e., em uma mesma sentença, um traço morfossintático ou semântico obrigatório que marca alta transitividade tende a não ocorrer com outro que marca baixa transitividade.

Cada parâmetro proposto por estes autores[2], essencialmente, captura algum tipo de diferença entre as unidade oracionais. O primeiro parâmetro proposto, denominado Número de Participantes, distingue a particularidade de orações que aparecem sem objeto sintático, como (5a), daquelas que têm um ou mais objetos sintáticos, como (5b).

(5) a. a intimação estourô
b. eu disse pra ela do medo

O segundo, o terceiro e o quarto parâmetros da transitividade analisados nesta pesquisa são a Cinese, a Telicidade e a Pontualidade da predicação. Esses parâmetros fazem parte da noção mais geral de aspecto e referem-se à forma de se conceptualizar a estrutura temporal interna de uma determinada situação. Cinese refere-se à

[1] Os exemplos apresentados aqui foram retirados do *corpus*. Na transcrição foram respeitados alguns padrões de pronúncia, tais como a ausência da fricativa glotal surda (/r/) nos infinitivos verbais, ausência de morfema de plural e reduções como "tá", para "está" ou "cê" para "você".

[2] Nesta pesquisa, foram incluídos todos os parâmetros de transitividade com exceção da Individuação do objeto, que se refere a um conjunto variado de traços, o que inclui aspectos da referencialidade e definição/indefinição do objeto sintático. Este parâmetro depende de um tratamento teórico inconcluso e Hopper & Thompson (1980) o operacionalizaram em uma escala própria, distinta dos demais.

distinção entre predicações que expressam ações, como em (6a) daquelas que expressam estados, como em (6b). Telicidade refere-se à presença ou à ausência do traço télico do evento, i.e., se o evento é conceptualizado como tendo um ponto de conclusão definido, como em (7a), ou sem este ponto de conclusão, como em (7b). Pontualidade distingue eventos pontuais, que não possuem fases intermediárias entre o seu início e o seu final, como em (8a), de eventos durativos, como em (8b).

(6) a. minha mãe me ligô esses dia
b. os médico daqui eles são muito bom

(7) a. tinha que fazê o relatório
b. aí eu andava bastante

(8) a. ele morreu
b. reformô minha casa toda

Com o parâmetro Modalidade, Hopper & Thompson (1980) fazem a distinção entre o modo *realis* e o *irrealis* do evento, marcando a oposição entre a forma indicativa e formas não assertivas, tais como o subjuntivo, o condicional, o hipotético etc. O parâmetro da Modalidade distingue, portanto, o grau de realização do evento linguístico expresso. Também o parâmetro Polaridade, definido como a distinção entre a forma negativa e a forma afirmativa da oração, relaciona-se com o grau de realização do evento.

Os parâmetros Agentividade e Intencionalidade referem-se ao grau de envolvimento do argumento externo na atividade expressa pelo verbo. A Agentividade refere-se ao elemento desencadeador do processo e Intencionalidade refere-se ao sentido de volição. Em (9) encontra-se um exemplo de unidade oracional com argumento externo não agentivos e não volitivos e em 10 o argumento externo agentivo e volitivo.

(9) a mente dele é de criança

(10) e eu falei pra ela

Por fim, a noção semântica de Afetação, último parâmetro analisado aqui, é tradicionalmente um dos critérios essenciais na definição de transitividade. Muitas línguas mostram um padrão de se codificarem argumentos fortemente afetados pelo evento verbal como objetos de construções transitivas, e argumentos não afetados, ou menos afetados, em outras posições sintáticas (Lepesqueur, 2017). Em (12) encontra-se um exemplo típico da presença do parâmetro Afetação na construção transitiva.

(11) eu podia amputá sua perna

No modelo de Hopper & Thompson (1980) a transitividade passou a ser definida não como uma característica do elemento verbal, mas como um conjunto de componentes ligados à unidade oracional que se relacionam de maneiras específicas. Essa mudança de perspectiva alimentou uma série de pesquisas que investigaram tanto a maneira como as línguas codificam formalmente os parâmetros da transitividade, quanto as motivações semânticas e pragmáticas da variação na morfossintaxe transitiva.

No português do Brasil, Lepesqueur (2017) mostrou que apenas um dos parâmetros propostos por Hopper & Thompson (1980), a saber, a Afetação do objeto sintático, é um preditor positivo, estatisticamente significativo, da sintaxe transitiva. Dito de outra forma, a presença da afetação do objeto na oração é um indicador de alta probabilidade da ocorrência da estrutura oracional transitiva. Os demais parâmetros de transitividade encontram-se distribuídos de maneira mais ou menos homogênea entre todas as estruturas oracionais, não compondo elementos distintivos da sintaxe transitiva. O autor sugere ainda que certos parâmetros, especialmente a Telicidade, podem estar associadas a padrões sintáticos não-transitivos. Isso aponta para certas particularidades da organização da estrutura transitiva no português e sugerem possibilidades de se repensar a associação entre a semântica e a sintaxe transitiva.

Como os parâmetros de transitividade, independentemente da estrutura da unidade oracional, se agrupam no português do Brasil? Existe apenas uma semântica não transitiva, ou podemos esperar diferentes padrões semânticos fora da transitividade? Partindo desse conjunto de questões, esse trabalho visa a apresentar uma metodologia estatística capaz de identificar padrões semânticos e sintáticos da oração, chegando a resultados semelhantes a categorias teoricamente conhecidas. O objetivo foi identificar grupos de orações que compartilham semelhanças em termos de um conjunto de traços morfossintáticos e semânticos ligados à transitividade. Esperamos que os resultados apresentados aqui possam elucidar as regularidades sintático-semânticas às quais os falantes estão expostos e sobre as quais emergem os fenômenos gramaticais.

3 Metodologia

Uma das principais dificuldades para a compreensão da transitividade é que estamos lidando em um campo de interface entre a estrutura formal e

a estrutura conceptual[3]. Apesar dos avanços recentes da linguística sobre a natureza dessa articulação, restam ainda muitas questões a respeito da maneira através da qual um item lexical se integra em uma sintaxe —e, mais ainda, em uma estrutura macrotextual e discursiva— e como isso pode produzir efeitos de significado.

Por consequência, a compreensão da transitividade depende, antes de tudo, de um tratamento de dois eixos distintos entre si: um eixo essencialmente semântico e outro essencialmente sintático. Por fim, além da descrição desses dois eixos, é preciso um modelo linguístico, talvez mais especificamente semiótico, que explique a maneira complexa e particular através da qual a sintaxe e a semântica transitiva interagem.

Em uma análise de interface entre sintaxe e semântica, o caminho tradicional de investigação da transitividade tem sido agrupar padrões morfossintáticos a fim de se analisar uma estrutura semântica subjacente. Assim, por exemplo, pode-se distinguir a estrutura formal transitiva da intransitiva para, em seguida, tentar-se identificarem as diferenças semânticas nesses grupos. Mas o caminho inverso também é possível: primeiro identificar grupos de orações semanticamente semelhantes e posteriormente analisar a relação desse grupo com padrões formais da língua.

Por diversas razões, a primeira opção tem sido o caminho canônico de investigação. Uma das principais questões é o fato da língua agrupar uma quantidade a princípio ilimitada de informações conceptuais em um número relativamente limitado de estruturas e regras gramaticais. Isso torna mais fácil agrupar as unidades oracionais a partir das suas características formais, que são em um número relativamente reduzido, do que agrupá-las a partir das suas distinções conceptuais. Desta perspectiva, vários teóricos têm tentado analisar padrões gramaticais (morfossintáticos e lexicais) buscando inferir uma estrutura conceptual subjacente. Este é o raciocínio básico por trás dos trabalhos de Hopper & Thompson (1980), Givón (2001) ou Næss (2007).

[3]Aqui utilizamos o termo conceptual, escrito com p, para destacar o caráter processual da estrutura semântica. Em geral, os teóricos da Linguística Cognitiva têm utilizado o termo *conceptualization* (traduzido normalmente como conceptualização) para se referir ao processo de construção de significado, destacando sua natureza dinâmica e processual. A conceptualização tem sido descrita como um processo imagético (em oposição à noção tradicional de estruturas proposicionais), interativo (porque envolve processos de negociação e interação entre os interlocutores), e imaginativo (porque envolve processos de simulação e mesclagens conceituais) (Broccias, 2013).

Um caminho alternativo é tomar a estrutura conceptual como um dado, perceptível pelos falantes, a fim de, posteriormente, estabelecerem-se relações simbólicas com a estrutura formal da língua. A proposta de Halliday et al. (2014), que compreende o sistema da transitividade como uma função gramatical organizadora, com seus próprios modelos e esquemas, é um exemplo da tentativa de focalizar, inicialmente, a maneira como a informação conceptual é estruturada para, posteriormente, identificar sua manifestação formal.

Este trabalho parte desta ultima via e tem como objetivo identificar grupos de orações que compartilham semelhanças em termos do conjunto de parâmetros de transitividade como um todo, de maneira parcialmente[4] independente da estrutura formal da oração e, posteriormente, tentar estabelecer uma relação entre os parâmetros e a estrutura sintática oracional do português do Brasil. Buscamos identificar agrupamentos naturais de unidades oracionais (grupos de orações que compartilham semelhanças em termos dos seus parâmetros) a partir de um conjunto de técnicas estatísticas de agrupamento. Essa metodologia mostrou-se capaz de analisar, empiricamente, traços semânticos ou morfossintáticos em dados reais da língua em uso, chegando a resultados semelhantes àqueles esperados teoricamente

3.1 Composição do *corpus*

O *corpus* desta pesquisa é composto de relatos orais produzidos por 23 participantes, publicados em Lepesqueur (2017)[5]. As narrativas orais produzidas por esses participantes foram gravadas e transcritas. Para facilitar a importação e o tratamento dos dados pelo programa computacional de análise estatística, cada linha da transcrição contém o trecho correspondente

[4]Dizemos parcialmente porque os parâmetros não são puramente semânticos. Por exemplo, o parâmetro Afetação refere-se a uma distinção semântica que ocorre em um certa posição sintática, a saber, a posição de objeto. Mas esse objeto pode ser, a princípio, preposicionado ou não, ou fazer parte de uma estrutura sintática transitiva ou bitransitiva.

[5]A pesquisa de Lepesqueur (2017) teve o objetivo de investigar o fenômeno da transitividade em uma população clínica. Parte do *corpus*, portanto, é comporto de entrevistas produzidas por pacientes com diagnóstico de esquizofrenia paranoide. O referido trabalho não identificou algum tipo de correlação especial intra-parâmetros na população clínica, apenas a maior probabilidade de ocorrer o parâmetro Afetação na fala dos pacientes. Uma vez que trata-se de um parâmetro pouco frequente no *corpus*, não há evidências de que os agrupamentos apresentados neste trabalho não possam ser generalizados.

a uma única unidade oracional, definida como uma predicação centralizada pela unidade verbal. As transcrições foram realizadas usando-se as convenções ortográficas, sem, no entanto, dar atenção especial às questões fonéticas, uma vez que não possuem relevância para a pesquisa.

Do total de 7939 unidades oracionais da transcrição, 5690 fizeram parte da análise, uma vez excluídos trechos do entrevistador, unidades oracionais abandonadas ou parcialmente incompreensíveis, expressões idiomáticas e estruturas não sentenciais.

Em um processo de amostragem simples sem reposição, foram selecionadas 690 unidades oracionais (30 por participante), analisadas em termos dos parâmetros de transitividade e sua sintaxe oracional. Os *corpus* é original da pesquisa de Lepesqueur (2017), que definiu o tamanho da amostra respeitando o número mínimo de observações sugeridas por Hair et al. (2009) para análise de regressão logística. O autor também considerou um desenho experimental balanceado em termos do número de observações por participantes. Os dados foram organizados em uma planilha eletrônica de forma a conter, para cada unidade oracional observada, a classificação dos parâmetros de Hopper & Thompson (1980) e Thompson & Hopper (2001), em termos de alta (1) ou baixa (0) transitividade[6]. Posteriormente, esses dados foram analisados no ambiente de programação estatística R (R Development Core Team, 2017).

3.2 Análise de clusters

A análise de *cluster* (também conhecida como análise de conglomerado ou de agrupamentos) é um conjunto de algoritmos e de técnicas analíticas multivariadas que visa a agrupar os elementos de uma amostra ou população a partir da similaridade desses elementos quando os comparamos em uma série de variáveis (Mingoti, 2017). O objetivo desse tipo de técnica é realizar agrupamentos que maximizem as semelhanças entre observações que pertençam a um mesmo grupo, o que torna o grupo mais homogêneo, ao mesmo tempo em que minimizem as semelhanças entre grupos diferentes, o que torna os grupos heterogêneos entre si. No campo dos estudos linguísticos a análise de agrupamentos tem sido utilizada para descrever uma ampla gama de fenômenos que vão desde diferenças dialetais até polissemias (Divjak & Fieller, 2014).

Uma questão central desse tipo de análise refere-se à métrica utilizada para se decidir o grau de similaridade (ou inversamente, de dissimilaridade) entre os elementos observados. No caso de variáveis qualitativas, tais como os parâmetros binários de transitividade analisados aqui, foi utilizado o coeficiente de concordância simples (s_{ij}) (Sokal & Sneath, 1963). Trata-se de um coeficiente simétrico, ou seja, que considera o mesmo peso para as concordâncias positivas ou negativas. O coeficiente é calculado pela soma do número total de concordâncias entre os atributos dos elementos i e j, dividido pelo número total de atributos.

$$s_{ij} = \frac{Número\ de\ atributos\ concordantes}{Número\ total\ de\ atributos}$$

O valor de s_{ij} pode variar entre 0 e 1. O coeficiente s_{ij} foi calculado, para cada i≠ j, através do coeficiente geral de Gower (1971) que permite integrar também, se necessário, variáveis quantitativas ou ordinais.

Considerando o coeficiente de similaridade s_{ij}, a matriz de dissimilaridade dos dados será composta pelo índice de dissimilaridade d_{ij}, calculado pelo complementar de s_{ij} para cada par de orações do *corpus*.

A partir dessa matriz de dissimilaridade foram utilizadas técnicas hierárquicas de agrupamento para encontrar a melhor partição dos dados. Optamos pelo uso das técnicas hierárquicas, em uma análise exploratória, uma vez que não temos um número pré-estabelecido de grupos e buscamos identificar uma estrutura natural dos dados. As análises foram conduzidas utilizando-se tanto técnicas hierárquicas aglomerativas[7] quanto a divisivas[8] (Rousseeuw & Kaufman, 1990).

A técnica aglomerativa começa com n grupos, sendo n o número de elementos no banco de dados. Cada observação é separada em um *cluster* específico e o algoritmo de agrupamento tenta encontrar os valores mais semelhantes para formar os grupos. Inversamente, a técnica divisiva assume inicialmente todos os elementos em um único grupo e inicia a divisão dos elementos

[6]Essa analise foi feita manualmente a partir dos critérios descritos em Lepesqueur (2017). O coeficiente de Kappa sugere uma concordância substancial entre dois avaliadores especialistas, previamente treinados (k=0.74, $p < 0.05$). Não existe hoje uma boa abordagem para automatizar a avaliação dos parâmetros de transitividade. Isso depende, antes de tudo, de uma maneira eficiente de tratar computacionalmente valores semânticos ligados especialmente ao elemento verbal. A esse respeito, um dos projetos piorneiros é o FrameNet, idelizado por Chales Fillmore, no campo da Semântica de *Frames*. Para mais detalhes ver sobre o projeto ver `https://framenet.icsi.berkeley.edu`

[7]Função *hclust*, do pacote *stats* do software R.

[8]Função *diana*, do pacote *cluster* do software R.

mais distantes em grupos diferentes. A similaridade entre dois conglomerados foi definida pelo método de ligação completa, ou seja, a partir da comparação da maior distância entre os pontos de dois grupos. Esse método tende a formar grupos mais compactos e sem a tendência de longas cadeias[9].

Para decidir sobre o número K de grupos da partição final dos dados analisados, utilizamos algumas medidas de avaliação da qualidade dos agrupamentos, analisando tanto a compacidade (a máxima similaridade intra-grupo) quanto a separabilidade (a mínima similaridade entre grupos).

Inicialmente utilizamos duas medidas de avaliação de todas as partições de 2 a 30 *clusters*[10], tanto na técnica aglomerativa quanto na divisiva. A primeira medida foi uma generalização da soma de quadrados dos desvios intra-*cluster* (tipicamente utilizada na métrica euclidiana) e a segunda medida a largura de silhueta (Rousseeuw, 1987).

A soma de quadrados dos desvios intra-*cluster* (SQ_k) é uma estimativa da compacidade de um dado *cluster* k e se refere, aqui, à metade da soma dos quadrados das dissimilaridades intra-*cluster* dividido pelo tamanho do *cluster*. SQ_k é definido como:

$$SQ_k = \frac{1}{2n_k} \sum_{i,j \in C_k} d_{ij}^2$$

onde n_k é o número de elementos no *cluster* C_k e d_{ij} é o valor da dissimilaridade entre o elemento i e j do *cluster* C_k. Quanto maior o valor de SQ_k, menor será a compacidade do *cluster* k. A soma de quadrados dos desvios intra-cluster é uma medida particular do *cluster* k. Na partição final com K *clusters*, cada um desses *clusters* apresenta um valor próprio de SQ_k. Para uma dada partição final, a soma de quadrados dos desvios intra-*cluster* desta partição é dada pela média dos valores de SQ_k de todos os K *clusters*.

A largura média de silhueta (L) oferece uma estimativa da separabilidade dos agrupamentos ao comparar a similaridade de uma observação amostral com as demais observações do próprio *cluster* e do seu vizinho mais próximo. A largura média de silhueta é calculada a partir do coeficiente de silhueta (S_i) da observação amostral i:

$$S_i = \frac{b_i - a_i}{MAX\,(a_i, b_i)}$$

onde a_i é a média da dissimilaridade (d_{ij}) da observação amostral i com todos os membros do *cluster* ao qual pertence e b_i, a dissimilaridade (d_{ij}) mínima da observação i com todos os demais dados que não pertencem ao seu *cluster*.

O coeficiente de silhueta S_i varia no intervalo de $[-1, 1]$ e se aproxima de -1 quando o elemento i está, em média, mais próximo dos elementos de um *cluster* vizinho do que dos elementos do seu próprio *cluster* (caso em que $b_i < a_i$.). O coeficiente aproxima-se de 0 na medida em que b_i seja semelhante a a_i, sugerindo que o elemento i encontra-se em um ponto intermediário entre dois *clusters*. O coeficiente aproxima-se de 1 quando o elemento i está em média mais próximo dos elementos do próprio *cluster* do que do *cluster* vizinho (caso em que $b_i > a_i$).

A largura de silhueta (S_k) de um *cluster* k é dada pela média dos coeficientes de silhueta de todas as observações pertencentes ao *cluster* k. Por suas vez a largura média de silhueta (L) da partição foi calculado pela média dos valores de S_k dos *clusters* que compõem aquela partição.

3.3 Escolha da melhor técnica de agrupamento

A partir dos índices apresentados, definimos a melhor partição obtida na técnica aglomerativa e a melhor partição obtida na técnica divisiva. Para auxiliar na comparação entre essas duas partições finais, iniciamos um segundo passo de análise das medidas da qualidade da partição a partir do Índice Dunn2 (Halkidi et al., 2001) e Índice WB. Ambos os índices são calculados a partir da dissimilaridades intra-*cluster* $d\,(C_k)$ e a dissimilaridade entre *clusters* $d\,(C_k, C_l)$.

Quanto menor as dissimilaridades intra-*cluster*, maior o compacidade da partição. A dissimilaridade intra-*cluster* do *cluster* k é dado por:

$$d\,(C_k) = \frac{2}{n_k\,(n_k - 1)} \sum_{i \in C_k, j \in C_k} d_{ij}$$

Quanto maior as dissimilaridades entre *clusters*, maior a separabilidade da partição final. A dissimilaridade entre o *cluster* k e l é dado por:

$$d\,(C_k, C_l) = \frac{1}{n_k n_l} \sum_{i \in C_k, j \in C_l} d_{ij}$$

O Índice WB (*whitin/between*), I_{wb}, é calculado pela razão entre as médias de $d\,(C_k)$ e

[9] Que ocorre quando um *cluster* incorpora, a cada interação, um único elemento próximo.

[10] A principio, não esperamos que haja mais de 30 grupos teoricamente importantes para explicar o fenômeno da transitividade. Mas não se trata de uma restrição da técnica. Ainda que computacionalmente demorado, é possível analisar até n-1 agrupamentos, sendo n é o número de observações no banco de dados.

$d(C_k, C_l)$ para todos os *clusters* da partição final. Quanto menor o índice WB, melhor a relação entre compacidade (numerador) e separabilidade (denominador).

O Índice Dunn2 é dado pela razão entre a menor dissimilaridade entre dois *clusters* e a maior dissimilaridade intra-*cluster* da partição final. Quanto maior o índice, melhor a relação entre a separabilidade (numerador) e a compacidade (denominador).

A partição final, depois de comparadas as técnicas aglomerativa e divisiva, foi representada graficamente utilizando a técnica de escalonamento multidimensional (MDS). O método MDS faz a decomposição espectral de uma matriz relacionada à matriz de dissimilaridade entre os elementos amostrais. Assim, ao se construir novas dimensões e grafar seus valores num gráfico de dispersão, conserva-se aproximadamente as dissimilaridades que os elementos amostrais apresentam entre si. Em suma, essa técnica permite representar espacialmente a matriz de dissimilaridade dos elementos sintetizando essa matriz em um certo número de componentes utilizadas como coordenadas de um gráfico de percepção. Neste gráfico, as relações geométricas correspondem, de maneira aproximada, às relações de dissimilaridade dos dados observados (Mingoti, 2017).

4 Resultados e discussão

As unidades oracionais do *corpus* foram analisadas em termos dos 9 parâmetros propostos por Hopper & Thompson (1980), sendo cada um destes parâmetros caracterizado como de alta (1) ou baixa (0) transitividade. A Figura 1[11] mostra a distribuição, no *corpus*, da frequência absoluta dos parâmetros, segundo o seu grau de transitividade. Uma vez que cada parâmetro soma 690 observações (tamanho da amostra), o gráfico representa também, visualmente, a proporção relativa dos traços de baixa e alta transitividade.

Alguns traços de transitividade são especialmente raros na amostra analisada: a afetação do objeto sintático (Afetação=Alta) ocorre em menos de 10% do *corpus* e a baixa polaridade da oração (Polaridade=Baixa) em apenas 12,6% A maior proporção de orações afirmativas é, provavelmente, uma consequência do gênero textual do *corpus*, composto por entrevistas orais.

[11]Na figura os parâmetros de transitividade foram abreviados e são apresentados na seguinte ordem: Telicidade (Telicid.), Pontualidade (Pont.), Polaridade (Pol.), Participante (Part.), Modalidade (Mod.), Intencionalidade (Intenc.), Cinese (Cin.), Agentividade (Agent.), Afetação (Afet.)

De maneira geral, a presença de traços de baixa transitividade são mais comuns na amostra do que traços de alta transitividade[12]. Esta característica já era esperada uma vez que a bibliografia especializada tem afirmado que o gênero conversação tende a ser de baixa transitividade, como sugerem Thompson & Hopper (2001), para o inglês, Rozas (2004), para o espanhol, Shahrokhi & Lotfi (2012), para o persa, e Lima (2013), para o português. Bois (2003), analisando a preferência no discurso pelo uso de certas configurações sintáticas, mostrou que, em diversas línguas (a saber, Hebrew, Sakapultek, Papago, Inglês e Goonyandi), 50 a 62% das unidades oracionais não possuem nenhum argumento nominal. De maneira geral, as orações de baixa transitividade parecem ser mais úteis no contexto de comunicação interpessoal e de aspectos subjetivos do que as orações de alta transitividade (Rozas, 2004).

4.1 Análise de agrupamento por técnica hierárquica

Para a investigação do melhor agrupamento dos dados, iniciamos com a análise do nível de fusão dos aglomerados. À medida que o número de *clusters* na partição aumenta, a média da dissimilaridade intra-*cluster* decresce. As Figuras 2 e 3, mostram os valores da média de SQ_k de todos os *clusters* que formam cada partição, tanto na técnica hierárquica aglomerativa quanto na divisiva.

Buscamos identificar nos gráficos mudanças significativas ("quedas bruscas") na média da soma de quadrados da dissimilaridade, o que representa ganhos importantes na homogeneidade ou compacidade dos agrupamentos. No uso da técnica aglomerativa, na Figura 2, destaca-se o ganho de consistência interna na partição $K = 13$. No uso da técnica divisiva, na Figura 3, a queda significativa na soma de quadrado dos desvios ocorre na partição $K = 3$.

As Figuras 4 e 5, a seguir, mostram a média das larguras de silhueta de todos os *clusters* que compõem as partições com 2 a 30 grupos. Buscamos aqui as partições com maiores médias da largura de silhueta, o que representa maior separabilidade entre grupos vizinhos. Com o uso da técnica aglomerativa, o salto na média da largura de silhueta ocorre novamente no agrupamento de $K = 13$, com ganhos poucos significativos depois

[12]Especialmente se retiramos o parâmetro Polaridade, que deixou de ser considerado relevante na descrição do fenômeno da transitividade em Thompson & Hopper (2001).

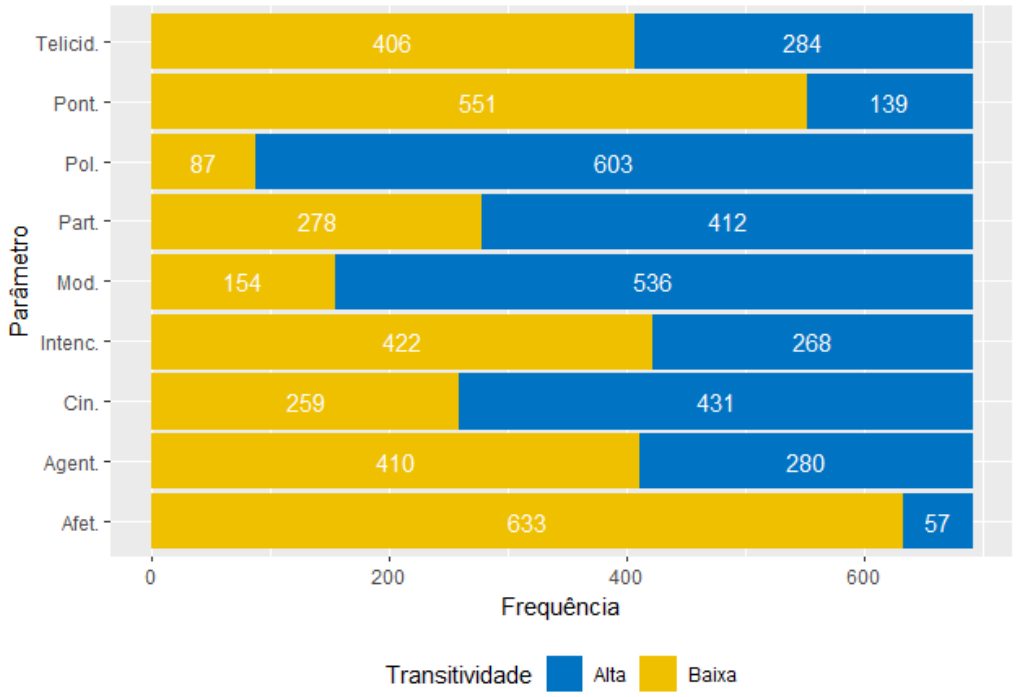

Figura 1: Gráfico de barras da frequência absoluta dos parâmetros de transitividade.

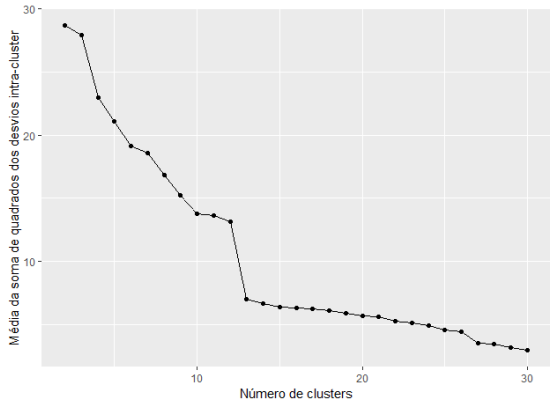

Figura 2: Média da soma de quadrados da dissimilaridade intra-*cluster* na Técnica Aglomerativa

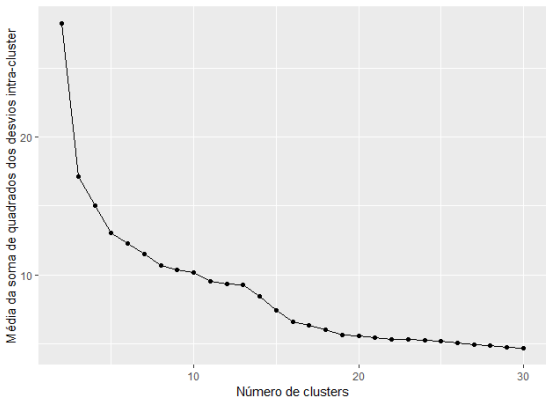

Figura 3: Média da soma de quadrados da dissimilaridade intra-*cluster* na Técnica Divisiva

dessa partição. Na técnica divisiva, o pico ocorre na partição com $K = 3$.

As duas medidas de avaliação da qualidade dos agrupamentos sugerem, portanto, uma partição com K=3, no uso da técnica divisiva, ou com K=13, no uso da técnica aglomerativa. A Tabela 1 apresenta a comparação das duas partições através de outros índices de avaliação da qualidade dos agrupamentos.

A Tabela 1 mostra um melhor desempenho da partição $K = 3$ (técnica divisiva) nos índices Dunn2 e média da largura de silhueta (sendo ambos os índices uma estimativa da relação entre compacidade e separabilidade), além do melhor desempenho na média da dissimilaridade entre *clusters* (Média de $d(C_k, C_l)$). Apesar da partição final com $K = 13$ apresentar menor dissimilaridade intra-*clusters* (Média de $d(C_k)$), e consequentemente, melhor desempenho na razão (I_{wb}), esse ganho não acompanha a perda em parcimônia no uso de um número tão grande de grupos. Optou-se, portanto, pela partição final com $K = 3$, utilizando-se a técnica hierárquica divisiva.

Utilizando a técnica de Escalonamento Multidimensional (MDS), é possível representar espacialmente, em um gráfico de percepção, a matriz de dissimilaridade dos elementos e a partição final dos dados. A Figura 6 representa as observações

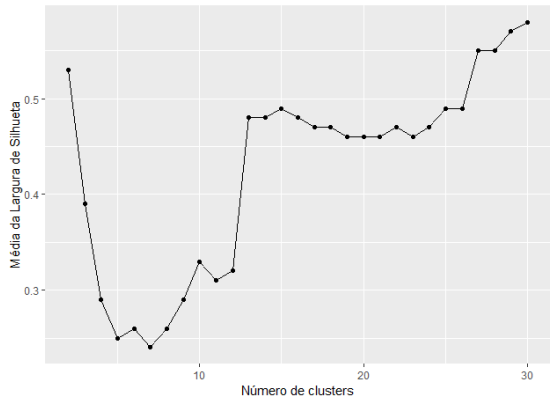

Figura 4: Média da Largura de silhueta na Técnica Aglomerativa

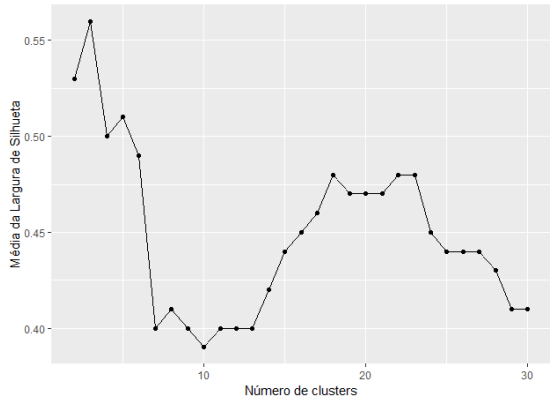

Figura 5: Média da Largura de silhueta na Técnica Divisiva

e o agrupamento final de maneira que a distância entre os pontos corresponde aproximadamente à dissimilaridade entre as observações.

A proporção da variância explicada pelas duas dimensões obtidas através do escalonamento é de 0,47. É importante notar que não pretendemos aqui realizar a análise de agrupamento a partir do MDS, mas apenas representar graficamente a distribuição dos *clusters*. O gráfico permite visualizar bem a correspondência entre a partição final obtida no uso da técnica divisiva e a representação espacial das observações.

Número de clusters	K=3	K=13
Média de SQ_k	17,16	7,01
Média de $d(C_k)$	0,18	0,09
Média de $d(C_k, C_l)$	0,50	0,43
I_{wb}	0,37	0,22
Índice Dunn2	1,75	0,93
Média da largura de silhueta	0,56	0,48

Tabela 1: Medidas de avaliação da qualidade dos agrupamentos (k=3, técnica divisiva; k=13, técnica aglomerativa).

Para melhor identificar as características típicas de cada *cluster*, apresentamos também a frequência relativa das variáveis em cada um deles. A Figura 7 mostra a frequência relativa de ocorrências dos traços de transitividade em cada grupo da técnica divisiva $K = 3$, destacando em verde quando o traço ocorre em aproximadamente 100% das unidades oracionais do grupo em questão e, em amarelo, quando os parâmetros correm em aproximadamente 0% das unidades oracionais pertencentes àquele grupo. A ordem de apresentação das categorias nesse gráfico foi escolhida de modo a facilitar a visualização dos conjuntos de traços mais frequentes (blocos em verde) e os menos frequentes (blocos em amarelo) em cada grupo.

Percebe-se que os parâmetros Polaridade (negativa e afirmativa) e Modalidade (*realis* e *irrealis*) são relativamente distribuídos de forma homogênea entre os grupos. Os demais parâmetros se agrupam de forma bem definida, mostrando um padrão semântico específico de cada *cluster*, representado nos blocos em verde e amarelo.

O *cluster* 3 possui uma estrutura aspectual bem definida. Por estrutura aspectual, entendemos as diferenças da estrutura temporal interna, não relacionais, do evento expresso na oração (Comrie, 1976). Este grupo apresenta predicados que expressam estados (parâmetro Cinese=Baixa), sendo durativos (parâmetro Pontualidade=Baixa) e atélicos (parâmetro Telicidade=Baixa). Eles são igualmente não agentivos e não intencionais (parâmetros agentividade e intencionalidade=Baixa). Em (12) encontramos um exemplo prototípico desse *cluster*.

Distintamente, os *clusters* 1 e 2 expressam ações (parâmetro Cinese=Alta). O *cluster* 1 expressa eventos não-agentivos e não-intencionais (parâmetros agentividade e intencionalidade=Baixa), tipicamente pontuais (parâmetro Pontualidade=Alta), enquanto o *cluster* 2 expressa eventos tipicamente agentivos, intencionais (parâmetros agentividade e intencionalidade=Alta) e durativos (parâmetro Pontualidade=Alta), podendo ou não apresentar um ponto télico (parâmetro Telicidade=Indefinido). Os exemplos (13) e (14) são prototípicos do *cluster* 1 e 2 respectivamente.

(12) ele é casado

(13) depois a intimação estourô

(14) eu tratava das criação

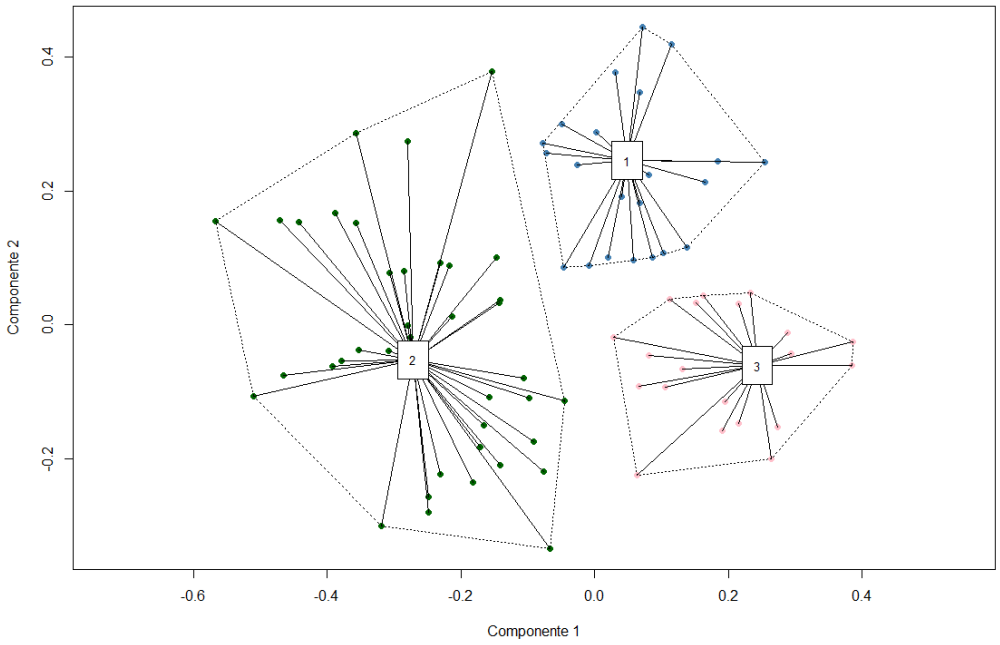

Figura 6: Gráfico de Percepção – Técnica divisiva $k = 3$.

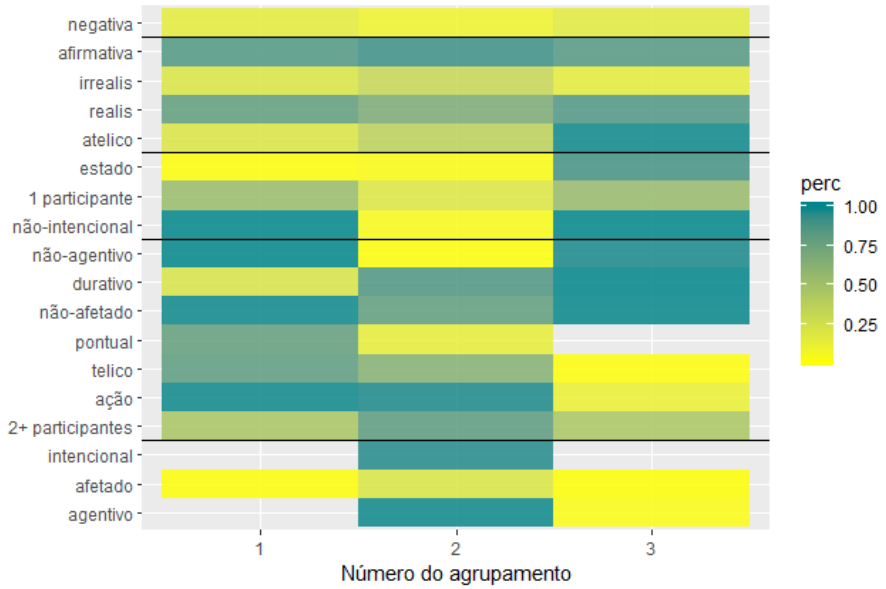

Figura 7: Frequência relativa dos parâmetros em cada *cluster* considerando a técnica divisiva.

4.2 Relação entre os agrupamentos e a estrutura sintática da oração

Especialmente no âmbito da Linguística Cognitiva, um dos conceitos chaves para a compreensão da estrutura semântica é a categorização. O processo de aquisição da linguagem envolve não apenas aprender quais categorias são relevantes para nós, em nosso ambiente, mas também aprender um número limitado de estruturas e regras gramaticais utilizadas para se expressar um número ilimitado de experiências (Divjak & Fieller, 2014).

A categorização é o resultado de uma capacidade cognitiva humana geral de realizar abstrações e reconhecer um núcleo comum de aspectos da experiência corpórea e social. Contrariamente à visão clássica que compreende os conceitos como representações de estados de um mundo objetivo e, portanto, não sujeitos à experiência subjetiva, estudos empíricos têm mostrado que

os conceitos são definidos e compreendidos dentro de um quadro conceitual que depende da natureza da experiência humana. Esta concepção denominada de actuação (*enaction*) ou corporeidade (*embodiment*) foi especialmente tratada por Johnson (2013) e por Varela et al. (1991), dentro das Ciências Cognitivas, e se resume na afirmação de que a cognição não pode ser compreendida fora de nossa história social e de ações corporalizadas. Por ação corporalizada, entende-se, primeiro, que a nossa cognição é inseparável da forma como experienciamos processos sensoriais e motores (percepção e ação) decorrentes de termos um corpo como o nosso e, segundo, que essa experiência encontra-se mergulhada em um contexto biológico, psicológico e cultural mais abrangente. A experiência envolve padrões recorrentes, ou *gestalts*, no sentido de uma organização coerente, que são fundamentais para o processo de significação e estão na origem de certos pontos de referência do nosso sistema conceitual (Johnson, 2013) (Lakoff, 2008). Em resumo, nosso sistema conceitual ancora-se em certos padrões de interação sensório-motoras, que servem de base para a significação.

O conjunto dos dados analisados neste trabalho revela que as características semânticas da transitividade podem ser agrupadas em padrões relativamente bem definidos em termos de suas características. Esses grupos parecem instanciar três cenas prototípicas ou microcenários narrativos sobre as quais a unidade oracional se organiza.

Nós reencontramos aqui um agrupamento semelhante à distinção tradicional das classes acionais de Vendler (1967). O primeiro *cluster* aproxima-se do que Vendler denominou de *achievements*: eventos pontuais que expressam tipicamente uma mudança, mais ou menos súbita, de um estado para outro. As orações desse grupo, no *corpus*, ocorrem tipicamente associadas a sujeitos sintáticos não-agentivos e não intencionais. O segundo *cluster* agrupa o que Vendler denominou de Atividade e *Accomplishment*. Essas duas classes denotam processos que se desenvolvem no tempo, seja sem ou com um ponto télico (um ponto final ou de culminância do evento). Nos dados, eles ocorrem tipicamente associados com sujeitos sintáticos agentivos e intencionais. O *cluster* 3 denota o que Vendler chama de estado, o que equivale a uma eventualidade que se mantém inalterada em um determinado intervalo temporal.

A partir dessa divisão, é possível verificar a frequência relativa da sintaxe oracional em cada *cluster*. A Figura 8 mostra que cada grupo pode

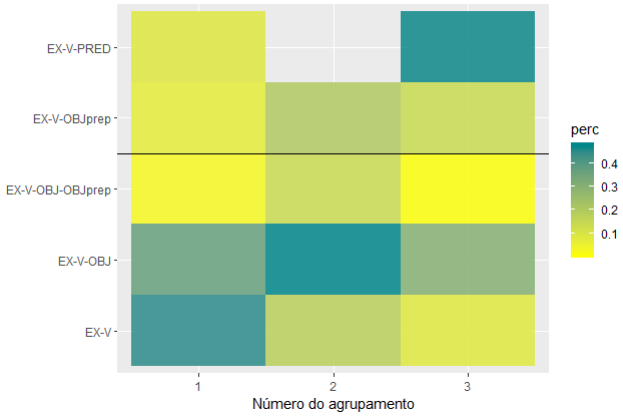

Figura 8: Frequência relativa da sintaxe oracional em cada *cluster*.

ser caracterizado pela predominância de determinadas formas sintáticas. Aqui a sintaxe é representada por um tipo de notação que agrupa unidades oracionais que compartilham certas características e comportamentos sintáticos semelhantes. Utilizou-se a notação "EXT" (argumento externo) como uma variável que representa o que é identificado classicamente como o sujeito sintático, independentemente da posição que ocupa na oração. Isso também inclui a desinência verbal, que em Português marca as noções gramaticais de sujeito, de pessoa e de número. O argumento externo pode representar também um sintagma fora do escopo da unidade oracional que tem um papel semântico associado ao verbo e à sua construção. O símbolo "V" representa uma unidade verbal, o que inclui não apenas o verbo, mas também perífrases aspectuais e modais, assim como construções compostas por verbos leves. Por fim, o símbolo "PRED" representa um sintagma predicativo, "OBJ" objeto direto não preposicionado e "OBJprep", objeto indireto ou preposicionado.

O primeiro *cluster* (eventos pontuais que expressam uma mudança, mais ou menos súbita, de um estado para outro, tipicamente não agentivos e não intencionais) apresenta predominantemente estruturas do tipo "EX-V" como em (15) e (16). Mas podem ocorrer também formas "EX-V-OBJ", como em (17), (18) e (19) especialmente envolvendo verbos de percepção (como ver e ouvir):

(15) quatro pessoas morreu

(16) depois a intimação estourô,

(17) ela também viu ele.

(18) Já ouví passá uma sombra

(19) Eu ganhei trinta mil reais

O segundo *cluster* (processos que se desenvolvem no tempo, com ou sem um ponto télico, tipicamente agentivos e intencionais) apresenta predominantemente estruturas do tipo "EX-V-OBJ" como em (20) e (21), mas ocorrem também com objetos preposicionados, como em (22) e (23). Em orações bitransitivas, como (23), o objeto preposicionado frequentemente marca o ponto télico do evento:

(20) fiquei apertando esse ossinho

(21) Aí eu preparei minhas mala toda.

(22) eu tratava das criação

(23) que ele me levô pro interior

Por fim, o terceiro *cluster* (eventualidade que se mantém inalterada em um determinado intervalo temporal, tipicamente não agentivas e não intencionais) é composto principalmente por orações com predicativos do sujeito como em (24) e (25), mas também com algumas ocorrências de estruturas do tipo "EX-V-OBJ", principalmente com o verbo "ter", como em (26) e certos verbos psicológicos como em (27):

(24) Eu tô doida

(25) E ele era evangélico,

(26) eu tenho marido,

(27) a psicóloga que sabe tudo,

5 Conclusão

Os resultados quantitativos apresentados nessa pesquisa mostram que as unidades oracionais, no português do Brasil, podem ser agrupadas em termos de parâmetros da transitividade, revelando a presença de três microcenários narrativos, semanticamente específicos, sobre os quais se desenrola o evento expresso. Apesar de não haver uma associação perfeita entre sintaxe e esses microcenários, é possível perceber a predominância relativa de certas estruturas sintáticas associadas a cada padrão semântico. Esse tipo de análise corrobora a hipótese adotada por diversos autores da Linguística Cognitiva (Brandt, 2004; Goldberg, 1995; Radden & Dirven, 2007) de que existe uma relação entre o núcleo conceitual de um determinado evento e a forma como ele é expresso em construções gramaticais.

Cada *cluster* analisado revela um tipo de significado protoconceptual, o que inclui traços aspectuais e actanciais próprios, que introduz as categorias lexicais da oração em uma cena ou cenário dinâmico. Essa noção de cenas predicativas, que vem desde Tesnière (1959), tem sido amplamente reconhecida no âmbito da Linguística Cognitiva.

Não existe um consenso na literatura, mesmo com o extenso debate produzido sobre o assunto, em relação a quais seriam essas cenas associadas à sintaxe oracional e como elas podem ser descritas em termos de valores semânticos. O desafio teórico é a demonstração de regras gerais das operações sintáticas, uma vez que os efeitos de significação que elas produzem são enormemente variados. Mas se tomarmos o caminho inverso, ao analisar a semântica de maneira relativamente independente da sintaxe, fica evidente que esses cenários micro-narrativos existem enquanto um grupo de certos traços associados. A questão central é que esses cenários aparecem correlacionados a certos padrões sintáticos, mas não são exclusivos destes últimos. Diferentes padrões sintáticos podem acomodar um mesmo padrão semântico geral, impondo a este ultimo, possivelmente, particularidades.

A metodologia estatística de análise de agrupamentos mostrou-se uma ferramenta útil para se captarem esses padrões semânticos, chegando a resultados semelhantes às categorias aspectuais teoricamente conhecidas e mostrando, além disso, como essas categorias aspectuais se relacionam com categorias actanciais de agentividade e intencionalidade. Sob uma nova perspectiva, esse tipo de metodologia pode ser útil na investigação dos padrões semânticos a que os falantes são expostos, de maneira relativamente independente da sintaxe, e sugere uma arquitetura específica sobre a qual a língua se organiza.

Referências

Bois, John W. Du. 2003. *Preferred argument structure: Grammar as architecture for function*. John Benjamins.

Brandt, Per Aage. 2004. Dynamic schematism and the cognitive semantics of language. `https://case.edu/artsci/cogs/larcs/documents`, accessed on 19 March, 2019.

Broccias, Cristiano. 2013. Cognitive grammar. Em Thomas Hoffmann e Graeme Trousdale (ed.), *The Oxford handbook of construction grammar*, 149–161. Oxford University Press.

Comrie, Bernard. 1976. *Aspect: An introduction to the study of verbal aspect and related problems*, vol. 2. Cambridge University Press.

Divjak, Dagmar & Nick Fieller. 2014. Cluster analysis: Finding structure in linguistic

data. Em Dylan Glynn & Justyna Robinson (eds.), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*, 405–441. John Benjamins Publishing Company. 10.1075/hcp.43.16div.

Givón, Talmy. 2001. *Syntax: An introduction*, vol. 1. John Benjamins Publishing.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Gower, John C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 857–871. 10.2307/2528823.

Hair, Joseph F., William C. Black, Barry J. Babin & Rolph E. Anderson. 2009. *Multivariate data analysis*. Prentice-Hall.

Halkidi, Maria, Yannis Batistakis & Michalis Vazirgiannis. 2001. On clustering validation techniques. *Journal of intelligent information systems* 17(2-3). 107–145. 10.1023/A:1012801612483.

Halliday, Michael Alexander Kirkwood, Christian Matthiessen & Michael Halliday. 2014. *An introduction to functional grammar*. Routledge.

Hopper, Paul J. & Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 251–299. 10.2307/413757.

Johnson, Mark. 2013. *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago Press.

Lakoff, George. 2008. *Women, fire, and dangerous things*. University of Chicago Press.

Lepesqueur, Marcus. 2017. *Transitividade na esquizofrenia: comparação dos relatos orais de eventos psicóticos entre grupos clínico e não clínico*: Universidade Federal de Minas Gerais. Tese de Doutoramento.

Lima, Lucia Chaves de Oliveira. 2013. *A transitividade na conversação: uma abordagem cognitivo-funcional*: Universidade Federal do Rio Grande do Norte. Tese de Mestrado.

Lucena, Nedja Lima & Maria Angélica Furtado Cunha. 2012. Relações de herança em orações transitivas: O mecanismo de extensão metafórica. *Letras & Letras* 27(1).

Mingoti, Sueli Aparecida. 2017. *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. UFMG.

Næss, Åshild. 2007. *Prototypical transitivity*, vol. 72. John Benjamins Publishing.

R Development Core Team. 2017. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. http://www.R-project.org. ISBN 3-900051-07-0.

Radden, Günter & René Dirven. 2007. *Cognitive English Grammar*, vol. 2. John Benjamins Publishing.

Rousseeuw, Peter J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20. 53–65. 10.1016/0377-0427(87)90125-7.

Rousseeuw, Peter J. & L. Kaufman. 1990. *Finding groups in data: An introduction to cluster analysis*. Wiley Online Library.

Rozas, Victoria Vázquez. 2004. Transitividad prototípica y uso. *Boletín de Lingüística* 92–115.

Shahrokhi, Mohsen & Ahmad Reza Lotfi. 2012. Manifestation of transitivity parameters in persian conversations: a comparative study. *Procedia-Social and Behavioral Sciences* 46. 635–642. 10.1016/j.sbspro.2012.05.176.

Sokal, Robert R. & Peter H. A. Sneath. 1963. *Principles of numerical taxonomy*. W. H. Freeman.

Tesnière, Lucien. 1959. *Eléments de syntaxe structurale*. Klincksieck.

Thompson, Sandra A & Paul J. Hopper. 2001. Transitivity, clause structure, and argument structure: evidence from conversation. Em *Frequency and the emergence of linguistic structure*, vol. 45, 27–60. John Benjamins.

Varela, Francisco, Evan Thompson & Eleonor Roch. 1991. *A mente corpórea: ciência cognitiva e experiência humana*. Instituto Piaget.

Vendler, Zeno. 1967. *Linguistics in philosophy*. Cornell.

Proposta recebida em Outubro de 2019 e aceite para publicação em Novembro de 2019.

Identificação automática de unidades de informação em testes de reconto de narrativas usando métodos de similaridade semântica

Automatic identification of information units in tests based on narrative retelling using semantic similarity methods

Leandro Borges dos Santos
Universidade de São Paulo
leandrobs@usp.br

Sandra Maria Aluísio
Universidade de São Paulo
sandra@icmc.usp.br

Resumo

Os diagnósticos da Doença de Alzheimer (DA) e do Comprometimento Cognitivo Leve (CCL) baseiam-se na análise das funções cognitivas do paciente pela administração de baterias de avaliação cognitiva e neuropsicológica. O emprego do reconto de narrativas é comum para auxiliar a identificação e quantificação do grau de demência: é atribuído um ponto para cada unidade recordada, e o escore final representa a quantidade de unidades recordadas. Avaliamos duas tarefas da área clínica: a identificação automática de quais elementos de uma narrativa recontada foram recordados; e a classificação binária da narrativa produzida por um paciente, tendo as unidades identificadas como atributos, visando uma triagem automática dos pacientes com comprometimentos cognitivos. Utilizamos dois conjuntos de dados de reconto transcritos que possuem as sentenças divididas e anotadas manualmente com as unidades de informação e os disponibilizamos publicamente. São eles: a Bateria Arizona para Desordens de Comunicação e Demência (*ABCD*) com narrativas de pacientes com CCL e Controles Saudáveis e a Bateria de Avaliação da Linguagem no Envelhecimento (BALE), com narrativas de pacientes com DA e CCLs, e Controles Saudáveis. Avaliamos dois métodos baseados em similaridade semântica, chamados de *STS* e *Chunking*, e transformamos o problema multirrótulo de identificação de elementos de uma narrativa recontada em problemas de classificação binária, encontrando um ponto de corte para o valor de similaridade de cada unidade de informação. Dessa forma, conseguimos superar dois *baselines* para os dois conjuntos de dados na métrica *SubsetAccuracy*, que é a mais punitiva para o cenário multirrótulo. Na classificação binária nem todos os seis métodos de aprendizado de máquina avaliados tiveram melhor desempenho do que os baselines de identificação de unidades de informação. Para a *ABCD*, os melhores métodos foram Árvores de Decisão e *KNN*, e para a BALE, o *SVM* com *kernel RBF*.

Palavras chave

testes neuropsicológicos, reconto de narrativas, métodos de similaridade semântica

Abstract

Diagnoses of Alzheimer's Disease (AD) and Mild Cognitive Impairment (CCL) are based on the analysis of the patient's cognitive functions by administering cognitive and neuropsychological assessment batteries. The use of retelling narratives is common to help identify and quantify the degree of dementia. In general, one point is awarded for each unit recalled, and the final score represents the number of units recalled. In this paper, we evaluated two clinical tasks: the automatic identification of which elements of a retold narrative were recalled; and the binary classification of the narrative produced by a patient, having the units identified as attributes, aiming at an automatic screening of patients with cognitive impairment. We used two transcribed retelling data sets in which sentences were divided and manually annotated with the information units. These data sets were then made publicly available. They are: the Arizona Battery for Communication and Dementia Disorders (*ABCD*) that contains narratives of patients with CCL and Healthy Controls and the *Avaliação da Linguagem no Envelhecimento* (BALE), which includes narratives of patients with AD and CCLs as well as Healthy Controls. We evaluated two methods based on semantic similarity, referred to here as STS and Chunking, and transformed the multi-label problem of identifying elements of a retold narrative into binary classification problems, finding a cutoff point for the similarity value of each information unit. In this way, we were able to overcome two baselines for the two datasets in the SubsetAccuracy metric, which is the most punitive for the multi-label scenario. In binary classification, however, not all six machine learning methods evaluated performed better than the baselines methods. For ABCD, the best methods were Decision Trees and KNN, and for BALE, SVM with RBF kernel stood out.

Keywords

neuropsychological tests, narrative retellings, semantic similarity methods

DOI: 10.21814/lm.11.2.304

1 Introdução

O envelhecimento da população é uma tendência social conhecida em países desenvolvidos e que tem se tornado cada vez mais pronunciada também nos países em desenvolvimento (Fichman et al., 2011). O Brasil, por exemplo, está mudando sua pirâmide etária, segundo os censos do Instituto Brasileiro de Geografia e Estatística (IBGE) de 2000 e 2010[1].

A maior expectativa de vida é um bem desejável, porém o envelhecimento pode ser acompanhado de doenças neurodegenerativas, como as demências, dentre as quais a Doença de Alzheimer (DA) é a mais proeminente, correspondendo a $50 - 80\%$ dos casos (Abbott, 2011). Assim, as demências são consideradas pela Organização Mundial de Saúde como um desafio para as próximas décadas, devido aos seus custos sociais e econômicos (Wortmann, 2012). Outra enfermidade que tem recebido atenção nos últimos anos é o Comprometimento Cognitivo Leve (CCL), que ocasiona declínio em funções cognitivas, podendo progredir para um quadro demencial. Assim, o CCL tem sido descrito como uma condição pré-clínica da DA (Clemente & Ribeiro-Filho, 2008; Frota et al., 2011). Mas, em alguns casos, o quadro pode se reverter para um estado normal e compatível com indivíduos da mesma faixa etária e nível de escolaridade, sendo melhor definido como uma condição heterogênea (Frota et al., 2011).

O diagnóstico das demências e síndromes relacionadas, comumente, baseia-se na análise das funções cognitivas do paciente, pela administração de baterias de avaliação cognitiva e neuropsicológica (McKhann et al., 2011; Frota et al., 2011; Mapstone et al., 2014; Hübner et al., 2019). As baterias avaliam as funções que são mais afetadas como diferentes tipos de memória, orientação, linguagem e resolução de problemas. Estas baterias são usadas antes, durante e depois de tratamentos, como diagnóstico, acompanhamento e direcionamento de tratamento (de Abreu et al., 2005). Como exemplos de baterias e testes temos: o teste Memória Lógica da Wechsler Memory Scale (Wechsler, 1997), a Bateria Montreal de Avaliação da Comunicação (Nasreddine et al., 2005), a Bateria Arizona para Desordens da Comunicação e Demência (ABCD) (Bayles & Tomoeda, 1993), o teste de Boston para o Diagnóstico da Afasia (Goodglass & Kaplan, 1983), dentre outros.

O emprego do reconto de narrativas é comum para auxiliar a identificar e quantificar o grau de demência. Em geral, as tarefas de reconto de narrativas utilizam uma história curta que é contada ao paciente, a quem se solicita que reconte a história imediatamente após ouvi-la com o máximo de detalhes. Em alguns casos, é solicitado ao paciente recontar novamente após 30 minutos. O reconto é gravado para posterior transcrição e análise.

Como pacientes com quadros demenciais tendem a possuir um vocabulário e usar estruturas sintáticas mais simples, são analisados os aspectos lexicais e sintáticos dos recontos. Também é possível mensurar a capacidade de memória de um paciente, por isso a narrativa é dividida em unidades de informação, podendo ser palavras ou orações. Em geral, é atribuído um ponto para cada unidade recordada, e o escore final representa a quantidade de unidades recordadas. As principais desvantagens dessa análise são: (i) a demanda de tempo, por ser uma tarefa de avaliação manual; (ii) a subjetividade do avaliador na checagem da presença das unidades de informação da narrativa no reconto. Assim, torna-se bem-vinda e importante a aplicação de métodos computacionais tanto para a automatização dessa tarefa, o que viabiliza sua aplicação em larga escala, como para a manutenção da uniformidade na correção.

Entretanto, há desafios computacionais também para a automatização do cálculo do escore por um sistema computacional. O sistema deverá resolver vários fenômenos que são comuns para essa tarefa, por exemplo: mudança da ordem de palavras da história original, uso de palavras similares às da história original, comentários que não estão relacionados com a história, e disfluências que tornam a história recontada bastante diferente da original.

Na Figura 1 (a), apresentamos a história do teste do reconto da bateria *ABCD*, traduzida para o português. Ela possui 5 sentenças e 61 palavras. Em (b), a mesma história é dividida em 17 unidades de informação, com possíveis alternativas entre parênteses, sendo 17 a sua pontuação máxima. Em (c), apresentamos um reconto imediato de um paciente com CCL, com pontuação 12, pois a avaliação manual de seu reconto contabilizou 12 unidades relembradas. Neste reconto, há trechos com disfluências ((1) e (3)), duplicação de unidades de informação recontadas ((1) e (3); (5) e (6)) e comentários não relacionados com a história ((2), (10) a (13)).

[1] https://censo2010.ibge.gov.br/sinopse/webservice/frm_piramide.php

(a) Enquanto uma senhora fazia compras, sua carteira caiu da bolsa, mas ela não viu. Quando ela foi ao caixa, não tinha como pagar as compras. Então, ela colocou as compras de lado e foi para casa. Assim que ela abriu a porta da casa, o telefone tocou e uma menininha disse-lhe que tinha achado a carteira. A senhora ficou muito aliviada.
(b) **Senhora (mulher)** // **estava fazendo compras (na loja, foi às compras, foi ao mercado)** // Sua carteira (seu porta-notas, sua moedeira) // **carteira caiu (derrubou a carteira, perdeu a carteira, perdeu a bolsa)** // da sua bolsa (da sua mochila, de sua pasta) // **Ela não viu a carteira cair (ela não notou)** // No caixa (quando ela foi pagar, guichê) // **não tem como pagar (ela não tinha dinheiro, não tinha sua carteira)** // Coloca as mercadorias de lado (coloca as mercadorias de volta) // **foi para sua casa (voltou para sua casa)** // Quando ela abriu a porta (quando ela chegou em casa, assim que ela entrou) // **telefone tocou (fone tocou, ela recebeu uma ligação)** // Pequena (jovem) // menina (garota) // lhe disse (falou, contou) // **ela achou a carteira (achou sua moedeira, achou o porta-notas)** // **Senhora aliviada (senhora estava feliz, senhora estava radiante, senhora estava agradecida)**
(c) (1) ahm uma senhora foi fazer compras no me foi no mercado. (2) não lembrava o local. (3) no me fazer compras. (4) e quando ela foi pagar a conta no caixa percebeu que estava sem a carteira. (5) aí ela foi deixou a mercadoria. (6) não levou a mercadoria. (7) voltou para casa. (8) chegando em casa toca o telefone. (9) era uma garotinha avisando ela que que tinha achado a carteira. (10) é isso. (11) tem mais coisa. (12) não cortei. (13) eu resumi o que eu ouvi.

Figura 1: (a) Narrativa original da bateria ABCD; (b) Narrativa original separada em unidades de informação; as nove unidades marcadas em negrito são as principais, o resto são detalhes; (c) Transcrição do reconto imediato de um paciente com CCL, segmentada manualmente em sentenças.

Existem poucos trabalhos na literatura que tratam da automatização da identificação de unidades de informação em recontos de narrativas. Podemos dividi-los em: métodos de busca de palavras (Pakhomov et al., 2010; Fraser et al., 2016), métodos de alinhamento (Prud'hommeaux & Roark, 2015), e métodos de *clustering* (Yancheva & Rudzicz, 2016; Fraser et al., 2019).

Neste artigo, avaliamos automaticamente quais elementos de uma narrativa recontada foram recuperados, utilizando dois métodos baseados em similaridade semântica. Esses elementos são usados como atributos para métodos de classificação binária da narrativa produzida por um paciente realizando um teste neuropsicológico baseado em reconto. No melhor do nosso conhecimento, não há trabalhos na literatura sobre a identificação das unidades de informação em recontos modelada com métodos de similaridade semântica.

O restante deste artigo é organizado do seguinte modo: na Seção 2 são apresentadas as principais características do diagnóstico da Doença de Alzheimer e do Comprometimento Cognitivo Leve (Seção 2.1) e uma descrição dos trabalhos sobre identificação automática de unidades de informação em recontos (Seção 2.2). Na Seção 3, são descritos os córpus utilizados neste estudo, os métodos de similaridade semântica propostos e as baselines; já na Seção 4, mostramos os resultados dos experimentos para a classificação das unidades de informações nos dois datasets avaliados neste artigo. Na Seção 5, são apresentados os resultados dos métodos de classificação automática de narrativas que se basearam nos atributos recuperados automaticamente pelos métodos descritos na Seção 3. Por fim, na Seção 6, trazemos as conclusões e apresentamos sugestões de trabalhos futuros.

2 Trabalhos relacionados

2.1 Diagnóstico de Demências e Síndromes Relacionadas

O envelhecimento acarreta algumas perdas de funcionalidades, como a capacidade motora, a diminuição dos mecanismos de defesa natural do organismo e da adaptação ao ambiente. Acarreta, também, a diminuição de funcionalidades cognitivas, como a linguagem, que tem um papel fundamental na vida das pessoas, possibilitando a comunicação e as demais atividades sociais. Essas modificações não ocorrem de forma isolada e sim estão relacionadas com as alterações na memória operacional, na atenção e nas habilidades visuoespacial (Freitas, 2010). Nos idosos, são notadas alterações dos padrões discursivos conforme o estímulo. Para tarefas nas quais é exigido o reconto de narrativas ouvidas recentemente, são obtidos textos curtos e simples. Ao contrário das tarefas em que é necessária a produção de narrativas livres, elicitadas por meio de estímulo visual de um livro de figuras, nas quais os idosos tendem a elaborar textos mais longos, mas contendo um número maior de informações irrelevantes e com baixa coesão (Garcia & Mansur, 2006). Nesta seção, apresentamos as principais características do diagnóstico da Doença de Alzheimer e do Comprometimento Cognitivo Leve.

2.1.1 Doença de Alzheimer

No Brasil, as recomendações para o diagnóstico da DA foram elaboradas em 2011, pelos membros do Departamento de Neurologia Cognitiva e do Envelhecimento da Academia Brasileira de Neurologia (Frota et al., 2011). Seguem abaixo os critérios clínicos principais para o diagnóstico de demência de qualquer tipo:

1. Demência é diagnosticada quando há sintomas cognitivos ou comportamentais (neuropsiquiátricos) que: (i) Interferem com a habilidade no trabalho ou em atividades usuais; (ii) Representam declínio em relação a níveis prévios de funcionamento e desempenho; (iii) Não são explicáveis por delirium (estado confusional agudo) ou doença psiquiátrica maior;
2. O comprometimento cognitivo é detectado e diagnosticado mediante combinação de (i) Anamnese com paciente e informante e (ii) Avaliação cognitiva objetiva, mediante exame breve do estado mental ou avaliação neuropsicológica;
3. Os comprometimentos cognitivos ou comportamentais afetam no mínimo dois dos seguintes domínios: memória, funções executivas, habilidades visuoespaciais, linguagem e personalidade/comportamento.

O declínio cognitivo progressivo é confirmado com exames sucessivos e a positividade de biomarcadores. Também são utilizados exames de imagens para exclusão de outros diagnósticos.

Mesmo que a perda de memória seja a característica mais frequente, alterações na linguagem também podem aparecer nos estágios iniciais da DA. Uma das formas de se avaliar a linguagem é a produção de narrativas, sendo observado que estas narrativas apresentam sentenças simples e curtas, maior número de proposições irrelevantes, vocabulário pobre, ruptura no desenvolvimento do tema, maior número de erros ortográficos e menor nível de complexidade sintática (Mansur et al., 2005).

2.1.2 Comprometimento Cognitivo Leve

Para a identificação do comprometimento cognitivo leve são utilizados testes neuropsicológicos, por serem mais sensíveis, embora não exista uma norma para o valor do ponto de corte. Essa dificuldade decorre pelo fato de ser uma situação entre o envelhecimento normal e a demência. Frota et al. (2011) sugerem as principais características utilizadas para identificação do CCL:

- Queixa de alteração cognitiva relatada pelo paciente ou informante próximo;
- Evidência de comprometimento cognitivo em um ou mais dos seguintes domínios: memória, função executiva, linguagem e habilidades visuoespaciais;
- Preservação da independência funcional;
- Não preenche critérios de demência.

Recentemente, há evidências de que indivíduos com CCL têm mais risco para desenvolver DA, devido a comprometimentos em múltiplos domínios, incluindo a linguagem. Por essa razão, é importante compreender a natureza do comprometimento de linguagem. Em Fleming & Harris (2008) foi realizado um estudo comparativo do discurso produzido por pacientes com CCL e idosos normais, observando que o discurso produzido por pacientes com CCL contém um número menor de palavras, e as suas características se comparam com os estágios iniciais de DA. Enquanto que em Chapman et al. (2002) foram comparadas as habilidades de compreensão, memória e expressão de texto discursivo extenso, identificando que a capacidade de fornecer informações detalhadas e realizar síntese de ideais a partir das narrativas estava comprometida quando comparadas com as habilidade dos pacientes normais, sendo muito similar à de pacientes acometidos por DA. Hodges et al. (1996) examinaram o desempenho de controles saudáveis e indivíduos com diversos graus de comprometimento de DA em tarefas de nomeação e geração de definições e reconheceram que a qualidade da definição produzia diferenças entre os grupos. Os estudos sobre descrição (oral e escrita) de figuras simples e complexas realizados por Forbes-McKay & Venneri (2005) também distinguem indivíduos com DA em grau leve e indivíduos saudáveis.

Em resumo, para avaliar a linguagem de indivíduos com CCL é importante dispor de instrumentos/testes sensíveis para detectar déficits sutis. Além disso, o monitoramento dessas dificuldades também carece de instrumentos acurados. A análise do discurso mostra-se interessante, pois abrange os diferentes componentes da linguagem, em uma perspectiva linguístico-cognitiva.

2.2 Métodos de Identificação Automática das Unidades de Informação em Recontos

Nesta seção, organizamos a descrição dos métodos de identificação automática das unidades de informação em recontos da literatura em três abordagens: métodos de busca de palavras (Pakhomov et al., 2010; Fraser et al., 2016), métodos de alinhamento (Prud'hommeaux & Roark, 2015), e métodos de *clustering* (Yancheva & Rudzicz, 2016; Fraser et al., 2019).

2.2.1 Métodos de Busca de Palavras

Pakhomov et al. (2010) compilaram uma lista com palavras e frases que representavam algum

conceito da cena do Roubo do Biscoito, que é uma subtarefa da Bateria de Boston (*Boston Diagnostic Aphasia Examination —BDAE*) (Goodglass & Kaplan, 1983). As narrativas foram divididas em *n-grams*, de 1 a 4, e para cada *n-gram* os autores realizaram uma busca na lista de palavras. Se o *n-gram* era encontrado, considerou-se que o paciente se relembrou dessa unidade de informação.

Os autores utilizaram 38 narrativas de idosos com Degeneração Lobar Frontotemporal, com o seguintes subtipos: Afasia Progressiva Primária, Demência Semântica, variante comportamental da Demência Frontotemporal, e Afasia Logopênica. Entretanto, não encontraram diferença estatisticamente significativa entre os grupos na contagem de unidades de informação recordadas.

Fraser et al. (2016) utilizaram uma lista de palavras para cada possível unidade de informação. Para as unidades de informação que representam uma ação, os autores utilizaram o *parser* de *Stanford* para identificar o verbo e o sujeito, e analisaram se essa combinação estava na lista de palavras. As unidades de informação foram utilizadas como atributos binários em conjunto com métricas de PoS, de complexidade sintática, psicolinguísticas, de diversidade lexical, de constituintes gramaticais, de repetividade de informações, e acústicas, totalizando 370 atributos. O objetivo dos autores foi distinguir narrativas de pacientes com Doença de Alzheimer e envelhecimento saudável no conjunto de dados *DementiaBank* (Becker et al., 1994), neste conjunto os pacientes são solicitados a descrever a cena do Roubo do Biscoito (Goodglass & Kaplan, 1983). Os autores utilizaram 233 narrativas de 97 participantes com envelhecimento saudável e 240 narrativas de 168 participantes com possível ou provável DA. Para a classificação final, usaram o algoritmo de Regressão Logística, *10-fold-cross-validation*, e a métrica acurácia para avaliação, dado que a classificação era binária. O melhor resultado foi 0,819 de acurácia, utilizando 35 atributos selecionados com o método de Correlação de Pearson.

2.2.2 Métodos de Alinhamento

Prud'hommeaux & Roark (2015) propuseram um método de alinhamento baseado em grafos, utilizando a técnica de passeios aleatórios (*Random Walks*) para automatizar o teste de reconto de narrativas do teste de Memória Lógica de Wechsler. Na abordagem proposta, cada palavra do reconto ou da narrativa original representa um nó do grafo e o alinhamento entre as palavras representa as arestas. São utilizadas narrativas de 235 pacientes, sendo 72 pacientes com CCL, 163 com envelhecimento saudável, e 48 narrativas de pacientes inelegíveis, i.e., que não se enquadraram em algum critério e não podem fazer parte dos grupos CCL ou Envelhecimento Saudável.

No método proposto, primeiramente, cada narrativa de reconto é alinhada com a narrativa original e as demais narrativas de reconto. Para obter os alinhamentos é utilizado o alinhador *Berkeley Aligner* (Liang et al., 2006). A partir desses alinhamentos é construído um grafo, em que é verificado se o alinhamento possui uma probabilidade maior que 0.5. Neste caso, é adicionado um vértice entre essas palavras. Desse modo, podem existir dois tipos de alinhamentos: o alinhamento com uma palavra da narrativa fonte, e o alinhamento com uma palavra da narrativa de reconto. Dada uma palavra da narrativa de reconto, esta é definida como o vértice inicial da caminhada aleatória. A cada passo da caminhada é gerado um valor aleatório, e caso este seja maior que um λ, é realizada uma transição para uma palavra da narrativa original; caso contrário, a transição é realizada para uma palavra da narrativa de reconto. Quando a caminhada aleatória atingir uma palavra da narrativa fonte, é proposto um novo alinhamento entre a palavra inicial e a palavra fonte, e a caminhada é encerrada.

Para cada palavra presente nas narrativas de reconto são realizados mil passeios aleatórios. O novo alinhamento, entre a palavra de reconto e a palavra fonte, é definido pelo alinhamento mais frequente dos passeios aleatórios. Após a obtenção dos alinhamentos, estes são utilizados como atributos para um classificador final; se alguma palavra da narrativa de reconto estiver alinhada com a narrativa original é considerado que o paciente se recordou desse trecho.

Na tarefa de classificação final (Envelhecimento Saudável *versus* CCL), Prud'hommeaux & Roark (2015) exploram duas representações para cada paciente: (i) o *Summary score* que é a quantidade de unidades de informações recordadas no reconto imediato e tardio; (ii) *Element scores* em que cada unidade de informação representa um atributo. É marcado se o paciente se recordou ou não dessa unidade de informação no reconto imediato e no tardio. O melhor resultado na classificação final, utilizando o método automático com o *Element scores* foi de $0,792$ de AUC, enquanto que utilizando os escores obtidos de forma manual o resultado foi de $0,813$.

2.2.3 Métodos de Clustering

Yancheva & Rudzicz (2016) e Fraser et al. (2019) automatizaram a análise de unidades de informação aplicando algoritmos de agrupamento, em que os *clusters* são considerados como um indicador (*proxy*) para as unidades de informação e são utilizados para extrair atributos. Os detalhes de cada método são explicados a seguir.

Yancheva & Rudzicz (2016) avaliaram o método no *DementiaBank*, utilizaram 241 narrativas de 98 participantes com envelhecimento saudável e 255 narrativas de 168 participantes com possível ou provável DA. Os verbos e os substantivos das transcrições são convertidos em uma representação densa com o método *GloVe* (Pennington et al., 2014); para cada grupo é aplicado o algoritmo *K-means* com a distância euclidiana e k igual a 10. A partir dos *clusters* são criados atributos baseados nas distâncias.

Na tarefa de classificação final, os autores optaram pelo classificador *Random Forest* via *10-fold-cross-validation*. A abordagem proposta foi comparada com o resultado da classificação utilizando uma lista de palavras para recuperar as unidades de informação. Os autores também adicionaram atributos de métricas linguísticas e acústicas. O classificador que utiliza os atributos do modelo de *cluster* do grupo de Envelhecimento Saudável e do grupo de DA obteve $0,74$ de F1, e quando combinado com as métricas linguísticas e acústicas obteve $0,80$. Observa-se que o *baseline* com lista de palavras de cada unidade de informação obteve $0,73$ de F1.

Fraser et al. (2019) substituíram o modelo *GloVe* pelo *FastText* (Bojanowski et al., 2017), possibilitando inferir palavras que não estão presentes no vocabulário do modelo de *embeddings*, e optarem pela distância do cosseno em vez da euclidiana. Foram utilizados três conjuntos de dados: o *DementiaBank*, com 97 participantes saudáveis e 19 participantes com CCL; o *Gothenburg* (Wallin et al., 2016), que é composto por transcrições da descrição da cena do Roubo do Biscoito de pacientes suecos, com 36 participantes saudáveis e 31 com CCL; e o *Karolinska*, que foi coletado por Cromnow & Landberg (2009), tendo sido solicitado a 96 indivíduos com envelhecimento saudável que produzissem uma descrição escrita da cena do Roubo do Biscoito em 5 minutos.

Os autores extraíram todos os verbos e os substantivos das transcrições. Em seguida, as palavras foram transformadas em vetores a partir do alinhamento das matrizes de *word embeddings* em Inglês e Sueco do *FastText*, em seguida aplicaram o algoritmo *k-means* com três variações do parâmetro k, sendo: 10, 23, e k_{sil}, onde $k_{sil} \in \{2, 3, \ldots, 30\}$. Para k_{sil} o valor é selecionado de forma automática pelo método da silhueta (Kaufman & Rousseeuw, 2009). Para cada configuração de k foram construídos modelos de agrupamento para o Inglês, Sueco, e uma versão multilíngue (Inglês e Sueco). Após a obtenção dos agrupamentos, foram extraídos atributos baseados nas distâncias em relação aos centroides.

Para a etapa de classificação, os autores utilizaram o *SVM* linear e o *leave-one-out*, e avaliaram acurácia no conjunto de dados do *DementiaBank* balanceado, e *Gothenburg*. O conjunto de dados *Karolinska* e 78 participantes saudáveis restantes do *DementiaBank* foram utilizados no treinamento dos modelos de agrupamentos. Para cada iteração do *leave-one-out*, os autores executaram um *inner-cross-validation* para selecionar os parâmetros do *SVM*, e selecionaram o modelo de agrupamento a partir de 10 execuções.

No *DementiaBank*, o melhor resultado foi o modelo multilíngue com k igual a 10, que obteve uma acurácia de $0,63$; para o modelo de agrupamento monolíngue a melhor acurácia foi de $0,47$ com k igual a 10 e k_{sil}. No *Gothenburg*, o melhor resultado foi de $0,72$ com o modelo de multilíngue, e k igual a 23, enquanto que o melhor resultado do modelo monolíngue foi de $0,55$ com k igual a 10. Além disso, os autores avaliaram no mesmo cenário de Yancheva & Rudzicz (2016), ou seja, identificação de pacientes com DA *versus* idosos saudáveis, para k igual a 10, e obtiveram F1 score de $0,83$, enquanto que Yancheva & Rudzicz (2016) obtiveram $0,74$.

3 Desenvolvimento

Na Seção 3.1, são apresentados os dois córpus utilizados neste trabalho, bem como a metodologia proposta para compilá-los. Na Seção 3.2, são descritos dois métodos de identificação automática de unidades de informação, baseados em métodos de similaridade semântica. Na Seção 3.3, são descritas as *baselines* para comparação de desempenho dos métodos da Seção 3.2.

3.1 Conjuntos de dados utilizados

Utilizamos dois conjuntos de dados de reconto que possuem as sentenças anotadas manualmente com as unidades de informação, descritos em Santos et al. (2019). Os conjuntos estão disponíveis para download no GitHub[2].

[2] https://github.com/lbsantos/ANAA-Dementia/tree/master/conjutos_de_dados

A Tabela 1 apresenta as estatísticas dos dois conjuntos, que trazem uma média do tamanho de sentenças bem próxima entre CCLs e Controles na Bateria Arizona para Desordens de Comunicação e Demência (*ABCD*), mas uma diferença maior dos grupos DA e CCL com o grupo de Controle da Bateria de Avaliação da Linguagem no Envelhecimento (BALE) (Hübner et al., 2019).

O primeiro conjunto de dados é formado por transcrições da *ABCD* que é composta de 17 subtestes. Nos interessa neste trabalho o subteste de reconto no qual é contada uma história ao paciente, e este tem quer recontar a história imediatamente e depois de 30 minutos. O teste do reconto foi aplicado em 23 idosos com CCL e 12 adultos com envelhecimento saudável (Controles), na Faculdade de Medicina da USP. Este teste possui 17 unidades de informação, apresentadas na Figura 1 (b), com possíveis alternativas entre parênteses; a sua pontuação máxima é 17.

O segundo conjunto de dados é formado por transcrições da BALE que possui diversas tarefas, sendo uma delas o reconto de uma história apresentada oralmente (História da Lúcia). A História da Lúcia possui originalmente 24 unidades de informação que foram reagrupadas neste trabalho, resultando em 21 unidades (Figura 2). O teste do reconto foi aplicado em 11 idosos com DA, 5 idosos com CCL e 53 adultos com envelhecimento saudável (Hübner et al., 2019).

Lúcia // **mora** // **interior** // do Paraná // Numa manhã de 2a feira // **ela saiu de casa** // **para buscar emprego (foi para uma uma entrevista, foi buscar trabalho)** // na capital do estado (em Curitiba) // **Foi para rodoviária** // **foi de carona (pegou carona)** // **com amigo Pedro (com Pedro)** // Estava chovendo // naquela manhã // O carro // passou (caiu) // por um buraco // **o pneu furou** // Pensou que ia perder (achou que ia perder) // o ônibus // **Pegou um táxi** // **conseguiu chegar chegou a tempo (chegou a tempo)**

Figura 2: Narrativa utilizada na BALE, separada em unidades de informação; as onze unidades principais são marcadas em negrito.

Nas Figuras 1 e 2, anotamos as unidades da macroestrutura em negrito, seguindo o modelo de análise de Kintsch & van Dijk (1978) em que as unidades de informação do texto são organizadas de forma hierárquica, sendo a macroestrutura correspondente às ideias principais e a microestrututura às ideias acessórias e detalhes.

Para cada conjunto de dados, o áudio do participante foi transcrito manualmente, seguindo os princípios do NURC / SP No 338 EF e 331 D (Preti, 2005) e segmentado manualmente em sentenças por um anotador experiente, usando conhecimento prosódico (pausas), sintático e semântico. Optamos por utilizar uma segmentação manual para isolarmos os efeitos de erros de um sistema de segmentação automática. Embora Treviso & Aluísio (2018) tenham desenvolvido um sistema de segmentação sentencial para narrativas elicitadas com estímulo visual (livros de figuras), este ainda não consegue generalizar para narrativas elicitadas com estímulos orais (recontos).

Para criarmos os conjuntos de dados anotados com as unidades de informação sobre as unidades de interesse (sentenças anotadas no pré-processamento), utilizamos o sistema de anotação *brat* (*brat rapid annotation tool*) (Stenetorp et al., 2012), realizando a anotação em duas fases. Na primeira fase, cada sentença da transcrição foi classificada de acordo com a lista de unidades de informação de cada bateria por um único anotador; na segunda fase, outro anotador revisou a anotação e os casos discordantes foram discutidos, visando obter uma anotação concordante.

A narrativa da *ABCD* foi mantida com as 17 unidades de informação originais, mas para a narrativa da BALE, realizamos algumas modificações nas unidades de informação (ora separando, ora juntando) para termos uma anotação manual uniforme, sem discrepâncias e possibilitar a aplicação de métodos automáticos. A partir dessas modificações, finalizamos com 21 unidades de informação (Figura 2) em vez das 24 unidades originais, com 11 delas sendo unidades macroestruturais.

3.2 Modelando a identificação de unidades de informação via similaridade semântica

Métodos de similaridade semântica textual (*STS*, *Semantic Textual Similarity*) e inferência textual (*RTE*, *Recognizing Textual Entailment*) têm aplicações em diversas tarefas de Processamento de Línguas Naturais como: recuperação de informação, sistemas de perguntas-repostas, avaliação de sistemas de tradução, dentre outras (Agirre et al., 2012, 2015).

Na tarefa de *STS* o objetivo é indicar o grau de similaridade entre dois textos, ou seja, dado um par de textos $(\mathbf{S}_i^1, \mathbf{S}_i^2)$, estamos interessados em atribuir um valor y_i em alguma escala, geralmente de 0 a 5 ou 1 a 5 (Agirre et al., 2012, 2015; Marelli et al., 2014; Fonseca et al., 2016). Essa gradação naturalmente captura as diferenças sutis de similaridade, como sentenças que possuem o mesmo significado (pontuação 5), possuem pequenas diferenças semânticas (pontuação 4), compartilham apenas alguns detalhes

Bateria	Grupo	Sujeitos	Média Sentenças (Desvio Padrão)	Média de palavras por sentença (Desvio Padrão)
ABCD	CCL	23	8,17 (1,92)	60,76 (17,39)
	Controle	12	7,67 (2,06)	58,96 (14,73)
BALE	DA	11	6,09 (2,63)	36,18 (17,10)
	CCL	5	6,00 (1,00)	36,40 (5,68)
	Controle	53	7,68 (2,67)	52,06 (19,18)

Tabela 1: Estatísticas dos Conjuntos de Dados.

(pontuação 3), sentenças não relacionadas, mas versam sobre o mesmo assunto (pontuação 2), ou mesmo que não possuem nada em comum (pontuação 1).

Já o *RTE* pode ser definido como uma relação direcional entre dois textos, em que dado um texto **T** permite-se que se conclua que uma hipótese **H** é verdadeira (Dagan et al., 2006; Marelli et al., 2014; Fonseca et al., 2016). Com essa definição é assumido que pessoas lendo o par $(\mathbf{T}, \mathbf{H})$ compartilham: (i) o conhecimento da língua em que os textos são formulados, e (ii) o mesmo conhecimento prévio sobre o tema (Dagan et al., 2006).

Uma das questões investigadas nesse artigo foi a possibilidade de utilizar métodos de *STS* para identificar as unidades de informação recordadas; abordagem inédita, até onde sabemos.

Para obter a similaridade semântica de duas sentenças neste artigo, utilizamos dois métodos:

1. O método de Hartmann (2016)[3], o qual obteve o melhor resultado de similaridade semântica na Avaliação de Similaridade Semântica e INferência textual (ASSIN) (Fonseca et al., 2016) — chamamos esse método de *STS*;
2. O método chamado de *Chunking*, proposto neste trabalho, explora a similaridade de representações vetoriais obtidas por *embeddings*.

Hartmann (2016) utilizou uma abordagem baseada no valor da similaridade do cosseno de duas representações vetoriais de cada sentença.

Na primeira representação, o autor soma os vetores de cada palavra obtidos pelo word2vec (Mikolov et al., 2013). Na segunda representação, é realizada uma expansão do vocabulário: para cada palavra de conteúdo são buscados os sinônimos no TEP (Thesaurus para o português do Brasil) (Maziero et al., 2008). Essa expansão é restrita apenas a palavras que possuam até um sinônimo, o que corresponde a 28% das entradas do TEP. Em seguida, os pares são transformados em uma representação esparsa, utilizando o *TF-IDF* (*frequency–inverse document frequency*), e é obtida a similaridade do cosseno.

Por fim, os valores dos cossenos entre as duas representações (*TF-IDF* e *word2vec*) de cada par são dados como entrada para um regressor linear que determina a similaridade do par.

Na Tabela 2, são apresentados alguns exemplos de sentenças das narrativas de reconto e as sentenças da narrativa original com os valores de similaridade. Com esses exemplos, é possível perceber a viabilidade da exploração de *STS* para a tarefa avaliada neste artigo. As duas primeiras linhas da Tabela 2 apresentam valores altos de similaridade semântica, sendo que, pela definição da anotação do ASSIN, os valores indicam que as sentenças são muito semelhantes, mas apresentam algumas informações exclusivas. Enquanto que a terceira e quarta sentenças apresentam valores de similaridade próximos de 3, sendo que esse valor indica que as sentenças possuem similaridade e podem se referir ao mesmo fato.

Dado que o sistema de *STS* recebe como entrada dois textos curtos, para possibilitar a aplicação desse sistema as narrativas originais de cada bateria foram sentenciadas e para cada sentença foram atribuídos seus respectivos rótulos. Nas Tabelas 3 e 4 são apresentados os resultados dessa etapa. Assim como nas sentenças dos pacientes, algumas sentenças possuem mais que um rótulo.

Como temos um problema multirrótulo, adotamos a abordagem de transformação de problema *Binary Relevance* (Tsoumakas et al., 2009), para reduzirmos o problema multirrótulo para vários problemas binários. Em cada problema queremos identificar se a sentença do reconto é uma respectiva unidade de informação ou não. Dessa forma, para cada sentença é criado um par $(\mathbf{S}_i^1, \mathbf{S}_{UI_j}^2)$, sendo que UI_j é a unidade de informação j, e para cada par é obtido o valor de similaridade. Na Tabela 5, são apresentados

[3] O autor gentilmente nos forneceu o código fonte e os modelos utilizados no sistema.

Similaridade	Sentença do reconto	Sentença da narrativa
4,17	e ela ficou aliviada.	a senhora ficou muito aliviada.
4,55	uma senhora fazia as compras no mercado.	uma senhora fazia compras.
3,09	e ai foi pegou um táxi pra chegar com tempo.	então ela pegou um táxi até a rodoviária.
2,93	ela pegou carona.	ela foi para a rodoviária de carona com seu amigo.

Tabela 2: Exemplos para os valores de similaridade semântica.

Sentenças	Rótulos
Lúcia mora no interior do Paraná	LUCIA MORA INTERIOR PARANA
numa manhã de segunda-feira	NUMA_MANHA_SEGUNDA
ela saiu de casa para mais uma entrevista de trabalho na capital do estado	SAIU_DE_CASA BUSCAR_EMPREGO NA_CAPITAL
ela foi para a rodoviária de carona com seu amigo Pedro	FOI_RODOVIARIA
estava chovendo naquela manhã	ESTAVA_CHOVENDO NAQUELA_MANHA
de repente o carro passou por um buraco	CARRO PASSOU_CAIU BURACO
e o pneu furou	PNEU_FUROU
Lúcia pensou que iria perder o ônibus	PENSOU_ACHOU_PERDER ONIBUS
então ela pegou um táxi até a rodoviária	PEGOU_TAXI
e conseguiu chegar a tempo	CONSEGUIU_CHEGAR_TEMPO

Tabela 3: Sentenças da narrativa original da BALE rotuladas com as unidades de informação.

os valores de similaridade para a sentença "*uma senhora fazia as compras no mercado*" contrastada com cada rótulo das sentenças da narrativa original, apresentada na Tabela 4.

Na Figura 3, são dispostos os histogramas e a estimação de densidade por *kernel* para cada unidade de informação da *ABCD*. É possível perceber a separação para algumas unidades de informação, como: SENHORA, ESTAVA_FAZENDO_COMPRAS, QUANDO_ELA_ABRIU_A_PORTA. Entretanto, outras não apresentam uma separação clara, como ELA_NAO_VIU_A_CARTEIRA_CAIR.

Dado o valor de similaridade semântica do par $(\mathbf{S}_i^1, \mathbf{S}_{UI_j}^2)$, é necessário definir um ponto de corte para transformar esse valor em uma resposta binária. Para encontrar o ponto de corte que maximizasse a medida $F1$ para classe UI_j, aplicamos um otimizador Bayesiano com a técnica *TPE* (*Tree-structured Parzen Estimator*) (Bergstra et al., 2013).

Os passos do segundo método avaliado e chamado de *Chunking* são elencados abaixo:

1. Um *tagger* probabilístico (López & Pardo, 2015) que atribui a classe gramatical mais frequente do conjunto de dados foi utilizado para filtrar as palavras de conteúdo das sentenças dos recontos, de forma semelhante aos trabalhos de Yancheva & Rudzicz (2016) e Fraser et al. (2019). Escolhemos esse *tagger*, pois as narrativas de reconto possuem ruídos que podem afetar o desempenho de *PoS taggers* treinados em córpus.
2. Em seguida, as palavras são convertidas para uma representação densa com o *FastText*, e calculamos a média dos vetores em $\mathbf{S}_{UI_j}^2$.
3. Lembrando que queremos identificar se a sentença é uma respectiva unidade de informação ou não, então para cada sentença é criado um par $(\mathbf{S}_i^1, \mathbf{S}_{UI_j}^2)$, sendo que UI_j é a unidade de informação j, e para cada par é obtido o valor de similaridade.
4. Dada uma sentença da narrativa de reconto, $\mathbf{S}_i^1$, esta é dividida em *n-grams*, variando de 1 a 3.
5. Para cada n-gram, é calculada a média dos vetores que compõem esse *n-gram*.
6. Por fim, calculamos a similaridade do cosseno desses vetores e retornamos o valor mais próximo de $\mathbf{S}_{UI_j}^2$.
7. Utilizamos um otimizador Bayesiano com a técnica *Tree-structured Parzen Estimator — TPE* para encontrar o ponto de corte.

Na Figura 4, apresentamos um exemplo da aplicação do método de *Chunking*. Dada uma sentença $\mathbf{S}_i^1$ contendo 4 palavras de conteúdo

Sentenças	**Rótulos**
uma senhora fazia compras	SENHORA ESTAVA_FAZENDO_COMPRAS
sua carteira caiu da bolsa	SUA_CARTEIRA CARTEIRA_CAIU DA_SUA_BOLSA
mas ela não viu	ELA_NAO_VIU_A_CARTEIRA_CAIR
quando ela foi ao caixa	NO_CAIXA
não tinha como pagar as compras	NAO_TEM_COMO_PAGAR
então ela colocou as compras de lado	COLOCA_AS_MERCADORIAS_DE_LADO
foi para casa	FOI_PARA_SUA_CASA
assim que ela abriu a porta da casa	QUANDO_ELA_ABRIU_A_PORTA
o telefone tocou	TELEFONE_TOCOU
uma menininha disse-lhe que tinha achado a carteira	PEQUENA MENINA LHE_DISSE ELA_ACHOU_A_CARTEIRA
a senhora ficou muito aliviada	SENHORA_ALIVIADA

Tabela 4: Sentenças da narrativa original da *ABCD* rotuladas com as unidades de informação.

Rótulos	**Similaridade**
SENHORA	4,55397
ESTAVA_FAZENDO_COMPRAS	4,55397
SUA_CARTEIRA	1,2814
CARTEIRA_CAIU	1,2814
DA_SUA_BOLSA	1,2814
ELA_NAO_VIU_A_CARTEIRA_CAIR	1,24164
NO_CAIXA	1,20222
NAO_TEM_COMO_PAGAR	2,71347
COLOCA_AS_MERCADORIAS_DE_LADO	2,48818
FOI_PARA_SUA_CASA	1,31341
QUANDO_ELA_ABRIU_A_PORTA	1,46262
TELEFONE_TOCOU	1,08743
PEQUENA	1,08353
MENINA	1,08353
LHE_DISSE	1,08353
ELA_ACHOU_A_CARTEIRA	1,08353
SENHORA_ALIVIADA	2,52263

Tabela 5: Valores de similaridade da sentença "uma senhora fazia as compras no mercado".

(P_1, P_2, P_3, P_4), esta é dividida em 9 *n-grams* (linhas da tabela à esquerda), e para cada *n-gram* calculamos a distância do cosseno para cada sentença que representa uma unidade de informação. Em seguida, utilizamos o valor máximo para cada classe (C_1, C_2, C_3) e aplicamos os pontos de corte para definir quais unidades de informação a sentença contém.

3.3 Baselines

Para comparar os métodos apresentados na Seção 3.2 na tarefa de identificação de unidades de informação, foram usadas duas estratégias.

A primeira, chamada de Casamento Exato, utiliza uma lista de palavras para identificar as unidades de informação, via casamento exato. Essa abordagem também foi utilizada em outros trabalhos (Prud'hommeaux & Roark, 2015; Pakhomov et al., 2010; Fraser et al., 2016). A segunda utiliza a saída do sistema *baseline* de inferência textual do ASSIN. Nessa abordagem, chamada aqui de *Inferência*, consideramos que a sentença contém a unidade de informação se o sistema de inferência retornar os rótulos *"Inferência"* e *"Paráfrase"*.

4 Avaliação de Métodos de Identificação de Unidades de Informação

Nesta seção, são apresentados os experimentos para identificação de unidades de informação, para os métodos descritos na Seção 3.2 e as *baselines* da Seção 3.3.

Os conjuntos foram separados em treino e teste, utilizando 70% para treino e 30% para teste de forma estratificada para cada grupo.

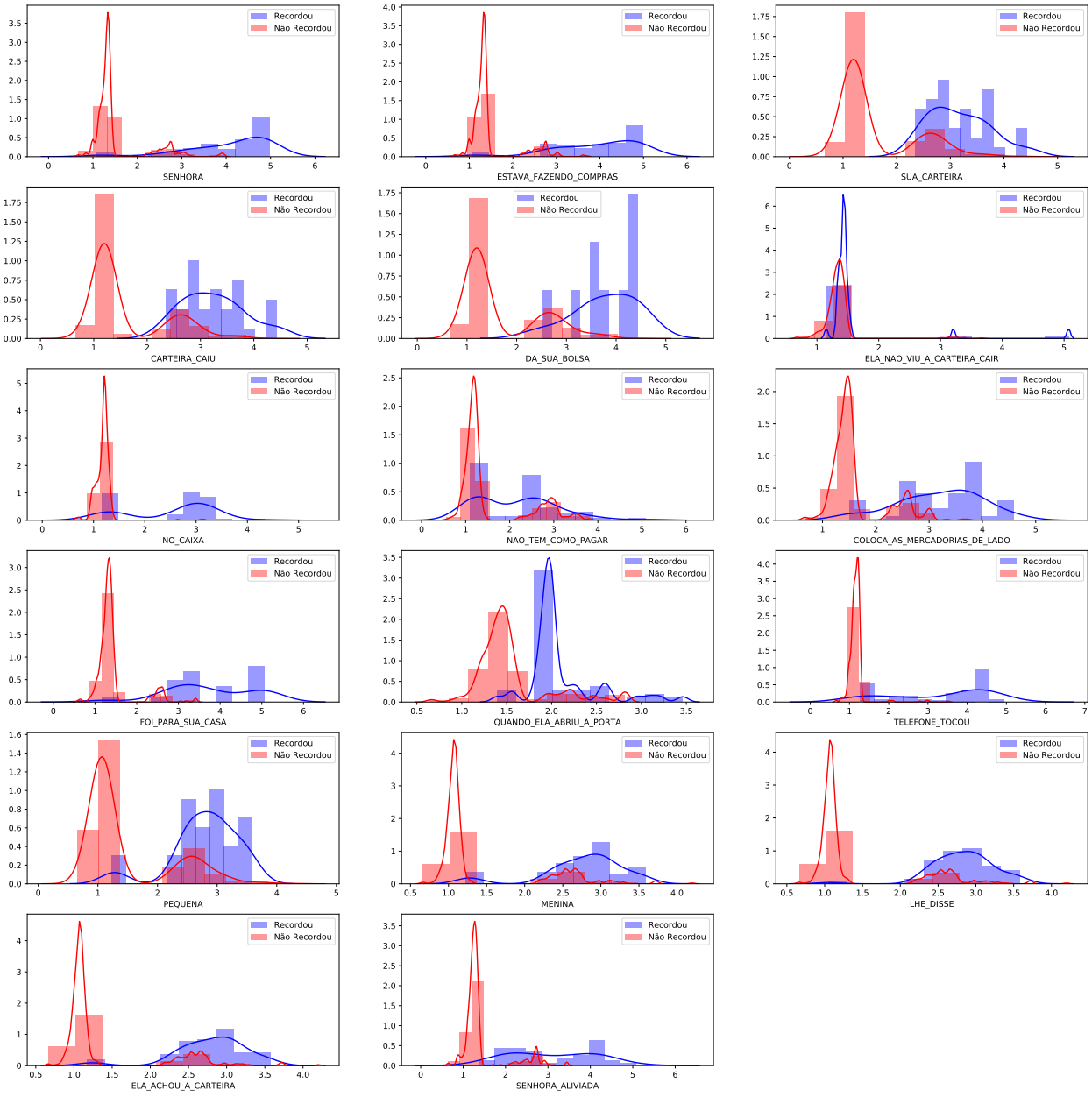

Figura 3: Histograma e distribuição acumulada para cada rótulo da *ABCD*.

S^1_i	C_1	C_2	C_3
P_1	0.55	0.10	0.01
P_1, P_2	0.60	0.15	0.30
P_1, P_2, P_3	0.30	0.13	0.40
P_2	0.40	0.18	0.35
P_2, P_3	0.30	0.30	0.55
P_2, P_3, P_4	0.15	0.20	0.70
P_3	0.20	0.15	0.48
P_3, P_4	0.30	0.12	0.60
P_4	0.15	0.08	0.05

Valor máximo para cada classe →

	C_1	C_2	C_3
S^1_i	0.60	0.30	0.70

Aplicação do ponto de corte para definir as classes finais →

$S^1_i = C_1, C_2$

Figura 4: Aplicação do método de *Chunking* para uma sentença com quatro palavras de conteúdo.

Na *ABCD*, cada participante produz duas narrativas, uma imediatamente após ouvir a história e a outra após 30 minutos. Para não enviesar a avaliação, os pares de narrativas produzidas foram alocados no conjunto de treino ou no conjunto de teste. Para a *BALE*, agrupamos as narrativas dos idosos com CCL e DA, dado o baixo número de idosos com CCL.

A Tabela 6 apresenta os resultados obtidos no conjunto de dados da *ABCD*. Por se tratar de um problema multirrótulo, reportamos a Precisão micro (Pr_{micro}), Precisão macro (Pr_{macro}), $F1$ micro ($F1_{micro}$), $F1$ macro ($F1_{macro}$), *SubsetAccuracy*, e o *HammingLoss*.

O método *baseline* Casamento Exato obteve os valores mais baixos, já o método *baseline* Inferência obteve os melhores resultados para precisão e o *HammingLoss*, mas foi superado pelo método de similaridade semântica *STS* nas outras medidas.

A Tabela 7 mostra os resultados obtidos no conjunto de dados da BALE. Os métodos *baselines* Casamento Exato e Inferência obtiveram os melhores resultados para precisão, enquanto o método *STS* obteve os melhores resultados em $F1$ e *SubetAccuracy*.

O objetivo final da pesquisa é criar um classificador para narrativas de testes neuropsicológicos de idosos saudáveis e idosos com comprometimento cognitivo (CCL e DA), para poder identificar os primeiros sinais de problemas cognitivos. Neste artigo, avaliamos o desempenho de classificadores utilizando unidades de informação como atributos. Na próxima seção, avaliamos se os métodos com medidas altas de $F1$ e *SubsetAccuracy* conseguem obter resultados de classificação próximos da anotação manual.

5 Classificação Automática de Narrativas visando uma Triagem Automática de Pacientes

Realizamos duas tarefas de classificação automática de narrativas, uma para cada conjunto de dados avaliado neste trabalho.

O conjunto *ABCD* possui as classes CCL e Controles Saudáveis e o conjunto BALE as classes CCL e DA, que foram agrupadas, e contrastadas com os Controles Saudáveis. Nessas duas tarefas, utilizamos as unidades de informação como vetores de atributos binários. Desta forma, cada narrativa da *ABCD* e da BALE possui 17 e 21 atributos, respectivamente. Avaliamos seis algoritmos de aprendizado de máquina: *SVM* (com kernel liner e RBF), *Naïve Bayes*, Árvores de decisão, *Gradient Boosting*, e *KNN*, implementados no *scikit-learn* (Pedregosa et al., 2011) versão 0.21.2, com os hiperparâmetros *default*.

Algumas particularidades destes métodos são destacadas abaixo:

Naïve Bayes é um dos algoritmos de aprendizado de máquina mais simples, pois assume que os atributos são independentes;

SVM é um algoritmo de classificação linear; sua função de otimização busca encontrar o hiperplano com margem máxima. Para esse algoritmo, utilizamos o *kernel* linear e o Radial Basis Function.

Árvores de decisão é um algoritmo que recursivamente particiona o espaço de entrada, geralmente de forma binária, definindo um modelo local em cada região resultante do espaço de entrada. É possível visualizar o modelo final na forma de uma árvore, em que cada partição representa um nó.

Gradient Boosting utiliza diversas árvores de decisão; cada árvore é treinada de forma sequencial para corrigir os erros da anterior.

KNN pertence à categoria *lazy*, pois não necessita de uma fase de treinamento para predizer um novo exemplo; busca no conjunto de treinamento os k exemplos mais similares e retorna o rótulo mais frequente.

Para as tarefas de classificação binária, os conjuntos de dados foram balanceados. Para a *ABCD*, utilizamos 12 idosos por grupo (Controle e CCL), sendo que cada idoso produziu 2 narrativas. O conjunto de dados final para a *ABCD* possui 48 narrativas. No caso da BALE, agrupamos os pacientes com CCL e DA (o grupo contém 16 narrativas), e selecionamos de forma randômica 16 narrativas do grupo de Controle. O conjunto de dados final para a BALE possui 32 narrativas.

Para a avaliação, utilizamos *10-fold-cross-validation* e a métrica de acurácia. Comparamos os métodos desenvolvidos para a identificação de unidades de informação e os *baselines* com a anotação manual, para analisar o impacto dos métodos na classificação final.

Os métodos *STS* e *Chunking* necessitam de uma busca de hiperparâmetros. Foi utilizada a metodologia *10-fold-cross-validation* para construir o novo conjunto de dados e este foi utilizado na avaliação dos classificadores.

Os resultados obtidos por todos os modelos na *ABCD*, em termos de acurácia, são apresentados na Tabela 8. Na segunda coluna da tabela,

Método	$\mathbf{Pr_{macro}}$		$\mathbf{Pr_{micro}}$		$\mathbf{F1_{macro}}$		$\mathbf{F1_{micro}}$		***SubsetAccuracy***		***Hamming Loss***	
	Treino	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**
Casamento Exato	0,570	0,469	0,903	0,758	0,337	0,283	0,246	0,233	0,246	0,169	0,078	0,081
Inferência	0,858	**0,793**	0,891	**0,873**	0,552	0,531	0,500	0,478	0,348	0,384	0,062	**0,062**
Chunking	0,705	0,699	0,587	0,577	0,640	**0,624**	0,668	**0,656**	0,668	**0,395**	0,076	0,076
STS	0,651	0,569	0,595	0,552	0,672	0,598	0,670	0,552	0,670	0,273	0,072	0,081

Tabela 6: Resultados da identificação de unidades de informação na *ABCD*

Método	$\mathbf{Pr_{macro}}$		$\mathbf{Pr_{micro}}$		$\mathbf{F1_{macro}}$		$\mathbf{F1_{micro}}$		***SubsetAccuracy***		***Hamming Loss***	
	Treino	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**	**Treino**	**Teste**
Casamento Exato	0,680	**0,740**	0,830	**0,930**	0,510	0,540	0,440	0,460	0,430	0,300	0,040	**0,040**
Inferência	0,625	0,690	0,808	0,781	0,485	0,519	0,359	0,404	0,447	0,324	0,042	0,050
Chunking	0,620	0,580	0,510	0,480	0,600	0,570	0,630	0,560	0,460	0,340	0,060	0,070
STS	0,680	0,640	0,670	0,650	0,740	**0,700**	0,720	**0,650**	0,520	**0,430**	0,030	**0,040**

Tabela 7: Resultados da identificação de unidades de informação na BALE

apresentamos os resultados para a anotação manual, que trouxe valores próximos para dois dos seis algoritmos de aprendizado (Árvores de Decisão e *Naïve Bayes*). Em geral, o classificador de Árvores de Decisão apresenta diferenças negativas maiores entre os valores da anotação manual e dos quatro modelos automáticos.

Dentre os dois métodos propostos para a identificação de unidades de informação (*STS* e *Chunking*), os melhores desempenhos para a *ABCD* foram do método de *Chunking*. Já o método de Inferência apresentou, em geral, resultados melhores do que o *STS*.

Na Tabela 9, são dispostos os resultados dos modelos na BALE. Para a anotação manual, tivemos quatro empates no desempenho de classificadores. O método *Chunking* apresenta diferenças negativas maiores entre os valores da anotação manual para todos os classificadores. Já o método *STS* apresenta as menores diferenças.

Em geral, o método *baseline* Casamento Exato superou os métodos automáticos propostos de identificação de unidades de informação.

Resumindo, os métodos com desempenhos adequados para a identificação de unidades de informação trazem resultados para classificação próximos da anotação humana, e as unidades de informação auxiliam mais na classificação final de pacientes com DA *versus* Controles Saudáveis (caso da BALE). Já para a *ABCD*, em que temos dois grupos balanceados de idosos saudáveis e com CCL, foi mais difícil separar as classes, corroborando com resultados da literatura (Santos et al., 2017; Fraser et al., 2019). Portanto, seguindo os trabalhos da literatura, há necessidade de trazer mais atributos para a classificação de pacientes saudáveis e com CCL, para melhorar o desempenho da classificação final dos pacientes.

6 Conclusões e Trabalhos Futuros

Este trabalho tratou de duas avaliações no cenário clínico: (i) avaliou métodos de similaridade semântica textual para a tarefa de identificação de unidades de informação em narrativas de recontos, e (ii) usou as unidades recuperadas como atributos para a classificação binária dos grupos idosos saudáveis *versus* idosos com comprometimento cognitivo, avaliando vários algoritmos de aprendizado de máquina.

Como observado na revisão dos trabalhos da literatura para identificação automática de unidades de informação em recontos, a grande dificuldade de utilizar listas de palavras para cada unidade de informação é a necessidade de um trabalho humano e subjetivo, pois nem sempre a lista possui todas as paráfrases/sinônimos possíveis. Métodos de *clustering* são úteis, pois conseguem criar automaticamente as unidades de informação, entretanto, podem gerar unidades pouco representativas ou não relacionadas às unidades que um dado teste neuropsicológico avalia.

Em estudos envolvendo análise de narrativas clínicas, geralmente a quantidade de dados é limitada, dado o alto custo de aquisição dos dados. Por se tratar de uma tarefa multirrótulo (identificação de unidades de informação), o cenário tratado neste artigo é ainda mais desafiador. Neste artigo, contornamos essas limitações aproximando a tarefa de identificação automática de unidades de informação em narrativas com similaridade semântica.

Avaliamos um método de similaridade semântica que explora a similaridade de representações vetoriais obtidas por *embeddings* e outro que se destacou na avaliação ASSIN, e transformamos o problema multirrótulo em

Método	Manual	Casamento Exato	Inferência	*Chunking*	*STS*
Árvore de Decisão	**0,638**	0,475	0,475	0,525	**0,538**
Gradient Boosting	0,538	0,463	**0,625**	0,550	0,500
KNN	0,575	**0,513**	0,525	**0,663**	0,413
SVM-Linear	0,500	0,475	0,588	0,525	0,488
SVM-RBF	0,563	0,363	0,463	0,463	0,488
Naïve Bayes	0,625	0,425	0,588	0,638	0,525

Tabela 8: Resultados da classificação utilizando as unidades de informações na *ABCD*

Método	Manual	Casamento Exato	Inferência	*Chunking*	*STS*
Árvore de Decisão	0,600	0,700	0,525	0,400	0,600
Gradient Boosting	**0,675**	0,725	0,450	0,400	0,575
KNN	0,650	0,575	0,475	0,525	0,625
SVM-Linear	**0,675**	0,625	0,600	0,550	0,625
SVM-RBF	**0,675**	0,525	**0,675**	**0,625**	**0,725**
NaiveBayes	**0,675**	**0,775**	0,550	0,550	0,700

Tabela 9: Resultados da classificação utilizando as unidades de informações na BALE

problemas de classificação binária, encontrando um ponto de corte para o valor de similaridade de cada unidade de informação. Desse forma, conseguimos superar ambos os *baselines* para os dois conjuntos de dados avaliados.

O uso de uma representação densa para sentenças é comum na literatura. Aqui, adotamos uma combinação com a média dos *embeddings* das palavras como proposto por Hartmann (2016), mas nos últimos anos essa abordagem vem sendo superada por métodos mais complexos como *ELMo* (*Embeddings from Language Models*) (Peters et al., 2018) ou *BERT* (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019). Como trabalhos futuros, pretendemos explorar esses modelos mais atuais, pois acreditamos que com sistemas melhores de similaridade semântica podemos obter métodos de identificação de unidades de informação também melhores.

Outro ponto para investigações futuras é a utilização de mais atributos para a classificação final, como os propostos em Santos et al. (2017). Observamos na avaliação dos classificadores finais deste trabalho que separar conjuntos com características próximas como os da *ABCD* (CCLs *versus* Controles Saudáveis) é mais difícil do que a classificação final de pacientes com DA *versus* Controles Saudáveis (caso da BALE). Novos atributos linguísticos e da representação de narrativas via redes complexas podem contribuir com essa tarefa.

Por fim, cabe também avaliar o desempenho dos métodos de identificação de unidades de informação em narrativas de recontos, usando métodos de segmentação automática das narrativas, como os explorados em Treviso et al. (2017a,b) e Treviso & Aluísio (2018), mas retreinados com os datasets de Testes Neuropsicológicos em Português do Brasil[4], disponibilizados publicamente recentemente.

Agradecimentos

O presente trabalho foi realizado com o apoio do CNPq, processos números 130100/2015-3, 155137/2015-8, e 153047/2016-0, e também contou com o apoio da Google via programa *Google Research Awards for Latin America*.

Referências

Abbott, Alison. 2011. A problem for our age. *Nature* 475(7355). S2–S4. doi 10.1038/475S2a.

de Abreu, Izabella Dutra, Orestes V. Forlenza & Hélio Lauar de Barros. 2005. Demência de alzheimer: correlação entre memóriaria e autonomia. *Revista de Psiquiatria Clınica* 32. 131–136. doi 10.1590/S0101-60832005000300005.

Agirre, Eneko, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea et al. 2015. Semeval-2015 task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. Em *9th International Workshop*

[4] https://github.com/nilc-nlp/DNLT-BP

on Semantic Evaluation (SemEval), 252–263. 10.18653/v1/S15-2045.

Agirre, Eneko, Mona Diab, Daniel Cer & Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. Em *1st Joint Conference on Lexical and Computational Semantics-Volume 1:*, 385–393.

Bayles, Kathryn & C. K. Tomoeda. 1993. *ABCD: Arizona battery for communication disorders of dementia*. Tucson, AZ: Canyonlands Publishing.

Becker, James T., François Boiler, Oscar L Lopez, Judith Saxton & Karen L. McGonigle. 1994. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology* 51(6). 585–594. 10.1001/archneur.1994.00540180063015.

Bergstra, James, Dan Yamins & David D. Cox. 2013. Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. Em *12th Python in Science Conference (SCIPY)*, 13–20.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5. 135–146. 10.1162/tacl_a_00051.

Chapman, Sandra Bond, Jennifer Zientz, Myron Weiner, Roger Rosenberg, William Frawley & Mary Hope Burns. 2002. Discourse changes in early Alzheimer disease, mild cognitive impairment, and normal aging. *Alzheimer disease & Associated Disorders* 16(3). 177–186. 10.1097/00002093-200207000-00008.

Clemente, Rená & Sergio Ribeiro-Filho. 2008. Comprometimento Cognitivo Leve: aspectos conceituais, abordagem clínica e diagnóstica. *Revista do Hospital Universitário Pedro Ernesto* 7(1). 68–77.

Cromnow, Karolina & Tove Landberg. 2009. *Skriftliga beskrivningar av bilden Kakstölden. Insamling av referensvärden från friska försökspersoner*: Division of Speech and Language Pathology, Karolinska institute. Tese de Mestrado.

Dagan, Ido, Oren Glickman & Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. Em Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini & Florence d'Alché Buc (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, 177–190. 10.1007/11736790_9.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.

Fichman, Helenice Charchat, Rosinda Martins Oliveira & Conceição Santos Fernandes. 2011. Neuropsychological and neurobiological markers of the preclinical stage of alzheimer's disease. *Psychology & Neuroscience* 4(2). 245–253. 10.3922/j.psns.2011.2.010.

Fleming, Valarie B. & Joyce L. Harris. 2008. Complex discourse production in mild cognitive impairment: Detecting subtle changes. *Aphasiology* 22(7-8). 729–740. 10.1080/02687030701803762.

Fonseca, Erick Rocha, Leandro Borges Santos, Marcelo Criscuolo & Sandra Maria Aluísio. 2016. Visão geral da avaliaçao de similaridade semântica e inferência textual. *Linguamática* 8(2). 3–13.

Forbes-McKay, K.E. & A. Venneri. 2005. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological Sciences* 26(4). 243–254. 10.1007/s10072-005-0467-9.

Fraser, Kathleen C., Kristina Lundholm Fors & Dimitrios Kokkinakis. 2019. Multilingual word embeddings for the assessment of narrative speech in mild cognitive impairment. *Computer Speech & Language* 53. 121–139. 10.1016/j.csl.2018.07.005.

Fraser, Kathleen C., Jed A. Meltzer & Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease* 49(2). 407–422. 10.3233/JAD-150520.

Freitas, Maria Isabel D'Ávila. 2010. *Habilidades linguísticas de pacientes com demência vascular: estudo comparativo com a doença de alzheimer*: Universidade de São Paulo. Tese de Doutoramento.

Frota, Norberto Anízio Ferreira, Ricardo Nitrini, Benito Pereira Damasceno, Orestes Forlenza, Elza Dias-Tosta, Amauri B da Silva, Emílio Herrera Junior & Regina Miksian Magaldi. 2011. Critérios para o diagnóstico de doença de Alzheimer. *Dementia & Neuropsychologia* 5(supl 1). 5–10. 10.1590/S1980-57642011DN05030002.

Garcia, Flavia Helena Alves & Letícia Lessa Mansur. 2006. Habilidades funcionais de comunicação: idoso saudável. *Acta fisiátrica* 13(2). 87–89.

Goodglass, Harold & Edith Kaplan. 1983. *The assessment of aphasia and related disorders*. Philadelphia: Lea & Febiger 2nd edn.

Hartmann, Nathan Siegle. 2016. Solo queue at ASSIN: Combinando abordagens tradicionais e emergentes. *Linguamática* 8(2). 59–64.

Hodges, John R., Karalyn Patterson, Naida Graham & Kate Dawson. 1996. Naming and Knowing in Dementia of Alzheimer's Type. *Brain and Language* 54(2). 302–325. 10.1006/brln.1996.0077.

Hübner, Lilian Cristine, Fernanda Loureiro, Bruna Tessaro, Ellen Siqueira, Gislaine Jerônimo & Anderson Smidarle. 2019. BALE: bateria de avaliação da linguagem no envelhecimento. Em Nicolle Zimmermann, François Delaere & Rochele Paz Fonseca (eds.), *Tarefas de avaliação neuropsicológica para adultos: memória e linguagem*, vol. 3, Memnon.

Kaufman, Leonard & Peter J. Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons.

Kintsch, Walter & Teun A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review* 85(5). 363–394. 10.1037/0033-295X.85.5.363.

Liang, Percy, Ben Taskar & Dan Klein. 2006. Alignment by agreement. Em *Human Language Technology Conference of the NAACL*, 104–111. 10.3115/1220835.1220849.

López, Roque & Thiago Pardo. 2015. Experiments on sentence boundary detection in user-generated web content. Em *Computational Linguistics and Intelligent Text Processing (CICLing)*, 227–237. 10.1007/978-3-319-18111-0_18.

Mansur, Letícia Lessa, Maria Teresa Carthery, Paulo Caramelli & Ricardo Nitrini. 2005. Linguagem e cognição na doença de Alzheimer. *Psicologia: reflexão e crítica* 18(3). 300–307. 10.1590/S0102-79722005000300002.

Mapstone, Mark, Amrita K. Cheema, Massimo S. Fiandaca, Xiaogang Zhong, Timothy R. Mhyre, Linda H. MacArthur, William J. Hall, Susan G. Fisher, Derick R. Peterson, James M. Haley et al. 2014. Plasma phospholipids identify antecedent memory impairment in older adults. *Nature Medicine* 20(4). 415–418. 10.1038/nm.3466.

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi & Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. Em *9th International Conference on Language Resources and Evaluation (LREC)*, 216–223.

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 *VI Workshop em Tecnologia da informação e da linguagem humana (TIL)*, 390–392.

McKhann, Guy M, David S Knopman, Howard Chertkow, Bradley T Hyman, Clifford R Jack Jr, Claudia H Kawas, William E Klunk, Walter J Koroshetz, Jennifer J Manly, Richard Mayeux et al. 2011. The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia* 7(3). 263–269. 10.1016/j.jalz.2011.03.005.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Em *International Conference on Learning Representations (ICLR)*, s/p.

Nasreddine, Ziad S, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings & Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society* 53(4). 695–699. 10.1111/j.1532-5415.2005.53221.x.

Pakhomov, Serguei V. S., Glenn E Smith, Dustin Chacon, Yara Feliciano, Neill Graff-Radford, Richard Caselli & David S. Knopman. 2010. Computerized analysis of speech and language to identify psycholinguistic correlates of frontotemporal lobar degeneration. *Cognitive and Behavioral Neurology* 23(3). 165–177. 10.1097/WNN.0b013e3181c5dde3.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot & E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. Glove: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee & Luke Zettlemoyer. 2018. Deep contextualized word representations. Em *Conference of the North American Chapter of the Association for Computational Linguistics*, 2227–2237. 10.18653/v1/N18-1202.

Preti, Dino (ed.). 2005. *O discurso oral culto*. São Paulo: Associação Editorial Humanitas 3rd edn. Projetos Paralelos. V.2.

Prud'hommeaux, Emily & Brian Roark. 2015. Graph-based word alignment for clinical language evaluation. *Computational Linguistics* 41(4). 549–578.

Santos, Leandro, Edilson Anselmo Corrêa Júnior, Osvaldo Oliveira Jr, Diego Amancio, Letícia Mansur & Sandra Aluísio. 2017. Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. Em *55th Annual Meeting of the Association for Computational Linguistics*, 1284–1296. 10.18653/v1/P17-1118.

Santos, Leandro, Lilian Cristiane Hübner, Anderson Dick Smidarle, Letícia Mansur & Sandra Aluísio. 2019. Anotação de unidades de informação em transcrições de fala na tarefa de reconto de narrativas em português. Em *XII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 253–261.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou & Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. Em *Demonstrations Session at European Association for Computational Linguistics (EACL)*, 102–107.

Treviso, Marcos, Christopher Shulby & Sandra Aluísio. 2017a. Evaluating word embeddings for sentence boundary detection in speech transcripts. Em *XI Brazilian Symposium in Information and Human Language Technology (STIL)*, 151–160.

Treviso, Marcos, Christopher Shulby & Sandra Aluísio. 2017b. Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. Em *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 315–325. Association for Computational Linguistics.

Treviso, Marcos Vinícius & Sandra Maria Aluísio. 2018. Sentence segmentation and disfluency detection in narrative transcripts from neuropsychological tests. Em *Computational Processing of the Portuguese Language (PROPOR)*, 409–418. 10.1007/978-3-319-99722-3_41.

Tsoumakas, Grigorios, Ioannis Katakis & Ioannis Vlahavas. 2009. Mining multi-label data. Em Maimon O. & Rokach L. (eds.), *Data Mining and Knowledge Discovery Handbook*, 667–685. Springer, Boston, MA.

Wallin, Anders, Arto Nordlund, Michael Jonsson, Karin Lind, Åke Edman, Mattias Göthlin, Jacob Stålhammar, Marie Eckerström, Silke Kern, Anne Börjesson-Hanson et al. 2016. The Gothenburg MCI study: design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow & Metabolism* 36(1). 114–131. 10.1038/jcbfm.2015.147.

Wechsler, David. 1997. *Wechsler memory scale - third edition.* San Antonio, TX: The Psychological Corporation.

Wortmann, Marc. 2012. Dementia: a global health priority-highlights from an ADI and World Health Organization report. *Alzheimer's Research & Therapy* 4(5). 40. 10.1186/alzrt143.

Yancheva, Maria & Frank Rudzicz. 2016. Vector-space topic models for detecting Alzheimer's disease. Em *54th Annual Meeting of the Association for Computational Linguistics*, 2337–2346. 10.18653/v1/P16-1221.

Novas Perspetivas

Propuesta recibida en octubre de 2018 y aceptada para publicación en noviembre de 2019.

Extracción y análisis de las causas de suicidio a través de marcadores lingüísticos en reportes periodísticos

Extraction and analysis of suicide causes through linguistic markers in news reports

José A. Reyes-Ortiz
Universidad Autónoma Metropolitana
jaro@azc.uam.mx

Mireya Tovar
Benemérita Universidad Autónoma de Puebla
mtovar@cs.buap.mx

Resumen

El análisis automático de información(textos) sobre el suicidio se ha convertido en un reto para el campo de investigación en lingüística computacional, cada vez más, son necesarias herramientas que ayuden a disminuir las tasas de suicidios, por ejemplo, extraer las causas para apoyar en su detección temprana. Los aspectos lingüísticos en los textos en Español, tales como frases clave o partes de la oración, pueden ayudar en dicha tarea. Por ello, en este artículo se presenta un enfoque computacional para la extracción y análisis de causas a partir de cabeceras de reportes periodísticos sobre el suicidio en español. La tarea de extracción automática de causas de suicidio es llevada a cabo mediante marcadores lingüísticos basados en verbos, conectores, preposiciones y conjunciones. Por su parte, el análisis de las causas de suicidio es realizado en dos enfoques: a) un análisis centrado en frases verbales y nominales, estudiando la presencia de la negación; b) un análisis centrado en la frecuencia de los unigramas y bigramas de palabras. Ambos análisis muestran resultados prometedores, los cuales son útiles para conocer los motivos de los suicidios reportados en México en un periodo determinado. Finalmente, se obtiene una colección de 581 causas del suicidio.

Palabras clave

análisis de causas, suicidio en reportes periodísticos, patrones lingüísticos, lingüística computacional

Abstract

The automatic analysis of suicide data(texts) has become a challenge for the computational linguistics research field, increasingly, tools are needed to help reduce suicide rates, for example, by extracting the suicide causes in order to support their early detection. Linguistic aspects in Spanish texts, such as cue phrases or parts of speech, can help in this task. Therefore, this paper presents a computational approach to the extraction and analysis of suicide causes from news reports in Spanish. The automatic extraction of suicide causes is carried out through linguistic markers based on verbs, connectors, prepositions and conjunctions. On the other hand, the analysis of the suicides causes is performed in two approaches: a) an analysis focused on verbal and noun phrases, studying the presence of the negation; b) an analysis on the frequency about unigrams or bigrams of words. Both analyzes show promising and correlated results, which are useful for recognizing the suicide causes reported in Mexico in a given period. Finally, a corpus is obtained with a collection of 581 suicide causes.

Keywords

cause analysis, suicide in news reports, linguistic patterns, computational linguistics

1 Introducción

El suicidio es definido por la Organización Mundial de la Salud (OMS), como un acto de quitarse deliberadamente la propia vida, el cual es iniciado y realizado por una persona en pleno conocimiento o expectativa de su desenlace fatal. En México, el Instituto Nacional de Estadística y Geografía (INEGI, 2015) define el suicidio como "la acción de matarse a sí mismo". Los suicidios son un problema presente en la sociedad mexicana, donde se suscitaron 5.2 suicidios por cada 100 mil habitantes en 2015 según el INEGI (2015). En ese sentido, se estima que el suicidio ocupa una de las primeras diez causas de muerte en México.

El suicidio se caracteriza como una muerte violenta o traumática (Hernández-Bringas & Flores-Arenales, 2011), el cual tiene una causa, motivo, razón o justificación. Esta causa resulta ser una de las características más importantes del evento, ya que proporciona información sobre su origen, que al extraerla podemos realizar un análisis con la finalidad de prevenir este evento.

La prevención del suicidio es un problema social muy importante ya que según Omer & Elitzur (2001) se necesitan esfuerzos en conjun-

DOI: 10.21814/lm.11.2.276

to entre organizaciones y personas para reunir la información suficiente con la cual caracterizarlo, por ejemplo sus causas. Por ello, el extraer y analizar las causas del suicidio se convierte en un reto importante que sería de gran ayuda para los analistas de noticias en dos aspectos: a) disminuyen los tiempos invertidos en la tarea tediosa de análisis manual de los textos; b) los analistas tienen conocimiento de las causas de los suicidios de manera automática para la toma de decisiones. Los textos de cabeceras de reportes periodísticos son una fuente importante para recabar dicha información. Estos textos resultan de gran utilidad ya que mediante señas lingüísticas relacionados a las causas del suicidio, se puede conocer el conjunto de ellas para conducir acciones hacia la prevención del mismo (Pestian et al., 2012). Esto se debe a que los periodistas reportan, entre otras cosas, las causas de haberse cometido un suicidio. La idea es contar con herramientas computacionales que apoyen a los analistas de noticias a realizar su actividad de manera rápida y contar con un apoyo en la detección temprana y prevención del suicidio. El problema radica en que existe una carencia de herramientas y recursos para el tratamiento de textos de suicidio en español, aunado a que resulta complicado tener acceso a una base de notas suicidas. Pero, es posible contar con los reportes periodísticos en línea, de los cuales sus cabeceras son de acceso público y pueden ser extraídas con facilidad a partir de la Web.

Los reportes periodísticos se han convertido en una fuente de datos muy poderosa, ya que nos brinda datos frescos sobre lo que está aconteciendo en una región o país y con una temporalidad determinada. En este artículo se presenta un enfoque de tratamiento automático de textos en español a partir de cabeceras de reportes periodísticos con la finalidad de detectar y analizar causas de suicidios en México haciendo uso de patrones lingüísticos. Este enfoque inicia con el reconocimiento de cabeceras de reportes periodísticos en español que traten sobre suicidio, utilizando el método presentado por Reyes-Ortiz & Bravo (2018); después, se extraen las causas, de manera automática, utilizando marcadores lingüísticos formados por verbos, preposiciones, conjunciones y conectores. Por último, tres enfoques de análisis de las causas son presentados: un análisis a nivel de frases tanto verbales como nominales, un análisis del impacto de la negación en las frases verbales y un análisis centrado en frecuencia de palabras. El conjunto de cabeceras de reportes periodísticos utilizado en este artículo se compone de la siguiente manera: a) como conjunto inicial, se utilizan 9 574 cabeceras para la tarea de clasificación; b) a partir del conjunto inicial, se identifican 1 347 pertenecientes a la categoría de suicidio; c) 581 causas son extraídas para su análisis a partir de las cabeceras sobre suicidio. Los resultados del análisis en términos de frecuencias de las palabras o frases son analizados mediante una nube de palabras para encontrar las causas más frecuentes y obtener la correlación de los resultados entre el enfoque de palabras y frases. Como producto final, un corpus de una colección de causas de suicidio es construido.

Las principales contribuciones de este artículo, se pueden resumir de la siguiente manera: a) la extracción automática de causas en las cabeceras de suicidios utilizando marcadores lingüísticos; b) el análisis de causas del suicidio usando dos enfoques: frecuencias de palabras y frecuencias de frases verbales y nominales; c) la creación de un corpus de causas de suicidio. Además, el enfoque propuesto resulta de gran utilidad para los analistas de noticias, apoyándolos en la tarea de recopilación de noticias sobre el suicidio y el análisis de sus causas, mediante su frecuencia, ya sea formadas por una palabra, un par de palabras, frases nominales o frases verbales.

El resto del artículo se organiza de la siguiente manera. En la Sección 2 se muestra el estado del arte relacionado a tres temas importantes: extracción automática de causas a partir de textos, aplicaciones de patrones lingüísticos para la extracción de información en diversos dominios y la extracción de cualquier tipo de información relacionada al suicidio. En la Sección 3 se expone la caracterización del suicidio y sus causas, así como la metodología de solución propuesta para la extracción y análisis de las causas. La Sección 4 presenta los marcadores lingüísticos utilizados para la extracción de causas en suicidios. La Sección 5 conduce un análisis de causas del suicidio a partir de los textos de cabeceras de reportes periodísticos. Finalmente, las conclusiones y el trabajo a futuro son presentados en la Sección 6.

2 Trabajo relacionado

La tarea de Extracción de Información es una subdisciplina de la Lingüística Computacional que consiste en identificar elementos o entidades de interés a partir de textos. Los patrones lingüísticos para esta tarea han sido utilizados como se describe a continuación. En el trabajo presentado por González-Gallardo et al. (2016) se hace uso de patrones sintácticos para la normalización de textos multilingüe extraídos de redes sociales con la finalidad de perfilar autores. La extracción de definiciones analíticas y relaciones semánticas de hiponimia-hiperonimia con

un sistema basado en patrones lingüísticos construidos manualmente es presentada por Dorantes et al. (2017). Los contextos definitorios son definidos por Sierra (2009) como un término y su definición introducida en el discurso de un texto de especialidad, en dicho trabajo se extraen estos contextos de manera automática con el apoyo del reconocimiento de patrones lingüísticos. La recuperación de patrones a partir de textos es una tarea que en el trabajo de Da Cunha et al. (2009) es abordada mediante un proceso de aprendizaje automático con la finalidad de generar resúmenes de textos especializados, es decir, de dominio específico. Roberto Rodríguez et al. (2013) presenta dos tipos de patrones (morfosintácticos y léxicos) para la clasificación automática de textos en registros lingüísticos en español, es decir, información sobre el perfil de los usuarios y sobre el contexto en sistemas de recomendación. Los patrones lingüísticos también pueden estar enfocados en el análisis automático de sentimientos en redes sociales mediante algoritmos de clasificación usando características lingüísticas de las partículas de los textos como las conjunciones ('but') y los condicionales ('if') (Chikersal et al., 2015) o bien, analizando las relaciones entre los conceptos usando patrones lingüísticos para obtener el tipo de sentimiento o polaridad en una sentencia (Poria et al., 2015). Adicionalmente, el uso de patrones lingüísticos es utilizado por Bertin et al. (2016) para la identificación de contextos de citas discerniendo las citas negativas.

La identificación automática de la causalidad a partir de textos se ha abordado desde un punto de vista de eventos. Los eventos tienen características como sus causas y efectos, aspectos que son identificados por Borsje et al. (2010), específicamente, para eventos financieros usando patrones semánticos y con la finalidad de enriquecer una ontología de dominio. En el dominio biomédico, un enfoque para la identificación automática de marcadores discursivos de causalidad usando aprendizaje automático con características semánticas a partir de textos biomédicos es presentado por Mihăilă & Ananiadou (2013). En el trabajo de Kang et al. (2017), se detecta las características causales haciendo uso de series de tiempo entre los *n-gramas*, temas, sentimientos y su composición extraídos a partir de textos. La clasificación de emociones es una tarea dentro del análisis de sentimientos, la cual es abordada desde un punto de vista lingüístico, por Li & Xu (2014), mediante la extracción de causas de eventos basada en patrones para ayudar en la clasificación de emociones como felicidad, tristeza, ira, sorpresa y disgusto, logrando resultados prometedores.

La idea de la prevención del suicidio se ha abordado como un análisis de información textual relacionada a los suicidios. De esta manera, diversos trabajos han abordado la extracción de información sobre suicidios a partir de textos como fuente de datos. Existen trabajos que han analizado las notas clínicas para predecir los riesgos de suicidios en los pacientes, como en el trabajo de Poulin et al. (2014) que utiliza un algoritmo de aprendizaje automático basado en programación genética para llevar a cabo esta tarea a partir de notas clínicas en inglés. El análisis de notas suicidas mediante técnicas de minería de sentimientos es un tema que ayuda en la prevención del suicidio, en la cual se han detectado trabajos que identifican de manera automática emociones (culpa, felicidad, agradecimiento, amor, información, desesperanza e instrucciones) en estas notas usando características de los textos con algoritmos de aprendizaje automático tales como: máquinas de soporte vectorial (Desmet & Hoste, 2013; Luyckx et al., 2012), campos aleatorios condicionales (Liakata et al., 2012) y un clasificador de máxima entropía (Wicentowski & Sydes, 2012). El aspecto lingüístico en el suicidio es de gran importancia. En esta línea, un análisis lingüístico de notas suicidas y textos sobre el suicidio aplicado al inglés y adaptado para el español ha sido presentado por Fernández-Cabana et al. (2015) para comparar las características sociodemográficas y forenses de ejemplos de víctimas de suicidio a partir de las notas suicidas dejadas por ellos, dicho estudio se realizó con notas suicidas en español comparando sus características de género, edad y nivel social. Los autores obtuvieron resultados sobre las características estudiadas y las diferentes frases lingüísticas utilizadas en las notas suicidas, tales como: longitud de las frases, uso de tiempos verbales, uso de pronombres y verbos. Finalmente, un estudio similar de Stirman & Pennebaker (2001) enfoca en el dominio de la poesía.

Con el estudio de los diversos temas y trabajos revisados en el estado del arte, se hace evidente que la investigación dirigida hacia el análisis lingüístico de la causalidad de los suicidios a partir de textos en español es un problema a resolver, existiendo trabajos como el de Reyes-Ortiz & Bravo (2018); Cook et al. (2016) que solo se enfocan en la clasificación del suicidio en español usando patrones o técnicas de aprendizaje supervisado. Esto muestra una carencia de herramientas y enfoques computacionales que utilicen técnicas de la Lingüística Computacional para el análisis de textos sobre las causas del suicidio en español. Este estudio, también, expone que la mayoría de los trabajos están enfocados

en el idioma inglés (Sawhney et al., 2018; Carson et al., 2019; Leiva & Freire, 2017) o en dos idiomas español-inglés (Cook et al., 2016), originando una necesidad creciente de contar con recursos de análisis de textos de causas del suicidio. Esto abre una ventana de desafíos y retos para llevar a cabo procesamiento automático de textos en español. Además, algunos trabajos revisados (Reyes-Ortiz & Bravo, 2018; Sawhney et al., 2018; Wicentowski & Sydes, 2012; Luyckx et al., 2012; Liakata et al., 2012; Poulin et al., 2014; Carson et al., 2019), se enfocan, solamente, en la identificación o clasificación del suicidio, mientras que nuestro trabajo añade la extracción y análisis de las causas del mismo. Por lo tanto, además de aportar una solución al problema de la extracción automática de causas del suicidio, brinda un panorama sobre los marcadores lingüísticos causales que son característicos de este tipo de textos en español y almacena las causas del suicidio reportadas en notas periodísticas.

3 Caracterización del suicidio y la causalidad

El suicidio como causa de muerte se encuentra dentro de la categoría de muertes violentas, ya que según Hernández-Bringas & Flores-Arenales (2011) se trata de muerte traumática, producidas por medios externos al organismo humano. Por otro lado, el Instituto Nacional de Estadística y Geografía (SSP & INEGI, 2012) ha definido al suicidio en México como "un evento que implica una conducta en la que una persona se priva de la vida por sí misma, involucrando sólo la intervención de una persona suicida", que se describe como:

- Suicida.
  Es alguien que se suicida por sí mismo.

Por lo tanto, un suicidio se caracteriza como un evento monovalente, es decir, que únicamente tiene un actor, el suicida. En la Figura 1 se muestra como se caracteriza un suicidio en el contexto del esquema actancial de la teoría de las valencias de eventos de Tesniére (1976). Donde "alguien" representa al actor que comete el suicidio.

Una causa expresa el argumento o la justificación, la cual es responsable de que suceda un evento o acción que Born (1949) formaliza como "la ocurrencia de una entidad B de cierta clase depende de la ocurrencia de una entidad A de otra clase", donde la entidad puede ser cualquier objeto físico, fenómeno, situación o evento. La Figura 2 muestra la representación de la causa de un evento suicida.

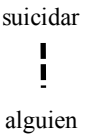

Figura 1: Esquema actancial de un evento suicida

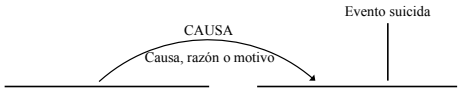

Figura 2: Caracterización de causalidad en eventos suicidas

Considerando la definición según Born (1949), un evento suicida tiene una micro-situación que expresa la causa. En este trabajo, se considera esta situación de causalidad para su análisis a partir de textos de cabeceras de reportes periodísticos en español.

En el contexto de datos textuales, la causalidad se manifiesta mediante una variedad de expresiones lingüísticas. Por lo tanto, en este artículo se aborda la extracción y el análisis de causas en eventos suicidas para el español. Este análisis se basa en un estudio sobre las construcciones de causas en las cuales intervienen marcadores lingüísticos y categorías gramaticales como, verbos, preposiciones y conjunciones. Para ello, se utilizan técnicas de la Lingüística Computacional para hacer posible el análisis automático de los textos en español relacionados al suicidio. La metodología de solución propuesta para la extracción y análisis de causas del suicidio en textos en español, es presentada en la Figura 3.

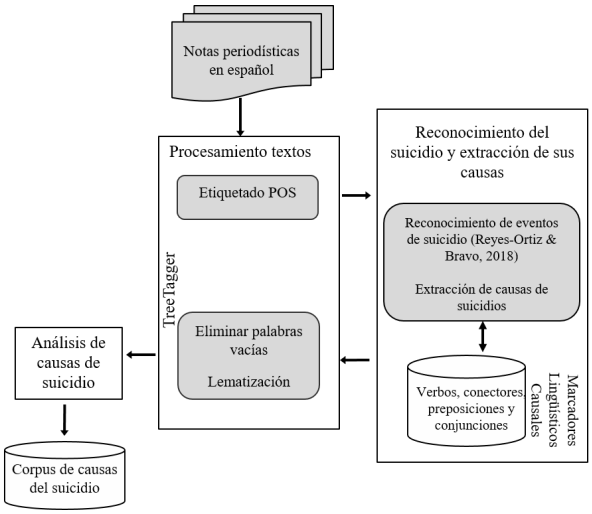

Figura 3: Metodología de solución para la extracción y análisis de causas del suicidio

Como se puede apreciar en la Figura 3 la metodología propuesta consiste de dos grandes etapas: extracción de causas del suicidio y el análisis estadístico de las mismas en México. Estas etapas son detalladas en las próximas secciones.

4 Extracción de causas del suicidio usando marcadores lingüísticos causales

En esta sección se presenta la etapa de la metodología correspondiente a la extracción automática de causas del suicidio utilizando expresiones lingüísticas tomadas de (Reyes-Ortiz et al., 2017) para reconocer las causas de suicidios a partir de cabeceras de reportes periodísticos en español. Es importante hacer notar que en (Reyes-Ortiz et al., 2017) se presentan marcadores lingüísticos para diversos tipos de eventos, sin embargo, en este trabajo se validan para el caso de eventos sobre el suicidio.

4.1 Marcadores lingüísticos causales

Un marcador lingüístico es una expresión formada por una o más palabras que tiene la función de conectar a los eventos con su argumento o justificación. En (Reyes-Ortiz et al., 2017) se presentan marcadores lingüísticos para identificar causas de cualquier evento, los cuales son llamados marcadores lingüísticos causales. Por ello, es que en este artículo, tomamos dichos marcadores y los aplicamos a los eventos de suicidios. Ellos están conformados por verbos causales, conectores conjuntivos, locuciones, preposiciones o conjunciones causales.

Para este trabajo se tomaron los marcadores lingüísticos causales presentados por Reyes-Ortiz et al. (2017) y se contrastaron con las construcciones lingüísticas presentadas en (Cano, 1981; Funes, 2010; Wunderlich, 1997). Además se consideran conectores causales obtenidos de la Real Académica de la Lengua Española (Española, 2009) para formar el siguiente conjunto de marcadores lingüísticos, los cuales serán utilizados para la extracción automática de causas del suicidio.

1. **Los verbos causales**. También llamados verbos causativos implícitos, verbos de carácter puramente causal o causativo propios, verbos básicamente causativos. Estos verbos están compuestos de una frase verbal y poseen un significado intrínsecamente causativo. La forma básica y representativa es el verbo *causar* y sus derivados como *provocar, originar, motivar, suscitar, desencadenar, promover, determinar, ocasionar, acarrear, producir, incitar, infundir, obrar* y *generar*. Su significado causal está supuesto por su forma semántica.
2. **Conectores causales**. Las causas también pueden estar representadas por conectores lingüísticos como: oraciones causales coordinadas que contiene nexos conjuntivos mediante los siguientes vocablos y locuciones *pues, pues que, porque, puesto que, a causa de, por esta razón, por eso, por ello, por esto, de esta manera, por lo cual, por lo que, debido a*; oraciones causales subordinadas, sus nexos conjuntivos son los vocablos: *de que, ya que*.
3. **Preposiciones causativas**. Las preposiciones son partículas lingüísticas que aportan gran significado a una oración. En el contexto de la causalidad, las preposiciones *tras*, *por* y *de* pueden expresar una causa. Según Funes (2010), esto se debe a que dichas preposiciones relacionan los eventos con su origen o argumentación.
4. **Conjunción causales**. Las conjunciones son partículas lingüísticas que funcionan como nexos entre segmentos de textos, las cuales pueden denotar una causalidad, como la conjunción *porque*, *pues* y *como*.

Estos marcadores lingüísticos son utilizados para la extracción automática de causas en suicidios a partir de textos en español de cabeceras de reportes periodísticos, con la finalidad de, posteriormente, llevar a cabo un análisis de las causas más frecuentes y detectar elementos relevantes.

4.2 Extracción automática de causas de suicidio

El proceso de extracción automática de causas del suicidio se compone de dos tareas: el reconocimiento de eventos reportados en cabeceras de reportes periodísticos centradas en suicidios y la extracción automática de causas basada en los marcadores lingüísticos presentados anteriormente. El contar con un conjunto de cabeceras de reportes periodísticos que contengan eventos de suicidio para la extracción de sus causas se vuelve indispensable. Para ello, se utiliza el enfoque presentado por Reyes-Ortiz & Bravo (2018) para el reconocimiento de eventos relacionados con la seguridad reportados en cabeceras periodísticas, entre ellos se encuentra el suicidio. Dicho enfoque utiliza patrones enriquecidos con información lingüística para reconocer y clasificar una cabecera de nota periodística, entre los patrones utilizados para la tarea de clasificación de una cabecera

en la categoría de suicidio se encuentran los siguientes: *X se quita la vida*, *X se suicida*, *X se ahorca*. Esta tarea considera un total de 9574 cabeceras y como resultado se obtienen 1 347 cabeceras de notas periodísticas que abordan el tema del suicidio.

Al conjunto de cabeceras filtradas (1 347) por el evento relacionado al suicidio se aplica el proceso de etiquetado POS y posteriormente, se aplican la regla mostrada en la Figura 4 con la finalidad de extraer causas en suicidios a partir de estos textos. Esta regla está basada en la gramática de JAPE (Cunningham & Tablan, 1999) que es una Máquina de Anotación de Patrones en Java basada en expresiones regulares. Esta regla hace uso de los marcadores lingüísticos causales presentados anteriormente. La frase que viene después del marcador y que denota la causa del suicidio (cs) puede ser un sintagma verbal (SV) o un sintagma nominal (SN). El etiquetado de partes de la oración denominado (POS) realizado con la herramienta denominada *TreeTagger* (Schmid, 1995), es necesario para el reconocimiento de sintagmas verbales y sintagmas nominales. El sintagma verbal es una palabra o grupo de palabras que constituyen una unidad sintáctica cuyo núcleo es un verbo. El sintagma nominal es una palabra o grupo de palabras que tiene como núcleo a un sustantivo.

```
Rule: ExtracciónCausaSuicidio
(EventoSuicidio)
(MarcadorLingüísticoCausal)
(SV|SN):cs -->
: cs.ExtracciónCausaSuicidio = {rule = "ExtracciónCausaSuicidio"}
```

donde cs expresa la causa del suicidio, SV expresa un sintagma verbal y SN es sintagma nominal.

Figura 4: Regla JAPE para extraer causas en suicidios

Esta tarea obtiene como resultado un corpus de 581 causas compuestas por sintagmas verbales y sintagmas nominales, donde el marcador lingüístico más representativo son los verbos causales.

El corpus obtenido es utilizado para un análisis de las causas en suicidios de la siguiente manera. Por un lado, los sintagmas se dejan en su forma original para hacer un análisis a nivel de frases. Mientras que, un procesamiento es aplicado al conjunto de sintagmas verbales y nominales que representan el argumento o justificación de los suicidios, con la finalidad de llevar a cabo un análisis preciso a nivel de palabras o entradas léxicas. El objetivo es extraer los términos (palabras o frases) relacionados al suicidio.

El procesamiento de las frases consiste en las siguientes tareas:

- Eliminar palabras vacías. El objetivo de esta tarea es quitar aquellas palabras que no aportan un significado en las frases que denotan las causas de suicidios, estas palabras están compuestas, principalmente, por artículos o determinantes (*el, las, los un, unos*), preposiciones (*a, ante, bajo, cabe, con, contra, de*), conjunciones (*y, o*) y algunos verbos, como el *ser/estar*, *tiene* que si bien tienen una frecuencia alta en las frases, no aportan relevancia en el estudio de causas. Aun cuando algunas preposiciones y conjunciones fueron útiles para extraer causas de suicidio, en esta etapa se suprimen para un análisis a nivel de términos relevantes como sustantivos, verbos o adjetivos.
- Lematización. Esta tarea tiene como objetivo normalizar las palabras resultantes contenidas en las causas para agrupar las palabras que provienen de la misma raíz. Este proceso consiste en obtener el lema correspondiente de cada palabra, eliminando tiempos verbales y conjugaciones en el caso de verbos, llevándolos a su forma en infinitivo. En el caso de sustantivos, adjetivos y adverbios, el lema corresponde a su forma en masculino singular. Por ejemplo, las formas *corrieron*, *corrió*, *correrán* es transformado a la forma verbal *correr*. La lematización de las causas obtenidas se ha llevado a cabo mediante el etiquetador *TreeTagger* (Schmid, 1995), el cual ha sido exitosamente utilizado para lematizar textos en Español. Esta tarea es indispensable para aplicar la regla propuesta sobre los textos de las cabecearas lematizados ya que los verbos causales están en su forma infinitivo.

Finalmente, después de la etapa de procesamiento de las frases, se crea una nube de palabras a partir de los términos o unidades léxicas resultantes. Esta nube de palabras ayuda a extraer con claridad las causas del suicidio a partir de los datos utilizados. Para generar esta nube de palabras fue considerada la frecuencia de aparición de cada palabra con el objetivo de descartar u otorgar menor importancia a las frases generadas por el significado polisémico de las preposiciones *por* y *de*, que en ocasiones no tienen un significado causativo.

La tarea de extracción automática de causas del suicidio ha demostrado que el tipo de mar-

cador lingüístico causal más representativo para esta tarea son los verbos causales en su forma de infinitivo. De esta manera, estos verbos causales caracterizan los textos de suicidios.

5 Análisis de causas del suicidio

En esta sección se presenta el análisis de causas extraídas con los marcadores lingüísticos presentados previamente, describiendo el conjunto de datos utilizado para este análisis. Primero, se utilizan los sintagmas nominales y verbales en su forma original con el objetivo de llevar a cabo un análisis a nivel de frases. Después, los componentes léxicos (palabras procesadas) de los sintagmas son utilizados para un análisis a nivel de una nube de palabras. El objetivo de esta nube de palabras es visualizar las causas de suicidios más frecuentes reportadas por periódicos mexicanos en un lapso de tiempo determinado.

El conjunto de datos utilizado para este estudio consiste en las 1 347 cabeceras de reportes periodísticos en español relacionadas con el suicidio, las cuales son extraídos de las cuentas en la red social Twitter de los principales periódicos y páginas en México que informan sobre noticias relacionadas con la seguridad, tales como: El Universal (*@El_Universal_Mx*), Milenio (*@Milenio*), Reforma (*@Reforma*), Excelsior (*@Excelsior*), Secretaria de Seguridad Publica de México(*@SSP_CDMX*), La Jornada (*@lajornadaonline*) y Noticias de Google México (*@googlenewsmx*). Las cabeceras de reportes periodísticos pertenecen al periodo de enero de 2017 a septiembre de 2018, las cuales son extraídas de manera automática con la *API* de Java denominada Twitter4J (Yamamoto, 2008). Esta herramienta ha permitido el filtrado de los mensajes por los parámetros de ubicación e idioma que nos ha permitido centrarnos en reportes periodísticos generados en la república mexicana y escritos en español.

A partir de estas 1 347 cabeceras de reportes periodísticos sobre suicidios, se lograron extraer 581 frases nominales y verbales, mediante la aplicación de los marcadores lingüísticos presentados. Entonces, se conservan esas frases para un primer análisis y después se eliminan palabras vacías y se lematizan para generar la nube de palabras.

5.1 Análisis de causas de suicidio centrado en frases

El primer análisis centrado en frases es realizado con los sintagmas nominales y verbales extraídas en su forma original. Se lleva a cabo una clasificación de frases en verbales y nominales cuyo resultado se muestra en la Tabla 1.

	Cantidad de frases	**Porcentaje**
Verbal	303	52.2 %
Nominal	278	47.8 %

Tabla 1: Resultados de la clasificación de frases.

Este análisis muestra una superioridad no muy marcada en la presencia de frases verbales para las causas de los suicidios. Las cinco frases verbales más frecuentes en las causas de suicidio son:

1. *ser víctima de violencia intrafamiliar*
2. *ser víctima de bullying*
3. *ser abusadas*
4. *ser separado de su familia*
5. *ser acusado de*

Es importante notar que la frase verbal *ser acusado de* tiene algunas variantes, entre las que destacan: *ser acusado de acoso sexual* y *ser acusado de violación*.

Las cinco frases nominales más frecuentes en las causas de suicidio son:

1. *bullying*
2. *problemas psicológicos*
3. *depresión*
4. *soledad*
5. *desesperación*

En el caso de las frases verbales (FV) se ha detectado la presencia de la negación, es decir, una frase verbal antecedida por una partícula semántica de negación, constituyendo una causa formal, como: $\neg FV$. Dada esta presencia, se incluye un estudio del impacto de este tipo de frases verbales utilizando su frecuencia en el conjunto de datos. De esta manera, dos ejemplo de una frase verbal negada y que han sido extraídas como causas de suicidios son: *no quería vivir con alzheimer* y *no recibir el regalo que quería*. La Tabla 2 muestra la distribución de este tipo de frases verbales presentes en el conjunto de causas obtenidas para el suicidio.

Las causas de suicidios tiene una gran tendencia a estar en forma afirmativa. Sin embargo, las causas expresadas con una frase verbal negativa no pueden ser descuidadas en una tarea de caracterización.

	Cantidad de frases	Porcentaje
$\neg FV$	32	10.6 %
FV	271	89.4 %

Tabla 2: Resultados de la distribución de la negación en frases verbales.

5.2 Análisis de causas de suicidio centrado en palabras

Este análisis de las causas de suicidios es llevado a cabo a nivel de palabras y mediante la generación de una nube de palabras. Para generar esta nube de palabras fue considerada la frecuencia de aparición de cada palabra en las causas, con ello se visualiza su importancia en la caracterización de causas de suicidios. Las palabras fueron obtenidas de las frases tanto verbales como nominales que representan las causas de suicidios. Estas palabras fueron lematizadas, y a partir de este conjunto, se eliminaron las palabras vacías. Dos escenarios son considerados para la generación de dos nubes de palabras: a) unigramas, causas como términos formados por una palabra; b) bigramas, causas como términos formados por un par de palabras. Esto con la finalidad de analizar la frecuencia de las causas con una y dos palabras.

En el primer caso se consideran unigramas de palabras a partir de las causas de suicidios detectadas. El resultado de esta nube de palabras es mostrado en la Figura 5, donde se aprecia una clara predominancia de la causa *bullying*.

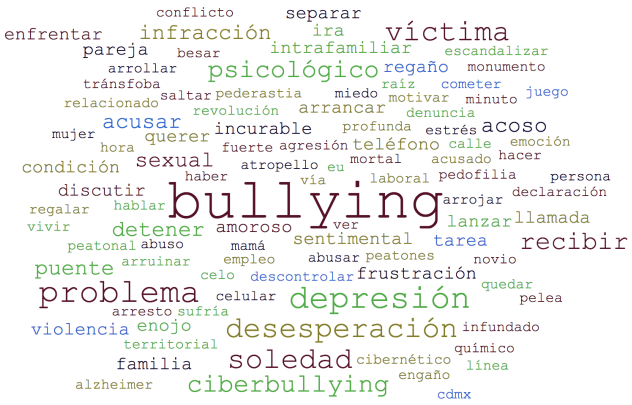

Figura 5: Nube de palabras de las causas de suicidios con unigramas

Para el segundo escenario se forman bigramas de palabras, es decir caso de causas de suicidios con número de palabras igual a dos. El resultado de esta nube de palabras es mostrado en la Figura 6.

Figura 6: Nube de palabras de las causas de suicidios con bigramas

Este análisis muestra que para el caso de los unigramas de palabras, las causas de suicidio más frecuentes son: *bullying*, *problemas*, *depresión*, *desesperación*, *soledad* y *ciberbullying*. Es importante notar que la causa *bullying* y *ciberbullying* en realidad tienen la misma naturaleza, la diferencia es que para el segundo caso, los ataques y ofensas (*bullying*) se llevan a cabo en Internet, en sitios como redes sociales, blogs o sistemas de mensajería electrónica. En el caso de los bigramas, las causas más frecuentes son: *violencia familiar* y *problemas psicológicos*.

El análisis presentado en esta sección revela que las frases verbales, frases nominales, unigramas y bigramas más frecuentes en las causas de suicidios pueden ser utilizadas como características lingüísticas para el análisis automático del suicidio a partir de cabeceras de reportes periodísticos en español. El caso de los unigramas revelan el uso de sustantivos formados por una palabra que son utilizados en las causas. Sin embargo, los bigramas serían de mayor utilidad en el caso de sustantivos compuestos, es decir, causas de suicidios expresadas con dos palabras, como es el caso de *problemas psicológicos*, que en términos de unigramas no aportarían un significado relevante a las causas. Si se desea realizar una caracterización de textos de suicidio, el uso de un solo tipo de característica no sería suficiente para abarcar todos los casos de causas de suicidios, lo más recomendable es un enfoque lingüístico que combine los diversos tipos de características.

6 Conclusiones

Este artículo ha presentado un enfoque para la extracción automática y, posterior, análisis de causas del suicidio en México utilizando los textos de cabeceras de reportes periodísticos en español correspondientes a un periodo determinado. El enfoque propuesto utiliza técnicas de la Lingüística Computacional para realizar el tra-

tamiento automático de los textos. El proceso completo consiste en una etapa de extracción automática de causas de suicidios a partir de texto en español y posteriormente, un análisis presentando estadísticas de frecuencia. El proceso de extracción utiliza marcadores lingüísticos causales basados en verbos, preposiciones, conectores y conjunciones para extraer frases verbales o nominales, negadas o afirmadas de los textos. Por su parte, el análisis es divido en dos escenarios, uno a nivel de frases y otro a nivel de palabras. Como salida se obtiene un corpus de 581 causas de suicidios extraídas y analizadas a partir de los textos.

El análisis centrado en frases verbales y nominales, muestra una ligera mayoría de frases verbales con un 52.2 % de las causas de suicidios. La presencia de la negación se ha encontrado en las frases verbales con una incidencia del 10.6 % de los casos.

El análisis centrado en palabras que se lleva a cabo mediante la formación de unigramas y bigramas de palabras exhibe cierta similitud a los resultados de frases, coincidiendo en el caso de las causas como el *bullying*, *problemas* y *depresión*.

En este artículo se presentaron las siguientes aportaciones: a) el método de extracción automática de causas del suicidio a partir de textos en español utilizando marcadores lingüísticos causales; b) el análisis de las causas de suicidio en los diversos escenarios: a nivel palabras o frases; c) la construcción del corpus con 581 causas de suicidios. Estas aportaciones disminuyen la carencia de recursos de análisis para textos en español, además de exponer las causas más frecuentes, ya sea formadas por una palabra (unigramas), un par de palabras (bigramas), frases nominales o frases verbales, negadas o afirmadas.

Con la tarea de extracción automática de causas se demuestra que el tipo de marcador lingüístico más representativo son los verbos causales, que fueron utilizados para extraer las frases verbales, frases nominales, unigramas y bigramas más frecuentes en las causas. Por lo tanto, estos marcadores se pueden utilizar para caracterizar textos sobre suicidios en español en notas periodísticas.

Aun cuando el análisis presentado es para un periodo determinado y utilizando cabeceras de reportes periodísticos, es de gran ayuda para los analistas de noticias, ya que tienen conocimiento de las causas del suicidio en México con la ayuda de un enfoque computacional, el cual es de gran utilidad al reducir el tiempo invertido en el análisis de noticias y reducir el esfuerzo humano. Con estos resultados los analistas pueden tomar decisiones como enfocar políticas de prevención del suicidio y conocer datos estadísticos sobre las causas del suicidio para evitarlas.

Como trabajo a futuro para este artículo, resulta enriquecedor extender el enfoque a periodos de tiempo mayores con reportes periodísticos completos de diversas fuentes en español. Además, la creación de un sistema informático para visualizar y consultar las causas del suicidio en México, sería de gran utilidad para los analistas de reportes periodísticos.

Agradecimientos

Este artículo ha sido apoyado por la Universidad Autónoma Metropolitana, unidad Azcapotzalco con el proyecto de investigación SI001-18. Los autores agradecen, también, a la Benemérita Universidad Autónoma de Puebla por el apoyo recibido y al CONACyT bajo el proyecto 257357.

Referencias

Bertin, Marc, Iana Atanassova, Cassidy R. Sugimoto & Vincent Lariviere. 2016. The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics* 109(3). 1417–1434. 10.1007/s11192-016-2134-8.

Born, Max. 1949. *Natural philosophy of cause and chance*. The Clarendon Press primary source ed.

Borsje, Jethro, Frederik Hogenboom & Flavius Frasincar. 2010. Semi-automatic financial events discovery based on lexico-semantic patterns. *International Journal of Web Engineering and Technology* 6(2). 115–140. 10.1504/IJWET.2010.038242.

Cano, Rafael. 1981. *Estructuras sintácticas transitivas en el español actual*, vol. 310. Gredos 1ª ed.

Carson, Nicholas, Brian Mullin, Maria Jose Sanchez, Frederick Lu, Kelly Yang, Michelle Menezes & Benjamin. Lê Cook. 2019. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. *PloS one* 14(2). e0211116. 10.1371/journal.pone.0211116.

Chikersal, Prerna, Soujanya Poria, Erik Cambria, Alexander Gelbukh & Chng Eng Siong.

2015. Modelling public sentiment in Twitter: using linguistic patterns to enhance supervised learning. En A. Gelbukh (ed.), *International Conference on Intelligent Text Processing and Computational Linguistics*, 49–65. Springer International Publishing. 10.1007/978-3-319-18117-2_4.

Cook, Benjamin L., Ana M. Progovac, Pei Chen, Brian Mullin, Sherry Hou & Enrique Baca-Garcia. 2016. Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid. *Computational and mathematical methods in medicine* 2016. 1–8. 10.1155/2016/8708434.

Cunningham, Diana, Hamish Maynard & Valentin Tablan. 1999. JAPE: a java annotation patterns engine.

Da Cunha, Iria, Juan Manuel Torres-Moreno, Patricia Velázquez-Morales & Jorge Vivaldi. 2009. Un algoritmo lingüístico-estadístico para resumen automático de textos especializados. *Linguamática* 1(2). 67–79.

Desmet, Bart & Véronique Hoste. 2013. Emotion detection in suicide notes. *Expert Systems with Applications* 40(16). 6351–6358. 10.1016/j.eswa.2013.05.050.

Dorantes, Miguel Alejandro, Alejandro Pimentel, Gerardo Sierra, Gemma Bel-Enguix & Claudio Molina. 2017. Extracción automática de definiciones analíticas y relaciones semánticas de hiponimia-hiperonimia con un sistema basado en patrones lingüísticos. *Linguamática* 9(2). 33–44. 10.21814/lm.9.2.257.

Española, Real Academia. 2009. *Nueva gramática de la lengua española*, vol. 2. Espasa Calpe 2ª ed.

Fernández-Cabana, Mercedes, Julio Jiménez-Féliz, María Teresa Alves-Pérez, Raimundo Mateos, Ignacio Gómez-Reino Rodríguez & Alejandro García-Caballero. 2015. Linguistic analysis of suicide notes in spain. *The European Journal of Psychiatry* 29(2). 45–155. 10.4321/S0213-61632015000200006.

Funes, Maria Soledad. 2010. La alternancia de las preposiciones por y de en las construcciones causales. *Revista de Estudios Hispanos* 1(1). 5–14.

González-Gallardo, Carlos, Juan-Manuel Torres-Moreno, Azucena Montes-Rendón & Gerardo Sierra. 2016. Perfilado de autor multilingüe en redes sociales a partir de n-gramas de caracteres y de etiquetas gramaticales. *Linguamática* 8(1). 21–29.

Hernández-Bringas, Héctor Hiram & René Flores-Arenales. 2011. El suicidio en méxico. *Papeles de población* 17(68). 69–101.

INEGI. 2015. Estadísticas de mortalidad. Accedido 05-09-2018. http://www.beta.inegi.org.mx/proyectos/registros/vitales/mortalidad/default.html.

Kang, Dongyeop, Varun Gangal, Ang Lu, Zheng Chen & Eduard Hovy. 2017. Detecting and explaining causes from text for a time series event. En *Conference on Empirical Methods in Natural Language Processing*, 2758–2767. 10.18653/v1/D17-1292.

Leiva, Victor & Ana Freire. 2017. Towards suicide prevention: Early detection of depression on social media. En *International Conference on Internet Science*, 428–436. 10.1007/978-3-319-70284-1_34.

Li, Weiyuan & Hua Xu. 2014. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications* 41(4). 1742–1749. 10.1016/j.eswa.2013.08.073.

Liakata, Maria, Jee-Hyub Kim, Shyamasree Saha, Janna Hastings & Dietrich Rebholz-Schuhmann. 2012. Three hybrid classifiers for the detection of emotions in suicide notes. *Biomedical informatics insights* 5. BII–S8967. 10.4137/BII.S8967.

Luyckx, Kim, Frederik Vaassen, Claudia Peersman & Walter Daelemans. 2012. Fine-grained emotion detection in suicide notes: a thresholding approach to multi-label classification. *Biomedical informatics insights* 5. BII–S8966. 10.4137/BII.S8966.

Mihăilă, Claudiu & Sophia Ananiadou. 2013. Recognising discourse causality triggers in the biomedical domain. *Journal of Bioinformatics and Computational Biology* 11(6). 1343008 (15 pages). 10.1142/S0219720013430087.

Omer, Haim & Avshalom C. Elitzur. 2001. What would you say to the person on the roof? a suicide prevention text. *Suicide and Life-Threatening Behavior* 31(2). 129–139. 10.1521/suli.31.2.129.21509.

Pestian, John P., Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K. Bretonnel Cohen, John Hurdle & Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights* 5. BII–S9042. 10.4137/BII.S9042.

Poria, Soujanya, Erik Cambria, Alexander Gelbukh, Federica Bisio & Amir Hussain. 2015. Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine* 10(4). 26–36. 10.1109/MCI.2015.2471215.

Poulin, Chris, Brian Shiner, Paul Thompson, Linas Vepstas, Yinong Young-Xu, Benjamin Goertzel, Bradley V.. Watts, Laura A. Flashman & Thomas W.. McAllister. 2014. Predicting the risk of suicide by analyzing the text of clinical notes. *PloS one* 9(1). e85733. 10.1371/journal.pone.0085733.

Reyes-Ortiz, José A. & Maricela Bravo. 2018. Enhancing patterns with linguistic information for criminal event recognition. *Journal of Intelligent and Fuzzy Systems* 34(5). 3027–3036. 10.3233/JIFS-169487.

Reyes-Ortiz, José. A., Maricela Bravo, Azecena Montes & Mireya Tovar. 2017. Event ontology enrichment with causal relations from spanish text. *International Journal of Computational Linguistics and Applications* 8(1). 1–16.

Roberto Rodríguez, John, Maria Salamó Llorente & Maria Antònia Martí Antonín. 2013. Clasificación automática del registro lingüístico en textos del español: un análisis contrastivo. *Linguamática* 5(1). 59–67.

Sawhney, Ramit, Prachi Manchanda, Raj Singh & Swati Aggarwal. 2018. A computational approach to feature extraction for identification of suicidal ideation in tweets. En *ACL 2018, Student Research Workshop*, 91–98. 10.18653/v1/P18-3013.

Schmid, Helmut. 1995. Treetagger: a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung* 43. 1–28.

Sierra, Gerardo. 2009. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática* 1(2). 13–37.

SSP & INEGI. 2012. Clasificación estadística del delito en méxico. Accedido 05-09-2018. http://www3.inegi.org.mx/sistemas/clasificaciones/delitos.aspx.

Stirman, Shannon Wiltsey & James W. Pennebaker. 2001. Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic medicine* 63(4). 517–522. 10.1097/00006842-200107000-00001.

Tesniére, Lucien. 1976. *Eléments de syntaxe structurelle*, vol. 2. Klincksieck 2ª ed.

Wicentowski, Richard & Matthew R. Sydes. 2012. Emotion detection in suicide notes using maximum entropy classification. *Biomedical informatics insights* 5. BII–S8972. 10.4137/BII.S8972.

Wunderlich, Dieter. 1997. Cause and the structure or verbs. *Linguistic Inquiry* 28(1). 27–68.

Yamamoto, Yusuke. 2008. Twitter4j. [Web; accedido el 19-10-2018]. http://twitter4j.org/en/index.

http://www.linguamatica.com/

linguaMATICA