Instituto Tecnológico de Costa Rica

Escuela de Ingeniería en Computación

Programa de Maestría en Computación

Arquitectura de desambiguación lingüística guiada por dominios de conocimiento distribuidos para traducción automática de lengua española a LESCO

Tesis sometida a consideración del Departamento de Computación,
para optar por el grado de Magíster Scientiae en Computación,
con énfasis en Ciencias de la Computación

Ing. Johan Manuel Serrato Romero

Profesor Asesor:
Ing. Mario Chacón Rivas, PhD.

Junio, 2017

Arquitectura de desambiguación lingüística guiada por dominios de conocimiento distribuidos para traducción automática de lengua española a LESCO

Resumen

La necesidad de los métodos estadísticos actuales de un conjunto grande de ejemplos de frases mapeadas entre dos lenguas y la investigación creciente en las lenguas de señas, implica la búsqueda de nuevas estrategias para flexibilizar la inclusión de nuevo vocabulario y producirlo correctamente de acuerdo al contexto del discurso. Esta tesis presenta el diseño e implementación de una arquitectura distribuida de hilos semánticos para un sistema de traducción de lengua española a la Lengua de Señas Costarricense (LESCO). Es mostrada la efectividad de un algoritmo de reconocimiento contextual propuesto para esta arquitectura a través de pruebas de textos divulgativos y el uso de corpus lingüísticamente validados.

Abstract

Current statistical methods necessity of large sets of phrase examples between two languages and the increasing research in sign languages, implies the search of new strategies to make more flexible inclusion of new vocabulary and produce it correctly according to the speech context. This thesis introduces the design and implementation of a distributed architecture of semantic threads for a Spanish language to Costa Rican Sign Language (LESCO) translation. It is showed the effectiveness of a contextual recognition algorithm proposed for this architecture through tests of informative texts and the use of linguistically validated corpora.

Programa de Maestría
en Computación
Teléfono: 2550-2402

APROBACIÓN DE LA TESIS

"Arquitectura de desambiguación lingüística guiada por dominios de conocimiento distribuidos para traducción automática de lengua española a LESCO"

TRIBUNAL EXAMINADOR

Dr. Mario Chacón Rivas
Profesor Asesor

Dr. Cesar Garita Rodríguez
Profesor Lector

Máster/DEA Luis Carlos Naranjo Zeledón
Profesor Externo

Dr. Roberto Cortés Morales
Coordinador del Programa
de Maestría en Computación

TEC | Tecnológico de Costa Rica
Maestría en Computación

Junio, 2017

Dedicatoria

Dedicado a
mi familia y amigos
más cercanos.

Agradecimientos

En primer lugar, al equipo humano del TEC Digital por su disponibilidad en pruebas y consejos brindados durante el desarrollo de la investigación del proyecto Traductor LESCO y todo el tiempo compartido en las actividades que potenciaron el proyecto para su consolidación.

A mi tutor, Mario Chacón Rivas, por su guía, consejos y experiencia aportada desde el inicio de las investigaciones sobre el tema de traducción a la LESCO, que han desembocado en la presente tesis y que continúan en el proyecto del Traductor LESCO.

A todos los investigadores y tutores que se acercaron a visualizar los distintos prototipos y aportaron observaciones oportunas sobre el papel que ocupa una herramienta como esta en la comunidad sorda: Ana Yarina Villanueva, Maria Infante, Eduardo Gómez Quijano, Christian Ramírez y Marcela Zúñiga.

Al grupo de investigación de Centro Nacional de Recursos para la Educación Inclusiva por sus tutorías, materiales, observaciones y apoyo incondicional: Tatiana Navarro y Rafael Martínez.

A Jaquelín Solís y Mauricio Ramírez por compartir su tiempo y experiencias sobre sus respectivas tesis de maestría.

Índice general

Índice de figuras

Índice de tablas

1 Introducción

La Lengua de Señas Costarricense (LESCO) es la lengua materna de la comunidad sorda costarricense, que como otras lenguas de señas en el mundo, posee una gramática y léxico muy diferentes a las de una lengua oral. Su origen y establecimiento son lingüísticamente recientes, y eso implica la poca cantidad de recursos tales como un vocabulario estandarizado, detallados corpus ejemplificando su uso o alguna clase de representación textual que la denoten completamente. Sin embargo, se han iniciado los primeros estudios sobre su gramática, como el diccionario de señas del Centro Nacional de Recursos para la Educación Inclusiva (CENAREC) (*CENAREC - Gramática - Proyecto de Descripción Básica de la LESCO*, s.f.), que asocia conceptos con una forma textual denominada glosa, lo que constituye un recurso inicial hacia estudios más completos y nuevas aplicaciones computacionales.

En el campo de *Machine Translation (MT)*[1], existen muchos enfoques abiertos sobre cómo transformar texto de una lengua a otra, principalmente por problemas relacionados a múltiples significados para un mismo término, contexto temático y generación de la lengua objetivo. Los investigadores en traducción automática hacia una lengua de señas aprovechan distintos esfuerzos realizados para la interpretación estructurada del texto de entrada, para después centrarse en encontrar la mejor forma de transferir esa interpretación inicial, llevarla a un medio de reproducción como un avatar tridimensional y hacerlo eficientemente, con el fin de ofrecer prototipos funcionales capaces de modificar y aumentar su vocabulario.

[1] *Machine Translation (MT)* se traduce en español como Traducción Automática. Para su referencia en el resto del documento, se utilizarán sus siglas en inglés.

El mayor problema actualmente es el acoplamiento entre el estado del arte en traducción automática y las necesidades de información para hacer funcional un traductor de señas, específicamente la definición de ejemplos de frases en un corpus de lengua de señas computacionalmente asequible, que aún debe ser generado y validado por la comunidad sorda, por lo que se debe considerar otros enfoques para generar resultados prácticos a corto plazo.

La propuesta de esta investigación consiste en elaborar una arquitectura para integrarla en el traductor de texto español hacia señas de la LESCO, desarrollado por el TEC Digital[2]. Dicha arquitectura se inspira en un modelo de procesamiento lingüístico mental concurrente, donde varios *gestores de contexto* (ver sección 4.1.2) reconocen patrones en la entrada y asignan un valor de pertenencia para su dominio lingüístico/temático, con posibilidad de generar una representación si el valor de pertenencia es alto. Entre los principales objetivos está abordar el problema de desambiguación contextual con el enfoque de distribuir su análisis entre estos módulos especializados y evaluar el grado de facilidad en la integración de nuevo vocabulario guiado por contextos temáticos en forma incremental.

El resto de este documento se encuentra conformado de la siguiente manera: el capítulo 2 presenta la revisión del estado del arte pertinente a la propuesta, que muestra de forma clasificada gran parte del conocimiento adquirido durante un proceso de mapeo sistemático, y que conforma el contexto de la investigación. El capítulo 3 explica el problema, su justificación en términos de innovación, impacto y profundidad, los objetivos que se quieren comprobar durante el desarrollo de la propuesta, y la definición de los alcances. El capítulo 4 explica la propuesta de arquitectura del traductor guiado por gestores de contexto. Luego, el capítulo 5 describe y muestra los resultados de los experimentos, seguido de este, se presenta el capítulo 6 con las principales conclusiones de la tesis y trabajos futuros con posibles líneas de investigación. El documento termina con el capítulo de referencias bibliográficas y anexos.

[2]La Unidad del TEC Digital se encarga de brindar soluciones de E-Learning a la comunidad docente y estudiantil del Instituto Tecnológico de Costa Rica. Es una dependencia de la Vicerrectoría de Docencia.

2 Contexto

En esta sección se exponen las definiciones, características y avances de los hitos más destacables encontrados en la etapa de investigación. Durante la revisión del material bibliográfico, se identificaron varios campos de estudio relevantes para los objetivos de investigación planteados. Los tres más importantes: traductores de lenguas orales a lenguas de señas, traducción automática y desambiguación lingüística, se explican en las siguientes secciones.

2.1. Traducción automática hacia lengua de señas

Una lengua de señas surge como una herramienta para la comunicación de la comunidad sorda[1], se deriva del lenguaje corporal y se complementa de forma natural con la oración de frases característica de la comunidad oyente, dentro del contexto cultural donde se origina. Es por esta razón que muchas señas son derivadas de gestos de apariencia intuitiva, para luego ser normalizadas cuando el lenguaje de señas alcanza un consenso de su práctica dentro de una región geográfica particular. Como consecuencia, tenemos que cada lengua de señas es diferente en cada país y a medida que la comunidad sorda aumenta, se fortalece más su distribución y evolución. Este crecimiento tiene que estar evidenciado también en las tecnologías de información, por lo que desde principios de la década de los 2010 hay un aumento en las investigaciones en computación para facilitar la comunicación bidireccional entre estas comunidades y los contenidos escritos y orales.

[1] Al referirnos a comunidad sorda, no sólo se trata de las personas con problemas del aparato auditivo, sino de todas las que se relacionan con ellas y que deben comunicarse a diario, potencialmente cualquier miembro de su país si la lengua de señas correspondiente es un idioma oficial.

La traducción automática de texto hacia una representación diferente como lo es una lengua de señas, se puede ver como un problema de representación de la salida en un traductor automático. Los problemas específicos que se observan son:

- Definir una representación computacional formal capaz de describir toda característica morfológica de la lengua de señas.
- Definir una metodología de generación capaz de usar el formato mencionado en el punto anterior, aplicando de forma válida la gramática de esta lengua.
- Obtener el vocabulario y ejemplos de uso que sean comparables con la lengua de entrada.

La mayoría de propuestas para traductores hacia lenguas de señas han delegado el procesamiento de la lengua de entrada a los analizadores de lenguaje natural, esto para enfocarse en las producciones visuales de las señas, con publicación de resultados sobre todo en los últimos diez años, donde el tema de accesibilidad a la información se ha venido insertando en tecnologías web, un mejor soporte y rendimiento gráfico por parte de los navegadores lo ha secundado y se dispone de una cantidad creciente de avances paralelos en investigación de traducción automática en general. Ejemplos recientes de trabajos sobre de traductores de texto a señas son:

- Un traductor de texto en griego a *Greek Sign Language (GSL)*[2] que orienta su uso a un ambiente de aprendizaje de la gramática de la lengua de señas, analizando sus ventajas de reducción de tiempo y costos ante la opción clásica de grabar videos (Fotinea, Efthimiou, Caridakis, y Karpouzis, 2008). Usa un sistema basado en reglas y un lexicón para aplicar reglas de frases básicas en GSL.

[2] *Greek Sign Language (GSL)* es la lengua de señas oficial de la República Helénica (Grecia). De aquí en adelante, se mencionará de acuerdo a sus siglas oficiales.

- Trabajos como (Jinghua, Baocai, Lichun, Dehui, y Yufei, 2012) y (Othman y Hamdoun, 2013) proponen notaciones para la especificación digital de las características fonológicas de las señas basándose en una descripción XML, importante para la representación interna de las señas del traductor y la interoperatividad de los reproductores de avatares 3D.

- Una metodología para la construcción de corpus paralelo entre el inglés y *American Sign Language (ASL)*[3], esto para alimentar a su traductor estadístico (Tmar, Othman, y Jemni, 2013).

Existen varios trabajos relacionados con la traducción de español hacia la Lengua de Señas Española (LSE) con un enfoque general de la arquitectura basada en reglas y el proceso secuencial de los análisis morfológico - sintáctico - semántico ((Baldassarri y Royo-Santas, 2009), (López-Ludeña y cols., 2012)). El trabajo de (Porta, López-Colino, Tejedor, y Colás, 2014) representa un traductor basado en reglas basado en un estudio de contraste de la gramática de esa lengua de señas con el castellano, logrando mejoras respecto a una versión estadística de la herramienta. Para la generación de las señas, (López-Colino y Colás, 2012) presenta un enfoque basado en un modelo fonológico flexible para facilitar ajustes manuales en la definición de las señas de su vocabulario.

Una propuesta sobresale por parte de (Filhol, Hadjadj, y Testu, 2015), que plantea un nuevo modelo de traductor basado en *triggering rules*, que serían definiciones de reglas de generación de elementos lingüísticos para una lengua de señas. Estas reglas serían ejecutadas por módulos intermedios de reconocimiento de patrones en el texto de entrada, de modo que ante la detección de un patrón específico, se hace la relación con la regla de generación correspondiente, minimizando con ello problemas relacionados con la complejidad morfológica de la lengua de señas y la falta de un corpus paralelo representativo para el desarrollo de la investigación. Del trabajo publicado, aún queda resolver el problema de construir una traducción coherente a partir de los resultados disparados por las reglas generadoras.

[3] *American Sign Language (ASL)* es la lengua de señas oficial de los Estados Unidos de América.

La siguiente sección es una revisión de las principales corrientes de investigación sobre traducción automática, al ser la base del análisis sobre el que trabajan la mayor parte de los trabajos de traductores de señas.

2.2. Traducción automática

Machine Translation (MT) es una rama de *Natural Language Processing (NLP)*[4] y es un tema central del campo de Computación Lingüística, este último enfocado en dar soluciones computacionales a problemas que surgen del uso y evolución de las lenguas. El objetivo de un sistema de MT es interpretar una representación textual de información en lenguaje natural y transformarla a otro sistema de representación, de modo que el resultado sea equivalente semánticamente.

Al plantear el problema de cómo traducir desde una lengua hacia otra, existe un entendimiento que es compartido por la mayoría de enfoques de investigación, tendiendo a dividir la solución en tres procesos secuenciales para llegar a la traducción final:

1. Procesar la forma física de la entrada. Lo típico es analizar directamente el texto en lenguaje natural, pero a este mismo proceso se le puede acoplar un submódulo de visión computacional para la detección de imágenes y movimiento, o sistemas de reconocimiento del habla para audio, esto como primer paso hacia una representación simbólica de la lengua de entrada que sea entendible por un sistema computacional.

[4] *Natural Language Processing (NLP)* se traduce en español como Procesamiento del Lenguaje Natural, entendido como un campo de estudio para la interpretación del lenguaje humano por parte de una computadora. De aquí en adelante, se usarán sus siglas en inglés cuando se referiere a este concepto.

2. Transferencia desde la lengua de entrada a la lengua de salida. Aquí se construye la relación entre los conceptos identificados de la entrada con su representación equivalente en la lengua de salida. Este paso puede incluir una representación intermedia o *interlengua* (Boitet, 2003) que llegue a facilitar el mapeo conceptual hacia la lengua destino, incluyendo detalles morfológicos (por ejemplo, terminaciones en palabras), sintácticos (el rol de las palabra dentro de una frase) y semánticos (el objeto, acción o atributo más probable al que apunta la palabra) detectados en el texto original.

3. Realización, que se encarga de tomar el sentido dado al discurso en el proceso anterior y sintetizar una representación correcta en la lengua destino, ya sea texto, audio o imágenes.

De los puntos anteriores, el segundo suele estar implícito en la práctica por metodologías que utilizan aprendizaje automático: el programador no puede intervenir directamente con el proceso de transferencia, sino que el proceso es un resultado de las configuraciones del aprendizaje, por ejemplo, como parte de la configuración de los pesos de una red neuronal.

En la actualidad, los principales problemas que enfrentan los sistemas MT están relacionados con generación incorrecta de la gramática, elección no conveniente de palabras sinónimas y estructuras desordenadas en la lengua destino. Las razones de estos casos tienen que ver con la elección de un contexto temático contrapuesto a el sentido más usado o popular, así como la falta de control específico en la selección de los ejemplos que alimentan a los algoritmos de aprendizaje. En las siguientes subsecciones, se dará un repaso por los principales enfoques para tratar de resolver el problema de la traducción automática y sus resultados.

2.2.1. Enfoque basado en reglas

Los primeros intentos formales de llevar la traducción automática a la realidad se enfocaron en sistemas basados en reglas, apoyándose en las teorías de gramáticas libres de contexto y clasificación de lenguajes en la jerarquía de Chomsky. Si bien estos sistemas están muy apegados a la teoría de reconocimiento y generación del lenguaje, su implementación pura contiene varios problemas:

- El enorme trabajo que implica validar lingüísticamente y codificar **todas las reglas gramaticales** tanto para el reconocimiento de la entrada como para la generación de la salida.
- El **manejo manual** de una colección de términos y reglas (diccionarios, lexicones u ontologías), cuyo mantenimiento y actualización implica un costo permanente durante el ciclo de vida del programa.
- La **falta de recursos lingüísticos** que describen las similitudes entre un par de lenguas, algo aún vigente de las llamadas lenguas fuertes (como ejemplos: inglés, chino, portugués, español) respecto a las débiles, como es el caso de las lenguas de señas y otras cuyo número de hablantes es mucho menor pero significativo.

Aún así, se han logrado resultados buenos en el análisis de textos muy especializados y en generación de lenguaje natural ((Hurtado Oliver, Costa, Segarra Soriano, García Granada, y Sanchis Arnal, 2016), (Lezcano, Guzmán, y Vélez, 2015)), cuyos formatos son poco variables con el tiempo y por ende sus reglas de producción son limitadas y manejables por la teoría de transductores ((Langkilde y Knight, 1998), (Vandeghinste y cols., 2013)) y gramáticas libres de contexto probabilísticas (Yuan, Wang, y Zhong, 2015).

2.2.2. Enfoque estadístico (basado en corpus)

A inicios de la década de los 90, toma importancia una serie de propuestas innovadoras de un grupo de investigadores de *IBM*[5] (Brown y cols., 1990), que rescatan el enfoque de conocimiento del traductor como un problema de recuperación de información postulado por varios pioneros en criptografía de la década de 1950. Su propuesta fue realizar un análisis estadístico acerca de la relación entre pares de frases de dos colecciones de texto correspondientes en significado, conocidos como *corpus paralelo*, donde una línea en una de las compilaciones tiene su correspondiente en la otra. Esta relación da lugar al alineamiento de frases, donde no sólo la frase corresponde, sino que se busca asociar cada fragmento de ella, sea una palabra o token de palabras, con su correspondiente parte en el corpus destino. Luego con las mejoras de eficiencia y otros avances en computación, estos enfoques de los años cincuenta ya eran factibles y se facilitaba el análisis de corpus con técnicas usadas en el propio campo de la traducción profesional, como es el uso de colecciones de ejemplos y memorias de traducción, tomando como métrica usual alguna forma de medir la frecuencia de palabras (Vargas Sierra, 2002).

Desde entonces, y potenciado por la cada vez más grande colección de textos disponibles en Internet a partir de la década de los años 2000, la traducción automática se concentró en encontrar modelos estocásticos que parten de la descripción de la probabilidad de que una frase destino sea la traducción correcta dado que se ha leído la frase de entrada: $p(destino|origen)$, dando origen a la clasificación del enfoque como *Statistical Machine Translation (SMT)*[6].

[5] *IBM: International Business Machines Corp.*

[6] *Statistical Machine Translation (SMT)* se traduce como Traducción Automática Estadística. De aquí en adelante, se usarán sus siglas en inglés para su referencia.

Se han realizado modelos entre pares de lenguas tales como inglés, francés, alemán, español, portugués, teniendo resultados muy buenos entre lenguas de una misma familia etimológica (por ejemplo, pares de lenguas romances como español y portugués), aunque todavía con problemas entre pares de lenguas de distinto origen, como las asiáticas y africanas, cuya estructura gramatical y morfológica es muy diferente, haciendo difícil su correcto alineamiento. Los cambios en la forma de alinear los componentes de los corpus estudiados por (Koehn, Och, y Marcu, 2003) resultaron en una mayor eficacia de los SMT que se enfocan prioritariamente a alinear palabras.

También han existido propuestas que tratan de unir ventajas lingüísticas de las técnicas basadas en reglas con las del enfoque estadístico, para contribuir en la alimentación de corpus más formalizados, tales como (Miyao, Ninomiya, y Tsujii, 2005) para la adquisición automática de reglas para gramáticas de frases.

2.2.3. Enfoque basado en redes neuronales

Con el impacto de la implementación de redes neuronales artificiales para el reconocimiento de patrones, se empezó a investigar cómo podrían mejorar los procesos existentes en SMT. Las redes neuronales proveen un marco de trabajo sobre el que se puede elaborar modelos conducidos por enormes cantidades de datos y algoritmos de aprendizaje, sin la intervención directa de un programador, lo cual ha sido aprovechado para generar implícitamente conjuntos de reglas gramaticales y cierto grado semántico en un vocabulario.

Desde 2010, existe un creciente número de estudios relacionados con el uso de las *Recurrent Neural Networks (RNN)*[7] para modelar una lengua (Mikolov, Karafiát, Burget, Cernocký, y Khudanpur, 2010). Un patrón de comportamiento de estas redes simula un sistema de memoria a corto plazo, teniendo una representación de los datos procesados con anterioridad y pudiéndose utilizar para análisis de referencia en las siguientes frases.

[7] *Recurrent Neural Networks (RNN)* en español es redes neuronales recurrentes, un tipo de red neuronal artificial. De aquí en adelante, se referenciará con sus siglas en inglés.

Se ha mostrado cómo se pueden usar las RNN para implementar el parseo del lenguaje natural (Socher, Bauer, Manning, y Ng, 2013), la utilidad de representar las palabras como vectores (Pennington, Socher, y Manning, 2014), hasta llegar a ser un tema de actualidad conocido como *Neural Machine Translation (NMT)*[8], donde se están probando las propuestas de sistemas de traducción basados enteramente en extensiones de las RNN ((Bahdanau, Cho, y Bengio, 2014), (Wu y cols., 2016)).

También se estudia sobre problemas de implementación como el rendimiento del aprendizaje en corpus enormes (Jean, Cho, Memisevic, y Bengio, 2015) y las palabras fuera del vocabulario (Out-Of-Vocabulary words) ((Luong, Sutskever, Le, Vinyals, y Zaremba, 2014), (Wu y cols., 2016)), caso que enfrenta los sistemas NMT cuando se detecta vocabulario desconocido y afecta a la realización final de la frase.

Actualmente están en progreso investigaciones para generar arquitecturas de traductores SMT *secuencia a secuencia*, con el objetivo de lograr un alineamiento de los datos de entrenamiento más adecuado para una NMT (Sutskever, Vinyals, y Le, 2014), orientando trabajos inspirados en análisis neurológicos que incluyen el uso de memoria dinámica (Kumar y cols., 2015).

2.3. Desambiguación lingüística

Word-Sense Dissambiguation (WSD)[9] es un tema de NLP que busca determinar el sentido semántico de los términos en un texto cuando estos pueden tener varios significados. A esto último se le llama polisemia, que es una característica inherente del lenguaje natural donde un mismo término o representación conceptual está asociado a diferentes significados, pero sólo uno es el correcto dentro en una frase contextualizada.

[8] *Neural Machine Translation (NMT)* se describe como un conjunto de técnicas que ajustan una red neuronal para que sirva como traductor. De aquí en adelante, se hará referencia por sus siglas.

[9] *Word-Sense Dissambiguation (WSD)* se llama en español como Desambiguación Lingüística. A partir de aquí, se hace referencia a este con sus siglas en inglés.

El tema de darle sentido a las palabras inició en los años 50, como una inevitable necesidad para procesar correctamente un lenguaje natural (Ide y Véronis, 1998). Ha sido considerada desde entonces una tarea intermedia para cualquier solución de NLP, hasta el punto de ser su propia rama de estudio. También es una tarea considerada como un problema *IA-completo*: para resolverla, se necesita primero hacerlo con todos los problemas difíciles en Inteligencia Artificial, tales como la representación del sentido común y en general encontrar implementaciones que minimicen el efecto de la "maldición de la dimensionalidad", que surge de la enorme cantidad de variables a considerar para una solución y que genera una intratabilidad del problema en la práctica.

Varios trabajos relevantes abordan la desambiguación especializada por temas y la elaboración de algoritmos de aprendizaje del vocabulario. El trabajo de (Justeson y Katz, 1995) propone un algoritmo para detectar términos técnicos relevantes por medio de patrones lingüísticos que funcionan como reglas gramaticales invariables, lo que resulta en buenos resultados acotados al dominio técnico. Una investigación sobre resúmenes de trabajos académicos (Saggion y Lapalme, 2000) utiliza la teoría de transductores para obtener párrafos con las principales ideas del documento, utilizando árboles AVL que contienen la frecuencia de los términos.

Un enfoque interesante de (Yarowsky, 1995) es de las primeras referencias de *ventana de contexto* utilizadas en algoritmos de desambiguación con aprendizaje no supervisado, basándose en la hipótesis de la posición de conceptos vecinos a una palabra semilla, algo que no se hace con las metodologías de conteo clásicas de *bag of words*[10], y se argumenta que una palabra tiene un significado consistente dentro de un discurso. Esto ha derivado más recientemente trabajos que aprovechan la raíz de este concepto, por ejemplo, basados en similitud de contexto por posicionamiento de palabras (Scruthi Sankar, Reghu Raj, y Jayan, 2016) o la adaptación de algoritmos de términos populares como algunas variantes del *PageRank* (Edmonds y Agirre, 2006).

[10]Conocido en español como *bolsa de palabras*, es un método de conteo de ocurrencias de palabras en un texto sin tomar en cuenta el orden ni el lugar en que aparecen.

Recientes trabajos se basan en el análisis estadístico de corpus y colecciones relacionadas como *WordNet* (*WordNet - About*, s.f.) y lexicones en otras lenguas disponibles en Internet ((Chen, Ding, Bowes, y Brown, 2009), (Liu y Sun, 2015)), agregando a su vez una nueva rama para la extracción del conocimiento desde los documentos, por ejemplo con un enfoque de algoritmos genéticos (AlSaidi, 2016).

Existen trabajos que argumentan ventajas con el uso de técnicas de desambiguación, o bien la inclusión de semántica, en métodos existentes de traducción automática. La introducción del análisis de corpus para la desambiguación es popularizada desde el trabajo de (Brown, Pietra, Pietra, y Mercer, 1991), donde propone una técnica para etiquetar el significado de las palabras con su representación en la otra lengua, logrando un resultado promisorio en el traductor pero limitado en la capacidad polisémica de la palabra etiquetada.

La propuesta de (Bangalore y Rambow, 2000) trata sobre un modelo estocástico basado en árboles de parseo para seleccionar el mejor lexema a la hora de generar la salida, usando a *WordNet* como fuente de información de las relaciones de sinonimia, logrando mejorar la eficacia respecto al modelo basado en bolsa de palabras. En (Seng Chan, Tou Ng, y Chiang, 2007), se argumenta la integración puntual de los resultados de un sistema WSD en un SMT.

Más recientemente, en (Hurtado Oliver y cols., 2016) se acota el contexto a un programa de atención automático para pasajeros de un servicio de trenes, probando metodologías de etiquetado semántico de palabras y una interlengua, para compararlas con las SMTs puras. Entre sus conclusiones, se mejora el resultado de traducción con la métrica *BLEU* (Papineni, Roukos, Ward, y Zhu, 2002), así como el detalle de ejemplos donde se penaliza injustamente traducciones correctas debido al uso de palabras diferentes pero con el mismo significado, esto sobre todo en sistemas que usan interlengua para la fase de transferencia.

Otros trabajos relacionados con determinar contextos y relaciones semánticas que mejoran las traducciones se mencionan en (Haque, Naskar, van den Bosch, y Way, 2011) y (Wong, Liu, y Bennamoun, 2012), en éste último se hace un análisis del estado del arte en la elaboración de ontologías a partir de texto, enumerando varios procesos de recuperación de información e inferencia de conocimiento usados en WSD y MT.

El objetivo de formar las ontologías es una opción a tomar en cuenta como modelo futuro para el intercambio de información en la red, y que puede ser aprovechado también para la Traducción Automática como un modelo del lenguaje con jerarquía semántica, con ejemplos de trabajos que pueden servir como extractores de conocimiento a partir de bancos de ontologías o mapas conceptuales en la web ((Eskridge, Hayes, y Hoffman, 2006), (Peng, 2010), (Caliusco y Stegmayer, 2010), (Simón y cols., 2006)).

3 Definición del problema

En el proceso de traducción entre dos lenguajes naturales, se debe aplicar un método de desambiguación semántica considerando el estado del arte. Sin embargo, existen diferentes enfoques al problema y cada uno implica modelos de datos y arquitecturas diferentes con resultados no definitivos respecto al problema de traducir bajo el contexto correcto.

Específicamente sobre los modelos de vocabulario, el estado de arte en SMT apunta hacia la construcción de estos modelos a partir de ejemplos representativos (corpus) en las lenguas de entrada y de salida, pero este recurso no está disponible en la Lengua de Señas Costarricense (LESCO), debido a que la formalización y divulgación de su gramática es muy reciente y no cuenta con una representación formal para usarla en sistemas informáticos.

Otro problema es que varias metodologías con algoritmos de aprendizaje sufren de *overfitting*[1], lo que hace perder generalización a un modelo entrenado previamente, limitando su posibilidad de modificación, lo cual es necesario de hacer frecuentemente para una lengua viva y más aún para lenguas poco estudiadas. Si bien se puede desechar el modelo de lenguaje aprendido y crear uno actualizado con las modificaciones, el tamaño del corpus a enseñar puede hacer que el entrenamiento tarde mucho tiempo, algo que en futuras aplicaciones de tiempo real será una clara desventaja para su adaptación automática.

[1] Este término se refiere a degradar la calidad de los resultados de una red neuronal al sobrepasar cierto número de entrenamientos con el mismo conjunto de datos.

3.1. Justificación

Parte de los entregables de esta propuesta es aplicar la arquitectura al Traductor LESCO, proyecto del TEC Digital del Instituto Tecnológico de Costa Rica, como producto de la línea de innovación e investigación que caracteriza a esta unidad. A continuación, se detallan aspectos puntuales acerca de la investigación que la justifican de manera estructurada (innovación, impacto, profundidad).

3.1.1. Innovación

En la rama de traducción automática hacia una lengua de señas tan joven como la LESCO, se aporta una nueva arquitectura para facilitar su desarrollo progresivo respecto a las formalizaciones de la gramática que se vayan publicando, a falta de un proyecto lingüístico que construya un corpus suficientemente representativo de la lengua y computacionalmente leíble.

En el área de traducción automática en general, se trata de un enfoque inspirado en la forma de construcción del pensamiento humano y abordar un poco el tema del llamado “sentido común”, donde cada componente de la arquitectura de gestores de domino está especializado para generar una interpretación contextualizada a partir del texto a traducir y el conocimiento especializado, restringiendo el problema de desambiguación a contextos conocidos.

3.1.2. Impacto

Esta arquitectura agilizará el desarrollo y validación del traductor, facilitando la delegación de tareas de investigación y programación en diferentes módulos de conocimiento, la alimentación sistemática de vocabulario y la publicación de avances de la herramienta en tiempo real.

Como aporte al estado del arte, se busca aumentar sistemáticamente la flexibilización de la investigación y desarrollo de la arquitectura de un programa traductor y su extensión para ser aplicado con otros pares de lenguas, sobre todo si estas carecen de una colección representativa de frases suficientemente general, como lo es en este caso la LESCO.

Cabe mencionar que en la actualidad existe una brecha de comunicación entre personas sordas usuarias de la LESCO y las personas oyentes, que genera grandes barreras, sobre todo cuando se requiere el acceso a un servicio, limitando el accionar de las personas que presentan esta condición de discapacidad e incurriendo en un acto de discriminación. Por esta razón, una herramienta de traducción no solo facilitaría la comunicación entre personas, sino que garantizaría un derecho establecido como fundamental, inherente a todo ser humano.

3.1.3. Profundidad

Se busca describir una arquitectura flexible y efectiva, explorando y midiendo sus resultados. Se espera que las pruebas experimentales sirvan de base comparativa para luego aplicar otras o futuras metodologías utilizadas en el campo de traducción automática y así evaluar una posible integración a la arquitectura desarrollada.

3.2. Objetivos

3.2.1. Objetivo general

Desarrollar una arquitectura distribuida para la traducción de un texto en español a LESCO, orientada por la concurrencia de módulos especializados en contextos temáticos y lingüísticos, capaces de reconocer y construir una expresión válida en su contexto a partir de una sección del texto de entrada.

3.2.2. Objetivos específicos

- Diseñar una metodología de inicialización (prueba de concepto de un aprendizaje supervisado), análisis y elección de contexto, usada para procesar el corpus correspondiente en cada gestor de contexto.
- Aplicar la arquitectura de dominios de conocimiento distribuidos como un módulo complementario del Traductor LESCO del TEC Digital, para la desambiguación del texto de entrada.
- Validar resultados del proceso de traducción con temas relevantes para miembros de la comunidad sorda, como salud, política, economía, cultura sorda y accesibilidad.
- Publicar los resultados más relevantes de los experimentos.

3.3. Alcance

- Durante el desarrollo de la arquitectura y pruebas de resultados de traducción, se considerará el uso y evaluación de herramientas relacionadas con traducción automática y de desambiguación del sentido de la palabra de código abierto (por ejemplo: *FreeLing*[2], *Moses*[3], *Word2Vect*[4]).
- La implementación de las bibliotecas de NLP y MT como parte o totalidad de un demonio está sujeta a su facilidad de uso y factores técnicos, así como el enfoque de los experimentos. De este modo, no se asegura que se utilicen todas las mencionadas en el punto anterior.

[2]Sitio web de *FreeLing*: http://nlp.lsi.upc.edu/freeling
[3]Sitio web de *Moses*: http://www.statmt.org/moses
[4]Sitio web de *Word2Vect*: https://deeplearning4j.org/word2vec

- No se profundizará en el análisis de entrenamientos de algoritmos de aprendizaje que no encajen con la arquitectura propuesta ni métodos de extracción de información.

- Los dominios de traducción estarán limitados como mínimo a cinco temas relevantes para la comunidad sorda costarricense con el fin de realizar demostraciones de casos reales.

- La representación final de las señas por parte del Traductor LESCO no contempla la totalidad del uso de reglas gramaticales específicas de la LESCO, al hacer falta más estudios de uso y formalidad en esta lengua, por lo que la medición está enfocada en qué tan correcta es la interpretación de un demonio acerca del texto que analiza, para luego dar un acercamiento de frase en LESCO a partir de una construcción semántica sin ambigüedades contextuales.

- No se estudiarán las características morfológicas de la LESCO, al ser esto parte del proceso de realización de las señas, que es trabajo del componente correspondiente del Traductor LESCO. Sólo se tratará hasta el nivel de representación de glosas publicado por el diccionario LESCO del Centro Nacional de Recursos para la Educación Inclusiva (CENAREC) (*CENAREC - Gramática - Proyecto de Descripción Básica de la LESCO*, s.f.).

- La representación de los conceptos en el resultado visual de la traducción dependerá de la cantidad de vocabulario visual manejado por el Traductor LESCO. Por ejemplo, textos muy especializados podrían presentar más casos de deletreo que una versión normalmente señada por la comunidad sorda. Se tratará de minimizar esto orientando los contextos a evaluar hacia un uso más frecuente del vocabulario oficial del diccionario de señas de la LESCO realizado por el CENAREC, en el que se basa del Traductor LESCO del TEC Digital.

- Como prueba de concepto, se propone una metodología sencilla de desambiguación que incluye una aproximación de árboles de términos (Saggion y Lapalme, 2000), cuya implementación se explica en la sección 4.1.2.

3.4. Entregables

A continuación se citan los productos completos que deberán ser entregados al final del proceso de desarrollo de la tesis:

- Diseño detallado de la arquitectura implementada en el Traductor LESCO (capítulo 4 del presente documento de tesis).
- Una implementación funcional de la arquitectura de traducción propuesta, como módulo del proyecto Traductor LESCO del TEC Digital.
- Documento de tesis con los resultados, hallazgos y conclusiones acerca del desarrollo de la metodología hacia la solución del problema y las pruebas experimentales en cada contexto temático seleccionado.
- Al menos una publicación en conferencia indexada o revista indexada relacionada con un resultado importante del proceso de tesis.

4 Método

Se propone el desarrollo de una arquitectura flexible y eficaz para abordar el problema de traducción de la lengua española a la LESCO, en una forma sistemática y que permita la integración de nuevo vocabulario de la forma más intuitiva y mantenible, con énfasis en ofrecer traducciones considerando el contexto correcto del mensaje de entrada.

La hipótesis es que la distribución del problema de traducción a contextos temáticos, con reconocimiento de patrones lingüísticos, aumente la calidad de traducción final al especializarse la generación de representaciones para la LESCO. Lo anterior corresponde con las observaciones de trabajos recientes en el campo de la traducción a lengua de señas tales como (Filhol y cols., 2015) y (Porta y cols., 2014), así como su relación con el estudio de contextos definitorios (Sierra, 2009).

La estrategia se trata de una idea intuitiva sobre la forma en que el cerebro humano procesa el lenguaje: durante la recepción de un discurso en forma secuencial, la mente está continuamente haciendo asociaciones de los fonemas que identifica, llegando a construir significados de forma paralela y subconsciente. Al terminar de interpretar cada frase, se habrán formado varios significados para esta, de los cuales la atención de la persona elige el que mejor se adapte al contexto del discurso hasta el momento. Si se requiere comunicar el significado comprendido, el mismo sector que generó el significado sabe cómo hacer las construcciones adecuadas para sintetizarlo de acuerdo a su contexto.

Este procesamiento del lenguaje es naturalmente paralelizable y modularizable, de modo que permite establecer jerarquías de redes, donde un módulo principal es el que toma la decisión final de la iterpretación de la entrada. Otra ventaja de esta modularización es la posibilidad de mezclar varias técnicas de enfoques de traducción especializados en contextos en los que funcionen muy bien, simplificando su implementación y mantenimiento.

En el caso de los traductores basados en redes neuronales, se puede reducir el costo computacional al repartir el vocabulario de entrenamiento, parecido a la idea de (Jean y cols., 2015), en que logran mejorar el rendimiento para el entrenamiento de un sistema neuronal de traducción por medio de la selección de partes del corpus de entrenamiento.

En la siguiente sección se detalla la idea general explicada anteriormente, describiendo los componentes fundamentales de la arquitectura.

4.1. Descripción teórica de la arquitectura del traductor

La arquitectura para traducción automática de lengua española a LESCO guiada por dominios de conocimiento distribuidos, ilustrada en la figura 4.1, consiste de las siguientes partes, cada una explicada en las subsecciones siguientes:

- Preprocesador de entrada
- Gestor de dominio (demonio)
- Bolsa de pensamientos
- Gestor actual (demonio atento)
- Generador de lengua destino

4.1.1. Preprocesador de la entrada

Esta primera etapa se encarga de convertir el texto de entrada a una representación computable que será consultada por los gestores de dominio. Se aplica como una biblioteca de procesamiento de lenguaje natural que se enfoca en la tokenización del texto de la entrada. Cada token se envía a cada uno de los demonios existentes en la arquitectura, de forma secuencial conforme a su aparición en el texto original.

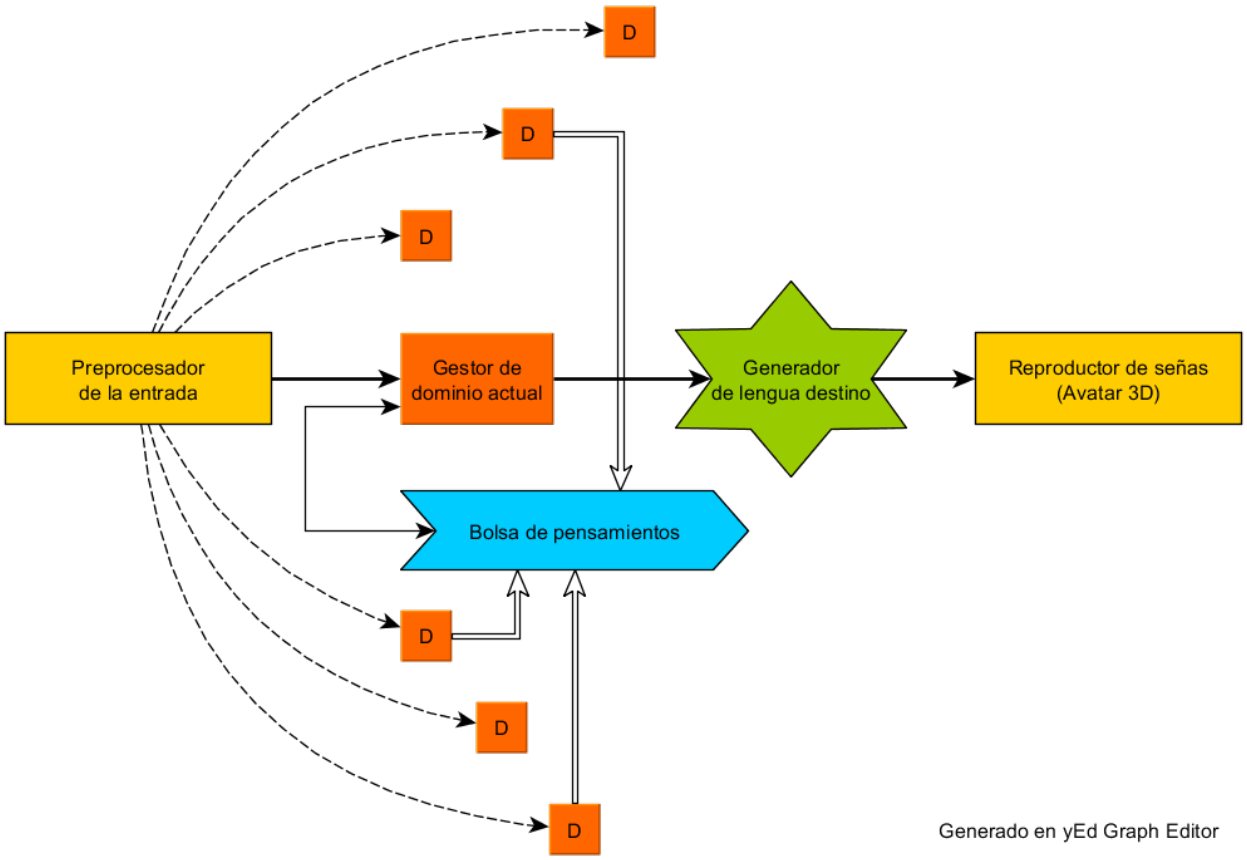

Figura 4.1: Arquitectura propuesta. El procesador envía a todos los gestores de dominio los *tokens* del texto a traducir, luego las flechas blancas indican los resultados de los dominios más significativos que van a terminar en la bolsa de pensamientos. El gestor de dominio actual interactúa con la bolsa para elegir el mejor contexto, y brinda el resultado al generador de la lengua destino, que se encarga de construir la descripción física de las señas que reproducirá el avatar.

El análisis puede brindar formatos de resultado de acuerdo a las necesidades del diseño de los demonios: textos cortos, árboles de parseo, clases de datos, vectores de palabras, incluso podría publicar varios de estos formatos en flujos simultáneos.

4.1.2. Gestor de dominio (demonio)

Se trata de un proceso programado para la detección de patrones en el flujo de entrada para un contexto específico. Existen varios de estos gestores ejecutándose concurrentemente y analizando la entrada para sugerir interpretaciones de su significado, esto al coincidir un patrón de frase característico en su dominio de conocimiento. El proceso es el siguiente:

1. Todos los demonios obtienen un token de la entrada a la vez.

2. Conforme analice cada token, el gestor de dominio elabora progresivamente una estructura de interpretación que denota su sentido semántico, esto de acuerdo con la implementación específica de cada demonio para evaluar su contexto. Esta estructura puede ser directamente la traducción hacia la lengua destino o una interlengua a ser procesada en el generador de la lengua de salida. Casos especiales de la lengua origen y destino, como las palabras *Out-Of-Vocabulary (OOV)*[1], se pueden tratar en este paso.

3. Al analizar todos los tokens de la frase, entonces el demonio calcula el valor semántico hasta el momento, y si alcanza un valor mínimo de aceptación, es decir, si la frase tiene un sentido aceptable en el contexto del demonio, se enviará esta estructura a la bolsa de pensamientos con el valor semántico con el que ha quedado respecto a su dominio.

[1] *Out-Of-Vocabulary (OOV)* son términos no existentes en el conocimiento manejado por el traductor.

- Si durante el proceso se llega a un valor semántico de aceptación sin haber procesado una frase completa, significa que la subfrase analizada hasta el momento tiene una estructura propia del contexto del demonio, hecho que se notifica inmediatamente a la bolsa de pensamientos[2].

4. Después del análisis de todos los gestores, el gestor actual toma en cuenta los resultados que podrían estar en la bolsa de pensamientos y el suyo propio para decidir la mejor traducción de la frase actual.

5. Se repite el proceso desde el paso 1 con la siguiente frase y así sucesivamente hasta terminar todo el documento de entrada.

Dada la libertad en la estrategia de detección de los gestores de dominio, estos pueden ser implementados desde formas muy directas para el tipo de token que están esperando, o pueden ser subsistemas de traducción completos que funcionen en una red de traductores automáticos. Lo que se debe asegurar es la consistencia de la metodología del dominio actual para discernir la mejor opción presentada en un determinado momento de la traducción. Además, un demonio podría utilizar resultados de otro que esté especializado en encontrar un subpatrón, teniendo la posibilidad de construir jerarquías de demonios, constituyendo una ontología de conocimiento por contexto temático y su representación gramatical.

4.1.3. Bolsa de pensamientos

El objetivo de esta estructura de datos es recolectar los resultados de identificación de los demonios con cierto grado de sentido semántico durante el análisis de una frase y evaluarlas en un momento puntual del proceso de traducción. Esta es una estructura compartida entre todos los demonios, aunque en principio sólo el gestor actual tendrá acceso para cuando se escoja el mejor contexto para esa frase.

[2]Este caso será analizado con profundidad en la implementación de la arquitectura.

4.1.4. Gestor actual (demonio consciente)

Se trata de un rol especial para uno de los gestores de contexto que centraliza el flujo de datos en un momento del proceso de traducción para una frase. Existe un procedimiento de “sentido común” usado por el demonio actual, que será utilizado para escoger la mejor opción presente en la bolsa de pensamientos en un momento dado. El criterio de elección tiene varios parámetros para sacar su decisión: el peso de pertenencia de la opción al demonio que lo propuso, la información del contexto del gestor actual y los contextos elegidos para traducir las frases anteriores.

4.1.5. Generador de lengua destino

Se trata de un módulo enfocado en la realización de la lengua destino a partir de los resultados de los gestores de dominio. Dependiendo de la implementación técnica de los gestores, el generador de lengua destino puede ser necesario luego de cada frase traducida o al final del proceso de “demonificación” del texto de entrada, por ejemplo, si los demonios producen una representación intermedia (interlengua), será necesario acoplar un módulo generador para la salida deseada.

Sin embargo, puede ser útil que cada demonio genere casos especiales de representación que, en un contexto más general para la lengua de salida, son difíciles de ubicar. Ejemplos son la representación de conceptos multipalabra, tipos de enumeraciones o el manejo de espacios mentales, de acuerdo al estudio de (*CENAREC - Gramática - Proyecto de Descripción Básica de la LESCO*, s.f.).

4.2. Implementación de la arquitectura del traductor

La arquitectura propuesta se implementa en C++ como un módulo de servicio para la desambiguación, adaptado a la estructura del prototipo del Traductor LESCO. Esta se compone de un servidor FreeLing 4.0 para el análisis lingüístico de la entrada en español (Padró y Stanilovsky, 2012), una base de datos *PostgreSQL 9.4* para la gestión de datos de vocabulario en español y LESCO, un *back-end PHP 5.6* con lógica de traducción y un *front-end* JavaScript incluyendo un avatar tridimensional programado en *Unity 5.4*. La figura 4.2 muestra la forma en que se comunican estos componentes.

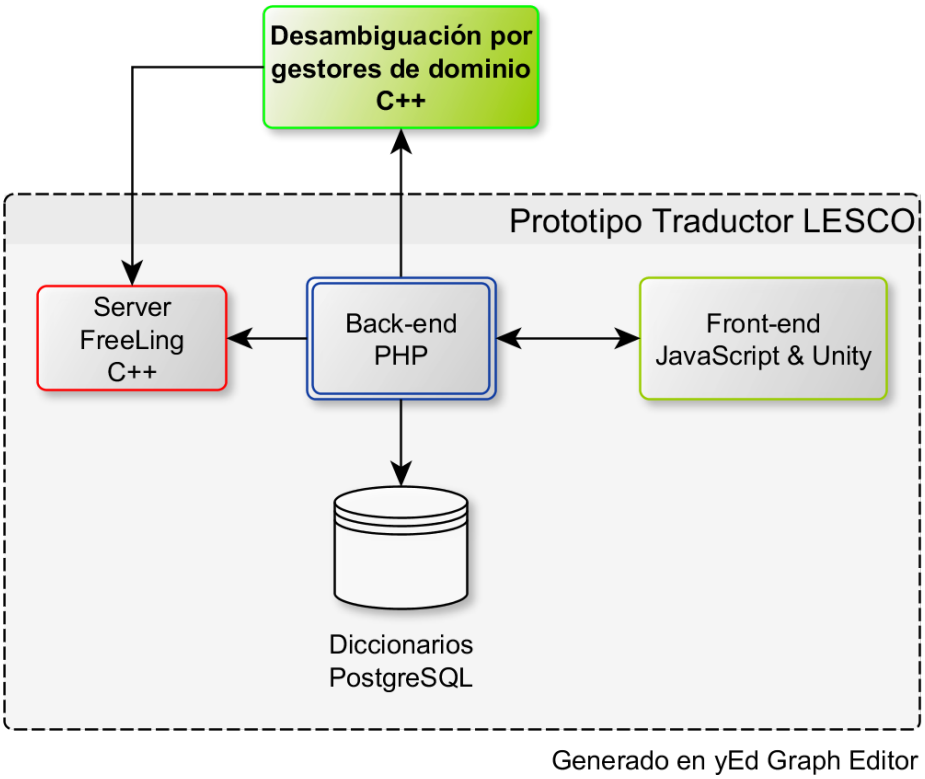

Figura 4.2: Integración de módulo desambiguador con el prototipo actual del Traductor LESCO.

A continuación se describe la forma en que cada parte de la arquitectura de gestores de domino fue construida para su utilización en el Traductor LESCO.

4.2.1. Implementación de preprocesador

El preprocesador de la entrada realiza dos tareas principales:

1. Analizar sintácticamente el texto de entrada con una herramienta de parseo morfosintáctico. Se utiliza FreeLing para esta tarea.

2. Publicar de forma secuencial cada frase, construyendo una estructura que es compartida por los hilos de ejecución que representan a los demonios. Cada frase incluye un análisis de dependencia con estructura de árbol sintáctico.

El análisis de dependencia de un texto por parte de FreeLing resulta en un conjunto de datos relacionados por cada frase identificada. Cada frase es descrita por una tabla con registros de cada palabra/token especial con un código de identificación. Este código luego es usado como referencia en una estructura adicional que representa el árbol de dependencia sintáctico, como el ejemplo de la figura 4.3.

En la biblioteca FreeLing, los textos son procesados en una secuencia de módulos de procesamiento textual y lingüístico: tokenización $\longrightarrow$ separador de frases $\longrightarrow$ analizador morfológico $\longrightarrow$ otros módulos como etiquetador de significado, desambiguación o etiquetadores Part-Of-Speech. Existe un programa en el paquete de FreeLing 4.0 llamado `analyze` que se utiliza como un servidor de muchas de estas funciones y las configura como módulos en un archivo, lo que permite cuáles seleccionar y cambiar su comportamiento específico según el tipo de análisis deseado.

El apéndice E muestra el archivo de configuración usado para este trabajo, donde el principal cambio es la activación del análisis de dependencias para la construcción de los árboles sintácticos. También se configuró para una salida en formato JSON para facilitar el intercambio de datos en el programa principal.

```
{ "id":"1",
  "tokens" : [
    { "id" : "t1.1", "form" : "Mi", "lemma" : "mi", "tag" : "DP1CSS", "ctag" : "DP", ...},
    { "id" : "t1.2", "form" : "nombre", "lemma" : "nombre", "tag" : "NCMS000", "ctag" : "NC", ...},
    { "id" : "t1.3", "form" : "es", "lemma" : "ser", "tag" : "VSIP3S0", "ctag" : "VSI", ...},
    { "id" : "t1.4", "form" : "Juan", "lemma" : "juan", "tag" : "NP00000", "ctag" : "NP", ...},
    { "id" : "t1.5", "form" : ".", "lemma" : ".", "tag" : "Fp", "ctag" : "Fp", ...}],
  ...,
  "dependencies" : [
    {"token" : "t1.3", "function" : "top", "word" : "es", "children" : [
      {"token" : "t1.2", "function" : "subj", "word" : "nombre", "children" : [
        {"token" : "t1.1", "function" : "spec", "word" : "Mi"}
      ]},
      {"token" : "t1.4", "function" : "attr", "word" : "Juan"},
      {"token" : "t1.5", "function" : "punc", "word" : "."}
    ]}]
}
```


dependencies

token	function	word	children
t1.3	top	es	token: t1.2, function: subj, word: nombre, children: [token: t1.1, function: spec, word: Mi]; token: t1.4, function: attr, word: Juan; token: t1.5, function: punc, word: .

Generado en json2table.com

Figura 4.3: Ejemplo de las estructuras de datos exportadas por el análisis de dependencia de Freeling. La frase analizada es *“Mi nombre es Juan.”*

4.2.2. Implementación de demonios

La implementación de los demonios se basa en un sistema concurrente con un hilo que toma el rol del gestor actual para centralizar los resultados de los demás. Se han programado gestores de tipo *contextual*, encargados de detectar las frases más representativas de un tema específico, tan general o específico como se requiera dependiendo del tamaño del corpus que lo alimente. Al terminar el procesamiento de cada frase, la sincronización se realiza con barreras de la biblioteca *pthreads*, donde cada uno de los hilos creados guarda su resultado en la estructura de la bolsa de pensamientos y prepararse para la siguiente frase.

Cada gestor de contexto ejecuta el mismo procedimiento genérico para el análisis de la frase, la diferencia está en los datos del modelo del contexto. La construcción y ejecución es así:

1. En la inicialización del programa, existe una carpeta `models` que contiene los archivos de modelo de contexto. Por cada uno de ellos se creará un hilo de gestor contextual.

2. Al terminar el análisis de la entrada, se procede a analizar el árbol de dependencia correspondiente a una frase por parte de los gestores y se busca un patrón de entrada en el modelo descrito por el archivo. Este recorrido otorga un *porcentaje de pertenencia* de dicha frase al contexto temático del gestor.

4.2.3. Algoritmo de construcción del grafo de términos

La preparación de los modelos para ser utilizados por los gestores de contexto, antes de la ejecución del módulo de desambiguación, se hace por medio de un programa auxiliar de construcción de grafos contextuales a partir de archivos de texto. Por esta tarea, a este programa se le denomina *teacher*.

El resultado devuelto por el *teacher* se trata de un archivo que define un grafo de términos dirigido o ***grafo contextual***, formando una estructura jerárquica gramatical. Así, el primer nivel lo constituye verbos que definen acciones características del dominio, seguido de relaciones de uso y complementos con el resto de componentes conceptuales de una frase característica del contexto.

La figura 4.4 ilustra un ejemplo de esta estructura en el contexto de *"ejercicio del cuerpo"*, con cada concepto acompañado del valor de conteo respecto al pequeño corpus de ejemplo, en el cuadro superior izquierdo.

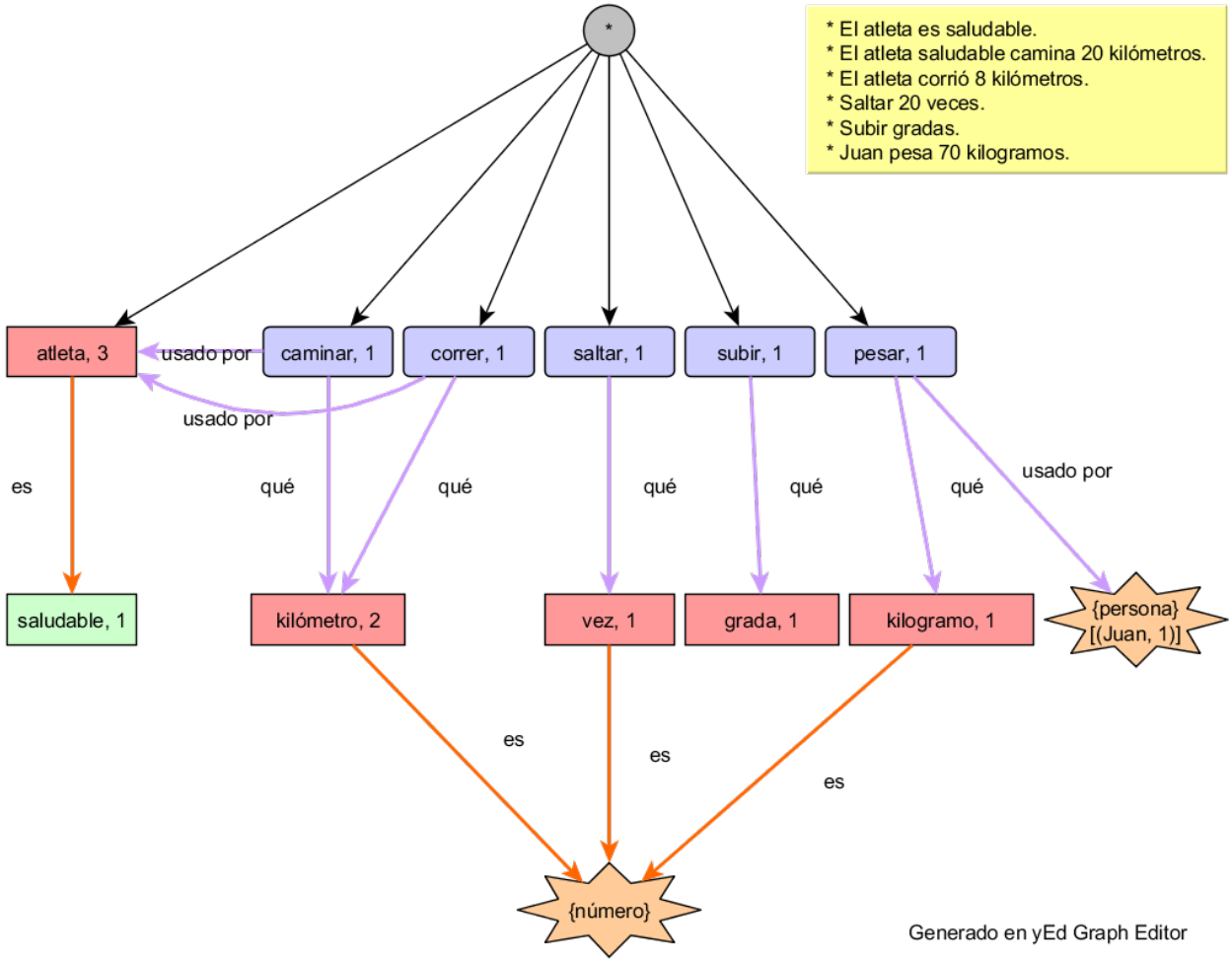

Figura 4.4: Grafo de términos para un contexto temático.

El proceso de adquisición de conocimiento en grafos contextuales es independientemente del proceso de traducción, al generar archivos de modelo para que los demonios construyan directamente los árboles sin tener que analizar el corpus original en tiempo de ejecución. Un ejemplo del formato de estos archivos se puede consultar en el apéndice F.

Para construir el grafo de términos de un gestor contextual, se realizan los siguientes pasos:

1. Construir un corpus de ejemplos de frases representativas del tema. Una frase representativa de un tema específico debe ser una oración alusiva al concepto que define el nombre dado al contexto (Barrada, 2007), de modo que su secuencia léxico-gramatical sea asociada siempre con este. Definir el grado de alusión es una tarea lingüística cuya validación queda fuera del alcance de este estudio, pero que en el análisis de los experimentos se comentará.

2. Analizar sintácticamente el corpus con la herramienta *Freeling*, obteniendo un árbol de dependencias por cada frase del corpus.

3. Cada árbol se recorre desde su raíz para colocar los lemas[3] en el grafo de términos para el demonio. Si hay algún concepto ya incluido en el grafo de términos, se creará un arco desde el nodo actual hacía ese término y se aumenta su marcador de conteo.

4. Al insertar un nuevo lema, se crea un marcador de conteo que denota su frecuencia de uso en el corpus, el cual aumenta cada vez que el término se utilice en otras frases (detalles en el apartado 4.2.3.1). Además, existe un contador de arcos general por cada tipo de concepto significativo que se inserte en el grafo contextual.

[3]Un lema es una representación neutral de un término sintáctico. Por ejemplo, *caminar* es lema de *caminado*; *amarillo* es lema de *amarilla*.

5. Al terminar de incluir todas las frases del corpus, el grafo contextual se recorrerá en su totalidad para calcular el porcentaje de pertenencia de cada nodo conceptual, de modo que refleje la importancia global de ese concepto dentro del contexto construido. Se utiliza un ponderado relativo al número total de arcos de entrada del nodo respecto a la cantidad de arcos del tipo conceptual del nodo. El resultado es el porcentaje de pertenencia de este nodo particular respecto al resto de los nodos dentro del grafo contextual.

De acuerdo a las formas de los árboles de dependencia, se espera que en el primer nivel después de la raíz del grafo, se encuentren los conceptos de entrada más distintivos del dominio de conocimiento, divididos en dos áreas: verbos que denoten acciones muy frecuentes en el tema, y sustantivos que denoten definiciones.

4.2.3.1. Métrica del porcentaje de pertenencia a un contexto

La métrica para calcular el porcentaje de pertenencia de cada nodo tiene las siguientes propiedades:

- El porcentaje de pertenencia de un nodo refleja la importancia del concepto para ese contexto, basado en la frecuencia de su utilización en el corpus y su tipo.
- Los pesos de cada tipo de concepto son proporcionales a 1 y se clasifican así:
  - Acciones (verbo) definido como $W_{Act} = 0{,}60$.
  - Entidades (sustantivo) definido como $W_{Ent} = 0{,}30$.
  - Atributos (adjetivo, adverbio) definido como $W_{Atr} = 0{,}09$.

  - Se calcularán con peso $W_{Stw} = 0{,}00$ palabras auxiliares del lenguaje de tipo *stopwords* que no aportan por sí solas un significado temático[4]. Tampoco contarán símbolos ortográficos como los de puntuación, interrogativos, exclamativos y paréntesis, y en el caso de cierto tipo de palabras que funcionen como *verbos atributivos*[5].
  - Otros tipos de palabra no contemplados tendrán peso $W_{Other} = 0{,}01$.

- El porcentaje de pertenencia para un nodo dentro de un grafo contextual, Per, se define como:

$$Per_i = \frac{k_i}{K_t} \cdot W_t$$

, donde k_i es el número total de arcos de entrada al nodo i, K_t es el total de arcos de tipo conceptual t para el nodo i y W_t es la constante de peso para el tipo conceptual t, con $W_t \in \{W_{Act}, W_{Ent}, W_{Atr}, W_{Other}, W_{Stw}\}$.

- La suma de todos los porcentajes de pertenencia de un grafo contextual es 1, de modo que:

$$\sum_{i=1}^{N} Per_i = 1$$

, donde N es el total de arcos en el grafo conceptual.

Durante el análisis de un texto, el índice de pertenencia de una frase es la suma de los nodos contextuales cuya conexión logra coincidir con toda o parte de la forma del árbol de dependencias de la frase. Conforme el análisis avanza, se incrementa un índice acumulado de pertenencia de un texto por cada uno de los contextos en un arreglo que funciona como el marcador global. Esto brinda una forma de medir cuánto de las frases del texto significan a cada contexto.

[4]Los ejemplos más frecuentes son los artículos (*el*, *la*, *un*, *unos*, *los*) y conjunciones (*y*, *o*).

[5]Estos son verbos que pueden ser omitidos y la frase sigue significando lo mismo: ser, estar.

En la figura 4.5 se observa un ejemplo de la secuencia del cálculo del índice de pertenencia para una frase en un sector de un grafo conceptual. La suma de los porcentajes de pertenencia de cada nodo comienza con la transición desde el nodo raíz al núcleo verbal de la frase, siguiendo el sintagma nominal compuesto por las palabras *"el"* y *"saludable"*, para luego hacer la suma con los complementos verbales.

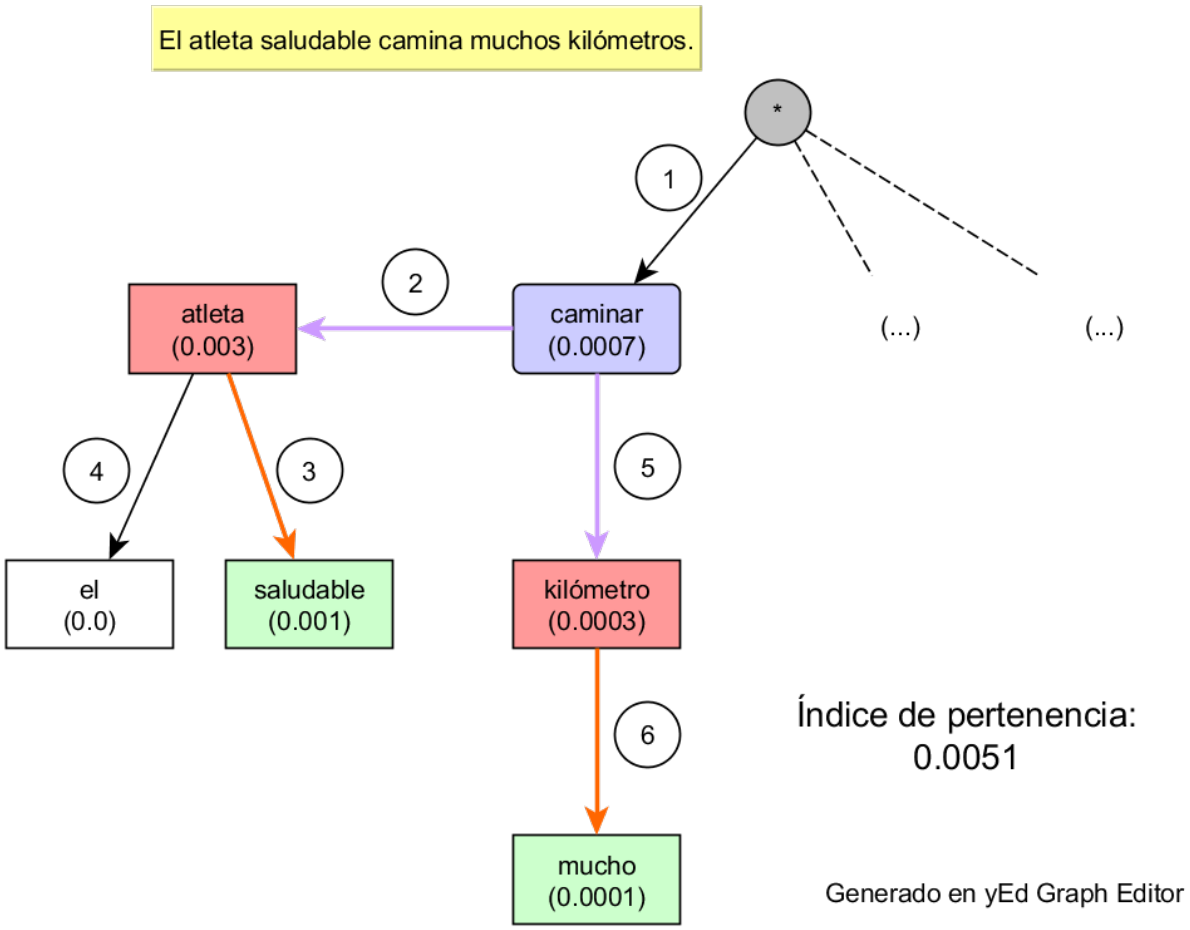

Figura 4.5: Secuencia del cálculo de índice de pertenencia de una frase en un grafo contextual.

4.2.4. Implementación de bolsa de pensamientos

La bolsa de pensamientos, ilustrada en la figura 4.6, consiste en un vector de referencia hacia los resultados de los gestores, donde existe el valor semántico de su interpretación por contexto. Cada posición del vector esta asociada al identificador entero del demonio, en una relación uno a uno.

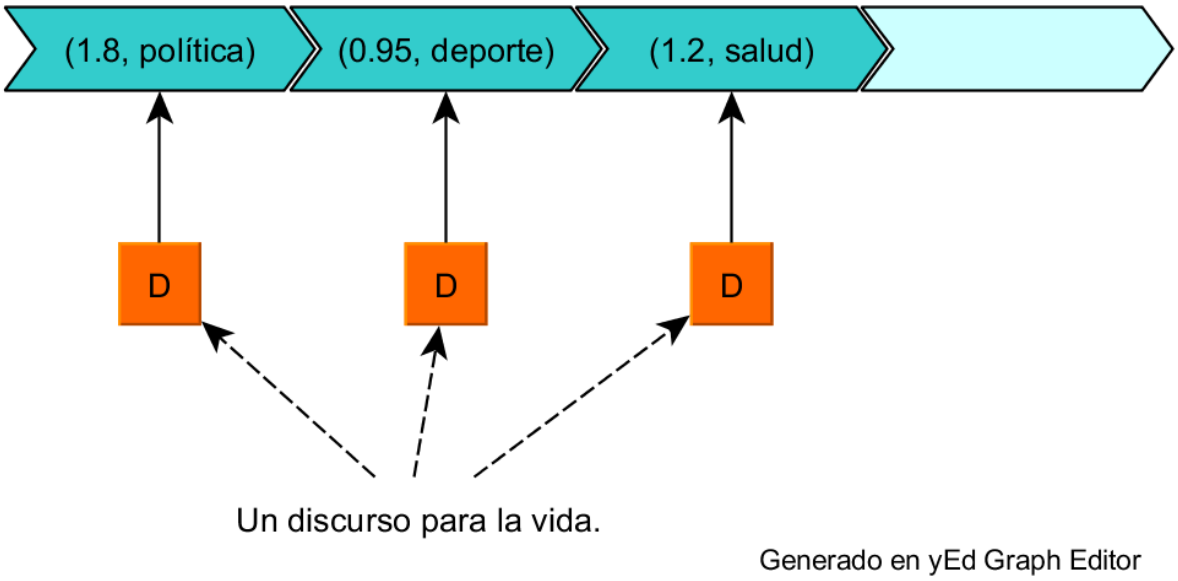

Figura 4.6: Ejemplo de una bolsa de pensamientos con tres interpretaciones de tres demonios con su correspondiente valor semántico.

Además, el *marcador de contextos* es otra estructura que lleva el valor acumulado de los índices de cada gestor contextual. Es muy similar a la bolsa, con la diferencia de que la bolsa se reinicia con valores 0 después de cada frase analizada, mientras el marcador no se borrará hasta que el análisis de todo el texto de entrada termine. La finalidad de este marcador es que sea una forma de medir qué contexto tiene mayor incidencia en los patrones del texto en un momento determinado, independientemente del contexto con mayor puntaje en una frase determinada. Esto puede ayudar a seleccionar el contexto ante casos como marcadores muy parecidos entre contextos.

4.2.5. Implementación de comportamiento del gestor actual

El gestor actual ejecuta una sección del algoritmo de análisis donde se escoge un contexto y se puntúa la frase en la bolsa de pensamientos. La implementación de la selección para el gestor actual es así:

- Al analizar la frase inicial, el demonio consciente es el primero en una lista indexada de hilos. Cada vez que se desconoce el contexto de una frase, es decir, todos los índices de pertenencia en 0, este hilo se selecciona como el gestor actual.
- Al finalizar el análisis de una frase, el demonio consciente revisa la bolsa de pensamientos para escoger el contexto con el índice de pertenencia más alto. Cada puntuación de frase se suma al marcador de contexto, que determina el valor semántico del documento analizado en los contextos que vayan surgiendo.
- El demonio consciente limpia la bolsa de pensamientos para proceder al análisis de la siguiente frase.

4.2.6. Interfaz con generación de LESCO

Para la generación se acoplan los resultados del proceso de desambiguación con el módulo de síntesis ya programado en el Traductor LESCO. Para este proceso, el módulo correspondiente recibe los datos en formato JSON con la secuencia de frases analizadas por el módulo desambiguador. A continuación, pasará por varias descripciones particulares para su reproducción en un avatar, incluyendo gestos manuales, velocidades, formas de mano y deletreos en caso de conceptos sin seña en la base de datos del generador. Un ejemplo de la estructura recibida por el generador se puede ver en el apéndice G.

4.3. Ejemplo de ejecución de la arquitectura

A continuación, se detalla un ejemplo de uso de la arquitectura analizador por pasos, desde el aprendizaje de contextos hasta el análisis contextual y producción del resultado de desambiguación, proceso que se puede visualizar en la figura 4.7.

1. Se realizan los archivos de corpus manualmente para cada contexto temático. Es importante asumir que toda palabra correspondiente a un concepto (acción, objeto, atributo) que se incluya es significativa para el contexto. Una mayor frecuencia de esta palabra denotará una mayor importancia, así que se debe tratar de encontrar un número de ejemplos adecuado para pesar la importancia, basado en las frases más frecuentes o utilizadas del contexto temático.

2. Con los corpus listos, se envía a procesar uno por uno los archivos al servidor de Freeling, en su configuración de "análisis de dependencia" y con salida en formato JSON. Cada uno de estos archivos contiene la lematización de los conceptos y los árboles de dependencia para cada oración del corpus.

3. Con los archivos producidos por el análisis de dependencia, se envía a procesar uno por uno en el módulo teacher para producir su archivo de modelo correspondiente. El teacher parsea cada archivo y arma un grafo conceptual general del tema con los árboles de dependencia de las oraciones del corpus, clasificando y filtrando de acuerdo a su función sintáctica y lemma. Los archivos de modelo se guardan en la carpeta `models` contiguo al ejecutable del analizador.

4. El analizador contextual se ejecuta, construyendo un hilo de gestor contextual por cada archivo de modelo válido encontrado en la carpeta de modelos. El analizador queda entonces como un servicio listo para recibir datos por medio de un pipe Linux.

5. Ya en el proceso de traducción, el analizador recibe un resultado de análisis de dependencia de Freeling desde el texto original y en formato JSON, parseándolo y estructurándolo para su lectura por parte de los gestores de dominio.

6. Cada gestor procesa una frase a la vez, de forma independiente entre ellos (aunque puede haber posibilidad de cooperación de resultados o delegación del análisis). El proceso consiste en un recorrido comparativo entre el árbol de dependencia de la oración de entrada y el grafo conceptual desde el nodo inicial, sumando la puntuación de cada concepto que corresponda.

7. Al finalizar el proceso de una frase, cada demonio guarda la puntuación en la bolsa de pensamientos, se cambia el rol del demonio consciente hacia aquel que tenga la mayor puntuación, y este procede a la generación de la frase traducida. Esto puede tomar en cuenta el marcador de contextos como criterio para la escogencia del demonio actual.

8. La generación de la frase se hace a partir del demonio actual que puede tener una implementación diferente acorde a la gramática de su tema. Se incluyen datos como lemma, texto original y etiqueta morfológica.

9. La frase generada queda en un arreglo secuencial que, al final de todo el proceso, se exporta como una estructura JSON hacia el pipe de salida de la aplicación. Este pipe sería usado por el módulo de generación del traductor LESCO.

10. Al recibir un token especial, el servicio limpia sus recursos (pipes y threads) y se cierra el proceso.

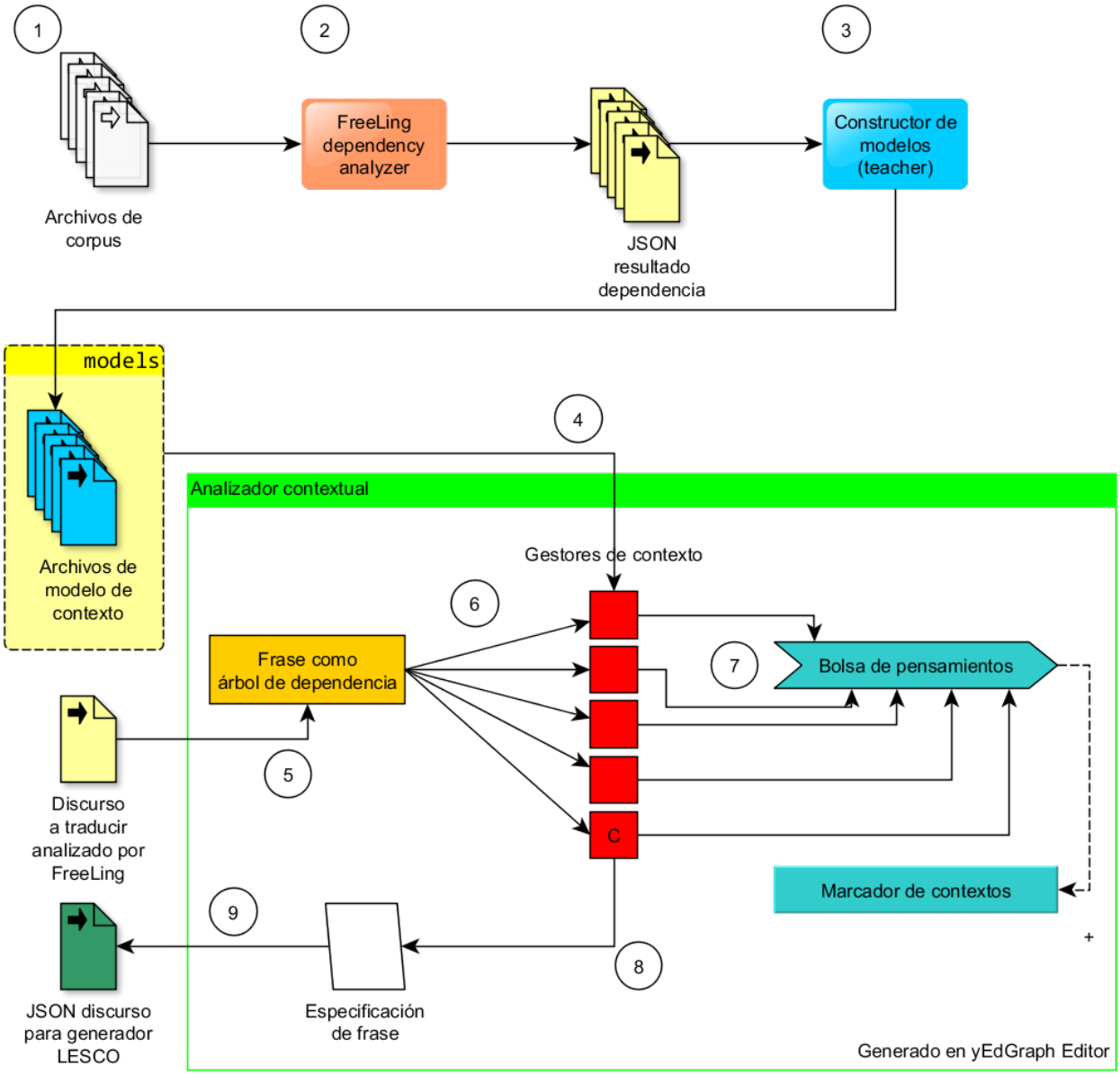

Figura 4.7: Ejemplo de ejecución de la arquitectura.

5 Experimentos

En este capítulo se detallan las pruebas hechas a la implementación de la arquitectura propuesta en el capítulo 4, así como la muestra y análisis de los resultados obtenidos.

La hipótesis es que se puede mejorar la calidad de las traducciones desde un punto de vista semántico, argumentando el uso de la especialización del contexto temático y su aplicación a un sistema distribuido con una entidad de atención que selecciona la mejor opción. Relacionado a esto, se han determinado variables de la configuración del módulo de análisis contextual que modifican los resultados del proceso, apreciables en la tabla 5.1.

Los experimentos están orientados al análisis de textos de carácter informativo para el público general, usando corpus especializados y validados lingüísticamente, donde se espera que estas variables se manifiesten en el análisis de los resultados y así ayudar a identificar criterios para la elección de un contexto temático, como es determinar índices numéricos de aceptación de una frase. Para comprobar esto con la arquitectura implementada, se realizan dos clases de pruebas: una con textos largos relacionados a una contexto temático y otra con análisis de tres grupos de oraciones individuales.

Para las pruebas de textos largos, se analizan los índices de pertenencia generados por el algoritmo de puntuación de cada contexto, sacando sus relaciones y determinando las condiciones para determinar el comportamiento de los porcentajes de pertenencia adecuados para aceptar un contexto.

Tabla 5.1: Variables del proceso para análisis contextual

Variable	**Efectos esperados**
Tamaño de corpus	Grande: menor porcentaje de pertenencia por nodo. Pequeño: mayor porcentaje de pertenencia por nodo.
Cantidad de contextos	Muchos: más gestores reaccionan ante oraciones desconocidas. Pocos: casi ninguna opción en contextos desconocidos.
Extensión de la entrada	Una entrada grande daría un resultado más exacto; un entrada muy pequeña mantendría un nivel de ambigüedad significativo.
Ambigüedad temática de la entrada	Un texto que trata de varios temas conocidos por el traductor puede contener patrones compartidos entre contextos y producir índices de pertenecia similares.
Estilo de redacción	Tanto para corpus como para la fuente a analizar, este criterio puede hacer la detección de patrones sea más o menos efectiva para textos generales. Por ejemplo: glosarios, diálogos, definiciones, texto técnico descriptivo, narraciones.

Los datos a probar para el primer grupo experimental están compuestos de cinco contextos temáticos específicos: economía, ambiente, derecho, informática y medicina. Los recursos textuales usados son de dos tipos:

- Para construir los modelos de contexto, se utilizaron los archivos en español del corpus temático del Instituto Universitario de Lingüística Aplicada (IULA)[1], una organización de postgrado especializada en estudiar el comportamiento de varias lenguas a partir de esta fuente de datos (*Proyecto Corpus. Corpus textual especializado plurilingüe*, s.f.). Este corpus plurilingüe ha sido analizado por especialistas de cada área de conocimiento mencionada anteriormente, de modo que su vocabulario, estructura gramatical y extensión sean suficientemente significativos para su contexto[2].

- Para las entradas de texto, se han escogido notas informativas sobre cada uno de los cinco temas que maneja el corpus de la IULA, además de una nota adicional sobre un tema desconocido (música). Los textos de estos ejemplos se pueden consultar en el apéndice A.

Las pruebas con oraciones específicas también utilizan los modelos generados de las pruebas de notas temáticas, y se componen de tres grupos de 20 oraciones cada uno, que evaluarán la efectividad del módulo de desambiguación para:

- *Oraciones en el contexto*. Son frases extraídas de cada corpus existente del módulo. Sus resultados sirven para medir el grado de integridad de cada modelo. Se espera que en todos los casos se elija el contexto correcto con un valor relativamente alto en el índice de pertenencia.

[1] El IULA es parte de la Universitat Pompeu Fabra (UPF), Barcelona, España.

[2] Cada contexto se creó a partir de toda o gran parte del corpus español, limitando las fuentes más extensos a ocho mil (8000) líneas (informática y medicina), debido a una limitante técnica descubierta en el parser JSON utilizado en el módulo de desambiguación.

- *Oraciones ambiguas*. Son frases que por su significado pueden pertenecer a dos de los contextos conocidos por el módulo. Se espera que hayan índices de aceptación similares para los dos contextos involucrados o que alguno de ellos logre el máximo índice en cada caso.
- *Oraciones desconocidas*. Son frases que no se consideran pertenecientes a ninguno de los contextos conocidos. Se esperan valores relativamente bajos para los índices de dependencia, donde el caso ideal es cero en todos los contextos.

Las oraciones utilizadas se pueden consultar en el apéndice C.

5.1. Resultados experimentales

A continuación, se describen los resultados de cada experimento de nota temática y un resumen de resultados del experimento de oraciones individuales, realizando el análisis respectivo en cada sección.

5.1.1. Ejecución sobre ejemplos de contextos

En estas pruebas se busca comprobar la efectividad del módulo de desambiguación y analizar el comportamiento de los gestores de contexto cuando analizan una nota informativa. Los textos analizados se pueden consultar en el apéndice A.

En la figura 5.1, se observa que *Economía* fue el contexto elegido, correspondiendo efectivamente al tema del artículo. Esto luego de elegir la mayor cantidad de oraciones (13) y además tener el mayor puntaje acumulado en el marcador de contextos (0.788). Notar que para este ejemplo, *Ambiente* fue el segundo en puntaje general, a pesar de que se eligió menos de la mitad de las oraciones (6) que *Economía*.

La figura 5.2 muestra el comportamiento de los resultados del índice de pertenencia de cada uno de los cinco contextos programados en el módulo de desambiguación, durante el análisis de 36 frases detectadas por FreeLing. Los picos de detección se situaron por debajo del 0.1, sobresaliendo casos alrededor del 0.05. Se notan dos valores importantes

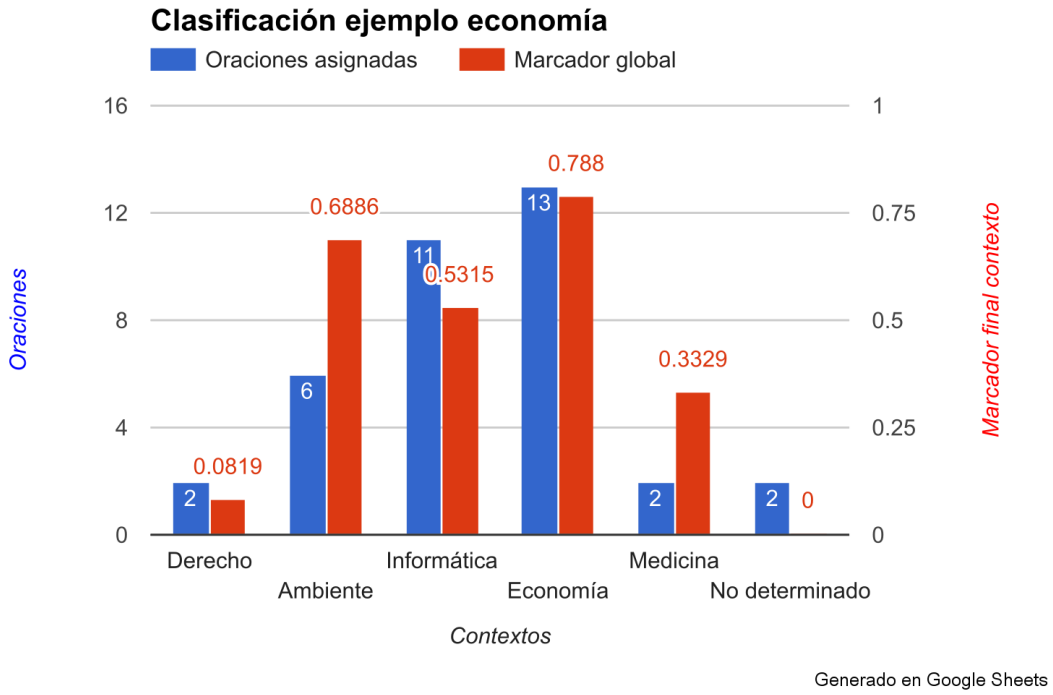

Figura 5.1: Clasificación para texto sobre el tema de *Economía*.

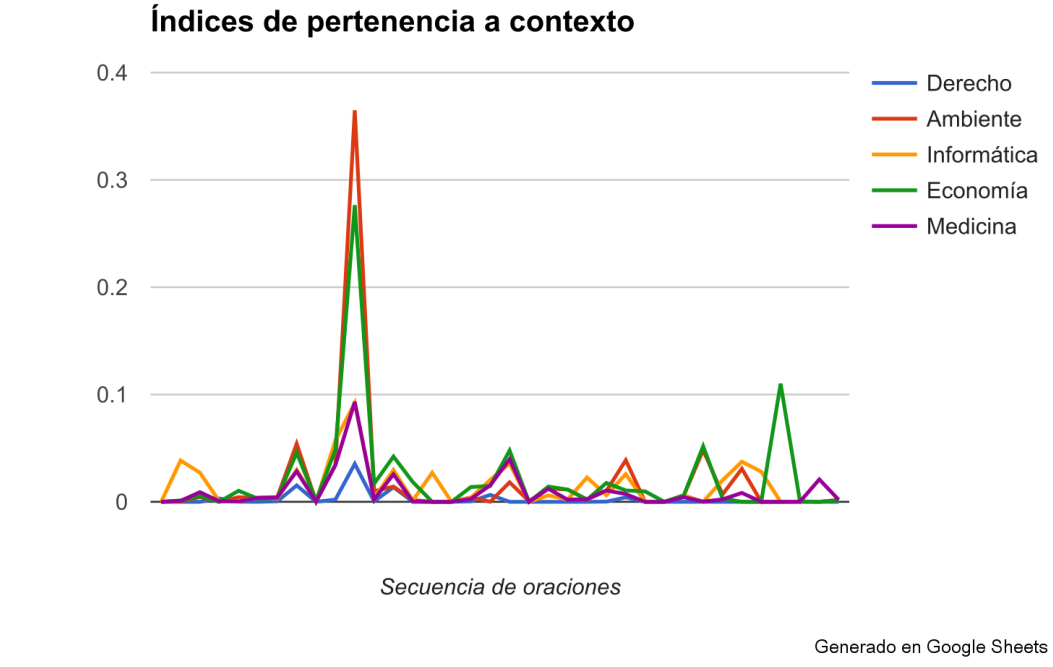

Figura 5.2: Comportamiento de gestores de contexto sobre el ejemplo de *Economía*.

del hilo de *Economía*: 0.2765 para la frase 11 y 0.1101 para la frase 33. El contexto *Ambiente* detectó con un valor de 0.3649 para la frase 11, mismo caso para el máximo del contexto *Economía*. La frase 11 del ejemplo es:

```
La cara oculta de la moneda ha sido, sin embargo, la inflación, que ha
saltado al 2,3% y podría llegar al 2,7% a fin de año, según un estudio
del Center for Economics and Business Research (CEBR).
```

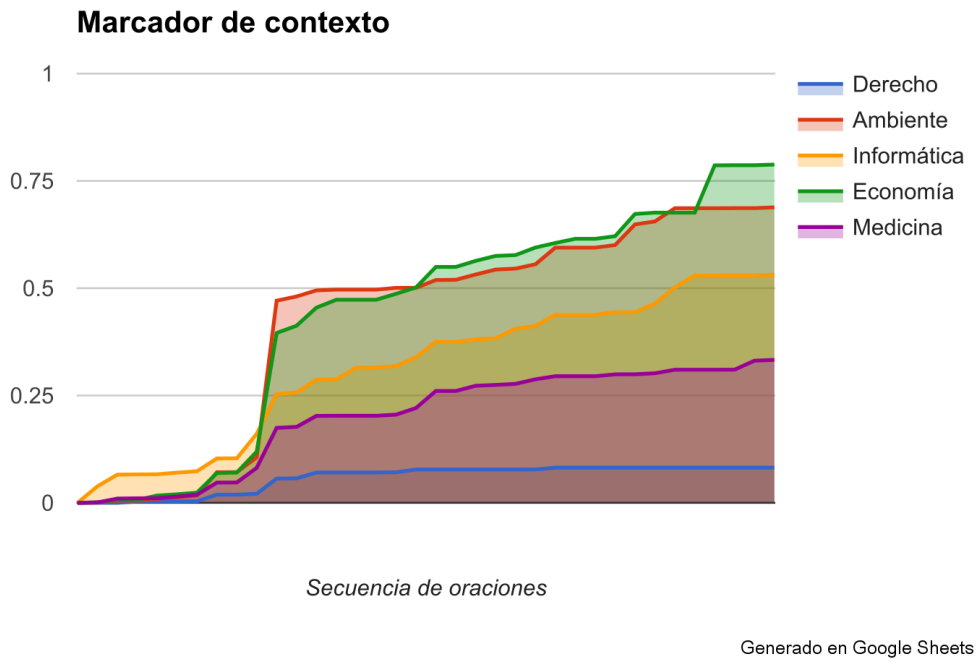

Figura 5.3: Marcador acumulado durante el análisis del ejemplo de *Economía*.

La figura 5.3 muestra el marcador global acumulado por cada contexto durante el análisis de las 36 frases. En este se nota un dominio del contexto *Informática* en el primer cuarto del análisis, para luego llegar al caso de la frase 11 visto en la figura 5.2, cuando por unas cuantas oraciones más, *Ambiente* es el contexto con más puntaje acumulado. El puntaje acumulado del contexto *Economía* sigue su crecimiento hasta ser el de mayor puntaje después de la mitad del análisis, y así prácticamente hasta el final, donde saca una diferencia importante con el pico de la oración 33.

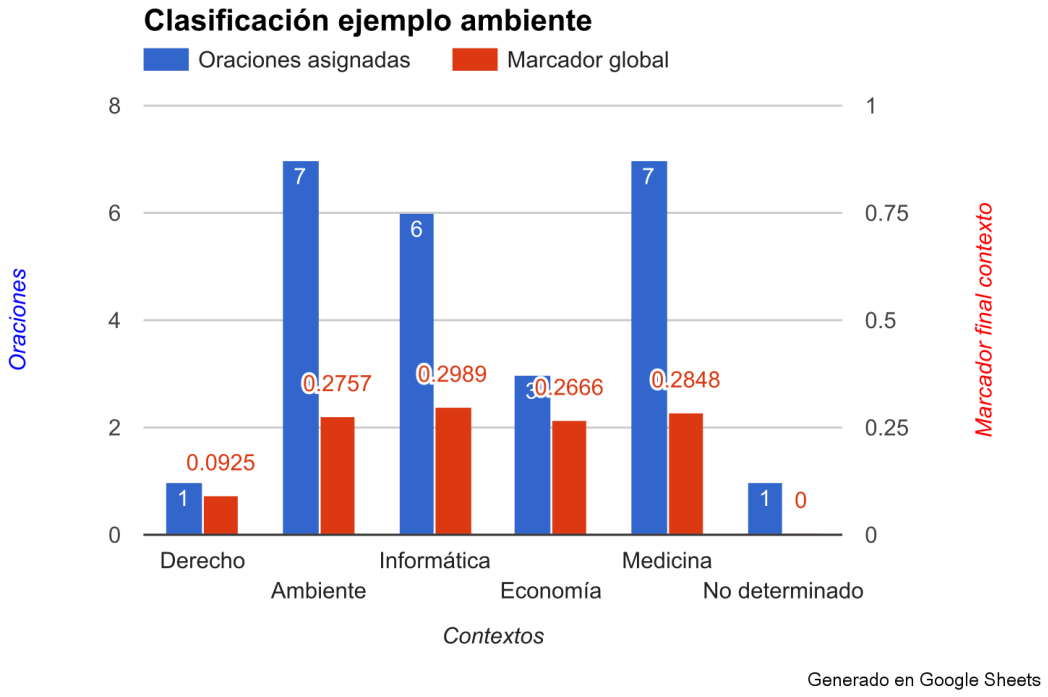

Figura 5.4: Clasificación para texto sobre el tema de *Ambiente*.

Para el experimento con el texto sobre *Ambiente*, se observa en la figura 5.4 que los casos donde se eligió *Ambiente* empataron con los del contexto *Medicina* (7 casos cada uno), uno más que el caso de *Informática*, que al final tuvo el mejor puntaje global (0.2989), todo lo anterior para un total de 25 frases detectadas por FreeLing.

Las figuras 5.5 y 5.6 muestran que para el ejemplo del texto de *Ambiente*, los picos de detección no superaron el 0.1, lo que se refleja en los marcadores acumulados que apenas pudieron superar el 0.25.

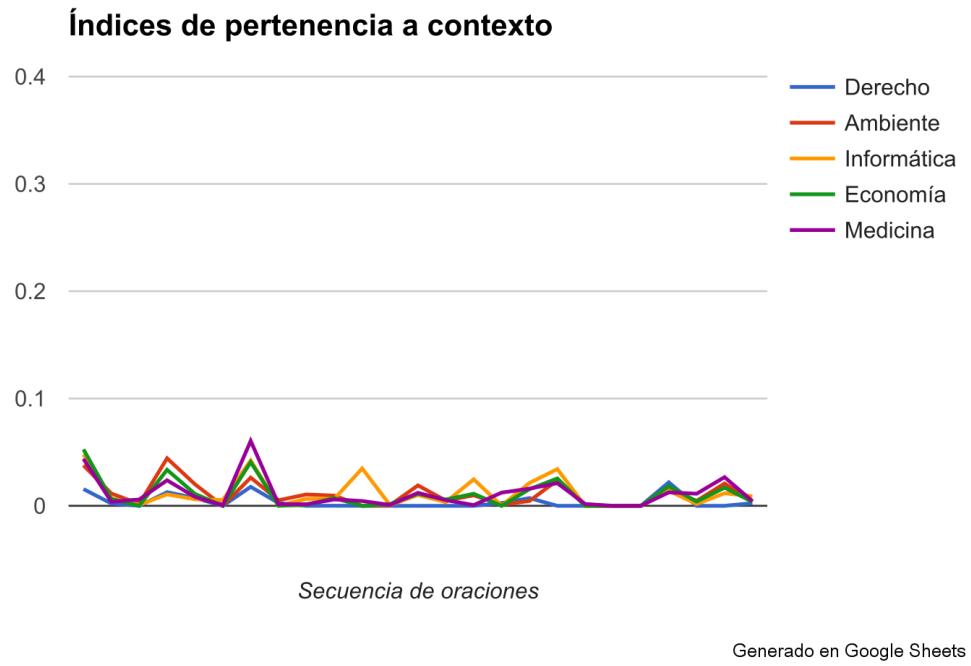

Figura 5.5: Comportamiento de gestores de contexto sobre el ejemplo de *Ambiente*.

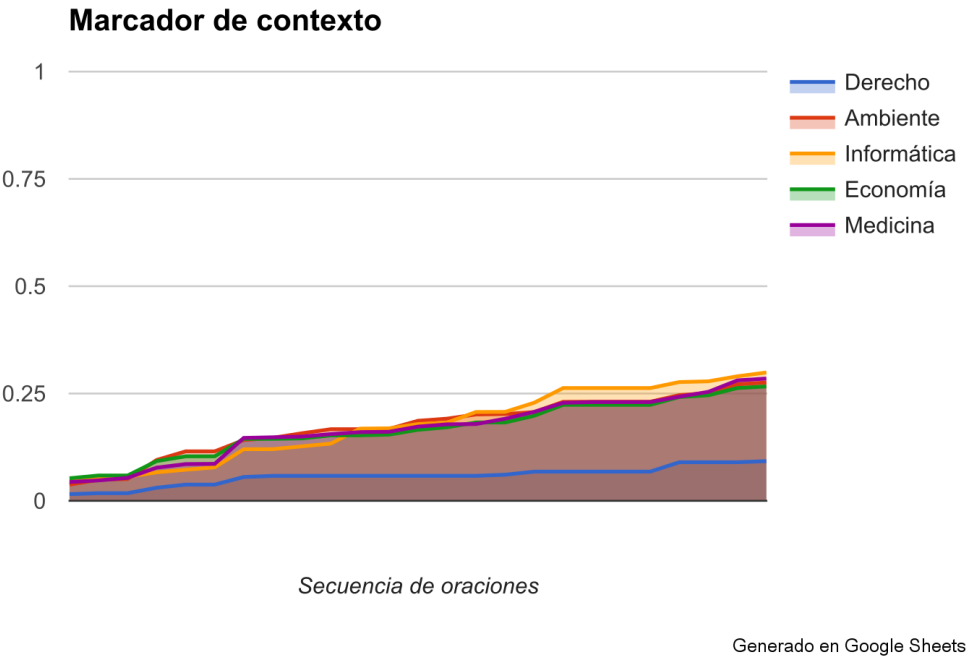

Figura 5.6: Marcador acumulado durante el análisis del ejemplo de *Ambiente*.

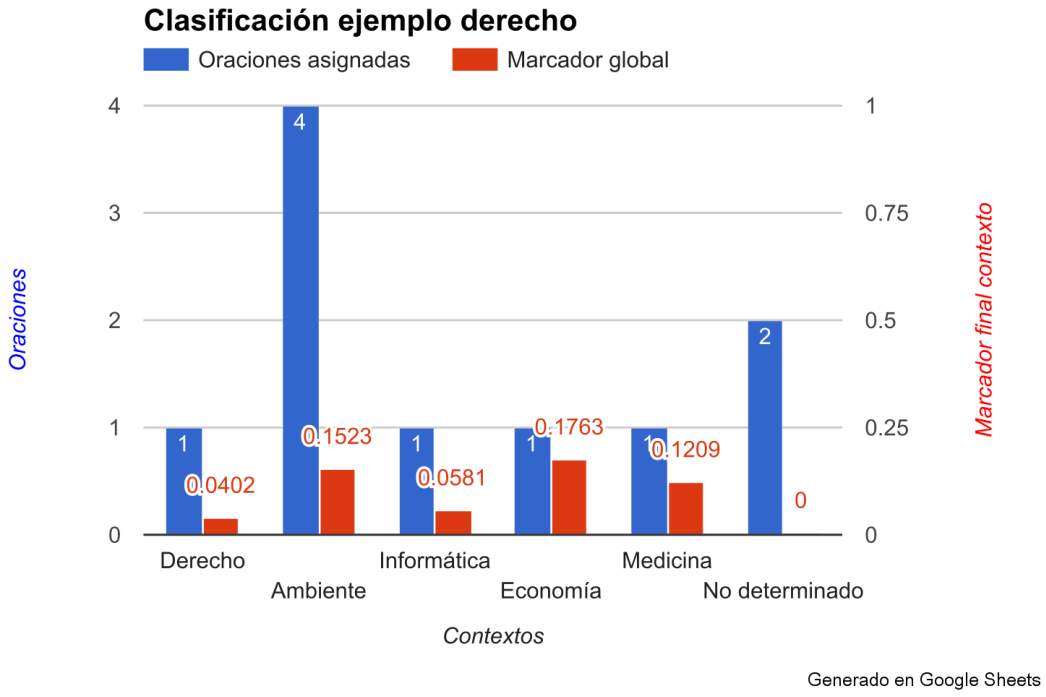

Figura 5.7: Clasificación para texto sobre el tema de *Derecho*.

Con 8 frases detectadas del ejemplo de *Derecho*, la figura 5.7 muestra la ventaja en la elección de frases como del dominio de *Ambiente*, llegando a haber 2 casos sin reconocer en contexto alguno y repartiendo una frase por cada uno de los restantes en los demás contextos. En el puntaje global, *Economía* logró un 0.1763 como el máximo puntaje acumulado del análisis, mientras *Derecho* fue el mínimo con 0.0402.

Revisando la figura 5.8, pocas frases superaron el 0.05 de índice de pertenencia, siendo el segundo caso superarior al 0.1 en el contexto de *Economía*. La frase es:

```
En nuestros tiempos, es innegable que la desaparición de las fronteras
nacionales en el desarrollo del comercio, requiere de normas uniformes
y de la armonización de su interpretación en los estados que, de una u
otra forma, quieran modernizarse e ingresar al selecto grupo de países
que han visualizado la importancia del derecho uniforme del comercio
internacional.
```

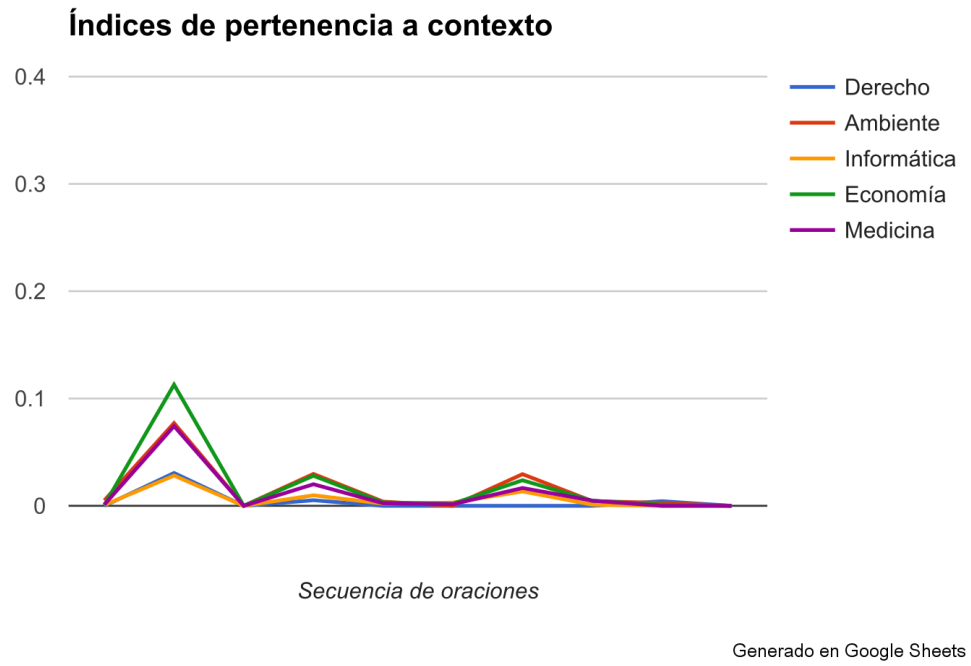

Figura 5.8: Comportamiento de gestores de contexto sobre el ejemplo de *Derecho*.

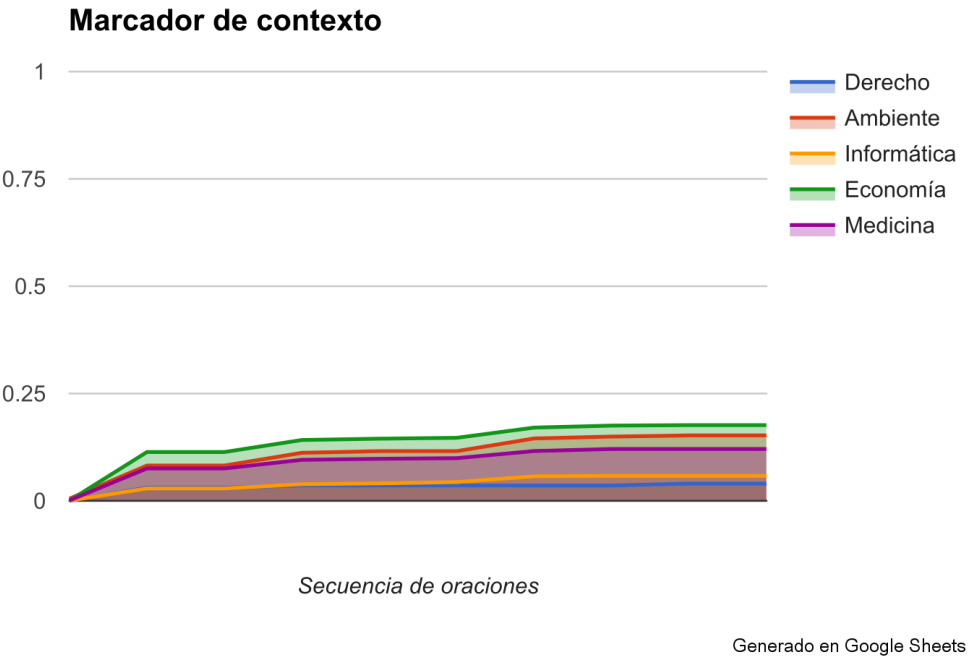

Figura 5.9: Marcador acumulado durante el análisis del ejemplo de *Derecho*.

Como consecuencia de los valores mínimos de los índices durante el análisis del ejemplo de *Derecho*, la figura 5.9 se nota más plana, no llegando ni a superar el 0.25 acumulado. El contexto *Economía* fue entonces el que tuvo el mayor valor acumulado durante el proceso.

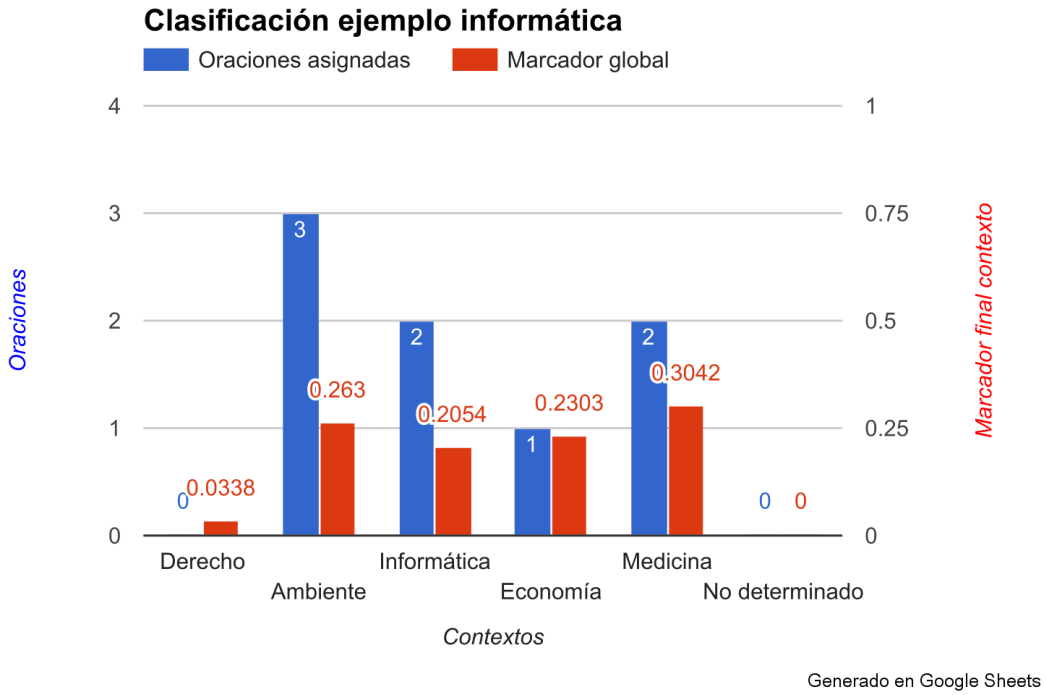

Figura 5.10: Comportamiento de gestores de contexto sobre el ejemplo de *Informática*.

El gráfico de la figura 5.10 presenta a *Ambiente* como el contexto con más frases seleccionadas (3), quedando *Informática* empatado con *Medicina* (2) de un total de 8 frases detectadas en el ejemplo. En el acumulado, *Medicina* logró el mayor puntaje (0.3042).

La figura 5.11 muestra que la frase 2 tuvo un gran impacto en 4 de los 5 contextos, llegando hasta el 0.2 en el contexto *Medicina*. La frase es:

```
El británico de 22 años, que prefiere mantener el anonimato, afirmó
a la BBC que "quizás no este fin de semana, pero con bastante
probabilidad el lunes por la mañana" comenzará un ataque similar.
```

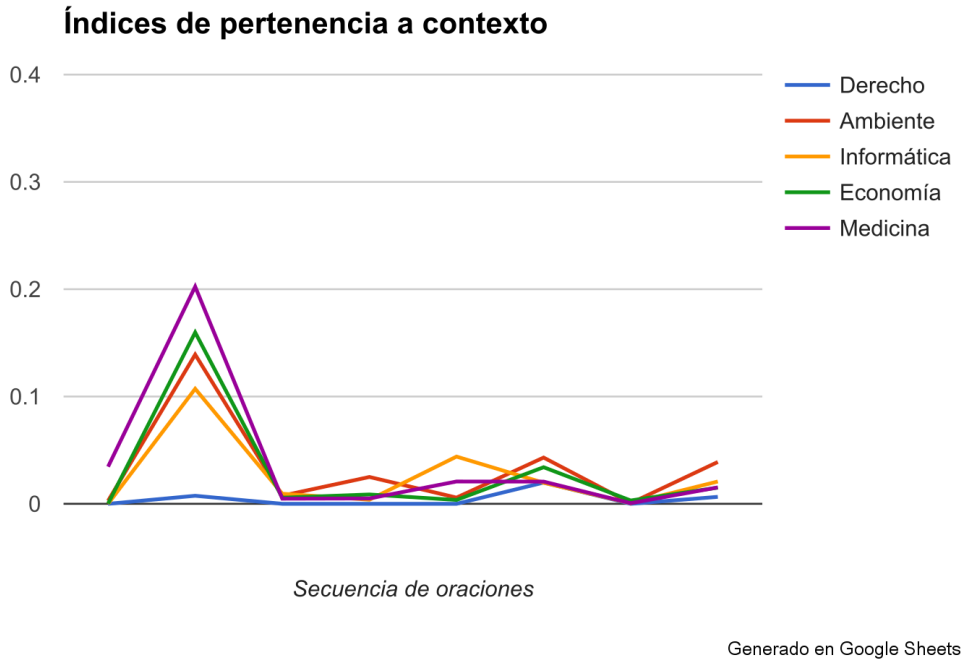

Figura 5.11: Clasificación para texto sobre el tema de *Informática*.

El contexto de *Informática* tuvo dos detecciones mayores (frase 2: 0.1072 y frase 5: 0.044).

En el marcador global, a partir de la frase 2, los 4 contextos aumentaron de forma casi lineal, siendo *Medicina* el que siempre supo ventaja y terminando con un marcador de 0.3042.

En el caso del ejemplo sobre *Medicina*, los resultados son altos comparados con los análisis previos. La figura 5.13 muestra que de los 18 casos de frases detectadas, *Informática* tuvo uno más que *Medicina* (6 vs 5), pero tuvo menor marcador global (0.8545 vs 0.8792). Ambiente y *Economía* también tuvieron un marcador global alto, superando el 0.7, aunque no tuvieron tantos casos seleccionados (3 y 2 respectivamente).

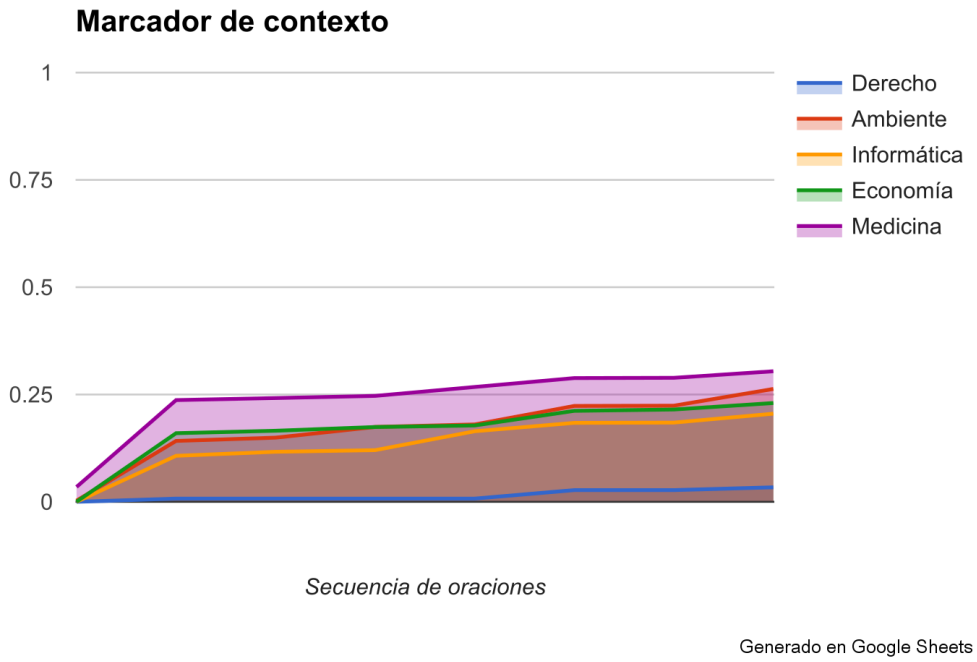

Figura 5.12: Marcador acumulado durante el análisis del ejemplo de *Informática*.

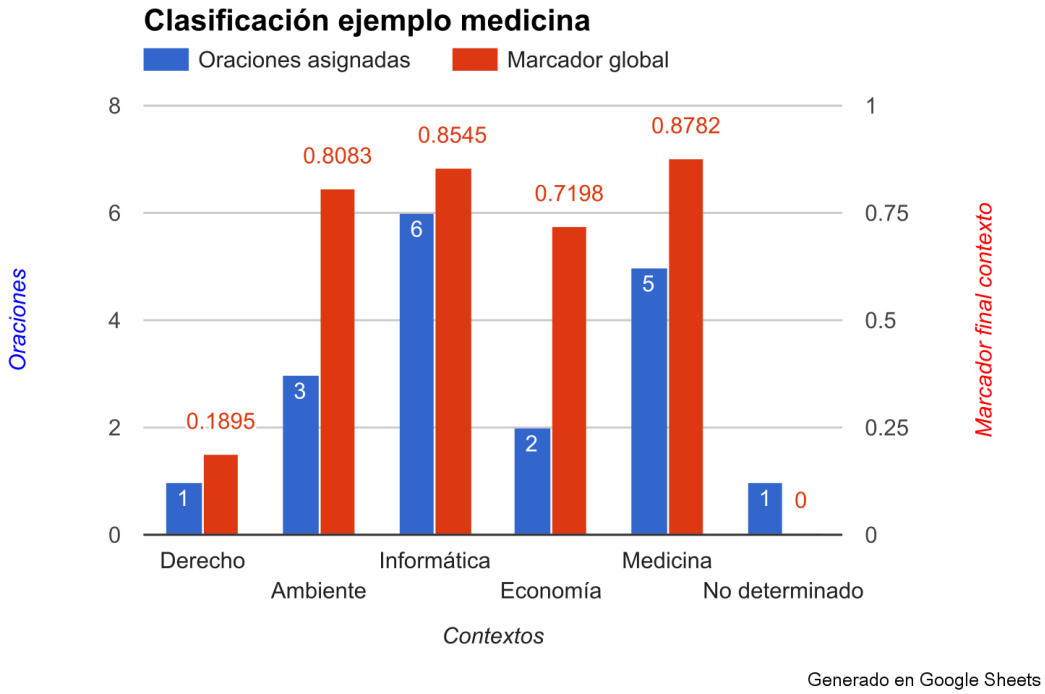

Figura 5.13: Clasificación para texto sobre el tema de *Medicina*.

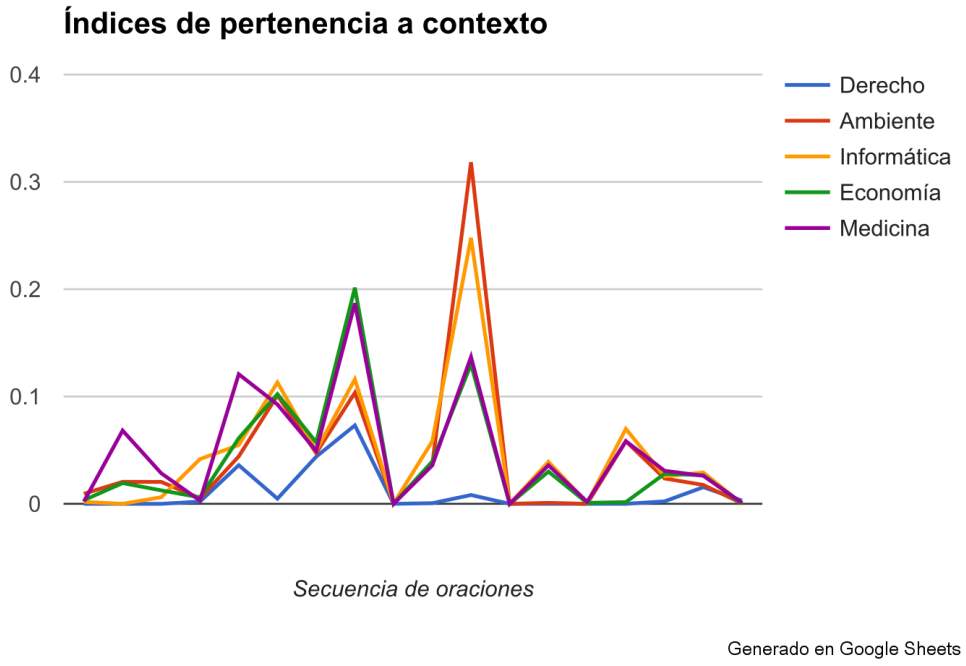

Figura 5.14: Comportamiento de gestores de contexto sobre el ejemplo de *Medicina*.

Los resultados de los gestores de contexto en la figura 5.14 refleja los altos resultados de pertenencia conseguidos respecto a los demás textos, muchos de los que tienen pico de *Medicina* entre el 0.05 y el 0.2. Se destaca la frase 11, donde *Ambiente* tuvo un pico de 0.3183 e *Informática* uno de 0.2479. La frase es:

```
Este polímero se podría utilizar para obtener nanofibras o nanopartículas.
```

La figura 5.15 denota que el acumulado de *Medicina* durante el proceso fue el más alto hasta el final, con un repunte de los contextos correspondientes en la frase 11 que no logra superarlo. Notar que el contexto *Derecho* no cambia mucho desde antes de la mitad del análisis hasta el final.

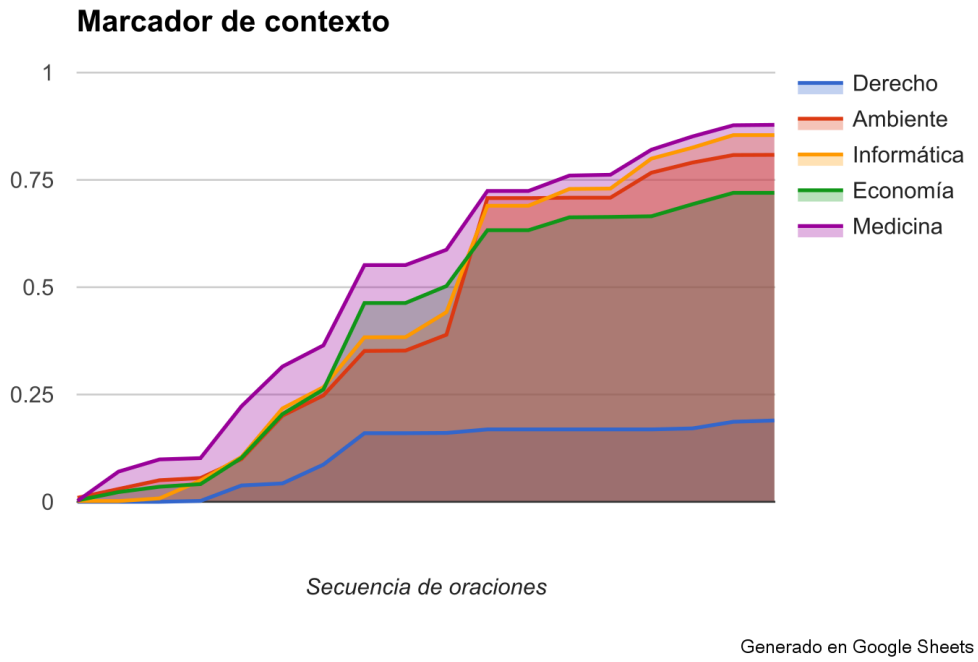

Figura 5.15: Marcador acumulado durante el análisis del ejemplo de *Medicina*.

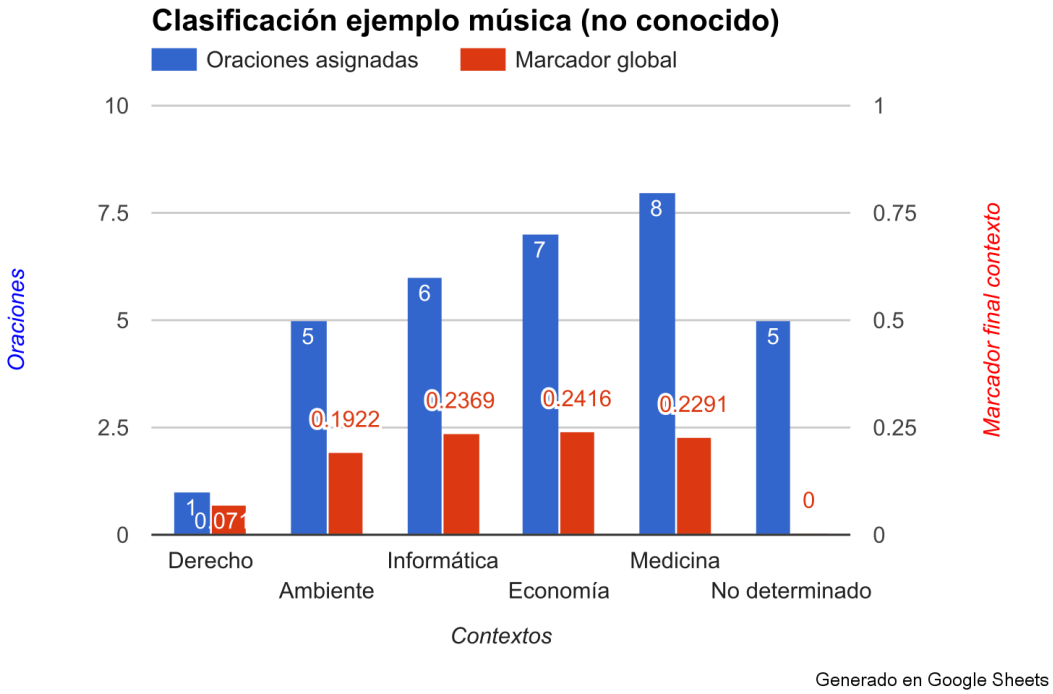

Figura 5.16: Clasificación para texto sobre el tema de *Música*.

La figura 5.16 muestra los resultados para un ejemplo adicional sobre un contexto desconocido para el módulo programado, en este caso, una nota sobre el contexto musical. En este caso, fueron 32 frases detectadas de las cuales 5 no se les asignó un contexto (0 en el índice de pertenencia en todos los contextos), mientras que el resto de frases se distribuyeron de una forma lineal al resto de contextos.

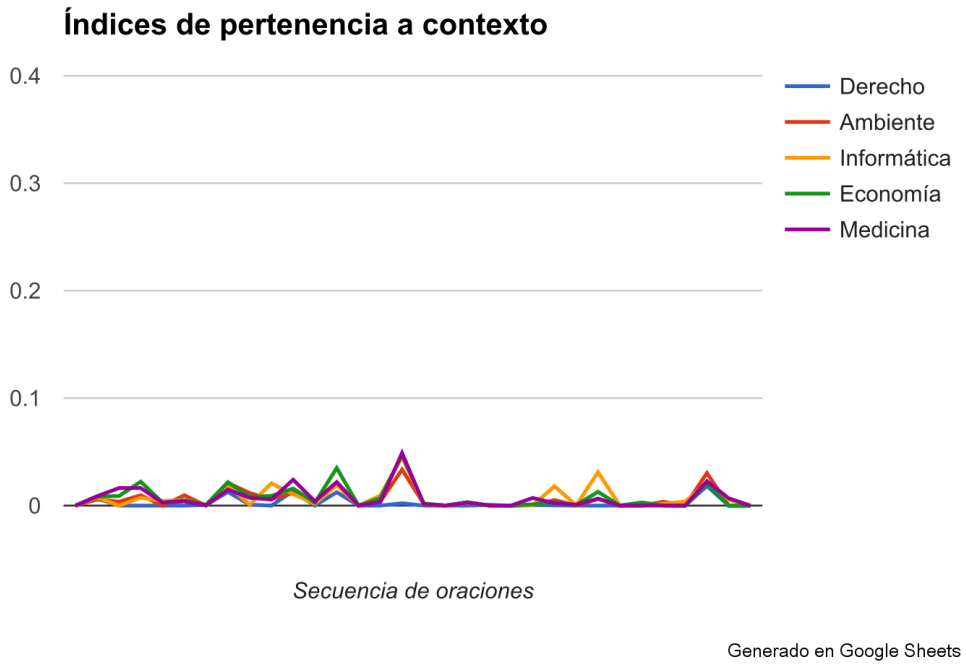

Figura 5.17: Comportamiento de gestores de contexto sobre el ejemplo de *Música*.

La elección de contextos se hizo con valores muy bajos de índices de pertenencia, ninguno superando el 0.05 de acuerdo a la figura 5.17.

En el marcador acumulado de la nota de contexto desconocido, el crecimiento de los puntajes fue muy lento durante el análisis y sin superar el 0.25, con una leve ventaja para el contexto *Economía* (0.2417), como se aprecia en la figura 5.18.

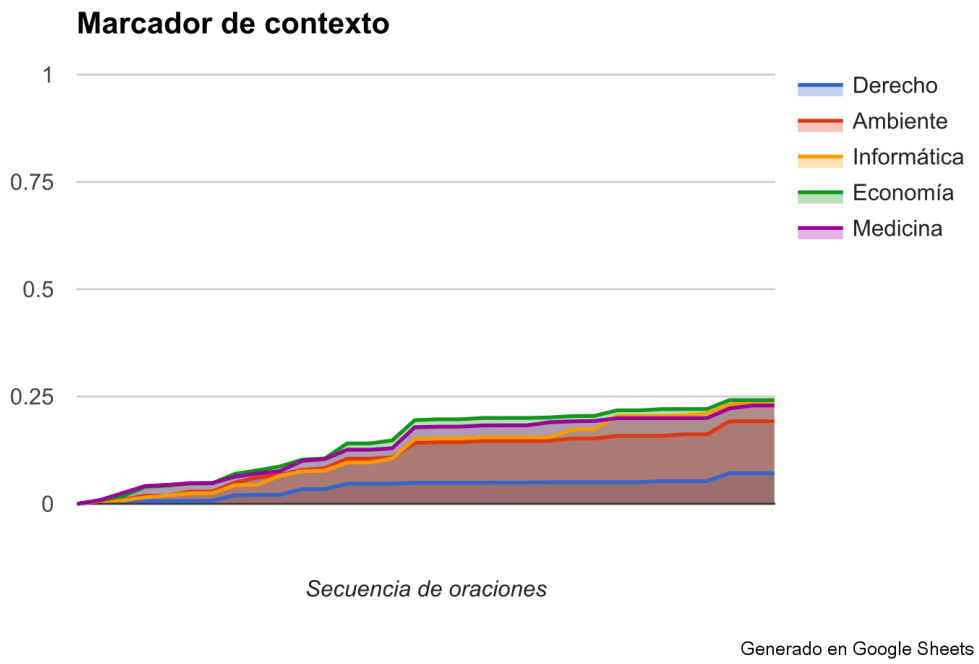

Figura 5.18: Marcador acumulado durante el análisis del ejemplo de *Música*.

5.1.2. Ejecución sobre oraciones individuales

Se realizaron pruebas del comportamiento del sistema analizando oraciones individuales (consultar apéndice C), con el fin de medir la cantidad de errores de tipo *falso positivo* ante diferentes contextos de oraciones cortas. La tabla 5.2 muestra los datos obtenidos luego de la ejecución del experimento.

Tabla 5.2: Resultado análisis de oraciones individuales

Oraciones	**Contexto acertado**			**Falsos positivos**		
	Casos	Índice promedio	Desviación estándar	Casos	Índice promedio	Desviación estándar
En el contexto	20/20	0.055750	0.028998	0/20	0	0
Ambiguas	12/20	0.020808	0.017180	8/20	0.008813	0.009893
Desconocidas	2/20*	0	0	18/20	0.010356	0.011747

El primer grupo de oraciones, tomadas directamente del corpus de cada contexto, se clasificaron correctamente en su totalidad, con un promedio de 0.05575 para el índice de pertenencia del gestor de contexto correspondiente. Este promedio tuvo una desviación estándar de 0.028998.

El segundo grupo, las oraciones ambiguas, clasificó de una forma ideal 12 casos, donde los restantes 8 fueron falsos positivos. El promedio del índice de pertenencia estuvo en 0.020808, con una desviación estándar de 0.01718, mientras los falsos positivos decidieron otro contexto con un promedio de 0.008813, con una desviación estándar de 0.009893.

En el tercer grupo con oraciones fuera de los contextos conocidos, ocurrieron 2 casos donde todos los gestores de domino dan 0 como índice de pertenencia, índice idealmente esperado, mientras que en los restantes casos (18) ganó un contexto con un índice promedio de 0.010356, con desviación estándar de 0.011747.

5.2. Análisis y discusión de resultados

Los índices de casos detectados y varios picos de detección corresponden con el índice promedio de oraciones propias del contexto, estando alrededor del 0.05, lo cuál evidencia una cota arrojada por el algoritmo de desambiguación propuesto.

Los corpus están orientados a un estilo de redacción que hace alusión a vocabulario de tecnológica y estadística. Lo que se encontró es que la redacción de los textos, particularmente los corpus de *Ambiente*, *Informática* y *Medicina*, incluyen propuestas de soluciones tecnológicas, con datos formales que involucran vocabulario de costos, tales como porcentajes, años y tecnicismos relacionados. Esto es interesante desde el punto de vista de la herramienta de desambiguación, que en este caso particular detecta un subconjunto temático dentro de estas tres áreas generales, altamente relacionado. Queda fuera del alcance de la investigación un análisis lingüístico sobre esto, pero el hecho de que se presente esta *ambigüedad temática* es un resultado esperado si se toma en cuenta el tamaño de los corpus utilizados. Una especialización de estos textos a subtemas debería resultar en detecciones más puntuales e implicaría un índice de pertenencia mayor por cada frase detectada.

Los dos textos informativos de mejores resultados fueron *Medicina* y *Economía*, donde en el primer contexto hubo picos de detección muy superiores al 0.05, mientras que en el marcador acumulado el contexto *Medicina* estuvo arriba en todo momento, a pesar de no obtener la mayoría de oraciones. En el caso de la nota sobre *Economía*, este contexto tuvo mayor cantidad de oraciones seleccionadas y al final obtuvo el mayor marcador acumulado, principalmente por el caso de la frase 11 compartido con *Ambiente*, superando por un margen muy superior a los demás contextos.

De los demás casos de textos informativos, y en parte el caso de *Economía*, los casos típicos de detección correcta estuvieron alrededor del 0.05. Esto se refleja en los gráficos de marcador acumulado, que cuando no se detecta un contexto con alguna diferencia significativa, no supera el 0.25. Estos datos son interesantes por el hecho que parece haber una cota para el marcador acumulado, posiblemente proporcional al número de contextos y el promedio de los falsos positivos de los picos de 0.01, evidenciados también en el caso del análisis de las oraciones individuales.

Varias frases que producen un efecto en todos los contextos simultáneamente contienen patrones gramaticales que coinciden en la redacción de varios de los textos de los corpus y que se repiten, como el uso de expresiones *"se puede utilizar para"* de la frase 11 del ejemplo de *Medicina*, *"afirmar a"*, *"pero con"* o la palabra *"probabilidad"*, son expresiones de la frase 2 del ejemplo de *Informática*.

En la tabla 5.2 notar el asterisco de la cuenta de casos tomados como falsos positivos para las oraciones de contexto desconocido. Esto es porque se incluyen casos cuyo índice de aceptación fue mayor a cero que podrían ser considerados no significativos si se tomara la cota de 0.05 como índice de aceptación. El caso ideal de que ningún contexto detecte nada es poco probable debido al vocabulario propio de la lengua, los vocablos compartidos entre contextos y el estilo de redacción de los corpus. Normalmente, estas detecciones están limitadas a un nodo del primer nivel de los grafos de contexto, con un porcentaje de pertenencia relativamente bajo y que en su mayoría cumplió con ser inferior a la cota del 0.05.

6 Conclusiones

Este capítulo presenta los principales hallazgos identificados a lo largo del proceso de investigación y validación de la propuesta.

Hallazgos derivados del análisis del contexto

Después de analizar la documentación relacionada con el proceso de traducción hacia una lengua de señas, se destaca una tendencia a iniciar las investigaciones utilizando el esquema clásico de traducción guiada por reglas. Se han hecho esfuerzos para realizar la traducción automática estocástica, pero la mayor limitación para abordar este enfoque ha sido la falta de corpus formales, muchas veces por el poco tiempo de estudio sobre la lengua de señas específica.

Tomando en cuenta el trabajo realizado en traducción automática y la identificación de contexto, los mejores resultados se centran en propuestas con temas especializados, lo que se puede aprovechar para abordar de forma incremental el problema de generación de nuevo vocabulario de las lenguas de señas jóvenes.

Hallazgos sobre la propuesta de solución

De las principales contribuciones en la construcción de la propuesta de solución a los problemas planteados, se han obtenido las siguientes conclusiones:

- La implementación de la arquitectura funciona y ofrece resultados esperados con corpus especializados validados lingüísticamente, ofreciendo la flexibilidad buscada para la generación de señas guiada por contextos y casos especiales de generación de la lengua de señas.

- En el algoritmo de detección de contexto propuesto, existe un umbral de acción alrededor del 0.05 de índice de pertenencia, que es superado por la mayor parte de las frases correctamente clasificadas. Debajo de este, suelen estar clasificaciones erróneas o no significativas si el contexto de la entrada de texto analizada tiene poco en común. Así mismo, existe un umbral de 0.25 para el marcador acumulado de los contextos, comportándose de forma análoga a los casos individuales. Este comportamiento ayuda a definir una cota mínima para la aceptación del resultado de un gestor de contexto.

- Para el algoritmo de detección propuesto, existen factores que permiten cambiar la calidad de las respuestas del sistema: tamaño de corpus, cantidad de contextos, extensión del texto a traducir y grado de ambigüedad temática del texto a traducir conocidos.

Hallazgos sobre las herramientas y recursos textuales utilizados

Los datos del corpus IULA utilizados para las pruebas se acoplaron al modelo sin problemas de formato de texto y se pudieron detectar frases de textos ajenos al corpus en su debido contexto. Los casos de frases en los que no se pudo escoger el contexto esperado tienen formas textuales compartidas entre los corpus que hacen alusión a lo estadístico y tecnológico.

Sobre la herramienta FreeLing 4.0, la función de exportación a formato JSON del análisis de dependencia presenta fallos y produce una estructura no válida. El problema se solucionó parcialmente con una función de preprocesamiento del texto analizado, pero aún existen casos por revisar donde el formato presenta más inconsistencias. Debido al enorme tamaño de los ejemplos donde ocurre esto (más de 100000 líneas), se opta por dejar estos casos afuera para posterior análisis e identificación del error y corregir en la misma función de preprocesamiento, o esperar por una versión corregida de FreeLing.

6.1. Resumen de la contribución al estado del arte

A continuación se citan de manera puntual los aportes realizados a partir de este trabajo de investigación de la arquitectura para traducción automática de lengua española a LESCO guiada por dominios de conocimiento distribuidos.

- Se demostró la factibilidad de un marco de trabajo flexible para el desarrollo iterativo de un traductor de lengua oral a lengua de señas guiado por contextos de conocimiento.
- Se ofreció un método de desambiguación determinista y configurable, compatible con la arquitectura de contextos distribuidos y que funciona como métrica de elección del contexto apropiado para un texto.
- Se midió la efectividad del método de desambiguación programado, brindando un análisis estadístico de los resultados para posteriores mejoras y comparaciones.
- Se identificó una serie de variables importantes aplicables al formato del conocimiento (corpus) de la arquitectura de contextos especializados y que deben ser evaluadas desde un punto de vista lingüístico: tamaño del corpus, cantidad de contextos, estilo de redacción.

6.2. Trabajos futuros

A partir de la investigación realizada se identificaron nuevas líneas de trabajo que ofrecen nichos sin explorar, o bien, que ofrecen oportunidades de mejorar en las diversas áreas de conocimiento involucradas en este proceso. En cuanto a las mejoras a la propuesta, se identificaron los siguientes elementos:

- **Usar resultados del marcador de contexto**: Se debe modificar la política de selección del gestor actual que tome en cuenta el valor del marcador global, para ello, hay que analizar dos posibilidades: tiempo real y final de análisis, cada caso tiene sus implicaciones. Por ejemplo, si una frase es muy significante para ese contexto (100 %), comparada a la mayor puntuación acumulada del mejor contexto del discurso, se elige esa opción y se cambia el gestor actual a el contexto ganador. El objetivo siempre será escoger el contexto más adecuado de la frase según el índice obtenido y el marcador de contextos.

- **Implementación de gestores de tipo lingüísticos**: La flexibilidad de la arquitectura de hilos programada deja la posibilidad de implementar otro tipo de gestores que detectan patrones de estructura importantes por su regularidad de generación en LESCO, por ejemplo: listas de términos, números, fechas y deletreo. Estos funcionarían como estados de encendido o apagado que añadirían una capa de análisis morfológico adicional a FreeLing y que tenga relación con la gramática de LESCO, permitiendo realizar casos muy específicos de señado que no se pueden detectar solamente dependiendo del análisis de la lengua española.

- **Implementación de nodos genéricos en el grafo contextual**: Existen casos especiales durante la construcción y recorrido de árboles de dependencia donde modificadores genéricos (números, orden cardinal, sistemas métricos) y sustantivos propios (nombres de persona o lugares) podrían ser factorizados en el grafo contextual si se determina que su tipo de palabra importa más que la instancia específica, por ejemplo, que en ciertas frases cierto nodo es un número y por la forma característica de la frase en la práctica no importe cual, por lo que debería contar para el marcador.

  Se plantea el uso de *nodos genéricos* que representen en sí el concepto general de una instancia específica, por ejemplo un nodo “{número}”. El nodo creado podría ser genérico en caso de que la palabra esté cumpliendo un rol sintáctico específico. Por ejemplo el caso de nombres propios, puede haber un nodo “{lugar}” si el análisis

sintáctico así lo postula, y se puede adjuntar al nodo una lista de frecuencia de términos que han pasado por esa construcción, de modo que si “Limón” es un lugar muy importante para el tema, se le den más puntos cuando se utilice en ese punto. Si el corpus es suficientemente representativo, “Limón” puede ser encontrado varias veces en el texto. Si es un caso aislado y puede ir perfectamente otro lugar, una sola aparición no representará un valor mayor para el significado de la frase total durante el análisis.

- **Fuentes de corpus**: Se plantea la necesidad de evaluar nuevas fuentes de corpus especializado, empezando con aquellos temas que tengan más vocabulario definido en LESCO y orientado a las prioridades de información de la comunidad sorda, como servicios públicos. También cabe la posibilidad de generar corpus y validarlos lingüísticamente según el proyecto lo requiera, con la ventaja de que pueden ser muy específicos y acelerar el proceso de obtención.

- **Mejoras de rendimiento**: Factorización de los datos usados en el algoritmo y el formato de salida del módulo de desambiguación son opciones a evaluar para acelerar los análisis de textos extensos. La omisión de nodos no significativos en el grafo contextual es una posibilidad que debe ser más estudiada debido a la posibilidad de aporte semántico que estos signos pueden dar para la generación de señas, pero que en muchos casos no son necesarios.

6.3. Posibles líneas de investigación

Aparte de las mejoras que se pueden aplicar al modelo propuesto, se identificaron otras áreas de interés para investigaciones futuras que podrían aportar conocimientos nuevos o oportunidades de mejoras en áreas relacionadas.

Generación gramatical en gestores de contexto

Este enfoque de implementación se basa en usar los demonios como transductores, es decir, incluir un proceso de generación directa, de modo que el módulo de generación de la arquitectura no exista como un ente sino que esté repartido entre los demonios. De esta forma, un demonio no sólo se encargaría de identificar el contexto, sino también de generarlo, y brindaría la ventaja de encapsular sus propias reglas gramaticales separadas de las demás, pudiendo realizar un sistema más escalable y portable.

Gestor de contexto como red neuronal

Dada la tendencia reciente de probar redes neuronales como sistemas de traducción, un sistema distribuido como el propuesto en este trabajo puede facilitar su implementación de forma más controlada para el programador, así como abrir diferentes alternativas de diseño de la solución.

Ontologías y jerarquía contextual

La naturaleza distribuida del procesamiento de los diversos contextos permite realizar jerarquías y dependencias de submódulos de reconocimiento de patrones que pueden ser explotadas para mejorar el proceso de reconocimiento y acelerar la generación de las oraciones con un alto grado de confianza. La teoría sobre ontologías puede aportar insumos valiosos para elaborar esta estructura.

Detección de similitudes lingüísticas entre distintos contextos

La arquitectura distribuida de los gestores de contexto puede ofrecer casos donde una frase compleja puede coincidir sintáctica y semánticamente en dos o más contextos. Esto se podría usar para descubrir coincidencias temáticas entre dos contextos en principio sin relación alguna, estableciendo analogías que pueden llevar a unificación de conocimiento. Por ejemplo, *“Marcar un gol.”* es una frase que puede estar en dos contextos de deporte,

fútbol y hockey, y la idea que engloba es la misma: anotar un punto en un partido del deporte correspondiente.

Variantes en algortimos y estructuras de datos existentes

Existen varias ideas que explorar derivadas de cambios específicos en los componentes del módulo desambiguador y que afectarían su comportamiento:

- Variantes en algoritmos de planificación de prioridades podrían servir para mantener actualizado la bolsa de pensamientos, de modo que se cargue con un factor alfa sobre los contextos más usados y provocar inanición para los contextos menos probables. El valor guardado en la bolsa de pensamientos puede ser útil si se implementa una política de mantenimiento de la bolsa que involucre algún criterio de rendimiento basado en prioridades. Esto se puede acompañar con un diseño para tiempo real del analizador, donde los demonios disparan su resultado tan pronto este alcance un umbral al analizar parte de la frase.

- Cambios y restricciones en la métrica del algoritmo de cálculo de porcentaje de pertenencia y el índice de pertenencia de frases. Por ejemplo, en un demonio debería de contar como 1 de probabilidad el hecho de tener la totalidad de términos de la frase analizada, independientemente de que sean pocos. Esto tiene lógica para una conversación donde se da un texto introductorio para contextualizar al receptor del mensaje, al inicio de la conversación, y con cada frase posterior, el contexto queda más claro. De este modo, solo en los casos que exista una palabra desconocida por el gestor de contexto y no genérica (fecha, número, lugar, nombre), el peso global de las palabras será el utilizado. De lo anterior, se debe definir un algoritmo apropiado para pesar el contexto del que se está hablando (en el marcador de puntuación contextos), de modo que sea posible corregir el sentido/traducción de las primeras frases del discurso analizado.

Referencias

AlSaidi, B. K. (2016). Automatic approach for word sense disambiguation using genetic algorithms. *(IJACSA) International Journal of Advanced Computer Science and Applications*, *7*(1), 41–44. Descargado de http://thesai.org/Publications/ViewPaper?Volume=7&Issue=1&Code=IJACSA&SerialNo=6 doi: 10.14569/IJACSA.2016.070106

Bahdanau, D., Cho, K., y Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. Descargado de http://cs224d.stanford.edu/papers/neural_machine.pdf

Baldassarri, S., y Royo-Santas, F. (2009). An automatic rule-based translation system to spanish sign language (lse). En A. J. Macías, A. Granollers Saltiveri, y M. P. Latorre (Eds.), *New trends on human–computer interaction: Research, development, new tools and methods* (pp. 1–11). London: Springer London. Descargado de http://dx.doi.org/10.1007/978-1-84882-352-5_1 doi: 10.1007/978-1-84882-352-5_1

Bangalore, S., y Rambow, O. (2000). Corpus-based lexical choice in natural language generation. En *Proceedings of the 38th annual meeting on association for computational linguistics* (pp. 464–471). Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dx.doi.org/10.3115/1075218.1075277 doi: 10.3115/1075218.1075277

Barrada, A. (2007, July). Intertextualidad y traducción: la alusión como elemento primordial en la traduccion de los textos literarios del árabe al español. *Revista Electrónica de Estudios Filológicos*(13). Descargado de https://www.um.es/tonosdigital/znum13/secciones/estudios_C_barrada.htm#_ftn10

Boitet, C. (2003, february). Automated translation. *Revue française de linguistique appliquée*, *8*, 99–121. Descargado de http://www.cairn.info/revue-francaise-de-linguistique-appliquee-2003-2-page-99.htm

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., ... Roossin, P. S. (1990, junio). A statistical approach to machine translation. *Comput. Linguist.*, *16*(2), 79–85. Descargado de http://dl.acm.org/citation.cfm?id=92858.92860

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., y Mercer, R. L. (1991). Word-sense disambiguation using statistical methods. En *Proceedings of the 29th annual meeting on association for computational linguistics* (pp. 264–270). Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dx.doi.org/10.3115/981344.981378 doi: 10.3115/981344.981378

Caliusco, M. L., y Stegmayer, G. (2010). Semantic web technologies and artificial neural networks for intelligent web knowledge source discovery. En Y. Badr, R. Chbeir, A. Abraham, y A.-E. Hassanien (Eds.), *Emergent web intelligence: Advanced semantic technologies* (pp. 17–36). London: Springer London. Descargado de http://dx.doi.org/10.1007/978-1-84996-077-9_2 doi: 10.1007/978-1-84996-077-9_2

CENAREC - Gramática - Proyecto de Descripción Básica de la LESCO. (s.f.). http://www.cenarec-lesco.org/index.php/gramar. (Fecha de acceso: Noviembre 2016)

Chen, P., Ding, W., Bowes, C., y Brown, D. (2009). A fully unsupervised word sense disambiguation method using dependency knowledge. En *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics* (pp. 28–36). Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dl.acm.org/citation.cfm?id=1620754.1620759

Edmonds, P., y Agirre, E. (2006). *Word Sense Disambiguation: Algorithms and applications*. Springer Verlag. Text, Speech and Language Technology Series. Descargado de http://artxiker.ccsd.cnrs.fr/artxibo-00080512

Eskridge, T., Hayes, P., y Hoffman, R. (2006, September 5-8). Formalizing the informal: a confluence of concept mapping and the semantic web. En A. J. Cañas

y J. D. Novak (Eds.), *Concept maps: Theory, methology, technology / second international conference on concept mapping (cmc2006)* (Vol. 1, pp. 247–254). San José, Costa Rica: Universidad de Costa Rica. Descargado de http://cmc.ihmc.us/cmc2006Papers/cmc2006-p199.pdf

Filhol, M., Hadjadj, M. N., y Testu, B. (2015). A rule triggering system for automatic text-to-sign translation. *Universal Access in the Information Society*, 1–12. Descargado de http://dx.doi.org/10.1007/s10209-015-0413-4 doi: 10.1007/s10209-015-0413-4

Fotinea, S.-E., Efthimiou, E., Caridakis, G., y Karpouzis, K. (2008). A knowledge-based sign synthesis architecture. *Universal Access in the Information Society*, *6*(4), 405–418. Descargado de http://dx.doi.org/10.1007/s10209-007-0094-8 doi: 10.1007/s10209-007-0094-8

Haque, R., Naskar, S. K., van den Bosch, A., y Way, A. (2011). Integrating source-language context into phrase-based statistical machine translation. *Machine Translation*, *25*(3), 239–285. Descargado de http://dx.doi.org/10.1007/s10590-011-9100-2 doi: 10.1007/s10590-011-9100-2

Hurtado Oliver, L. F., Costa, I., Segarra Soriano, E., García Granada, F., y Sanchis Arnal, E. (2016, september). Traducción automática usando conocimiento semántico en un dominio restringido. *Procesamiento del Lenguaje Natural*(57), 101–108. Descargado de http://hdl.handle.net/10045/57757

Ide, N., y Véronis, J. (1998, marzo). Introduction to the special issue on word sense disambiguation: The state of the art. *Comput. Linguist.*, *24*(1), 2–40. Descargado de http://dl.acm.org/citation.cfm?id=972719.972721

Jean, S., Cho, K., Memisevic, R., y Bengio, Y. (2015, July). On using very large target vocabulary for neural machine translation. En *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1–10). Beijing, China: Association for Computational Linguistics. Descargado de

http://www.aclweb.org/anthology/P15-1001

Jinghua, L., Baocai, Y., Lichun, W., Dehui, K., y Yufei, W. (2012, Nov). Diversified gesture generation for chinese sign language animation. En *Digital home (icdh), 2012 fourth international conference on* (p. 179-183). doi: 10.1109/ICDH.2012.44

Justeson, J. S., y Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, *1*(01), 9–27. Descargado de https://www.cambridge.org/core/journals/natural-language-engineering/article/technical-terminology-some-linguistic-properties-and-an-algorithm-for-identification-in-text/D5F076938C4E3F24B11EDC2E831216AF# doi: 10.1017/S1351324900000048

Koehn, P., Och, F. J., y Marcu, D. (2003). Statistical phrase-based translation. En *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology - volume 1* (pp. 48–54). Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dx.doi.org/10.3115/1073445.1073462 doi: 10.3115/1073445.1073462

Kumar, A., Irsoy, O., Su, J., Bradbury, J., English, R., Pierce, B., ... Socher, R. (2015). Ask me anything: Dynamic memory networks for natural language processing. *CoRR*, *abs/1506.07285*. Descargado de http://arxiv.org/abs/1506.07285

Langkilde, I., y Knight, K. (1998). Generation that exploits corpus-based statistical knowledge. En *Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics - volume 1* (pp. 704–710). Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dx.doi.org/10.3115/980845.980963 doi: 10.3115/980845.980963

Lezcano, L. A., Guzmán, J. A., y Vélez, C. A. (2015). Un modelo para la identificación de elementos kaos (especificación automática de adquisición de conocimientos) a partir de la especificación de requisitos en lenguaje natural. *Información tecnológica*, *26*, 129–138. Descargado de http://www.scielo.cl/scielo.php?script=sci

_arttext&pid=S0718-07642015000600015&nrm=iso

Liu, U., y Sun, H. (2015, november). Word sense disambiguation for chinese based on semantics calculation. *Mathematical Problems in Engineering*, *2015*(1), 1–6. Descargado de http://dx.doi.org/10.1155/2015/235096 doi: 10.1155/2015/235096

López-Colino, F., y Colás, J. (2012). Hybrid paradigm for spanish sign language synthesis. *Universal Access in the Information Society*, *11*(2), 151–168. Descargado de http://dx.doi.org/10.1007/s10209-011-0245-9 doi: 10.1007/s10209-011-0245-9

López-Ludeña, V., San-Segundo, R., Montero, J. M., Córdoba, R., Ferreiros, J., y Pardo, J. M. (2012, june). Automatic categorization for improving spanish into spanish sign language machine translation. *Computer Speech and Language*, *26*(3), 149–167. Descargado de http://dx.doi.org/10.1016/j.csl.2011.09.003 doi: 10.1016/j.csl.2011.09.003

Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., y Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*. Descargado de http://cs224d.stanford.edu/papers/addressing.pdf

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., y Khudanpur, S. (2010). Recurrent neural network based language model. En *Interspeech* (Vol. 2, p. 3). Descargado de http://www.fit.vutbr.cz/research/groups/speech/publi/2010/mikolov_interspeech2010_IS100722.pdf

Miyao, Y., Ninomiya, T., y Tsujii, J. (2005). Corpus-oriented grammar development for acquiring a head-driven phrase structure grammar from the penn treebank. En K.-Y. Su, J. Tsujii, J.-H. Lee, y O. Y. Kwong (Eds.), *Natural language processing – ijcnlp 2004: First international joint conference, hainan island, china, march 22-24, 2004, revised selected papers* (pp. 684–693). Berlin, Heidelberg: Springer Berlin Heidelberg. Descargado de http://dx.doi.org/10.1007/978-3-540-30211-7_72 doi: 10.1007/978-3-540-30211-7_72

Othman, A., y Hamdoun, R. (2013, Oct). Toward a new transcription model in xml for sign language processing based on gloss annotation system. En *Fourth international conference on information and communication technology and accessibility (icta)* (pp. 1–5). doi: 10.1109/ICTA.2013.6815317

Padró, L., y Stanilovsky, E. (2012, May). Freeling 3.0: Towards wider multilinguality. En *Proceedings of the language resources and evaluation conference (lrec 2012).* Istanbul, Turkey. Descargado de http://nlp.lsi.upc.edu/publications/papers/padro12.pdf

Papineni, K., Roukos, S., Ward, T., y Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318). Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dx.doi.org/10.3115/1073083.1073135 doi: 10.3115/1073083.1073135

Peng, Y. (2010). *Ontology mapping neural network: An approach to learning and inferring correspondences among ontologies* (Tesis Doctoral, School of Information Sciences). Descargado de http://d-scholarship.pitt.edu/6832/

Pennington, J., Socher, R., y Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, *12*. Descargado de http://nlp.stanford.edu/pubs/glove.pdf

Porta, J., López-Colino, F., Tejedor, J., y Colás, J. (2014, mayo). A rule-based translation from written spanish to spanish sign language glosses. *Comput. Speech Lang.*, *28*(3), 788–811. Descargado de http://dx.doi.org/10.1016/j.csl.2013.10.003 doi: 10.1016/j.csl.2013.10.003

Proyecto corpus. corpus textual especializado plurilingüe. (s.f.). https://www.iula.upf.edu/corpus/corpuses.htm. (Fecha de acceso: Mayo 2017)

Saggion, H., y Lapalme, G. (2000). Concept identification and presentation in the context of technical text summarization. En *Proceedings of the 2000 naacl-anlp workshop*

on automatic summarization (pp. 1–10). Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dl.acm.org/citation.cfm?id=1567564.1567565

Scruthi Sankar, K. P., Reghu Raj, P. C., y Jayan, V. (2016, december). Unsupervised approach to word sense disambiguation in malayalam. *Procedia Technology*(24), 1507–1513. Descargado de https://www.researchgate.net/publication/305110060_Unsupervised_Approach_to_Word_Sense_Disambiguation_in_Malayalam doi: 10.1016/j.protcy.2016.05.106

Seng Chan, Y., Tou Ng, H., y Chiang, D. (2007, june). Word sense disambiguation improves statistical machine translation. En Omnipress (Ed.), *Proceedings of the 45th annual meeting of the association for computational linguistics* (pp. 33–40). 2600 Anderson Street, Madison, WI 53704, USA: Omnipress.

Sierra, G. (2009, December). Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. *Linguamática*, *1*(2), 13–37. Descargado de http://linguamatica.com/index.php/linguamatica/article/view/38

Simón, A., Ceccaroni, L., Willmott, S., Rosete, A., Estrada, V., y Lara, V. (2006, September 5-8). Modelo unificado de representación de conocimiento en mapas conceptuales y ontologías. En A. J. Cañas y J. D. Novak (Eds.), *Concept maps: Theory, methology, technology / second international conference on concept mapping (cmc2006)* (Vol. 1, pp. 440–448). San José, Costa Rica: Universidad de Costa Rica. Descargado de http://cmc.ihmc.us/cmc2006Papers/cmc2006-p153.pdf

Socher, R., Bauer, J., Manning, C. D., y Ng, A. Y. (2013). Parsing with compositional vector grammars. En *In proceedings of the acl conference.* Descargado de http://nlp.stanford.edu/pubs/SocherBauerManningNg_ACL2013.pdf doi: 10.1.1.387.6840

Sutskever, I., Vinyals, O., y Le, Q. V. (2014). Sequence to sequence learning with neural networks. En *Advances in neural information processing systems* (pp. 3104–3112). Descargado de http://cs224d.stanford.edu/papers/seq2seq.pdf

Tmar, Z., Othman, A., y Jemni, M. (2013, March). A rule-based approach for building an artificial english-asl corpus. En *Electrical engineering and software applications (iceesa), 2013 international conference on* (pp. 1–4). doi: 10.1109/ICEESA.2013.6578458

Vandeghinste, V., Martens, S., Kotzé, G., Tiedemann, J., Van den Bogaert, J., De Smet, K., ... van Noord, G. (2013). Parse and corpus-based machine translation. En P. Spyns y J. Odijk (Eds.), *Essential speech and language technology for dutch: Results by the stevin-programme* (pp. 305–319). Berlin, Heidelberg: Springer Berlin Heidelberg. Descargado de http://dx.doi.org/10.1007/978-3-642-30910-6_17 doi: 10.1007/978-3-642-30910-6_17

Vargas Sierra, C. (2002). Utilización de los programas de concordancias en la traducción especializada. En C. E. S. de Traducción (Ed.), *El español, lengua de traducción [recurso electrónico] : Actas del i congreso internacional, almagro, 12-14/05/2002* (pp. 468–483). Descargado de http://hdl.handle.net/10045/13587

Wong, W., Liu, W., y Bennamoun, M. (2012, septiembre). Ontology learning from text: A look back and into the future. *ACM Comput. Surv.*, *44*(4), 20:1–20:36. Descargado de http://doi.acm.org/10.1145/2333112.2333115 doi: 10.1145/2333112.2333115

WordNet - About. (s.f.). https://wordnet.princeton.edu. (Fecha de acceso: Noviembre 2016)

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... Dean, J. (2016, sep). Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv e-prints*.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. En *Proceedings of the 33rd annual meeting on association for computational linguistics* (pp. 189–196). Stroudsburg, PA, USA: Association for Computational Linguistics. Descargado de http://dx.doi.org/10.3115/981658.981684 doi: 10.3115/981658.981684

Yuan, C., Wang, X., y Zhong, Z. (2015). Stochastic language generation using situated pcfgs. En *Proceedings of the 4th ccf conference on natural language processing and chinese computing - volume 9362* (pp. 64–75). New York, NY, USA: Springer-Verlag New York, Inc. Descargado de http://dx.doi.org/10.1007/978-3-319-25207-0_6 doi: 10.1007/978-3-319-25207-0_6

A Notas temáticas utilizadas en experimentos

Nota para contexto *Derecho*

Fuente:
http://www.elfinancierocr.com/economia-y-politica/Congreso-abogados-comercio_internacional_0_1166283374.html

Legales: Compromisos del derecho con el comercio global

Si usted posee una pyme, se encuentra vinculado al comercio internacional como abogado o empresario o, simplemente, quiere conocer el futuro del derecho de los contratos en Costa Rica, este artículo le interesa.
En nuestros tiempos, es innegable que la desaparición de las fronteras nacionales en el desarrollo del comercio, requiere de normas uniformes y de la armonización de su interpretación en los estados que, de una u otra forma, quieran modernizarse e ingresar al selecto grupo de países que han visualizado la importancia del derecho uniforme del comercio internacional.
Desde hace 50 años, la Comisión de las Naciones Unidas para el Derecho Mercantil Internacional (CNUDMI o Uncitral, por sus siglas en inglés), principal órgano de las Naciones Unidas en el ámbito del derecho mercantil, se ha decantado por dictar normas armónicas y sencillas, que pretenden enriquecer y facilitar la interacción y el desarrollo de las economías mundiales en temas ya tan cotidianos, como lo son el arbitraje, las garantías mobiliarias, el transporte internacional de mercaderías, el comercio electrónico y la compraventa internacional de mercaderías.
Uno de los instrumentos de su autoría que marcó un antes y un después en el comercio mundial, es la Convención de las Naciones Unidas sobre los contratos de compraventa internacional de mercaderías -conocida como la Convención de Viena de 1980 (CISG)-, la cual, regula el 80% de las ventas transfronterizas de todo el orbe.

Un paso para Costa Rica
Luego de muchos años, Costa Rica se puso una flor en el ojal y, el pasado 3 de marzo de 2017, publicó en La Gaceta la Ley No. 9421, que incorpora a nuestro derecho interno dicho tratado.
Este simple actuar nos coloca no solo como el país 86 en asumir como suya esta convención, sino que, su adopción traerá el fortalecimiento de las pymes y de los comerciantes que, en general, se dediquen a la importación y exportación de bienes.
El paso es especial, porque la mayoría de los socios comerciales de Costa Rica, entre ellos, Estados Unidos, Canadá, Colombia, México, China, Japón y 24 de los 28 países que integran la Unión Europea (UE), también la han adoptado en el pasado.
Ahora bien, la implementación de la CISG lleva aparejada una inmensa responsabilidad, sea, la de capacitar a abogados y, a personas que en general, se vean inmersas en el comercio internacional.

Más información
Por ese motivo, los días 9 y 10 de mayo de 2017, se estará celebrando en el Colegio de Abogados y Abogadas de Costa Rica el IV Congreso Internacional sobre Derecho Uniforme y Derecho del Comercio Electrónico, que será de índole interdisciplinario y en el que se abordarán los puntos más relevantes de la CISG.
Al evento, organizado en forma conjunta por la CNUDMI, Procomer, la Sala Primera del Poder judicial y el Colegio de Abogados, asistirán ponentes internacionales del más alto nivel, muchos de la Universidad Carlos III de Madrid, España, deseosos de trasladarnos sus conocimientos en esta materia tan ampliamente desarrollada en otras latitudes.

Nota para contexto *Ambiente*

Fuente:
http://cnnespanol.cnn.com/2017/05/15/estos-son-los-10-alimentos-que-mas-perjudican-al-medio-ambiente/

Estos son los 10 alimentos que más perjudican al medio ambiente

APÉNDICE A. NOTAS TEMÁTICAS UTILIZADAS EN EXPERIMENTOS

El Consejo de Defensa de los Recursos Naturales, una organización ambiental internacional sin ánimo de lucro, fundada en 1970, acaba de publicar un informe sobre comida y contaminación por calentamiento global en Estados Unidos, que incluye una lista con los 10 alimentos que se considera son los más perjudiciales para el clima, en relación con la cantidad de dióxido de carbono liberada por cada kilogramo producido de comida.
El estudio analizó 197 productos alimenticios -monitoreados por el Servicio de Investigación Económica del Departamento de Agricultura de Estados Unidos (USDA, por sus siglas en inglés)- y registró la contaminación por calentamiento global que acumularon entre el 2005 y el 2014.
1 de 10. La carne es ampliamente reconocida como el alimento más dañino para el clima. Un nuevo estudio sobre comida y contaminación por calentamiento global en Estados Unidos publicado por el Consejo de Defensa de los Recursos Naturales afirma que cada kilogramo de carne produce 26,5 kilogramos de emisiones de dióxido de carbono (CO2), la cantidad más alta de toda la investigación, que es cinco veces mayor a la del pollo y el pavo. La agricultura animal es responsable del 14,5% de las emisiones de gases de efecto invernadero en todo el mundo, según la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO). Mira a través de esta galería para ver más alimentos que son perjudiciales para el medio ambiente.
2 de 10. Otro rumiante, el cordero, está en el segundo lugar de la lista de los 10 alimentos que más daño le hacen al medio ambiente, lo que confirma que la carne roja es de las comidas que más recursos utilizan y, por ende, de las más nocivas para el clima. Por cada kilo consumido de cordero, se emiten 22,9 kilos de CO2 a la atmósfera, según el Consejo de Defensa de los Recursos Naturales.
3 de 10. Un poco más lejos, pero igualmente nociva para el medio ambiente, aparece la mantequilla: cada kilogramo de mantequilla equivale a 12 kilogramos de dióxido de carbono, casi la mitad de los que produce la carne de res. La mantequilla es el producto lácteo más perjudicial para el clima porque su preparación implica muchos pasos que consumen muchísima energía.
4 de 10. Cada kilo de mariscos le puede costar al medio ambiente 11,7 kilos

de CO2, según la investigación hecha en Estados Unidos por el Consejo de Defensa de los Recursos Naturales, una organización ambiental internacional sin ánimo de lucro, fundada en 1970. Según el Consejo, los estadounidenses han venido reduciendo el consumo de mariscos desde el 2005.

5 de 10. Otro producto lácteo, el queso, aparece en el quinto lugar de los 10 alimentos más dañinos para el medio ambiente, con 9,8 kilos de emisiones de CO2 por cada kilo producido. Y la cosa empeora cuando los quesos requieren refrigeración para ser transportados y llevados por vía aérea a otros países, lo que tiene impactos mayores para el clima.

6 de 10. En la lista de los 10 alimentos que más daño le hacen al clima hay un vegetal: el espárrago. Por cada kilo de espárrago que se produce se emiten 8,9 kilos de CO2. La razón está, sobre todo, en las millas aéreas. Casi todo el espárrago que se consume en Estados Unidos viene de América Latina, lo que significa que para llegar al plato de la mesa se han hecho gigantescas emisiones de gases de efecto invernadero.

7 de 10. Cada kilo de cerdo que se produce libera 7,9 kilos de dióxido de carbono, según el estudio del Consejo de Defensa de los Recursos Naturales, que analizó 197 productos alimenticios. Otra carne más en la lista.

8 de 10. Otro alimento que pertenece a la cadena de suministro del ganado vacuno y los lácteos: la ternera. Sin embargo, tiene un impacto menor que la carne de res, porque los terneros son sacrificados cuando son muy jóvenes, normalmente cuando tienen 20 semanas, frente a los 18 meses que en promedio se espera para matar una vaca. Cada kilo de ternera produce 7,8 kilos de CO2.

9 de 10. Aunque comer menos pollo es uno de los factores que más ha permitido reducir las emisiones per cápita de gases de efecto invernadero relacionadas con la comida en Estados Unidos, los productos avícolas siguen en el top 10 de los más perjudiciales para el clima, con alrededor de 5 kilogramos de CO2 por cada kilogramo producido de ese alimento.

10 de 10. Y cerrando la lista aparece el pavo, que tiene la misma huella de carbono que el pollo, con cerca de 5 kilos de CO2 por cada kilo producido.

Nota para contexto *Informática*

Fuente:
http://www.abc.es/tecnologia/informatica/abci-ciberataque-sin-precedentes-informatico-ayudo-desactivar-virus-advierte-otros-ataques-inminentes-201705141014_noticia.html

El informático que ayudó a desactivar el virus advierte de otros ataques inminentes

Un experto informático conocido como "MalwareTech", que ayudó a limitar el alcance del ciberataque global que afectó a cerca de cien países el viernes, ha alertado este domingo de que otros ataques similares podrían desencadenarse de manera inminente.

El británico de 22 años, que prefiere mantener el anonimato, afirmó a la BBC que "quizás no este fin de semana, pero con bastante probabilidad el lunes por la mañana" comenzará un ataque similar.

"Es muy importante que la gente proteja sus sistemas ahora", señaló el informático, después de que un software malicioso bloqueara más de 125.000 ordenadores el viernes en países como el Reino Unido, España, Francia y Rusia, según la cadena británica.

"MalwareTech" y expertos de la firma de seguridad Proofpoint desactivaron el virus al comprar un dominio de internet con el que el software trataba de comunicarse, lo que sirvió como un "interruptor" para detener la propagación de un "malware" que pedía un rescate económico para restaurar el sistema.

"Hemos detenido este, pero llegará otro y no podremos hacerlo. Hay mucho dinero en esto. No hay razón para que dejen de hacerlo. No cuesta mucho esfuerzo modificar el código y empezar de nuevo", explicó el británico.

Darien Huss, de la firma Proofpoint, coincidió en que "dada la enorme cobertura que está recibiendo este incidente" en los medios, "probablemente ya hay gente trabajando" para crear virus similares.

El ataque informático del pasado viernes ha afectado al menos a 200.000 víctimas de 150 países, según el último recuento elaborado por la agencia policial de la UE, Europol, que advierte que la cifra podría seguir aumentando este lunes, cuando se reanuden las actividades laborales en muchas empresas.

El director de Europol, Rob Wainwright, ha explicado este domingo en declaraciones a la cadena británica ITV que la particularidad de este ataque es que el software para pedir rescate (ransomware) fue utilizado en combinación con una «funcionalidad de gusano» para que la infección se extendiera automáticamente.
«El alcance global no tiene precedentes. El último recuento es de más de 200.000 víctimas en al menos 150 países y esas víctimas son en muchos casos empresas, algunas de ellas grandes corporaciones», ha explicado. «En este momento afrontamos una escalada de la amenaza. Las cifras crecen. Estoy preocupado por que las cifras puedan seguir aumentando cuando se vuelva al trabajo y enciendan sus máquinas en la mañana del lunes», ha apuntado.

Nota para contexto *Economía*

Fuente:
http://www.elmundo.es/economia/empresas/2017/05/15/5915926646163ffa7f8b4694.html

'Brexit' rico, 'Brexit' pobre

Cara y cruz del Brexit. Para unos ha sido un boom, para otros un apretón. Los mil más ricos del Reino Unido han visto crecer un 14% sus fortunas en el último año, hasta llegar a los 98.000 millones de euros. Mientras, millones de británicos se ajustan el cinturón ante la congelación de los salarios y la inflación, con una pérdida estimada de 600 euros por familia en 2017.
Es la doble realidad a la que se enfrentan los británicos, acuciada por la ruptura con la Unión Europea, que amenaza con acercar su economía a los parámetros de Singapur y con acentuar cada vez más la desigualdad económica.
Paradójicamente, el triunfo del Brexit -interpretado como un voto de protesta contra las élites- ha servido de momento para enriquecer a los millonarios como Arron Banks (promotor de la campaña Leave.eu) o Sir James Dyson, el inventor/inversor que ha trepado hasta el número 14 en lista de los más ricos de The Sunday Times.
Al menos, 28 de los 100 más ricos del Reino Unido son generosos donantes del Partido Conservador de Theresa May. Entre las 20 mayores fortunas británicas amasan el equivalente a 227.000 millones de euros, más que los

presupuestos combinados de Educación y Salud. La sanidad pública hace, entre tanto, aguas: las escuelas se enfrentan a recortes y los precios de los alimentos y de las necesidades básicas se disparan.

«Mientras la mayoría de nosotros nos preocupábamos por el resultado del referéndum, los más ricos del Reino Unido mantuvieron la calma y siguieron haciendo miles de millones», asegura Rober Watts, editor de la celebérrima Rich List de The Sunday Times. La bonanza de los mercados ha beneficiado a los más privilegiados, que han visto revalorizadas sus acciones, mientras que otros han sacado partido a la caída de la libra.

La cara oculta de la moneda ha sido, sin embargo, la inflación, que ha saltado al 2,3% y podría llegar al 2,7% a fin de año, según un estudio del Center for Economics and Business Research (CEBR). La subida de los salarios se estima en el 2,2%, de modo que el pistoletazo de salida del Brexit se traducirá en una disminución del poder adquisitivo de las familias (de 35.300 a 34.800 libras como promedio).

«Este año va suponer un severo ajuste para los británicos», advierte Nina Skero, analista del CEBR. «La inflación seguirá aumentando, y se va a notar, sobre todo, en el precio de los alimentos, del transporte y de la vivienda. A pesar del bajo nivel de desempleo (4,7%), el aumento de los salarios va a ser débil. El índice de confianza de los consumidores está ya por debajo de los niveles anteriores al Brexit: la gente está notando que los precios suben y son más cautos a la hora de gastar. Esto es un gran reto para nuestra economía, cuyo crecimiento depende del gasto de los consumidores».

Las alertas han saltado en el Banco de Inglaterra, que esta misma semana ha revisado a la baja la perspectiva del crecimiento de la economía británica, del 2% al 1,9%. «Se avecinan tiempos duros para los hogares», ha advertido el gobernador Mark Carney. «Los salarios no podrán crecer en la misma medida que los precios».

La economía británica amenaza con adquirir definitivamente una doble velocidad. Londres sigue siendo la capital mundial de las grandes fortunas, con 86 ilustres habitantes por encima del listón de los mil millones de libras. Los hoteles de cinco estrellas están llenos (y más desde la caída de la libra) y Oxford Street es una pasarela diaria de millonarios árabes y chinos. Pero, la caída del poder adquisitivo de la clase media está ya

haciendo mella en el mercado inmobiliario.

«Los ricos siguen despegándose del resto de nosotros, mientras los más pobres ven como sus bolsillos encogen», recalca por su parte Wanda Wyporska, directora ejecutiva de la fundación The Equality Trust. «Esta es una economía que sigue funcionando para muy pocos y no para la mayoría... Con las elecciones a la vista, los políticos deberían decidir qué tipo de país quieren construir: uno en el que la prosperidad funcione para todos o uno en el que todos acabemos recogiendo las migajas de los ricos».

Con los 98.000 millones de euros de ganancias de los mil multimillonarios se podrían haber pagado las facturas de la luz de todo el país durante dos años y medio, según sus estimaciones. El mismo dinero habría servido para dar de comer durante 56 años al millón de británicos que dependen de los bancos de alimentos. También habría servido para erradicar, a corto plazo, la pobreza que afecta al 30% de la población infantil (cuatro millones de niños).

La creciente desigualdad económica se está convirtiendo en la patata caliente de la campaña electoral. Para la mayoría, no para unos pocos, es el lema con el que el líder de la oposición laborista, Jeremy Corbyn, se ha lanzado a la campaña, anunciando la mayor intervención estatal del último medio siglo, para dar el volantazo a la economía «y transformar las vidas de millones de británicos».

Nick Clegg, portavoz del Partido Liberal Demócrata para asuntos europeos, se ha lanzado al ruedo con el apretón del Brexit como bandera. «La salida de la UE está costando 500 libras por familia y haciendo daño a la gente más vulnerable», declaró Clegg. «La gran pregunta es: '¿Le haremos pagar a Theresa May por el daño que va a causar al Reino Unido?'».

«La devaluación significa más inflación», recalcó el ex viceprimer ministro. «La mitad de los alimentos en nuestros supermercados son importados, y se estima que el aumento será del 3% por la caída de la libra. Los precios han subido ya en la factura de la luz, ropa, vino, y en los aparatos electrónicos. Incluso en el té. Irse de vacaciones a España es un 17% más caro».

Clegg destacó cómo el apretón, que están notando ya las empresas, está poniendo también en una situación crítica a las arcas del Estado. «En un momento en que necesitamos desesperadamente dinero para los servicios

públicos, el Gobierno está destinando millones al agujero negro del Brexit», advirtió. «En los próximos cinco años, la salida de la UE podrá haberse llevado 59.000 millones de libras (70.000 millones de euros) que podrían haber servido para paliar las necesidades de la sanidad pública, de la asistencia social y de nuestras escuelas».

Los sindicatos han puesto también el dedo en llaga. «Los trabajadores están notando la doble presión de los precios que suben y el estancamiento de los salarios», señala Frances O'Grady, secretaria general del Trade Union Congress (TUC). «Pero May parece conformarse ante el hecho de que avanzamos hacia una crisis del nivel de vida. El Gobierno no puede tardar más en pasar a la acción». Finalmente, ha prometido intervenir el precio de la electricidad y el gas para evitar subidas abusivas (la factura de la luz ha crecido un 158% en los últimos 15 años). La medida, robada al ex líder laborista Ed Miliband, ha sido calificada como necesaria pero insuficiente.

«Hay claras señales de que las familias están empezando a sufrir por el apretón de los ingresos reales y esa sensación se va a intensificar», reconoce Chris Hare, analista de Investec. «El gasto de los hogares británicos supone aproximadamente dos tercios de la demanda en el Reino Unido, y ésa es la principal razón por la que la economía se ha ralentizado este año».

«Los británicos pueden sufrir hasta 15 años de pérdida del poder adquisitivo», advierte en The Guardian el director del Instituto de Estudios Ficales (IFS), Paul Johnson. «Nuestras estimaciones proyectan unos salarios medios en 2022 no superiores a los de 2007. Esto es algo sin precedentes».

Otro estudio, auspiciado por la Resolution Foundation, va aún más allá y asegura que estamos «en la peor década para el crecimiento de los salarios desde las guerras napoleónicas (1803-1815)», con una pérdida real de los salarios de 14.500 euros entre 2010 y 2020. Según este estudio, el Brexit va a servir como excusa para perpetuar las políticas de austeridad hasta bien entrada la década de 2020.

Nota para contexto *Medicina*

Fuente:
http://diarioelzondasj.com.ar/politica/cientificos-sanjuaninos-aplican-la-nanotecnologia-a-la-medicina

Científicos sanjuaninos aplican la nanotecnología a la medicina

El Instituto de Investigaciones en Ciencias Químicas de la Universidad Católica de Cuyo, bajo la dirección del Dr. en Ciencias Químicas, Diego Kassuha, está trabajando en nanotecnología aplicada a la medicina con importantes proyectos. Con la llegada de un moderno equipamiento a esta casa de altos estudios, ya los científicos pudieron avanzar en sus investigaciones que pretenden generar partículas a una escala nano con la posibilidad de ser utilizadas en el ámbito de la medicina y de la farmacoterapia.

"Las nanopartículas son interesantes porque se pueden aplicar en varios campos. En Farmacia, en la regeneración de tejidos, en el ambiente o alimentos, en la química inorgánica con muchas aplicaciones. Es importante que se puedan manipular estas partículas pequeñas porque tienen acceso a muchos lugares donde normalmente una partícula macroscópica o de tamaño micro no podía acceder", explicó la Dra. en ingeniería química y biomolecular, Sandra Noriega.

Uno de los proyectos en los cuales se trabaja en la Universidad Católica está referido a la industria farmacéutica. "En farmacia uno toma un fármaco y se distribuye en todo el cuerpo pero con una nanopartícula tendría la capacidad de poder liberarla o llevarla al lugar donde uno la necesite", explicó Diego Kassuha.

Con el uso de las nanopartículas "uno puede hacer que el medicamento se libere más rápido o más lento o a un pH determinado, o se podría colocar un anticuerpo y teledirigirlo a un tumor o zona donde uno quiere que llegue la particular y no a otros tejidos, evitando así los efectos adversos en otros tejidos u órganos".

Según explicó el doctor Kassuha, una de las tesistas del Instituto y becaria doctoral del CONICET, Virna Martín Giménez trabaja en desarrollar un fármaco

para disminuir la hipertensión arterial renal sin provocar efectos adversos en otros órganos. "El fármaco que se utiliza tiene efectos adversos a nivel del sistema nervioso y su idea es obtener nanopartículas que lo liberen a nivel renal y no pase a la sangre ni al sistema nervioso. Mi tema de investigación es mejorar propiedades desfavorables de fármacos como la escasa solubilidad, estas particulas nano son mucho mas pequeñas, tienen mayor superficie de contacto con los solventes y se disuelven más fácil y aumentan la solubilidad y la chance de llegar a distribuirse, en ese caso es para mejorar la eficacia de los fármacos", explicó el director del Instituto.
Por su parte la doctora Sandra Noriega detalló las características del proyecto en el que ella trabaja."Estamos trabajando en un proyecto para producir un polímero biodegradable a partir de residuos de la industria agrícola de la región, a partir del escobajo de uva o de residuos de la industria olivícola como el alperujo. Este polímero se podría utilizar para obtener nanofibras o nanopartículas. Mi interés es obtenerlo en fibras para hacer aplicación en regeneración de tejidos en un futuro para humanos y trasplantes", indicó.
"Esto es muy interesante porque en lugar de usar un polímero comercial empleando técnicas que no son muy amigables con el ambiente, se soluciona un problema que es el que generan los residuos de la industria y por otro lado se genera un producto que se puede utilizar en aplicaciones biomédicas", añadió Kassuha.
"La aplicación inmediata de este polímero es pensando en envases para la industria alimenticia. Un polímero biodegradable se degrada con el tiempo y en este momento los plásticos son un problema grave y por eso estamos tratando de resolver algo de este problema", añadió Noriega.
Según informó el doctor Kassuha no hay muchos equipos del tipo que adquirió la Universidad Católica de Cuyo, para utilizar mediante el método electrospinning. "El equipo es bastante simple: tiene una fuente de alto voltaje y una bomba jeringa que puede mover fluidos a bajo caudal. Uno coloca a la jeringa la solución con el polímero y esa bomba va a propulsar el líquido a través de la jeringa. Cuando sale la gota en lugar de tener la forma de gota se deforma y genera un cono que rompe en un jet y que puede generar nanopartículas o nanofibras que impactan contra una superficie

metálica donde colectamos las partículas.
De acuerdo con los científicos, el equipo permite ajustar el voltaje y así, de acuerdo a la concentración que hay del polímero en la solución, la distancia a la placa metálica colectora, el tipo de polímero utilizado y la viscosidad de la solución generada, el resultado serán fibras o partículas siempre en la escala nano.
El equipo de profesionales y científicos que desarrolla proyectos en el Instituto de Investigación en Ciencias Químicas en el Laboratorio de Control de Calidad, Dr. Alberto Graffigna de la UCCuyo, está integrado por el doctor en Ciencias Químicas Diego Kassuha y la Doctora en ingeniería Química y biomolecular Sandra Noriega, además por el bioquímico Gerardo Castro y la ingeniera Cecilia Bustos. También participan la Farmacéutica y becaria de CONICET, Virna Martín Giménez y la becaria del CONICET Carla Groff, además de la alumna Mariana Albarracín.

Nota para contexto desconocido (*Música*)

Fuente:
http://cultura.elpais.com/cultura/2017/05/14/actualidad/1494752888_070782.html

El último genio toca el clave

Londres desempeñó un papel capital en la revolución interpretativa de la música antigua, hasta el punto de prestar su nombre a varios grupos pioneros, como The Early Music Consort of London, la visionaria creación de David Munrow, o Pro Cantione Antiqua of London, que holló tantos caminos hasta entonces inexplorados de la mano de Bruno Turner, ambos fundados a finales de los años sesenta. Aquellos músicos inconformistas y amantes de la experimentación crearon a su vez un público ávido de escuchar el repertorio medieval, renacentista y barroca sin incómodas adherencias anacrónicas y con una generosa amplitud de miras.
A mediados de los ochenta, estas interpretaciones con una conciencia histórica, como se decía entonces, gozaban ya de una excelente salud y fue entonces cuando, también en Londres, nació un festival de música barroca bautizado con el nombre de Lufthansa, la compañía aérea alemana que creyó

en las bondades y el potencial renovador de aquella revolución y que se animó a financiar aquella iniciativa de un entonces jovencísimo Ivor Bolton, el actual director musical del Teatro Real, y de la musicóloga Tess Knighton. Hace tres años que Lufthansa se desvinculó del festival, pero el empeño ha seguido adelante, ya con Lindsay Kemp como director artístico, con el nombre mucho más neutral de Festival de Música Barroca de Londres. En la presente edición recuerda especialmente a Claudio Monteverdi, nacido en 1567, y a Georg Philipp Telemann, fallecido exactamente dos siglos después. Ambos representan, para Kemp, ambas orillas del Barroco, de ahí el título de Baroque at the Edge que encabeza carteles y programas. La idea admite, claro, otras lecturas, otros sesgos, observarse desde otros ángulos, y la segunda jornada del festival ha ejemplificado al menos uno de ellos a las mil maravillas.

El primero de los tres conciertos del sábado fue un recital de música francesa protagonizado por Jean Rondeau, un jovencísimo clavecinista francés que triunfó en 2012 con tan solo veintiún años en el concurso internacional más famoso de cuantos se convocan para su instrumento, el de Brujas, y que desde entonces no ha dejado de asombrar no ya solo por su virtuosismo, sino fundamentalmente por su madurez y, cuando llega el momento para ello, por su iconoclastia. No fue aún el caso de este primer recital, celebrado en la iglesia de St Peter, en Eaton Square, con piezas de Jean-Philippe Rameau Joseph-Nicolas-Pancrace Royer.

El enhiesto peinado de Rondeau de hace unos años, que parecía electrificado y apuntando hacia el cielo, ha dado ahora paso a una melena lacia y una larguísima barba que le presta el aspecto de un eremita recién salido del desierto de la Tebaida. Cuando se pone a tocar, sin embargo, la heterodoxia desaparece y escuchamos versiones canónicas, pero geniales, del inconfundible repertorio barroco francés para clave, pródigo en esas miniaturas de títulos fragantes y no siempre fácilmente comprensibles: La Majestueuse, L'Incertaine, La Remouleuse, Les Tendres Sentiments, La Sensible o Le Vertigo, por citar solo algunos de los de Royer, mucho menos conocidos que los de Rameau. Jean Rondeau parecía predestinado por su apellido a tocar justamente esta música, ya que varias de estas piezas son, precisamente, rondeaux.

APÉNDICE A. NOTAS TEMÁTICAS UTILIZADAS EN EXPERIMENTOS

Hacía mucho tiempo que no surgía un clavecinista con la pulsación -nítida y delicada- y la fantasía -desbordante e irresistible- de este joven talento francés, capaz de convertir cada una de estas miniaturas en un mundo perfectamente cerrado sobre sí mismo. Impresiona especialmente, claro, su virtuosismo en las piezas plagadas de exigencias técnicas, como la Gavotte avec les doubles de la Gavotte de Rameau o La Marche des Scythes de Royer, ambas con auténticos vendavales de notas que Rondeau traduce con una insólita transparencia y precisión. Pero también en las piezas breves (el Prélude de Rameau que abrió el recital o La Sensible de Royer) se percibe a un músico con mayúsculas, en la estela de sus más grandes predecesores, con Gustav Leonhardt a la cabeza. Un momento culminante de su concierto fue la interpretación de L'Enharmonique, sobre cuyas audacias armónicas previno el propio Rameau al decir que "puede que no sean de inmediato del gusto de todo el mundo. Sin embargo (...) es posible llegar a apreciar toda su belleza una vez superada la aversión inicial". Rondeau subrayó genialmente todas esas disonancias lacerantes, al igual que recreó con humor y picardía las onomatopeyas de La Poule. Ningún pequeño detalle de esta sucesión de maravillas, vertidas siempre con el fraseo justo, con el tempo adecuado, le pasó inadvertido. El público lo percibió, le aplaudió con insistencia y Rondeau regaló a modo de propina una pieza del segundo libro de François Couperin que, hablando de títulos, lleva uno difícil de olvidar: Les Baricades Mistérieuses. Otro rondeau.

Por la noche, Rondeau -con mayúscula- unió fuerzas con otro joven portentoso (el laudista Thomas Dunford, que tocó hace poco en L'Orfeo que representaron Les Arts Florissants en los Teatros del Canal de Madrid) y con el percusionista iraní Keyvan Chemirani. Sin partitura alguna, y con un contacto visual y auditivo permanente entre los tres músicos, ofrecieron un concierto nocturno en el que se dieron la mano con naturalidad, y sin tonterías ni excentricidades, el Barroco y la música tradicional persa. Las improvisaciones -individuales y colectivas- se sucedieron sin descanso, tanto sobre piezas occidentales como orientales. Entre las primeras, hubo una secuencia genial en la que sonaron con ropajes insólitos, pero llenas de sentido, una chaconne de Robert de Visée, una ciaccona de Bernardo Storace y los dos ground basses del lamento de Dido y Music for a while, de Henry

APÉNDICE A. NOTAS TEMÁTICAS UTILIZADAS EN EXPERIMENTOS

Purcell. La fidelidad a la letra del concierto anterior dio paso aquí a la conservación del espíritu, dejando un amplísimo margen para crear música en una atmósfera de libertad y verdadera camaradería. Fuera de programa interpretaron Les Sauvages, de Rameau, que había tocado previamente verbatim Jean Rondeau en su recital de la tarde. La comparación entre el original y la reinvención fue, efectivamente, como cruzar de una orilla a otra del mismo río.

Entre ambos conciertos, también en la sede principal del festival de St John's Smith Square, pudo escucharse al grupo Florilegium, que tocó el quinto Concierto de Brandeburgo y dos obras de ultimísima época de Telemann. No fueron interpretaciones geniales, ni perfectas, pero sí fue uno de esos conciertos en los que el qué supera con mucho en interés al cómo. Y el atractivo fundamental radicaba aquí en la presencia al final del programa de la inusual y extraordinaria cantata Ino, de un Telemann muy inspirado y ya octogenario, cantada con entusiasmo por la soprano Elin Manahan Thomas. Flanqueado como estuvo por dos conciertos mayúsculos, casi se agradeció este pequeño remanso en unas pocas horas con semejante despliegue de emociones concentradas.

B Resultados detallados de notas temáticas

Los resultados están registrados por cada una de las frases detectadas en el texto correspondiente.

Resultados contexto *Derecho*

Frase	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f1	0	0.0051	0.0007	0.0008	0.0011
f2	0.0306	0.0771	0.0282	0.1129	0.0743
f3	0	0	0	0	0
f4	0.0052	0.0297	0.0098	0.028	0.0201
f5	0	0.0039	0.002	0.0031	0.0024
f6	0	0	0.003	0.002	0.0015
f7	0	0.0295	0.0134	0.0238	0.0166
f8	0	0.0044	0.0011	0.0047	0.0048
f9	0.0043	0.0027	0	0.0011	0
f10	0	0	0	0	0
Total	0.0402	0.1523	0.0581	0.1763	0.1209

Resultados contexto *Ambiente*

Frase	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f1	0.0156	0.0377	0.0478	0.0526	0.0438
f2	0.0022	0.0115	0.007	0.0064	0.0038
f3	0	0.0017	0.0009	0	0.0059
f4	0.0127	0.0443	0.0107	0.0336	0.0238
f5	0.0073	0.0201	0.0061	0.0111	0.0083
f6	0	0	0.0054	0	0.0005
f7	0.0177	0.0263	0.0423	0.0407	0.0605
f8	0.0026	0.0052	0.0002	0	0.0015
f9	0	0.0106	0.0065	0.001	0.0015
f10	0.0001	0.0095	0.0065	0.0072	0.0059
f11	0	0	0.0348	0	0.0045
f12	0	0.0006	0.0006	0.0017	0.0008
f13	0	0.019	0.0101	0.0112	0.0121
f14	0	0.0049	0.0034	0.0059	0.0053
f15	0	0.0098	0.0247	0.0112	0.0007
f16	0.0026	0.0011	0.0002	0	0.0123
f17	0.0072	0.0045	0.0213	0.0155	0.016
f18	0	0.0238	0.0342	0.0255	0.0211
f19	0	0.0003	0	0	0.0017
f20	0	0	0	0	0
f21	0	0	0	0	0.0001
f22	0.0219	0.0153	0.014	0.0183	0.0125
f23	0	0.0046	0.0017	0.004	0.0113
f24	0	0.0208	0.0117	0.0169	0.0267
f25	0.0025	0.004	0.0089	0.0037	0.0043
Total	0.0925	0.2757	0.2989	0.2666	0.2848

Resultados contexto *Informática*

Frase	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f1	0	0.0028	0	0.0002	0.0346
f2	0.0075	0.1391	0.1072	0.1597	0.2024
f3	0	0.0076	0.0094	0.0057	0.0048
f4	0	0.025	0.0038	0.0087	0.005
f5	0	0.0058	0.044	0.0036	0.0208
f6	0.0198	0.0431	0.0196	0.0341	0.0207
f7	0	0.0006	0.0006	0.0031	0.0006
f8	0.0065	0.039	0.0208	0.0151	0.0153
Total	0.0338	0.263	0.2054	0.2303	0.3042

Resultados contexto *Economía*

Frase	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f1	0	0	0	0	0
f2	0	0.0005	0.0385	0.0013	0.0009
f3	0	0.0045	0.0272	0.0054	0.009
f4	0.0031	0	0.0005	0	0.0006
f5	0.0001	0.0043	0.0003	0.0103	0.0003
f6	0	0.0034	0.0039	0.0029	0.0038
f7	0.0005	0.0044	0.0032	0.0038	0.0042
f8	0.0153	0.0541	0.0298	0.0457	0.0284
f9	0	0.0002	0.0003	0.0012	0.0002
f10	0.0021	0.0346	0.0572	0.0486	0.0341
f11	0.0354	0.3649	0.0922	0.2765	0.0931
f12	0.0005	0.0099	0.0037	0.0169	0.0023
f13	0.0136	0.0142	0.0297	0.0422	0.0257
f14	0.0001	0.0017	0.0013	0.0184	0.0002
f15	0	0	0.0271	0	0
f16	0	0	0	0	0
f17	0.0004	0.0041	0.0038	0.0138	0.0027
f18	0.0065	0.0001	0.0205	0.0148	0.0154
f19	0	0.0182	0.0357	0.0478	0.0397
f20	0	0.0006	0	0	0
f21	0	0.0123	0.006	0.0142	0.0121
f22	0	0.0115	0.0022	0.0114	0.002
f23	0	0.0022	0.0227	0.002	0.0025
f24	0.0002	0.0101	0.0062	0.0176	0.0106
f25	0.0041	0.0388	0.0258	0.0106	0.0072

Frase	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f26	0	0	0	0.0097	0
f27	0	0	0	0	0
f28	0	0.0059	0.0061	0.0061	0.0043
f29	0	0.0482	0.0001	0.0518	0.0002
f30	0	0.0069	0.0201	0.0031	0.0023
f31	0	0.0308	0.0374	0	0.0083
f32	0	0	0.0281	0	0
f33	0	0	0	0.1101	0
f34	0	0.0002	0.0002	0.0004	0.0001
f35	0	0	0	0	0.0209
f36	0	0.0018	0.0018	0.0014	0.0019
Total	0.0819	0.6886	0.5315	0.788	0.3329

Resultados contexto *Medicina*

Frase	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f1	0	0.0094	0.0018	0.0034	0.0021
f2	0	0.0205	0	0.0193	0.0683
f3	0	0.0204	0.0062	0.0125	0.0284
f4	0.0021	0.0051	0.0417	0.006	0.003
f5	0.036	0.0442	0.0548	0.0611	0.1208
f6	0.0048	0.1011	0.1131	0.1022	0.0926
f7	0.0439	0.0473	0.0496	0.0573	0.0494
f8	0.0731	0.1035	0.1161	0.2014	0.1869
f9	0	0.0008	0.0003	0	0
f10	0.0006	0.037	0.0583	0.0395	0.0358
f11	0.0082	0.3183	0.2479	0.1301	0.1369
f12	0	0	0	0	0
f13	0	0.0009	0.0391	0.03	0.0359
f14	0	0	0.0009	0.0009	0.0019
f15	0	0.0583	0.0698	0.0016	0.0584
f16	0.0023	0.0236	0.0256	0.028	0.0308
f17	0.0155	0.0176	0.0292	0.0266	0.0261
f18	0.003	0.0004	0	0	0.001
Total	0.1895	0.8083	0.8545	0.7198	0.8782

Resultados contexto desconocido (*Música*)

Frase	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f1	0	0	0	0	0
f2	0.0066	0.0057	0.0077	0.0084	0.0088
f3	0	0.0036	0	0.0089	0.0163
f4	0	0.0096	0.0073	0.0223	0.0163
f5	0	0	0.0043	0.0028	0.0028
f6	0	0.0096	0.0052	0.0049	0.004
f7	0.0006	0	0.0002	0.0008	0.0002
f8	0.0127	0.0212	0.0187	0.0218	0.0149
f9	0.0009	0.0117	0.001	0.0082	0.0073
f10	0	0.006	0.0207	0.009	0.0056
f11	0.0135	0.0117	0.0107	0.0159	0.0241
f12	0	0.004	0.0016	0.0028	0.0039
f13	0.0125	0.0219	0.019	0.035	0.0218
f14	0	0	0	0	0
f15	0	0.0036	0.0094	0.0069	0.0034
f16	0.0022	0.0335	0.0451	0.0472	0.049
f17	0	0.0015	0.0012	0.002	0.0013
f18	0	0	0	0	0.0001
f19	0	0.0031	0.0028	0.0031	0.0028
f20	0.0006	0	0	0	0.0003
f21	0	0	0	0	0
f22	0.0006	0.0003	0.0002	0.0012	0.0071
f23	0.0002	0.0052	0.018	0.0032	0.0022
f24	0	0	0.0004	0.0006	0.0008
f25	0	0.0062	0.0311	0.0128	0.0065

Frase	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f26	0	0	0	0	0
f27	0.0027	0	0	0.0029	0
f28	0	0.0036	0.0017	0.0001	0.0001
f29	0	0	0.0038	0	0
f30	0.0182	0.0303	0.0212	0.0209	0.0227
f31	0	0	0.0057	0	0.0068
f32	0	0	0	0	0
Total	0.0713	0.1922	0.2369	0.2416	0.2291

C Oraciones para experimento por oración individual

Grupo de oraciones del contexto

1. La revisión del uso del cloro ya no queda limitada a los grupos ecológicos sino que abarca los gobiernos y las entidades internacionales.
2. Una de las aportaciones más significativas al debate en torno al cloro en años recientes ha sido la "Campaña para la Eliminación del Cloro" de Greenpeace International.
3. En cuanto al uso del cloro en la producción del óxido de propileno, el estudio sugiere que el proceso alternativo con oxiran no sólo es viable económicamente sino que es ventajoso desde el punto de vista ecológico.
4. Se han expresado recientemente inquietudes sobre la posibilidad de que los productos químicos presentes en el medio ambiente sean capaces de imitar la acción de las hormonas animales y humanas, los estrógenos.
5. El pueblo catalán proclama como valores superiores de su vida colectiva la libertad, la justicia y la igualdad, y manifiesta su voluntad de avanzar por una vía de progreso que asegure una digna calidad de vida para todos los que viven y trabajan en Cataluña.
6. Una Ley del Parlamento regulara la organización territorial de Cataluña de acuerdo con el presente Estatuto, garantizando la autonomía de las distintas entidades territoriales.

APÉNDICE C. ORACIONES PARA EXPERIMENTO POR ORACIÓN INDIVIDUAL

7. El gobierno y la administración autónoma de las provincias estarán encomendados a Diputaciones u otras corporaciones de carácter representativo.

8. Todos los españoles tienen los mismos derechos y obligaciones en cualquier parte del territorio del Estado.

9. Existen marcadas diferencias en el ritmo de integración de los países a la economía mundial.

10. Por último, una característica de los países que se han integrado con rapidez es su elevado nivel de inversiones en infraestructura, por ejemplo, producción de energía eléctrica, pavimentación de caminos e instalación de líneas telefónicas.

11. El éxito de la integración de muchos países en desarrollo dependerá de la liberalización del comercio, para lo cual es necesario adoptar medidas de política drásticas que a menudo entrañan un elevado costo real.

12. El otro obstáculo grave de índole externa, a saber, los contingentes del Acuerdo Multifibras, constituye un elevadísimo impuesto obligatorio a las exportaciones de textiles y vestuario de los países en desarrollo.

13. La configuración se basa en las tarjetas configuradas y en los protocolos habilitados y configurados.

14. La configuración del servidor implica cierta planificación y toma de decisiones sobre la marcha.

15. Para evitar que el servidor notifique paquetes de notificación de servicios destinados a zonas particulares de la red.

16. Toda la red debe utilizar la misma máscara de subred.

17. El objetivo general es el desarrollo de las bases científicas y técnicas necesarias para evaluar nuevos medicamentos, en particular los destinados al tratamiento de las enfermedades neurológicas y mentales, inmunológicas y víricas.

18. La cartografía genética y el análisis del genoma, incluyendo la construcción de mapas de transcripción integrados, la secuenciación de regiones cromosómicas específicas y la explotación de enfoques comparativos.

19. La investigación en ética biomédica, de naturaleza horizontal, se interesará por las normas generales del respeto a la dignidad humana y de la protección del individuo en el contexto de la investigación biomédica y sus aplicaciones clínicas.

20. El cerebro es el órgano que mueve los músculos.

Grupo de oraciones ambiguas

21. Una tecnología que alumbra un floreciente negocio cuyos ingresos crecen a un ritmo anual del 55%.

22. Las máquinas que piensan como humanos contribuirán a mejorar la productividad impulsado con ello el crecimiento económico.

23. Desde hace años hemos sido testigos de la profunda transformación de la sociedad producida al compás de singulares avances tecnológicos.

24. Esta nueva estructura de vida descansa en un mundo digital donde a pesar de los indudables beneficios existen ciertas atribuciones negativas, este actual entorno también viene acompañado de nuevos y poderosos riesgos a la postre del progreso tecnológico, fundamentalmente en lo que atañe a

la seguridad en general de medios y dispositivos, y consecuentemente a la privacidad de las personas, pero no son los únicos.

25. En este nuevo entorno permanentemente conectado digitalmente, las empresas han tomado partida de sus beneficios.

26. La constitución de una comunicación continua impulsada por el empleo de dispositivos tecnológicos como herramientas para el desarrollo de las funciones encomendadas a los trabajadores, propician una nueva estructura organizativa empresarial.

27. Las tecnologías móviles se han convertido en herramientas de desarrollo para las empresas, ya que pueden contribuir a aumentar la productividad de sus empleados y sus ingresos, fomentan la innovación y aportar nuevas formas de atender a sus clientes.

28. El 84 por ciento de las empresas que aplican la tecnología móvil afirman haber aumentado la productividad en el último año. Podemos definir la conexión digital laboral como aquella situación en la que los empleados se ven permanentemente conectados al trabajo, sin lugar a desconexión, mediante el uso de herramientas tecnológicas empleadas para el desarrollo de las funciones encomendadas, de suerte que se ven obligados a atender multitud de tareas a través de dichas herramientas que conlleva a que el tiempo de trabajo quede dilatado a espacios de tiempo delimitados para el descanso y el disfrute de la vida personal.

29. Si bien es cierto que dicho marco normativo no proporciona una definición del derecho a la desconexión digital, por el contrario, proporciona la obligación de las empresas en adoptar políticas de actuación tendentes a asegurar el tiempo dedicado al descanso digital.

30. Las empresas deberán implantar sistemas tecnológicos que limiten o impidan

el acceso de los trabajadores a sus dispositivos digitales fuera del horario de trabajo.

31. El derecho forestal es una rama especial del derecho público ambiental que puede ser definido como el conjunto de principios y normas jurídicas que tienen por objeto la preservación, conservación, incremento, manejo y aprovechamiento sostenible de los ecosistemas forestales.

32. A pesar del reconocimiento internacional en nuestra época contemporánea, la denominación de Derecho médico no existía.

33. El Derecho Médico se define como la rama del Derecho que trata de la relación y aplicación de las leyes comunes y estatutarias a los principios y procedimientos de la higiene, ciencias de la salud y administración pública.

34. Se ha intentado, en su concepción, limitar el Derecho médico a la mera sustanciación de un juicio legal (civil o penal), contra un prestador de servicios de salud en una demanda judicial.

35. El uso de páginas wiki ha comenzado a atraer a la informática médica, considerando que la calidad de los artículos de Wikipedia ha ido aumentando y la gran necesidad que hay de difundir los conocimientos médicos de esta especialidad y de las ciencias de la salud en general.

36. La conciencia de que un medio ambiente, un medio laboral y doméstico deteriorado produce enfermedades supone un nuevo escenario para cometer su estudio y mejoramiento.

37. Entre las positivas es que la tecnología debe comprometerse a seguir procesos que no atenten contra el medio ambiente y así evitar el deterioro de los recursos naturales y ambientales.

38. Los residuos electrónicos, en cuyo grupo están incluidos los residuos informáticos, son un tipo de residuo formado por piezas o partes electrónicas rotas, o que ya no se desean usar más.

39. De la economía de la salud surge la Farmacoeconomía o evaluación económica de intervenciones en salud aplicada al uso de fármacos.

Grupo de oraciones desconocidas

41. Vivo en una casa adosada en Nueva Zelanda que tiene muchas habitaciones.

42. El albaricoque es una fruta de origen asiático.

43. La reconstrucción hecha por los canadienses indicó que, tras la colisión con el protoplaneta, la atmósfera terrestre estaba formada por vapores hirvientes con capacidad de disolver las rocas más superficiales, "más o menos como el azúcar se disuelve en el café".

44. Según los geólogos la cadena de eventos reconstruida en laboratorio determinó que sobre la Tierra, en un tiempo más bien breve, se registró la aparición de condiciones tales para favorecer el origen de la vida.

45. El Atlético de Madrid publica una nueva infografía con detalles del aspecto que tendrá su nuevo estadio, el Wanda Metropolitano.

46. Rafa Nadal mantiene esta temporada la posibilidad de firmar la temporada perfecta sobre tierra batida.

47. Escritas con arrojo y rabia, sus páginas rezuman sensualidad y autenticidad, su estilo es sugerente y original, cada párrafo posee tensión y chispa.

48. El autobús saldrá de la estación a las ocho de la noche.

49. En la cima de la montaña, se encuentra la bandera de nuestro país.
50. Los discos compactos reemplazaron a los cassettes como medio de almacenamiento de la música a mediados de los años ochenta.
51. La nota blanca dura dos notas negras en una partitura.
52. La lotería nacional se juega los domingos por la noche con gran expectación.
53. Las señas en LESCO se realizan con varias partes del cuerpo, no sólo las manos.
54. Caminante no hay camino, se hace camino al andar.
55. La franquicia Star Wars sigue asombrando en las pantallas de los cines a nivel mundial.
56. El fénix es un ave mitológica que puede revivir de sus cenizas.
57. La sal es un compuesto químico que se llama cloruro de sodio.
58. La tarea de taladrar siempre es algo que infunde mucho respeto, sobre todo para aquellas personas que no están muy iniciadas en el bricolaje.
59. Cocinar arroz es sencillo pero al mismo tiempo cosa seria, pues quien sabe hacer arroz correctamente se puede considerar un experto en la cocina.
60. La mañana de este lunes se pudo observar una columna de vapor que se extiende a unos 500 metros de altura sobre el volcán Turrialba, en Cartago.

D Resultados detallados experimento de oraciones individuales

Ejemplo	**Esperado**	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f1	Ambiente	0	0.0352	0	0.0024	0
f2	Ambiente	0.0016	0.0258	0.0227	0.0096	0.008
f3	Ambiente	0	0.0746	0	0.0075	0.0124
f4	Ambiente	0	0.0342	0.0002	0.0146	0.0282
f5	Derecho	0.1307	0.0142	0.006	0.0257	0
f6	Derecho	0.0723	0	0	0	0.0013
f7	Derecho	0.0366	0.0095	0.0261	0.027	0.0152
f8	Derecho	0.0598	0.031	0.0214	0.0303	0.023
f9	Economía	0	0.0044	0.004	0.0253	0.013
f10	Economía	0.0038	0.0508	0.0211	0.0717	0.0241
f11	Economía	0	0.0393	0.0056	0.0453	0.0126
f12	Economía	0.0105	0.0206	0.0008	0.0635	0.0076
f13	Informática	0.0019	0.0083	0.0511	0.0021	0.0051
f14	Informática	0.0012	0.0013	0.026	0.0045	0.0019
f15	Informática	0	0.0115	0.0504	0.0014	0.0017
f16	Informática	0.0001	0.0001	0.0519	0	0
f17	Medicina	0.0144	0.033	0.0146	0.0314	0.0447
f18	Medicina	0.0083	0.0469	0.0359	0.0301	0.1087
f19	Medicina	0	0	0	0.045	0.0934
f20	Medicina	0.0051	0	0	0	0.0138

Ejemplo	**Esperado**	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f21	Economía o Informática	0	0.0011	0	0	0
f22	Economía o Informática	0	0.0041	0	0.0025	0.003
f23	Derecho o Informática	0	0.0096	0.0182	0.0179	0.0081
f24	Derecho o Informática	0	0.0037	0.004	0.004	0.0005
f25	Economía o Informática	0	0.0031	0.008	0.0091	0.0019
f26	Economía o Informática	0	0.0006	0	0.0003	0
f27	Economía o Informática	0	0.0076	0.0283	0.0081	0.0027
f28	Economía o Informática	0	0.0074	0.0065	0.001	0.0059
f29	Economía o Informática	0	0.0077	0.0458	0.0655	0.0099
f30	Derecho o Informática	0	0.0028	0.0091	0.0022	0.0026
f31	Economía o Informática	0	0	0	0	0.0017
f32	Derecho o Ambiente	0.0107	0.0195	0	0.0004	0.0067
f33	Derecho o Medicina	0	0.0039	0.0035	0.0062	0.0035
f34	Derecho o Medicina	0	0.0433	0.0265	0.0411	0.0447
f35	Derecho o Medicina	0	0.0003	0.0188	0.0039	0.0016
f36	Informática o Medicina	0	0	0	0.0009	0.0173
f37	Medicina o Ambiente	0	0.0046	0.0017	0.004	0.0131
f38	Informática o Ambiente	0	0.0047	0.0017	0.0013	0.0017
f39	Informática o Ambiente	0.0002	0.0114	0.0162	0.0112	0.0079
f40	Economía o Medicina	0	0.0306	0.0155	0.027	0.0147

Ejemplo	**Esperado**	**Derecho**	**Ambiente**	**Informática**	**Economía**	**Medicina**
f41	*No definido*	0	0.0014	0	0.0026	0.0011
f42	*No definido*	0	0	0	0.0022	0
f43	*No definido*	0	0.0019	0.0023	0.0238	0.002
f44	*No definido*	0.0102	0.0081	0.0046	0.0087	0.0071
f45	*No definido*	0.0112	0.0009	0.0004	0.0068	0
f46	*No definido*	0	0.004	0.0044	0.0054	0.0039
f47	*No definido*	0	0	0	0	0
f48	*No definido*	0	0.0007	0.0081	0	0.0032
f49	*No definido*	0.0013	0.0029	0.0041	0.0043	0.0047
f50	*No definido*	0	0	0	0.0001	0.0005
f51	*No definido*	0	0	0.0008	0	0.0014
f52	*No definido*	0	0.0005	0	0.0004	0.0002
f53	*No definido*	0	0.0047	0.0074	0.0107	0.0125
f54	*No definido*	0	0.0001	0.0183	0.0107	0.0001
f55	*No definido*	0	0	0	0	0
f56	*No definido*	0	0.0005	0.0007	0.0005	0.0006
f57	*No definido*	0	0.0011	0	0	0
f58	*No definido*	0.0024	0.0064	0.0064	0.0057	0.0024
f59	*No definido*	0	0.0399	0.007	0.0078	0.006
f60	*No definido*	0	0.0309	0.0003	0.0369	0.0341

E Archivo de configuración para FreeLing 4.0

```
##
#### default configuration file for Spanish analyzer
##

#### General options
Lang=es
Locale=default

### Tagset description file, used by different modules
TagsetFile=$FREELINGSHARE/es/tagset.dat

#### Trace options. Only effective if we have compiled with -DVERBOSE
#
## Possible values for TraceModule (may be OR'ed)
#define SPLIT_TRACE         0x00000001
#define TOKEN_TRACE         0x00000002
#define MACO_TRACE          0x00000004
#define OPTIONS_TRACE       0x00000008
#define NUMBERS_TRACE       0x00000010
#define DATES_TRACE         0x00000020
#define PUNCT_TRACE         0x00000040
#define DICT_TRACE          0x00000080
#define SUFF_TRACE          0x00000100
#define LOCUT_TRACE         0x00000200
#define NP_TRACE            0x00000400
#define PROB_TRACE          0x00000800
#define QUANT_TRACE         0x00001000
#define NEC_TRACE           0x00002000
#define AUTOMAT_TRACE       0x00004000
#define TAGGER_TRACE        0x00008000
#define HMM_TRACE           0x00010000
#define RELAX_TRACE         0x00020000
#define RELAX_TAGGER_TRACE  0x00040000
#define CONST_GRAMMAR_TRACE 0x00080000
#define SENSES_TRACE        0x00100000
#define CHART_TRACE         0x00200000
#define GRAMMAR_TRACE       0x00400000
#define DEP_TRACE           0x00800000
```

APÉNDICE E. ARCHIVO DE CONFIGURACIÓN PARA FREELING 4.0

```
#define UTIL_TRACE          0x01000000

TraceLevel=3
TraceModule=0x0000

## Options to control the applied modules. The input may be partially
## processed, or not a full analysis may me wanted. The specific
## formats are a choice of the main program using the library, as well
## as the responsability of calling only the required modules.
## Valid input/output formats are: plain, token, splitted, morfo, tagged, parsed
InputLevel=text

 # Líneas cambiadas de la configuración por omisión
OutputLevel=dep
OutputFormat=json

# consider each newline as a sentence end
AlwaysFlush=no

#### Tokenizer options
TokenizerFile=$FREELINGSHARE/es/tokenizer.dat

#### Splitter options
SplitterFile=$FREELINGSHARE/es/splitter.dat

#### Morfo options
AffixAnalysis=yes
CompoundAnalysis=yes
MultiwordsDetection=yes
NumbersDetection=yes
PunctuationDetection=yes
DatesDetection=yes
QuantitiesDetection=yes
DictionarySearch=yes
ProbabilityAssignment=yes
DecimalPoint=,
ThousandPoint=.
LocutionsFile=$FREELINGSHARE/es/locucions.dat
QuantitiesFile=$FREELINGSHARE/es/quantities.dat
AffixFile=$FREELINGSHARE/es/afixos.dat
CompoundFile=$FREELINGSHARE/es/compounds.dat
ProbabilityFile=$FREELINGSHARE/es/probabilitats.dat
DictionaryFile=$FREELINGSHARE/es/dicc.src
PunctuationFile=$FREELINGSHARE/common/punct.dat
```

APÉNDICE E. ARCHIVO DE CONFIGURACIÓN PARA FREELING 4.0

```
ProbabilityThreshold=0.001

# NER options
NERecognition=yes
NPDataFile=$FREELINGSHARE/es/np.dat
## comment line above and uncomment one of those below, if you want
## a better NE recognizer (higer accuracy, lower speed)
#NPDataFile=$FREELINGSHARE/es/nerc/ner/ner-ab-poor1.dat
#NPDataFile=$FREELINGSHARE/es/nerc/ner/ner-ab-rich.dat
# "rich" model is trained with rich gazetteer. Offers higher accuracy but
# requires adapting gazetteer files to have high coverage on target corpus.
# "poor1" model is trained with poor gazetteer. Accuracy is splightly lower
# but suffers small accuracy loss the gazetteer has low coverage in target corpus.
# If in doubt, use "poor1" model.

## Phonetic encoding of words.
Phonetics=no
PhoneticsFile=$FREELINGSHARE/es/phonetics.dat

## NEC options. See README in common/nec
NEClassification=no
NECFile=$FREELINGSHARE/es/nerc/nec/nec-ab-poor1.dat
#NECFile=$FREELINGSHARE/es/nerc/nec/nec-ab-rich.dat

## Sense annotation options (none,all,mfs,ukb)
SenseAnnotation=none
SenseConfigFile=$FREELINGSHARE/es/senses.dat
UKBConfigFile=$FREELINGSHARE/es/ukb.dat

#### Tagger options
Tagger=hmm
TaggerHMMFile=$FREELINGSHARE/es/tagger.dat
TaggerRelaxFile=$FREELINGSHARE/es/constr_gram-B.dat
TaggerRelaxMaxIter=500
TaggerRelaxScaleFactor=670.0
TaggerRelaxEpsilon=0.001
TaggerRetokenize=yes
TaggerForceSelect=tagger

#### Parser options
GrammarFile=$FREELINGSHARE/es/chunker/grammar-chunk.dat

#### Dependence Parser options
```

APÉNDICE E. ARCHIVO DE CONFIGURACIÓN PARA FREELING 4.0

```
DependencyParser=txala
DepTxalaFile=$FREELINGSHARE/es/dep_txala/dependences.dat
DepTreelerFile=$FREELINGSHARE/es/dep_treeler/dependences.dat

#### Coreference Solver options
CorefFile=$FREELINGSHARE/es/coref/relaxcor/relaxcor.dat
SemGraphExtractorFile=$FREELINGSHARE/es/semgraph/semgraph-SRL.dat
```

F Formato de archivo de modelo

Cinco oraciones sobre el tema de *viajes* en un modelo contextual.

```
CTX-MODEL Viajes
* 0 * ir|pasar|cobrar
, 0 Fc
. 0 Fp
? 0 Fit
a 0.00277778 SP estados_unidos
avión 0.025 NCMS000
año 0.025 NCMS000 el|venir|,
cobrar 0.06 VMIP3S0 ¿|cuánto|por|en|uno|habitación
cuánto 0 PT0MS00
cámping 0.025 NCMS000
de 0.000555556 SP cámping
día 0.025 NCMS000 uno
el 0.00166667 DA0MP0
en 0.00277778 SP avión|nueva_zelanda|francia|españa
españa 0.025 NP00000
estados_unidos 0.025 NP00000 el
francia 0.025 NP00000
habitación 0.025 NCFS000 singular|?
ir 0.36 VMN0000 ir|a|en|.|año|de
mes 0.025 NCMS000 uno
noche 0.025 NCFS000
nueva_zelanda 0.025 NP00000
pasar 0.12 VMN0000 ir|a|usted|mes|en|.|semana|día
por 0.000555556 SP noche
próximo 0.045 AQ0FS00
que 0 PR0CN00
semana 0.025 NCFS000 el|próximo|,
singular 0.045 AQ0CS00
uno 0.00166667 DI0MS0
usted 0 PP2CP0P
venir 0.06 VMIP3S0 que
¿ 0 Fia
```

Corpus

```
Van a ir a los Estados Unidos en avión.
Ustedes van a pasar un mes en Nueva Zelanda.
El año que viene, vamos a ir de cámping en Francia.
La semana próxima, voy a pasar un día en España.
¿Cuánto cobra por noche en un habitación singular?
```

G Formato de salida del módulo desambiguador hacia generador de señas

Ejemplo resumido de formato del archivo resultado para el texto *“El veloz murciélago hindú comía feliz cardillo y kiwi. La cigüeña tocaba el saxofón detrás del palenque de paja.”*


```
[{
  "concepts":[
    {
      "conceptId":"el",
      "lemma":"el",
      "source":"El",
      "tag":"DA0MS0",
      "unknown":true
    },
    {
      "conceptId":"veloz",
      "lemma":"veloz",
      "source":"veloz",
      "tag":"AQ0CS00",
      "unknown":true
    },

    ...

    {
      "conceptId":".",
      "lemma":".",
      "source":".",
      "tag":"Fp",
      "unknown":true
    }
    ],
  "context":"Derecho"
},
```

```
{
  "concepts":[
    {
      "conceptId":"el",
      "lemma":"el",
      "source":"La",
      "tag":"DA0FS0",
      "unknown":true
    },
    {
      "conceptId":"cigüeña",
      "lemma":"cigüeña",
      "source":"cigüeña",
      "tag":"NCFS000",
      "unknown":true
    },

    ...

    {
      "conceptId":".",
      "lemma":".",
      "source":".",
      "tag":"Fp",
      "unknown":true
    }
    ],
  "context":"Derecho"
}]
```