PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
FACULTAD DE LETRAS
DEPARTAMENTO DE CIENCIAS DEL LENGUAJE

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

DISEÑO Y DESARROLLO DE UN MODELO DE DESAMBIGUACIÓN LÉXICA AUTOMÁTICA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

Tesis presentada como requisito parcial para obtener el grado de Doctor en Lingüística

FREDY NÚÑEZ TORRES

Profesor director:
Dr. Carlos González Vergara
Profesor codirector:
Dr. Carlos Periñán-Pascual (Universitat Politècnica de València, España)

Profesora informante interna:
Dra. Luciana Pissolato de Oliveira
Profesores informantes externos:
Dra. Beatriz Pérez Cabello de Alba (Universidad Nacional de Educación a Distancia, España)
Dr. Walter Koza Orellana (Pontificia Universidad Católica de Valparaíso, Chile)

Septiembre de 2021

A la memoria de mi padre,
Fredy Eduardo Núñez Gómez (1958-2018).
No todo termina con la muerte.

Resumen

La presente investigación doctoral tiene como objetivo general desarrollar un modelo más robusto de medida para la similitud y relación semántica que los disponibles actualmente para resolver el problema de la desambiguación léxica automática, aplicado al procesamiento del lenguaje natural (PLN). Para esto, se realizó una revisión del fenómeno lingüístico de la ambigüedad léxica, junto con los métodos para la desambiguación léxica automática más representativos y que han sido aplicados en PLN: de relación semántica, de similitud semántica, y basados en conocimiento contextual. Luego se expuso una panorámica cronológica de la utilización del corpus en el análisis lingüístico, junto con una caracterización de los llamados recursos lingüísticos informatizados. Como aspecto central de la propuesta, se estableció una metodología para la aplicación de los subtipos generales de procesamiento de datos en aprendizaje automático, con sus respectivas tareas de procesamiento. Posteriormente se ejecutó un experimento de desambiguación léxica automática basado en el corpus SENSEVAL-3 (*Evaluating Word Sense Disambiguation Systems*), utilizando un método de aprendizaje automático supervisado. Este experimento permitió consolidar la metodología para la ejecución un nuevo experimento, diseñado a partir del montaje de un corpus basado en una submuestra de CODICACH (*Corpus Dinámico del Castellano de Chile*), que consideró unidades léxicas polisémicas seleccionadas desde la base de conocimiento FunGramKB. Posteriormente, se reportaron los resultados de los sistemas de desambiguación basados en aprendizaje automático, junto con las críticas al modelo. Este proceso permitió desarrollar un modelo de desambiguación léxica automática basado en una medida híbrida, y fundamentado tanto lingüística como estadísticamente en la interacción de dos enfoques de exploración taxonómica: distancia entre rutas y contenido de información, a través de la incorporación de FunGramKB como inventario de sentidos. En cuanto a la evaluación, la medida de similitud propuesta $SIM_{\text{híbrida}}(c_i, c_j)$ logró resultados consistentemente eficientes desde un punto de vista lingüístico en el proceso de desambiguación léxica automática.

Palabras clave: ambigüedad léxica, desambiguación léxica automática, procesamiento del lenguaje natural, recursos lingüísticos informatizados, procesamiento de datos textuales.

Tanto mis estudios doctorales como el desarrollo de esta tesis fueron patrocinados por la Agencia Nacional de Investigación y Desarrollo (ANID) del Ministerio de Ciencia, Tecnología, Conocimiento e Innovación del Gobierno de Chile, en el marco del Programa de Formación de Capital Humano Avanzado, Beca de Doctorado Nacional 2016 (folio Nº 21160361).

Los resultados de esta tesis doctoral están vinculados con el desarrollo del módulo de PLN en el proyecto de investigación "Planificación y gestión de recursos hídricos a partir de análisis de datos de IoT (WATERoT)" (RTC 2017-6389-5), financiado por el Ministerio de Economía, Industria y Competitividad (MINECO), Agencia Estatal de Investigación (AEI) y el Fondo Europeo de Desarrollo Regional (FEDER).

Índice de contenidos

Resumen **03**
Índice de contenidos **05**
Lista de tablas 09
Lista de figuras 11
Lista de anexos 13
Agradecimientos **14**
Capítulo 1. Introducción **16**
Capítulo 2. Estado de la cuestión **21**
2.1 El aporte de las ciencias del lenguaje al PLN 21
2.2 El problema lingüístico de la ambigüedad léxica 27
2.2.1 El enfoque descriptivo 30
2.2.2 El enfoque relacional 31
2.2.3 El enfoque generativo 32
2.3 El problema computacional de la ambigüedad léxica 35
2.3.1 Definición computacional de la ambigüedad léxica 37
2.3.2 La desambiguación léxica en el ámbito del PLN 38
2.4 Métodos para la desambiguación léxica automática 42
2.4.1 Métodos de relación semántica (*word relatedness*) 44
2.4.1.1 Tipo Lesk (1986; 1987) 44
2.4.1.2 Desambiguación estadística (Cantos-Gómez, 1996) 48
2.4.1.3 Tipo Lesk adaptado (Banerjee & Pedersen, 2002) 50
2.4.2 Métodos de similitud semántica (*word similarity*) 51
2.4.2.1 Medidas de distancia entre rutas 51
2.4.2.1.1 Wu & Palmer (1994) 52
2.4.2.1.2 Leacock & Chodorow (1998) 53
2.4.2.2 Medidas de contenido de información 54
2.4.2.2.1 Resnik (1995) 56
2.4.2.2.2 Jiang & Conrath (1997) 56
2.4.2.2.3 Lin (1998) 57

2.4.3 Métodos basados en conocimiento contextual 57
2.4.3.1 Aprendizaje automático supervisado 58
2.4.3.1.1 Algoritmo bayesiano ingenuo (*Naïve Bayes*) 59
2.4.3.2 Aprendizaje automático no supervisado 63
Capítulo 3. Conceptos fundamentales 66
3.1 La utilización de corpus en el análisis lingüístico 66
3.2 Desde el corpus hacia los recursos lingüísticos informatizados 68
3.3 La base de conocimiento léxico-conceptual-gramatical FunGramKB 71
3.3.1 Aspectos generales de la Gramática del Papel y la Referencia 74
3.3.2 Lenguaje de Representación Conceptual (COREL) 75
3.4 El lexicón mental 79
3.5 El entorno de trabajo DAMIEN (*Data Mining Encountered*) 83
3.6 Herramientas informáticas para el tratamiento de datos textuales 84
3.6.1 Expresiones regulares (regex) 85
3.6.2 Lenguaje de etiquetado extensible (XML) 88
3.6.3 Lenguaje de consulta estructurada (SQL) 89
Capítulo 4. Pregunta de investigación y objetivos 92
4.1 Objetivo general 92
4.2 Objetivos específicos 92
Capítulo 5. Metodología 93
5.1 Subtipos de procesamiento de datos en aprendizaje automático 93
5.1.1 Preprocesamiento 93
5.1.2 Procesamiento 93
5.1.3 Evaluación 94
5.2 Experimento piloto utilizando el corpus de prueba SENSEVAL-3 96
5.2.1 Selección del corpus de prueba SENSEVAL-3 96
5.2.2 Resultados del experimento piloto utilizando el corpus de prueba SENSEVAL-3 99
5.3 Proceso de elección de conceptos en la subontología #ENTITY 102
5.3.1 La polisemia en «cabeza» y sus representaciones conceptuales en FunGramKB 102

5.3.2 La polisemia en «cara» y sus representaciones conceptuales en FunGramKB 103
5.3.3 La polisemia en «carta» y sus representaciones conceptuales en FunGramKB 104
5.4 Corpus Dinámico del Castellano de Chile (CODICACH) 104
5.4.1 Submuestra de CODICACH y colecciones de documentos 105
5.5 Procedimiento en DAMIEN para experimentos de aprendizaje automático 107
5.5.1 Tareas de preprocesamiento 108
5.5.2 Tareas de procesamiento 110
5.5.3 Tarea de minería textual 114
5.5.4 Tareas de evaluación 115
Capítulo 6. Experimentos de base para la desambiguación léxica automática 118
6.1 Resultados de los sistemas de desambiguación en aprendizaje automático 118
6.1.1 Resultados del sistema de desambiguación automática para la unidad léxica «cabeza» 118
6.1.2 Resultados del sistema de desambiguación automática para la unidad léxica «cara» 122
6.1.3 Resultados del sistema de desambiguación automática para la unidad léxica «carta» 123
6.2 Críticas al modelo de aprendizaje automático para la desambiguación léxica 126
Capítulo 7. Modelo para la desambiguación léxica automática basado en una medida híbrida 129
7.1 Una medida híbrida: fundamentos de la propuesta 129
7.2 Relaciones taxonómicas e información conceptual en FunGramKB 132
7.2.1 Taxonomía y postulados de significado para «cabeza» 132
7.2.2 Taxonomía y postulados de significado para «cara» 136
7.2.3 Taxonomía y postulados de significado para «carta» 139
7.3 Componentes de la medida híbrida 142

Capítulo 8. Evaluación de la medida híbrida **148**
8.1 Valores de profundidad, hipónimos y contenido de información 148
8.2 Evaluación de casos de similitud semántica basados en la medida híbrida 152
8.2.1 Evaluación de similitud semántica para el caso «órgano» y los sentidos de «cabeza» 155
8.2.2 Evaluación de similitud semántica para el caso «superficie» y los sentidos de «cara» 159
8.2.3 Evaluación de similitud semántica para el caso «documento» y los sentidos de «carta» 161
8.3 Resultados para la evaluación del desempeño de la medida híbrida 162
Capítulo 9. Conclusiones y futuras investigaciones **165**
Referencias bibliográficas **170**
Anexos **186**

Lista de tablas

Tabla 1. Sentidos y definiciones para «pine» 45
Tabla 2. Sentidos y definiciones para «cone» 45
Tabla 3. Puntajes para el solapamiento entre los conceptos «pine» y «cone» 46
Tabla 4. Dos sentidos de «ball» y sus ocurrencias 49
Tabla 5. Conteo de ocurrencias para dos sentidos de «ball» en un corpus 49
Tabla 6. Metacaracteres del tipo *clase* y sus definiciones 86
Tabla 7. Metacaracteres del tipo *frontera* y sus definiciones 86
Tabla 8. Metacaracteres del tipo *cuantificador* y sus definiciones 87
Tabla 9. Matriz de confusión para «partido.1» 99
Tabla 10. Matriz de confusión para «partido.2» 100
Tabla 11. Resultados del sistema de desambiguación automática para «partido» 100
Tabla 12. Polisemia en «cabeza» y sus representaciones en FunGramKB 103
Tabla 13. Polisemia en «cara» y sus representaciones en FunGramKB 103
Tabla 14. Polisemia en «carta» y sus representaciones en FunGramKB 104
Tabla 15. Ejemplos de organización de la submuestra para «cabeza» 105
Tabla 16. Ejemplos de organización de la submuestra para «cara» 106
Tabla 17. Ejemplos de organización de la submuestra para «carta» 106
Tabla 18. Descripción cuantitativa para cada colección de documentos desde la submuestra de CODICACH 106
Tabla 19. Tareas de preprocesamiento para experimentos de aprendizaje automático 110
Tabla 20. Tareas de procesamiento para experimentos de aprendizaje automático 113
Tabla 21. Tarea de minería textual para experimentos de aprendizaje automático 115
Tabla 22. Tareas de evaluación para experimentos de aprendizaje automático 117
Tabla 23. Matriz de confusión para el sentido +HEAD_00 de «cabeza» 119
Tabla 24. Matriz de confusión para el sentido +CHIEF_00 de «cabeza» 119
Tabla 25. Matriz de confusión para el sentido +LEADER_00 de «cabeza» 120
Tabla 26. Matriz de confusión para el sentido +INTELLIGENCE_00 de «cabeza» 121
Tabla 27. Resultados del sistema de desambiguación automática para «cabeza» 121
Tabla 28. Matriz de confusión para el sentido +FACE_00 de «cara» 122

Tabla 29. Matriz de confusión para el sentido +SIDE_00 de «cara» 123
Tabla 30. Resultados del sistema de desambiguación automática para «cara» 123
Tabla 31. Matriz de confusión para el sentido +LETTER_00 de «carta» 124
Tabla 32. Matriz de confusión para el sentido +CARD_00 de «carta» 124
Tabla 33. Matriz de confusión para el sentido $MENU _00 de «carta» 125
Tabla 34. Resultados del sistema de desambiguación automática para «carta» 125
Tabla 35. Macropromedios para los sistemas de desambiguación automática 126
Tabla 36. Componentes de la medida híbrida de similitud semántica 146
Tabla 37. Relaciones *IS-A* para los conceptos en la taxonomía de #ENTITY 149
Tabla 38. Valores *depth*, *hypo* e IC para los nodos correspondientes a los sentidos de «cabeza» 150
Tabla 39. Valores *depth*, *hypo* e IC para los nodos correspondientes a los sentidos de «cara» 151
Tabla 40. Valores *depth*, *hypo* e IC para los nodos correspondientes a los sentidos de «carta» 152
Tabla 41. Valores para los parámetros de optimización α y β 154
Tabla 42. Valores normalizados de IC para «cabeza» y «órgano» 156
Tabla 43. Valores normalizados de PB para la combinatoria de «cabeza» y «órgano» 158
Tabla 44. Resultados de aplicación de la medida híbrida para $SIM_{\text{híbrida}}$(*cabeza*, *órgano*) 158
Tabla 45. Valores normalizados de IC para «cara» y «superficie» 159
Tabla 46. Valores normalizados de PB para la combinatoria de «cara» y «superficie» 159
Tabla 47. Resultados de aplicación de la medida híbrida para $SIM_{\text{híbrida}}$(*cara, superficie*) 160
Tabla 48. Valores normalizados de IC para «carta» y «documento» 161
Tabla 49. Valores normalizados de PB para la combinatoria de «carta» y «documento» 161
Tabla 50. Resultados de aplicación de la medida híbrida para $SIM_{\text{híbrida}}$(*carta, documento*) 162
Tabla 51. Resultados de aplicación de $SIM_{\text{híbrida}}(c_i, c_j)$ para la combinación tres 163

Lista de figuras

Figura 1. Selección de técnicas en ciencia de datos (tomada de Choi *et al*., 2020) 23
Figura 2. Procedimiento general de métodos basados en solapamiento 47
Figura 3. Taxonomía de los conceptos HILL y COAST 52
Figura 4. Hiperónimo común más cercano para los conceptos CARD y MAGICIAN en la taxonomía ENTITY de FunGramKB 55
Figura 5. El planeta cognitivo para el nivel conceptual de FunGramKB 73
Figura 6. Niveles conceptuales en FunGramKB 74
Figura 7. Jerarquía para la subontología de entidades en FunGramKB 77
Figura 8. Proceso de lexicalización o codificación léxica 81
Figura 9. Matriz (binaria) de confusión 94
Figura 10. Extracción de la ventana contextual en DAMIEN 108
Figura 11. Generación de una colección de documentos (sin anotar) en DAMIEN 109
Figura 12. Extracción de etiquetas *senseID* en DAMIEN 109
Figura 13. Generación de una matriz *N-grama/documento* en DAMIEN 111
Figura 14. Generación de una lista de inicio en DAMIEN 112
Figura 15. Generación de una matriz *N-grama/documento* con lista de inicio en DAMIEN 112
Figura 16. Secuencia *join right* para generar una matriz filtrada y anotada en DAMIEN 113
Figura 17. Validación cruzada en DAMIEN 114
Figura 18. Aplicación del algoritmo bayesiano ingenuo en DAMIEN 115
Figura 19. Generación de una matriz de confusión en DAMIEN 116
Figura 20. Ejemplo de matriz de confusión en DAMIEN 116
Figura 21. Propuesta de modelo de desambiguación léxica automática 132
Figura 22. Taxonomía de la subontología #ENTITY para los sentidos de «cabeza» 133
Figura 23. Información conceptual para +HEAD_00 en FunGramKB 134
Figura 24. Información conceptual para +LEADER_00 en FunGramKB 135
Figura 25. Información conceptual para +CHIEF_00 en FunGramKB 135
Figura 26. Información conceptual para +INTELLIGENCE_00 en FunGramKB 136
Figura 27. Taxonomía de la subontología #ENTITY para los sentidos de «cara» 137
Figura 28. Información conceptual para +FACE_00 en FunGramKB 138

Figura 29. Información conceptual para +SIDE_00 en FunGramKB 139
Figura 30. Taxonomía de la subontología #ENTITY para los sentidos de «carta» 139
Figura 31. Información conceptual para +LETTER_00 en FunGramKB 140
Figura 32. Información conceptual para +CARD_00 en FunGramKB 141
Figura 33. Información conceptual para $MENU_00 en FunGramKB 142

Lista de anexos

Anexo 1. Minidiccionario para los sentidos de «partido» correspondiente al corpus SENSEVAL-3 186
Anexo 2: Selección de 120 instancias para la unidad léxica «partido» extraída desde el corpus SENSEVAL-3 187
Anexo 3: Selección de 120 instancias para la unidad léxica «cabeza» 217
Anexo 4: Selección de 120 instancias para la unidad léxica «cara» 221
Anexo 5: Selección de 120 instancias para la unidad léxica «carta» 225
Anexo 6: Matrices de confusión para cada *dataset* del sentido +HEAD_00 de «cabeza» 229
Anexo 7: Matrices de confusión para cada *dataset* del sentido +CHIEF_00 de «cabeza» 230
Anexo 8: Matrices de confusión para cada *dataset* del sentido +LEADER_00 de «cabeza» 231
Anexo 9: Matrices de confusión para cada *dataset* del sentido +INTELLIGENCE_00 de «cabeza» 232
Anexo 10: Matrices de confusión para cada *dataset* del sentido +FACE_00 de «cara» 234
Anexo 11: Matrices de confusión para cada *dataset* del sentido +SIDE_00 de «cara» 235
Anexo 12: Matrices de confusión para cada *dataset* del sentido +LETTER_00 de «carta» 236
Anexo 13: Matrices de confusión para cada *dataset* del sentido +CARD_00 de «carta» 237
Anexo 14: Matrices de confusión para cada *dataset* del sentido $MENU _00 de «carta» 238

Agradecimientos

Bendita ambigüedad del lenguaje, de la que vivimos desde hace tantos años.
Marcos Mundstock, integrante de Les Luthiers (2017).

Al terminar este proceso siento una alegría profunda, solemne, también un tanto melancólica. Tal vez un poco aturdido aún por el final pienso que, probablemente, este planteamiento de un modelo para la desambiguación léxica automática propuesto por un lingüista en el extremo sur del mundo no constituirá un avance definitorio para la disciplina. Sin embargo, tengo la certeza de que ha sido una de las más desafiantes, impredecibles, edificantes y satisfactorias experiencias que he vivido. Sé positivamente que de ninguna manera hubiese podido solo. Les agradezco por perseverar conmigo:

A Andrea, quien supo rebatir con lucidez, paciencia y ternura las tres razones que le expuse para renunciar. Durante mucho tiempo tuve la certeza de que el trabajo de la muerte era más fuerte que el amor presente en mi historia. Gracias a ella he llegado a comprender lo equivocado que estaba.

A mi profesor director de tesis, Carlos González Vergara. Hubo momentos clave en los que fui sostenido e inspirado por Carlos. Su influencia no solo me ha ayudado a crecer, primero como estudiante y ahora como lingüista, sino que también me ha enseñado a ser un mejor profesor, puesto al servicio de otros y otras. Gracias por todas esas horas de trabajo, conversaciones y aprendizaje.

A mi profesor codirector de tesis, Carlos Periñán Pascual, quien me recibió en una estancia de investigación en la Universitat Politècnica de València. Carlos es una de las personas más generosas que he conocido. Gracias a su orientación y sabiduría he podido adquirir conocimientos y vivir experiencias que han sido fundamentales para mi desarrollo, tanto personal como académico.

A la profesora Teresa Oteíza Silva, quien fuera directora del Doctorado en Lingüística UC durante la mayor parte de mi tiempo en el programa. Esta tesis no hubiese sido posible sin la confianza que ella depositó en mí y en mi trabajo cuando más lo necesité.

A los profesores Ricardo Mairal-Usón, Beatriz Pérez Cabello de Alba, y Ángel Felices Lago. Siempre que pienso en España lo primero que recuerdo son nuestras reposadas caminatas por el Parque del Retiro, las inigualables lecciones de historia del arte en La Cartuja, o la aventura de probar la mejor tortilla de patatas de todo Madrid. Gracias por su apoyo constante, por las oportunidades que me han brindado y, sobre todo, por su amistad, que es para mí un regalo enorme.

A mis padres, Fredy y Verónica, por todo el amor con el que construyeron nuestra familia. Gracias también a Eduardo, Ana Julia, Eugenio y Óscar. Ellos representan a mis familiares, amigos, amigas y todos/as quienes me han acompañado, de diversas maneras y en distintos momentos, con su cariño y preocupación. De manera especial, quisiera expresar mi gratitud y admiración por mis compañeros de generación, tremendos lingüistas y mejores personas, con quienes compartí durante estos años en el doctorado: Claudio, Karina, Dania y Diana.

Para terminar, espero con mucha ilusión que esta tesis pueda ser un aporte para el fomento del trabajo interdisciplinar y la generosidad intelectual entre lingüistas e informáticos.

Fredy Núñez Torres, julio 5 de 2021.

Capítulo 1
Introducción

La lingüística computacional surge durante la década de 1940 como un ámbito de trabajo orientado hacia el desarrollo o construcción de sistemas de traducción automática. Se definía a sí misma como el estudio de los procedimientos computacionales necesarios para la comprensión y generación automática de las lenguas naturales. Según Espunya & Prat (1994), el objetivo de los lingüistas computacionales es escribir programas que puedan manejar (comprender o generar) tanto material de lenguaje natural como sea posible. Estos programas serían soluciones eficientes pero aproximadas; no podrían ocuparse de todas las estructuras de una lengua natural, aunque se hicieran cargo de las construcciones más comunes e interesantes. Desde esta perspectiva, el lenguaje humano era concebido como un proceso comunicativo en el que un emisor y un receptor son capaces de procesar determinada información en función de un conocimiento lingüístico común. La idea que subyace era que un objeto matemático y computable podría constituir una modelización de un fenómeno lingüístico y, por lo tanto, corresponder a un sustituto de una lengua natural; es decir, a una imitación parcial de la manifestación de una capacidad cognitiva humana, innata y propia de nuestra especie. Según lo anterior, las lenguas naturales:

> "[…] se adquieren «de forma espontánea, inconsciente e inadvertida», sin que sea necesaria una ayuda pedagógica dirigida; este aprendizaje «totalmente natural, sin estudio consciente», sin embargo, solo se produce durante un determinado período crítico de maduración […] y sus hablantes «nunca se plantean ninguna cuestión relativa a la estructura gramatical de su idioma» […] dependen de las posibilidades y limitaciones de la anatomía y fisiología humanas." (Moreno-Cabrera, 2013: 50).

En este sentido, el objetivo más relevante de la lingüística computacional es reproducir los patrones con los que funcionan la mente y el lenguaje humanos, para transferirlos a la relación entre humanos y máquinas.

Más adelante, durante los años 50 del siglo XX, logran desarrollar los primeros sistemas de traducción automática, que fueron abordados inicialmente desde una perspectiva únicamente informática, sin tomar en cuenta aportes desde la lingüística, lo que no tuvo buenos resultados. Esto llevó a la necesidad de integrar modelos lingüísticos que permitieran resolver estos problemas. Otros hitos que sucedieron a este fueron el desarrollo de sistemas de diálogo interactivos, métodos de

aprendizaje automático y sistemas capaces de inferir conocimiento lingüístico a partir de algoritmos (Periñán-Pascual, 2012). Desde este punto de vista, que es por cierto más actual, la lingüística computacional se entiende como una disciplina científica e interdisciplinaria que estudia el lenguaje desde una perspectiva computacional, considerando el desarrollo de sistemas automatizados para la comprensión y producción de las lenguas naturales.

El procesamiento del lenguaje natural (en adelante PLN), por su parte, es una rama aplicada de la lingüística computacional (o de la inteligencia artificial). Específicamente, el PLN tiene como fin estudiar, diseñar y aplicar sistemas informáticos que faciliten la comunicación entre personas y entre personas y máquinas, logrando que esta comunicación sea flexible, eficiente y fluida (Periñán-Pascual & Mairal-Usón, 2009; Choi *et al*., 2020). Su objetivo es imitar artificialmente algunos de los aspectos de la capacidad humana para el lenguaje, lo que se traduce en procesos de producción y comprensión. Desde una perspectiva predominantemente informática, el PLN se puede definir como el análisis de datos lingüísticos utilizando métodos computacionales. Según lo anterior, el objetivo del PLN es generalmente construir una representación del texto que agregue estructura al lenguaje natural no estructurado, aprovechando los conocimientos de la lingüística. Algunos de sus productos podemos verlos actualmente en sistemas de traducción automática (*Google Translate*, *DeepL*), o asistentes virtuales (*Siri*, *Cortana*, *Alexa*, etc.).

Un debate tradicional, pero que continúa hasta nuestros días, es a qué ámbito de estudios pertenece el PLN: a la informática o a la lingüística, lo que ha tenido consecuencias en la manera en que estas dos diferentes perspectivas han abordado sus problemas. Por un lado, el PLN se considera como un ámbito de la lingüística aplicada, en la medida que permite desarrollar aplicaciones orientadas hacia mejorar algún aspecto de la vida social a través del conocimiento y/o la interacción lingüística. Por otro lado, se considera el PLN como una rama aplicada de la inteligencia artificial, en la medida que integra un proceso de modelización matemática para tratar computacionalmente una lengua natural, y así generar procesos de comunicación entre máquinas y humanos.

Otra polémica bastante actual tiene que ver con los espacios de colaboración entre informáticos y lingüistas. Conocida es la anécdota protagonizada por Frederick Jelinek (1932-2010), quien fuera director del Grupo de Investigación de IBM para el Reconocimiento Continuo de la Voz (*IBM Continuous Speech Recognition Group*). Cierto día, uno de los lingüistas del equipo renunció. Entonces, Jelinek decidió reemplazarlo no por otro lingüista, sino por un ingeniero. Poco después, notó que el rendimiento de su sistema mejoró significativamente. Luego alentó a otro lingüista a

buscar un empleo alternativo y, efectivamente, el rendimiento mejoró nuevamente. Fue ahí cuando, según varios testigos, pronunció su célebre frase: "Cada vez que despedimos a un lingüista, la eficiencia de nuestro sistema mejora"[1]. El fundamento de este problema radica en las diferentes maneras que ambas disciplinas tienen para acercarse al mismo fenómeno. La informática, típicamente, pone el énfasis en los procedimientos necesarios para la solución de un problema específico y, por tanto, esas soluciones se transforman en el foco de atención y evaluación científica. Por el contrario, la lingüística ha sido capaz de construir modelos teóricos robustos, que se focalizan en distintos aspectos del lenguaje humano. Por tanto, el centro de la discusión no estará puesto en la manera de abordar un problema, ya sea teórico o aplicado, sino en la naturaleza misma del problema, en su descripción y explicación.

Precisamente, en el ámbito de la traducción automática, el objetivo de cualquier sistema es determinar la opción correcta de significado para una palabra, a partir de un número finito de significados posibles previamente almacenados. Las computadoras no poseen la capacidad inherente de procesar lenguas naturales y, por lo tanto, no pueden reconocer casos de ambigüedad a menos que cuenten con mecanismos específicos para llevar a cabo esta tarea. De acuerdo con lo anterior, cualquier sistema de desambiguación implica el proceso de hacer coincidir el contexto de la aparición de la palabra, ya sea con la información de una fuente externa de conocimiento o con la información sobre los contextos de casos previamente desambiguados de la palabra que se derivan de un corpus. En resumen, esta variedad de métodos de asociación se utiliza para determinar la mejor coincidencia entre el contexto dado y una de estas fuentes de información, con el objetivo de asignar un significado a cada ocurrencia.

La motivación principal de esta propuesta de investigación es el abordaje de la ambigüedad léxica desde una perspectiva lingüística, para luego proponer el desarrollo de un modelo formal para la desambiguación léxica automática que pueda ser aplicado a una herramienta para el PLN. No es nuestra intención que esta tesis se configure como un esfuerzo únicamente lingüístico, o exclusivamente informático, sino que hemos apostado por el aporte que constituye para el quehacer lingüístico el conocimiento técnico en las áreas de las ciencias de la computación, la estadística y la minería textual. De esta manera, es posible favorecer el análisis de grandes volúmenes de datos textuales, a la vez que se fomenta la colaboración de lingüistas en investigaciones interdisciplinares,

[1] "Every time we fire a linguist, the performance of our system goes up" (Frederik Jelinek).

junto con el desarrollo de soluciones de ingeniería lingüística que puedan ser aplicadas en distintos ámbitos de la vida social.

Según lo anterior, el objetivo principal de esta tesis es desarrollar un modelo más robusto de medida para la similitud y relación semántica que los disponibles actualmente para la desambiguación léxica automática, aplicado al PLN.

La organización de nuestro trabajo es la siguiente:

En el capítulo dos se expone un panorama general de los aportes de las ciencias del lenguaje en el ámbito del PLN. Específicamente, se realiza una revisión del problema lingüístico de la ambigüedad léxica, junto con los métodos para la desambiguación léxica automática más representativos, y que han sido aplicados en el área del PLN.

En el capítulo tres se ofrece una revisión de algunos conceptos fundamentales relacionados con el tratamiento y procesamiento de datos textuales, mediante una panorámica cronológica de la utilización del corpus y los recursos informatizados en el análisis lingüístico. Además, se revisa la arquitectura de representación del conocimiento que propone FunGramKB (*Functional Grammar Knowledge Base*). Por último, se describe la tecnología informática básica que suele emplearse en la construcción y exploración de recursos lingüísticos informatizados, y que de hecho se utilizó para el montaje del corpus durante el desarrollo de esta investigación: expresiones regulares, lenguaje de etiquetado extensible (XML), y lenguaje de consulta estructurada (SQL).

En el capítulo cuatro se establece la pregunta de investigación que guiará esta tesis, junto con el objetivo general y sus respectivos objetivos específicos.

En el capítulo cinco se presentan los aspectos metodológicos de la investigación. En primer lugar, se establecen los subtipos generales de procesamiento de datos en aprendizaje automático, con sus respectivas tareas de procesamiento. Luego se reporta el experimento preliminar, basado en el corpus SENSEVAL-3 (*Evaluating Word Sense Disambiguation Systems*). En tercer lugar, se expone el procedimiento para el montaje de una colección de documentos basada en la selección conceptual de FunGramKB y en una submuestra del *corpus* CODICACH (Corpus Dinámico del Castellano de Chile). Por último, se describen las tareas de procesamiento específicas para la realización de experimentos de aprendizaje automático utilizando la herramienta DAMIEN (*Data Mining Encountered*).

En el capítulo seis se reportan los resultados de los experimentos de aprendizaje automático basados en la colección de documentos elaborada desde de la submuestra del c*orpus* CODICACH, a partir de la metodología expuesta en el capítulo cinco.

En el capítulo siete se presenta la propuesta de un modelo de desambiguación léxica automática basado en una medida híbrida, fundamentado en la interacción de dos enfoques de exploración taxonómica: distancia entre rutas y contenido de información.

Finalmente, en el capítulo ocho se presenta la evaluación del modelo de desambiguación léxica automática propuesto, Primero, se exponen las variables necesarias para el cálculo de la medida híbrida, cuyos valores han sido extraídos desde la base de conocimiento FunGramKB. Segundo, se proponen tres casos de evaluación junto con sus respectivos resultados, considerando todos los pasos necesarios para la aplicación de la medida de similitud $SIM_{\text{híbrida}}(c_i, c_j)$.

Capítulo 2
Estado de la cuestión

El objetivo del siguiente capítulo es exponer un panorama de los aportes de las ciencias del lenguaje en el ámbito del PLN. Específicamente, se realiza una revisión del problema lingüístico de la ambigüedad léxica, junto con los métodos para la desambiguación léxica automática más representativos y que han sido aplicados en el área del PLN.

2.1 El aporte de las ciencias del lenguaje al PLN

Desde los inicios de los estudios en inteligencia artificial ha existido una tensión en cuanto a la integración de las disciplinas de la informática y la lingüística. La informática ha propiciado esencialmente el desarrollo de modelos estocásticos (o probabilísticos), que se caracterizan por la aplicación de técnicas matemáticas sobre un gran volumen de datos textuales con el objetivo de inferir conocimiento lingüístico (Espunya i Prat, 1994; Allen, 1995; Cantos-Gómez, 1996). Estas implementaciones no almacenan conocimiento lingüístico (o conceptual), sino que aplican determinadas técnicas matemáticas sobre corpus textuales con el fin de extraer conocimiento. Así, estos sistemas estocásticos son capaces de inferir conocimiento lingüístico a través de la utilización de algoritmos. En definitiva, se trata de una construcción automatizada del conocimiento, basada en una aproximación computacional al análisis de textos como volúmenes de información computable.

Este tipo de modelos consideran las lenguas naturales como un conjunto de sucesos que presentan una determinada frecuencia. Esto quiere decir que cada unidad lingüística (sea un morfema, palabra, sintagma o cualquier tipo de categoría, ya sea morfológica o sintáctica), presenta una probabilidad específica de manifestarse o aparecer en un contexto oracional acotado (Chomsky, 1988; Jackendof, 2002). Así, dado que esta perspectiva depende de la calidad y cantidad de la información almacenada en un corpus lingüístico, mientras mayor sea el número de datos utilizados, mejor se comportará el modelo, cualquiera sea este. Un ejemplo representativo y bastante actual de la aplicación de los métodos estadísticos son los modelos basados en el aprendizaje automático (*machine learning*). Se trata de un ámbito de estudio de la inteligencia artificial cuyo objetivo es el desarrollo de algoritmos que sean capaces de representar eficientemente determinados conjuntos de datos (Choi *et al*., 2020). Específicamente, se espera poder determinar automáticamente la probabilidad de que a una unidad lingüística se le asigne un valor predefinido como correcto. Los métodos de aprendizaje

automático que se utilizan con mayor frecuencia son: supervisado, no supervisado, y de aprendizaje por refuerzo.

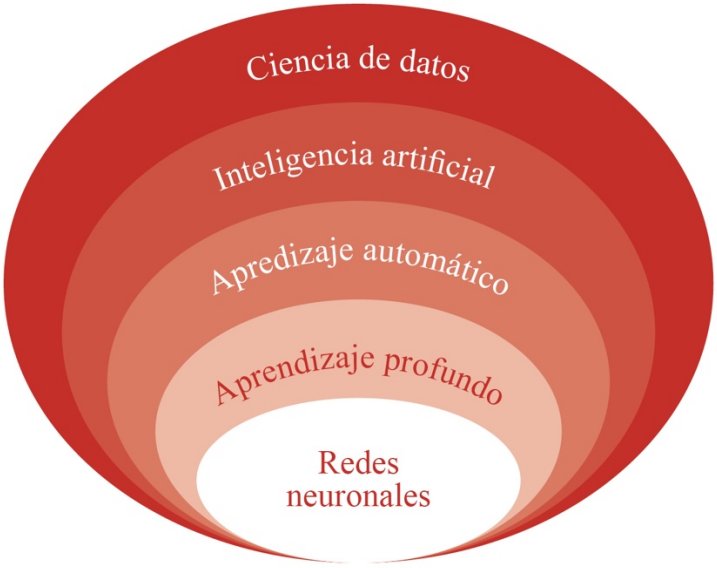
El aprendizaje supervisado determina patrones en un corpus de entrenamiento, con el objetivo de mapear atributos que servirán como conjunto de datos a partir de los cuales realizar predicciones en un nuevo corpus. Entonces, se le denomina supervisado porque el modelo es capaz de inferir información a partir de un algoritmo y un conjunto de datos previamente etiquetado, y transferir sus características a una predicción (Moor, 2006).

En el caso del aprendizaje no supervisado, la detección de patrones en un conjunto de datos se realiza a través de un algoritmo sin la necesidad de haber incorporado información previa. Esta técnica se emplea predominantemente para las tareas de agrupación, asociación y detección de anomalías (Hastie *et al*., 2009; James *et al*., 2013). Una explicación más detallada de las técnicas de aprendizaje supervisado y no supervisado se propone hacia el final del capítulo, en coherencia con la naturaleza de los experimentos desarrollados a lo largo de esta investigación.

En cuanto al aprendizaje por refuerzo, se trata de una técnica en que la máquina, o agente, recibe una valoración basada en el desempeño de la tarea que ha realizado. En términos generales, se establece un proceso de decisión en el que el agente debe ser capaz de seleccionar una acción determinada que pueda maximizar la obtención de una recompensa, a partir de un determinado comportamiento o estrategia (Sutton & Barto, 1998; Russel & Norvig, 2002). Específicamente, se espera que el sistema explore su entorno y observe los resultados de determinadas acciones que le permitirán obtener datos a partir de los cuales aprender, puesto que este proceso se realiza sin conocimiento previo de una opción correcta, es decir, mediante el ensayo y error (López-Boada *et al*., 2005). Una de las ventajas de esta técnica es que, a diferencia del aprendizaje supervisado y el no supervisado, no se limita a realizar tareas de clasificación o predicción, sino que puede utilizarse para propiciar la interacción en diferentes ambientes. Por ejemplo, el aprendizaje por refuerzo se aplica frecuentemente en videojuegos, sistemas de navegación, servicios de retransmisión de contenido multimedia, etc., para predecir las preferencias de sus usuarios a partir de los datos proporcionados por su registro de opciones.

La siguiente figura, tomada de (Choi *et al*., 2020), representa el conjunto de técnicas de aprendizaje automático en el marco de la ciencia de datos. De esta forma, la inteligencia artificial como disciplina contiene las diferentes técnicas de aprendizaje automático, incluidos el aprendizaje profundo y el modelo de redes neuronales.

Figura 1. Selección de técnicas en ciencia de datos (tomada de Choi *et al*., 2020).



Para ilustrar lo anterior de manera eficiente, volvamos nuevamente al caso de la desambiguación léxica automática. Esta requiere una automatización del proceso de asociación entre cierta palabra en un texto, y una definición o significado entre varios significados potencialmente atribuibles a esa palabra. Esta tarea se puede dividir en dos subprocesos: la discriminación de sentidos y el etiquetado de sentidos. En el primer subproceso, el objetivo es dividir las ocurrencias de una palabra en un número determinado de clases para dos ocurrencias, pertenezcan ellas al mismo sentido o no. En el segundo subproceso, la meta es asignar un sentido a cada clase previamente determinada, para luego etiquetar la aparición de cada una de las palabras ambiguas (Yarowsky, 1995; Mooney, 1996; Schütze, 1998). En el caso del aprendizaje automático supervisado, se trata del desarrollo de algoritmos que utilizan como fuente de conocimiento un corpus entrenado o etiquetado previamente. Los algoritmos comúnmente utilizados son: método de los *k*-vecinos más próximos (Cunningham & Delany, 2007; Zhang, 2016), algoritmo bayesiano ingenuo (Langley *et al*., 1992; Gale *et al.*, 1992; Manning & Schütze, 1999; Widlak, 2004), árboles de decisión (Hancock *et al*., 1996; Rokach & Maimon, 2002), regresión lineal (Bishop, 2006; Van Le *et al*., 2018), y máquinas de vectores de soporte (Joachims, 2005; Dien *et al*., 2019). Los experimentos de base para la evaluación de la desambiguación léxica automática que propone este trabajo se han desarrollado utilizando el algoritmo bayesiano ingenuo, que será explicado de manera pormenorizada más adelante.

En cuanto al tratamiento de un corpus lingüístico para ser utilizado en aprendizaje automático, este debe ser etiquetado/anotado a partir de determinadas características. En el caso de la información lingüística, típicamente se selecciona una palabra potencialmente ambigua, además de otras palabras que ocurren en el mismo contexto oracional. También es necesario, en casi todos los casos, la definición de una ventana contextual de palabras (por ejemplo, el conteo de un número determinado de palabras hacia adelante y hacia atrás a partir de la posición de la palabra que constituye el objeto del estudio). Dado que esto solo da cuenta superficialmente de un contexto a nivel léxico, se considera una representación de bolsa de palabras (*bag of words*). Cabe señalar que esta opción metodológica no toma en consideración la estructura sintáctica, puesto que el objetivo es ejecutar técnicas de minería textual para la extracción rápida y sencilla de predicciones, a partir de datos textuales no estructurados (Salton & McGill, 1986; Zhang *et al*., 2015; Soumya & Shibili, 2014; Deepu *et al*., 2016). Así, en la medida que cada instancia de la palabra en análisis haya sido previamente desambiguada en un corpus entrenado, un conteo de frecuencia de cada palabra en el contexto oracional servirá de pista para automatizar la desambiguación. Finalmente, utilizando como insumo estos datos etiquetados, el algoritmo calculará probabilidades condicionales que determinen el sentido más probable para la palabra en análisis.

Desde otro punto de vista, las ciencias del lenguaje han favorecido el desarrollo de modelos simbólicos (también conocidos como axiomáticos o lingüísticos) que, a diferencia de los estocásticos, almacenan conocimiento lingüístico en forma de esquemas de representación del conocimiento. Desde esta perspectiva, la mente humana se entiende como una máquina modular, es decir, como un conjunto de estrategias cognitivas de comprensión y pensamiento. El objetivo es reproducir los patrones con los que funciona la mente humana mediante un sistema basado en una teoría lingüística. El enfoque simbólico se especializa en la construcción o desarrollo de aplicaciones, para la automatización de tareas específicas, que almacenan conocimiento lingüístico (i.e. esquemas de representación del conocimiento). Esta perspectiva ha propiciado el desarrollo de teorías orientadas a la descripción, explicación y formalización de las propiedades lingüísticas de diferentes lenguas, como recursos para la representación del conocimiento lingüístico. Algunos ejemplos representativos de estos modelos son: la Gramática Generativa Transformacional (Chomsky, 1957; 1964) y la Semántica de Marcos (Fillmore, 1976; 1982; Fillmore & Atkins, 1994), como antecedente teórico del proyecto FrameNet (Baker *et al*., 1998; Petruck & de Melo, 2012); la Gramática de Construcciones (Fillmore, 1988; Goldberg, 1995; Kay & Fillmore, 1999) y, específicamente, la Gramática de Construcción

Incorporada (Feldman *et al*., 2010), en el marco del desarrollo de un analizador automático de construcciones (Bergen & Chang, 2003); por último, la Gramática del Papel y la Referencia (Van Valin & LaPolla, 1997; Van Valin, 2005) y el Modelo Léxico-construccional (Ruiz de Mendoza & Mairal-Usón, 2008; Mairal-Usón & Ruiz de Mendoza, 2009), integrados en la base de conocimiento léxico conceptual *FunGramKB* (Periñán-Pascual & Arcas-Túnez, 2004; 2010; Periñán-Pascual, 2012a). Este último modelo, que además ha sido pionero en el ámbito de la representación del conocimiento, se abordará de manera detallada en los capítulos posteriores.

Este tipo de aproximaciones pretenden reflejar de manera formal, unívoca y bien delimitada, la estructura lógica del lenguaje. Para esto, desarrollan sistemas compuestos por unidades y por reglas que establecen las combinaciones posibles entre las unidades. Así, se postulan las relaciones que se pueden establecer entre los distintos elementos. El producto de esta última perspectiva ha sido el desarrollo de lexicones automatizados y ontologías. En efecto, para trabajar en el ámbito de estos modelos, es necesario desarrollar un lexicón vinculado con representaciones semánticas, que pueda a su vez relacionarse, mediante algún algoritmo de enlace, con representaciones sintácticas. En otras palabras, se trata de poder determinar estructuras semánticas que permitan predecir estructuras sintácticas. Para ejecutar esta serie de mecanismos es necesaria la utilización de un lenguaje de programación que permita al humano describir una serie de instrucciones y reglas (algoritmos) a la máquina, tanto para extraer como para ingresar información. De hecho, en los sistemas simbólicos, parte de la información conceptual se ingresa manualmente.

Si bien los métodos estocásticos han demostrado ser altamente eficientes, no son consecuentes con el mayor objetivo de la inteligencia artificial: imitar el lenguaje, entendido como una capacidad cognitiva (Grishman, 1986; Chowdhury, 2005). Esta falencia se hace evidente al tratar de mantener un diálogo fluido y espontáneo con un asistente virtual como *Siri*, *Alexa* o *Cortana* (Hoy, 2018). Desde una perspectiva simbólica, la mayor crítica que se puede hacer a los métodos estocásticos es que reducen las capacidades lingüísticas a decisiones probabilísticas basadas en algoritmos, lo que no parece coincidir con la manera en que funciona el lenguaje humano. Según Manaris (1998), existen diferentes habilidades lingüísticas que una máquina debería desplegar en la interacción entre humanos y sistemas artificiales: ser, por ejemplo, capaz de establecer relaciones entre el contenido conceptual que se encuentra en las diferentes palabras, de inferir información a partir de esas relaciones, y de actuar en diferentes entornos y con múltiples propósitos.

Típicamente, los lingüistas han tenido integraciones periféricas en proyectos relacionados con el PLN. Este tipo de colaboración ha sido constante desde los primeros intentos por desarrollar tanto traductores como diccionarios electrónicos, hasta estos días. Los profesionales del lenguaje han sido relegados a convertirse en proveedores de información para poblar de manera eficiente diversas bases de datos léxicas o sintácticas. En este sentido, el aporte más sistemático y constante que han realizado los lingüistas en el desarrollo de proyectos de PLN ha sido la recopilación de corpus lingüísticos. Con la irrupción de las ciencias informáticas aplicadas a la recopilación de información lingüística a partir de la década de 1960, y contemporáneo a la hegemonía del programa generativista en la teoría lingüística, comienza el desarrollo del primer corpus legible por la máquina: el *Corpus Brown* (Francis, 1965)[2].

Convencionalmente, se define corpus como una muestra amplia de lengua escrita o hablada que se considera representativa de alguna variante diatópica de una lengua, o bien de un período histórico determinado (McEnery & Wilson, 1996; Meyer, 2002). Esta aproximación se puede expandir hacia el ámbito computacional como una colección finita de textos legibles por la máquina que son representativos de una lengua. La proliferación de sistemas informatizados para el análisis lingüístico derivó, a partir de la década de 1970, en la utilización de corpus para investigaciones de PLN que involucraron modelar el comportamiento lingüístico. Junto con esto, el potencial computacional para la recopilación de información lingüística dio paso a diferentes tipos de recursos lingüísticos informatizados (Amsler, 1986; Amsler & Whim, 1979; Wilks *et al*., 1988). Estos tipos corresponden a corpus (monolingüe o bilingüe), lexicón, glosario, taxonomía y ontología[3].

Esta perspectiva sugiere entonces que el corpus, dada su aplicación en el desarrollo de herramientas computacionales, se establezca en realidad como un término parcial, o subtipo de una categoría más inclusiva y apropiada: los recursos lingüísticos informatizados. Estos, a su vez, se definen como cualquier conjunto de datos lingüísticos, oral o escrito, que se construya en un formato ya sea legible o además tratable por un ordenador. En esta línea, los recursos lingüísticos informatizados pueden pertenecer a determinados tipos de conjuntos de datos lingüísticos, entendidos

[2] El *Corpus Brown* (*The Brown University Standard Corpus of Edited Present-Day American English*) es el primer corpus textual del inglés americano. Fue publicado entre 1963 y 1964 por Winthrop Nelson Francis y Henry Kučera, en el Departamento de Lingüística de la Universidad de Brown. Contiene un millón de palabras (500 muestras de más de 2.000 palabras cada una) de texto en prosa editado e impreso en Estados Unidos durante el año 1961.

[3] Una revisión pormenorizada para cada uno de los tipos de conjuntos de datos lingüísticos, entendidos como recursos lingüísticos informatizados, se realiza en el punto 3.2 del capítulo 3: Conceptos fundamentales.

como formas de clasificación de conocimiento. En los capítulos posteriores de este trabajo se planteará una clasificación de los recursos lingüísticos informatizados.

Actualmente, basándonos en Periñán-Pascual (2012a), los aportes del lingüista al desarrollo de sistemas del PLN estarían posicionados principalmente en tres grandes ámbitos. Primero, la descripción formalizada de la estructura morfológica, sintáctica y/o semántica de las lenguas naturales. Este tipo de aporte consiste en un ámbito de trabajo que no considera una implicación directa del lingüista en el diseño o desarrollo de aplicaciones; es decir, se trata de un entorno de investigación reducido a la formalización de regularidades gramaticales que puedan ser luego expresadas por medio de reglas o metalenguajes.

Segundo, el montaje y/o anotación de diversos recursos lingüísticos informatizados, que comprende la construcción de lexicones computacionales que permiten organizar la formalización del conocimiento gramatical de una lengua natural, junto con el etiquetado de diversos recursos lingüísticos informatizados con información relevante para el desempeño de un sistema automatizado.

Tercero, y como ámbito más relevante de colaboración, la construcción de ontologías entrega al lingüista la posibilidad de explotar la organización terminológica y la representación del conocimiento semántico, en el marco de modelos de análisis lingüístico con potencial computacional. Así, el tránsito hacia a la integración plena del lingüista y la perspectiva interdisciplinar del ingeniero del conocimiento implica un diálogo entre distintos ámbitos de conocimiento, predominantemente la lingüística y la informática, para lograr un balance entre los métodos estocásticos y simbólicos (Hatzivassiloglou, 1994; Meeter & Gish, 1994). Esto requiere, por parte de los lingüistas, una formación técnica en aspectos del PLN: lenguajes de programación, gestión de bases de datos, y otros conocimientos informáticos básicos que le permitan desempeñar un papel relevante en el diseño de aplicaciones. Asimismo, los informáticos requieren conocimiento analítico de las lenguas naturales en un nivel que les permita hacer más eficientes sus propuestas para la representación del conocimiento lingüístico.

2.2 El problema lingüístico de la ambigüedad léxica

Desde una perspectiva general, la ambigüedad lingüística correspondería a "[l]a propiedad de una expresión que puede tener más de un significado" (Escandell, 2004: 337). En este sentido, la ambigüedad describe el problema de la correspondencia entre unidades léxicas u oraciones, con sus respectivas condiciones de verdad. Desde un punto de vista más específico, el problema de la

ambigüedad ha sido abordado por la semántica como "[…] el hecho de que las situaciones en las que se emplea una expresión puedan encauzar su significación en direcciones diferentes" (Ducrot & Todorov, 1972: 275). Según lo anterior, se considerará una proposición como ambigua si es que contiene dos o más sentidos que hacen referencia a estados de cosas del mundo diferentes. A continuación, según los planteamientos de Del Teso (2002) y Escandell (2004), se exponen tres tipos de ambigüedad y sus correspondientes ejemplos. Por una parte, la *ambigüedad sintáctica* (o estructural), que ocurre debido a la existencia de dos análisis sintácticos diferentes para una misma realización lingüística, de la que se infieren dos interpretaciones semánticas diferentes. Por ejemplo:

[S [NP El atleta] [VP [V subió] [PP a [NP el podio] [AP adornado [PP con [NP flores]]]]]][4]

En la oración anterior, la frase adjetiva corresponde a la estructura que propicia la ambigüedad, el proceso de desambiguación depende de una asignación de referencia y esta, a su vez, estará dada en principio por la cercanía de las estructuras. Entonces, la estructura [AP adornado [PP con [NP flores]]] determina a [NP el podio] por su cercanía. Por el contrario, la representación coherente de un estado de cosas del mundo en el que «el atleta» haya estado «adornado con flores» debería responder a este patrón de asignación, con la expresión [[NP El atleta] [AP adornado [PP con [NP flores]]]]. En este caso particular, la subordinación del complemento está asignada al nominal «atleta». Al respecto, Murphy (2010) señala que, a diferencia de la explicación anterior, en el proceso de desambiguación los componentes semánticos de un lexema son capaces de proyectar significados hacia una estructura determinada. Estos significados particulares estarían especificados por la compatibilidad de la entrada léxica para ser modificada. En este caso, el proceso de desambiguación dependerá de la compatibilidad del significado léxico, que es necesaria para la elección de una estructura sintáctica particular por sobre otra. Por ejemplo, en el mismo caso referido, la interpretación estará dada por la compatibilidad, en mayor o menor medida, de los significados nominales de [NP El atleta] y [NP el podio]. De esta manera, en un estado de cosas del mundo relativamente estable, puede ser más frecuente que sea el «podio» el que se encuentre «adornado con flores» al momento en que un «atleta» sube. Esta perspectiva tiene una estrecha relación con el efecto de los prototipos, en cuanto propicia la asignación de significado a partir de estructuras típicamente relacionadas.

[4] Esta notación es una adaptación del modelo de árboles sintácticos de Jackendoff (2002).

Por otra parte, la *ambigüedad semántica* ocurre debido a la existencia de dos o más representaciones semánticas para una misma representación lingüística; como por ejemplo en «Todos los chicos de una clase están enamorados de una chica», donde es posible determinar que todos los chicos de una clase están enamorados de una chica diferente, o bien que hay una única chica de la que están enamorados todos los chicos de la clase. Para la comprensión de este fenómeno, es fundamental considerar el principio de composicionalidad como un dispositivo que regula las interpretaciones posibles tanto de una unidad léxica como de la manera en la que se combinan distintas unidades para formar expresiones complejas. Este principio de composicionalidad (Escandell, 2004; Garrido-Medina, 1994) propone que existe también un significado presente en la estructura sintáctica, y no sólo en cada unidad léxica; es decir, el significado de una expresión compleja corresponderá a la función del significado de cada una de las unidades léxicas que la compongan, más el significado de la combinatoria de ellas. Por lo tanto, es posible acceder al significado de una expresión compleja a partir de la interpretación sucesiva de sus componentes y de las elecciones combinatorias que se manifiestan mediante determinadas reglas.

Finalmente, la *ambigüedad léxica* se define como el fenómeno en el que una entrada léxica puede contener dos o más significados diferentes. En particular, esta investigación se focaliza en la resolución de la ambigüedad léxica para los casos de homonimia y polisemia, que serán explicados más adelante, como la caracterización del fenómeno en análisis desde una perspectiva informática. Como ejemplo preliminar, la ambigüedad léxica ocurre debido a la existencia de dos significados denotativos para una misma etiqueta lingüística, como en «La gata está sobre el cobertizo», donde «gata» puede referir tanto al felino doméstico como a la herramienta levantapesos.

En cuanto a la perspectiva tradicional para la descripción de la ambigüedad léxica, Ide & Véronis (1998) la caracterizan como la asociación entre un ítem léxico determinado en un texto o discurso y un sentido que pueda ser distinguible de otros significados potencialmente atribuibles a ese ítem léxico. Según lo anterior, algunos de los autores clásicos en el ámbito de la semántica y la lingüística teórica, como Lyons (1977) y Cruse (1986), distinguen dos tipos generales de ambigüedad léxica en tanto se le reconoce como un fenómeno no uniforme: homonimia y polisemia. Estas distinciones han sido actualizadas principalmente por Pustejovsky (1991; 1995), y Pustejovsky & Boguraev (1996). En primer lugar, la homonimia corresponde a un fenómeno en el que un ítem léxico contiene accidentalmente dos o más significados. También se le ha llamado *ambigüedad contrastiva.*

Un ejemplo para este caso es la unidad «velas» en las siguientes proposiciones:

(1)

a. El contramaestre mandó izar las *velas*.
b. Los monjes lograron encender las *velas*.

En el ejemplo (1a), «velas» se refiere a los 'instrumentos que capturan la fuerza del viento en una embarcación'; o bien, en (1b), a los 'artefactos hechos de cera con una mecha que sirven para proveer de iluminación'. En segundo lugar, la polisemia[5] se refiere a los casos en los que un mismo lexema puede tener múltiples significados relacionados. Un ejemplo de esto es la relación entre los distintos usos del verbo «abrir», como en las siguientes construcciones:

(2)

a. El niño *abrió* el debate.
b. El niño *abrió* la lata.

En el primer caso (2a), la apertura supone el inicio de una acción comunicativa, mientras que en el segundo (2b) implica la apertura de un objeto físico. Si bien en ambos casos «el niño» funciona como movilizador de la acción, o agente, los sentidos de «abrir» para las construcciones sintácticas propuestas difieren.

A continuación, se abordan distintos enfoques para la descripción de la ambigüedad léxica basados en la representación de los sentidos de una unidad léxica cuyas relaciones pueden ser formalizables y, por tanto, potencialmente compatibles con descripciones a la vez lógicas y lingüísticas.

2.2.1 El enfoque descriptivo

La mayoría de los diccionarios monolingües son representantes de este enfoque convencional y de carácter intuitivo para la definición de los sentidos de palabras. Los significados, a partir de esta aproximación, se definen a través de condiciones necesarias y suficientes que son construidas a partir

[5] Llamada también polisemia complementaria en los trabajos de Weinreich (1964), y en las primeras versiones de los trabajos de Pustejovsky (1991; 1995).

de un número determinado de rasgos previamente definidos, mientras que el sentido corresponde al modo de presentación de un referente, que a su vez depende de las relaciones que posee un a expresión lingüística determinada con otras expresiones en el sistema lingüístico. Ambos conceptos, tanto sentido como significado, se relacionan a partir de la designación de una referencia, entendida como el acto que lleva a cabo un hablante para referir a una entidad o situación por medio de una expresión lingüística (Frege, 1973 [1892]; Putnam, 1975; Rivano, 2002). Si estas condiciones, propias de una definición componencial, se cumplen para una entidad, es posible acceder a la información lexicalizada de una palabra. Este enfoque, no obstante, presenta un sesgo relevante dada la condición estática de los rasgos que constituyen el repertorio discreto de sentidos de una palabra. Por lo tanto, los significados se reducen a listas de rasgos descontextualizados de su definición y del contexto oracional en el que esas unidades léxicas se utilizan frecuentemente. Por ejemplo, el *Longman Dictionary of Contemporary English* restringe las palabras que se emplean en sus definiciones a un vocabulario de aproximadamente 2.000 palabras. Además, utiliza una gramática controlada para la sintaxis de sus definiciones (Procter, 1978). A pesar de lo anterior, el enfoque descriptivo es eficiente y ampliamente utilizado para la representación general de sentidos de palabras.

2.2.2 El enfoque relacional

El enfoque relacional define los significados de los conceptos en términos de relaciones semánticas. Según este enfoque, los significados de los signos lingüísticos se caracterizan como conceptos léxicos; es decir, como una estructura conceptual que corresponde a una unidad léxica en una lengua determinada (McRae *et al*., 1997; Viglioco & Wilson, 2005). Esta aproximación se basa en la premisa de que el léxico lingüístico está organizado por relaciones de implicación y no sólo por la oposición de rasgos semánticos. En este sentido, la propuesta relacional surge a partir de la influencia del paradigma conexionista para la descripción del lexicón mental, propuesto originalmente por Collins & Loftus (1975), y McClelland & Rumenhart (1981). Según estos autores, los sentidos disponibles de una unidad léxica estarían basados en una jerarquía conceptual; es decir, en su coocurrencia dentro de un corpus representado como sistema de categorías que están relacionadas unas con otras. Si bien el enfoque relacional, al igual que el descriptivo, define el significado como una agrupación finita de información, se trata de una perspectiva mucho más explícita en cuanto a la caracterización de las relaciones que se establecen entre determinadas agrupaciones de conceptos. Esta aproximación al problema del significado se ha aplicado en el ámbito de la lingüística computacional a partir de los

modelos de semántica distribucional (Harris, 1954; Firth, 1957), como una representación computacional del significado en espacios vectoriales multidimensionales, donde cada unidad léxica funciona como un vector, a través del análisis estocástico de los contextos oracionales en que cada unidad léxica ocurre (Martí-Antonín, 2018).

2.2.3 El enfoque generativo

Este enfoque está basado en el modelo del Lexicón Generativo propuesto por Pustejovsky (1991; 1995). Según esta teoría, el léxico se define como un componente dinámico del sistema lingüístico, cuya función es ser depósito de la potencialidad significativa y creativa del lenguaje. Se trata de categorías gramaticales entendidas desde su capacidad sígnica para intervenir en diferentes estructuras sintácticas, así como en distintos mecanismos de composición semántica. Específicamente, Pustejovsky (1995) propone caracterizar la ambigüedad como una proyección bivalente de significado que podría expresarse en distintos niveles: un lexema, una construcción sintáctica o la proyección de cualquiera de estas estructuras en la representación de un estado de cosas del mundo particular. De esta manera, la ambigüedad léxica se puede definir como la asociación entre un ítem léxico con más de un significado en el sistema de la lengua (Chiernia & McConnel-Ginet, 1990). Este es el caso del ítem «cordero» en las siguientes proposiciones:

(4)

a. El *cordero* corre por el campo.
b. Pedro comió *cordero* en la cena.

En estos ejemplos se evidencia que ambos sentidos estarían vinculados lógicamente y, por tanto, formarían parte de la capacidad del ítem léxico en análisis para extenderse en otros sentidos derivados. En ambos casos se trata de unidades léxicas que son capaces de proyectar diferentes sentidos en la medida que interactúan con determinados contextos oracionales. Como antecedente de análisis de aplicación del enfoque generativo, se presenta el trabajo de Brouillon & Busa (2001), cuyo objetivo era plantear una representación semántica para el verbo francés «attendre» (*esperar*), que diera cuenta de sus especificaciones contextuales. De esta manera, se propone una explicación para los aspectos tanto semánticos como sintácticos que constituyen la polisemia en contextos de uso particulares. Según estos autores, la polisemia corresponde a un fenómeno particular de comportamiento

lingüístico que puede ser explicado mediante una interpretación composicional de los ítems léxicos. Se trata de un mecanismo que describe a un ítem léxico o estructura sintáctica que admite más de una aplicación funcional.

Para explicar el fenómeno en términos generales e introductorios, se recurre al ejemplo del verbo «bake» (*cocinar/cocer*), a partir de los siguientes casos:

(5)

a. John *baked* the potato.
   'John cocinó la/una papa'.
b. John *baked* the cake.
   'John cocinó la/una torta'.

Según lo anterior, «bake» presenta polisemia en tanto (5a) significa un cambio de estado del padecedor. Por tanto, el estado resultante de la papa es haber sido cocinada o 'estar cocinada/cocida'. Luego, en (5b), el significado apunta a la creación de un producto por parte un agente. En este caso particular de polisemia, la ambigüedad está dada por un lexema que tiene la capacidad de cargar con múltiples significados, dependiendo de la estructura a la que esté modificando. De esta manera se define el mecanismo generativo de la co-composición.

El trabajo ya citado de Brouillon & Busa (2001) registra seis casos de polisemia para el verbo/evento «attendre». Los separa, preliminarmente, en dos grupos. Esta distinción está dada tanto por las distintas lecturas con las que se asocia el verbo en análisis, como con las características sintácticas de los complementos que propician los diferentes sentidos. En primer lugar, se identifican los casos en los que existe una cláusula que a su vez contiene una frase verbal. Estos casos se relacionan con la espera por parte del emisor de que el estado de cosas descrito llegue a ser verdadero, o a existir. Como sigue:

(6)

a. J´attends [que tu partes] CLAUSE
   'Estoy esperando (a) que te vayas'.
   Complemento del tipo *evento (proceso)*.

b. J´attends [de savoir la vérité] CLAUSE
   ‘Estoy esperando (a) saber la verdad’.
   Complemento del tipo *evento (estado)*.

En segundo lugar, se agrupan los casos en los que el complemento de «attendre» es una frase nominal. Los autores señalan que las distinciones semánticas aquí son menos transparentes, pues el verbo adopta diferentes interpretaciones a partir de cada contexto posible. Como sigue:

(7)

a. J´attends [le concert] CD(FN)
   ‘Estoy esperando el concierto’.
b. J´attends [le bus] CD(FN)
   ‘Estoy esperando el bus’.
c. J´attends [son prochain livre] CD(FN)
   ‘Estoy esperando su nuevo libro’.

En el caso (7a), «attendre» denota la espera del inicio de un evento; en (7b) el verbo significa que alguien está esperando a que un objeto físico llegue a existir en el ámbito pertinente para el hablante; por último, en (7c) la interpretación del verbo refiere a alguien que está esperando a que un libro sea publicado. Así, en (7a) y (7c) el complemento es un ‘evento’, mientras que en (7b) corresponde a un ‘artefacto’ (objeto físico).

Los autores logran sistematizar estos aspectos según la información que es requerida por el verbo para su interpretación. De esta manera, según un primer acercamiento al fenómeno, «attendre» se constituiría como un ítem léxico funcionalmente ambiguo, tanto en los casos en los que su objeto directo es una cláusula, como en aquellos en los que es una frase nominal. Sin embargo, más adelante se verá que la propuesta de los autores niega lo anterior. De esta manera, se puede establecer que ambos casos seleccionan cierto tipo de argumentos: “the individual who is waiting (WAITER), the event that the individual expects to take place (the WAITED), and the event that will occur if the WAITED is true (RESULT)” (Brouillon & Busa, 2001: 153).

Finalmente, y a modo de síntesis de los aspectos más relevantes que aportan los tres enfoques revisados, cualquier proceso de desambiguación, a nivel cognitivo, debería comprender tres etapas. Siguiendo a Cottrell (1984):

a. Decodificar el *input* y parearlo con las unidades que son léxicamente ambiguas.
b. Acceder a la información léxica y semántica de la unidad ambigua.
c. Integrar la información con el contexto sintáctico.

El problema de los enfoques para la descripción del fenómeno de la ambigüedad, según Hong (2015), ha sido discutido predominantemente en los ámbitos de la semántica, la psicolingüística y la lingüística computacional. En este sentido, es importante considerar que el estudio de la ambigüedad implica la interacción de distintos niveles de análisis lingüístico: léxico, sintáctico, semántico y/o fonológico. Así, es posible observar diferentes hipótesis para la interacción de estos niveles en las etapas del proceso de desambiguación: por un lado, un tratamiento modular en el que la integración de información dependerá del *output* producido por niveles anteriores o, por otro, una perspectiva interactiva, en la que el procesamiento de un nivel afectará el procesamiento de otros niveles adyacentes.

2.3 El problema computacional de la ambigüedad léxica

El siguiente apartado tiene por objetivo exponer algunas definiciones relevantes en el ámbito de la lingüística computacional en general y el PLN en particular para el tratamiento de la ambigüedad léxica.

El lenguaje humano integra dos tipos complementarios de información: la información conceptual, que es de naturaleza semántica, y la información computacional[6], que es de naturaleza metadiscursiva. Esta última permite, en un sentido amplio, organizar representacionalmente el contenido conceptual para convertirlo en conocimiento accesible (Mairal-Usón *et al*., 2013; Geeraerts, 2010; Pinker, 2001).

A modo de ejemplo, se puede caracterizar la información conceptual como un depósito de conocimiento, cuyo acceso es posible mediante el enlace entre representaciones y formas lingüísticas. En este sentido, la información computacional permite integrar estos elementos tanto para incorporar como para extraer información conceptual.

[6] Cabe destacar que, en este contexto, "computacional" no refiere estrictamente a la máquina, sino que apunta más bien a los cómputos cognitivos.

Según lo anterior, la lingüística computacional tiene como objetivo reproducir los patrones con los que funcionan la mente y el lenguaje humanos, para transferirlos a la relación entre humanos y máquinas. Según Periñán-Pascual & Mairal-Usón (2009), este problema se aborda específicamente en el marco del PLN. Este último corresponde a una rama aplicada de la inteligencia artificial que se focaliza en el estudio y diseño de sistemas computacionales para facilitar la comunicación entre personas y máquinas. En definitiva, busca entregar soluciones a los problemas concretos que son planteados por la lingüística computacional. Algunas de las tareas específicas que aborda el PLN son las siguientes:

a. Desarrollo de sistemas de diálogo: según Llisterri (2006), definidos como sistemas conversacionales que constituyen una "[…] tecnología concebida para facilitar la interacción natural mediante el habla entre un humano y un ordenador" (p. 11). El objetivo principal de estos sistemas es optimizar la precisión del proceso de reconocimiento acústico-fonético por parte de la máquina, incorporando modelos lingüísticos (Gerbino *et al*., 1995).
b. Extracción y recuperación de información: de acuerdo con Manning *et al.* (2009) y Savoy & Gaussier (2010), se trata de un ámbito de investigación dedicado al desarrollo de procedimiento informáticos para encontrar documentos que contienen datos textuales no estructurados (es decir, que carecen de una estructura semánticamente abierta y fácil de usar por una máquina), con el objetivo para satisfacer una necesidad de información dentro de una colección de documentos almacenada en un computador.
c. Traducción mecánica (ya sea automática o semiautomática): corresponde a una subdisciplina de la lingüística computacional que investiga y desarrolla programas especializados para traducir datos textuales desde una lengua origen hacia una lengua meta (Sinhal & Gupta, 2017). La definición anterior abarca una variedad de herramientas, con diferentes niveles de desarrollo y basadas en distintas técnicas de PLN, como la traducción automática basada en reglas o la traducción automática estadística (Sutopo & Hastuti, 2020).

En cuanto a los enfoques de PLN, se identifican predominantemente dos, en coherencia con Periñán-Pascual (2012a). En primer lugar, el enfoque estocástico (también llamado estadístico o computacional) propone la construcción de sistemas que no almacenan conocimiento lingüístico (o conceptual), sino que aplican determinadas técnicas matemáticas sobre corpus textuales con el fin de extraer conocimiento. Estos sistemas estocásticos son capaces de inferir conocimiento lingüístico a

través de la utilización de algoritmos mediante los que se realiza una construcción automatizada del conocimiento. En definitiva, se trata de una aproximación computacional al análisis de textos entendidos como volúmenes de información computable, y que caracteriza la mente humana como un fenómeno individual y aislado de variables externas.

En segundo lugar, el enfoque simbólico se especializa en la construcción o desarrollo de sistemas que almacenan conocimiento lingüístico; esto es, esquemas de representación del conocimiento para la automatización de tareas específicas. Esta perspectiva ha propiciado el desarrollo de teorías orientadas a la descripción y explicación de las propiedades lingüísticas de las construcciones de diferentes lenguas a partir de formalizaciones. Para ejecutar esta serie de mecanismos es necesaria la utilización de un lenguaje de programación que permita a un humano entregar una serie de instrucciones y reglas (algoritmos) a la máquina, tanto para extraer como para ingresar información. De hecho, en sistemas simbólicos parte de la información conceptual se ingresa manualmente. Para trabajar en el ámbito del PLN desde este enfoque, es necesario desarrollar un lexicón vinculado con representaciones semánticas, que pueda a su vez relacionarse con representaciones sintácticas mediante algún algoritmo de enlace. En otras palabras, el objetivo del sistema es determinar estructuras semánticas que permitan predecir estructuras sintácticas.

2.3.1 Definición computacional de la ambigüedad léxica

El fenómeno de la ambigüedad léxica se ha abordado a partir de una aproximación simplificada si se la compara con la descripción lingüística. En el ámbito del PLN, los trabajos de Yurafsky & Martin (1998) han sistematizado las convenciones conceptuales con las que se hace referencia a ciertos rasgos de las unidades léxicas. En primer lugar, se define la *forma de palabra*, o morfo (*wordform*), que corresponde a la forma flexionada de una unidad léxica, tal y como aparece en el cotexto, o contexto oracional.

En segundo lugar, durante el desarrollo de esta investigación hemos restringido el uso del concepto *palabra* a la noción de palabra ortográfica (también llamada *palabra gráfica*), por lo que un N-grama corresponde a una secuencia de *n* palabras. Luego utilizaremos el concepto de *unidad léxica* para hacer referencia a una unidad funcional de significado que se realiza lingüísticamente mediante una o más palabras. Así, el criterio que hemos establecido para reconciliar las definiciones de los conceptos de *palabra* y *unidad léxica* implica que todas las unidades léxicas están compuestas por N-gramas, pero no todos los N-gramas pueden considerarse unidades léxicas; es decir, los N-gramas, en

este caso sintácticos, se pueden definir como una secuencia de unidades léxicas de extensión *N*. Así, una secuencia como «Miguel», donde *N=1* corresponde a un unigrama; una secuencia como «El Quijote», donde *N=2*, corresponde a un bigrama; y una secuencia como «Miguel de Cervantes», donde *N=3*, corresponde a un trigrama.

En tercer lugar, se establece como lema la forma citada que presenta la misma base léxica de una forma de palabra. Por ejemplo, la unidad léxica «banks» corresponde a un *morfo*, mientras que «bank» es su lema[7].

Por otra parte, el sentido corresponde a una representación discreta del significado de una unidad léxica. Finalmente, la ambigüedad léxica se puede manifestar en dos casos: homonimia o polisemia. Por ejemplo:

a. Homonimia: coincidencia de dos lemas en su escritura o pronunciación, aún cuando difieren en sus sentidos. Presenta, a su vez, dos tipos:
    i. Homógrafo: coincidencia gráfica en la escritura.
       p. ej. *can* (poder - lata)
    ii. Homófono: coincidencia en la pronunciación, diferencia en la escritura.
       p. ej. *write* (escribir) | *right* (derecho).
b. Polisemia: coincidencia de dos lemas en su escritura y pronunciación, cuyos sentidos se encuentran relacionados con el mismo significado:
    i. p. ej. *bank* (institución) | *bank* (edificio de la institución).

2.3.2 La desambiguación léxica en el ámbito del PLN

Desde la perspectiva del PLN, la desambiguación léxica se define como una habilidad computacional, y por lo tanto automática, para activar una interpretación posible de una unidad léxica en un contexto oracional determinado. Este problema es de suma importancia, ya que no solamente implica una adecuada descripción del fenómeno, sino que requiere del desarrollo de un algoritmo que sea capaz de asignar sentidos a unidades léxicas que funcionan en contextos lingüísticos auténticos. El procedimiento supone etiquetar un corpus y proveer un número consistente de reglas para la selección de sentidos de una unidad léxica en un contexto oracional determinado. En este sentido, el desarrollo

[7] El tratamiento de estos conceptos, desde el ámbito de la lexicografía, establece que una unidad léxica es seleccionada como lema, y se posiciona por tanto como el criterio desde el que se organizan las definiciones en los diccionarios semasiológicos.

de algoritmos para hacer más eficientes distintos procedimientos de extracción de información desde bases de datos léxicas es el componente fundamental de los métodos estocásticos.

El tratamiento de la polisemia presente en estos ejemplos presenta la dificultad de seleccionar adecuadamente los parámetros del *output* que una búsqueda pueda arrojar a partir de un determinado *input*. Estos sistemas deben comprender una alta eficiencia computacional para procesar grandes volúmenes de información. En efecto, se sitúa la polisemia como uno de los problemas más importantes que enfrenta actualmente el desarrollo de sistemas para la recuperación de información, sobre todo aquellos que se focalizan en sistemas que permiten desambiguar términos en contextos de conocimiento especializado, como PubMed[8], o generales como FrameNet[9]. A partir de la década de 1950 ha existido interés, en el contexto del tratamiento computacional del lenguaje, por la desambiguación del sentido de las palabras como una labor necesaria e intermedia para lograr determinadas tareas de PLN. Las máquinas, o computadores, no poseen la habilidad inherente para procesar lenguas naturales y, por tanto, son incapaces de reconocer casos de polisemia a menos que se les provea de mecanismos específicos para lograr esta tarea. Es fundamental comprender el problema de la desambiguación léxica como un aspecto del desarrollo de la inteligencia artificial que solo puede ser solucionado en la medida que sean abordados otros dos problemas: (a) la representación del conocimiento enciclopédico y (b) la representación del conocimiento que proviene del sentido común. En este sentido, el proceso de desambiguación léxica se define, según Nevzorova *et al*. (2015), como la habilidad de la máquina para identificar el significado de un ítem léxico en determinados contextos a partir de procedimientos computacionales.

En general, se asume que la desambiguación léxica constituye un procedimiento de asignación de significado a una unidad léxica, que está basado en el contexto en el que esa unidad ocurre (Patwardhan *et al*., 2003). Esto es coherente con la idea de que un significado adecuado para un determinado ítem léxico será seleccionado a partir de un inventario de sentidos, que definen a su vez el rango de posibilidades para esa unidad léxica en un contexto particular. Existen ámbitos del PLN en los que la resolución del problema de la desambiguación es crítica para establecer de manera eficiente la comunicación hombre-máquina. A continuación, se presentan algunos ejemplos para los tipos de tareas en PLN cuyas aplicaciones requieren de métodos para abordar la desambiguación léxica. Posteriormente se exponen, en términos generales y según el criterio de aquellos que han sido

[8] https://www.ncbi.nlm.nih.gov/pubmed

[9] https://framenet.icsi.berkeley.edu/fndrupal

más relevantes, algunos de los métodos para tratar el problema de la desambiguación léxica. En ambos casos, la siguiente exposición se basa predominantemente en el trabajo de Ide & Veronis (1998). En cuanto a las tareas de PLN, se reconoce el desarrollo de sistemas que puedan abordar y resolver el problema de la desambiguación léxica como una tarea intermedia para procedimientos computacionales complejos. Por ejemplo:

a. Traducción mecánica: traducción adecuada al contexto en caso de ambigüedad léxica predominantemente (Vickrey *et al*., 2005; Ríos *et al.*, 2017). El procedimiento de desambiguación léxica automática aplicado a la traducción mecánica incluye, típicamente, dos procesos (Kaur-Sidhu & Kaur, 2013). Primero, la comprensión de la lengua origen a partir de una fuente de conocimiento lingüístico; segundo, la traducción a la lengua meta. En ambos procesos, tanto las unidades léxicas de la lengua origen como las de la lengua meta serán potencialmente ambiguas, puesto que la asignación de los sentidos correctos dependerá de su aparición en diferentes contextos oracionales.
b. Extracción o recuperación de información: precisión de ocurrencias para resultados de búsquedas por palabras clave tanto en documentos, bases de datos o motores de búsqueda (Stokoe *et al*., 2003). Específicamente, según Chifu *et al.* (2014), los diversos métodos de desambiguación léxica automática se pueden aplicar para ayudar a los sistemas de recuperación de información a determinar qué documentos o datos textuales deben recuperarse en relación con una consulta potencialmente ambigua.
c. Procesamiento de habla: síntesis de voz o sistemas de fonetización de palabras en los que es necesaria la segmentación y la discriminación de homofonía en reconocimiento de voz. En particular, el trabajo de Yarowsky (1997) expone algunos casos en los que la pronunciación de determinadas unidades léxicas no se puede determinar sin considerar los aspectos tanto sintácticos como semánticos del contexto oracional. Por ejemplo, un tipo de ambigüedad de pronunciación ocurre cuando dos homófonos constituyen diferentes componentes de la estructura sintáctica, como en «three lives were lost», que se traduce como «se perdieron tres vidas», donde «lives» [‘laɪvz] funciona como sustantivo; en contraposición a «one lives to eat», que se traduce como «uno vive para comer», donde «lives» [‘lɪvz] funciona como verbo. Este tipo de ambigüedad se puede resolver mediante la identificación de patrones sintácticos locales.

Asimismo, la desambiguación léxica requiere de un proceso que lleve a cabo la asociación entre cierto ítem léxico en un texto (o discurso) y una definición o significado (sentido) entre varios significados potencialmente atribuibles a esa palabra. La traducción mecánica, por su parte, tiene por objetivo determinar la opción correcta de significado para un ítem léxico, a partir de un número finito de significados posibles. Este método se utiliza principalmente en la traducción de textos técnicos, es decir, aquellos que pertenecen a dominios reducidos y/o especializados. Esta distinción podría facilitar la desambiguación en tanto se trata de información contextual relevante para el proceso.

Como otras tareas en el ámbito del PLN, la desambiguación léxica automática está sometida al llamado problema del cuello de botella en la adquisición del conocimiento, abordado por Buchanan & Wilkins (1993), Wagner (2006), Aussenac-Gilles & Gandon 2013, y Pasini (2020). Para comprenderlo, es fundamental reconocer que la representación del conocimiento (en este caso, entendido como conocimiento de mundo proveniente desde el sentido común) se realiza mediante una base de conocimiento que opera a través de un sistema de inteligencia artificial. Entonces, cualquier sistema de PLN basado en conocimiento debería funcionar mediante la relación entre la representación del conocimiento, una base de conocimiento y un motor de inferencia. Así, según Wagner (2006), el problema del cuello de botella en la adquisición del conocimiento consiste en el límite que se plantea para la representación del conocimiento, dado que el desarrollo de una ontología generalmente requiere de expertos que puedan contribuir a poblar la base de conocimiento. Estos expertos, a su vez, emprenden manualmente una tarea que, evidentemente, es costosa en términos de tiempo y recursos. Además, no serían capaces de dar cuenta de todo el conocimiento disponible en las fuentes de información, pues en la medida que una base de conocimiento crece, también lo hace el requisito de mantenimiento y actualización.

Finalmente, este problema afecta a los diferentes métodos propuestos para la desambiguación léxica automática, particularmente aquellos que están basados en conocimiento y aprendizaje automático, puesto que cada unidad léxica en un lexicón contiene un conjunto de potenciales sentidos dependientes de un contexto oracional particular. Así, la eficiencia de un sistema de desambiguación dependerá de la calidad y cantidad de información que contenga un recurso lingüístico informatizado que constituirá el corpus de entrenamiento en el caso del aprendizaje supervisado, o bien las instancias en análisis en el caso de las métricas de desambiguación. Por ejemplo, según el caso extraído de Pasini (2020), si se considera una lengua con un total aproximado de 200.00 sentidos distribuidos en determinadas palabras, se podría establecer la pertinencia de un conjunto de instancias < 2.000.000,

que permita proporcionar, al menos, diez ejemplos para cada sentido. Además, la situación se agravaría al considerar la granularidad de los sentidos de las palabras, y la manera en la que esos sentidos se distribuyen al interior de un corpus.

2.4 Métodos para la desambiguación léxica automática

A continuación, se presenta una exposición de los métodos para la desambiguación léxica automática disponibles en el ámbito del PLN. La clasificación para estos métodos tiene que ver con el tipo de conocimiento que utilizan como recurso lingüístico informatizado desde el que se extraen los casos de ambigüedad. En primer lugar, los métodos basados en fuentes externas de conocimiento utilizan como fuentes de conocimiento exógeno distintos tipos de recursos lingüísticos informatizados, con especial énfasis en las taxonomías y las ontologías. Algunos ejemplos representativos de este tipo de recursos lingüísticos informatizados son: *WordNet* (Miller, 1985; Miller *et al*., 1993; Fellbaum, 1998), *MultiWordNet* (Pianta *et al*., 2002), *Spanish FrameNet* (Subirats & Petruck, 2003; Subirats, 2004), *CYC Proyect* (Matuszek *et al*., 2006) y *FunGramKB* (Periñán-Pascual & Arcas-Túnez, 2007; 2010; Periñán-Pascual, 2012b). Específicamente, el trabajo de Curtis *et al*. (2006), en el marco del Proyecto CYC, realiza una propuesta para determinar una medida de cercanía semántica basada en las relaciones taxonómicas entre conceptos formalizados en la ontología de CYC, con el objetivo de aplicar este proceso para resolver la desambiguación léxica automática.

De esta manera, los sistemas de desambiguación pueden utilizar información construccional, léxica, sintáctica o semántica durante el proceso de clasificación de los sentidos. Por otra parte, los métodos basados en conocimiento contextual[10], utilizan el mismo corpus en análisis, típicamente controlado a partir del contexto oracional o ventana contextual, para derivar información predominantemente estocástica. Dados estos dos mecanismos y sus distintas tareas de procesamiento, que serán explicadas más adelante, los métodos basados en fuentes externas pueden presentar el problema del exceso de conocimiento, pues cuentan con una cantidad de información tan extensa que podría dificultar el análisis; mientras que los métodos basados en conocimiento contextual padecen un defecto de conocimiento, pues el cotexto no necesariamente constituye un conjunto de datos textuales suficiente para lograr que un sistema automático derive clasificaciones eficientes.

[10] También llamado *cotexto, instancia o ventana de palabras* (en la que aparece la palabra objetivo). Estas tres maneras de referir a los componentes léxicos que rodean a la palabra ambigua serán utilizados alternadamente a lo largo de este trabajo. Se trata de la bolsa de palabras adyacentes a la palabra objetivo al interior de un corpus.

Luego los métodos basados en fuentes externas de conocimiento presentan a su vez una distinción metodológica fundamental para comprender el desarrollo de estos sistemas: relación semántica, *word relatedness* (Lesk, 1986; 1987; Brown *et al*., 1991; Banerjee & Pedersen, 2002; 2003), y similitud semántica, *word similarity* (Wu & Palmer, 1994; Resnik, 1995; Jiang & Conrath, 1997; Leacock & Chodorow, 1998; Lin, 1998). Por un lado, las palabras similares corresponden a los llamados sinónimos cercanos (*near-synonyms*), mientras que las palabras relacionadas son aquellos sentidos que se encuentran dentro del mismo campo semántico; es decir, que comparten ciertos aspectos de su significado, y frecuentemente coocurren en el mismo contexto oracional. Por ejemplo, las unidades léxicas «auto» y «bicicleta» son sinónimos cercanos (o palabras similares) porque, entre sus características, ambos significados presentan una relación léxica de significado correspondiente a la hiponimia de la categoría 'medios de transporte'. Por el contrario, «auto» y «rueda» no son similares, sino que se trata de palabras relacionadas, porque comparten algunos aspectos de su significado basados una relación metonímica. De esta manera, las palabras similares se pueden definir como aquellas que frecuentemente comparten el mismo cotexto, entendido a su vez como una instancia en la cual aparece (o que contiene) la palabra objetivo. Por ejemplo, «doctor» es similar a «salud», porque comparten un campo semántico entendido como un conjunto de unidades léxicas relacionadas en virtud de la presencia en todas ellas de ciertas notas de significado común (Escandell, 2007).

Los criterios para determinar la similitud están dados por las características o atributos (*features*) que presenta una determinada instancia o *wordform*, y que corresponden a su vez a cualquier palabra de contenido (en oposición a las palabras funcionales), presente en el grupo de palabras adyacentes a la palabra ambigua, o bien los sustantivos contenidos en su definición al interior de un diccionario legible por la máquina. Se entenderán como características la o las palabras de contenido, cualquiera sea su clase (dejando fuera entonces a las palabras funcionales, o *stopwords*), que sean adyacentes a las unidades léxicas referidas en la definición de la palabra objetivo en un diccionario legible por la máquina, que a su vez se utilice como fuente de conocimiento lingüístico.

Además, la mayoría de los métodos que se presentan utilizan la taxonomía de WordNet como fuente de conocimiento. WordNet (Miller, 1985; Miller *et al*., 1993; Fellbaum, 1998) es un recurso lingüístico informatizado creado por el departamento de Ciencias Cognitivas de la Universidad de Princeton. Si bien se lo define como un diccionario legible por la máquina, se trata más bien de una base de datos léxico-conceptual estructurada en forma de red semántica, cuyo objetivo es constituir

un modelo del conocimiento léxico-conceptual de los hablantes de inglés. WordNet almacena exclusivamente palabras de clase abierta; esto es, sustantivos, adjetivos, verbos y adverbios (por tanto, deja fuera palabras de clase cerrada, como pronombres, conjunciones y preposiciones). Sus relaciones semánticas básicas son la sinonimia y la hiponimia-hiperonimia, en tanto que constituye una jerarquía conceptual. Cada palabra, o etiqueta, se agrupa en conjuntos de sentidos de palabras. Estos sentidos son, a su vez, los significados disponibles que se asocian a cada etiqueta; es decir, un conjunto de sinónimos para cada concepto, o *synsets*. WordNet integra > 126.000, organizadas en 91.000 *synsets*. Por ejemplo, el sustantivo «board» aparece en los siguientes *synsets*: (1) {*board*, *plank*} para el sentido 'tabla' o 'plancha'; (2) {*board*} para 'consejo'; y {*board*, *table*} para 'mesa'. Esto, siguiendo a Climent (1999), representa que *board-1* es sinónimo de «plank»; y *board-3* es sinónimo de «table». Finalmente, la exposición que se presenta a continuación no pretende revisar exhaustivamente todos los métodos disponibles en la actualidad para la desambiguación léxica automática, sino que se trata de una selección de los métodos que se relacionan específicamente con los objetivos que hemos propuesto en esta investigación, y que se pueden clasificar como métodos de relación semántica, de similitud semántica, y basados en conocimiento contextual.

2.4.1 Métodos de relación semántica (*word relatedness*)

2.4.1.1 Tipo Lesk (1986; 1987)

Este método, desarrollado inicialmente por Michael Lesk (1986), corresponde, en términos generales, a una medida basada en la información contenida en un diccionario legible por la máquina, a partir de la cual se calcula el solapamiento de palabras entre dos palabras objetivo[11] y las glosas de sus respectivas definiciones. El mayor mérito de este procedimiento es su capacidad para encontrar la combinación de los sentidos de palabra que maximice la relación total entre los sentidos de las palabras objetivo, de modo que estas cumplan con el criterio de la relación semántica.

En términos lingüísticos, uno de los fundamentos para este método es la idea de que el significado de un concepto se expresa mediante una agrupación de palabras y, por tanto, la manera más eficiente para cuantificar esta relación semántica entre las unidades léxicas que componen una definición sería el solapamiento de palabras. Según lo anterior, se hace imprescindible la utilización de definiciones de diccionario, como una fuente de conocimiento externa que permitirá establecer los

[11] Se utilizará de aquí en adelante la expresión *palabra objetivo* para referirnos a aquella unidad léxica que se establece como referencia para la definición de la ventana de palabras durante el análisis de un texto de entrada.

parámetros de coocurrencia para el solapamiento de glosas. En cuanto a su formalización, y según la explicación de Navigli (2009), dadas dos palabras objetivo (W_1, W_2), se calculará un puntaje para cada par de sentidos de palabra (S_1, S_2), donde $S_1 \in sentidos(W_1)$, y $S_2 \in sentidos(W_2)$. De esta forma, se establece la intersección entre las glosas de las definiciones de cada palabra objetivo:

$$puntaje_{LESK}(S_1, S_2) = glosa(S_1) \cap glosa(S_2)$$

De esta forma, la glosa de cada sentido de palabra consistiría en una bolsa de palabras correspondiente a la definición textual del sentido (S_i) para cada palabra (W_i).

El ejemplo más representativo para la explicación del procedimiento de solapamiento de glosas es el cálculo de la relación semántica entre los sentidos de las palabras «pine» y «cone», propuesto por Lesk (1986). Para esto, se utilizaron las definiciones provenientes del diccionario *Oxford Advanced Learner* (Hornby *et al*., 1974), en el que se incluyen cuatro sentidos para «pine» y tres sentidos para «cone». Lo anterior se expone en las tablas 1 y 2:

Tabla 1. Sentidos y definiciones para «pine».

PINE	
Sentido	**Glosa**
S_1 = *tree*	*glosa*(S_1) = seven kinds of evergreen tree with needle-shaped leaves 2 pine
S_2 = *pine*	*glosa*(S_2) = pine
S_3 = *waste*	*glosa*(S_3) = waste away through sorrow or illness
S_4 = *something*	*glosa*(S_4) = pine for something, pine to do something

Tabla 2. Sentidos y definiciones para «cone».

CONE	
Sentido	**Glosa**
S_1 = *body*	*glosa*(S_1) = solid body which narrows to a point
S_2 = *shape*	*glosa*(S_2) = something of this shape, whether solid or hollow
S_3 = *fruit*	*glosa*(S_3) = fruit of certain evergreen trees (fir, pine)

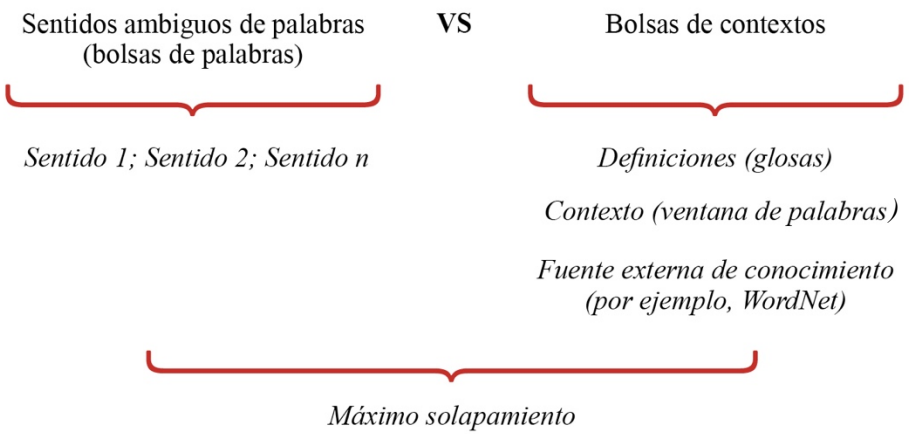
Posteriormente, se evaluó el solapamiento de cada una de las glosas correspondientes al conjunto de pares para los sentidos de «pine» y «cone» registrados en la fuente de conocimiento, para así establecer un puntaje que expresa el número de coocurrencias. Lo anterior se expresa de la siguiente forma:

Tabla 3. Puntajes para el solapamiento entre los conceptos «pine» y «cone».

Solapamiento	Puntaje
$cone_{glosa(S_1)} \cap pine_{glosa(S_1)}$	= 0
$cone_{glosa(S_2)} \cap pine_{glosa(S_1)}$	= 0
$cone_{glosa(S_3)} \cap pine_{glosa(S_1)}$	= 0
$cone_{glosa(S_4)} \cap pine_{glosa(S_1)}$	= 0
$cone_{glosa(S_1)} \cap pine_{glosa(S_2)}$	= 0
$cone_{glosa(S_2)} \cap pine_{glosa(S_2)}$	= 0
$cone_{glosa(S_3)} \cap pine_{glosa(S_2)}$	= 1
$cone_{glosa(S_4)} \cap pine_{glosa(S_2)}$	= 0
$cone_{glosa(S_1)} \cap pine_{glosa(S_3)}$	**= 2**
$cone_{glosa(S_2)} \cap pine_{glosa(S_3)}$	= 1
$cone_{glosa(S_3)} \cap pine_{glosa(S_3)}$	= 0
$cone_{glosa(S_4)} \cap pine_{glosa(S_3)}$	= 1

Como resultado, el máximo puntaje de solapamiento entre todas las posibles combinaciones de sentidos se obtiene para la intersección entre la glosa del sentido tres de «pine» y la glosa del sentido uno de «cone», en el que se evidencian dos coocurrencias. En este caso, se solapan las palabras «evergreen» y «tree». Por lo tanto, la medida de Lesk seleccionará estos sentidos y los asignará en el proceso de desambiguación cada vez que las palabras «pine» y «cone» coincidan en la misma ventana contextual. Finalmente, si eso ocurre, entonces probablemente el sentido de «pine», entendido como un concepto relacionado semánticamente con «cone», sería el de un 'tipo de árbol de hoja perenne con hojas en forma de aguja', mientras que el de «cone» correspondería al de un 'cuerpo sólido que se estrecha hasta un punto determinado'. Este tipo de procedimiento podría generalizarse en la siguiente figura:

Figura 2. Procedimiento general de métodos basados en solapamiento.



Diversos estudios acerca del desarrollo y progresión de los métodos para la desambiguación automática de sentidos de palabra publicados a partir de la década de 1990 (Ide & Veronis, 1998; Navigli, 2009; Vihdu & Abirami, 2014), indican que el método Lesk es uno de los primeros algoritmos para la desambiguación léxica automática considerado como exitoso, y que logró de manera eficiente la consolidación del criterio de incorporar una fuente de conocimiento externa para abordar este tipo de procesamiento. Esto es particularmente relevante desde la perspectiva de la inclusión de un incipiente razonamiento lingüístico dentro de una propuesta íntegramente estocástica. A pesar de la aprobación de la que goza el método Lesk original, debido a la simpleza de su algoritmo y la efectividad que presenta en varios de los casos experimentales más canónicos, las críticas para este método son bastante consistentes en la mayor parte de la literatura especializada (Wilks *et al*., 1989; Cowie *et al*., 1992).

Se han establecido dos grandes deficiencias para este método. En primer lugar, se ve afectado directamente por el tamaño de la fuente de conocimiento (i.e. el diccionario legible por la máquina), y por la exactitud de las palabras utilizadas en las glosas de las definiciones. Este problema implica que, dado que las definiciones son típicamente breves, la bolsa de palabras de cada una no permite establecer una relación entre los sentidos en análisis, porque la frecuencia de los casos de solapamiento es muy baja. En segundo lugar, se trata de un método que, en la medida que intenta encontrar la relación más significativa de sentidos de palabra, se ve sometido al problema de la explosión

combinatoria. Esta limitación implica que, mientras más sentidos en análisis sean seleccionados, y junto con ellos aumente el volumen de cada bolsa de palabras para sus correspondientes palabras objetivo, entonces se reduce la probabilidad de encontrar una relación óptima entre una palabra y su correspondiente sentido.

2.4.1.2 Desambiguación estadística (Cantos-Gómez, 1996)

En términos prácticos, la analogía de Weaver (1955) explica con claridad el procedimiento que aborda la desambiguación estadística. Según este autor, sería imposible acceder al significado de todas las palabras en un libro solamente mirando a través de una máscara negra con un pequeño agujero que permitiera leer. Por el contrario, si ese agujero se agranda o aumenta de manera sucesiva, entonces sería posible relacionar progresivamente un número cada vez mayor de palabras. El objetivo entonces es que el agujero sea tan grande como para poder observar el mayor número de palabras posibles y relacionadas entre sí o con una palabra central que requiere de un proceso de desambiguación. En definitiva, es necesario responder a la pregunta: ¿Qué valor mínimo de *n* conducirá, al menos en una fracción tolerable de casos, a la elección correcta del significado de la palabra objetivo? Esta perspectiva para el abordaje de la desambiguación ha resultado ser particularmente predictiva en cuanto a los significados adecuados de las palabras en dominios expertos (Brown *et al*., 1991). En efecto, se trata de una aproximación que se utiliza ampliamente en tareas como la traducción automática o la extracción de información. Un ejemplo ilustrativo para comprender la manera en la que trabaja la desambiguación estadística es el que propone Cantos-Gómez (1996).

Consideremos el caso de un corpus lingüístico en el que se registran 16.585 apariciones para la unidad léxica del inglés «ball». De estas apariciones, se intenta desambiguar a partir de dos sentidos posibles, que se podrían relacionar a su vez con estructuras sintácticas del tipo:

(8)

a. The *ball* was kicked by the forward.
b. The *ball* might consider all kind of music.

En el caso de (8a), el significado de «ball» se podría homologar en español al de un «balón» o «pelota», que corresponde a un ‘objeto sólido con forma esférica que se utiliza para juegos’. Por otra parte, en (8b) la unidad «ball» refiere al español «baile», como un ‘tipo de evento establecido

socialmente en el que se ejecuta la acción de bailar'. A partir de estos dos sentidos de «ball», se obtiene la siguiente información estadística:

Tabla 4. Dos sentidos de «ball» y sus ocurrencias (adaptada de Cantos-Gómez, 1996).

$sentido_1$	Esfera sólida o hueca que se utiliza típicamente en juegos	Ocurrencias = 16.400
$sentido_2$	Reunión social formal cuyo objetivo es bailar	Ocurrencias = 185
		Total de ocurrencias = 16.585

Después de obtener esta evidencia, en la que el *sentido*$_1$ para «ball» ocurre consistentemente en el 98% de los casos, se concluye que, a pesar de que el corpus podría no constituir una instanciación suficientemente representativa para las unidades léxicas en análisis, esta técnica entregará la respuesta correcta el 98% de las veces. Además, el cálculo de la probabilidad se puede mejorar con la adición de algunas variables del contexto de aparición; por ejemplo, si se seleccionan algunas colocaciones, o palabras que típicamente aparecerán con el ítem en análisis cuando corresponda a uno u otro de sus sentidos, o bien cuando aparezcan algunas de las palabras que componen la definición. El objetivo, en términos generales, consiste en calcular la probabilidad de que los sentidos de una palabra relativos a una porción del texto focalizado en esa palabra sean efectivamente correspondientes en un contexto de aparición determinado. Luego si se considera un nuevo conteo de ocurrencias para un conjunto de instancias específicas del texto, se puede estimar una nueva probabilidad dependiente del contexto de aparición, como se representa en la siguiente tabla:

Tabla 5. Conteo de ocurrencias para dos sentidos de «ball» en un corpus de 15 millones de palabras, adaptada de Cantos-Gómez (1996).

	ball / sentido$_1$	*ball / sentido*$_2$
tennis	685	1
dancing	1	88
the	15854	166
Ocurrencias totales	**16400**	**185**

Así, es posible estimar la probabilidad de cada sentido correspondiente a partir de la dependencia del cotexto. En este caso, la probabilidad condicional de que «tennis» se relacione a *sentido*$_1$, en relación con sus ocurrencias, se podría calcular dividiendo el número de ocurrencias de «tennis» para el sentido 'esfera sólida o hueca […]' por el número de ocurrencias totales:

$$\text{Prob}_{\text{n}}\left(\frac{tennis \mid ball}{\text{sentido}_1}\right) = \frac{685}{16.400} = 0{,}04176$$

En comparación, la probabilidad condicional de que «tennis» se relacione con el sentido 'reunión formal social […]' es más baja a partir de sus ocurrencias divididas por el número de ocurrencias totales:

$$\text{Prob}_{\text{n}}\left(\frac{tennis \mid ball}{\text{sentido}_2}\right) = \frac{1}{185} = 0{,}00540$$

Si bien el anterior se trata de un ejemplo simple a partir de una fórmula simplificada que pretende establecer un cálculo estadístico para la probabilidad de ocurrencias del sentido particular de una palabra puesta en contexto, permite explicar adecuadamente el principio que regula este tipo de métodos.

2.4.1.3 Tipo Lesk adaptado (Banerjee & Pedersen, 2002)

Esta propuesta de adaptación de la medida de Lesk (1986) surge como una solución para el problema de la limitación de las glosas de las definiciones de diccionario. Dado que las definiciones típicamente tienden a ser breves, la identificación de sentidos relacionados según la medida de solapamiento original se ve imposibilitada porque la información contenida en la fuente de conocimiento es insuficiente. Asimismo, la medida original es redundante, altamente iterativa e ineficiente dado el problema de la explosión combinatoria.

En los trabajos de Banerjee & Pedersen (2002; 2003) se propone una adaptación a la medida original de Lesk, mediante la incorporación del contexto sintáctico de la palabra objetivo. La crítica de estos autores a la medida original, además de las debilidades antes expuestas, es que el proceso de activación (*spreading activation*) del solapamiento es limitado, dado que no es un mecanismo suficiente para dar cuenta de relaciones más indirectas entre las palabras en análisis. Así, se proponen dos soluciones preliminares: (1) expandir las definiciones de diccionario para aumentar las

posibilidades de encontrar coocurrencias, y (2) considerar ventanas contextuales específicas, al mismo tiempo que se incorporan las glosas de los sentidos para aquellos conceptos que se encuentran relacionados léxica o semánticamente con la taxonomía de WordNet (Miller, 1985; Miller *et al*., 1993; Fellbaum, 1998). Esta adaptación de la medida original considera un cotexto, correspondiente a *n tokens* de WordNet a la izquierda y a la derecha de la palabra objetivo; es decir, una ventana contextual de $2n + 1 = N$, donde $2n$ corresponde al número de palabras adyacentes. Si la palabra objetivo se encuentra al inicio o al final de la instancia a considerar en la ventana contextual, se adhieren palabras en la dirección opuesta que corresponda. Acerca del criterio para la definición de la ventana contextual, Choueka & Lusignan (1985) afirman que, desde una perspectiva cognitivista, el ser humano realiza decisiones de desambiguación basadas en intervalos de información estrechos que rodean a la palabra objetivo, usualmente una o dos palabras en cada dirección. Luego $W_i \leq i \leq N$; palabras de WordNet en la ventana contextual, donde cada palabra contiene uno o más posibles sentidos, y W_i corresponde al número de sentidos posibles para cada palabra. Según esto, se establece un puntaje de combinación producto de la evaluación de todas las posibles combinatorias para la asignación de sentidos en la ventana contextual, mediante la comparación de glosas para los pares de palabras. Este puntaje corresponde a la siguiente ecuación:

$$\prod_{i=1}^{N} |W_i|$$

donde la multiplicatoria de todos los valores de W_i implica que i varía desde 1 hasta el valor total de la ventana contextual. Esta técnica permite que el sentido desambiguado corresponda a aquella combinación que se activa con mayor fuerza, en relación con las palabras adyacentes en la ventana contextual.

2.4.2 Métodos de similitud semántica (*word similarity*)

2.4.2.1 Medidas de distancia entre rutas

Las medidas de distancia entre rutas utilizan como método de desambiguación la clasificación direccional de relaciones taxonómicas. Utilizan como fuente de conocimiento la taxonomía de WordNet, que contiene relaciones taxonómicas del tipo *IS-A* (vertical) y *HAS-PART* (horizontal). Desde ahí es posible establecer la distancia de las rutas taxonómicas mediante etiquetas numéricas.

La pregunta central de este tipo de medidas es: ¿Qué tan cercanas son estas unidades léxicas en cuanto a su posición en una jerarquía conceptual? Así, dos conceptos serán similares, o tendrán sentidos similares, mientras más cerca se encuentren el uno del otro en una jerarquía taxonómica. A continuación, se presentan tres medidas basadas en la distancia entre rutas.

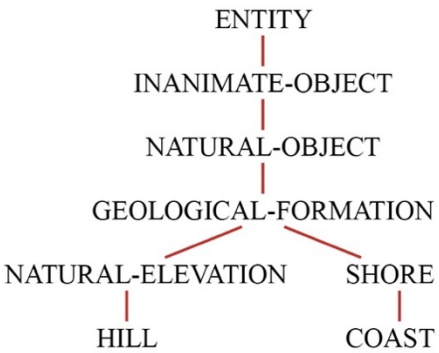
2.4.2.1.1 Wu & Palmer (1994)

La propuesta de Wu & Palmer (1994) consiste en una medida de similitud semántica basada en las variables de distancias y profundidades en una taxonomía. De esta forma, se considera la distancia de la ruta entre dos conjuntos de sentidos, S_1 y S_2, y su hiperónimo común más cercano, S_3 (*lowest common subsumer*, de aquí en adelante LCS), así como la distancia entre el LCS y la raíz de la taxonomía en la que se encuentran los conjuntos de sentidos. Lo anterior se puede expresar de la siguiente manera:

$$Sim_{Wu\ \&\ Palmer}(S_1, S_2) = \frac{2 \times\ Dist_{ruta}(S_3, raiz)}{Dist_{ruta}(S_1, S_3) + Dist_{ruta}(S_2, S_3) + 2 \times\ Dist_{ruta}(S_3, raiz)}$$

A modo de ejemplo, si se considera la siguiente jerarquía extraída desde WordNet, es posible establecer la similitud entre dos conceptos como una función entre la longitud de la ruta que los pone en relación *IS-A*, y la posición de esos conceptos en la taxonomía:

Figura 3. Taxonomía de los conceptos HILL y COAST, tomado de Jurafsky & Martin (1998).



Luego, la medida de similitud corresponderá al número de enlaces *N* presentes en la distancia de la ruta desde un concepto a otro:

$$Sim_{Wu\ \&\ Palmer}(S_1, S_2) = \frac{2 \times N_3}{N_{1+}N_2 + 2 \times N_3}$$

donde N_1 y N_2 corresponden al número de enlaces *IS-A* desde el concepto S_1 hasta el concepto S_2, en relación con su LCS, considerando N_3 como el número de enlaces *IS-A* desde el LCS hasta el nodo raíz de la taxonomía. De esta forma, según los valores de la taxonomía de ENTITY para HILL y COAST:

$$Sim_{Wu\ \&\ Palmer}(hill, coast) = \frac{2 \times 3}{2 + 2 + 2 \times 3} = 0{,}6$$

En este caso, la medida de similitud para las unidades «hill» y «coast» requiere establecer GEOLOGICAL-FORMATION como la superclase específica común. Luego N_1 y N_2 contabilizan dos enlaces hasta GEOLOGICAL-FORMATION, respectivamente, y N_3 contabiliza tres enlaces desde GEOLOGICAL-FORMATION hasta el nodo raíz ENTITY.

2.4.2.1.2 Leacock & Chodorow (1998)

Se trata de una medida basada en la longitud de la distancia entre rutas; es decir, la similitud entre dos conceptos correspondería a una función entre la longitud de la ruta que relaciona esos conceptos en una jerarquía conceptual de tipo *IS-A*, y su posición respecto a otros hiperónimos en la taxonomía. En este caso, para las definiciones de sustantivos se utiliza la jerarquía conceptual de WordNet. La medida se formaliza de la siguiente manera (Leacock & Chodorow, 1998):

$$Sim_{Leacock\ \&\ Chodorow}(S_1, S_2) = -\log\left(\frac{dist_{node}(S_1, S_2)}{2 \times depth}\right)$$

Para comprender esta medida es necesario considerar, con especial relevancia, dos variables. Primero, el valor de la profundidad de la taxonomía (*depth*) entre los conceptos en análisis, que corresponde a la longitud del camino más corto entre cada *synset* y el nodo raíz de la taxonomía. Segundo, el valor

de la distancia entre los nodos, que corresponderá en este caso al LCS entre ambos conceptos. Por ejemplo, si se considera la distancia entre rutas para los conceptos SHORE y HILL, según la figura dos, la profundidad máxima es igual a 5, mientras que la distancia entre sus nodos tiene un valor de 3. Luego, el puntaje de similitud corresponderá a:

$$Sim_{Leacock\ \&\ Chodorow}(hill, coast) = -\log\left(\frac{3}{2 \times 5}\right) = -\log(0{,}3) = 1{,}20$$

La ventaja de esta medida es que pone en relación la profundidad junto con el concepto que contiene los atributos comunes a los conceptos en análisis. Sin embargo, presenta la dificultad de que el puntaje final se ve afectado significativamente por la presencia o ausencia del hiperónimo común más cercano.

2.4.2.2 Medidas de contenido de información

Las medidas basadas en el contenido de información (*information content*, de aquí en adelante IC). constituyen una propuesta que surge a partir de la noción de LCS, donde *P*(*c*) es la probabilidad de que una palabra elegida aleatoriamente sea una instancia del concepto *c*. Así, la frecuencia de la instancia de un concepto *c* en la taxonomía, o *freq*(*c*), se puede calcular a partir de la sumatoria de aparición de las palabras que son hipónimos del nodo al que pertenece *c*:

$$freq(c) = \sum_{w \in w(c)} count\ (w)$$

Luego, la probabilidad de que una palabra elegida aleatoriamente desde un corpus sea una instancia del concepto *c* corresponderá a la frecuencia de *c* dividida por el número total de palabras en el corpus:

$$P(c) = \frac{freq(c)}{N}$$

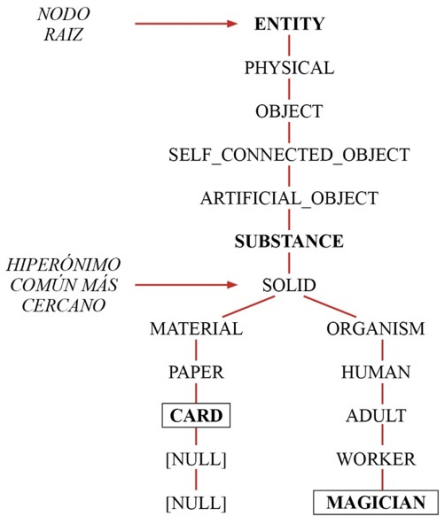
De esta forma, el IC puede ser representado matemáticamente como el logaritmo negativo de esa probabilidad, puesto que, cuando *P*(*c*) aumenta, el valor de *IC*(*c*) disminuye:

$$IC(c) = -\log P(c)$$

Finalmente, un puntaje de IC alto se puede interpretar como el hecho de que un concepto representa un significado conceptual altamente específico cuando ocurre en un texto, mientras que un puntaje de IC bajo corresponde a un significado conceptual general. Así, el indicador de IC correspondería a la medición de la especificidad de un concepto en cuanto a su significado.

Un ejemplo para comprender el concepto de LCS, como una de las variables más relevantes para determinar el IC de un concepto y su posterior comparación con el objetivo de resolver automáticamente la ambigüedad léxica, se puede extraer considerando la base de conocimiento FunGramKB[12] como fuente exógena de conocimiento. En la siguiente figura se expone un extracto de la subontología de entidades, con el objetivo de establecer el LCS para los conceptos básicos +CARD_00 y +MAGICIAN_00. El objetivo es determinar la similitud semántica de ambos conceptos, dada la frecuencia con la que tienden a aparecer en determinadas estructuras sintácticas presentes en el corpus en análisis. Este ejercicio considera como ambigua la lexicalización del concepto en español «carta», y su polisemia presente en tanto carta como 'cartulina con dibujos utilizada para juegos' o 'papel escrito que una persona envía a otra para comunicarse'.

Figura 4. Hiperónimo común más cercano para los conceptos CARD y MAGICIAN en la taxonomía ENTITY de FunGramKB.



[12] Para una comprensión pormenorizada de la base conocimiento léxico-conceptual-gramatical FunGramKB (*Functional Grammar Knowledge Base*), se recomienda revisar el punto 3.3 del capítulo 3: Conceptos fundamentales.

Según el ejemplo anterior, SOLID es el LCS para los conceptos CARD y MAGICIAN. En cuanto a la distancia entre rutas, desde CARD hasta SOLID hay tres nodos, mientras que desde MAGICIAN hasta SOLID se observan cinco enlaces. A continuación, se presentan tres medidas basadas en el IC.

2.4.2.2.1 Resnik (1995)

La propuesta de Resnik (1995) es la precursora del concepto de IC. Según esto, se establece que el valor asociado a cada concepto en una jerarquía se basa íntegramente en la evidencia disponible en corpus. A partir de esta aproximación, los conceptos se relacionan de tal manera que un IC alto está ligado a conceptos mayormente especificados. Por otra parte, un IC bajo se relaciona con conceptos menos especificados, o generales. Por ejemplo, el concepto HERRAMIENTA tendrá un IC bajo; mientras que un tipo específico de herramienta, como MARTILLO, tendrá un IC alto, que seguirá aumentando en la medida que existan más subtipos de martillos. La formalización de la medida de Resnik es la siguiente:

$$Sim_{resnik}(S_1, S_2) = IC\big(LCS(C_1, C_2)\big)$$

En términos generales, la medida de Resnik (1995) considera el IC como una cuantificación de la información que dos conceptos tienen en común. Luego se incorpora el número de ocurrencias que un concepto tiene en un corpus en relación con el número de sentidos disponibles para cada concepto. Además, integra el valor del LCS entre ambos conceptos, para determinar la información solapada según su IC. Así, según esta medida, la frecuencia en la que un concepto aparezca en el corpus incluirá la frecuencia de todos sus conceptos subordinados. Entonces, el conteo de los conceptos específicos aporta al resultado de los conceptos genéricos. Esto tiene un impacto relevante en el valor de su probabilidad asociada: a mayor probabilidad de una frecuencia de aparición alta, tendrán un valor bajo de IC y, por tanto, se tratará de conceptos generales, o cercanos a metaconceptos.

2.4.2.2.2 Jiang & Conrath (1997)

En la propuesta de Jiang & Conrath (1997), primero se utiliza el algoritmo de Resnik (1995) para calcular el IC, y luego se incorpora el cálculo de la longitud de distancia entre rutas:

$$Sim_{Jiang\&Conrath}(S_1, S_2) = IC(S_1) + IC(S_2) - 2(IC(LCS(S_1, S_2))$$

Se trata de una medida híbrida en cuanto a los criterios que selecciona para calcular la similitud del *synset*, puesto que incluye la diferencia entre el contenido de información (i.e. $IC(S_1) + IC(S_2)$) y el hiperónimo común más cercano (i.e. $2(IC(LCS(S1, S_2)))$).

2.4.2.2.3 Lin (1998)

Si bien Lin (1998) concuerda con las definiciones de Resnik (1995) en cuanto a las medidas de similitud, propone que la similitud entre dos conceptos no solamente depende de aquello que tienen en común, sino también de sus diferencias. Por tanto, la medida de Lin propone que, mientras más diferencias existan entre un S_1 y S_2, entonces menos similares serán. Luego se establecen dos conceptualizaciones para dar cuenta de los tipos de relaciones que se pueden establecer entre dos *synsets*:

a. *Commonality*: mientras más en común tengan A y B, más similares serán.
b. *Difference*: mientras más diferencias entre A y B existan, menos similares serán.

Según Torres-Ramos (2012), la medida de Jiang & Conrath (1997) es bastante parecida a la propuesta de Lin (1998), en cuanto a la definición del concepto de similitud que subyace a ellas. No obstante, aunque ambas se basan en el cálculo de la cantidad de información necesaria para escribir la información común entre ambos conceptos, fueron postuladas y publicadas de manera independiente[13]. Luego, la formalización de la medida de Lin (1998) es la siguiente:

$$Sim_{Lin}(S_1, S_2) = \frac{2 \times \left(IC\left(LCS(S_1, S_2)\right)\right)}{IC(S_1) + IC(S_2)}$$

Específicamente, esta medida está diseñada para representar la similitud como una función entre el IC dado el LCS, dividido por la suma del IC de ambos.

2.4.3 Métodos basados en conocimiento contextual

Los métodos basados en conocimiento contextual utilizan recursos lingüísticos informatizados previamente anotados o etiquetados para derivar reglas o modelos que realicen el proceso de

[13] Si bien el trabajo de Lin (1998) no hace referencia a la propuesta de Jiang & Conrath (1997), considera como punto de partida fundamentalmente los trabajos ya citados de Resnik (1995), y de Wu & Palmer (1994).

desambiguación léxica automática. En términos generales, este tipo de métodos selecciona un conjunto de muestras en lenguaje natural a partir de un corpus, considerando las distintas clasificaciones de cada elemento. Luego, deben ser capaces de identificar regularidades asociadas a cada elemento, para así generalizar patrones de reglas que serán aplicadas para generalizar los elementos nuevos. A continuación, se exponen dos clasificaciones para los métodos basados en conocimiento contextual: aprendizaje automático supervisado y aprendizaje automático no supervisado. Ambos métodos están basados en el principio del aprendizaje automático, esto es, que la máquina pueda aprender automáticamente a partir de la observación de instancias o datos textuales, con el objetivo de predecir un determinado comportamiento de estos.

2.4.3.1 Aprendizaje automático supervisado

En el aprendizaje supervisado, se dispone de un corpus de entrenamiento previamente etiquetado con el sentido correspondiente para cada instancia de una palabra objetivo. Por ejemplo, en la oración «en la mesa estaba la carta que el mago adivinó», la palabra objetivo corresponderá a la unidad léxica en análisis que presenta ambigüedad. En este caso, «carta» considerando los sentidos posibles: 'papel escrito que una persona envía a otra con la intención de comunicarse' o 'cartulinas rectangulares con ilustraciones que se utilizan en los juegos de azar'. Luego, las palabras de contenido adyacentes conformarán el cotexto o conjunto de palabras contextuales {*mesa*, *estaba*, *mago*, *adivinó*}. Posteriormente, los algoritmos de aprendizaje automático se aplican al corpus de entrenamiento con características contextuales extraídas desde instancias del corpus, considerando los sentidos individuales como clases discretas. La mayoría de los métodos supervisados tradicionales tienen, al menos, cuatro fases de aplicación en común:

a. Selección de un conjunto de datos textuales que muestre las diferentes clasificaciones para cada elemento (valores, atributos, características).
b. Identificación de los patrones asociados con cada elemento.
c. Generalización de patrones.
d. Aplicación de patrones para clasificar nuevos elementos no presentes en el conjunto de datos textuales inicial.

2.4.3.1.1 Algoritmo bayesiano ingenuo (*Naïve Bayes*)

En términos generales, el algoritmo bayesiano ingenuo es un clasificador probabilístico. El objetivo de este tipo de modelos matemáticos es extraer información a partir de conjuntos de datos previamente etiquetados o entrenados, para que la máquina pueda etiquetar automáticamente conjuntos de datos nuevos, o corpus de prueba. Así, el corpus de entrenamiento representa las etiquetas esperadas, mientras que el corpus de prueba es desde el cual se establecen las etiquetas predichas. Este enfoque se denomina 'ingenuo' porque la presencia o ausencia de una característica particular no estará relacionada necesariamente con la presencia o ausencia de cualquier otra, dada la variable original.

Un caso de uso de este algoritmo es el llamado aprendizaje textual. Por ejemplo, luego de determinar una probabilidad previa o *a priori* basada en un corpus de entrenamiento previamente etiquetado, se establece que la probabilidad de que dos sujetos sean clasificados como autores de un correo electrónico es de un 50% cada uno; sean: $P(Jack) = 0{,}5$; y $P(Jill) = 0{,}5$. Siguiendo la lógica bayesiana, a partir de esta probabilidad previa se calculará la probabilidad posterior o a posteriori, que corresponde a la probabilidad de que ocurra el evento A dado que el evento B ha ocurrido; es decir, el resultado de una probabilidad condicional. Luego, la máquina podría enfrentar el siguiente problema: Ante un correo electrónico que contiene el texto «love deal», determinar la probabilidad, tanto para *Jack* como para *Jill*, de su autoría. El procedimiento que, en términos generales, subyace a la resolución de este problema es la incorporación de evidencia de prueba a la probabilidad previa, para obtener así una probabilidad posterior.

Para abordar este problema utilizando las palabras objetivo y las palabras presentes en las instancias disponibles en el texto, es decir, las unidades «love» y «deal», es necesario extraer información de un corpus de entrenamiento previamente etiquetado. En este corpus, la frecuencia de aparición de las palabras objetivo, i.e., su relevancia estadística, es la siguiente:

$$Jack: love = 0.1;\ deal = 0.8$$
$$Jill: love = 0.5;\ deal = \ 0.2$$

Entonces, la probabilidad posterior (i.e. su probabilidad condicional) será una función entre el peso estadístico de cada palabra objetivo y la probabilidad previa de cada posible autor:

$$postprob = freq_{obj1} \times freq_{obj2} \times priorprob$$

Lo anterior, se aplica como sigue para cada caso:

$$0.04 = 0.1 \times 0.8 \times 0.5$$
$$por\ lo\ \text{tan}to, P(Jack \mid \text{«}love\ deal\text{»}) = 0.04$$

$$0.05 = 0.5 \times 0.2 \times 0.5$$
$$por\ lo\ \text{tan}to, P(Jill \mid \text{«}love\ deal\text{»}) = 0.05$$

En conclusión, existe un 50% de probabilidad de que *Jill* sea la autora del correo cuyo contenido textual es «love deal», que se puede leer como la probabilidad del *Jill* ante el evento «love deal». De esta forma, se ha identificado, a partir de una fuente de texto, si la etiqueta *Jack* o la etiqueta *Jill* es la más probable, ignorando el orden de las palabras y considerando la frecuencia como una manera de establecer la clasificación.

La aplicación del algoritmo bayesiano ingenuo en la desambiguación léxica automática, implementado por primera vez por Gale *et al.* (1992), está basada en un modelo matemático de dependencias entre los sentidos de palabra y un conjunto de características presentadas en un recurso lingüístico informatizado; es decir, cada una de sus características constituye una probabilidad independiente. La afirmación anterior tiene dos consecuencias relevantes: la primera es que se ignora la sintaxis y el carácter lineal de las palabras dentro del cotexto, lo que deriva en el llamado modelo de bolsa de palabras. La segunda es que la presencia de una palabra en esta bolsa es independiente de la presencia de otra, lo que no es cierto en el caso de las lenguas naturales. Sin embargo, a pesar de estos supuestos simplificadores, y como se señala en los trabajos de Manning & Schütze (1999), se ha demostrado que este modelo es bastante efectivo desde una perspectiva cognitiva, que es adecuado en el caso de un problema relacionado con el PLN.

En efecto, el enfoque para la desambiguación léxica automática en el que se basa el modelo bayesiano ingenuo representa un enfoque teórico relevante en el ámbito del procesamiento estadístico del lenguaje. La idea del clasificador bayesiano en el contexto de la desambiguación léxica automática es observar palabras objetivo alrededor de una bolsa de palabras contigua en una determinada ventana contextual. Así, cada palabra de contenido dentro de la ventana contextual aportará información relevante acerca del sentido de la palabra ambigua. Este algoritmo se utiliza ampliamente debido a su eficiencia y su capacidad para combinar evidencia de una gran cantidad de características (Escudero

et al., 2000; Aung *et al*., 2011; Fulmari & Chaldak, 2014; Gamallo *et al*., 2014; Gosal, 2015). Es aplicable si el estado de cosas del mundo en el que se basa una clasificación se describe como una serie de características o atributos utilizados para la descripción, y que a su vez son condicionalmente independientes.

Finalmente, el proceso de desambiguación se realiza utilizando la regla de decisión de *Bayes*: se calcula la puntuación de cada sentido de una palabra ambigua y decide el sentido más apropiado para una palabra específica en la oración de prueba. de la siguiente manera:

$$P(sense|feature) = \frac{P(sense) \times P(feature|sense)}{P(feature)}$$

Según Fulmari & Chaldak (2014), mediante la aplicación del supuesto de ingenuidad, el algoritmo se reduce a:

$$\underset{S_1 \in senses(w)}{argmax} P\left(S_i \prod_{j=1}^{m} P\left(f_j|S_i\right)\right),$$

donde f_j representa el vector de características o atributos, mientras que S_i representa el sentido de una palabra en particular. Por lo tanto, el sentido correcto de una palabra será el sentido con el valor de probabilidad condicional más alto. El algoritmo bayesiano, en este caso, demuestra ser ingenuo porque ignora el orden de las palabras; es decir, no logra incorporar realmente las variables del contexto oracional que se utilizan como *input*. A pesar de esto, se trata de un modelo que ha demostrado ser sencillo de implementar y eficiente para el procesamiento de recursos lingüísticos informatizados extensos. No obstante, la calidad de la clasificación estará supeditada a la necesidad de incorporar una mayor cantidad de fuentes de información lingüística. Aún así, el proceso se basa en parámetros exclusivamente estocásticos, y, al realizar la clasificación, la máquina solamente es capaz observar la frecuencia tanto de las palabras objetivo como del contexto oracional.

Mooney (1996) realiza una comparación experimental entre diferentes algoritmos de aprendizaje automático que se utilizan para resolver la desambiguación léxica. Distingue a los algoritmos a partir de las diversas técnicas que utilizan para la clasificación de palabras considerando el contexto oracional: algoritmo bayesiano ingenuo, de redes neuronales, árboles de decisión, basados

en reglas, y basados en casos. El método bayesiano, junto con las redes neuronales, muestra un desempeño más eficiente que otras técnicas para la desambiguación de palabras objetivo en un corpus de prueba. En efecto, se demostró que el algoritmo bayesiano ingenuo obtuvo un promedio de precisión del 71% para una tarea de desambiguación automática de seis sentidos de la unidad léxica «list», con 1.200 ejemplos como corpus de entrenamiento. En base a estos resultados el estudio señala que, si bien se consideraron diferentes algoritmos de aprendizaje automático que pueden funcionar de manera bastante similar, todos presentan sesgos específicos en su representación conceptual. Según lo anterior, el método bayesiano ingenuo en particular requiere de un esfuerzo significativo durante el proceso de entrenamiento del corpus y de identificación de los rasgos necesarios para la clasificación.

La propuesta de Carpuat & Wu (2005) presenta una prueba para la evaluación de métodos de desambiguación léxica automática basados en el aprendizaje automático, con el objetivo de ser utilizados en un traductor automático chino-inglés. Los autores valoran positivamente la utilización de los inventarios de sentidos o corpus de entrenamiento, y la inclusión de información lingüística mediante *rasgos* o características. El estudio se basa en corpus entrenado. A partir de este, se comparan cuatro modelos estocásticos: clasificador bayesiano ingenuo, modelo de máxima entropía, *Boosting model*, y *Kernel PCA-based model*. El algoritmo bayesiano ingenuo, a pesar de su ya mencionada simpleza matemática, resultó ser significativamente más eficiente para las tareas específicas de desambiguación léxica automática en el ámbito de la traducción automática.

En la revisión de métodos para la desambiguación léxica automática, Widlak (2004) declara que el algoritmo bayesiano ingenuo es altamente competitivo en el ámbito y, en algunos casos, logra superar a otros algoritmos de aprendizaje automático. En efecto, se trata de un algoritmo que ha sido utilizado con éxito para tareas generales de clasificación textual. Como se mencionó anteriormente, esta técnica está basada en la suposición de independencia condicional: la ocurrencia de una palabra en un texto, dada una clase *x*, es independiente de la ocurrencia de cualquier otra palabra en el mismo texto dada la misma clase *x*. La crítica fundamental a esta suposición proviene desde la lingüística teórica y la ciencia cognitiva (Eberhardt & Danks, 2011) pues, en la realidad de las lenguas naturales, la suposición de independencia condicional es incorrecta en tanto las palabras dependen unas de otras y se influencian mutuamente en distintos niveles de análisis lingüístico. No obstante, la misma suposición ha demostrado ser, desde el punto de vista probabilístico, una ventaja (Rish, 2001).

En resumen, la mayoría de las propuestas de métodos de desambiguación léxica automática basadas en el algoritmo bayesiano ingenuo exponen puntos en común relevantes, también entendidos como ventajas en relación con los métodos basados en métricas. Se pueden resumir en tres puntos:

a. Es un algoritmo llamado sencillo o simplista, dentro la gama de posibilidades para el aprendizaje automático.
b. Es posible incluir un alto número de *rasgos* o características para poder capturar información lingüística que sea necesaria en el proceso de elección de probabilidad; es decir, este método no se limita a la información que provee el cotexto, y puede considerar criterios de análisis provistos por un lingüista.
c. Tiene un desempeño consistentemente sobresaliente, pero que depende de las características del corpus y del número de rasgos.

2.4.3.2 Aprendizaje automático no supervisado

A diferencia de su contraparte, el aprendizaje automático no supervisado no depende de un recurso lingüístico informatizado que haya sido previamente anotado con las estructuras o rasgos que el algoritmo de clasificación pretende producir como valores de salida. En este caso, el algoritmo es provisto solamente con los datos que provienen del conocimiento contextual; es decir, de las instancias en análisis A partir de esa información, debe analizar estructuras lingüísticas mediante la identificación de patrones textuales de distribución y de propiedades de agrupación de los rasgos que emergen desde los datos (Popescu & Hristea, 2010; Ustalov *et al*., 2018).

El surgimiento de estos métodos de aprendizaje automático no supervisado, principalmente a partir de la década de 1990, se debe principalmente a que los corpus etiquetados manualmente, o previamente entrenados por un humano, agregan una dificultad significativa al procedimiento. Por lo tanto, este costo debe compararse con la precisión que proporcionaría la utilización de métodos supervisados. En la medida en que los algoritmos no supervisados no incurran en estos costos, ofrecen una ventaja importante solo si son capaces de mantener un nivel aceptable de rendimiento en las aplicaciones para las que están diseñados.

Según lo anterior, el cuello de botella de la adquisición del conocimiento en desambiguación léxica automática se ha mantenido como un problema relevante en el ámbito de la representación del conocimiento y el desarrollo de métodos de aprendizaje automático. En efecto, la principal dificultad al comparar métodos de aprendizaje supervisado y no supervisado es que los algoritmos supervisados

a menudo necesitan de una gran cantidad de instancias previamente etiquetadas con sentidos relevantes para llevar a cabo el proceso de desambiguación. No obstante, y como ya se mencionó, el procedimiento de etiquetar manualmente corpus extensos con información de sentidos de palabras requiere mucho tiempo, entendido este como un recurso relevante para el esfuerzo investigativo, a la vez que resulta altamente propenso a errores o sesgos. Por ejemplo, en el trabajo de Yarowsky (1995) se reporta que el investigador demoró aproximadamente tres años en etiquetar a mano 37.232 ejemplos, mientras que Ng (1997) incluso proyectó un intervalo de dieciséis años como el tiempo en el que un humano demoraría en etiquetar con sentidos de palabras todo el *Corpus Brown*, que cuenta con aproximadamente 3,2 millones de entradas (*tokens*).

En los métodos de aprendizaje automático no supervisados, el hecho de que no se utilice información etiquetada manualmente de manera previa también plantea una serie de desafíos. Estos se dan sobre todo en torno a la evaluación de sus resultados al implementar algoritmos basados en agrupaciones (*clustering*). Según Jurafsky & Martin (2009) las siguientes corresponden a las desventajas más relevantes para los métodos no supervisados:

a. Es altamente probable que no se conozcan los sentidos correctos de las instancias utilizadas en los enfoques supervisados.
b. Los grupos que derivan de la clasificación tienden a ser heterogéneos con respecto a los sentidos de las instancias contenidas en el corpus.
c. El número de grupos resultante es, en la mayoría de los casos, diferente del número de sentidos de las palabras objetivo que se desambiguan.

Según lo revisado en este capítulo, el problema lingüístico de la ambigüedad léxica se puede abordar, en primer lugar, desde una perspectiva cognitivista y holística, que entiende la mente como un proceso de estrategias cognitivas de comprensión y pensamiento. Así, la desambiguación formaría parte de un sistema dependiente de las opciones del hablante, por un lado, y la manera en la que estas opciones se relacionan con el conocimiento de mundo y el sentido común, por otro. En segundo lugar, el enfoque computacional aborda este problema desde la definición de la mente como una máquina modular, que reduce la manifestación de la capacidad cognitiva a patrones probabilísimos. En este sentido, el proceso de desambiguación léxica se encontraría aislado de variables externas, en la medida que las lenguas naturales pueden ser descritas como modelos formales con el potencial de ser manipulados a partir de la aplicación de métodos estadísticos.

En conclusión, para el PLN, el problema de la desambiguación léxica automática se ha estado abordando, y con relativo éxito, desde la década de 1980. Sin embargo, aunque se trate de un problema antiguo, aún existe un campo de desarrollo relevante en el que se proponen distintas maneras de mejorar el proceso de automatización. Por otra parte, para la lingüística, es el problema mismo de la desambiguación léxica automática lo que aporta una perspectiva novedosa desde este ámbito aplicado e interdisciplinar, puesto que los modelos teóricos que son capaces de describir y explicar el fenómeno ya son bastante satisfactorios, aunque inaplicables por sí solos computacionalmente.

Este ámbito interdisciplinar supone una perspectiva altamente valiosa, puesto que, si bien los métodos estocásticos son muy eficientes para tareas de PLN, no necesariamente estarían abordando la reproducción de los patrones con los que funciona la mente humana. Por el contrario, los modelos probabilísticos no representan la manifestación de una capacidad cognitiva humana como el lenguaje, sino que en realidad corresponden a un conjunto de métodos que facultan a las máquinas para examinar las lenguas naturales, con el objetivo de imitar, y no reproducir, la capacidad humana de comprender el lenguaje. Esta paradoja, en definitiva, es una motivación para el trabajo colaborativo tanto de lingüistas como de ingenieros informáticos, cuyos objetivos de investigación comunes nos permitirán comprender de manera más exhaustiva la mente humana, en el esfuerzo de reproducir artificialmente sus habilidades cognitivas. Un ejemplo de esto es, sin dudas, el problema de la desambiguación léxica automática.

Capítulo 3
Conceptos fundamentales

El propósito de este tercer capítulo es ofrecer una revisión de ciertos conceptos fundamentales relacionados con el tratamiento y procesamiento de datos textuales. Primero, se establece una panorámica cronológica de la utilización del corpus en el análisis lingüístico, junto con una caracterización de los llamados recursos lingüísticos informatizados. Luego se revisa la fundamentación teórica y la arquitectura de representación del conocimiento que propone FunGramKB, junto con su correspondiente lenguaje de representación conceptual. Posteriormente, se establece una síntesis de los aspectos centrales del lexicón mental, como un modelo que vincula los estudios cognitivos con los patrones lingüísticos mediante los que se representa el conocimiento. Por último, se describe la tecnología informática básica que suele emplearse en la construcción y exploración de recursos lingüísticos informatizados, y que de hecho se utilizó para el montaje del corpus durante el desarrollo de esta investigación: expresiones regulares, lenguaje de etiquetado extensible (XML), y lenguaje de consulta estructurada (SQL).

3.1 La utilización de corpus en el análisis lingüístico

Los antecedentes del concepto de corpus y su utilización para la investigación lingüística, anteriores al siglo XIX, se focalizaron en la recolección de muestras. Una definición preliminar de corpus, según Villayandre (2008), sería un conjunto de textos escritos, cuya finalidad es el estudio de lenguas muertas, como el latín o el sánscrito. En el campo de la lingüística de corpus, una definición más satisfactoria y específica para el ámbito es provista por Santalla (2005):

> Un corpus es un conjunto de textos de lenguaje natural e irrestricto, almacenados en un formato electrónico homogéneo, y seleccionados y ordenados, de acuerdo con criterios explícitos, para ser utilizados como modelo de un estado o nivel de lengua determinado, en estudios o aplicaciones relacionados en mayor o menor medida con el análisis lingüístico (p. 45-46).

Por otra parte, la observación de datos para el análisis lingüístico con alcances metodológicos comenzó a desarrollarse a partir de la primera mitad del siglo XX. En efecto, con el advenimiento de

la lingüística moderna se aborda la descripción de la lengua[14] como norma de todas las manifestaciones lingüísticas, y al mismo tiempo como objeto de estudio de las ciencias del lenguaje (Saussure, 2003 [1916]). Según esta perspectiva, una aproximación exhaustiva a este componente del lenguaje debía estar basada en información extraída de muestras, tanto orales como transcritas, a partir de las que fuera posible derivar información fonético-fonológica y morfosintáctica de una lengua. A partir de esta propuesta, el corpus constituía el único acceso posible a la observación y descripción de los componentes de un sistema lingüístico particular.

Más adelante, con la irrupción del generativismo, que se sitúa en la publicación de *Syntactic Structures* (Chomsky, 1957) y particularmente *Current Issues in Linguistic Theory* (1964), comienzan a cuestionarse los métodos de la propuesta estructuralista para el análisis lingüístico. Particularmente, la metodología empirista o basada en corpus se ve desprestigiada y reemplazada por el método racionalista, fundamentado en las intuiciones lingüísticas del analista y la validación de datos lingüísticos justificados a partir de su propia competencia como hablante. En términos generales, la crítica hacia el enfoque de corpus está basada en una concepción del lenguaje como la manifestación de una capacidad cognitiva accesible mediante una orientación internista y naturalista. Por tanto, los datos para describir el comportamiento de las estructuras lingüísticas debían centrarse en la identificación de principios. Este foco en el conocimiento interiorizado de la lengua invalida la utilización de corpus, en tanto los considera como una manifestación externa de la lengua sujeta a variaciones y desviaciones de la norma. Además, en términos prácticos, se critica la parcialidad de los corpus como recurso finito y estático, debido a que no pueden dar cuenta de la naturaleza potencialmente infinita de las lenguas naturales, en coherencia con su capacidad generativa.

Con el advenimiento de las ciencias informáticas aplicadas a la recopilación de información lingüística a partir de la década de 1960, y contemporáneo a la hegemonía del programa generativista en la teoría lingüística, comienza el desarrollo del primer corpus informatizado: el *Corpus Brown*[15]. Este se define como un diccionario legible por la máquina en lengua inglesa. Si bien esta innovación fue resistida en principio por la comunidad científica, por no adscribir a la idea de que la única fuente legítima del conocimiento gramatical de una lengua se encontraba en las intuiciones de los hablantes

[14] Desde una perspectiva estructuralista, la lengua entendida como un componente del lenguaje: sistema de signos lingüísticos con valor relativo.

[15] Una descripción pormenorizada de los componentes del *Corpus Brown*, que incluye sus categorías y subdivisiones, se encuentra disponible en: http://korpus.uib.no/icame/manuals/brown/index.html.

nativos, varios lingüistas sostuvieron la utilización de corpus tanto para los estudios teóricos como para el análisis descriptivo de las lenguas (Meyer, 2002).

Esta disputa, a la vez teórica y metodológica, fue decisiva para, finalmente, marcar una tendencia hacia el futuro respecto a la importancia de que el análisis lingüístico se llevase a cabo a partir de instancias de habla o de escritura auténticas, y no basada en datos inventados o proporcionados de manera *ad-hoc* por el lingüista.

3.2 Desde el corpus hacia los recursos lingüísticos informatizados

Convencionalmente, se define corpus como "una muestra amplia de lengua escrita o hablada que se considera representativa o bien del estándar o de alguna variante diatípica o diatópica, o de algún período histórico determinado" (Lavid, 2005: 62). Esta aproximación se puede expandir hacia el ámbito computacional, según McEnery & Wilson (1996), como una colección finita de textos legibles por la máquina que son representativos de una lengua, o de un estado particular de una lengua. La proliferación de sistemas informatizados para el análisis lingüístico derivó, a partir de la década de 1970, en la utilización de corpus para investigaciones de PLN que involucraron modelar el comportamiento lingüístico. Junto con esto, el potencial computacional para la recopilación de información lingüística dio paso a diferentes tipos de recursos lingüísticos informatizados, que serán definidos más adelante. A saber: corpus (monolingüe o bilingüe), lexicón, glosario, taxonomía y ontología.

Desde la lingüística de corpus, para la creación y validación de un corpus que se utilizará en investigación lingüística debe tenerse en cuenta el cumplimiento de ciertos estándares (Dash, 2010), entre los cuales destacan:

a. Cantidad: debe ser una muestra de gran tamaño como suma de sus componentes, ya sea escrito, oral o transliterado.
b. Calidad (equivalente a autenticidad): las instancias, o textos, deben ser obtenidas de muestras de habla o de escritura auténticas; es decir, sin considerar aquellas emisiones que puedan ser producto de condiciones experimentales o circunstancias artificiales.
c. Representatividad: es necesario que se incluyan instancias provenientes de una amplia diversidad de textos. Lo anterior implica que el corpus debe balancearse para cubrir diferentes usos lingüísticos que sean representativos de un ámbito y/o lengua particular.

d. Recuperabilidad: los datos proporcionados, así como los ejemplos y las referencias, deben ser de fácil acceso por parte de los usuarios. Este estándar sugiere la necesidad de preservar y unificar determinadas técnicas para almacenar, compartir y presentar datos lingüísticos en formato electrónico o asistido por computadora.

A partir de estos estándares, queda en evidencia que el concepto de corpus se ha definido predominantemente a partir de la descripción de las características, la validez y el tratamiento de los datos que una colección de textos es capaz de almacenar y proporcionar. Sin embargo, actualmente las definiciones de corpus más influyentes dentro de las ciencias del lenguaje se orientan hacia los entornos informatizados. Así, según Leech (1997), corpus se define como:

> "[…] a body of naturally-occurring (authentic) language data which can be used as a basis for linguistic research. This body of data may consist of written texts, spoken discourses, or samples of spoken and/or written language. Often it is designed to represent a particular language or language variety. In the past thirty-five years, the term corpus has been increasingly applied to a body of language material which exists in electronic form, and which may be processed by computer for various purposes such as linguistic research and language engineering" (p. 1).

Esta aproximación sugiere entonces que el corpus, dada su aplicación hacia herramientas computacionales, se establezca en realidad como un término parcial, o subtipo de una categoría más inclusiva y apropiada: los recursos lingüísticos informatizados. Estos, a su vez, se definen como cualquier conjunto de datos, ya sea oral o escrito, que se construya en un formato legible o además tratable por un ordenador.

En términos generales, los recursos lingüísticos informatizados pueden pertenecer a dos categorías que, a su vez, integrarán cada uno de los subtipos de recurso. En primer lugar, nos referiremos a la distinción entre recursos legibles para la máquina (RLPM) y recursos tratables para la máquina (RTPM). En el primer caso, los RLPM consideran cualquier recurso que contenga conocimiento lingüístico, seleccionado y organizado por un lexicógrafo, y que sea legible por una computadora para su posterior acceso y consulta mediante una interfaz gráfica de usuario. Algunos ejemplos de esto son las versiones en línea del *Longman Dictionary of Contemporary English* (Procter, 1978), o el *Diccionario de la Lengua Española* (Real Academia Española, 2014).

Por otra parte, un RTPM se define como un conjunto de datos lingüísticos que, dado el formato de su colección de documentos, puede ser utilizado directamente en tareas de PLN, y cuyo formato permitirá la realización de tareas de investigación lingüística mediante la utilización de algunas herramientas técnicas. Algunos ejemplos de estas herramientas técnicas son el lenguaje de etiquetado XML, el lenguaje de consulta estructurada SQL o las expresiones regulares.

La aproximación anterior, de carácter más amplio, está basada en la distinción tradicional entre diccionarios legibles por la máquina y diccionarios tratables por la máquina, según los trabajos de Amsler (1982), Amsler & Whim (1979), y Wilks *et al*. (1989). En esta, tanto los RLPM como los RTPM pueden pertenecer a determinados tipos de conjuntos de datos lingüísticos, entendidos como formas de clasificación del conocimiento:

a. Corpus (mono o bilingüe): conjunto de textos o instancias de lenguaje natural que han sido seleccionados a partir de criterios determinados y explícitos, con el fin de ser representativos de un estado particular de la lengua. Pueden ser monolingües; esto es, especializados en la lengua meta, o bien bilingües, como una representación de dos o más lenguas. Uno de los ejemplos más representativos en lengua española es el CREA: *Corpus de Referencia del Español Actual* (Real Academia Española, 2008), y en lengua inglesa el ya citado *Corpus Brown* (Francis & Kučera, 1964). Además, como se indicó en el capítulo anterior, en esta investigación se utilizará una muestra del CODICACH: *Corpus Dinámico del Castellano de Chile*, desarrollado por Sadowsky (2006).
b. Lexicón: colección de toda la información lingüística que se encuentra contenida en una palabra, entendida a su vez como objeto de una lengua natural que contiene determinados significados y propiedades, tanto morfológicas como sintácticas. Un ejemplo de lexicón corresponde a la *Base de Datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español* (ADESSE), desarrollada por García-Miguel *et al*. (2010).
c. Glosario: repositorio de palabras y sus respectivas definiciones. Puede ser de dominio general o específico, como glosarios terminológicos o técnicos. Dos ejemplos representativos de glosarios son el proyecto IATE: *Interactive Terminology for Europe*, correspondiente a un glosario terminológico plurilingüe (Unión Europea, 2018), y el *Diccionario de la Lengua Española*, en su edición digital publicada por la Real Academia Española (2014).

d. Taxonomía: colección de información que constituye una jerarquía conceptual. Se organiza en niveles a partir de determinados nodos y puede proveer de relaciones léxicas de significado, como la hiponimia/hiperonimia, también llamadas relaciones del tipo *IS-A*. Ejemplos de taxonomías son *WordNet* (Miller, 1985; Miller *et al*., 1993; Fellbaum, 1998), *MultiWordNet* (Pianta *et al*., 2002), y *Spanish FrameNet* (Subirats & Petruck, 2003; Subirats, 2004).
e. Ontología: en términos generales, refiere a un catálogo de conocimiento derivado por la experiencia de mundo, con un alto grado de prototipicidad. En cuanto a la relación entre lenguaje e informática, se define como un modelo jerárquico para describir el mundo a partir de una base de conocimiento que determina propiedades y relaciones entre las unidades conceptuales de una lengua natural. Algunos ejemplos de ontologías corresponden a *CYC Project* (Matuszek *et al*., 2006) y *FunGramKB* (Periñán-Pascual & Arcas-Túnez, 2007; 2010; Periñán-Pascual, 2012b).

3.3 La base de conocimiento léxico-conceptual-gramatical FunGramKB

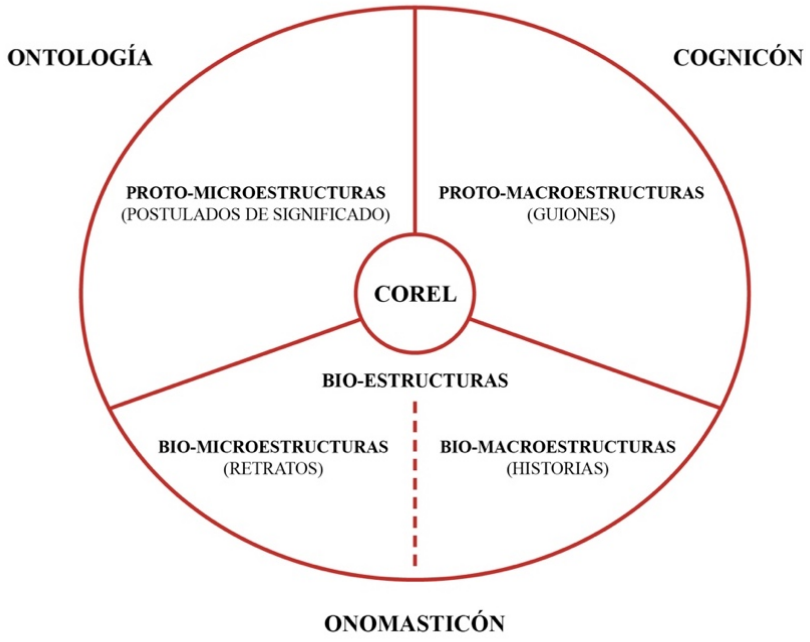
FunGramKB, *Functional Grammar Knowledge Base* (Periñán-Pascual & Arcas-Túnez, 2007; 2010; Periñán-Pascual, 2012b), es una base de conocimiento léxico-conceptual-gramatical multipropósito, cuyo objetivo es el procesamiento de las lenguas naturales. FunGramKB tiene como punto de partida, según Periñán-Pascual & Mairal-Usón (2009), la premisa de que los sistemas complejos para PLN deben cumplir con el objetivo de administrar e interpretar información lingüística. Según lo anterior, estos sistemas deben estar fundamentados en teorías lingüísticas que sean capaces de capturar eficientemente generalizaciones sintáctico-semánticas. De este modo, nos encontramos con que la llamada ingeniería del conocimiento puede servirse de la lingüística en tanto requiere de modelos sólidos para el desarrollo de implementaciones computacionales.

FunGramKB se puede categorizar como una base de conocimiento, en oposición a una base de datos léxica, porque almacena el conocimiento a través de un lenguaje formal, en este caso el lenguaje de representación conceptual COREL, que será explicado más adelante. Es decir, todas las convenciones para la descripción de significados pueden ser controladas mediante una formalización unívoca y bien delimitada, que establezca a su vez relaciones jerárquicas de herencia, y posea un sistema de notación autónomo. Además, una base de conocimiento como esta puede contener conocimiento conceptual, declarativo y procedimental. En este punto, los esquemas conceptuales de FunGramKB, correspondientes a su vez a un enfoque conceptualista; es decir, que transita desde el

concepto a la referencia, están basados en el modelo de memoria a largo plazo de Tulving (1985). Específicamente, en su categorización para la adquisición del conocimiento que proviene del sentido común. Tulving clasifica la memoria en tres tipos: la memoria semántica, que almacena cognitivamente todos los rasgos constitutivos del lexema; la memoria episódica, que almacena el conocimiento biográfico, ya sea propio o ajeno; y la memoria procedimental, que almacena el cómo se llevan a cabo ciertos procesos cotidianos, y cómo los percibimos.

En cuanto a la aplicación de estos alcances teóricos en FunGramKB, existen tres niveles para el almacenamiento de los conceptos. Cada uno de estos niveles contiene a su vez diferentes módulos conceptuales. Primero, el nivel léxico lo integran el lexicón, el morficón y el constructicón. Estos módulos corresponden a las unidades léxicas en sus relaciones sintácticas y pragmáticas, así como todas sus flexiones gramaticales. El segundo es el nivel conceptual. Si bien en general cada módulo es específico para cada lengua, el nivel conceptual es una representación lingüística abstracta entendida como un marco semántico aplicable para todas las lenguas. En este nivel, los tres tipos de conocimiento de Tulving tienen su correlato en cada uno de los tres módulos conceptuales: ontología, cognicón y onomasticón. La ontología almacena el conocimiento que viene dado por la experiencia de mundo en un alto nivel de prototipicidad. Contiene eventos, entidades y cualidades. Luego el cognicón contiene el conocimiento procedimental como instancias de eventos o entidades. Finalmente, el onomasticón incluye el conocimiento episódico, es decir, la especificación de eventos y personajes particulares. En la siguiente figura, adaptado de Periñán-Pascual & Mairal-Usón (2010a), se presenta un esquema del nivel conceptual:

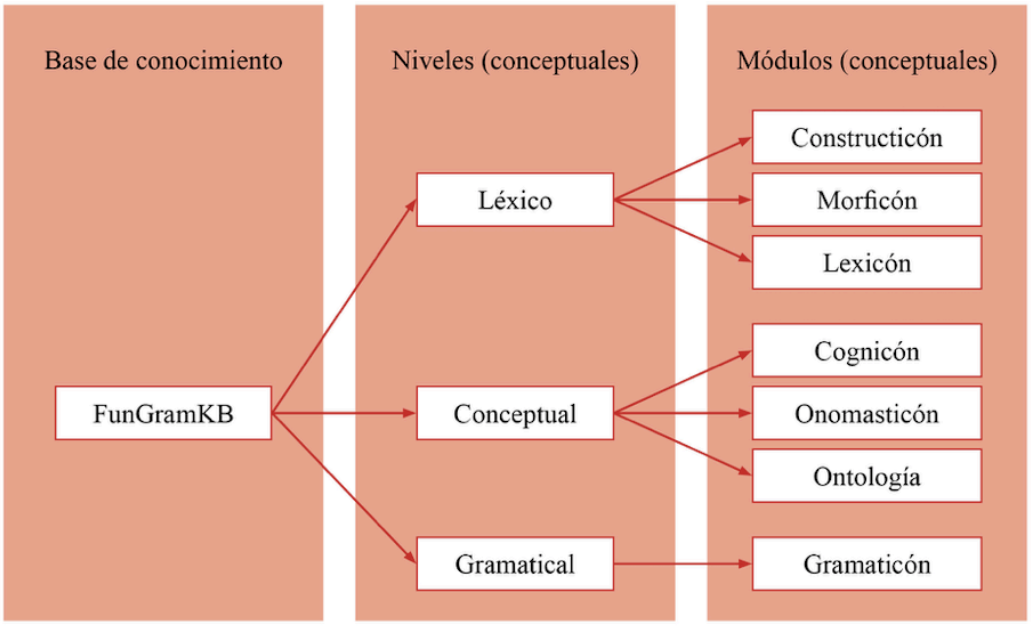
Figura 5. El planeta cognitivo para el nivel conceptual de FunGramKB, adaptado de Periñán-Pascual & Mairal-Usón (2010a).



En resumen, la ontología almacena postulados de significado (proto-microestructuras), el cognicón almacena guiones (proto-macroestructuras), y el onomasticón almacena retratos e historias particulares (bio-microestructuras y bio-macroestructuras).

El tercer nivel es el gramatical. Este nivel contiene el módulo conceptual gramaticón, que sistematiza y clasifica gramaticalmente cada ítem léxico a partir de los postulados de la Gramática del Papel y la Referencia (*Role and Reference Grammar*, abreviado RRG). La base de conocimiento pretende ser multilingüe, es decir, aun cuando se base solamente en lenguas indoeuropeas, a razón de buscar homogeneidad en cuanto a la cosmología y ontología, se aspira a un procesador multifuncional. Por lo pronto, son el español y el inglés las lenguas más pobladas, pero se aspira a la incorporación de cualquier lengua. Lo anterior se puede sistematizar en:

Figura 6. Niveles conceptuales en FunGramKB, adaptado de Periñán-Pascual & Mairal-Usón (2009).



3.3.1 Aspectos generales de la Gramática del Papel y la Referencia

La Gramática del Papel y la Referencia (Van Valin & La Polla, 1997; González Vergara, 2006; Pavey, 2010) constituye un modelo de corte estructural-funcionalista, que define el fenómeno del lenguaje como un sistema de acción comunicativa y de carácter social. Lo anterior recupera la formalización del sistema lingüístico como un sistema de signos regulados por el valor y, por tanto, anclado en una perspectiva renovada del estructuralismo. Asimismo, la perspectiva comunicativa de acción social establece un vínculo explícito con el funcionalismo, en tanto el lenguaje se entiende como un medio orientado hacia un fin. En este sentido, la definición de lenguaje de la RRG no determina una reducción de conceptos orientados al discurso o el contexto. Además, este modelo no considera a la sintaxis como un componente autónomo de la gramática, sino que establece una perspectiva de interfaces en la cual el sistema lingüístico supone una interacción entre los niveles sintáctico, semántico y pragmático.

Además, la RRG se plantea como un modelo multilingüe. Esto se manifiesta en los intereses por la adecuación tipológica y por desarrollar descripciones y explicaciones del comportamiento de distintas lenguas. De hecho, uno de los planteamientos iniciales fue la de proveer una teoría cuyo punto de partida no fuese el inglés, sino lenguas originarias o no consideradas en el análisis lingüístico tradicional. Esto ha sido de vital importancia para demostrar la adecuación explicativa del modelo y

su descripción de las propiedades que gobiernan tanto el lenguaje en general como las distintas lenguas en particular. También, se plantea una perspectiva comunicativa y cognitiva para comprender las lenguas humanas como medios de comunicación. Esto implica que las estructuras gramaticales solamente pueden ser comprendidas y explicadas en términos de las condiciones impuestas por el uso, pero sin reducir el lenguaje a la explicación del discurso.

En definitiva, se resumen tres rasgos propios de la RRG que implican la superación de la tensión expuesta entre lo formal y lo funcional. En primer lugar, la RRG es un modelo en el que las estructuras morfosintácticas, las reglas gramaticales y las representaciones semánticas son siempre explicadas en relación con su función comunicativa. Lo anterior implica que la lengua se entiende como un sistema abstracto, pero desde una interfaz sintáctica-semántica-pragmática. En segundo lugar, la RRG propone una teoría en la que los componentes sintáctico y semántico se encuentran directamente conectados a través de un algoritmo de enlace. Esto implica que la sintaxis no se establece como el aspecto fundamental del lenguaje ni lo especifica como un componente autónomo. En tercer lugar, la RRG es un modelo teórico que plantea un criterio externo de validación teórica. Por lo tanto, se enmarca como un modelo no anglocéntrico que surge a partir del análisis de lenguas con estructuras gramaticales diversas. Esto quiere decir que el aparato teórico de la RRG debiera satisfacer las realizaciones de distintos sistemas comunicativos, sin alterar los componentes formales del modelo.

3.3.2 Lenguaje de Representación Conceptual (COREL)

Conceptual Representation Language **(**COREL, por sus siglas en inglés) es el sistema de notación para la base de conocimiento FunGramKB. Este lenguaje de representación es una actualización del modelo del Lexicón Generativo, específicamente de la estructura de *qualia*, a partir de los postulados de la Gramática del Papel y la Referencia. Es decir, al igual que los *qualia* de Pustejovsky, también en COREL aparecen los papeles temáticos como antecedente teórico común. La estructura de *qualia* es una diferenciación estructural de la fuerza predicativa o relacional de un ítem léxico, en la que se relacionan las estructuras de evento y de argumento. Este modelo puede dar cuenta de las diferentes lecturas o significados que puede capturar un ítem léxico dependiendo de la estructura sintáctica. Podemos entender los *qualia* como "[…] un conjunto de propiedades o eventos asociados a un ítem léxico que mejor explican qué quiere significar una palabra." (Pustejovsky, 1995: 77). La estructura de *qualia* está definida por cuatro *qualia* o papeles: (1) *quale* formal: es la categoría básica que

distingue un objeto con relación a un dominio más amplio. Da cuenta de la orientación, magnitud, forma, dimensión, color y/o posición; (2) *quale* constitutivo: la relación entre el objeto y sus partes constituyentes. Da cuenta del material, peso y otras partes o componentes del objeto; (3) *quale* télico: el propósito y la función del objeto, si es que la hay. Puede dar cuenta tanto del propósito que un agente lleva a cabo en una acción, como de la meta que la ejecución de cierta actividad involucra; y (4) *quale* agentivo: los factores involucrados en el origen o devenir del objeto. Da cuenta del creador, artefacto, naturaleza y/o cadena causal del objeto.

En cuanto a la definición de FunGramKB, se trata de una aplicación computacional centrada en sistematizar el conocimiento proveniente del sentido común: "Todas estas definiciones sirven de aducto a un motor de razonamiento que permite a la máquina simular los patrones de razonamiento humanos y así sacar conclusiones utilizando el mismo conocimiento no especializado sobre las cosas cotidianas de la vida […]" (Periñan-Pascual & Mairal-Usón, 2010b: 8). La base de conocimiento está poblada por tres tipos de unidades conceptuales. Primero, los metaconceptos muestran los grandes ámbitos conceptuales para cada unidad, que en la mayoría de los casos coinciden con las unidades ontológicas del nivel superior. Cada metaconcepto posee un marco temático que heredará a sus conceptos hijo o conceptos básicos. Los metaconceptos están precedidos por el símbolo (#), como en #ENTITY, #EVENT y #QUALITY. Luego, los conceptos básicos son conceptos específicos que se utilizan para definir otros conceptos. Cada concepto básico posee un marco temático y un postulado de significado. Han sido identificados desde el *Longman Dictionary of Contemporary English* (Procter, 1978), como resultado de "[…] varias fases de proyección: conceptualización, jerarquización, remodelación y refinamiento" (Periñán-Pascual & Mairal-Usón, 2010a: 16). Los conceptos básicos están precedidos por el símbolo (+), como en +FEEL_00 para eventos, +ATTRIBUTE_00 para entidades, y +FAST_00 para cualidades.

Finalmente, los conceptos terminales son aquellos conceptos que difieren de los conceptos básicos en tanto especifican algunos de sus aspectos. En este sentido, los conceptos básicos pueden ser definidos, pero no definidores. Además, conceptos básicos y terminales también pueden contener subconceptos. Los subconceptos corresponden a conceptos cuyas preferencias de selección alteran o especifican uno o más participantes del marco temático del respectivo concepto básico o terminal. Se entiende que un marco temático otorga las características de la unidad, integrando aquellas que hereda de sus conceptos padres. Luego el postulado de significado detallará esas características mediante la asignación de una serie predeterminada de papeles. Los conceptos terminales están representados por

el símbolo ($), como en $EXCHANGE_00 para eventos, $SCARECROW_00 para entidades, y $FRIENDLY_N_00 para cualidades. Además, cada una de las unidades conceptuales en COREL está contenido en una subontología, ya sea para entidades, eventos o cualidades. Cada subontología da cuenta de los respectivos metaconceptos, conceptos terminales y conceptos básicos. De este modo, los metaconceptos de distribuyen en tres subontologías.

En cuanto a la subontología para entidades, que se configura como la más relevante para este trabajo, se establece una gran división entre entidades abstractas y físicas. Por un lado, las entidades abstractas se subdividen a partir de: sus atributos, como color o peso; su proposición, que refiere al contenido cognitivo; y su cantidad, ya sea en términos de tiempo o de espacio. Por otra parte, las entidades físicas se subdividen en cuanto a si corresponden a un objeto, como un grupo, una parte de algo, o un lugar; o a un proceso, entendido como cualquier hecho que ocurre. Lo anterior, se puede esquematizar como sigue:

Figura 7. Jerarquía para la subontología de entidades en FunGramKB.

#ENTITY
- 1.1 #ABSTRACT
  - 1.1.1 #ATTRIBUTE
    - 1.1.1.1 #PHYSICAL_ATT
    - 1.1.1.2 #PSYCHOLOGICAL_ATT
      - 1.1.1.2.1 #BEHAVIOURAL_ATT
      - 1.1.1.2.2 #COGNITIVE_ATT
      - 1.1.1.2.3 #EMOTIONAL_ATT
      - 1.1.1.2.4 #PERCEPTUAL_ATT
  - 1.1.2 #PROPOSITION
  - 1.1.3 #QUANTITY
    - 1.1.3.1 #TIME
- 1.2 #PHYSICAL
  - 1.2.1 #OBJECT
    - 1.2.1.1 #COLLECTION
    - 1.2.1.2 #FEATURE
    - 1.2.1.3 #REGION
    - 1.2.1.3 #SELF_CONNECTED_OBJECT
  - 1.2.2 #PROCESS

COREL consta de su propia semántica y sintaxis. Por ejemplo, si analizamos el postulado de significado del concepto terminal $EXCHANGE_00, podríamos explicarlo desde varios niveles: Primero, el propio postulado de significado del concepto terminal $EXCHANGE_00, se lee como un evento *e1* en el que un agente *x1*, transfiere un tema *x2*, desde un origen *x3* hacia una meta *x4*. Como especificación de *e1*, o predicación satélite *f1*, se marca un *e2*, en el que un agente *x4* que equivale a

la meta en *e1*, tiene el propósito de transferir un nuevo tema *x5*, desde un nuevo origen *x6*, hacia la meta *x1*, que en *e1* correspondía al agente. De este modo, la relación que se establece es la de 'dar a alguien algo, y con el propósito de recibir de ese alguien otra cosa'.

+(e1: +TRANSFER_00 (x1)Agent (x2)Theme (x3)Origin (x4)Goal
(f1: (e2:+TRANSFER_00 (x4)Agent (x5)Theme (x6)Origin (x1)Goal))Purpose)

Segundo, $EXCHANGE_00 se jerarquiza desde la subontología de #EVENT. Es decir, es un tipo de transferencia, que corresponde a un tipo de movimiento, que es un tipo material de *hacer* (+DO_00), que corresponde a su vez a un evento.

#EVENT >> #MATERIAL >> +DO_00 >> +MOVE_00 >> +TRANSFER_00 >> $EXCHANGE_00

Finalmente, el marco temático de $EXCHANGE_00 es heredado desde el concepto superordinal +TRANSFER_00. Posee cuatro papeles: agente (*x1*), tema (*x2*), origen (*x3*) y meta (*x4*). Como sigue:

(x1: +HUMAN_00) Agent
(x2) Theme
(x3) Origin
(x4: +HUMAN_00) Goal

Otro ejemplo es el postulado de significado para el concepto terminal $ILLUSTRATION_00. Leemos la representación de este postulado de significado como una entidad que tiene de la característica de *ser* (+BE_00) una ilustración *x1*, que a su vez corresponde al papel tema. El referente *x2* es que esa ilustración corresponda a una imagen. Bajo el símbolo (*) se propone una predicación rebatible. En este caso, el tema corresponde a *ser-en/estar-en* (+BE_02) un libro. En este caso, el *ser* referido corresponde a un atributo de la entidad. Es decir, se propone que una ilustración se encuentra, generalmente, en una locación *x3* libro, entendido a su vez como un conjunto de hojas, típicamente de papel, impresas o manuscritas, que se encuentran unidas por uno de sus lados y encuadernadas, formando un volumen.

+(e1: +BE_00 (x1: $ILLUSTRATION_00) Theme (x2: +IMAGE_00)Referent) *(e2: +BE_02 (x1)Theme (x3: +BOOK_00)Location)

Finalmente, $ILLUSTRATION_00 se jerarquiza desde la subontología de #ENTITY. Es decir, una ilustración es un tipo de imagen, que a su vez es un tipo de objeto de información, que corresponde a un objeto autoconectado, inserto en la categoría objetos, y que se clasifica como una entidad física.

#ENTITY >> #PHYSICAL >> #OBJECT >> #SELF_CONNECTED_OBJECT >> +INFORMATION_OBJETC_00 >> +IMAGE_00 >> $ILLUSTRATION_00

3.4 El lexicón mental

En la mayoría de los modelos de análisis lingüístico que entienden el lexicón como un componente mental de información que posee alcances computacionales, se define el lexicón como un conjunto estático de sentidos de palabras, etiquetados a su vez con rasgos distintivos que representan información sintáctica, morfológica y semántica (Pustejovsky, 1991). Según los trabajos de Aitchinson (1987), es bastante común otorgar al lexicón las propiedades de un diccionario semasiológico. Es decir, el léxico se organizaría en la mente según criterios lexicográficos, y almacenaría además todas las posibles derivaciones de las palabras y sus respectivos significados. Luego el usuario realiza el trabajo de buscar y extraer información desde un depósito de unidades lingüísticas que le permitirán referir a una realidad en particular.

Sin embargo, la organización del léxico mental es considerablemente más compleja que la de los diccionarios, pues no incorpora el requisito primordial del orden. En efecto, el léxico lingüístico que se organiza en la mente desborda en complejidad al compararlo con la organización de cualquier diccionario de una lengua conocida, en tanto integra características particulares que lo acercan más a la comparación con un sistema interconectado de información. Se trata de un repositorio que no tiene, como el ya mencionado diccionario, un número fijo de palabras que responden a las características predeterminadas de un corpus. Esta propiedad definitoria implica que los hablantes pueden adherir, crear y/o intervenir libremente las unidades léxicas en cualquier lengua.

Además, el conocimiento de mundo compartido por los hablantes permitirá vincular de manera eficiente distintos mecanismos lexicogenésicos con nuevos significados. Esta productividad lingüística postula que es posible establecer un número ilimitado de combinaciones para un número limitado de unidades en el sistema. Por ejemplo, una expresión como «golear» podrá ser comprendida

por un hablante siempre y cuando se cumplan determinadas condiciones: primero, si el uso es compatible con el contexto; segundo, si ambos hablantes comparten conocimientos respecto al significado del evento GOL*;* tercero, si emisor y receptor comprenden (por convención) que típicamente el sufijo «-ear» forma verbos derivados de sustantivos y añade un significado iterativo. De esta manera, «golear» puede significar que, en un partido de fútbol, se ha producido repetida e indefinidamente el evento GOL por parte de un equipo participante y en desmedro del otro.

El lexicón mental se define entonces, según Bonin (2004), como un repositorio mental de todas las representaciones que están intrínsecamente relacionadas con palabras. En efecto, es posible identificar distintas operaciones que están involucradas en los procesos léxicos asociados al lexicón. En relación con el nivel fonético/fonológico, etiquetar estímulos visuales, leer y/o escuchar palabras corresponden a procesos vinculados con la decodificación fonológica, cuya correspondiente conciencia fonológica se desarrollará progresivamente durante los primeros años de vida como una competencia para prestar atención, identificar y manipular las unidades del lenguaje oral. Por otra parte, en el plano de la escritura, las representaciones mentales de las palabras típicamente están vinculadas a la correspondencia con una secuencia de grafemas en particular. Las representaciones semánticas, por su parte, son producto de la movilización o mediación de un significado a partir del reconocimiento visual o auditivo de una palabra. Estas representaciones implican un conocimiento de mundo específico del hablante y el reporte de ciertas categorías para categorizar un estado de cosas del mundo. En cuanto a los aspectos morfológicos, existen diversos procedimientos para la producción de palabras que funcionan como reglas que componen significados, como por ejemplo la morfología derivativa.

Por otra parte, el lexicón mental se constituye como una compleja red de relaciones de significado, equivalente a un tesauro. Las palabras, en efecto, no se relacionan solamente por vínculos opositivos, sino también por inclusión de clases según el principio de implicación lógica. Aitchinson (1987) afirma que el lexicón es capaz de almacenar información sumamente detallada con respecto a entidades, cualidades y eventos, además de sus respectivas instancias o especificaciones. Así, la mente humana se convierte en un enorme banco de palabras con sus respectivos contextos de uso, que utiliza además un brevísimo tiempo para los procesos de reconocimiento y localización de palabras. En este sentido, la comparación entre el lexicón mental y cualquier obra lexicográfica es una manera eficiente de mostrar que no es posible realizar deducciones acerca de la organización lingüística en nuestra mente a partir de los métodos mediante los que se organizan las palabras en los libros.

Finalmente, Elman & McClelland (1984) y Miranda-García (1993), proponen la explicación de ficheros de almacenamiento. Según esta, el lexicón mental contendría toda la información de cada palabra, según los siguientes contenidos:

a. Contenido fónico (transcripción, acento y separación silábica).
b. Contenido grafémico.
c. Contenido semántico (rasgos semánticos estrictos y rebatibles).
d. Contenido morfosintáctico (clase gramatical, reglas de colocación y combinatoria).
e. Contenido léxico (relaciones léxicas de significado: sinonimia, antonimia, hiponimia/hiperonimia; campos semánticos).

El atributo principal y definitorio de una palabra es su contenido de información, comúnmente llamado significado. Estos significados, o conjuntos de sentidos de palabra, cuando aún no están asociados con una forma léxica, pueden ser genéricamente definidos como conceptos, y por tanto concebidos como categorías mentales que integran un contenido de información. Así, siguiendo a Elman & McClelland (1984) y Ježek (2016), se puede decir que los significados existen independientemente de la lengua. Desde este punto de vista, la asociación entre un concepto y una forma léxica constituye un proceso que puede definirse como lexicalización. No obstante, la noción de lexicalización puede interpretarse de varias formas. Según lo anterior, la lexicalización corresponderá a cualquier proceso en el que, para una lengua determinada, un contenido de información se asocie a una forma léxica específica. Entonces, todas las palabras de una lengua son el resultado de un proceso de lexicalización, como se especifica en la siguiente figura:

Figura 8. Proceso de lexicalización o codificación léxica.

Desde el punto de vista del significado, las palabras que componen el léxico de una lengua se dividen en dos grandes grupos: palabras de contenido y palabras funcionales (Lyons, 1968). La distinción canónica, y particularmente extendida desde la gramática descriptiva, para estas categorías, es que verbos, sustantivos, adjetivos y adverbios pertenecen a las llamadas categorías léxicas mayores; mientras que los determinantes, pronombres, conjunciones y preposiciones se incluyen en las llamadas categorías léxicas menores. Esta agrupación respecto a las clases de palabras es, por cierto, imperfecta pero útil, porque resalta la forma diferente en la que estos dos tipos de palabras contribuyen al significado de las estructuras oracionales en las que aparecen. Finalmente, según Lyons (1968), mientras que las palabras de contenido proporcionan significado conceptual, las palabras funcionales permitirían explicitar los tipos de relaciones que se establecen entre las distintas proposiciones introducidas por las palabras de contenido.

El significado de las palabras de contenido corresponde al significado léxico, mientras que el de las palabras funcionales recibe, típicamente, el nombre de significado gramatical. Un ejemplo bastante ilustrativo de lo anterior es la verificación de que una preposición es capaz de proyectar significados potencialmente diferentes dependiendo de la relación que puede establecer en la combinatoria de unidades. Este es el caso de la polisemia complementaria que propicia la preposición «en» para las instancias que se presentan a continuación:

(9)

c. En la mañana.
d. En la biblioteca.

En el caso (9a), la preposición «en» ocurre en el cotexto del sustantivo «mañana» y, por tanto, permite que se haga referencia a un periodo de tiempo. Por otra parte, la preposición «en» ocurre, en (9b), en el cotexto de «biblioteca». En ese caso, se establece que la referencia es un espacio físico. Entonces, «en» para (9a) carga con el significado de ‘durante’, mientras que en (9b) carga con el de ‘dentro de’.

Una diferencia adicional entre las palabras de contenido y las palabras funcionales es que las primeras constituyen una categoría abierta de elementos; es decir, un conjunto en el que se pueden incluir o descartar nuevos elementos. Por su parte, las palabras funcionales corresponden a categorías cerradas, puesto que el número de elementos permanece constante. Si bien la adquisición de nuevas palabras de contenido en el léxico es un fenómeno común en todas las lenguas naturales, la aparición

de una palabra con una nueva función es un fenómeno aislado que se presenta con menor frecuencia. Como se mencionó anteriormente, el criterio que establece una correspondencia entre la clase de palabra y el tipo de significado, léxico o gramatical, es una explicación bastante consistente, pero aun así imperfecta desde el punto de vista de la teoría lingüística. No obstante, se trata de una categorización que es altamente eficiente computacionalmente, pues permite establecer criterios para determinar los aspectos que aportan mayor relevancia tanto semántica como morfosintáctica a la descripción de un fenómeno particular en una lengua natural. Esta distinción, en el ámbito el PLN, recibe la nomenclatura de *content words* para las palabras de contenido, y *stop words* para las palabras funcionales o palabras excluidas, en la medida que la eliminación de palabras funcionales permite que el contenido textual sea procesado de una manera más eficiente.

3.5 El entorno de trabajo DAMIEN (*Data Mining Encountered*)

DAMIEN (*DAta MIning ENcountered*) es un entorno informático que puede integrar técnicas de múltiples disciplinas dentro de análisis de texto (i.e. lingüística de corpus, estadística y minería textual) para apoyar la investigación lingüística de manera más efectiva. A continuación, se realiza una descripción de los componentes y características más relevantes de DAMIEN, según Periñán-Pascual, (2017), que serán relevantes para la implementación de los ejercicios propuestos más adelante. DAMIEN contiene cuatro interfaces especializadas en tareas específicas:

a. *Corpus*: tareas relacionadas con la exploración, preprocesamiento y procesamiento de un corpus o colección de textos.
b. *Statistics*: tareas relacionadas con la descripción e interpretación de datos a partir de la aplicación de parámetros estadísticos.
c. *Mining*: tareas relacionadas con minería de datos y métodos de predicción (como clasificadores o métodos de agrupamiento para el aprendizaje automático).
d. *Evaluation*: tareas relacionadas con la aplicación de medidas de evaluación.

En el ámbito del PLN en general, y el tratamiento de recursos lingüísticos informatizados en particular, DAMIEN propicia que los lingüistas puedan alcanzar sus objetivos de investigación de manera más efectiva mediante la integración de métodos y técnicas provenientes de varios campos dentro del análisis de datos textuales. Actualmente, los programas computacionales disponibles de manera gratuita, como *TextSTAT* y *AntConc* para lingüística de corpus, *R commander* para estadística, *WEKA*

para minería de datos o *GATE* para ingeniería lingüística, no son capaces de reunir en una sola interfaz gráfica de usuario todos los requerimientos necesarios para la investigación basada en corpus lingüísticos. Por el contrario, DAMIEN logra integrar en un mismo entorno de trabajo las diferentes herramientas y técnicas que pueden ser aplicadas en el análisis de datos textuales basados en corpus. Estas técnicas provienen principalmente de:

a. Lingüística de corpus (i.e. listas de frecuencia, procesamiento de XML (XSL), administración de bases de datos y consultas SQL, búsqueda por expresiones regulares, etc.).
b. Estadística (i.e. estadística descriptiva e inferencial, representación gráfica de datos).
c. Procesamiento del lenguaje natural (i.e. extracción de N-gramas, derivación, análisis morfológico o sintáctico, etiquetado POS, etc.).
d. Minería de textos (i.e. clasificación y métodos de agrupamiento).

En resumen, DAMIEN permite a las y los investigadores resolver tres grandes grupos de tareas. Estas tareas, a su vez, representan aquellos pasos necesarios para planificar, montar y evaluar experimentos de análisis de corpus:

a. Administración de colecciones de datos: visualización, edición, aleatorización, búsqueda y extracción de información.
b. Análisis de datos: convergencia de la estadística con la lingüística de corpus a través de la estadística descriptiva (medidas de posición y dispersión) y la estadística inferencial (pruebas estadísticas de distinto tipo, como correlación, regresión, o análisis multivariante).
c. Presentación de datos: transferencia a texto y representaciones gráficas a partir del análisis de datos, para la difusión científica en diferentes formatos.

3.6 Herramientas informáticas para el tratamiento de datos textuales

A continuación, se presentan tres herramientas informáticas para la construcción y exploración de recursos lingüísticos informatizados, junto con el análisis de datos textuales: expresiones regulares, lenguaje de etiquetado extensible (XML) y lenguaje de consulta estructurada (SQL). Se trata de tecnologías relevantes para el desarrollo de investigaciones lingüísticas basadas en datos textuales. Cada herramienta se aborda desde una definición, una exposición de sus ámbitos de uso, y una breve descripción de su funcionamiento.

3.6.1 Expresiones regulares (regex)

Las expresiones regulares son patrones utilizados para encontrar una determinada combinación de caracteres dentro de una cadena de texto. Según el tutorial de expresiones regulares de Goyvaerts (2007), se trata de una cadena de texto especial que describe un patrón de búsqueda; es decir, una herramienta que permite analizar y validar cadenas de texto(s), que representan a su vez grandes volúmenes de información, para la búsqueda de patrones textuales determinados. Específicamente, mediante la utilización de expresiones regulares es posible extraer, editar, reemplazar o eliminar subcadenas de texto, como también agregar las cadenas extraídas a una colección.

Existen dos tipos de caracteres que es necesario conocer para operar con expresiones regulares: (1) caracteres literales: uno o más componentes o elementos textuales alfanuméricos, o patrones literales; y (2) metacaracteres (o caracteres especiales): operadores que se pueden utilizar en la consulta para determinar qué es lo que debe concordar usando la comprobación expresión regular (delimitadores, agrupadores, cuantificadores, etc.). Si bien es posible establecer búsquedas a partir de patrones íntegramente ejecutados mediante caracteres literales, el potencial de las expresiones regulares está dado por la utilización de caracteres especiales para enriquecer el proceso de búsqueda.

El listado de metacaracteres que se presenta a continuación no es una lista exhaustiva o final, sino que corresponde a aquellos considerados como básicos o fundamentales. Se trata de catorce caracteres que aportan funciones especiales al patrón requerido por la expresión regular: la barra invertida (\ ), el símbolo de intercalación (^), el signo de dólar ($), el punto (.), la barra vertical (|), el signo de interrogación (?), el asterisco (*), el signo más (+), el paréntesis de apertura (( ), el paréntesis de cierre ( )), el corchete de apertura ([ ), el corchete de cierre ( ]), la llave de apertura ({), y la llave de cierre (}). Estos metacaracteres se pueden agrupar en cuatro tipos:

a. Los metacaracteres que distinguen clases, los cuales se utilizan para realizar la búsqueda de cualquier carácter dentro de un grupo, o bien identificar caracteres específicos dentro de un grupo:

Tabla 6. Metacaracteres del tipo *clase* y sus definiciones.

Metacaracter	Definición
.	Coincide con cualquier caracter.
[ ]	Coincide con un caracter incluido en el conjunto.
[^]	Coincide con un caracter no incluido en el conjunto.
[-]	Coindice con un caracter dentro de un rango o intervalo de caracteres (valores).
\w	Coincide con un caracter alfanumérico.
\W	Coincide con un caracter no alfanumérico.
\s	Coincide con un caracter de espacio en blanco (según *regex*, los espacios en blanco son tratados como caracteres literales).
\S	Coincide con un caracter que no es espacio en blanco.
\d	Coincide con un dígito.
\D	Coincide con un no dígito.

b. Los metacaracteres que determinan fronteras; es decir, aquellos que establecen un límite entre uno o más caracteres literales:

Tabla 7. Metacaracteres del tipo *frontera* y sus definiciones[16].

Metacaracter	Definición
\b	Coincide con una frontera de palabra (esto es, entre un caracter \w y uno \W).
^	También llamado 'símbolo de intercalación' o 'salto de línea', se utiliza para indicar el inicio de una nueva línea.
$	También llamado 'final de línea', se utiliza para indicar el término de una cadena en la línea actual.

c. Los metacaracteres que sirven como cuantificadores, es decir, permiten hacer coincidir instancias de un carácter literal, un grupo o clase de caracteres en una cadena:

[16] Si una expresión regular completa está delimitada por un metacaracter de intercalación (^) y uno de final de línea ($), coincide con una línea completa. Entonces, para hacer coincidir todas las cadenas que contengan un solo caracter, se puede utilizar la expresión regular *^. $*.

Tabla 8. Metacaracteres del tipo *cuantificador* y sus definiciones.

Metacaracter	Definición
X?	Coincide con el caracter X cero o una vez.
X*	Coincide con el caracter X cero o más veces.
X+	Coincide con el caracter X una o más veces.
X{n}	Coincide con el caracter X exactamente *n* veces.
X{n,}	Coincide con el caracter X al menos *n* veces.
X{n, m}	Coincide con el caracter X al menos *n* veces, pero no más de *m* veces.

d. Otros metacaracteres relevantes son los paréntesis de agrupación ( ); la barra de alternación (|); y el escape (/), que se utiliza para tratar un metacaracter como un carácter literal.

Por ejemplo, para realizar la búsqueda del patrón 'palabras que comiencen con un carácter en mayúscula', se debería realizar el siguiente razonamiento:

a. Frontera de palabra (es un espacio en negro entre dos espacios en blanco).
b. Las mayúsculas constituyen un rango de caracteres posibles.
c. Linealmente, una vez que aparezca la mayúscula, la expresión regular debe seleccionar cualquier carácter que aparezca cero o más veces hasta el final de la palabra.

Luego, la expresión regular correspondiente sería:

\b[A-Z]\w*\b

Además, considerando las vocales tildadas como un nuevo carácter literal, el resultado final correspondería a:

\b[A-ZÁÉÍÓÚ]\w*\b

En otro ejemplo, se intenta buscar en un conjunto de datos las cadenas de texto «rat» «cat» y «bat». Para esto, es necesario considerar el siguiente razonamiento:

a. Frontera de palabra (es un espacio en negro entre dos espacios en blanco).
b. Linealmente, la expresión regular debe considerar las secuencias que son comunes y las que son diferentes.
c. Es posible (y necesario) establecer agrupaciones de caracteres para evitar la redundancia.

Luego, la expresión regular resultante sería:

\b[cbt]at\b

3.6.2 Lenguaje de etiquetado extensible (XML)

La función de almacenamiento de datos se realiza mediante un lenguaje extensible de marcado, o XML (*eXtensible Markup Language*). Es un lenguaje en términos informáticos, porque se establece como un conjunto de reglas (a pesar de que no sirva para ejecutar ninguna acción) para describir etiquetas con metainformación mediante las cuales es posible almacenar datos de manera jerárquica. En este sentido, XML ha sido diseñado para estructurar, almacenar e intercambiar información, no para desplegarla o visualizarla. Por ejemplo, dado el siguiente conjunto de datos textuales no estructurado:

Carlos
Bustos Gómez
Cruzados #3028
46730

cada uno de los componentes podría corresponder a cualquier tipo de información. Es necesario marcar los datos mediante la incorporación jerarquizada de metadatos; es decir, etiquetas que describen los datos o información. Entonces, un archivo XML podría corresponder al siguiente conjunto:

```
<trabajador>
        <nombre>Carlos</nombre>
        <apellidos>Bustos Gómez</apellidos>
        <dirección>Cruzados #3028</dirección>
        <código postal>46730</código postal>
</trabajador>
```

En este ejemplo, se almacena la información personal de un trabajador a través de la incorporación de etiquetas (metadatos) para describir dicha información (datos). Entonces, la primera etiqueta, en orden jerárquico, que incluye a las demás, es la de <trabajador>. Esta etiqueta debe abrirse < > y cerrarse </> cuando corresponda. En su interior se encuentran, ordenados, los datos requeridos con sus respectivas etiquetas.

Los archivos con extensión *.xml* se almacenan en formato de texto plano. Esto quiere decir que, independientemente del software o el hardware, pueden ser leídos y procesados por diferentes tipos de aplicaciones y sistemas operativos. Así, la creación, almacenamiento y/o conversión de repositorios de datos en formato XML facilitan el intercambio de información entre sistemas informáticos cuyos formatos son incompatibles.

En el ámbito de la lingüística de corpus, la utilización de XML se encuentra ampliamente extendida en la gestión de documentación y en la organización de conocimiento lingüístico, puesto que las etiquetas pueden ser creadas, reutilizadas y modificadas por el analista. Según lo anterior, en el siguiente ejemplo se establecen algunas etiquetas para la clasificación semántica del evento «jugar», basadas en el corpus ADESSE (García-Miguel *et al*., 2010). En este caso el lexema «jugar» es el único evento disponible para la categoría 'clasificación semántica', en la que se despliegan los atributos tipo de proceso y valencia, junto con la incorporación de un ejemplo de uso.

```
<clasificación semántica>
        <eventos>
                <jugar>
                        <proceso>actividad</proceso>
                        <valencia>3</valencia>
                        <ejemplo>Juan juega con María</ejemplo>
                </jugar>
        </eventos>
</clasificación semántica>
```

De esta forma, las etiquetas permiten estructurar la información jerárquicamente a la vez que incorporan nombres ad hoc para facilitar el almacenamiento y distribución de conjuntos de datos por parte del investigador. Por lo tanto, XML se ha consolidado como un estándar ampliamente aceptado para el intercambio de datos entre sistemas informáticos de cualquier tipo.

3.6.3 Lenguaje de consulta estructurada (SQL)

El lenguaje de consulta estructurada, o *Structured Query Language* (SQL), es un lenguaje estándar que se utiliza para almacenar, administrar y extraer información desde una base de datos (según el modelo de gestión de bases de datos relacionales). Actualmente, es el sistema de gestión de bases de datos relacionales más utilizado (Keibrich, 2010). Fue desarrollado por IBM a comienzos de la década de 1970, y tuvo su primera versión comercial en 1979. Los sistemas de bases de datos relacionales se utilizan para almacenar registros definidos por el usuario en tablas con grandes volúmenes de

información. Además del almacenamiento y la gestión de datos, un motor de base de datos puede procesar comandos de consulta complejos que combinan datos de varias tablas para generar informes y resúmenes de datos.

Las funciones básicas, o iniciales, para operar con SQL son tres: SELECT (selecciona o consulta registros de una base de datos específica; en este caso, el nombre de una columna), FROM (especifica el nombre de la tabla desde la que se extraerán los registros seleccionados) y WHERE (determina una condición para los registros seleccionados). Por tanto, para seleccionar una columna desde una tabla donde se cumpla una condición, se puede ejecutar el siguiente comando:

SELECT *nombre_columna* FROM *nombre_tabla* WHERE *condición*

O bien, si el requerimiento consistiese en seleccionar dos columnas desde una tabla:

SELECT *nombre_columna_01, nombre_columna_02* FROM *nombre_tabla*

Estas instrucciones de base se pueden complementar con diferentes tipos de cláusulas, como GROUP BY (separa los registros seleccionados en grupos específicos), ORDER BY (ordena los registros seleccionados según un criterio específico), DISTINCT (identifica solo los valores que son diferentes en los registros seleccionados) COUNT (cuenta el número de registros de una columna); AVG (calcula el promedio de los valores seleccionados en los registros), SUM (calcula la suma de los valores seleccionados en los registros), MAX (determina el valor más alto dentro de los registros seleccionados), o MIN (determina el valor más bajo dentro de los registros seleccionados). Por ejemplo, si el requerimiento fuese seleccionar una columna desde una tabla con sus respectivos registros ordenados según otra columna, la sentencia SQL adoptaría la siguiente forma:

SELECT *nombre_columna* FROM *nombre_tabla* ORDER BY *nombre_columna*

O bien, si un nuevo requerimiento consintiese en determinar el número de registros distintos de una columna desde una tabla:

SELECT COUNT (DISTINCT *nombre_columna*) FROM *nombre_tabla*

Otros operadores relevantes que se pueden ejecutar para seleccionar información son: AND (conjunción que evalúa dos condiciones y entrega un valor verdadero sólo si ambas son verdaderas), OR (disyunción que evalúa dos condiciones y devuelve un valor de verdad si alguna de las dos es verdadera), NOT (negación lógica que devuelve el valor contrario de la expresión), BETWEEN (selecciona un intervalo de valores dentro de los registros seleccionados), y LIKE (compara los valores de un registro según el criterio de otro registro seleccionado). Por ejemplo, si el requerimiento consistiese en seleccionar una columna, cualquiera sea (*) , desde una tabla, con la condición de que sus registros correspondan a un patrón determinado (anteponiendo el símbolo porcentaje (%)) , la sentencia SQL sería la siguiente:

SELECT * FROM *nombre_tabla* WHERE *nombre_columna* LIKE '*%*'

Según Godoc (2010), el lenguaje SQL presenta, en términos generales, dos ventajas. En primer lugar, debido a que obtuvo una enorme difusión, es empleado en la mayoría de los sistemas informáticos actuales en el ámbito comercial. En segundo lugar, es capaz de optimizar el tiempo de realización de las operaciones relacionadas con la minería de datos, en comparación al esfuerzo que requeriría llevarlas a cabo por medio de lenguajes de programación tradicionales.

Capítulo 4
Pregunta de investigación y objetivos

La pregunta que orienta esta propuesta de investigación es la siguiente: ¿qué grado de integración entre modelos estadísticos y simbólicos es necesario para el abordaje del problema de la desambiguación léxica en el ámbito del PLN a partir de un modelo de medida para la similitud y relación semántica? Según lo anterior, esta investigación se propone tratar la representación formal de un procedimiento computacional para poder resolver, a partir de un enfoque interlingüe y basado en el conocimiento, la desambiguación léxica mediante el desarrollo de una herramienta para PLN.

4.1 Objetivo general

El objetivo general de esta propuesta de investigación se resume como sigue: *Desarrollar un modelo más robusto de medida para la similitud y relación semántica que los disponibles actualmente para la desambiguación léxica automática, aplicado al PLN.*

4.2 Objetivos específicos

Los objetivos específicos son los siguientes:

a. Caracterizar el comportamiento del fenómeno lingüístico de la ambigüedad en general y de la ambigüedad léxica en particular.
b. Caracterizar el problema computacional de la desambiguación léxica automática, los métodos disponibles, y sus implicancias en el ámbito del PLN.
c. Compilar un corpus que integre instancias auténticas de ambigüedad léxica en español de Chile.
d. Ejecutar experimentos de base para el testeo de métodos de desambiguación léxica basados en conocimiento contextual, utilizando el corpus citado en el objetivo específico (c).
e. Representar formalmente un procedimiento computacional para poder resolver la desambiguación léxica automática, basado en la propuesta de una medida de desambiguación híbrida.
f. Evaluar la medida híbrida de desambiguación léxica automática propuesta en el objetivo específico (e), a partir del análisis de unidades léxicas derivadas de una fuente de conocimiento lingüístico.

Capítulo 5
Metodología[17]

A continuación, se presentan los aspectos metodológicos de la investigación. En primer lugar, se establecen los subtipos generales de procesamiento de datos en aprendizaje automático, con sus respectivas tareas de procesamiento. Luego se reporta el experimento preliminar, basado en el corpus SENSEVAL-3, cuyos resultados permitieron consolidar la metodología tanto para la selección del corpus en análisis como para los pasos para la aplicación de los experimentos posteriores de aprendizaje automático. En tercer lugar, se describe el procedimiento para el montaje del corpus a partir de un *subcorpora* de CODICACH, constituido por un conjunto de instancias en las que se incluyen las unidades léxicas polisémicas seleccionadas desde la base de conocimiento FunGramKB. Por último, se exponen las tareas de procesamiento específicas para la realización de los experimentos de aprendizaje automático utilizando la herramienta DAMIEN.

5.1 Subtipos de procesamiento de datos en aprendizaje automático

5.1.1 Preprocesamiento

Los pasos que se realizaron en la fase de preprocesamiento fueron los siguientes:

i. Creación de una colección de documentos derivados del corpus, correspondiente a 120 archivos en formato *.txt*, que contienen cada uno el ítem léxico en análisis o palabra objetivo, y su respectiva ventana contextual.
ii. Anotación de la muestra anterior en formato en formato *.csv*, con los sentidos etiquetados para cada uno de los textos de entrada incluidos en la colección de documentos.

5.1.2 Procesamiento

Los pasos que se llevaron a cabo en la fase de procesamiento fueron los siguientes:

i. Creación de una matriz *N-grama/documento* para cada ítem léxico, considerando la frecuencia absoluta de aparición en cada documento.
ii. Aplicación de una lista de inicio (que será explicada más adelante) para filtrar las palabras estadísticamente significativas en la recopilación de datos.

[17] Todos los archivos correspondientes tanto al montaje del corpus como a los resultados de esta investigación, que serán expuestos desde este capítulo en adelante, se encuentran disponibles de manera permanente en el siguiente repositorio: https://github.com/fredyrodrigors/tesis-phd#readme.

iii. Creación de una nueva matriz *N-grama/documento*, considerando el filtro generado con la lista de inicio.

iv. Aplicación de una validación cruzada para resolver el problema que implica la selección de conjuntos de datos previamente etiquetados. Este procedimiento permite extraer instancias agrupadas aleatoriamente desde la matriz etiquetada que constituye el corpus de entrenamiento, para que la máquina pueda etiquetar automáticamente nuevos conjuntos de datos desconocidos (Yung & Hu, 2015). Así, se realiza un procedimiento de distribución aleatoria y uniforme de *k-veces*, donde $k = 3$, a partir el conjunto de datos de las 120 instancias en análisis. De esta forma, se generaron tres conjuntos de datos con 40 instancias cada uno, que constituyen los *datasets* de prueba.

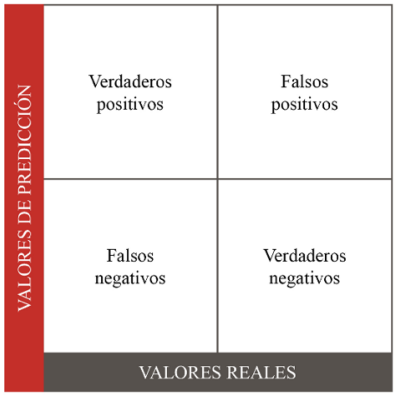
v. Aplicación del algoritmo de clasificación bayesiano ingenuo para los conjuntos de datos de cada sentido.

5.1.3 Evaluación

En la fase de evaluación se realizó el siguiente paso:

i. En el procedimiento de evaluación se empleó una matriz de confusión, con la que se realizó una medición del rendimiento de un sistema de clasificación basado en el aprendizaje automático. Cada resultado se expresó en una tabla de doble entrada con cuatro combinaciones diferentes de valores para un problema de clasificación de aprendizaje automático en el que la salida puede ser de dos o más clases, con cuatro combinaciones diferentes de valores de predicción (positivo o negativo) y valores reales (verdadero o falso), como se muestra en la siguiente figura:

Figura 9. Matriz (binaria) de confusión.



ii. En cuanto a las medidas para evaluar el rendimiento de los sistemas de desambiguación, se consideraron los valores de precisión, cobertura y media armónica (*puntaje F*). La definición y los parámetros para cada medida son los siguientes:

   a. Precisión: medida que representa la dispersión del conjunto de valores obtenidos. Así, cuanto menor es la dispersión, mayor es la precisión. Este valor se expresa como una proporción entre el número de predicciones correctas, sean estas positivas o negativas, y el total de predicciones. Lo anterior se formaliza como sigue:

$$Precisión = \frac{verdadero\ positivo}{verdadero\ positivo\ +\ falso\ positivo}$$

   b. Cobertura: medida que representa la proporción de las clases que fueron correctamente etiquetadas. Se expresa mediante la tasa de verdaderos positivos:

$$Cobertura = \frac{verdadero\ positivo}{verdadero\ positivo\ +\ falso\ negativo}$$

   c. Puntaje F: estadístico que permite integrar la precisión y la cobertura en una sola medida. Su relevancia radica en la posibilidad de representar de manera eficiente la distribución de las clases, para así establecer un puntaje armónico para el desempeño del sistema cuyo valor máximo posible es igual a 1. Lo anterior se formaliza como sigue:

$$Puntaje\ F = 2 \times \frac{precisión\ \times\ cobertura}{precisión\ +\ cobertura}$$

iii. Los resultados generales se determinan considerando el promedio de todos los conjuntos de datos para cada sentido, junto con el macropromedio del sistema de desambiguación. Las posibles combinaciones y sus interpretaciones son las siguientes:

   a. Precisión y cobertura altas: el sistema clasifica eficientemente la clase en análisis.
   b. Precisión alta y cobertura baja: el sistema no clasifica eficientemente la clase en análisis. Sin embargo, cuando logra clasificarla es altamente confiable.

c. Precisión baja y cobertura alta: el sistema clasifica de una manera parcialmente eficiente, pues la clasificación incluye casos de clases diferentes.
d. Precisión y cobertura bajas: el sistema no logra clasificar la clase en análisis de manera eficiente, o bien no realiza ninguna clasificación a partir de los datos.

5.2 Experimento piloto utilizando el corpus de prueba SENSEVAL-3

En una primera instancia, se decidió trabajar a partir de la metodología propuesta para el montaje del corpus SENSEVAL-3 (*Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*). El objetivo de este certamen es la evaluación de métodos y técnicas de desambiguación automática del sentido de palabras[18]. En términos generales, se trata de sistemas computacionales diseñados para la asignación automática de sentidos, aplicables a tareas ya sea generales o especializadas de PLN. A partir de estas aproximaciones preliminares, se establecieron los recursos lingüísticos necesarios para las tareas de procesamiento, y se realizó el pilotaje para los experimentos de desambiguación léxica automática.

5.2.1 Selección del corpus de prueba SENSEVAL-3

El corpus utilizado para la tarea de muestra léxica del español en SENSEVAL-3 está formado por 12.625 ejemplos etiquetados, que cubren 25.875 frases y 1.506.233 palabras en total. El cotexto considerado para cada ejemplo incluye la palabra objetivo más una ventana contextual. Todos los ejemplos han sido extraídos desde el corpus del año 2000 de la Agencia Española de Noticias EFE, que incluye 289.066 noticias (2.814.291 frases y 95.344.946 palabras), de enero a diciembre de 2000 (Márquez *et al*., 2004). Para cada palabra, un mínimo de 200 ejemplos ha sido etiquetados manualmente por tres anotadores humanos expertos independientes. Los casos de desacuerdo han sido resueltos por otro lexicógrafo (asignando un sentido único a cada ejemplo).

Para la ejecución del experimento de prueba de aprendizaje automático utilizando el algoritmo bayesiano ingenuo, se seleccionaron 120 instancias de la muestra léxica para la palabra objetivo «partido», extraída desde el corpus SENSEVAL-3. Los sentidos fueron seleccionados desde el lexicón

[18] Esta versión del certamen se celebró en la ciudad de Barcelona (España), entre marzo y abril de 2004. Fue organizada por la ACL (*Association for Computational Linguistics*) en colaboración con la Universidad del Norte de Texas. La competición incluyó incluyó 14 tareas diferentes para la desambiguación automática del sentido de las palabras, así como la identificación de papeles semánticos, anotaciones multilingües y formas lógicas. Más información acerca del desarrollo de SENSEVAL-3 se encuentra disponible en http://web.eecs.umich.edu/~mihalcea/senseval/.

de WordNet, donde «partido.1» corresponde a 'organización política cuyos miembros comparten la misma ideología', y «partido.2» a 'prueba deportiva en la que se enfrentan dos equipos o jugadores'. Más abajo se presenta esta información en un fragmento del esquema XML correspondiente al archivo etiquetado como *partido_minidir_senseval.xml*[19]:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<lexelt>
        <sense id="partido.1" definition="Organización política cuyos miembros comparten la misma
        ideología" used="yes">
                <example text="el principal partido del país"/>
                <example text="el partido en la oposición"/>
                        <collocation text="partido comunista"/>
                        <collocation text="partido conservador"/>
                        <collocation text="partido político"/>
                <synset wordnet="1.5" id="05259394n"/>
        </sense>
        <sense id="partido.2" definition="Prueba deportiva en la que se enfrentan dos equipos o
        jugadores" used="yes">
                <example text="partido de baloncesto"/>
                <example text="el mejor partido de la temporada"/>
                        <collocation text="partido amistoso"/>
                        <collocation text="partido de fútbol"/>
                        <collocation text="partido de ida"/>
                <synset wordnet="1.5" id="04780657n"/>
        </sense>
</lexelt>
```

Posteriormente, cada una de las instancias, junto con sus respectivas etiquetas o *senseID*, se almacenaron en el archivo *partido_instancecorpus_senseval.xml*. A continuación, se muestra un ejemplo de este corpus de instancias, que considera tanto la palabra objetivo como su ventana contextual[20]:

[19] El documento *.xml* que contiene el minidiccionario para los sentidos de «partido», correspondientes al corpus SENSEVAL-3, está disponible en https://github.com/fredyrodrigors/tesis-phd/blob/main/experimento_senseval-3/partido_minidir_senseval.xml, y al final de este trabajo como Anexo 1.

[20] El documento *.xml* que contiene 120 instancias para cada uno de los sentidos de «partido» se ha incluido al final de este trabajo como Anexo 2. Además, se encuentra disponible en https://github.com/fredyrodrigors/tesis-phd/blob/main/experimento_senseval-3/partido_instancecorpus_senseval.xml.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<corpus lang="Spanish">
        <lexelt item="partido.n">
                <instance id="partido.n.128" docsrc="efe_996_2000/12/01">
                        <cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
                        <cat scheme="IPTC" code="11000000"/>
                        <answer instance="partido.n.128" senseid="partido.1"/>
                        <context>
                        <previous> La votación del informe de gestión, en la que 199 delegados se
                        pronunciaron a favor y tres se abstuvieron, cerró la primera jornada del congreso
                        de los socialistas extremeños, que fue inaugurado por el secretario general del
                        PSOE, José Luis Rodríguez Zapatero. </previous>
                        <target> En la exposición de su informe de gestión, Rodríguez Ibarra afirmó que
                        si de este congreso regional y de los próximos provinciales sale un
                        <head>partido</head> dividido internamente "estaríamos ante una estafa" y "una
                        falta de respeto" al electorado, al que se ofreció un proyecto y un partido "unido
                        y sólido".</target>
                        <following> Para Rodríguez Ibarra, que prefirió mantener la incógnita por el
                        momento sobre cual será el cartel electoral del PSOE en las próximas elecciones
                        autonómicas, aseguró que este congreso debe servir para dar muestra a la
                        sociedad de nuestra solidez y madurez. </following>
                        </context>
                </instance>
                <instance id="partido.n.2" docsrc="efe_4178_2000/01/08">
                        <cat scheme="ANPA" code="DEP: DEPORTES, FUTBOL"/>
                        <cat scheme="IPTC" code="15000000"/>
                        <answer instance="partido.n.2" senseid="partido.2"/>
                                <context>
                                <previous> Manzano auguró que Shoji Jo se aclimatará al fútbol español
                                porque tiene "gran rapidez de movimientos" y que lo poco que puede
                                faltarle lo adquirirá con un mínimo periodo de adaptación.</previous>
                                <target> La incorporación de Jo ha propiciado que el club vallisoletano
                                haya recibido numerosas peticiones de la camiseta del jugador desde
                                Japón, desde donde se han solicitado también acreditaciones para ver
                                los <head> partidos </head> del equipo. </target>
                                <following> Unos treinta periodistas nipones estarán presentes el
                                próximo lunes en la presentación del jugador, según confirmaron a Efe
                                fuentes del club. </following>
                                </context>
                </instance>
        </lexelt>
</corpus>
```

Las tareas de procesamiento fueron aplicadas para la información contenida dentro de la etiqueta */corpus/lexelt/instance/context/target*. La extracción de este conjunto de datos textuales y su traspaso a un documento *.csv* se llevó a cabo mediante el desarrollo y ejecución de una plantilla *.xsl*, a través del motor XSL (*Extensible Stylesheet Language Transformations*), integrado en DAMIEN:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output method="text" encoding="iso-8859-1"/>
        <xsl:template match="/">
        <xsl:for-each select="corpus/lexelt[@item='*']/instance/context">
                <xsl:value-of select="target"/>
        </xsl:for-each>
        </xsl:template>
</xsl:stylesheet>
```

5.2.2 Resultados del experimento piloto utilizando el corpus de prueba SENSEVAL-3

Los resultados del experimento piloto[21], luego de aplicar cada una de las tareas de procesamiento, fueron los siguientes:

Tabla 9. Matriz de confusión[22] para «partido.1».

Ítem léxico	**partido**						
SenseID	**partido.1**						
Dataset	**VP**	**FP**	**VN**	**FN**	**Precisión**	**Cobertura**	**Puntaje F**
1	8	15	2	1	0,8	0,888	0,842
2	10	12	2	2	0,833	0,833	0,833
3	10	13	1	2	0,909	0,833	0,869
4	8	15	1	2	0,888	0,8	0,842
5	13	14	1	1	0,928	0,928	0,928
				Promedio	**0,872**	**0,856**	**0,863**

[21] Los resultados para todos los experimentos de pilotaje correspondientes al corpus SENSEVAL-3, que incluyen pruebas para las distintas medidas de reducción de la dimensión que fueron consideradas preliminarmente, como *chi*-cuadrado, ganancia de información e información mutua, se encuentran disponibles en el siguiente directorio: https://github.com/fredyrodrigors/tesis-phd/tree/main/experimento_senseval-3.

[22] Para todas las matrices de confusión que se presentan en este capítulo, las siglas corresponden a las siguientes variables: verdaderos positivos (VP); falsos positivos (FP); verdaderos negativos (VN); y falsos negativos (FN).

Tabla 10. Matriz de confusión para «partido.2».

Ítem léxico	partido						
SenseID	partido.2						
Dataset	VP	FP	VN	FN	Precisión	Cobertura	Puntaje F
1	15	8	1	2	0,937	0,882	0,909
2	12	10	2	2	0,857	0,857	0,857
3	13	10	2	1	0,866	0,928	0,896
4	15	8	2	1	0,882	0,937	0,909
5	14	13	1	1	0,933	0,933	0,933
				Promedio	**0,895**	**0,907**	**0,901**

Tabla 11. Resultados del sistema de desambiguación automática para «partido».

	M Precisión	M Cobertura	M Puntaje F
«partido.1»	0,872	0,856	0,864
«partido.2»	0,895	0,907	0,901
Macropromedio del sistema	**88,33%**	**88,19%**	**88,25%**

En el caso de «partido.1», el promedio del puntaje F fue más bajo [$\overline{X}$ = 0,864; *DE* = 0,038][23] que en «partido.2» [$\overline{X}$ = 0,901; *DE* = 0,027]. No obstante, ambos puntajes constituyen un macropromedio de 88,25% para el sistema de desambiguación basado en el corpus y la metodología SENSEVAL-3 utilizando el algoritmo bayesiano ingenuo, y considerando una medida de información mutua, que se explicará más adelante, para la reducción de la dimensión de los datos.

Respecto al corpus, se utilizó el repertorio para una de las tareas propuestas en SENSEVAL-3, basado en una muestra léxica. En esta modalidad se seleccionan ventanas contextuales en las que se busca desambiguar solamente una palabra objetivo. En términos generales, la construcción de un corpus de muestra léxica en el contexto del testeo de tareas de PLN sigue criterios que adhieren a principios de muestreo e inferencia estadísticos. Este criterio metodológico basado en el muestreo estadístico no es necesariamente aplicable para la construcción de un corpus lingüístico. En el caso de SENSEVAL-3, se trata de una muestra léxica que no delimita la población total de forma rigurosa. En efecto, debido al gran tamaño de la población a la que aspira, siempre será posible

[23] Donde $\overline{X}$ equivale al promedio y *DE* a la desviación estándar.

demostrar que alguna característica no está adecuadamente representada en la muestra. A pesar de lo anterior, y al verificar las características de los corpus desarrollados en el ámbito de la lingüística computacional, típicamente se aceptan los resultados de cada implementación como si cualquier montaje de datos textuales se hubiese llevado a cabo de una manera metodológicamente correcta. De esta manera, se intentan prever posibles objeciones relacionadas con criterios provenientes desde el marco teórico de la lingüística de corpus. Sin embargo, se trata de un modo de proceder que puede provocar una tendencia a maximizar los efectos del error experimental, dado el problema que plantean tanto la representatividad como el balance del corpus. Si bien en el ámbito de la lingüística de corpus se establece que toda muestra está sesgada de alguna manera, los corpus desarrollados para implementaciones en PLN no son evaluados como una variable que puede ser capaz de modificar los resultados experimentales de cualquier sistema de procesamiento de datos textuales.

Según lo anterior, la evaluación de los resultados obtenidos por un sistema de desambiguación léxica automática basado en el corpus SENSEVAL-3 debe necesariamente cuestionar los parámetros metodológicos desde los que se obtuvo la muestra léxica en análisis, y si a partir de ella las conclusiones alcanzadas son válidas o presentarían un sesgo de confirmación relevante. Así, el desafío para la desambiguación en un corpus auténtico dependerá de la aparición de determinadas unidades léxicas para establecer el peso estadístico de una u otra clasificación, en desmedro de casos en los que las palabras adyacentes a la palabra objetivo no aparezcan. Si bien se observan resultados de un sistema de desambiguación eficiente en cuanto a los parámetros de precisión y cobertura, se trataría de un comportamiento altamente dependiente de las características del corpus.

En relación con la metodología, el procedimiento propuesto de tareas de procesamiento demuestra ser eficiente para ser aplicado en los experimentos de aprendizaje automático posteriores. Ahora bien, según los resultados y las observaciones del experimento inicial con el corpus SENSEVAL-3, de aquí en adelante se ha decidido trabajar con la medida de información mutua para la reducción de la dimensión del corpus, dado que los datos más relevantes de las instancias en análisis corresponden, precisamente, a los valores de la frecuencia absoluta para la información mutua en una matriz *N-grama/documento.*

Para solventar estos problemas derivados del corpus de prueba en el montaje del corpus basado en las unidades léxicas y sus correlatos conceptuales en FunGramKB, se realizó una selección semiautomática de unidades léxicas (sustantivos) potencialmente polisémicas, correspondientes con conceptos de la base de conocimiento FunGramKB, junto con sus respectivos contextos oracionales.

En total, se seleccionaron 120 instancias para cada unidad léxica, extraídas desde el *subcorpora* de prensa escrita chilena perteneciente al corpus CODICACH (Sadowsky, 2006). Tras filtrar esta selección, se decidió trabajar con las siguientes unidades léxicas: los sustantivos «cabeza», «cara» y «carta». A continuación, se presenta el proceso pormenorizado de selección de unidades léxicas, junto con sus representaciones conceptuales en la base de conocimiento FunGramKB. Luego se establecen los parámetros para el montaje del corpus en análisis.

5.3 Proceso de selección de conceptos en la subontología #ENTITY

El proceso de selección de conceptos básicos es el primer paso para el montaje del corpus en análisis. Su objetivo es verificar en la base de conocimiento FunGramKB cuáles son los conceptos que serán considerados como casos de polisemia, junto con sus respectivos postulados de significado. Además, cada concepto integra una lexicalización que, a su vez, será la unidad léxica que corresponderá con la palabra objetivo en el corpus. Este procedimiento requiere ser llevado a cabo antes de la fase de consolidación del corpus, puesto que el sistema de desambiguación propuesto dependerá del conocimiento almacenado en la ontología. Si se ejecutara el proceso inverso; es decir, desde el corpus hacia la base de conocimiento, se correría el riesgo de incorporar demasiada información nueva en la ontología, lo que supondría un esfuerzo adicional considerable, a la vez que amplificaría el problema del cuello de botella en la representación el conocimiento. Los conceptos candidatos, entonces, cumplen con el requisito de presentar dos o mas sentidos, correspondientes a otros conceptos ya ingresados en la base de conocimiento.

5.3.1 La polisemia en «cabeza» y sus representaciones conceptuales en FunGramKB

En el primer caso, la polisemia de «cabeza» está relacionada con cinco conceptos básicos en la base de conocimiento, como se muestra en la Tabla 12:

a. Como parte superior del cuerpo en humanos, o anterior en algunas especies animales, en la que típicamente se encuentra el cerebro, y que podría considerar además la parte superior de una cosa, correspondiente al concepto básico +HEAD_00.
b. Como la habilidad para elaborar pensamientos o para ejercer el juicio de manera acertada, correspondiente al concepto básico +INTELLIGENCE_00.
c. Como persona que preside, gobierna o está formalmente a cargo de un grupo, comunidad o corporación, correspondiente al concepto básico +CHIEF_00.

d. Como persona que dirige, inspira o es el punto de referencia para otras personas, correspondiente al concepto básico +LEADER_00[24].

Tabla 12. Polisemia en «cabeza» y sus representaciones en FunGramKB[25].

Unidad léxica	Concepto (sentidos)	Descripción en FunGramKB
Cabeza (*head*)	+HEAD_00	The upper or front part of the body in animals; contains the face and brains; "he stuck his head out the window"
	+INTELLIGENCE_00	Your ability to think, feel, and imagine things
	+CHIEF_00	A person who is in charge; "the head of the whole operation"
	+LEADER_00	A person who rules or guides or inspires others

5.3.2 La polisemia en «cara» y sus representaciones conceptuales en FunGramKB

En el segundo caso, la polisemia de «cara» se vincula con dos conceptos básicos en la base de conocimiento, como se muestra en la tabla 13:

a. Como la parte frontal de la cabeza humana, en la que se perciben órganos como la boca, nariz, ojos, etc., correspondiente al concepto básico +FACE_00.
b. Como la superficie de una cosa o de un lugar, correspondiente al concepto básico +SIDE_00.

Tabla 13. Polisemia en «cara» y sus representaciones en FunGramKB.

Unidad léxica	Concepto (sentidos)	Descripción en FunGramKB
Cara (*face*)	+FACE_00	The front of the head from the forehead to the chin and ear to ear; "he washed his face"; "I wish I had seen the look on his face when he got the news"
	+SIDE_00	A surface forming part of the outside of an object; "he examined all sides of the crystal"; "dew dripped from the face of the leaf"

[24] El sentido de «cabeza», si bien está capturado en la base de conocimiento como la parte superior de cualquier entidad inanimada, correspondiente al concepto básico +TOP_00, no fue incluido en la selección de conceptos dado que no presentaba casos de aparición en la submuestra de CODICACH.

[25] Se han incluido las descripciones para cada sentido en inglés, que es la lengua en la que se reportan en la base de conocimiento FunGramKB.

5.3.3 La polisemia en «carta» y sus representaciones conceptuales en FunGramKB

En el tercer caso, la polisemia de «carta» esta relacionada con dos conceptos básicos y un concepto terminal en la base de conocimiento, como se muestra en la tabla 14:

a. Como una misiva o papel escrito que una persona envía a otras personas, correspondiente al concepto básico +LETTER_00.
b. Como una cartulina rectangular con números y dibujos que se utiliza para juegos, correspondiente al concepto básico +CARD_00.
c. Como la lista de platos y bebidas que se ofrecen en un restaurante, correspondiente al concepto terminal $MENU_00.

Tabla 14. Polisemia en «carta» y sus representaciones en FunGramKB.

Unidad léxica	Concepto (sentidos)	Descripción en FunGramKB
Carta (*letter*)	+LETTER_00	A written message addressed to a person or organization; "wrote an indignant letter to the editor"
	+CARD_00	A small piece of thick stiff paper with numbers or pictures on them, used to play a particular game
	$MENU_00	A list of dishes available at a restaurant; "the menu was in French"

5.4 Corpus Dinámico del Castellano de Chile (CODICACH)

Luego de la revisión de la muestra léxica basada en el corpus SENSEVAL-3, se estableció la necesidad de contar con un corpus desarrollado y consolidado a partir de criterios lingüísticos fundamentados en los estándares de la lingüística de corpus, y no en el criterio metodológico basado en el muestreo estadístico, que evidenció ser deficiente en cuanto al potencial explicativo para los resultados de los experimentos piloto. Según lo anterior, se utilizó una submuestra extraída desde CODICACH. Este corpus corresponde a una muestra representativa sincrónica del español escrito de Chile, desarrollado por Sadowsky (2006). Se trata de un recurso lingüístico informatizado compuesto por aproximadamente 800 millones de unidades léxicas registradas en 1,3 millones de documentos, y divididas en 102 *subcorpora* organizados según la fuente textual.

5.4.1 Submuestra de CODICACH y colecciones de documentos

Para esta investigación se seleccionó una submuestra desde el *subcorpora* Periodismo, con un conteo de 534.921.215 unidades léxicas disponibles. Cada una de las columnas a partir de las que se organizó el corpus corresponde a las variables de *corpusID* (identificador de cada instancia en un archivo digital del corpus CODICACH); *source* (fuente desde la que se extrae la instancia en el corpus, correspondiente a un medio de comunicación escrito chileno, como periódico o revista); *context* (ventana de palabras en la que aparece la palabra objetivo); *senseID* (etiqueta para el sentido de la palabra objetivo en la ventana contextual correspondiente, que a su vez se relaciona con el concepto en COREL extraído desde la base de conocimiento FunGramKB, con el potencial de ser utilizado como clave primaria). Todos los sentidos para las 120 instancias correspondientes a cada una de las unidades léxicas en análisis fueron etiquetados manualmente. Un extracto de la organización del corpus para cada una de las palabras objetivo en análisis se puede ver en las siguientes tablas:

Tabla 15. Ejemplos de organización de la submuestra para «cabeza»[26].

CorpusID	Source	Context	SenseID
280160900	Capital	Se supone que a esas alturas a el pájaro se le ha acabado el oxígeno y ha desarrollado un músculo en la «cabeza» que lo obliga a picar y salir.	HEAD_00[27]
504560692	El Mercurio	En cualquier caso, todavía no hay ninguna «cabeza» visible para el grupo, sabido el gusto de Miranda por operar fuera de las luces.	LEADER_00
632975648	La Tercera	Por ello, el 12 de diciembre debes votar con la «cabeza» y con el corazón.	INTELLIGENCE_00
842196090	Qué Pasa	La «cabeza» de la empresa de telecomunicaciones más importante de España no dejó que su obsesión descansara.	CHIEF_00

[26] La selección de las 120 instancias que componen la submuestra para la unidad léxica «cabeza», en formato *.csv*, está disponible en https://github.com/fredyrodrigors/tesis-phd/blob/main/corpus_seleccion_codicach/cabeza_corpus_seleccion.csv. Además, se ha incluido al final de este trabajo como Anexo 3.

[27] Durante la ejecución del procedimiento de validación cruzada, DAMIEN no reconoce el atributo de clase cuando la etiqueta del concepto lleva el signo +. Por esta razón, en *senseID* se decidió no incluirlo.

Tabla 16. Ejemplos de organización la submuestra para «cara»[28].

CorpusID	Source	Context	SenseID
280194893	Capital	El objetivo es que los estadounidenses adquieran videos, ropa y regalos con la «cara» de Mickey y preparen sus vacaciones en los parques de atracciones a través de internet.	FACE_00
478396484	El Mercurio	La «cara» poniente es ondeada y allí están todas las áreas de servicios de los bomberos.	SIDE_00

Tabla 17. Ejemplos de organización de la submuestra para «carta»[29].

CorpusID	Source	Context	SenseID
652923447	La Tercera	Paulatinamente ha ido desapareciendo la costumbre de comunicarse a través de una simple «carta»	LETTER_00
442976051	LUN	Con ese juego de «cartas» sólo faltará Jorge Hevia para tener el póker completo	CARD_00
515600079	El Mercurio	El actual chef Cristián Rebolledo ofrece una «carta» claramente inspirada en esa cocina, pero con aportes personales interesantes	MENU_00

En términos cuantitativos, esta submuestra léxica cuenta con un total de 11.866 palabras (*tokens*), distribuidos de la siguiente manera para cada colección de documentos según palabras objetivo:

Tabla 18. Descripción cuantitativa para cada colección de documentos desde la submuestra de CODICACH.

Colección	Palabras	Densidad léxica	Palabras/Oración
Cabeza	3.900	40%	29,0
Cara	3.772	39%	30,0
Carta	4.190	39%	31,8
Promedio	**3.954**	**0,39**	**30,26**

[28] La selección de 120 instancias que componen la colección de documentos para la unidad léxica «cara», en formato *.csv*, está disponible en https://github.com/fredyrodrigors/tesis-phd/blob/main/corpus_seleccion_codicach/cara_corpus_seleccion.csv. Esta selección se ha incluido al final del trabajo como Anexo 4.

[29] La selección de 120 instancias que componen la colección de documentos para la unidad léxica «carta», en formato *.csv*, está disponible en https://github.com/fredyrodrigors/tesis-phd/blob/main/corpus_seleccion_codicach/carta_corpus_seleccion.csv. Además, al final de este trabajo se ha incluido como Anexo 5.

De acuerdo con la información presentada en las tablas 15, 16, 17 y 18, se observa que las tres colecciones de documentos, para las palabras objetivo «cabeza», «cara» y «carta» respectivamente, se diseñaron considerando específicamente los estándares de calidad, representatividad y recuperabilidad (Dash, 2010). Primero, en cuanto al estándar de calidad, todas las instancias fueron obtenidas a partir de muestras de escritura auténticas provenientes de medios de prensa escrita chilenos, sin ninguna intervención que pueda ser considerada como una circunstancia de producción artificial. Segundo, para cumplir con el estándar de representatividad, si bien se determinó un ámbito específico correspondiente al *subcorpora* Periodismo de CODICACH (Sadowsky, 2006), esta selección se realizó resguardando que todos los medios de prensa disponibles en el *subcorpora* referido tuvieran por lo menos una aparición durante el proceso semiautomático de identificación de casos en análisis. Otro punto relevante en cuanto a la aplicación de este estándar, de acuerdo con la tabla 18, es que se planificó una variación mínima entre los valores para los indicadores de densidad léxica, y la proporción de palabras por oración.

En el caso de la densidad léxica, también llamada *types/tokens ratio*, esta se calculó considerando el número de palabras únicas (*types*) dividido por el número de palabras totales (*tokens*). Este indicador arrojó un promedio del 39% de densidad léxica [*DE* = 0,005], junto con un promedio de 30,26 palabras por oración [*DE* = 1,41], llamado *tokens/sentences ratio*, correspondiente a su vez a la proporción promedio de palabras totales dividida por el número estimado de segmentos entre signos de puntuación «.» (*sentences*). Por último, para resguardar el criterio de recuperabilidad, el resultado de estos procedimientos se almacenó en tres archivos unificados con extensión *.csv,* que incluyeron las variables *corpusID*, *source, context* y *senseID*. Además, se compilaron tres documentos en formato *.txt* que incluyeron solo las 120 instancias en análisis. Estas, su vez, incorporaron cada palabra objetivo junto con su respectiva ventana contextual, correspondiente a la columna *context* para la submuestra léxica.

5.5 Procedimiento en DAMIEN para experimentos de aprendizaje automático

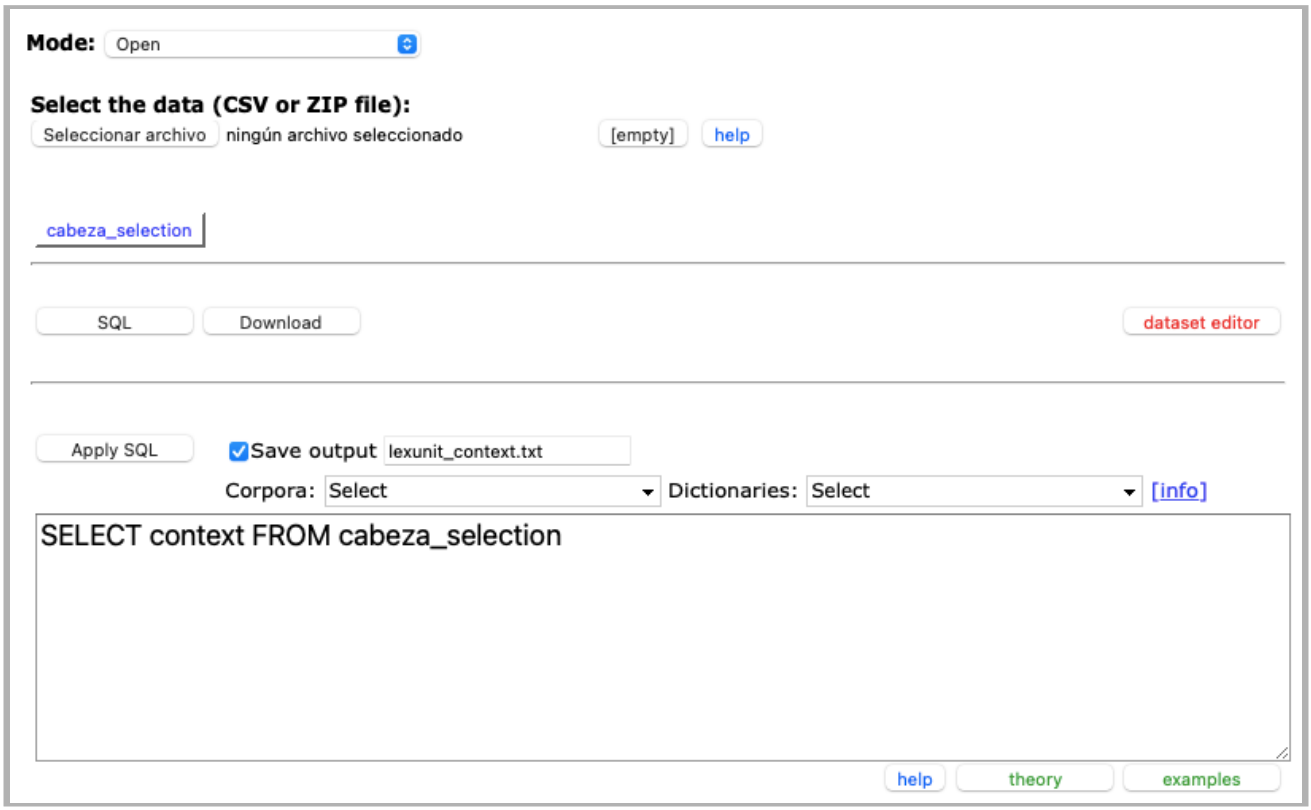
A continuación, se presenta el procedimiento específico para la ejecución en DAMIEN de los experimentos de aprendizaje automático aplicando el algoritmo bayesiano ingenuo, que se divide en cuatro subtipos de procedimientos: preprocesamiento, procesamiento, minería textual y evaluación. Posteriormente, para cada uno de estos subtipos de procesamiento se estableció un número determinado de tareas. Finalmente, para el desarrollo estandarizado de los experimentos se requirió la

ejecución de diez tareas de procesamiento con sus respectivas secuencias de pasos en DAMIEN. El archivo de partida es un documento *.csv* con cuatro columnas correspondientes a la submuestra de CODICACH ya descrita en el apartado anterior: *corpusID*, *source*, *context*, *senseID*; utilizando como separador el signo | (correspondiente al archivo *lexunit_selection.csv* en nuestro repositorio).

5.5.1 Tareas de preprocesamiento

Se establecieron tres tareas de preprocesamiento junto con sus respectivas tareas en DAMIEN. La primera tarea de preprocesamiento consistió en extraer los datos textuales correspondientes a la ventana contextual para cada una de las instancias seleccionadas. Específicamente, se realizó un procedimiento para la extracción de un archivo *.txt* que almacenara la información de la columna *context* correspondiente al documento *lexunit_selection.csv*. La secuencia en DAMIEN se ejecutó en el editor de *Corpus* mediante la aplicación de un comando SQL para la selección de la columna *context*, como se ve en la siguiente figura:

Figura 10. Extracción de la ventana contextual en DAMIEN.

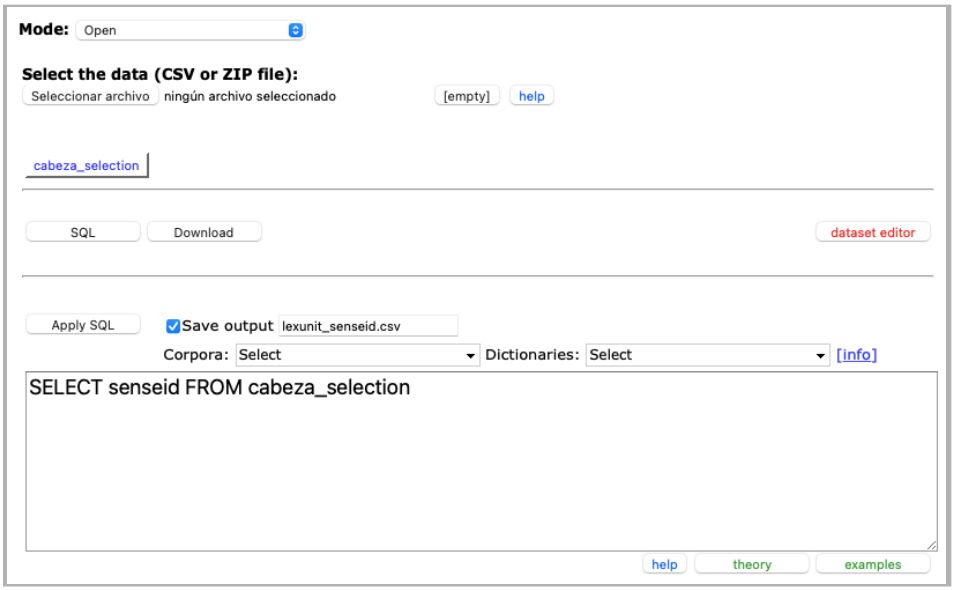


La segunda tarea de preprocesamiento consistió en generar una colección de documentos sin anotar. Para esto, la secuencia que se ejecutó en DAMIEN fue el cambio de tamaño (*file resizing*) del archivo *.txt,* mediante la aplicación de una expresión regular *\n* para dividir el contenido de cada cotexto en 120 documentos, como se evidencia en la siguiente figura:

Figura 11. Generación de una colección de documentos (sin anotar) en DAMIEN.

La tercera tarea de preprocesamiento fue extraer las etiquetas *senseID*, correspondientes a cada uno de los sentidos seleccionados para las palabras objetivo en la colección de documentos. Al igual que en la primera tarea, este procedimiento se llevó a cabo mediante la ejecución de un comando SQL, como se muestra en la siguiente figura:

Figura 12. Extracción de etiquetas *senseID* en DAMIEN.



Finalmente, las tres tareas de preprocesamiento anteriores y sus respectivas secuencias en DAMIEN se presentan de manera pormenorizada en la siguiente tabla:

Tabla 19. Tareas de preprocesamiento para experimentos de aprendizaje automático.

Nº	Tarea	Secuencia en DAMIEN
1	Extracción de la ventana contextual	1. Extraer un archivo *.txt* para la columna *context* desde lexunit_selection.csv 2. Corpus > Open 3. Cargar un *.zip* con la tabla en *.csv* 4. Aplicar comando SQL: SELECT context FROM lexunit_selection (el resultado se genera en *.csv*) 5. Guardar como *.txt* 6. Guardar resultado = *lexunit_context.txt*
2	Generación de una colección de documentos (sin anotar)	1. Cargar un archivo .txt 2. Corpus > Pre-process > File resizing > Split (regex \n) 3. Guardar resultado = *lexunit_collection.zip*
3	Extracción de etiquetas *senseID*	1. Extraer un archivo .csv para *senseID* desde lexunit_selection.csv 2. Corpus > Open 3. Comando SQL: SELECT senseID FROM *.csv 4. Guardar resultado = *lexunit_senseid.csv*

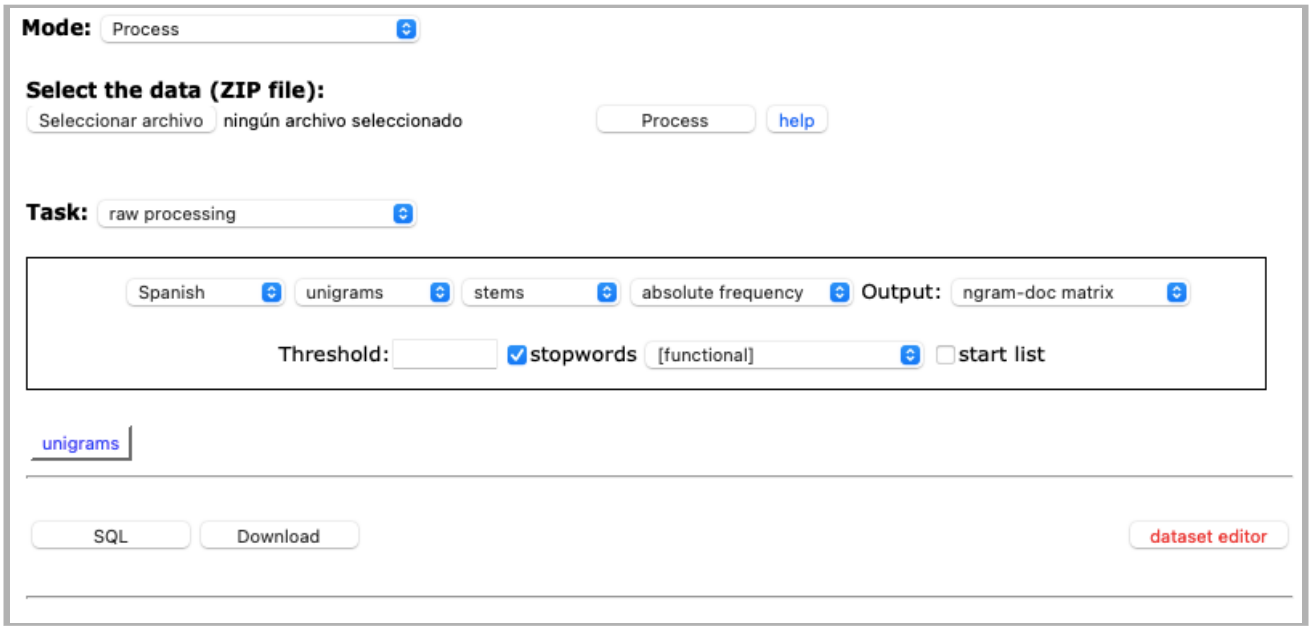
5.5.2 Tareas de procesamiento

Se establecieron cuatro tareas de procesamiento y sus respectivas secuencias en DAMIEN. La primera tarea (cuarta en la secuencia general) consistió en generar la primera matriz *N-grama/documento*. Para la ejecución de esta tarea fue necesario procesar la colección de documentos correspondiente a los cotextos que contuvieran cada palabra objetivo, para así establecer un análisis de la frecuencia de aparición de unidades léxicas en cada una de las instancias en análisis. Las condiciones para la implementación de esta secuencia fueron: (1) que el procesamiento se realice en lengua española; y (2) que la captura corresponda a unigrama representados por cada *stem*[30] y su frecuencia absoluta. Además, se estableció una lista de palabras vacías o palabras de función (*stopwords)*; es decir, no

[30] En cuanto al concepto de *stem*, según Bauer (2004), se trata de un término que se utiliza para designar una unidad interna siempre presente en un lexema, que permanece estable cuando se han extraído todos los afijos presentes a la vez que incluye los potenciales morfemas flexivos. Así, se ha decidido trabajar con la variable *stem* debido a la frecuencia y relevancia en el corpus de las distintas flexiones de número para cada una de las unidades léxicas en análisis.

fueron consideradas en la matriz las palabras funcionales, con el objetivo de eliminar el ruido documental. La secuencia en DAMIEN se puede revisar en la siguiente figura:

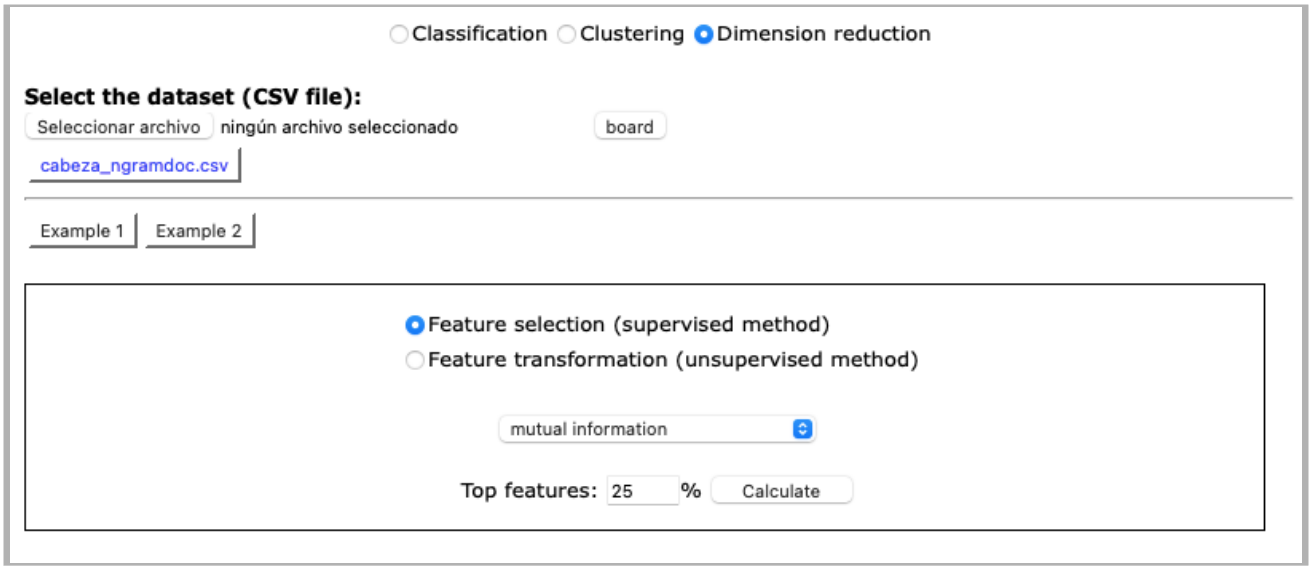
Figura 13. Generación de una matriz *N-grama/documento* en DAMIEN.



La segunda tarea de procesamiento (quinta en la secuencia general), consistió en la generación de una lista de inicio (etiquetada como *startlist*) con el objetivo de filtrar aquellas palabras con mayor peso estadístico dentro del corpus, mediante la aplicación de la medida de información mutua.

La medida de información mutua determina la reducción de la incertidumbre de una variable dado el valor conocido de otra variable; es decir, calcula la cantidad de información que se puede obtener a partir de una variable aleatoria, considerando un valor conocido. En términos estadísticos, esta medida se utiliza para calcular la dependencia entre dos variables aleatorias. Específicamente, se para esta tarea de procesamiento se midió la cantidad promedio de información que ciertas unidades léxicas transmiten o proyectan sobre otras unidades léxicas presentes en cada colección de documentos. El resultado fue una lista de las palabras que presentaron una mayor significancia estadística luego del análisis de frecuencia proporcionado por la matriz *N-grama/documento*. Finalmente, se utilizó como filtro el 25% superior de esta lista. La siguiente figura muestra, en la interfaz de DAMIEN, la secuencia antes descrita.

Figura 14. Generación de una lista de inicio en DAMIEN.



En tercera tarea de procesamiento (sexta en la secuencia general) se generó una segunda matriz *N-grama/documento*, considerando el filtro de la *startlist* creada a partir de la aplicación de la medida de información mutua:

Figura 15. Generación de una matriz *N-grama/documento* con lista de inicio en DAMIEN.

La cuarta tarea de procesamiento, y séptima en la secuencia general, fue fundamental para la ejecución del proceso de desambiguación léxica automática. Consistió en la generación de una matriz *N-grama/documento* filtrada y anotada con los sentidos correspondientes para cada una de las palabras objetivos presentes en los documentos. Para esto, se incluyó en la matriz una nueva columna con los

sentidos que fueron extraídos en un paso anterior, mediante la aplicación del comando de unión a la derecha (*join right*), incluyendo el símbolo barra | como separador. En la siguiente figura se presenta la interfaz en DAMIEN para esta secuencia:

Figura 16. Secuencia *join right* para generar una matriz filtrada y anotada en DAMIEN.

Finalmente, en la siguiente tabla se presentan las cuatro tareas de procesamiento y sus respectivas secuencias en DAMIEN, que incluyen cada uno de los pasos y configuraciones de manera pormenorizada:

Tabla 20. Tareas de procesamiento para experimentos de aprendizaje automático.

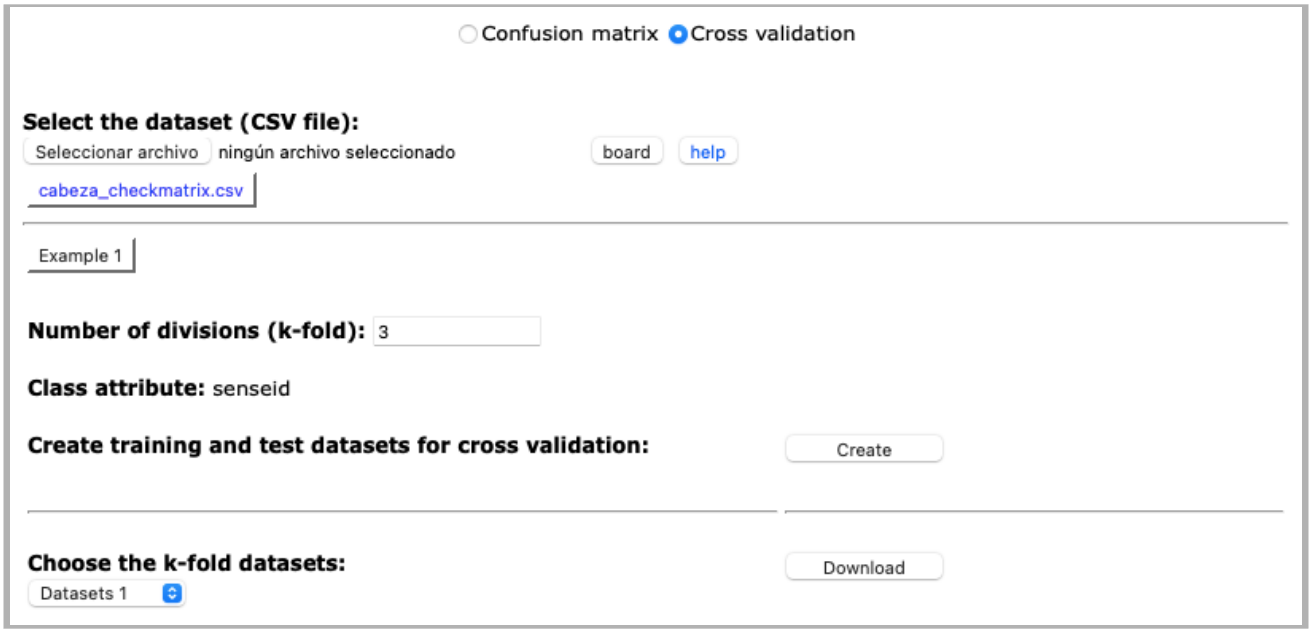
Nº	Tarea	Secuencia en DAMIEN
4	Generación de la primera matriz *N-grama/documento*	1. Corpus > Process > Task > Raw processing 2. Settings = Spanish; unigrams; stems; absolute frequency 3. Output = doc-ngram matrix 4. Stopwords [functional] 5. Guardar resultado = *lexunit_ngramdoc.csv*
5	Generación de una *startlist* para filtrar palabras con mayor peso estadístico en el corpus	1. Mining > Dimension reduction 2. Feature selection = mutual information 3. Top Features = 25 4. Guardar resultado = lexunit_startlist.csv

6	Generación de matriz *N-grama/documento* con *startlist*	1. Aplicar el filtro = lexitem_collection_startlist.zip 2. Corpus > Process > Task > Raw processing Settings = Spanish; unigrams; stems; absolute frequency 3. Output = doc-ngram matrix 4. Startlist = lexunit_startlist 5. Guardar resultado = *lexunit_ngramdoc_startlist.csv*
7	Generación de matriz *N-grama/documento* filtrada y anotada	1. Aplicar comando JOIN (right) para incluir una columna con los *senseID* de cada documento 2. Guardar resultado = *lexunit_checkmatrix.csv*

5.5.3 Tarea de minería textual

La tarea de minería textual, correspondiente a la octava tarea en la secuencia general, consiste en la validación cruzada de las 120 instancias procesadas en la matriz *N-grama/documento* filtrada y anotada; es decir, en la división aleatoria de los datos textuales en una cantidad determinada de grupos del mismo tamaño considerando la columna *senseID* como el atributo de clase. En este caso, se generaron aleatoriamente tres *trainings sets* y tres *test sets*, cada uno con 40 instancias. Este procedimiento en DAMIEN se puede ver en la siguiente figura:

Figura 17. Validación cruzada en DAMIEN.



La secuencia pormenorizada en DAMIEN para la tarea de validación cruzada se expone en la siguiente tabla:

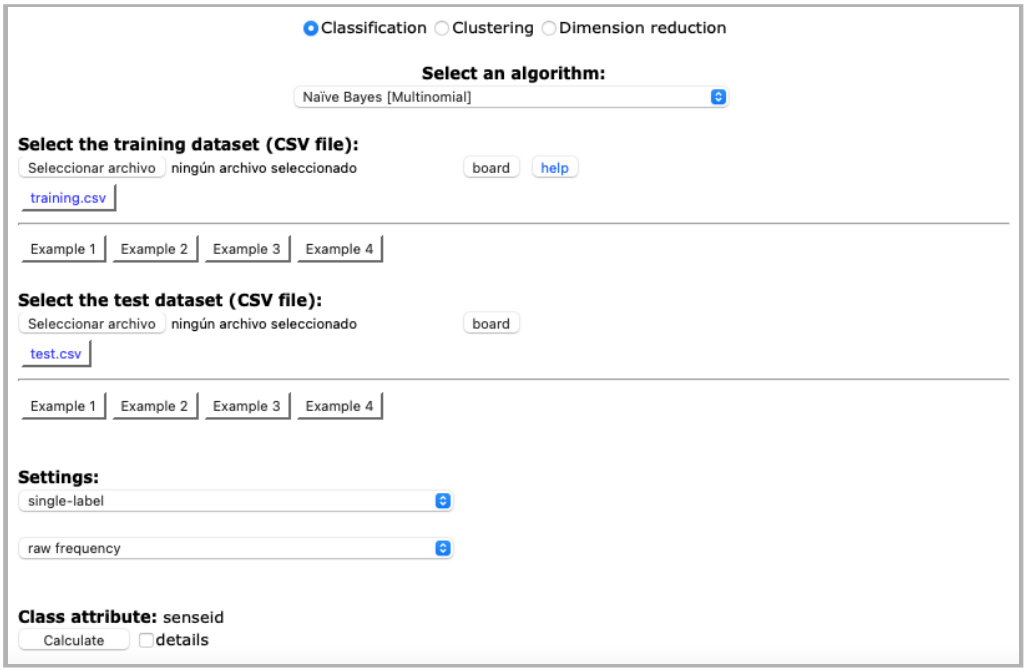
Tabla 21. Tarea de minería textual para experimentos de aprendizaje automático.

Nº	Tarea	Secuencia en DAMIEN
8	Validación cruzada	1. Evaluation > Cross Validation Settings: *k*-fold = 3 2. El resultado corresponde a la creación de tres carpetas con los archivos *training.txt*, *test.txt* y *predicted.txt*.

5.5.4 Tareas de evaluación

Se determinaron dos tareas de evaluación con las que concluyeron los experimentos de aprendizaje automático. La primera tarea (novena en la secuencia general) consistió en a la aplicación del algoritmo *Naïve Bayes* para la clasificación de cada *test dataset* basado en el respectivo *training dataset*. Para realizar esta secuencia, se cargaron los conjuntos de datos generados en la validación cruzada. Finalmente, el correspondiente atributo de clase fue la columna *senseID*, identificado automáticamente. En la siguiente figura se puede ver la interfaz de DAMIEN:

Figura 18. Aplicación del algoritmo bayesiano ingenuo en DAMIEN.



La segunda tarea de evaluación, correspondiente a la décima en la secuencia general, es la generación de una matriz de confusión para la evaluación de cada sistema de desambiguación. Este procedimiento requiere crear carpetas con los datos para cada *senseID* por *dataset*. Luego en cada carpeta se deben incluir los correspondientes documentos con los valores tanto esperados como predichos. La interfaz DAMIEN para esta tarea, junto con un ejemplo de matriz de confusión, se exponen en las siguientes figuras:

Figura 19. Generación de una matriz de confusión en DAMIEN.

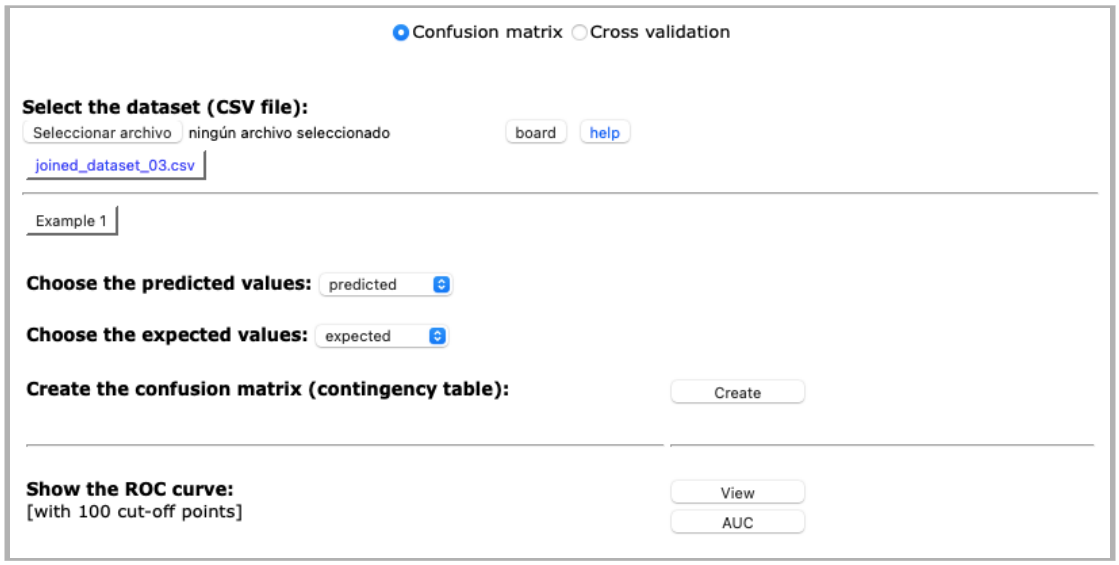
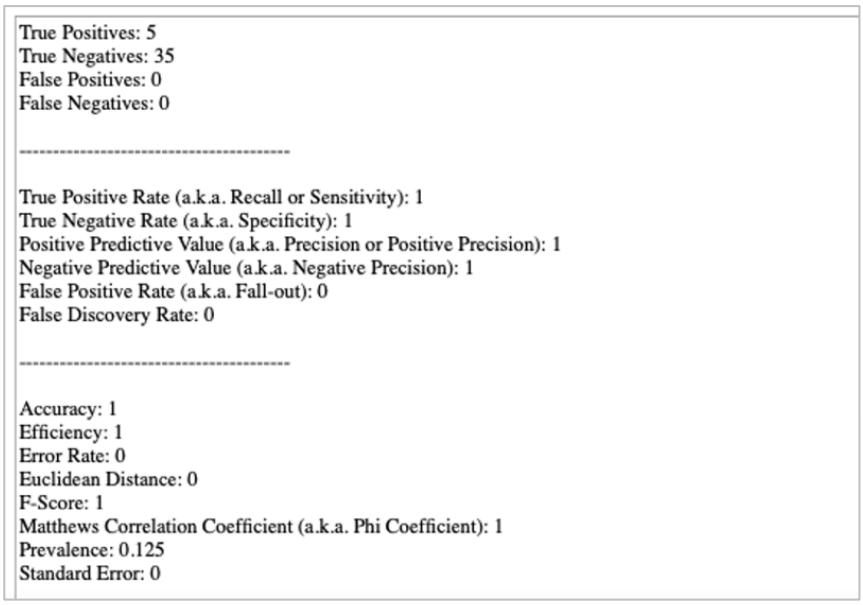


Figura 20. Ejemplo de matriz de confusión en DAMIEN.



Finalmente, las tareas de evaluación y su secuencia de pasos en DAMIEN se muestran de manera detallada en la siguiente tabla:

Tabla 22. Tareas de evaluación para experimentos de aprendizaje automático.

Nº	Tarea	Secuencia en DAMIEN
9	Aplicación de algoritmo bayesiano ingenuo	1. Mining > Classification > *Naïve Bayes* (multinomial) *2.* Seleccionar y cargar cada *training dataset* en *.csv* 3. Seleccionar y cargar *test dataset* en *.csv* Settings = single-label; raw frequency 4. Class attribute (lo identificará automáticamente desde la tabla) = *senseid* 5. Los resultados se deben guardar como un archivo *predicted.csv* en las carpetas correspondientes para cada *dataset*.
10	Generación de una matriz de confusión para la evaluación del sistema	1. Crear carpetas con los datos para cada *senseID* por *dataset.* En cada carpeta se deben incluir los correspondientes documentos *a_expected.txt* y *b_predicted.txt* 2. Aplicar comando JOIN (*right*) para generar la tabla de evaluación. 3. Reemplazar las etiquetas de s*enseID* por los valores de 0 y 1, correspondientes con cada uno de los sentidos en evaluación. 4. Reemplazar los nombres de las columnas, de izquierda a derecha, por las etiquetas *expected* y *predicted.* 5. Evaluation > Confusion matrix Guardar resultado = *lexunit_conmatrix_0x.txt*

Capítulo 6
Experimentos de base para la desambiguación léxica automática

A continuación, se exponen los resultados de los experimentos de aprendizaje automático basados en el corpus diseñado a partir de la submuestra de CODICACH. La metodología descrita en el capítulo cinco fue aplicada utilizando la herramienta DAMIEN. Así, este apartado desarrolla dos objetivos relevantes para la investigación: primero, ejecutar experimentos de base para el testeo de un método de desambiguación, en este caso el aprendizaje automático mediante el algoritmo bayesiano ingenuo; y segundo, implementar las tareas de procesamiento necesarias para llevar a cabo estos experimentos utilizando el entorno de trabajo DAMIEN.

6.1 Resultados de los sistemas de desambiguación en aprendizaje automático

Los sistemas de desambiguación para aprendizaje automático para las unidades léxicas en análisis «cabeza», «cara» y «carta», se desarrollaron considerando los sentidos disponibles en la base de conocimiento FunGramKB. Todos estos sentidos fueron testeados a partir de un corpus de entrenamiento etiquetado manualmente con 120 instancias, que incluyeron cada palabra objetivo junto con una ventana contextual. Este conjunto de instancias, para efectos de la evaluación, fue dividido en tres *datasets* aleatorizados para cada uno de los sentidos en análisis. Luego de la ejecución del algoritmo bayesiano ingenuo, se realizó una evaluación mediante una matriz de confusión para determinar, en última instancia, el macropromedio correspondiente a la media armónica (o puntaje F), que establecería los resultados del sistema mediante una función entre los valores promedio de precisión y cobertura para cada sentido.

6.1.1 Resultados del sistema de desambiguación automática para la unidad léxica «cabeza»[31]

En el caso de la unidad léxica «cabeza», se consideraron los sentidos +HEAD_00, +CHIEF_00, +LEADER_00 y +INTELLIGENCE_00. A continuación, se presentan las tablas con el resumen de la matriz de confusión para cada sentido, junto con los resultados para el sistema de desambiguación léxica automática.

[31] Todas las tareas de procesamiento para los experimentos de desambiguación léxica basados en aprendizaje automático, correspondientes a las unidades léxicas «cabeza», «cara» y «carta», según los pasos declarados en la metodología, están disponibles en el siguiente repositorio: https://github.com/fredyrodrigors/tesis-phd/tree/main/tareas_de_procesamiento.

Tabla 23. Matriz de confusión para el sentido +HEAD_00 de «cabeza»[32].

Ítem léxico	CABEZA						
SenseID	+HEAD_00						
Dataset	VP	FP	VN	FN	Precisión	Cobertura	Puntaje F
1	14	11	11	4	0,56	0,777	0,651
2	15	8	11	6	0,652	0,714	0,681
3	11	9	13	7	0,55	0,611	0,578
				Promedio	**0,587**	**0,701**	**0,637**

En cuanto al sentido +HEAD_00, en los tres *datasets* se mantuvo un rendimiento homogéneo, considerando específicamente los resultados de la dispersión para el promedio de la media armónica [$\overline{X}_F$ = 0,637; *DE* = 0,05], equivalente a un 64%. Los desempeños particulares se caracterizaron por una mayor precisión [*P* = 0,652] y cobertura [*C* = 0,714], correspondientes a un 68% de rendimiento para el *dataset* dos [*F* = 0,681].

Tabla 24. Matriz de confusión para el sentido +CHIEF_00 de «cabeza»[33].

Ítem léxico	CABEZA						
SenseID	+CHIEF_00						
Dataset	VP	FP	VN	FN	Precisión	Cobertura	Puntaje F
1	3	1	29	7	0,75	0,3	0,428
2	1	4	30	6	0,2	0,166	0,181
3	3	3	26	8	0,5	0,272	0,352
				Promedio	**0,483**	**0,246**	**0,32**

[32] Las matrices de confusión completas para cada *dataset* del sentido +HEAD_00 se han incluido al final de este trabajo como Anexo 6. Además, se encuentran disponibles en:
dataset 1 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/head_conmatrix_dataset_01.csv;
dataset 2 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/head_conmatrix_dataset_02.csv;
dataset 3 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/head_conmatrix_dataset_03.csv;

[33] Las matrices de confusión completas para cada *dataset* del sentido +CHIEF_00 se han incluido al final de este trabajo como Anexo 7. Además, se encuentran disponibles en:
dataset 1 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/chief_conmatrix_dataset_01.csv;
dataset 2 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/chief_conmatrix_dataset_02.csv;
dataset 3 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/chief_conmatrix_dataset_03.csv.

En el caso del sentido +CHIEF_00, se observó un resultado bajo [< 50%] para el promedio de la media armónica [$\overline{X}_F$ = 0,32; *DE* = 0,126], equivalente a un 32% de rendimiento. Si bien los *datasets* uno y tres mostraron un desempeño similar, del 43% y 53% respectivamente, el *dataset* dos obtuvo resultados de precisión [*P* = 0,2] y cobertura [*C* = 0,166] particularmente deficientes, equivalentes a un 18%.

Tabla 25. Matriz de confusión para el sentido +LEADER_00 de «cabeza»[34].

Ítem léxico	CABEZA						
SenseID	+LEADER_00						
Dataset	VP	FP	VN	FN	Precisión	Cobertura	Puntaje F
1	4	6	26	4	0,4	0,5	0,444
2	1	4	28	7	0,2	0,125	0,153
3	4	7	27	2	0,363	0,666	0,470
				Promedio	**0,321**	**0,430**	**0,356**

Los resultados para el sentido +LEADER_00 corresponden a un desempeño del 36% considerando el promedio para la media armónica [$\overline{X}_F$ = 0,356; *DE* = 0,175]. Si bien se trata de un resultado bastante homogéneo en cuanto a la dispersión, el *dataset* dos obtuvo un rendimiento bajo para los indicadores de precisión [*P* = 0,2] y cobertura [*C* = 0,125], con puntaje F equivalente a un 15%.

[34] Las matrices de confusión completas para cada *dataset* del sentido +LEADER_00 se han incluido al final de este trabajo como Anexo 8. Además, se encuentran disponibles en:
dataset 1 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/leader_conmatrix_dataset_01.csv;
dataset 2 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/leader_conmatrix_dataset_02.csv;
dataset 3 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/leader_conmatrix_dataset_03.csv.

Tabla 26. Matriz de confusión para el sentido +INTELLIGENCE_00 de «cabeza»[35].

Ítem léxico	CABEZA						
SenseID	+INTELLIGENCE_00						
Dataset	VP	FP	VN	FN	Precisión	Cobertura	Puntaje F
1	0	1	35	4	0	0	*NaN*
2	2	5	30	3	0,285	0,4	0,333
3	0	3	32	5	0	0	*NaN*
				Promedio	**0,095**	**0,133**	**0,333**

Los resultados para el sentido +INTELLIGENCE_00 fueron los que arrojaron el rendimiento más irregular del sistema, considerando que el promedio para la media armónica [$\overline{X}_F = 0{,}333$; $DE = 0{,}192$], equivalente a un 33%, es una expresión válida solamente para los valores de precisión [$P = 0{,}285$] y cobertura [$P = 0{,}4$], correspondientes al *dataset* dos. Lo anterior se justifica dado que, para todas las casillas correspondientes a valores numéricos en los que se indique *NaN* (*not a number*), no ha sido posible establecer un resultado y, por tanto, esa variable no fue considerada en el análisis.

Tabla 27. Resultados del sistema de desambiguación automática para «cabeza».

	M Precisión	M Cobertura	M Puntaje F
+HEAD_00	0,587	0,701	0,637
+CHIEF_00	0,483	0,246	0,320
+LEADER_00	0,321	0,430	0,356
+INTELLIGENCE_00	0,095	0,133	0,333
Macropromedio del sistema	**37,15%**	**37,75%**	**41,15%**

El sistema de desambiguación automática para la unidad léxica «cabeza», que considera cuatro sentidos disponibles en la base de conocimiento, logra un rendimiento promedio del 41,15% [$DE = 0{,}151$], a partir de cuya dispersión se puede establecer un desempeño homogéneo, con la

[35] Las matrices de confusión completas para cada *dataset* del sentido +INTELLIGENCE_00 se han incluido al final de este trabajo como Anexo 9. Además, se encuentran disponibles en:
dataset 1 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/intelligence_conmatrix_dataset_01.csv;
dataset 2 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/intelligence_conmatrix_dataset_02.csv;
dataset 3 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cabeza/intelligence_conmatrix_dataset_03.csv.

excepción del sentido +HEAD_00 que presenta un promedio de la media armónica [$F = 0{,}637$] superior en 0,3 puntos al puntaje F más alto para el resto de los sentidos, correspondiente a +LEADER_00 [$F = 0{,}356$]. Estos resultados, además, indican que la cantidad de sentidos fue en desmedro de la capacidad del sistema para clasificar correctamente las instancias en análisis. Según lo anterior, el rendimiento más bajo fue el sentido +CHIEF_00, con un 32% [$F = 0{,}32$].

6.1.2 Resultados del sistema de desambiguación automática para la unidad léxica «cara»

La unidad léxica «cara» consideró los sentidos +FACE_00, y +SIDE_00, correspondientes a la base de conocimiento FunGramKB. Los resultados para el sistema de desambiguación automática son los siguientes:

Tabla 28. Matriz de confusión para el sentido +FACE_00 de «cara»[36].

Ítem léxico	CARA						
SenseID	+FACE_00						
Dataset	VP	FP	VN	FN	Precisión	Cobertura	Puntaje F
1	16	6	3	15	0,727	0,516	0,603
2	23	5	2	10	0,821	0,696	0,754
3	24	9	3	4	0,727	0,857	0,786
				Promedio	**0,758**	**0,690**	**0,714**

Los resultados para el sentido +FACE_00 se caracterizan por un alto desempeño del sistema, correspondiente a un 71% de rendimiento promedio según la media armónica, con una dispersión homogénea [$\overline{X}_F = 0{,}714$; $DE = 0{,}097$]. Específicamente, el *dataset* dos alcanzó el valor más alto de precisión [$P = 0{,}821$], mientras que el *dataset* tres obtuvo el valor más alto de cobertura [$C = 0{,}857$]. Estos resultados representaron, en términos generales, el desempeño más alto para los sistemas de desambiguación propuestos.

[36] Las matrices de confusión completas para cada *dataset* del sentido +FACE_00 se han incluido al final de este trabajo como Anexo 10. Además, se encuentran disponibles en:
dataset 1 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cara/face_conmatrix_dataset_01.csv;
dataset 2 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cara/face_conmatrix_dataset_02.csv;
dataset 3 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cara/face_conmatrix_dataset_03.csv.

Tabla 29. Matriz de confusión para el sentido +SIDE_00 de «cara»[37].

Ítem léxico	CARA						
SenseID	+SIDE_00						
Dataset	VP	FP	VN	FN	Precisión	Cobertura	Puntaje F
1	3	15	16	6	0,166	0,333	0,222
2	2	10	23	5	0,166	0,285	0,210
3	3	4	24	9	0,428	0,25	0,315
				Promedio	**0,253**	**0,289**	**0,249**

En cuanto al sentido +SIDE_00, el rendimiento promedio equivale a un 25% para la media armónica [$\overline{X}_F = 0{,}249$; $DE = 0{,}05$]. Si bien este resultado es bajo, el *dataset* tres mostró una precisión más alta [$P = 0{,}428$], aunque aún < 50%. No obstante, la dispersión promedio se muestra homogénea.

Tabla 30. Resultados del sistema de desambiguación automática para «cara».

	M Precisión	M Cobertura	M Puntaje F
+FACE_00	0,758	0,690	0,714
+SIDE_00	0,253	0,289	0,249
Macropromedio del sistema	**50,55%**	**48,95%**	**48,15%**

El sistema de desambiguación automática para la unidad léxica «cara» obtuvo un rendimiento promedio del 48,15% [$\overline{X}_F = 0{,}481$; $DE = 0{,}328$]. No obstante, hubo una diferencia de 0,465 puntos en el desempeño de los sentidos en análisis. Por tanto, se establece que el sistema no logra realizar la tarea de clasificación de manera eficiente, sobre todo para el caso de +SIDE_00.

6.1.3 Resultados del sistema de desambiguación automática para la unidad léxica «carta»

Para el análisis de la unidad léxica «carta», se consideraron los sentidos +LETTER_00, +CARD_00, y $MENU_00. A continuación, se presentan las tablas con el resumen de los resultados para el sistema de desambiguación automática.

[37] Las matrices de confusión completas para cada *dataset* del sentido +SIDE_00 se han incluido al final de este trabajo como Anexo 11. Además, se encuentran disponibles en:
dataset 1 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cara/side_conmatrix_dataset_01.csv;
dataset 2 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cara/side_conmatrix_dataset_02.csv;
dataset 3 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_cara/side_conmatrix_dataset_03.csv.

Tabla 31. Matriz de confusión para el sentido +LETTER_00 de «carta»[38].

Ítem léxico	CARTA						
SenseID	+LETTER_00						
Dataset	VP	FP	VN	FN	Precisión	Cobertura	Puntaje F
1	14	6	11	9	0,7	0,608	0,651
2	17	7	11	5	0,708	0,772	0,739
3	20	4	10	6	0,833	0,769	0,8
				Promedio	**0,747**	**0,716**	**0,730**

El sentido +LETTER_00 obtiene un promedio para la media armónica correspondiente a un 73% de rendimiento [$\overline{X}_F = 0{,}730$; $DE = 0{,}074$]. Estos resultados muestran un desempeño alto y homogéneo según la dispersión de los datos. En cuanto a los resultados particulares, la precisión más alta la alcanzó el *dataset* tres [$P = 0{,}833$], mientras que la cobertura más alta correspondió al *dataset* dos [$C = 0{,}772$].

Tabla 32. Matriz de confusión para el sentido +CARD_00 de «carta»[39].

Ítem léxico	CARTA						
SenseID	+CARD_00						
Dataset	VP	FP	VN	FN	Precisión	Cobertura	Puntaje F
1	5	9	23	3	0,357	0,625	0,454
2	4	6	27	3	0,4	0,571	0,470
3	2	4	29	5	0,333	0,285	0,307
				Promedio	**0,363**	**0,494**	**0,410**

[38] Las matrices de confusión completas para cada *dataset* del sentido +LETTER_00 se han incluido al final de este trabajo como Anexo 12. Además, se encuentran disponibles en:
dataset 1 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_carta/letter_conmatrix_dataset_01.csv;
dataset 2 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_carta/letter_conmatrix_dataset_02.csv;
dataset 3 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_carta/letter_conmatrix_dataset_03.csv.

[39] Las matrices de confusión completas para cada *dataset* del sentido +CARD_00 se han incluido al final de este trabajo como Anexo 13. Además, se encuentran disponibles en:
dataset 1 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_carta/card_conmatrix_dataset_01.csv;
dataset 2 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_carta/card_conmatrix_dataset_02.csv;
dataset 3 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_carta/card_conmatrix_dataset_03.csv.

En cuanto al sentido +CARD_00, el promedio para la media armónica equivale a un 41% de rendimiento [$\bar{X}_F = 0{,}410$; $DE = 0{,}089$]. En términos generales, los resultados indican una dispersión homogénea, lo que se traduce en un desempeño consistentemente < 50%. El rendimiento más alto lo alcanzó el *dataset* dos, tanto para el valor de precisión [$P = 0{,}4$] como de cobertura [$C = 0{,}571$], con un puntaje F equivalente al 47%.

Tabla 33. Matriz de confusión para el sentido $MENU _00 de «carta»[40].

Ítem léxico	CARTA						
SenseID	$MENU_00						
Dataset	VP	FP	VN	FN	Precisión	Cobertura	Puntaje F
1	4	2	29	5	0,666	0,444	0,533
2	6	0	29	5	1	0,545	0,705
3	5	5	28	2	0,5	0,714	0,588
				Promedio	**0,722**	**0,568**	**0,609**

El sentido $MENU_00 presentó un promedio para la media armónica equivalente a un 61% de rendimiento [$\bar{X}_F = 0{,}609$; $DE = 0{,}087$]. La precisión del *dataset* dos alcanzó el puntaje máximo [$P = 1$], lo que indica que el desempeño de las predicciones correctas fue de un 100%. Se trata de resultados eficientes, pero que evidencian una dispersión más alta que otros sentidos, considerando las diferencias entre los resultados para la media armónica de los tres *datasets*.

Tabla 34. Resultados del sistema de desambiguación automática para «carta».

	Precisión	Cobertura	Puntaje F
+LETTER_00	0,747	0,716	0,730
+CARD_00	0,363	0,494	0,410
$MENU_00	0,722	0,568	0,609
Macropromedio del sistema	**61,07%**	**59,27%**	**58,3%**

[40] Las matrices de confusión completas para cada *dataset* del sentido $MENU_00 se han incluido al final de este trabajo como Anexo 14. Además, se encuentran disponibles en:
dataset 1 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_carta/menu_conmatrix_dataset_01.csv;
dataset 2 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_carta/menu_conmatrix_dataset_02.csv;
dataset 3 https://github.com/fredyrodrigors/tesis-phd/blob/main/matrices_confusión/sentidos_carta/menu_conmatrix_dataset_03.csv.

El sistema de desambiguación automática para la unidad léxica «carta» alcanzó un desempeño promedio del 58,3% para su media armónica [$\overline{X}_F = 0{,}583$; $DE = 0{,}161$]. Se trata de un sistema de clasificación eficiente en términos generales, pero que muestra una dispersión más alta que los sistemas de «cabeza» y «cara». Eso se debe al impacto que tienen sobre el macropromedio los valores bajos para el sentido +CARD_00, por un lado, y el alto desempeño para el sentido $MENU_00, por otro. Finalmente, los resultados comparados para los macropromedios correspondientes a los sistemas de desambiguación automática aplicando el algoritmo bayesiano ingenuo son los siguientes:

Tabla 35. Macropromedios para los sistemas de desambiguación automática.

	Macropromedios		
Sistema	**Precisión**	**Cobertura**	**Puntaje F**
Cabeza	37,15%	37,75%	41,15%
Cara	50,55%	48,95%	48,15%
Carta	61,07%	59,27%	58,3%

En la tabla anterior se observa que dos de los tres sistemas alcanzan un desempeño > 50% para el promedio de la media armónica. En el caso de «cabeza», con un 41,15% de rendimiento para el promedio del puntaje *F*, si bien hubo una dispersión baja en los resultados pormenorizados, existiría cierta proporcionalidad entre los valores para los errores en la clasificación y la alta cantidad de sentidos disponibles; a saber, cuatro. En el caso de «cara», si bien se establecieron dos sentidos disponibles, los resultados del *dataset* dos fueron tan deficientes que impactaron en el puntaje *F* obtenido como macropromedio. En el caso de «carta», por el contrario, se evidencia un desempeño > 50% en el promedio para la media armónica, lo que indica una proporción eficiente para el número de aciertos en la clasificación.

6.2 Críticas al modelo de aprendizaje automático para la desambiguación léxica

Dentro de las diferentes técnicas para la desambiguación léxica automática, el aprendizaje automático supervisado ha sido ciertamente una de las más utilizadas en diferentes sistemas que realizan tareas de PLN. Específicamente, según Márquez *et al*. (2006), la implementación del modelo bayesiano ingenuo se posiciona como un clasificador simple dentro de toda la gama de sistemas disponibles para el aprendizaje automático supervisado (método de los *k*-vecinos más próximos, árboles de decisión, regresión lineal, máquinas de vectores de soporte, entre otros sistemas referenciados en el capítulo

dos). No obstante, una de las ventajas que presenta para la investigación lingüística es su simplicidad y rapidez, además de la posibilidad de incluir un gran número de atributos o características, entendidas a su vez como palabras de contenido para la captura de la información necesaria durante el proceso de clasificación.

La implementación del algoritmo bayesiano ingenuo que se presenta en este capítulo utiliza un número reducido de características correspondientes a palabras que ocurren dentro de una ventana contextual definida, en la que se encuentra la palabra objetivo para el proceso de desambiguación. Así, se trata de un modelo que selecciona un número restringido de palabras para disminuir el número de características utilizadas, con el objetivo de aumentar el rendimiento del proceso de desambiguación. Este método, entonces, se limita a la información proporcionada por el contexto sintáctico y su relación con el corpus de entrenamiento. A partir de lo anterior, se pueden establecer diferentes críticas, a las que hemos llamado 'inadecuaciones'. Estas no solo son relevantes para discutir el modelo bayesiano aplicado a la desambiguación léxica automática en particular, sino también para establecer una crítica a los métodos supervisados en general:

a. Inadecuación epistemológica: Si bien los modelos para el aprendizaje automático han demostrado ser altamente eficientes en la aplicación de diversas tareas de PLN, particularmente en sistemas expertos, se trata de algoritmos cuyos fundamentos evidencian una mínima comprensión del fenómeno del lenguaje humano. En este sentido, el ámbito de la inteligencia artificial tiene por objetivo reproducir los patrones mediante los que funcionan la mente y el lenguaje. Dada esa definición, los métodos estocásticos no logran simular artificialmente la manifestación de una capacidad cognitiva humana, sino que descansan exclusivamente en la eficiencia de los resultados y en la inferencia estadística.
b. Inadecuación explicativa: Los modelos estocásticos en general y bayesiano ingenuo en particular fundamentan su desempeño en los datos de entrenamiento y en la información que provee el contexto oracional. Sin embargo, el análisis de los resultados, sobre todo para un sistema de desambiguación léxica, no permite acceder a ningún mecanismo que explique el comportamiento de los datos, excepto la posibilidad de añadir más información con el objetivo de provocar resultados diferentes. Este problema también es conocido como la opacidad de los algoritmos en inteligencia artificial, y plantea el problema de que el conocimiento de la manera en la que funcionan los modelos estadísticos y su potencial explicativo es limitado, pero sus resultados son altamente eficientes y, por tanto, validados.

c. Inadecuación lingüística: La única variable que puede provocar cambios en los resultados de un sistema de aprendizaje automático supervisado es el volumen de datos. En este sentido, el funcionamiento de cualquier modelo requerirá de una cantidad alta de datos para poder ser ejecutado con resultados que puedan ser válidos. Esta dependencia del volumen del recurso lingüístico informatizado que se utilice, tanto como corpus de entrenamiento como de prueba, puede resultar inviable en muchos casos.

Capítulo 7
Modelo para la desambiguación léxica automática basado en una medida híbrida

En este capítulo, se presenta la propuesta de un modelo de desambiguación léxica automática basado en una medida híbrida. Esta implementación se fundamenta en la interacción de dos enfoques de exploración taxonómica: distancia entre rutas y contenido de información. Además, se utiliza la base de conocimiento FunGramKB como un inventario de sentidos que permite explotar dicha interacción. A lo largo de este capítulo, en primer lugar, se exponen los fundamentos para la propuesta de una medida híbrida y su posicionamiento en el modelo de desambiguación. En segundo lugar, se describe la naturaleza de la información conceptual disponible en la base de conocimiento FunGramKB, a partir de ejemplos específicos de las unidades léxicas que fueron seleccionadas para los experimentos de aprendizaje automático, junto con sus correlatos conceptuales. Finalmente, se establecen los distintos criterios para la selección de los componentes de la medida híbrida.

7.1 Una medida híbrida: fundamentos de la propuesta

Tanto la exposición inicial acerca del estado de la cuestión para los métodos de desambiguación léxica automática, como el posterior desarrollo de los experimentos de aprendizaje automático y sus correspondientes resultados, han mostrado que en todos los sistemas actuales de desambiguación léxica automática se utilizan, sin excepción, recursos lingüísticos informatizados de diversa naturaleza técnica. No obstante, una pregunta crítica aún permanece sin respuesta en el ejercicio de comparar estos sistemas: si bien los algoritmos de aprendizaje automático más populares, como *Naïve Bayes*, junto con las métricas de similitud semántica, logran resultados de desambiguación léxica similares, ¿por qué no alcanzan una concordancia perfecta o suficiente como para declarar la resolución del problema de la desambiguación léxica automática de manera definitiva? Esto se debe, probablemente, a que todos los métodos computacionales diseñados para tareas de desambiguación automática realizan predicciones basándose en conjuntos de características previamente manipuladas.

Según lo anterior, una de las conclusiones más relevantes que hemos podido establecer hasta este punto de la investigación es que los sistemas de desambiguación léxica automática son altamente dependientes de las fuentes de conocimientos que utilizan, en el caso de los métodos de aprendizaje automático, y de los inventarios de sentidos, en el caso de los métodos de similitud semántica. Los resultados expuestos en los capítulos anteriores, tanto para el corpus SENSEVAL-3 como para el

corpus basado en CODICACH, sugieren que los métodos se adaptan mejor a determinados recursos de información lingüística. En este sentido, la propuesta de una medida híbrida debe ser capaz de explotar el potencial de un método que integre la medida de la distancia entre rutas, por un lado, y la medida del contenido de información, por otro, con una fuente de conocimiento o inventario de sentidos. Esta perspectiva estaría basada en la disponibilidad de recursos léxicos y de conocimiento tanto lingüístico como conceptual particulares, como podría ser FunGramKB, que contiene la información necesaria para eliminar la ambigüedad a partir de sus representaciones conceptuales. Esta aproximación sería menos restrictiva que los enfoques de aprendizaje automático, por ejemplo, pues no dependería de la generalización de patrones para grandes corpus, o de la consistencia que alcancen quienes etiqueten un corpus de entrenamiento.

En cuanto a la fundamentación desde una perspectiva psicolingüística, Moreno-Sandoval (1998), basado en los aportes de Bob & Scha (1996), propone tres argumentos a favor de los métodos basados en datos y su relación con los procesos cognitivos de procesamiento lingüístico:

a. Los hablantes registran frecuencias de uso y diferencias entre frecuencias.
b. Los hablantes prefieren los análisis que ya han experimentado a análisis que tienen que construir por primera vez.
c. La preferencia está influida por la frecuencia de aparición de los análisis, de manera que prefieren los análisis más frecuentes a los menos frecuentes.

En coherencia con lo anterior, durante el desarrollo de esta investigación hemos puesto énfasis en la importancia de los modelos basados en regularidades estadísticas, estos argumentos ponen en evidencia la necesidad de contar con una fuente de conocimiento o inventario de sentidos que sea capaz de interactuar de manera más eficiente con datos textuales. En este sentido, las características que debe tener esta propuesta de medida, en coherencia con las conclusiones derivadas de la fase de experimentación, son las siguientes:

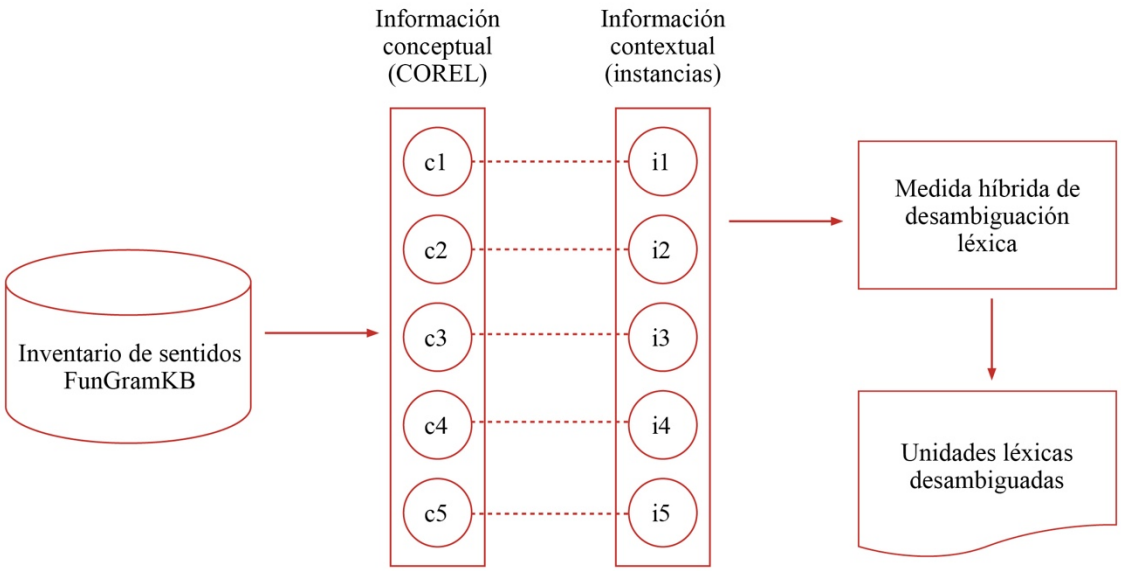
a. Debe ser una medida aplicada a tareas de procesamiento de datos textuales, coherente con la definición de Rada *et al*. (1989), en la que se establece como una función matemática, basada en la medición de la distancia conceptual. Específicamente, se trata de la medición de la distancia geométrica de puntos que representan conceptos o información conceptual.

b. Las propiedades que debe satisfacer una función $f(x, y)$ son las siguientes:
   i. $f(x, x) = 0$; la distancia entre los puntos es cero, solo si los puntos son iguales.
   ii. $f(x, y) = f(y, x)$; la función es simétrica.
   iii. $f(x, y) \geq 0$; la distancia entre los puntos es positiva (también llamada no-negatividad).
   iv. $f(x, y) + f(y, z) \geq f(x, z)$; se aplica la desigualdad triangular para espacios vectoriales.
c. Debe ser evaluable matemáticamente, y devolver un valor cuyo rango se encuentre entre [0, 1], como valor normalizado.
d. Debe facilitar la comparación entre dos conceptos.
e. Debe poder comparar unidades léxicas que pertenezcan a diferentes partes de la oración[54], como sustantivos, verbos y adjetivos, cuyos correlatos conceptuales corresponden a entidades, eventos y cualidades en FunGramKB.

En consecuencia, el modelo de desambiguación propuesto considera, en primer lugar, la base de conocimiento FunGramKB como un inventario de sentidos aplicable a la evaluación de un sistema de desambiguación léxica automática basado en una medida de similitud semántica. En segundo lugar, la desambiguación se basa en una comparación entre la información contextual, que contiene instancias en las que aparece una palabra objetivo, y la información conceptual provista por relaciones taxonómicas en FunGramKB y los postulados de significado que caracterizan cada sentido potencial de la palabra por medio de información lingüística. En tercer lugar, se aplica una medida híbrida de desambiguación, que integra el enfoque basado en la distancia entre rutas, y el enfoque basado en el contenido de información. Finalmente, es posible obtener un resultado numérico, dada una función $f(c_i, c_j)$, que establecerá el puntaje para la similitud de dos palabras objetivo puestas en un contexto oracional específico. Este modelo de desambiguación automática se ilustra como diagrama de flujo en la siguiente figura:

[54] Las etiquetas *part-of-speech* (POS) corresponden una lista de etiquetas de partes del discurso, también llamadas etiquetas de partes de la oración o etiquetas gramaticales, que se utilizan en el ámbito de la lingüística de corpus y el PLN para marcar determinadas categorías gramaticales presentes en un conjunto de datos textuales dentro de un corpus o recurso lingüístico informatizado. Típicamente, a cada componente léxico se le asignará una etiqueta gramatical. Una lista estandarizada de etiquetas POS, junto con otros recursos para realizar etiquetado de categorías morfosintácticas, ha sido elaborada en el marco del Proyecto *Penn Treebank* del Departamento de Computación y Ciencias Informáticas de la Universidad de Pennsylvania. Se trata del desarrollo de un corpus que cuenta con más de 4.5 millones de palabras en inglés norteamericano. Entre 1989 y 1992 este corpus fue anotado con categorías gramaticales. El listado se encuentra disponible en: https://www.ling.upenn.edu/~beatrice/annotation-for-audio-aligned-corpora/list_of_tags.html#pos.

Figura 21. Propuesta de modelo de desambiguación léxica automática.



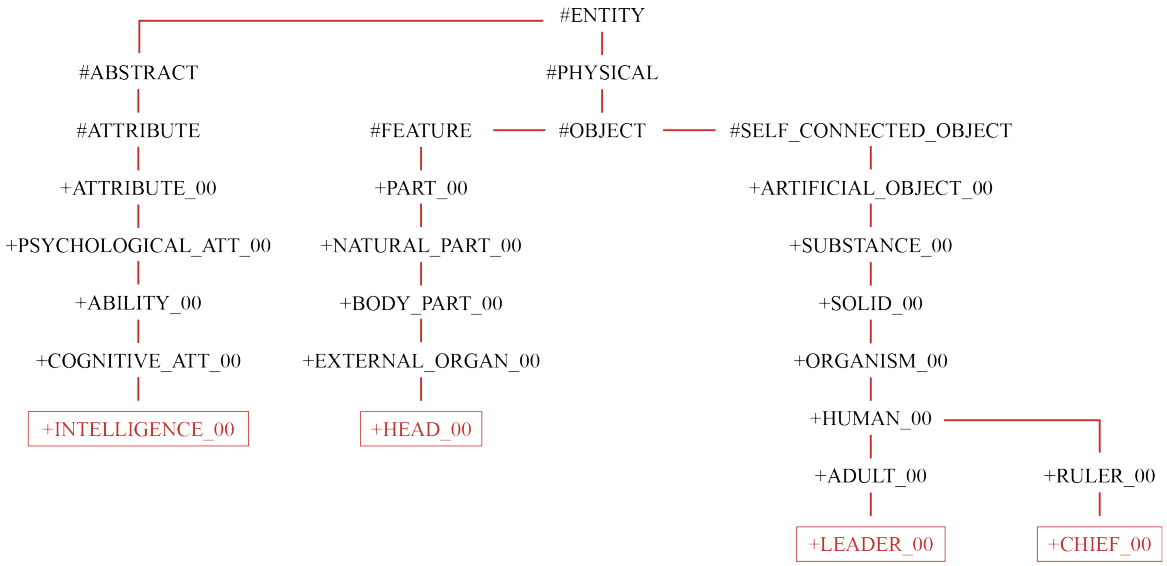
7.2 Relaciones taxonómicas e información conceptual en FunGramKB

La aplicación del modelo de desambiguación léxica automática basado en una medida híbrida para cada una de las unidades léxicas propuestas en el desarrollo experimental de esta investigación depende de la siguiente información acerca de las relaciones taxonómicas y la información conceptual observables en FunGramKB como inventario de sentidos.

7.2.1 Taxonomía y postulados de significado para «cabeza»

En primer lugar, una representación parcial de la subontología #ENTITY para los sentidos de la unidad léxica «cabeza» se expone en la siguiente figura:

Figura 22. Taxonomía de la subontología #ENTITY para los sentidos de «cabeza».



A partir de la figura anterior, se observa una relación semántica bastante estrecha entre los conceptos +LEADER_00 y +CHIEF_00, dado que su hiperónimo común más cercano, y con una profundidad de dos niveles, es +HUMAN_00. Luego esta relación se difumina hacia el hiperónimo común más cercano para +LEADER_00, +CHIEF_00 y +HEAD_00, correspondiente a +OBJECT_00, con una profundidad máxima de siete niveles taxonómicos. Recordemos que, a diferencia de WordNet, la taxonomía de FunGramKB se estructura exclusivamente a través de la relación de subsunción, o relación *IS-A*.

Finalmente, +INTELLIGENCE_00 solamente se relaciona con el resto de los sentidos desde el metaconcepto #ENTITY, lo que lo hace menos similar a ellos en cuanto a su información conceptual. A continuación, se exponen cada uno de los postulados de significado, junto con la información conceptual para los +HEAD_00, +LEADER_00 y +CHIEF_00.

El postulado de significado para concepto (básico)/sentido +HEAD_00 corresponde a:

+(e1: +BE_00 (x1: +HEAD_00)Theme (x2: +EXTERNAL_ORGAN_00)Referent)
+((e2: +BE_02 (x3: 1 +FACE_00)Theme (x4: +FRONT_00)Location)(e3: +BE_02 (x4)Theme (x1)Location)) *((e4: +BE_02 (x5: +HAIR_01)Theme (x6: +TOP_00)Location)(e5: +BE_02 (x6)Theme (x1)Location)(e6: +COMPRISE_00 (x7: +HUMAN_00)Theme (x1)Referent)) *(e7: +BE_02 (x8: 1 +BRAIN_00)Theme (x1)Location (f1: +IN_00)Position) *(e8: +BE_02 (x9: 2 +EAR_00)Theme (x1)Location)

Según la formalización anterior, «cabeza» es un 'tipo de órgano externo del cuerpo de los seres vivos, con una cara en la parte frontal'. Típicamente hay cabello en la parte superior de la cabeza de los humanos; dentro de ella está contenido el cerebro; y presenta orejas. La información conceptual completa contenida en la base de conocimiento se expone en la siguiente figura:

Figura 23. Información conceptual para +HEAD_00 en FunGramKB.

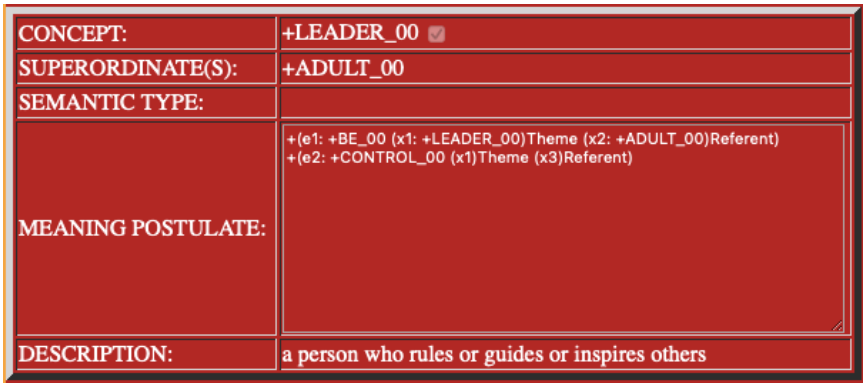
CONCEPT:	+HEAD_00
SUPERORDINATE(S):	+EXTERNAL_ORGAN_00
SEMANTIC TYPE:	+rigid, +identical, +unique
MEANING POSTULATE:	+(e1: +BE_00 (x1: +HEAD_00)Theme (x2: +EXTERNAL_ORGAN_00)Referent) +((e2: +BE_02 (x3: 1 +FACE_00)Theme (x4: +FRONT_00)Location)(e3: +BE_02 (x4)Theme (x1)Location)) *((e4: +BE_02 (x5: +HAIR_01)Theme (x6: +TOP_00)Location)(e5: +BE_02 (x6)Theme (x1)Location)(e6: +COMPRISE_00 (x7: +HUMAN_00)Theme (x1)Referent)) *(e7: +BE_02 (x8: 1 +BRAIN_00)Theme (x1)Location (f1: +IN_00)Position) *(e8: +BE_02 (x9: 2 +EAR_00)Theme (x1)Location)
DESCRIPTION:	the upper or front part of the body in animals; contains the face and brains; "he stuck his head out the window"

Para el concepto (básico)/sentido +LEADER_00, el postulado de significado corresponde a:

+(e1: +BE_00 (x1: +LEADER_00)Theme (x2: +ADULT_00)Referent)
+(e2: +CONTROL_00 (x1)Theme (x3)Referent)

Según la formalización anterior, «cabeza» es un tipo de adulto que ejerce control sobre un referente. Esta caracterización del concepto básico +CONTROL_00 está orientada hacia la 'capacidad de un individuo de inspirar o guiar a otro(s)'. La información conceptual del concepto básico +LEADER_00 es la siguiente:

Figura 24. Información conceptual para +LEADER_00 en FunGramKB.

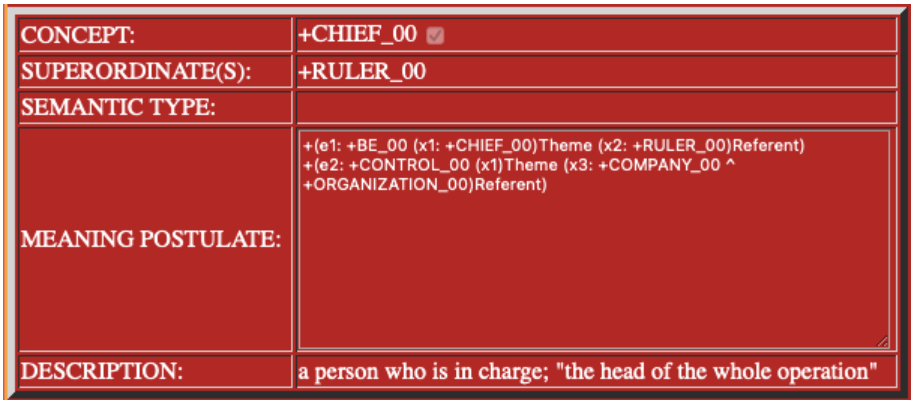


El postulado de significado para el concepto (básico)/sentido +CHIEF_00 es el siguiente:

+(e1: +BE_00 (x1: +CHIEF_00)Theme (x2: +RULER_00)Referent)
+(e2: +CONTROL_00 (x1)Theme (x3: +COMPANY_00 ^ +ORGANIZATION_00)Referent)

A partir del postulado de significado anterior, el concepto básico +CHIEF_00 corresponde a un 'tipo de gobernante (entendido desde un sentido más amplio que el aplicado exclusivamente al concepto de nación), que ejerce control en una compañía u organización'. En cuanto a su descripción, se establece la «cabeza» como una persona que se encuentra a cargo o es responsable de la operación dentro de una organización. A continuación, se presenta la información conceptual correspondiente:

Figura 25. Información conceptual para +CHIEF_00 en FunGramKB.



Para el concepto(básico)/sentido +INTELLIGENCE_00, el postulado de significado es:

+(e1: +BE_00 (x1: +INTELLIGENCE_00)Theme (x2: +COGNITIVE_ATT_00)Referent) *(e2: +THINK_00 (x3)Theme (x4)Referent (f1: x1)Means)

Según la información anterior, el concepto básico +INTELLIGENCE_00 se puede describir como 'un atributo cognitivo, en el que típicamente se realiza la acción de pensar en algo con orientación hacia un objetivo'. Según lo anterior, la «cabeza» tiene que ver con la demostración de la habilidad de pensar en los recursos necesarios para la consecución de un fin. Esta información conceptual se expone en la siguiente figura:

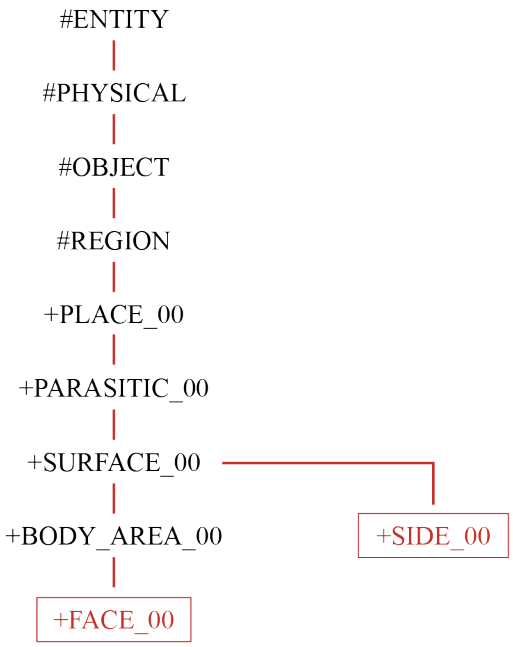
Figura 26. Información conceptual para +INTELLIGENCE_00 en FunGramKB.

CONCEPT:	+INTELLIGENCE_00
SUPERORDINATE(S):	+COGNITIVE_ATT_00
SEMANTIC TYPE:	
MEANING POSTULATE:	+(e1: +BE_00 (x1: +INTELLIGENCE_00)Theme (x2: +COGNITIVE_ATT_00)Referent) *(e2: +THINK_00 (x3)Theme (x4)Referent (f1: x1)Means)
DESCRIPTION:	your ability to think, feel, and imagine things

7.2.2 Taxonomía y postulados de significado para «cara»

En segundo lugar, la siguiente figura muestra una representación parcial de la subontología #ENTITY para los sentidos de la unidad léxica «cara»:

Figura 27. Taxonomía de la subontología #ENTITY para los sentidos de «cara».



En cuanto a los sentidos «cara», se trata de conceptos básicos que presentan una alta relación semántica. Específicamente, la distancia entre ellos es de tres nodos, considerando +SURFACE_00 como el hiperónimo común más cercano. Este comportamiento en la taxonomía está relacionado con la información que heredan desde la información conceptual de una superficie como la parte más externa de un cuerpo. Esta definición, efectivamente, puede aplicarse tanto a la «cara» como parte de la cabeza de un ser vivo, o bien a «cara» como un tipo de superficie para cosas o lugares. En cuanto a los resultados de los experimentos de aprendizaje automático para los sentidos de «cara», fueron más eficientes que para el resto de las unidades léxicas en análisis. A continuación, se exponen cada uno de los postulados de significado, junto con la información conceptual para los conceptos +FACE y +SIDE_00.

El postulado de significado para el concepto(básico)/sentido +FACE_00 es el siguiente:

+(e1: +BE_00 (x1: +FACE_00)Theme (x2: +BODY_AREA_00)Referent)
*(e2: +BE_02 (x3: 2 +CHEEK_00 & 1 +CHIN_00 & 2 +EYE_00 & 1 +NOSE_00 &
1 +FOREHEAD_00)Theme (x1)Location)

Para el caso del concepto básico +FACE_00, se trata de un 'área del cuerpo de los seres vivos, particularmente humanos que, típicamente, presenta mejillas, mentón, ojos, nariz y frente'. A continuación, se presenta la información conceptual completa para +FACE_00:

Figura 28. Información conceptual para +FACE_00 en FunGramKB.

CONCEPT:	+FACE_00
SUPERORDINATE(S):	+BODY_AREA_00
SEMANTIC TYPE:	+rigid, -identical, +dependent, +unique
MEANING POSTULATE:	+(e1: +BE_00 (x1: +FACE_00)Theme (x2: +BODY_AREA_00)Referent) *(e2: +BE_02 (x3: 2 +CHEEK_00 & 1 +CHIN_00 & 2 +EYE_00 & 1 +NOSE_00 & 1 +FOREHEAD_00)Theme (x1)Location)
DESCRIPTION:	the front of the head from the forehead to the chin and ear to ear; "he washed his face"; "I wish I had seen the look on his face when he got the news"

Para el concepto (básico)/sentido +SIDE_00, su postulado de significado corresponde a:

+(e1: +BE_00 (x1: +SIDE_00)Theme (x2: +SURFACE_00)Referent)

Este postulado de significado es bastante abierto, dado que se establece +SIDE_00 como 'un tipo de superficie'. En este sentido, «cara» podría aplicarse tanto a la superficie específica de un lugar, como a la parte diferenciada de un objeto. A continuación, se expone la información conceptual de +SIDE_00 en la base de conocimiento:

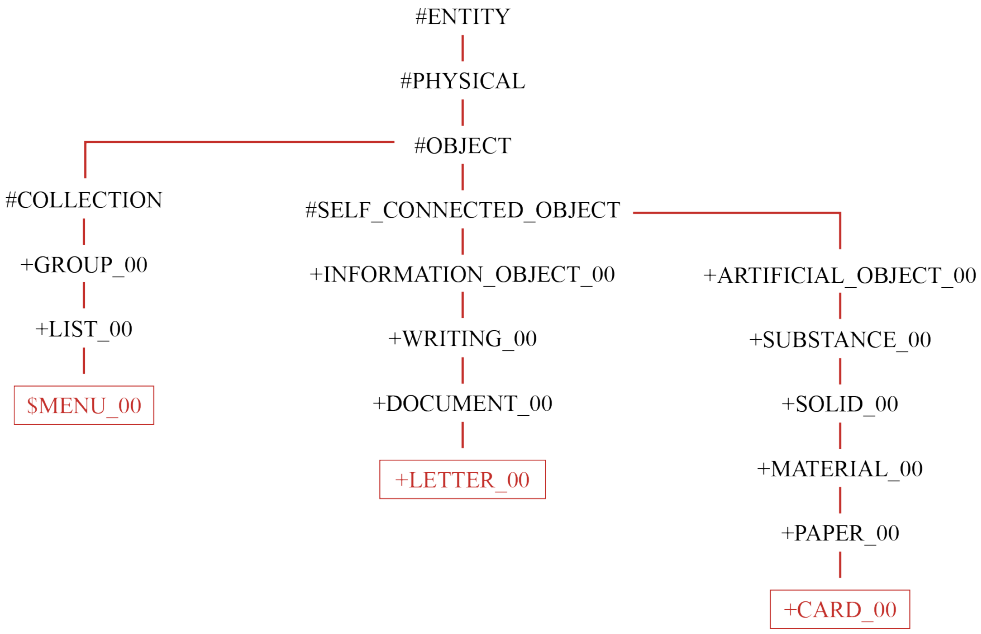
Figura 29. Información conceptual para +SIDE_00 en FunGramKB.

CONCEPT:	+SIDE_00
SUPERORDINATE(S):	+SURFACE_00
SEMANTIC TYPE:	+rigid, -identical, +dependent, ~unique
MEANING POSTULATE:	+(e1: +BE_00 (x1: +SIDE_00)Theme (x2: +SURFACE_00)Referent)
DESCRIPTION:	a surface forming part of the outside of an object; "he examined all sides of the crystal"; "dew dripped from the face of the leaf"

7.2.3 Taxonomía y postulados de significado para «carta»

En tercer lugar, en la siguiente figura se expone una representación parcial de la subontología #ENTITY para los sentidos de la unidad léxica «carta»:

Figura 30. Taxonomía de la subontología #ENTITY para los sentidos de «carta».



En el caso de la unidad léxica «carta», se observan tres conceptos/sentidos disponibles: los conceptos básicos +LETTER_00 y +CARD_00, y el concepto terminal $MENU_00. El hiperónimo común más cercano es el metaconcepto #OBJECT. Esto implica que se trata de conceptos alejados en la taxonomía. Por un lado, dada su transitividad, es posible describir el sentido de $MENU_00 como un tipo de colección, el sentido de +LETTER_00 como un tipo de objeto autoconectado/objeto de información, y el sentido de +CARD_00 como un tipo de objeto artificial. Esto implica una diferencia relevante para la unidad léxica «carta», que puede ser tratada como una lista con información, como un producto de escritura (no material), o bien como un objeto sólido hecho de papel. A continuación, se exponen cada uno de los postulados de significado, junto con la información conceptual para los conceptos +LETTER_00, +CARD_00 y $MENU_00.

El postulado de significado para el concepto(básico)/sentido +LETTER_00 es el siguiente:

> +(e1: +BE_00 (x1: +LETTER_00)Theme (x2: +DOCUMENT_00)Referent)
> +(e2: +WRITE_00 (x3: +HUMAN_00)Theme (x1)Referent) *(e3: +PUT_00 (x3)Agent (x1)Theme (x4)Origin (x5: +ENVELOPE_00)Goal (f1: +IN_00)Position (f2: (e3: +SEND_00 (x3)Agent (x1)Theme (x6)Origin (x7)Goal))Purpose)

En este caso, +LETTER_00 se define como un 'tipo de documento escrito por un humano'. Además, se establece que, típicamente, una carta se introduce dentro de un sobre para ser enviada con un propósito particular, desde un origen hasta una meta. La información conceptual completa para este concepto es la siguiente:

Figura 31. Información conceptual para +LETTER_00 en FunGramKB.

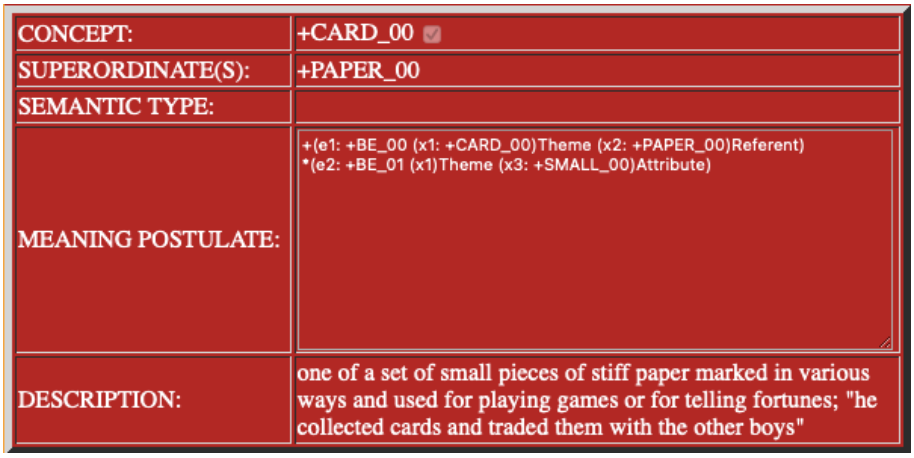
CONCEPT:	+LETTER_00
SUPERORDINATE(S):	+DOCUMENT_00
SEMANTIC TYPE:	
MEANING POSTULATE:	+(e1: +BE_00 (x1: +LETTER_00)Theme (x2: +DOCUMENT_00)Referent) +(e2: +WRITE_00 (x3: +HUMAN_00)Theme (x1)Referent) *(e3: +PUT_00 (x3)Agent (x1)Theme (x4)Origin (x5: +ENVELOPE_00)Goal (f1: +IN_00)Position (f2: (e3: +SEND_00 (x3)Agent (x1)Theme (x6)Origin (x7)Goal))Purpose)
DESCRIPTION:	a written message addressed to a person or organization; "wrote an indignant letter to the editor"

Para el concepto (básico)/sentido +CARD_00, el postulado de significado es:

+(e1: +BE_00 (x1: +CARD_00)Theme (x2: +PAPER_00)Referent)
*(e2: +BE_01 (x1)Theme (x3: +SMALL_00)Attribute)

Según su postulado de significado, +CARD_00 corresponde a un 'tipo de papel que se utiliza para un fin y tiene el atributo típico de ser pequeño'. En cuanto a la descripción provista por la información conceptual que se presenta más adelante, se caracteriza además como una pieza de papel que se utiliza en juegos o adivinaciones. En este caso, el postulado de significado está infraespecificado en la base de conocimiento, en tanto no da cuenta de la complejidad de esta descripción en una lengua natural, a la vez que requiere capturar algunas preferencias de selección que podrían ser relevantes durante el proceso de desambiguación automática.

Figura 32. Información conceptual para +CARD_00 en FunGramKB.



El postulado de significado para el concepto (básico)/sentido $MENU _00 es el siguiente:

+(e1: +BE_00 (x1: $MENU_00)Theme (x2: +LIST_00)Referent)
+(e2: +KNOW_00 (x3: +HUMAN_00)Theme (x4: (e3: +SELL_00 (x5: +RESTAURANT_00)Agent (x6: +FOOD_00)Theme (x5)Origin (x3)Goal))Referent (f1: x1)Instrument)

En cuanto a $MENU_00, se trata de un 'tipo de lista'. En este caso, es interesante que la segunda predicación estricta propone que se trata de un instrumento que contiene un tipo de conocimiento mediante el cual un humano realiza la acción de vender comida en un restaurante. Se trata, dada la

naturaleza de este concepto terminal, de un postulado de significado altamente especificado. A continuación se presenta la información conceptual completa para $MENU_00:

Figura 33. Información conceptual para $MENU_00 en FunGramKB.

CONCEPT:	$MENU_00
SUPERORDINATE(S):	+LIST_00
SEMANTIC TYPE:	-rigid, -identical, ~unique
MEANING POSTULATE:	+(e1: +BE_00 (x1: $MENU_00)Theme (x2: +LIST_00)Referent) +(e2: +KNOW_00 (x3: +HUMAN_00)Theme (x4: (e3: +SELL_00 (x5: +RESTAURANT_00)Agent (x6: +FOOD_00)Theme (x5)Origin (x3)Goal))Referent (f1: x1)Instrument)
DESCRIPTION:	a list of dishes available at a restaurant; "the menu was in French"

7.3 Componentes de la medida híbrida

El modelo de desambiguación léxica automática integra como componente central una medida híbrida, cuyo objetivo es asignar un valor numérico para la función f (c_i, c_j), equivalente a la interacción entre la información contextual y aquella que se puede extraer desde la exploración taxonómica, considerando a la base de conocimiento FunGramKB como inventario de sentidos.

El primer componente que considerar es la medida de la distancia entre rutas (*path base*, en adelante PB), basado en la propuesta de Leacock & Chodorow (1998). En la exploración taxonómica es relevante integrar la perspectiva de la relación semántica basada en PB, entendida como una función entre la longitud de la ruta que relaciona dos conceptos (*IS-A*), y la posición de esos conceptos en la taxonomía. Esta integración será particularmente importante considerando la jerarquía conceptual específica para cada una de las subontologías de la base de conocimiento. Este algoritmo es el siguiente:

(1)

$$PB_{Leacock\ \&\ Chodorow}(c_i, c_j) = -\log\left(\frac{len(c_i, c_j)}{2 \times depth\_max}\right)$$

correspondiente a un valor que establece la cercanía de los nodos entre dos conceptos en la taxonomía. Para incluirlo en la medida, es necesario determinar dos valores:

i. *len*(c_i, c_j), equivale a la longitud del camino más corto entre los conceptos c_i y c_j.
ii. *depth_max* = 14, equivale al valor máximo de profundidad de la taxonomía. En el caso de la subontología #ENTITY, este número es un valor constante para los niveles taxonómicos en FunGramKB.

El segundo componente para la medida híbrida corresponde a la medida del contenido de información. Uno de los problemas más relevantes que considera esta propuesta es que las métricas basadas en IC, dado que se establecen a partir WordNet como inventario de sentidos, no consideran únicamente la relación taxonómica *IS-A*. Por tanto, en esos casos no siempre se podrá identificar un ancestro común más cercano. En efecto, una de las críticas que plantean Gangemi *et al*. (2003), es que los enlaces léxicos de WordNet son de naturaleza heterogénea, aunque la mayoría de ellos corresponden a la subsunción habitual *IS-A*. Por el contrario, en las subontologías de FunGramKB, dado que utiliza exclusivamente la relación taxonómica *IS-A*, no siempre se podrá establecer un ancestro común más cercano que sea relevante semánticamente. Esto se debe a que se entiende la ponderación de la relación semántica entre dos unidades léxicas como un proceso de comprensión, en el que interviene un proceso inferencial. Sin embargo, aunque esta relación de subsunción provoca que no se puedan conectar directamente las cualidades con las entidades o con los eventos, sí se pueden relacionar a través de los postulados de significado.

Específicamente, la máquina debe encontrar una inferencia necesaria o vínculo perdido[55]. Según Croft & Cruse (2008), se puede afirmar que "[…] para que *X es un tipo de Y* sea aceptable, basta con que los *Xs* prototípicos se hallen dentro de los límites categoriales de *Y*. Esto supone asumir que *X* e *Y* son elementos léxicos y que denotan categorías conceptuales fijas" (p. 194). Adicionalmente, se establece que un enunciado hiponímico se juzgaría como aceptable si existen conceptualizaciones de *X* y de *Y* fácilmente aceptables y basadas en interpretaciones contextuales no anómalas. Esta inferencia necesaria es fundamental para poner en relación las dos unidades léxicas en análisis, puesto que establece un tránsito desde la palabra explícita (oída o leída), hasta el conocimiento tácito. Por ejemplo, si se consideran las unidades léxicas «manzana» y «pera», el vínculo perdido es que 'manzanas y peras pertenecen a la categoría fruta'; es decir, una relación *IS-A*.

[55] También llamado *enlace omitido* desde el marco conceptual de la semántica cognitiva.

Por otra parte, si se consideran las unidades léxicas «automóvil» y «viaje», el vínculo perdido corresponde al hecho de que 'el automóvil es uno de los medios de transporte que se puede utilizar para viajar'; es decir, se establece una relación semántica más cercana a un tipo de clase (llamada también taxonimia). Así, los vínculos perdidos expresan información que forma parte del conocimiento conceptual representado en marcos, entendidos como:

> "[…] estructuras más o menos variables que se asocian de forma estable con elementos léxicos, lo que permite la existencia de conceptualizaciones con límites variables, presumiblemente en términos de idoneidad de ajuste requerido entre la realidad que se percibe y los distintos aspectos del marco" (Croft & Cruse, 2008: 133).

Según lo anterior, en principio se consideró la propuesta adaptada de Seco *et al*. (2004), en la que se establece un algoritmo capaz de determinar un valor para el IC en tanto que, cuanto más abajo se encuentre un concepto en la taxonomía, mayor será el valor para el IC con respecto a sus conceptos superordinados en la ruta *IS-A*. Este valor demostraría un potencial favorable para establecer el vínculo perdido entre dos conceptos:

(2)

$$IC_{Seco\ et\ al.}(c) = 1 - \frac{\log\left(hypo(c) + 1\right)}{\log\left(\mathrm{C}\right)}$$

donde *hypo*(*c*) establece el número de hiperónimos de *c*, y *C* corresponde al número total de conceptos que se encuentran en la dimensión metacognitiva a la que pertenece *c*. No obstante, en coherencia con el potencial del contenido conceptual alojado en la ontología de la base de conocimiento FunGramKB, y sus diferencias con la taxonomía de WordNet, en la que por cierto se basan propuestas como la de Seco *et al*. (2004), es necesario seleccionar un algoritmo para establecer el valor de IC, que considere que la estructura taxonómica está organizada y regulada por determinados principios a partir de los cuales los hipónimos proporcionan información conceptual más específica que los hiperónimos, al mismo tiempo que es posible explotar las relaciones de herencia e inferencia para la información conceptual. Según esto, el trabajo de Zhou *et al*. (2008), que toma como antecedente la propuesta Seco *et al*. (2004), propone un algoritmo que supera el método establecido como convencional para obtener el IC, en el que se establece una función entre el conocimiento de la estructura taxonómica y las instancias derivadas de un corpus. El nuevo algoritmo considera la información que heredan los

hipónimos, junto con la profundidad de la estructura taxonómica. Así, el IC se establece como un valor más dependiente del inventario de sentidos que del contexto oracional en el que aparece la palabra objetivo. Los resultados de esta medida presentan, entonces, la ventaja de que no está basados en el análisis de las características del corpus, y por tanto es posible evitar el problema de la escasez o la imprecisión de los datos textuales disponibles, presente en modelos basados en información contextual, tal y como se pudo evidenciar a partir de los experimentos de aprendizaje automáticos basados en SENSEVAL-3. Luego, la medida para establecer el IC, adaptada de Zhou *et al*. (2008), es la siguiente:

(3)

$$IC_{Zhou\ et\ al.}(c) = k(1 - \frac{\log(hypo(c) + 1)}{\log\ (node\_max)}) + (1 - k)(\frac{\log(depth(c) + 1)}{\log\ (depth\_max)})$$

que equivale al logaritmo del número de hipónimos de *c*; esto es, de cada uno de los conceptos en análisis, dividido por el número máximo de conceptos en la taxonomía. Además, se considera el logaritmo de la profundidad del concepto c en relación con el nodo raíz, dividido por el logaritmo de la profundidad máxima de la taxonomía, como valor constante. En cuanto a los componentes, se integra el parámetro *k*, de adaptación manual, mediante el cual es posible ajustar el peso para cada uno de los dos componentes de la ecuación.

El tercer componente es el valor del hiperónimo común más cercano (*lowest common subsumer*, en adelante LCS) para dos conceptos, que se establecerá como una manera eficiente de incorporar la información heredada desde un concepto básico que contiene la información conceptual para la palabra objetivo. En definitiva, la ventaja de una propuesta híbrida basada en las funciones matemáticas propuestas por Leacock & Chodorow (1998) en (1), y por Zhou *et al*. (2008) en (3), es que facilitan la interacción de dos componentes para el cálculo de un valor numérico establecido como una función entre PB e IC. Estos componentes se resumen en la siguiente tabla:

Tabla 36. Componentes de la medida híbrida de similitud semántica.

	Componente	Algoritmo	Fuente
1	Distancia entre rutas	$PB(c_i, c_j) = -\log\left(\frac{len(c_i, c_j)}{2 \times depth\_max}\right)$	Leacock y Chodorow (1998).
2	Contenido de información	$IC(c) = k(1 - \frac{\log(hypo(c)+1)}{\log(node\_max)}) + (1-k)(\frac{\log(depth(c)+1)}{\log(depth\_max)})$	Zhou *et al*. (2008)

A partir de lo anterior, el primer componente, correspondiente a $PB(c_i, c_j)$, se puede reducir de la siguiente manera:

(4)

$$PB(c_i, c_j) = -log\left(\frac{len(c_i, c_j)}{2 \times depth\_max}\right) = log(2 \times depth\_max) - \log\left(len(c_i, c_j)\right)$$

Luego, para el segundo componente, se establece el valor de la variable k como una constante donde $k = 0{,}5$, de tal manera que no tendría ningún impacto en la medida. Se reduce de esta manera:

(5)

$$IC(c) = (1 - \frac{\log(hypo(c)+1)}{\log(node\_max)}) + \frac{\log(depth(c)+1)}{\log(depth\_max)}$$

Para integrar cada uno de estos componentes de la medida es necesario normalizar sus valores. Según Han *et al*. (2012), la normalización de valores tiene el objetivo de entregar a todas las variables de la medida el mismo peso estadístico. De esta manera, se espera evitar la dependencia derivada de la elección de una u otra unidad de medida, lo que provocaría un mayor peso o efecto de una determinada variable. En términos matemáticos, el procedimiento consiste en la transformación de los datos para que se encuentren dentro de un rango común, en este caso [0, 1].

Según lo anterior, para el caso de la medida híbrida propuesta, y siguiendo la metodología de Periñán-Pascual (2015), utilizaremos una técnica de normalización de valores en la que se establece una transformación lineal de los datos originales, pero conservando sus relaciones de magnitud. Esta

normalización, entonces, se expresará como *norm*(*PB*(c_i, c_j)) para el componente de distancia entre rutas, y *norm*(*IC*(*c*)), para el componente de contenido de información, como se expone en los algoritmos (6) y (7) respectivamente:

(6)

$$norm\left(PB\left(c_i, c_j\right)\right) = 1 - \frac{1}{log_2(2 + PB\left(c_i, c_j\right))}$$

(7)

$$norm\left(IC(c)\right) = 1 - \frac{1}{log_2(2 + \left(IC(c)\right))}$$

Posteriormente, se integran a la medida los parámetros de optimización α y β, con el objetivo de evaluar el impacto que tendría cada componente. De acuerdo con Daelemans & Hoste (2002), la optimización de parámetros se define como el proceso mediante el cual se ajustan los valores de los atributos de un algoritmo en un sistema de PLN, con el objetivo de configurar las condiciones óptimas para el desempeño en una tarea particular. En el caso de nuestra propuesta de medida híbrida, se integrarían dos parámetros de optimización, donde $\alpha + \beta = 1$, de tal manera que el valor resultante todavía sería un valor normalizado. Finalmente, la propuesta de similitud semántica basada en una medida híbrida se expone en (8):

(8)

$$SIM_{híbrida}\left(c_i, c_j\right) = \alpha\ (norm\left(PB\left(c_i, c_j\right)\right)) + \beta\ ((norm\left(IC(c_i)\right) + (norm\left(IC\left(c_j\right)\right))$$

En definitiva, se trata de una propuesta de medida híbrida que está basada en la medición del IC como una dimensión que permite una comprensión eficiente de la semántica de un concepto y su relación con su ámbito metaconceptual, a partir de la estimación de su grado ya sea de generalidad o de concreción. Esta aproximación considera que la única relación taxonómica válida en FunGramKB, y que en definitiva permite establecer el valor de PB, es la subsunción, a diferencia de otros recursos similares, como los que fueron descritos en los capítulos anteriores.

Capítulo 8
Evaluación de la medida híbrida

A continuación, se presenta la evaluación del modelo de desambiguación léxica automática propuesto, basado en una medida híbrida fundamentada en la interacción de los enfoques de exploración taxonómica de distancia entre rutas y contenido de información. En primer lugar, se exponen las variables necesarias para el cálculo de la medida híbrida, cuyos valores han sido extraídos desde la base de conocimiento FunGramKB: profundidad, hipónimos y contenido de información. En segundo lugar, se proponen tres casos de evaluación junto con sus respectivos resultados, considerando todos los pasos necesarios para la aplicación de $SIM_{\text{híbrida}}(c_i, c_j)$.

8.1 Valores de profundidad, hipónimos y contenido de información

En coherencia con la propuesta descrita en el capítulo siete, la medida híbrida para la desambiguación léxica automática requiere establecer, en primer lugar, valores de profundidad (*depth*), hipónimos (*hypo*) y contenido de información (IC) para cada uno de los nodos que constituyen las unidades tanto léxicas como conceptuales en análisis, extraídas desde la ontología de la base de conocimiento FunGramKB. Esta información sistematiza los valores referidos para la ruta taxonómica que se organiza desde el concepto hipónimo más profundo para cada unidad conceptual, hasta el nodo raíz #ENTITY, y que conforman su campo semántico en la subontología. Con el objetivo de determinar los valores de profundidad o número de nodos de la taxonomía de #ENTITY, junto con el número de hipónimos para cada uno de los conceptos, se utilizó una tabla con la información taxonómica correspondiente a la subontología de entidades de la base de conocimiento. Esta tabla propone una organización de la información basada en las relaciones *IS-A* en dos columnas. La columna uno con hipónimos (*concept1*), y la columna dos con hiperónimos (*concept2*). Un ejemplo parcial de lo anterior se expone en la siguiente tabla:

Tabla 37. Relaciones *IS-A* para los conceptos en la taxonomía de #ENTITY[56].

concept1	concept2
+CELL_00	+PLANT_PART_00
+HUMAN_00	+ORGANISM_00
+BEAR_00	+MAMMAL_00

A partir de esta tabla, se ejecutó un comando SQL para extraer, en una nueva tabla, la profundidad y los hipónimos de cada uno de los conceptos, mediante una selección de datos de manera recursiva que pudiera derivar la estructura taxonómica a partir el nodo raíz #ENTITY:

```
WITH RECURSIVE new_table_name(concept1, level) AS (
        VALUES ('concepto', 0)
        UNION ALL
        SELECT isa_clean.concept1, new_table_name.level +1
        FROM isa_clean JOIN new_table_name ON isa_clean.concept2 = new_table_name.concept1
        ORDER BY new_table_name.level +1 DESC
)
SELECT concept1 AS hypo, level FROM new_table_name;
```

A partir de esta información, se ha aplicado la siguiente ecuación con el objetivo de establecer los valores de IC para cada uno de los nodos:

(1)

$$IC(c) = (1 - \frac{\log(hypo(c) + 1)}{\log(node\_max)}) + \frac{\log(depth(c) + 1)}{\log(depth\_max)}$$

Además, para el cálculo de todas las ecuaciones para las que se consideren estas variables de aquí en adelante, se han establecido los siguientes valores constantes para la subontología #ENTITY de la base de conocimiento FunGramKB:

i. $node\_max_{\#ENTITY} = 1344$

ii. $depth\_max_{\#ENTITY} = 14$

[56] Por razones de confidencialidad y restricciones de uso tanto de la base de conocimiento FunGramKB como de su arquitectura cognitiva, no es posible publicar íntegramente la taxonomía *IS-A* que se presenta de manera parcial en esta tabla. No obstante, sí fue posible utilizar la información completa para los fines investigativos de esta tesis, con la autorización del director del proyecto FunGramKB. Para más información, se recomienda revisar los términos de uso de FunGramKB en: http://www.fungramkb.com/terms.aspx.

Según lo anterior, la taxonomía correspondiente a un extracto de la subontología #ENTITY para la unidad léxica «cabeza» hace referencia a la figura 22, «cara» a la figura 23, y finalmente «carta» a la figura 24, todas ellas presentadas en el capítulo siete. De esta manera, los valores para la unidad léxica «cabeza», y sus correspondientes sentidos capturados en los postulados de significado de los conceptos +INTELLIGENCE_00, +HEAD_00, +LEADER_00, y +CHIEF_00, son los siguientes:

Tabla 38. Valores *depth*, *hypo* e IC para los nodos correspondientes a los sentidos de «cabeza».

Nodos	depth	hypo	IC
#ENTITY	1	1343	0,263
#ABSTRACT	2	234	0,658
#PHYSICAL	2	1107	0,443
#ATTRIBUTE	3	75	0,924
#FEATURE	3	157	0,822
#OBJECT	3	964	0,571
#SELF_CONNECTED_OBJECT	3	623	0,632
+ATTRIBUTE_00	4	74	1,010
+PART_00	4	95	0,976
+ARTIFICIAL_OBJECT_00	4	585	0,725
+PSYCHOLOGICAL_ATT_00	5	33	1,189
+NATURAL_PART_00	5	81	1,067
+SUBSTANCE_00	5	583	0,795
+ABILITY_00	6	7	1,449
+BODY_PART_00	6	63	1,160
+SOLID_00	6	546	0,862
+COGNITIVE_ATT_00	7	5	1,539
+EXTERNAL_ORGAN_00	7	41	1,269
+ORGANISM_00	7	180	1,066
+INTELLIGENCE_00	**8**	**1**	**1,736**
+HEAD_00	**8**	**0**	**1,833**
+HUMAN_00	8	124	1,162
+ADULT_00	9	50	1,327
+RULER_00	9	7	1,584
+LEADER_00	**10**	**7**	**1,620**
+CHIEF_00	**10**	**3**	**1,716**

De acuerdo con la tabla 38, la unidad léxica «cabeza» y sus sentidos presentan los siguientes valores de contenido de información: $IC(cabeza_{\text{intelligence}}) = 1{,}736$; $IC(cabeza_{\text{head}}) = 1{,}833$; $IC(cabeza_{\text{leader}}) = 1{,}620$; y $IC(cabeza_{\text{chief}}) = 1{,}716$.

Los valores para la unidad léxica «cara», y sus correspondientes sentidos capturados en los postulados de significado de los conceptos +FACE_00 y +SIDE_00, son los siguientes:

Tabla 39. Valores *depth*, *hypo* e IC para los nodos correspondientes a los sentidos de «cara».

Nodos	depth	hypo	IC
#ENTITY	1	1343	0,263
#PHYSICAL	2	1107	0,443
#OBJECT	3	964	0,571
#REGION	4	133	0,930
+PLACE_00	5	132	1,000
+PARASITIC_PLACE_00	6	60	1,167
+SURFACE_00	7	36	1,287
+BODY_AREA_00	8	25	1,380
+SIDE_00	**8**	**7**	**1,544**
+FACE_00	**9**	**1**	**1,776**

La tabla 39 muestra que la unidad léxica «cara» y sus sentidos presentan valores de IC equivalentes a $IC(cara_{\text{side}}) = 1{,}544$ y $IC(cara_{\text{face}}) = 1{,}776$.

Finalmente, los valores para la unidad léxica «carta», junto con sus correspondientes sentidos capturados en los postulados de significado de los conceptos +LETTER_00, +CARD_00, y $MENU_00 se muestran a continuación:

Tabla 40. Valores *depth*, *hypo* e IC para los nodos correspondientes a los sentidos de «carta».

Nodos	depth	hypo	IC
#ENTITY	1	1343	0,263
#PHYSICAL	2	1107	0,443
#OBJECT	3	964	0,571
#COLLECTION	4	51	1,061
#SELF_CONNECTED_OBJECT	4	623	0,716
+ARTIFICIAL_OBJECT_00	4	585	0,725
+GROUP_00	5	50	1,133
+INFORMATION_OBJECT_00	5	35	1,181
+SUBSTANCE_00	5	583	0,795
+LIST_00	6	4	1,514
+WRITING_00	6	24	1,290
+SOLID_00	6	546	0,862
$MENU_00	**7**	**0**	**1,788**
+DOCUMENT_00	7	6	1,518
+LETTER_00	**8**	**0**	**1,833**
+MATERIAL_00	8	33	1,343
+PAPER_00	9	2	1,720
+CARD_00	**10**	**0**	**1,909**

Según la tabla 40, la unidad léxica «carta» presenta los siguientes valores para el contenido de información relativo a sus respectivos sentidos: $IC(carta_{letter})$ = 1,833; $IC(carta_{card})$ = 1,909; y $IC(carta_{menu})$ = 1,788.

8.2 Evaluación de casos de similitud semántica basados en la medida híbrida

A continuación, exponemos tres casos para el cálculo de la similitud semántica, correspondientes a cada una de las tres unidades léxicas en análisis. Estos casos permitirán evaluar el potencial de la medida híbrida de similitud. El criterio para la selección, tanto de unidades léxicas como de sus correlatos conceptuales en la base de conocimiento FunGramKB, fue la verificación de que su contenido conceptual estuviese incluido en la ruta taxonómica para cada uno de los sentidos de «cabeza», «cara» y «carta», respectivamente, y que además se pudiese clasificar objetivamente a partir de un sentido en particular.

Así, en primer lugar, para «cabeza» seleccionamos como hiperónimo la unidad léxica «órgano», cuyo significado está capturado por el postulado de significado correspondiente al concepto básico +EXTERNAL_ORGAN_00, definido a su vez como 'un órgano que está situado sobre o cerca de la superficie del cuerpo':

+(e1: +BE_00 (x1: +EXTERNAL_ORGAN_00)Theme (x2: +BODY_PART_00)Referent)
*(e2: +BE_02 (x1)Theme (x3: +BODY_00)Location (f1: +OUT_00)Position)

En segundo lugar, para el caso de «cara» hemos escogido como hiperónimo la unidad léxica «superficie», que se define como 'el límite exterior bidimensional extendido de un objeto tridimensional'. El significado correspondiente está capturado por el postulado de significado del concepto básico +SURFACE_00:

+(e1: +BE_00 (x1: +SURFACE_00)Theme (x2: +PARASITIC_PLACE_00)Referent)
+(e2: +BE_02 (x1)Theme (x3: +PLACE_00 ^ +NATURAL_OBJECT_00 ^
+ARTIFICIAL_OBJECT_00)Location)

En tercer lugar, para el caso de «carta» seleccionamos como hiperónimo la unidad léxica «documento», cuyo significado está capturado por el postulado de significado del concepto +DOCUMENT_00, correspondiente a un 'tipo de escrito que contiene información':

+(e1: +BE_00 (x1: +DOCUMENT_00)Theme (x2: +WRITING_00)Referent)
+(e2: +CONTAIN_00 (x3: +INFORMATION_00)Theme (x1)Location)

Finalmente, el objetivo de la evaluación es determinar el valor más alto de similitud semántica entre los pares (*cabeza*$_{\text{head}}$, *órgano*), (*cabeza*$_{\text{intelligence}}$, *órgano*), (*cabeza*$_{\text{chief}}$, *órgano*), y (*cabeza*$_{\text{leader}}$, *órgano*), correspondientes a los sentidos de la unidad léxica «cabeza»; entre los pares (*cara*$_{\text{face}}$, *superficie*) y (*cara*$_{\text{side}}$, *superficie*), que constituyen los sentidos de la unidad léxica «cara»; y entre los pares (*carta*$_{\text{letter}}$, *documento*), (*carta*$_{\text{card}}$, *documento*), y (*carta*$_{\text{menu}}$, *documento*), que conforman los sentidos de la unidad léxica «carta», aplicando la medida de similitud híbrida que hemos propuesto.

Finalmente, para cada una de las evaluaciones de los casos en análisis, se incorporaron los siguientes valores de α y β como parámetros de optimización:

Tabla 41. Valores para los parámetros de optimización α y β.

Combinación	Valores α	Valores β
1	0	1
2	0,1	0,9
3	0,2	0,8
4	0,3	0,7
5	0,4	0,6
6	0,5	0,5
7	0,6	0,4
8	0,7	0,3
9	0,8	0,2
10	0,9	0,1
11	1	0

Como se expuso en el capítulo siete, los parámetros de optimización se definen como una selección de valores que mejoran el desempeño de un sistema de PLN en un aspecto específico. Cada parámetro representa un determinado peso de una variable en la medida, lo que en definitiva provocará un sesgo tanto en la interpretación de los datos como en el funcionamiento de la medida. Este sesgo permitirá determinar cuáles son las variables que más influyen en la eficiencia del sistema de desambiguación. En este sentido, los trabajos de Hoste *et al.* (2002), Van Gompel & Van Den Bosch (2013), y Lu *et al*. (2019), entre otros, han demostrado que la inclusión de parámetros de optimización para modificar las variables disponibles en algoritmos de aprendizaje automático produce mejoras significativas en la precisión de las clasificaciones que puede realizar un sistema de desambiguación léxica automática.

En el caso de nuestra medida híbrida, los parámetros de optimización le otorgarán, para cada una de las combinaciones expuestas en la tabla anterior, un peso específico ya sea al valor del contenido de información o al valor de la distancia entre rutas. Esto conducirá a determinar los valores a partir de los cuales la medida obtiene resultados más eficientes.

8.2.1 Evaluación de similitud semántica para el caso «órgano» y los sentidos de «cabeza»

El objetivo de este caso es determinar la similitud semántica más alta entre los pares (*cabeza*$_{\text{head}}$, *órgano*), (*cabeza*$_{\text{intelligence}}$, *órgano*), (*cabeza*$_{\text{chief}}$, *órgano*), y (*cabeza*$_{\text{leader}}$, *órgano*), mediante la comparación de los valores de $SIM_{\text{híbrida}}(c_i, c_j)$, considerando las variables que constituyen tanto la distancia entre rutas como el contenido de información para los conceptos +HEAD_00, +INTELLIGENCE_00, +CHIEF_00, +LEADER_00, y +EXTERNAL_ORGAN_00. Según lo anterior, la hipótesis que se intentará comprobar es que, dado el campo semántico de la unidad léxica «cabeza», el sentido (*cabeza*$_{\text{head}}$) sería el que alcance el puntaje de similitud más alto tras la aplicación del modelo de desambiguación léxica automática basado en nuestra medida híbrida.

Para el cálculo de $SIM_{\text{híbrida}}(c_i, c_j)$, primero se deben establecer los valores normalizados de $IC(c_i)$, según la ecuación (2), para los sentidos (*cabeza*$_{\text{head}}$) en la ecuación (3), (*cabeza*$_{\text{intelligence}}$) en la ecuación (4), (*cabeza*$_{\text{chief}}$) en la ecuación (5), (*cabeza*$_{\text{leader}}$) en la ecuación (6), y (órgano) en la ecuación (7):

(2)

$$norm\big(IC(c_i)\big) = 1 - \frac{1}{log_2(2 + IC(c_i))}$$

(3)

$$norm\big(IC(cabeza_{head})\big) = 1 - \frac{1}{log_2(2 + 1{,}833)} = 0{,}484$$

(4)

$$norm\left(IC\big(cabeza_{intelligence}\big)\right) = 1 - \frac{1}{log_2(2 + 1{,}736)} = 0{,}474$$

(5)

$$norm\left(IC\big(cabeza_{chief}\big)\right) = 1 - \frac{1}{log_2(2 + 1{,}716)} = 0,$$

(6)

$$norm\big(IC(cabeza_{leader})\big) = 1 - \frac{1}{log_2(2 + 1{,}620)} = 0{,}461$$

(7)

$$norm\big(IC(organo)\big) = 1 - \frac{1}{log_2(2 + 1{,}269)} = 0{,}415$$

Según los resultados anteriores, los valores normalizados de IC para la unidad léxica «cabeza», considerando sus cuatro sentidos disponibles, y «órgano», son los siguientes:

Tabla 42. Valores normalizados de IC para «cabeza» y «órgano».

Variable	Valor
norm(*IC*(*cabeza*$_{\text{head}}$))	0,484
norm(*IC*(*cabeza*$_{\text{intelligence}}$))	0,474
norm(*IC*(*cabeza*$_{\text{chief}}$))	0,472
norm(*IC*(*cabeza*$_{\text{leader}}$))	0,461
norm(*IC*(*órgano*))	0,415

En segundo lugar, se deben establecer los valores de distancia entre rutas *PB*(c_i, c_j), según la ecuación (8), aplicada a los pares (*cabeza*$_{\text{head}}$, *órgano*) en la ecuación (9), (*cabeza*$_{\text{intelligence}}$, *órgano*) en la ecuación (10), (*cabeza*$_{\text{chief}}$, *órgano*) en la ecuación (11), y (*cabeza*$_{\text{leader}}$, *órgano*) en la ecuación (12):

(8)

$$PB\big(c_i, c_j\big) = log(2 \times depth\_max) - log\Big(len\big(c_i, c_j\big)\Big)$$

(9)

$$PB(cabeza_{head}, organo) = log(2 \times 14) - log(1) = 1{,}447$$

(10)

$$PB\big(cabeza_{intelligence}, organo\big) = log(2 \times 14) - log(14) = 0{,}301$$

(11)

$$PB\big(cabeza_{chief}, organo\big) = log(2 \times 14) - log(13) = 0{,}333$$

(12)

$$PB(cabeza_{leader}, organo) = log(2 \times 14) - log(13) = 0{,}333$$

En tercer lugar, se realizó el proceso de normalización de *PB*(c_i, c_j), mediante la aplicación de la ecuación (13). Este proceso tuvo como resultado la obtención de los siguientes valores normalizados para la distancia entre rutas: *norm*(*PB*(*cabeza*$_{\text{head}}$, *órgano*)) en la ecuación (14), *norm*(*PB*(*cabeza*$_{\text{intelligence}}$, *órgano*)) en la ecuación (15), *norm*(*PB*(*cabeza*$_{\text{chief}}$, *órgano*)) en la ecuación (16), y *norm*(*PB*(*cabeza*$_{\text{leader}}$, *órgano*)) en la ecuación (17):

(13)

$$norm\left(PB\left(c_i, c_j\right)\right) = 1 - \frac{1}{log_2(2 + \ PB\left(c_i, c_j\right))}$$

(14)

$$norm\left(PB(cabeza_{head}, organo)\right) = 1 - \frac{1}{log_2(2 + 1{,}447)} = 0{,}440$$

(15)

$$norm\left(PB\left(cabeza_{intelligence}, organo\right)\right) = 1 - \frac{1}{log_2(2 + 0{,}301)} = 0{,}168$$

(16)

$$norm\left(PB\left(cabeza_{chief}, organo\right)\right) = 1 - \frac{1}{log_2(2 + 0{,}333)} = 0{,}182$$

(17)

$$norm\left(PB(cabeza_{leader}, organo)\right) = 1 - \frac{1}{log_2(2 + 0{,}333)} = 0{,}182$$

Según los resultados anteriores, los valores normalizados de PB para los pares correspondientes a la combinatoria de los sentidos de «cabeza» y la unidad léxica «órgano» se muestran a continuación:

Tabla 43. Valores normalizados de PB para la combinatoria de «cabeza» y «órgano».

Variable	Valor
$norm(PB(cabeza_{\text{head}}, órgano))$	0,440
$norm(PB(cabeza_{\text{intelligence}}, órgano))$	0,168
$norm(PB(cabeza_{\text{chief}}, órgano))$	0,182
$norm(PB(cabeza_{\text{leader}}, órgano))$	0,182

Finalmente, se aplicó para los pares en análisis la medida $SIM_{\text{híbrida}}(c_i, c_j)$, expresada en la ecuación (18), considerando cada una de las combinaciones para los parámetros de optimización (α y β):

(18)

$$SIM_{híbrida}(c_i, c_j) = \alpha\ (norm\left(PB(c_i, c_j)\right)) + \beta\ ((norm(IC(c_i)) + \ (norm\left(IC(c_j)\right))$$

Los resultados se exponen en la siguiente tabla:

Tabla 44. Resultados de aplicación de la medida híbrida para $SIM_{\text{híbrida}}(cabeza, órgano)$.

Comb.	$(cabeza_{\text{head}}, órgano)$	$(cabeza_{\text{intelligence}}, órgano)$	$(cabeza_{\text{chief}}, órgano)$	$(cabeza_{\text{leader}}, órgano)$
1	0,899	0,889	0,887	0,876
2	0,853	0,848	0,817	0,807
3	0,807	0,806	0,746	0,737
4	0,761	0,765	0,676	0,668
5	0,715	0,723	0,605	0,598
6	0,670	0,682	0,535	0,529
7	0,714	0,729	0,553	0,547
8	0,578	0,599	0,394	0,390
9	0,532	0,557	0,323	0,321
10	0,486	0,516	0,253	0,251
11	0,440	0,474	0,182	0,182
Prom.	**0,678**	**0,690**	**0,543**	**0,537**

Según los resultados expuestos en la tabla 44, el puntaje de $SIM_{\text{híbrida}}(cabeza_{\text{head}}, órgano)$ promedio [$\overline{X}$ = 0,678; DE = 0,152] presenta sus valores más altos en las combinaciones uno [= 0,899], dos [= 0,853], y tres [= 0,807]. Estas combinaciones incorporan un peso nulo o mínimo para la variable de la distancia entre rutas. Por otra parte, el puntaje promedio para $SIM_{\text{híbrida}}(cabeza_{\text{intelligence}}, órgano)$

se convirtió en el resultado más alto para esta evaluación [$\overline{X} = 0{,}690$; $DE = 0{,}138$] y presentó, en la mayoría de sus combinaciones, valores en los que el peso de la distancia entre rutas fue el más preponderante.

8.2.2 Evaluación de similitud semántica para el caso «superficie» y los sentidos de «cara»

El objetivo de este caso en evaluación es determinar la similitud semántica más alta entre los pares ($cara_{face}$, *superficie*) y ($cara_{side}$, *superficie*), mediante la comparación de los valores de $SIM_{híbrida}(c_i, c_j)$, considerando las variables que constituyen tanto la distancia entre rutas como el contenido de información para los conceptos +FACE_00, +SIDE_00, y +SURFACE_00. Así, la hipótesis a demostrar será que, dado el campo semántico de la unidad léxica «cara», el sentido ($cara_{side}$), alcanzará el puntaje de similitud más alto tras la aplicación del modelo de desambiguación léxica automática basado en nuestra medida híbrida.

Para el cálculo de $SIM_{híbrida}(c_i, c_j)$, se establecieron los valores normalizados de $IC(c_i)$, según la ecuación (2), para los sentidos: ($cara_{face}$), ($cara_{side}$), y (*superficie*):

Tabla 45. Valores normalizados de IC para «cara» y «superficie».

Variable	Valor
$norm(IC(cara_{face}))$	0,478
$norm(IC(cara_{side}))$	0,452
$norm(IC(superficie))$	0,418

Posteriormente, se calcularon, mediante la aplicación de las ecuaciones (8) y (13), los valores normalizados de PB para los pares correspondientes a la combinatoria de los sentidos de «cara» y la unidad léxica «superficie», como se expone en la siguiente tabla:

Tabla 46. Valores normalizados de PB para la combinatoria de «cara» y «superficie».

Variable	Valor
$norm\ (PB(cara_{face}, superficie))$	0,395
$norm(PB(cara_{side}, superficie))$	0,440

Finalmente, se aplicó la medida $SIM_{\text{híbrida}}(c_i, c_j)$, expresada en la ecuación (18), para cada uno de los pares en análisis, considerando las combinaciones para los parámetros de optimización α y β presentadas en la tabla 41. Lo anterior se presenta en la siguiente tabla:

Tabla 47. Resultados de aplicación de la medida híbrida para $SIM_{\text{híbrida}}(cara, superficie)$.

Comb.	(*cara*$_{\text{face}}$, *superficie*)	(*cara*$_{\text{side}}$, *superficie*)
1	0,896	0,870
2	0,846	0,827
3	0,796	0,784
4	0,746	0,741
5	0,696	0,698
6	0,646	0,655
7	0,685	0,699
8	0,545	0,569
9	0,495	0,526
10	0,445	0,483
11	0,395	0,440
Prom.	0,654	0,663

La tabla 47 muestra que los valores de similitud semántica para (*cara*$_{\text{side}}$, *superficie*) presentan un promedio de $[\overline{X} = 0{,}663; DE = 0{,}141]$. El grupo de las combinaciones cinco a once fue consistentemente más alto, alcanzando su valor máximo en la combinación cinco [= 0,698], correspondiente a los valores de $[\alpha = 0{,}4]$ para la distancia entre rutas, y $[\beta = 0{,}6]$ para el contenido de información. No obstante, se observa que los valores seis a once corresponden a la asignación del parámetro de optimización cuyo mayor peso estuvo en la variable de la distancia entre rutas. Este caso en particular resultó problemático debido a que tanto el concepto +FACE_00 como el concepto +SIDE_00 presentan al concepto +SURFACE_00 como hiperónimo común más cercano. Esto provoca que, al asignarle mayor peso al parámetro de PB siempre se seleccione el sentido +FACE_00, y al establecer un mayor peso al parámetro de IC, se seleccione el sentido +SIDE_00. Según lo anterior, se presenta la dificultad de que el puntaje final de la medida se ve afectado significativamente por la presencia o ausencia del hiperónimo común más cercano, dado que esta información se obtiene desde el conocimiento intrínseco en FunGramKB.

8.2.3 Evaluación de similitud semántica para el caso «documento» y los sentidos de «carta»

El objetivo de este ejemplo es determinar la similitud semántica más alta entre los pares (*carta*$_{\text{letter}}$, *documento*), (*carta*$_{\text{card}}$, *documento*) y (*carta*$_{\text{menu}}$, *documento*), mediante la comparación de los valores de $SIM_{\text{híbrida}}(c_i, c_j)$, considerando las variables que constituyen tanto la distancia entre rutas como el contenido de información para los conceptos +LETTER_00, +CARD_00, $MENU_00, y +DOCUMENT_00. La hipótesis que se intentará comprobar en este ejemplo es que, dado el campo semántico de la unidad léxica «carta» en su sentido (*carta*$_{\text{letter}}$), será este último el sentido que alcance el puntaje de similitud más alto tras la aplicación del modelo de desambiguación automática basado en nuestra medida híbrida.

Para el cálculo de la similitud basada en la medida híbrida propuesta, en primer lugar, se deben establecer los valores normalizados de IC según la ecuación (2) para los sentidos (*carta*$_{\text{letter}}$), (*carta*$_{\text{card}}$), (*carta*$_{\text{menu}}$) y (*documento*). Estos resultados se reportan en la tabla 48:

Tabla 48. Valores normalizados de IC para «carta» y «documento».

Variable	Valor
norm(*IC*(*carta*$_{\text{letter}}$))	0,484
norm(*IC*(*carta*$_{\text{card}}$))	0,492
norm(*IC*(*carta*$_{\text{menu}}$))	0,480
norm(*IC*(*documento*))	0,449

En segundo lugar, se establecieron los valores de distancia entre rutas $PB(c_i, c_j)$, según las ecuaciones (8) y (13), aplicadas para los pares (*carta*$_{\text{letter}}$, *documento*), (*carta*$_{\text{card}}$, *documento*) y (*carta*$_{\text{menu}}$, *documento*). Lo anterior se expone en la siguiente tabla:

Tabla 49. Valores normalizados de PB para la combinatoria de «carta» y «documento».

Variable	Valor
norm(*PB*(*carta*$_{\text{letter}}$, *documento*))	0,440
norm(*PB*(*carta*$_{\text{card}}$, *documento*))	0,241
norm(*PB*(*carta*$_{\text{menu}}$, *documento*))	0,258

Finalmente, para cada uno de los pares en análisis se aplicó la medida $SIM_{\text{híbrida}}(c_i, c_j)$, presentada en la ecuación (18), considerando las combinaciones para los parámetros de optimización α y β presentadas en la tabla 41. Eso se expone en la tabla 50:

Tabla 50. Resultados de aplicación de la medida híbrida para $SIM_{\text{híbrida}}$(*carta, documento*).

Comb.	(*carta*$_{\text{letter}}$, *documento*)	(*carta*$_{\text{card}}$, *documento*)	(*carta*$_{\text{menu}}$, *documento*)
1	0,933	0,941	0,929
2	0,884	0,871	0,862
3	0,834	0,801	0,795
4	0,785	0,731	0,728
5	0,736	0,661	0,661
6	0,687	0,591	0,594
7	0,731	0,615	0,619
8	0,588	0,451	0,459
9	0,539	0,381	0,392
10	0,489	0,311	0,325
11	0,440	0,241	0,258
Prom.	0,695	0,600	0,602

En cuanto a los resultados de la tabla 50, se observa que el valor promedio más alto para los valores de similitud corresponde a (*carta*$_{\text{letter}}$, *documento*), con un [$\overline{X} = 0{,}695$; $DE = 0{,}163$]. En este caso, los valores de similitud para cada combinación fueron consistentemente más altos desde la combinación dos hasta la once, alcanzando su resultado más eficiente en la combinación dos [= 0,884], correspondiente a los valores de [$\alpha = 0{,}1$] para la distancia entre rutas, y [$\beta = 0{,}9$] para el contenido de información. Cabe señalar que los valores más altos se encontraron precisamente en las combinaciones dos a cinco, que presentaban un mayor peso en el parámetro correspondiente al contenido de información.

8.3 Resultados para la evaluación del desempeño de la medida híbrida

Como se estableció en el capítulo anterior, la propuesta de medida híbrida para el cálculo de la similitud semántica $SIM_{\text{híbrida}}(c_i, c_j)$ considera las relaciones de subsunción en la taxonomía de la base de conocimiento FunGramKB para establecer el valor de la distancia entre rutas y el contenido de información. Se trata de una medida que, si bien se fundamenta en la complementariedad de los valores de IC y PB, establece predominantemente el valor de IC como una dimensión que permite una comprensión eficiente de la semántica de un concepto y la relación con su ámbito metaconceptual, a

partir de la estimación de su grado ya sea de generalidad o de concreción. En cuanto a los resultados, los valores más altos para los casos de las unidades en análisis en los que se alcanza la desambiguación según los sentidos esperados para cada una de las hipótesis propuestas, basados en la aplicación de la medida híbrida, corresponden a aquellos en los que se evidenció un mayor peso en el parámetro de optimización correspondiente a la variable IC. En este sentido, y desde un punto de vista lingüístico, un puntaje de IC alto se puede interpretar como el hecho de que un concepto representa un significado conceptual altamente específico cuando ocurre en un texto. Esta variable, al ser complementada con el valor de PB como la medición de la especificidad de un concepto en cuanto a su significado, es capaz de describir de manera eficiente la naturaleza de la similitud entre dos conceptos. En efecto, la organización taxonómica de FunGramKB representa las unidades conceptuales como un sistema de relaciones que es consistente con la manera en la que los hablantes organizan su lexicón mental; es decir, es capaz de incorporar el conocimiento del mundo al conocimiento léxico. Esto es coherente con la preponderancia del valor del contenido de información en los resultados, dado que la similitud semántica como método aplicado a la desambiguación léxica se basa en la información léxico-semántica y en la relación taxonómica entre los sentidos disponibles de las unidades léxicas.

Según lo anterior, los resultados permiten determinar que las combinatorias más eficientes para la medida híbrida de similitud se encuentran, típicamente, en el intervalo de los valores uno a cinco, en los que el valor de PB en α aumenta progresivamente desde 0, mientras que el valor de IC en β disminuye desde 1. Específicamente, es la combinatoria tres, correspondiente a los valores de [$\alpha = 0{,}2$] para la distancia entre rutas, y [$\beta = 0{,}8$] para el contenido de información, la que obtiene un desempeño eficiente y más consistente para los casos en evaluación. Lo anterior se muestra en la siguiente tabla para los pares (*cabeza*$_{\text{head}}$, *órgano*), (*cara*$_{\text{side}}$, *superficie*) y (*carta*$_{\text{letter}}$, *documento*):

Tabla 51. Resultados de aplicación de $SIM_{\text{híbrida}}(c_i, c_j)$ para la combinación tres.

Comb. Tres	(*cabeza*$_{\text{head}}$, *órgano*)	(*cara*$_{\text{side}}$, *superficie*)	(*carta*$_{\text{letter}}$, *documento*)
[$\alpha = 0{,}2$; $\beta = 0{,}8$]	0,807	0,784	0,834

De acuerdo con estos resultados, la versión consolidada para la medida híbrida de desambiguación léxica automática, luego de la selección y aplicación de los parámetros de optimización, corresponde a la siguiente ecuación:

(18)

$$SIM_{híbrida}(c_i, c_j) = 0{,}2(norm\left(PB(c_i, c_j)\right)) + \ \ 0{,}8((norm(IC(c_i)) + (norm\left(IC(c_j)\right))$$

Finalmente, la medida híbrida propuesta ha mostrado ser eficiente desde le punto de vista de los métodos de similitud semántica, considerando el conocimiento lingüístico que provee la organización taxonómica de la base de conocimiento FunGramKB.

Capítulo 9
Conclusiones y futuras investigaciones

El objetivo general que hemos planteado para esta investigación es desarrollar un modelo más robusto de medida para la similitud y relación semántica que los disponibles actualmente para la desambiguación léxica automática, aplicado al PLN. Este desarrollo, en términos específicos, implicó el diseño de un modelo formal para la desambiguación léxica que puede ser aplicado a una herramienta de PLN. Así, en esta tesis hemos propuesto una medida híbrida para la evaluación de la similitud semántica, basada en la interacción de dos enfoques de exploración taxonómica: distancia entre rutas (Leacock y Chodorow, 1998) y contenido de información (Zhou *et al*., 2008), considerando la arquitectura cognitiva de la base de conocimiento FunGramKB (Periñán-Pascual & Mairal-Usón, 2010a) como un inventario de sentidos que permite explotar dicha interacción.

Como punto de partida, en el capítulo dos hemos revisado las diferencias entre las distintas descripciones del fenómeno de la ambigüedad léxica que ha propuesto la teoría lingüística, por un lado, y la aproximación operacional que se ha establecido desde la perspectiva informática, por otro. En términos lingüísticos, hemos definido la ambigüedad léxica como la asociación entre un ítem léxico determinado en un texto o discurso y un sentido que pueda ser distinguible de otros significados potencialmente atribuibles a ese ítem léxico, mientras que, desde el punto de vista computacional, la ambigüedad léxica se reduce a los fenómenos de la homonimia y la polisemia. Si bien esta última definición no corrige ni niega la aproximación lingüística, se trata de una explicación general, libre de modelo, y que permite su aplicación experimental en sistemas de PLN. Según lo anterior, hemos definido la desambiguación léxica como un procedimiento computacional de asignación de significado a una unidad léxica, que está basado en el contexto en el que esa unidad ocurre. Esta aproximación ha resultado coherente con la premisa de que un significado adecuado para un determinado ítem léxico será seleccionado a partir de un inventario de sentidos, que definen a su vez el rango de posibilidades para esa unidad léxica en un contexto particular (Patwardhan *et al*., 2003; Nevzorova *et al*., 2015).

No obstante, es relevante considerar que la descripción del fenómeno de la ambigüedad léxica implica la interacción de distintos niveles de análisis lingüístico en el proceso de desambiguación: léxico, sintáctico, semántico y/o fonológico. Según esto, y desde una perspectiva formal, el proceso en el que el hablante es capaz de determinar el sentido específico de una unidad léxica en un contexto

oracional particular puede tratarse como una integración modular de información, en la que el procesamiento de un nivel afectará el procesamiento de otros niveles adyacentes.

Uno de los objetivos específicos más relevantes que hemos abordado durante el desarrollo de la investigación ha sido compilar un corpus en español de Chile que integre instancias auténticas de ambigüedad léxica. A partir de este objetivo, el diseño del corpus en análisis se llevó a cabo considerando específicamente los estándares de calidad, representatividad y recuperabilidad propuestos por Dash (2010). Así, todas las instancias fueron obtenidas a partir de muestras de escritura auténticas provenientes de medios de prensa escrita chilenos correspondientes a una submuestra extraída desde el *corpus* CODICACH (Sadowsky, 2006), sin ninguna intervención que pueda ser considerada como una circunstancia de producción artificial, y resguardando una variación mínima entre los valores para los indicadores de densidad léxica y proporción de palabras por oración.

En cuanto a los métodos para la desambiguación léxica automática, se revisaron los de relación semántica (Lesk, 1986; 1987; Cantos-Gómez, 1996; Banerjee & Pedersen, 2002), de similitud semántica (Wu & Palmer, 1994; Leacock & Chodorow, 1998; Resnik, 1995; Jiang & Conrath, 1997; Lin, 1998), y basados en conocimiento contextual (Gale *et al.*, 1992; Manning & Schütze, 1999; Fulmari & Chaldak, 2014; Carpuat & Wu; 2005; Popescu & Hristea, 2010; Ustalov *et al*., 2018). Estas clasificaciones corresponden al tipo de conocimiento que cada una utiliza como recurso lingüístico informatizado desde el que se extraen los casos de ambigüedad, y la manera en la que se relacionan sus variables. Si bien hemos observado que en la actualidad se utilizan predominantemente los métodos basados en conocimiento contextual, ya sean de aprendizaje automático supervisado como de aprendizaje automático no supervisado, hemos optado por desarrollar un modelo basado en la exploración taxonómica como fuente intrínseca de conocimiento. Así, luego de realizar los experimentos de base expuestos en el capítulo seis, hemos podido establecer tres críticas a la aplicación del algoritmo bayesiano ingenuo como método para la desambiguación léxica automática basado en conocimiento contextual.

En primer lugar, como inadecuación epistemológica, hemos observado que los algoritmos de aprendizaje automático no logran simular artificialmente la manifestación de la capacidad lingüística humana, sino que descansan exclusivamente en la eficiencia de los resultados y en la inferencia estadística. En segundo lugar, como inadecuación explicativa, hemos demostrado el problema de la opacidad de los algoritmos de inteligencia artificial, en tanto que los modelos estocásticos no permiten acceder a ningún mecanismo que explique el comportamiento de los datos, excepto la posibilidad de

añadir más información con el objetivo de provocar resultados diferentes. En tercer lugar, y con mayor relevancia, como inadecuación lingüística hemos observado que la única variable que puede provocar cambios en los resultados de un sistema de aprendizaje automático supervisado es el volumen de datos.

Una de las conclusiones más relevantes que hemos podido establecer hasta este punto de la investigación es que los sistemas de desambiguación léxica automática son altamente dependientes de las fuentes de conocimientos que utilizan, en el caso de los métodos de aprendizaje automático, y de los inventarios de sentidos, en el caso de los métodos de similitud semántica. Esta afirmación ha puesto en evidencia la necesidad de contar con una fuente de conocimiento o inventario de sentidos que sea capaz de interactuar de manera más eficiente con datos textuales. A partir de lo anterior, en el capítulo siete hemos abordado el objetivo específico de representar formalmente un procedimiento computacional para poder resolver la desambiguación léxica automática, basado en la propuesta de una medida de desambiguación híbrida. Nuestro modelo de desambiguación léxica automática, cuyo componente central es esta medida híbrida, ha cumplido con tres propiedades fundamentales: (1) considera la base de conocimiento FunGramKB como inventario de sentidos; (2) se basa en una comparación entre la información contextual, que contiene instancias en las que aparece una palabra objetivo, junto con la información conceptual provista por relaciones taxonómicas en FunGramKB y los postulados de significado que caracterizan cada sentido potencial de la palabra por medio de información lingüística; y (3) integra el enfoque basado en la distancia entre rutas, y el enfoque basado en el contenido de información, a partir de los que es posible obtener un resultado numérico, dada una función f (c_i, c_j), que establece el puntaje para la similitud de dos palabras objetivo puestas en un contexto oracional específico.

Finalmente, en el capítulo ocho hemos realizado la evaluación del modelo de desambiguación léxica automática, basado en la medida híbrida propuesta, para la unidad léxica «cabeza» y sus sentidos capturados en los postulados de significado de los conceptos +INTELLIGENCE_00, +HEAD_00, +LEADER_00, y +CHIEF_00; la unidad léxica «cara» y sus correspondientes sentidos capturados en los postulados de significado de los conceptos +FACE_00 y +SIDE_00; y la unidad léxica «carta», junto con sus sentidos capturados en los postulados de significado de los conceptos +LETTER_00, +CARD_00, y $MENU_00. Además, se integraron parámetros de optimización (Daelemans & Hoste, 2002) con el objetivo de configurar las condiciones óptimas para el desempeño de la tarea de desambiguación considerando el peso específico de cada una de las variables. En el caso

de nuestra propuesta de medida híbrida, se integraron dos parámetros de optimización, donde $\alpha + \beta = 1$. Luego de este proceso, y considerando los resultados, la combinatoria correspondiente a los valores de [$\alpha = 0{,}2$] para la distancia entre rutas, y [$\beta = 0{,}8$] para el contenido de información, es la que obtuvo un desempeño eficiente y más consistente para los casos en análisis. Con estas definiciones, la versión consolidada de la medida híbrida es la siguiente:

(1)

$$SIM_{híbrida}(c_i, c_j) = 0{,}2(norm\left(PB(c_i, c_j)\right)) + \ 0{,}8((norm(IC(c_i)) + (norm\left(IC(c_j)\right))$$

donde la similitud semántica para c_i y c_j será igual a la suma entre el parámetro 0,2 multiplicado por el valor normalizado de la distancia entre rutas del par (c_i, c_j), y el parámetro 0,8 multiplicado por el valor normalizado del contenido de información de c_i más el valor normalizado del contenido de información de c_j. Esta medida híbrida ha mostrado ser eficiente desde el punto de vista de los métodos de similitud semántica, considerando el conocimiento lingüístico que provee la organización taxonómica de la base de conocimiento FunGramKB.

Además de los objetivos general y específicos antes descritos, y las fortalezas presentadas de esta tesis, hemos detectado algunos aspectos mejorables en futuras investigaciones. En primer lugar, las tres colecciones de documentos que conforman el corpus incluyeron 120 instancias. Cada una de estas instancias, a su vez, incorporaron las palabras objetivo en análisis junto con sus correspondientes contextos oracionales. De acuerdo con el planteamiento de Dash (2010), esta dimensión del corpus no logra cumplir con el estándar de tamaño. Si bien se ha podido satisfacer la necesidad de contar con un recurso informatizado desarrollado a partir de los estándares de la lingüística de corpus, y no desde un criterio metodológico basado exclusivamente en el muestreo estadístico, este no presenta un número de instancias que puedan establecerse como una muestra representativa y sincrónica del español de Chile. Este aspecto metodológico es altamente relevante para futuras investigaciones en el ámbito del PLN, dada la necesidad de superar la conceptualización del corpus como una bolsa de palabras, hacia la estandarización de recursos lingüísticos generalistas o diversificados. En segundo lugar, no ha sido posible, debido al problema de la explosión combinatoria, exponer resultados de aplicación para $SIM_{híbrida}(c_i, c_j)$, que pudieran ser comparables en términos de precisión y cobertura con todas las unidades léxicas presentes en cada una de las instancias en análisis. La complejidad de este procedimiento implicaría identificar automáticamente, en cada corpus, las palabras de contenido

para cada una de las 120 instancias, luego determinar los valores de similitud para las palabras de contenido y todos los sentidos disponibles de cada unidad léxica, para posteriormente establecer el puntaje de similitud total por cada instancia. No obstante, dados los alentadores resultados de la implementación de la medida híbrida, este procedimiento puede ser automatizado mediante la utilización de un lenguaje de programación que pueda implementar la medida en un conjunto de datos textuales, considerando FunGramKB como fuente de conocimiento.

Finalmente, el modelo que hemos presentado ha puesto en evidencia, al igual que aquellos expuestos como antecedente teórico, que una medida basada en la información léxico-semántica y en relaciones taxonómica disponibles, puede funcionar de manera altamente eficiente cuando se proporciona una fuente de conocimiento apropiada y coherente con criterios de análisis lingüísticos. En definitiva, esta tesis contribuye en el diseño de un modelo coherente tanto lingüística como matemáticamente, que es capaz de determinar el valor numérico que representa un puntaje de similitud semántica para la comparación entre la información tanto contextual como conceptual entre dos unidades léxicas, considerando la caracterización de cada sentido potencial de la palabra en análisis por medio de información lingüística.

Referencias bibliográficas

AITCHINSON, Jean, 1987: *Words in the Mind: an introduction to the mental lexicon*, Londres: Blackwell.

ALLEN, James, 1995: *Natural Language Understanding*, Redwood City: The Benjamin Cummings Publishing Company.

AMSLER, Robert, 1982: "Computational Lexicology: A Research Program"*,* comunicación presentada en The National Computer Conference (AFIPS).

AMSLER, Robert, y John WHIM, 1979: "Development of a Computational Methodology for Deriving Natural language Semantic Structures via Analysis of Machine-Readable Dictionaries"*, NSF Technical Report MCS 77-01315.*

AUNG, Nyein Thwet, Khin Mar SOE, y Ni Lar THEIN, 2011: "A Word Sense Disambiguation System Using Naïve Bayesian Algorithm for Myanmar Language", *International Journal of Scientific & Engineering Research*, volumen 2, número 9, 1-7.

AUSSENAC-GILLES, Nathalie, y Fabien GANDON, 2013: "From the knowledge acquisition bottleneck to the knowledge acquisition overflow: A brief French history of knowledge acquisition", *International Journal of Human-Computer Studies*, volumen 71, número 2, 157-165.

BAKER, Collin, Charles FILLMORE, y John LOWE, 1998: "The Berkeley FramNet Project", *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, volumen 1, 86-90.

BANERJEE, Satanjeev, y Ted PEDERSEN, 2002: "An adapted Lesk algorithm for word sense disambiguation using WordNet", comunicación presentada en The Third International Conference on Intelligent Text Processing and Computational Linguistics.

BANERJEE, Satanjeev, y Ted PEDERSEN, 2003: "Extended gloss overlaps as a measure of semantic relatedness", comunicación presentada en The 18th International Joint Conference on Artificial Intelligence (IJCAI).

BAUER, Laurie, 2004: *English Word-formation*, Cambridge: Cambridge Textbooks in Linguistics.

BERGEN, Benjamin, y Nancy CHANG, 2003: "Embodied Construction Grammar in Simulation-Based Language Understanding", en Jan-Ola ÖSTMAN and Mirjam FRIED (eds.), 2003: *Construction Grammar(s): Cognitive and Cross-Language Dimensions*, Amsterdam: John Benjamins.

BISHOP, Christopher, 2006: *Pattern Recognition and Machine Learning*, Nueva York: Springer.

BOD, Rens, y Remko SCHA, 1996: "Data-Oriented Language Processing: An Overview", *ILLC Technical Report LP-96-13*, Universidad de Amsterdam.

BONIN, Patrick, 2004: *Mental lexicon: Some words to talk about words*, Nueva York: Nova Science Publishers.

BOUILLON, Pierrete, y Federica BUSA, 2001: "Qualia and the Structuring of Verb Meaning", en *The Language of Word Meaning*, Cambridge: Cambridge University Press, 149-167.

BROWN, Phillip Shigeo, 2006: *A Small-scale exploration into the relationship between word-association and learner`s lexical development*. Tesis de maestría, Universidad de Birmingham.

BROWN, Peter, Jennifer LAI y Robert MERCER, 1991: "Aligning sentences in parallel corpora", comunicación presentada en The 29th Annual Meeting of the Association for Computational Linguistics.

BUCHANAN, Bruce, y David WILKINS, 1993: *Readings in Knowledge Acquisition and Learning: Automating the Construction and Improvement of Expert Systems*, San Mateo CA: Morgan Kaufmann.

CANTOS-GÓMEZ, Pascual, 1996: *Lexical ambiguity, dictionaries and corpora*, Murcia: Servicio de Publicaciones Universidad de Murcia.

CARPUAT, Marine, y Dekai WU, 2005: "Evaluating the Word Sense Disambiguation Performance of Statistical Machine Translation", comunicación presentada en The Second International Joint Conference on Natural Language Processing (IJCNLP).

CHIERNIA, Gennaro, y Sally MCCONNEL-GINET, 1990: *Meaning and Grammar, An Introduction to Semantics*, Cambridge: The MIT Press.

CHIFU, Adrian-Gabriel, Florentina HRISTEA, Josiane MOTHE, y Marius POPESCU, 2014: "Word sense discrimination in information retrieval: a spectral clustering-based approach", *Information Processing & Management*, volumen 51, número 2, 16-31.

CHOI, Rene, Aaron COYNER, Jayasheree KALPATHY-CRAMER, Michael CHIANG, y Peter CAMPBELL, 2020: "Introduction to Machine Learning, Neural Networks, and Deep Learning", *Translational Vision Science & Technology*, volumen 9, número 2, artículo 14, 1-12.

CHOMSKY, Noam, 1964: *Current Issues in Linguistic Theory*, New York: Moutin & Co.

CHOMSKY, Noam, 1978 [1957]: *Estructuras Sintácticas*, México DF: Siglo XXI.

CHOMSKY, Noam, 1988: *Language and Problems of Knowledge*, Cambridge: MIT Press.

CHOUEKA, Yaakov, y Serge LUSIGNAN, 1985: "Disambiguation by short contexts", *Computer and the Humanities*, número 19, 147-157.

CHOWDHURY, Gobinda, 2005: "Natural Language Processing", *Annual Review of Information Science and Technology*, volumen 37, número 1, 51-89.

CLIMENT, Salvador, 2000: "Individuación e Información parte-todo, Representación para el Procesamiento Computacional del Lenguaje", *Estudios de Lingüística Española (ELiEs)*, volumen 8.

COLLINS, Allan, y Elizabeth LOFTUS, 1975: "A spreading-activation theory of semantic processing", *Psychological Review*, volumen 82, número 6, 407-428.

COTTRELL, Garrison, 1984: "A model of lexical access of ambiguous words", comunicación presentada en The Association for the Advancement of Artificial Intelligence.

COWIE, Jim, Joe GUTHRIE, y Louise GUTHRIE, 1992: "Lexical disambiguation using simulated annealing", comunicación presentada en The 14th International Conference on Computational Linguistics (COLING-92).

CROFT, William, y David Alan CRUSE, 2008: *Lingüística Cognitiva*, Madrid: AKAL.

CRUSE, David Allan, 1986: *Lexical semantics*, Cambridge: Cambridge University Press.

CUNNINGHAM, Padraig, y Sarah Jane DELANY, 2007: "k-Nearest neighbour classifiers", Technical Report UCD-CSI-2007-4, 1-17.

CURTIS, Jon, John CABRAL, y David BAXTER, 2006: "On the Application of the Cyc Ontology to Word Sense Disambiguation", *Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference*, 11-13.

DAELEMANS, Walter, y Véronique HOSTE, 2002: "Evaluation of Machine Learning Methods for Natural Language Processing Tasks", *Proceeding of the Third International Conference on Language Resources and Evaluation*, *The University of Las Palmas de Gran Canaria*, disponible en http://www.lrec-conf.org/proceedings/lrec2002/pdf/94.pdf.

DASH, Niladri, 2010: "Corpus Linguistics: A General Introduction", comunicación presentada en The Workshop on Corpus Normalization (LDCIL).

DEL TESO, Enrique, 2002: *Compendio y Ejercicios de Semántica I*, Madrid: Arcolibros.

DEEPU, S., Raj PETHURU, y S. RAJARAAJESWARI, 2016: "A Framework for Text Analytics using the Bag of Words (BoW) Model for Prediction", *International Journal of Advanced Networking*

& Applications (IJANA), 1st International Conference on Innovations in Computing & Networking (ICICN16), 320-323.

DIEN, Tran Thanh, Bui Huu LOC, y Nguyen THAI-NGHE, 2019: "Article classification using natural language processing and machine learning", comunicación presentada en *International Conference on Advanced Computing and Applications (ACOMP)*, 78-84.

DUCROT, Oswald, y Tzetan TODOROV, 1972: *Diccionario Enciclopédico de las Ciencias del Lenguaje*, Buenos Aires: Siglo XXI.

EBERHARDT, Frederik, y David DANKS, 2011: "Confirmation in the Cognitive Sciences: The Problematic Case of Bayesian Models", *Minds and Machines*, volumen 21, número 3, 389-410.

ELMAN, Jeffrey, y James L. MCCLELLAND, 1984: "Speech perception as a cognitive process: The interactive activation model", en Norman LASS (ed.), 1984: *Speech and language: Advances in Basic Research & Practice*, Nueva York: Academic Press.

ESCANDELL, María Victoria, 2004: *Fundamentos de Semántica composicional*, Barcelona: Ariel.

ESCANDELL, María Victoria, 2007: *Apuntes de Semántica Léxica*, Madrid: Editorial UNED.

ESCUDERO, Gerard, Lluís MÀRQUEZ, German RIGAU, 2000: "A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation", *Proceedings of The Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*, 31-36.

ESPUNYA I PRAT, Anna, 1994: "Computational linguistics: a brief introduction", *Links & Letters*, número 1, 9-23.

FELLBAUM, Christiane (ed.), 1998: *WordNet: An Electronic Lexical Database*, Cambridge: MIT Press.

FELDMAN, Jerome, Ellen DODGE, y John BRYANT, 2010: "Embodied Construction Grammar", en Bernd HEINE y Heiko NARROG (eds.), 2010: *The Oxford Handbook of Linguistic Analysis*, 121-146.

FILLMORE, Charles, 1976: "Frame Semantics and the nature of language", *Annals of the New York Academy of Sciences*, número 280, 20-32.

FILLMORE, Charles, 1982: "Frame Semantics", en LINGUISTIC SOCIETY OF KOREA (ed.), 1982: *Linguistics in the Morning Calm*, Seúl: Hanshin.

FILLMORE, Charles, 1988: "The Mechanisms of Construction Grammar", comunicación presentada en *The Annual Meeting of the Berkeley Linguistics Society*, 35-55.

FILLMORE, Charles, y Beryl ATKINS, 1994: "Starting where the dictionaries stop: the challenge of Computational Lexicography", en Beryl ATKINS y Antonio ZAMPOLLI (eds.), 1994: *Computational Approaches to the Lexicon*, Oxford: Oxford University Press.

FIRTH, John Rupert, 1957: *Papers in Linguistics: 1934-1951*, Londres: Oxford University Press.

FRANCIS, Winthrop Nelson, 1965: "A Standard Corpus of Edited Present-Day American English", *College English*, volumen 26, número 9, 267-273.

FRANCIS, Winthrop Nelson, y Henry KUČERA, 1964: *A Standard Corpus of Edited Present-Day American English for use with Digital Computers,* [version revisada y ampliada en 1979], disponible en https://www.sketchengine.eu/brown-corpus/.

FREGE, Gottlob, 1973 [1892]: "Sobre sentido y referencia", *Estudios sobre Semántica,* Barcelona: Ariel Lingüística.

FULMARI, Abhishek, y Manoj CHANDAK, 2014: "An Approach for Word Sense Disambiguation using modified Naïve Bayes Classifier", *International Journal of Innovative Research in Computer and Communication Engineering Organization*, volumen 2, número 4, 3867-3870.

GANGEMI, Aldo, Nicola GUARINO, Claudio MASOLO, y Alessandro OLTRAMARI, 2003: "Sweetening WordNet with Dolce", *Ai Magazine of The American Association for Artificial Intelligence*, volumen 24, 13-24.

GALE, William, Kenneth Ward CHURCH, y David YAROWSKY, 1992: "A method for disambiguating word senses in a large corpus", *Computers and the Humanities*, número 26, 415–439.

GARCÍA-MIGUEL, José Manuel, Fita GONZÁLEZ DOMÍNGUEZ, Gael VAAMONDE, 2010: "ADESSE. A Database with Syntactic and Semantic Annotation of a Corpus of Spanish", *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, disponible en http://adesse.uvigo.es.

GARRIDO-MEDINA, Joaquín, 1994: *Lógica y Lingüística*, Madrid: Síntesis.

GEERAERTS, Dirk, 2010: *Theories of Lexical Semantics,* Oxford: Oxford University Press.

GERBINO, Elisabetta, Paolo BAGGIA, Edigio GIACHIN, y Claudio RULLENT, 1995: "Analysis and Evaluation of Spontaneous Speech Utterances in Focused Dialogue Contexts", *ESCA Workshop on Spoken Dialogue Systems*, ISCA Archive,185-188.

GODOC, Eric, 2014: *SQL: Los fundamentos del lenguaje*, Barcelona: Ediciones ENI.

GOLDBERG, Adele, 1995: *Constructions: A Construction Grammar Approach to Argument Structure*, Chicago: University of Chicago Press.

GONZÁLEZ-VERGARA, Carlos, 2006: "La Gramática del Papel y la Referencia: una aproximación al modelo", *Onomazein,* número 14 (2006/2), 101-140.

GOSAL, Gurinder, 2015: "A Naïve Bayes Approach for Word Sense Disambiguation", *International Journal of Advanced Research in Computer Science and Software Engineering*, volumen 5, número 7, 336-340.

GOYVAERTS, Jan, 2007: *Regular Expressions, The Complete Tutorial*, disponible en https://www.regular-expressions.info/print.html.

GRISHMAN, Ralph, 1986: *Computational Linguistics*, Cambridge: Cambridge University Press.

HAN, Jiawei, Micheline KAMBER, y Jian PEI, 2012: *Data Mining: Concepts and Techniques (3rd edition)*, Waltham: Morgan Kaufmann.

HANCOCK, Thomas, Tao JIANG, Ming LI, y John TROMP, 1996: "Lower Bounds on Learning Decision Lists and Trees", *Information and Computation*, volumen 2, número 126, 114-122.

HARRIS, Zellig, 1954: ''Distributional structure'', *Word*, volumen 10, número 23, 146-162.

HASTIE, Trevor, Robert TIBSHIRANI, y Jerome FRIEDMAN, 2009: *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, segunda edición, Nueva York: Springer-Verlag.

HATZIVASSILOGLOU, Vasileios, 1994: "Do we Need Linguistics When We Have Statistics? A Comparative Analysis of the Contributions of Linguistic Cues to a Statistical Word Grouping System", en Judith KLAVANS y Philip RESNIK (eds.), 1996: *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, Cambridge: The MIT Press.

HONG, Fei, 2015: "Previous Research on Lexical Ambiguity and Polysemy" en Fei HONG: *Verb Sense Discovery in Mandarin Chinese: A Corpus based Knowledge-Intensive Approach*, Berlin Heidelberg: Springer-Verlag.

HORNBY, Albert Sidney, Anthony Paul COWIE y Windsor LEWIS, 1974: *Oxford Advanced Learner's Dictionary of Current English*, Londres: Oxford University Press.

HOSTE, Véronique, Iris HENDRICKX, Walter DAELEMANS, y Antal VAN DEN BOSCH, 2002: "Parameter optimization for machine-learning of word sense disambiguation", *Natural Language Engineering*, volumen 8, número 4, 311-325.

HOY, Matthew, 2018: "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants", *Medical Reference Services Quarterly*, volumen 37, número 1, 81-88.

IDE, Nancy, y Jean VÉRONIS, 1998: "Introduction to the special issue on word sense disambiguation: the state of the art", *Computational Linguistics*, volumen 24, número 1, 1-40.

JACKENDOFF, Ray, 2002: *Fundamentos del Lenguaje: Mente, Significado, Gramática y Evolución*, México D.F.: Fondo de Cultura Económica.

JAMES, Gareth, Daniela WITTEN, Trevor HASTIE, y Robert TIBSHIRANI, 2013: *An Introduction to Statistical Learning with Applications in R*, Nueva York: Springer-Verlag.

JEŽEK, Elisabetta, 2016: *The Lexicon: An Introduction*, Nueva York: Oxford University Press.

JIANG, Jay, y David W. CONRATH, 1997: "Semantic similarity based on corpus statistics and lexical taxonomy", comunicación presentada en The International Conference on Research in Computational Linguistics (ROCLING X).

JOACHIMS, Thorsten, 1998: "Text categorization with Support Vector Machines: Learning with many relevant features", en Claire NÉDELLEC y Celine ROUVEIROL (eds.), 1998: *Machine Learning ECML-98: Lecture Notes in Computer Science*, Berlin Heidelberg: Springer.

JUNG, Yoonsuh, y Jianhua HU, 2015: "A *K*-fold Averaging Cross-validation Procedure", *Journal of Nonparametric Statistics*, volumen 27, número 2,1-13

JURAFSKY, Daniel, y James MARTIN, 2009: *Speech and language processing: an introduction to natural language processing, speech recognition, and computational linguistics*, New Jersey: Prentice Hall.

KAUR-SIDHU, Gurleen, y Navjot KAUR, 2013: "Role of Machine Translation and Word Sense Disambiguation in Natural Language Processing", *IOSR Journal of Computer Engineering (IOSR-JCE)*, volumen 11, número 3, 78-83.

KAY, Paul, y Charles FILLMORE, 1999: "Grammatical constructions and linguistic generalizations: The what's x doing y? construction", *Language*, número 75, 1-33.

KREIBICH, Jay, 2010: *Using SQLite*, O'Reilly Media, Inc., disponible en: https://www.oreilly.com/library/view/using-sqlite/9781449394592/.

LAVID, Julia, 2005: *Lenguaje y nuevas tecnologías: Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*, Madrid: Cátedra.

LANGLEY, Pat, Wayne IBA, y Kevin THOMPSON, 1992: "An analysis of Bayesian classifiers", *Proceedings of the Tenth National Conference on Artificial Intelligence*, 399-406.

LEACOCK, Claudia, y Martin CHODOROW, 1998: "Combining local context and WordNet similarity for word sense identification", en Christiane FELLBAUM (ed.), 1998: *WordNet: An electronic lexical database*, Cambridge: MIT Press, 265–283.

LEACOCK, Claudia, Martin CHODOROW y George MILLER, 1998: "Using corpus statistics and WordNet relations for sense identification", *Computational Linguistics*, volumen 24, número 1, 147-165.

LEECH, Geoffrey, 1997: "Introducing corpus annotation", en Roger GARSIDE, Geoffrey LEECH, y Toni MCENERY (eds.), 1997: *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Londres: Longman.

LESK, Michael, 1986: "Automatic sense disambiguation: How to tell a pinecone from an ice cream cone", comunicación presentada en The ACM SIGDOC Conference.

LESK, Michael, 1987: "Can Machine-Readable Dictionaries Replace a Thesaurus for Searches in Online Catalogs?", comunicación presentada en The 3rd Annual Conference of the UW Centre for the New OED: The Uses of Large Text Databases.

LIN, Dekang, 1998: "An information-theoretic definition of similarity", comunicación presentada en The 15th International Conference on Machine Learning.

LLISTERRI, Joaquim, 2006: "Introducción a los sistemas de diálogo", en Joaquim LLISTERRI y María Jesús MACHUCA (eds.), 2006: *Los sistemas de diálogo*, Bellaterra-Soria: Universitat Autònoma de Barcelona y Fundación Duques de Soria.

LÓPEZ-BOADA, María Jesús, Beatriz LÓPEZ-BOADA, y Vicente DÍAZ-LÓPEZ, 2005: "Algoritmo de aprendizaje por refuerzo continuo para el control de un sistema de suspensión semi-activa", *Revista Iberoamericana de Ingeniería Mecánica*, volumen 9, número 2, 77-91.

LU, Wenpeng, Fanqing MENG, Shoujin WANG, Guoqiang ZHANG, Xu ZHANG, Antai OUYANG, y Xiaodong ZHANG, 2019: "Graph-Based Chinese Word Sense Disambiguation with Multi-Knowledge Integration", *Computers, Materials & Continua*, volumen 61, número 1, 197-212.

LYONS, John, 1968: *Introduction to Theoretical Linguistics*, Cambridge: Cambridge University Press.

LYONS, John, 1977: *Semantics*, Cambridge: Cambridge University Press.

MAIRAL-USÓN, Ricardo, y Francisco RUIZ DE MENDOZA, 2009: "Levels of description and explanation in meaning construction", en Christopher BUTLER y Javier MARTÍN (eds.), 2009: *Deconstructing Constructions*, Amsterdam: John Benjamins.

MAIRAL-USÓN, Ricardo, Sandra PEÑA, Francisco CORTÉS-RODRÍGUEZ y Francisco RUIZ DE MENDOZA, 2013: *Teoría Lingüística: Métodos, Herramientas y Paradigmas*, Madrid: Editorial Universitaria Ramón Areces.

MANARIS, Bill, 1998: "Natural Language Processing: A Human-Computer Interaction Perspective", en Marvin ZELKOWITZ (ed.), 1998: *Advances in Computers*, volumen 47, 1-66, Nueva York: Academic Press.

MANNING, Christopher, y Hinrich SCHÜTZE, 1999: *Foundations of Statistical Natural Language Processing*, Cambridge: The MIT Press.

MANNING, Christopher, Prabhakar RAGHAVAN, y Hinrich SCHUTZE, 2009: *An Introduction to Information Retrieval*, Cambridge: Cambridge University Press.

MÁRQUEZ, Lluis, Mariona TAULÉ, Antonia MARTÍ, Nuria ARTIGAS, Mar GARCÍA, Francis REAL y Dani FERRÉS, 2004: "SENSEVAL-3: The Spanish lexical sample task", comunicación presentada en The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3), Association for Computational Linguistics.

MÁRQUEZ, Lluis, Gerard ESCUDERO, David MARTÍNEZ, y German RIGAU, 2006: "Supervised Corpus-Based Methods for WSD", en Eneko AGIRRE y Philip EDMONDS (eds.), 2006: *Word Sense Disambiguation: Algorithms and Applications*, Dordrech: Springert.

MARTÍ-ANTONÍN, María Antonia, 2018: "Modelos de semántica distribucional", *Actas do XIII Congreso Internacional de Lingüística Xeral*, Vigo: Universidad de Vigo, 16-22.

MATUSZEK, Cynthia, John CABRAL, Michael WITBROCK, y John DEOLIVEIRA, 2006: "An Introduction to the Syntax and Content of CyC", *AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, 1-6.

MCCLELLAND, James, y David RUMELHART, 1981: "An interactive activation model of context effects in letter perception: part 1. An account of basic findings", *Psychological Review*, número 88, 375-407.

MCENERY, Tony, y Andrew WILSON, 1996: *Corpus Linguistics*, Edimburgo: Edinburgh University Press.

MCRAE, Ken, Virginia DE SA, y Mark SEIDENBERG, 1997: "On the nature and scope of featural representations of word meaning", *Journal of Experimental Psychology: General*, número 126, 99-130.

METEER, Marie, y Herbert GISH, 1994: "Integrating Symbolic and Statistical Approaches in Speech and Natural Language Applications", en Judith KLAVANS y Philip RESNIK (eds.), 1996: *The*

Balancing Act: Combining Symbolic and Statistical Approaches to Language, Cambridge: The MIT Press.

MEYER, Charles, 2002: *English Corpus Linguistics: An Introduction*, Cambridge: Cambridge University Press.

MILLER, George, 1985: 'Wordnet: A Dictionary Browser", *Proceedings of the First Conference of the UW Centre for the New Oxford Dictionary*, University of Waterloo.

MILLER, George, Richard BECKWITH, Christiane FELLBAUM, Derek GROSS, y Katherine MILLER, 1991: "Introduction to WordNet: An On-line Lexical Database", *International Journal of Lexicography*, volumen 3, número 4.

MIRANDA-GARCÍA, Antonio, 1993: "Modelo teórico del lexicón mental", *Revista de Lingüística Cauce*, número 16, 91-100.

MOONEY, Raymond, 1996: "Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning and Bias Learning to Disambiguate Word Senses", comunicación presentada en Conference on Empirical Methods in Natural Language Processing (EMNLP-96).

MOOR, James, 2006: "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years", *AI Magazine*, volumen 27, número 4, 87.

MORENO CABRERA, Juan Carlos, 2013: *Cuestiones clave de la Lingüística*, Madrid: Síntesis.

MORENO SANDOVAL, Antonio, 1996: *Lingüística Computacional*, Madrid: Síntesis.

MURPHY, Lynne, 2010: *Lexical Meaning*, Cambridge: Cambridge University Press.

NAVIGLI, Roberto, 2009: "Word sense disambiguation: a survey", *ACM Computing Surveys (CSUR)*, volumen 41, número 2, 1-69.

NEVZOROVA, Olga, Alfiya GALIEVA y Vladimir NEVZOROVC, 2015: "Sentence context and resolving lexical ambiguity for special groups of words on the base of corpus data", *Procedia: Social and Behavioral Sciences*, número 198, 359-366.

NG, Hwee Tou, 1997: "Getting serious about word sense disambiguation", comunicación presentada en The ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? (Washington D.C.).

PASINI, Tommaso, 2020: "The Knowledge Acquisition Bottleneck Problem in Multilingual Word Sense Disambiguation", comunicación presentada en The Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20).

PATWARDHAN, Siddharth, Satanjeev BANERJEE y Ted PEDERSEN, 2003: "Using measures of semantic relatedness for word sense disambiguation", comunicación presentada en The 4th International Conference on Intelligent Text Processing and Computational Linguistics.

PAVEY, Emma, 2010: *The Structure of Language: An Introduction to Grammatical Analysis*, Cambridge: Cambridge University Press.

PERIÑÁN-PASCUAL, Carlos, 2012a: "En defensa del procesamiento del lenguaje natural fundamentado en la lingüística teórica", *Onomázein*, número 26, 13-48.

PERIÑÁN-PASCUAL, Carlos, 2012b: "The situated common-sense knowledge in FunGramKB", *Review of Cognitive Linguistics*, volumen 10, número 1, 184-214.

PERIÑÁN-PASCUAL, Carlos, 2015: "The underpinnings of a composite measure for automatic term extraction: the case of SRC", *Terminology*, volumen 21, número 2, 151-179.

PERIÑÁN-PASCUAL, Carlos, 2017: "Bridging the gap within text-data analytics: a computer environment for data analysis in linguistic research", *Revista de Lenguas para Fines Específicos*, volumen 23, número 2, 111-132.

PERIÑÁN-PASCUAL, Carlos, y Francisco ARCAS-TÚNEZ, 2004: "Meaning postulates in a lexico-conceptual knowledge base", comunicación presentada en The 15th International Workshop on Databases and Expert Systems Applications.

PERIÑÁN-PASCUAL, Carlos, y Francisco ARCAS-TÚNEZ, 2007: "Cognitive modules of an NLP knowledge base for language understanding", *Procesamiento del Lenguaje Natural*, número 39, 197-204.

PERIÑÁN-PASCUAL, Carlos, y Francisco ARCAS-TÚNEZ, 2010: "The architecture of FunGramKB", en *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, European Language Resources Association (ELRA), 2667-2674.

PERIÑÁN-PASCUAL, Carlos, y Ricardo MARIAL-USÓN, 2009: "Bringing Role and Reference Grammar to natural language understanding", *Procesamiento del Lenguaje Natural*, número 43, 265-273.

PERIÑÁN-PASCUAL, Carlos, y Ricardo MARIAL-USÓN, 2010a: "La gramática de COREL: un lenguaje de representación conceptual", *Onomazein*, número 21, 11-45.

PERIÑÁN-PASCUAL, Carlos, y Ricardo MARIAL-USÓN, 2010b: "Teoría lingüística y representación del conocimiento: una discusión preliminar", en Dolores GARCÍA-PADRÓN y María FUMERO-PÉREZ (coords.), 2010: *Tendencias en lingüística general y aplicada*, Madrid: Peter Lang.

PETRUCK, Miriam, y Gerard DE MELO, 2012: "Precedes: A semantic relation in FrameNet, *Proceedings of the Workshop on Language Resources for Public Security Applications*, 45-49.

PIANTA, Emanuele, Luisa BENTIVOGLI, y Christian GIRARDI, 2002: "MultiWordNet: Developing and Aligned Multilingual Database", Proceedings of the First International Conference on Global WordNet, 293-302.

PINKER, Steven, 2001: *Cómo funciona la mente*, Barcelona: Imago Mundi.

POPESCU, Marius, y Florentina HRISTEA, 2010: "State of the art versus classical clustering for unsupervised word sense disambiguation", *Artificial Intelligence Review*, número 35, 241-264.

PROCTER, Paul, 1978: *Longman Dictionary of Contemporary English*, [6ta edición en línea], Londres: Longman Group Limited, disponible en https://www.ldoceonline.com/es-LA/.

PUSTEJOVSKY, James, 1991: "The Generative Lexicon"*, Computational Linguistics*, número 17, 409-41.

PUSTEJOVSKY, James. 1995: *The Generative Lexicon*, Cambridge: The MIT Press.

PUSTEJOVSKY, James, y Brad BOGURAEV (eds.), 1996: *Lexical Semantics: The Problem of Polysemy*, Oxford: Oxford University Press.

PUTNAM, Hilary, 1975: "El significado de «significado»", en Luis VALDÉS VILLANUEVA (ed.), 1991: *La búsqueda del significado: lecturas de filosofía del lenguaje*, Madrid: Universidad de Murcia.

QUILLIAN, Ross, 1968: "Semantic Memory", en Marvin MINSKY (ed.), 2003: *Semantic Information Processing*, Cambridge: The MIT Press.

RADA, Roy, Hafedh MILI, y María BLETTNER, 1989: "Development and Application of a Metric on Semantic Nets", *IEEE Transactions on Systems Man and Cybernetics*, volumen 19, número 1, 17-30.

REAL ACADEMIA ESPAÑOLA, 2008: *Banco de datos CREA: Corpus de referencia del español actual*, [versión 3.2 en línea], disponible en http://corpus.rae.es/creanet.html.

REAL ACADEMIA ESPAÑOLA, 2014: *Diccionario de la Lengua Española*, 23.ª ed., [versión 23.4 en línea], disponible en https://dle.rae.es.

RESNIK, Phillip, 1995: "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", comunicación presentada en The 14th International Joint Conference on Artificial Intelligence.

RÍOS, Annette, Laura MASCARELL, y Rico SENNRICH, 2017: "Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings", *Proceedings of the Conference on Machine Translation (WMT)*, volumen 1, 11-19.

RISH, Irina, 2001: "An Empirical Study of the Naive Bayes Classifier", *The IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, volumen 3, número 22, 41-46.

RIVANO, Emilio, 2002: "Fuga en Frege: Proposiciones singulares como artefactos lingüísticos silogísticos de construcción de verdad", *Revista de Lingüística Teórica y Aplicada*, número 40, 161-176.

ROKACH, Lior, y Oded MAIMON, 2002: "Top-Down Induction of Decision Trees Classifiers: A Survey", *IEEE Transactions on Systems, Man and Cybernetics*, volumen 1, número 11, 1-12.

RUIZ DE MENDOZA, Francisco, y Ricardo MAIRAL-USÓN, 2008: "Levels of description and constraining factors in meaning construction: an introduction to the Lexical Constructional Model", *Folia Linguistica*, volumen 42, número 2, 355-400.

RUSSEL, Stuart, y Peter NORVIG, 2002: *Artificial Intelligence: A Modern Approach*, segunda edición, Nueva Jersey: Prentice Hall.

SADOWSKY, Scott, 2006: *Corpus Dinámico del Castellano de Chile (Codicach)*, base de datos electrónica, disponible en http://sadowsky.cl/codicach.html.

SALTON, Gerard, y Michael MCGILL, 1986: *Introduction to Modern Information Retrieval*, Nueva York: McGraw-Hill.

SANTALLA del Río, María Paula, 2005: "La elaboración de corpus lingüísticos", en Mario CAL, Paloma NÚÑEZ, e Ignacio PALACIOS (eds.), 2005: *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de Lenguas*, Santiago de Compostela: Servizo de Publicacións e Intercambio Científico Universidad de Santiago de Compostela.

SAUSSURE, Ferdinand de, 2003 [1916]: *Curso de Lingüística General*, Buenos Aires: Losada.

SAVOY, Jacques, y Eric GAUSSIER, 2010: "Information Retrieval", en Nitin INDURKHYA y Fred DAMERAU (eds.), 2010: *Handbook of Natural Language Processing*, Nueva York: Chapman and Hall.

SCHÜTZE, Hinrich, 1998: "Automatic word-sense discrimination", *Computational Linguistics*, volumen 24, número 1, 97-123.

SECO, Nuno, Tony VEALE, y Jer HAYES, 2004: "An Intrinsic Information Content Metric for Semantic Similarity in WordNet", comunicación presentada en The 16th European Conference on Artificial Intelligence (ECAI).

SINHAL, Ruchika, y Kapil GUPTA, 2014: "Machine Translation Approaches and Design Aspects", *IOSR Journal of Computer Engineering (IOSR-JCE)*, volumen 16, número 1, 22-25.

SOUMYA, George, y Joseph SHIBILY, 2014: "Text Classification by Augmenting Bag of Words (BOW): Representation with Co-occurrence Feature", *IOSR Journal of Computer Engineering (IOSR-JCE)*, volumen 16, número 1, 34-38.

STOKOE, Christopher, Michael OAKES, y John TAIT, 2003: "Word Sense Disambiguation in Information Retrieval Revisited", *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 159-166.

SUBIRATS, Carlos, 2004: "FrameNet Español. Una red semántica de marcos conceptuales", comunicación presentada en el VI Congreso Internacional de Lingüística Hispánica de la Universidad de Leipzig.

SUBIRATS, Carlos, y Miriam PETRUCK, 2003: "Surprise. Spanish FrameNet", comunicación presentada en The International Congress of Linguists: Workshop on Frame Semantics.

SUTOPO, Anam, y Diyah HASTUTI, 2020: "The Role of Machine Translators in Academic Translation Teaching", *International Journal of Scientific Engineering and Science*, volumen 4, número 5, 29-31.

SUTTON, Richard, y Andrew BARTO, 1998: *Reinforcement learning: An introduction*, Cambridge: The MIT Press.

TORRES-RAMOS, Sulema, 2012: "Estudio sobre métodos tipo Lesk usados para la desambiguación de sentidos de palabras", *Research in Computer Science*, número 47, 139-148.

TULVING, Endel, 1985: "How many memory systems are there?", *American Psychologist*, número 40, 385-398.

UNIÓN EUROPEA, 2018: IATE ('Interactive Terminology for Europe'), [versión 2.20.1 en línea], disponible en https://iate.europa.eu.

USTALOV, Dmitry, Denis TESLENKO, Alexander PANCHENKO, Mikhail CHERSNOSKUTOV, Chris BIEMANN, y Simone PONZETTO, 2018: "An Unsupervised Word Sense Disambiguation System for Under-Resourced Languages", *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.

VAN GOMPEL, Maarten, y Antal VAN DEN BOSCH, 2013: "WSD2: Parameter optimisation for Memory-based Cross-Lingual Word-Sense Disambiguation", *Second Joint Conference on Lexical and Computational Semantics (*SEM),* volumen 2, Seventh International Workshop on Semantic Evaluation (SemEval 2013), 183-187.

VAN LE, Duy, James MONTGOMERY, Kenneth KIRKBY, y Joel SCANLAN, 2018: "Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting", *Journal of Biomedical Informatics*, número 86, 49-58.

VAN VALIN, Robert, 2005: *Exploring the syntax-semantics interface*, Cambridge: Cambridge University Press.

VAN VALIN, Robert, y Randy LAPOLLA, 1997: *Syntax: structure, meaning and function,* Cambridge: Cambridge University Press.

VICKREY, David, Luke BIEWALD, Marc TEYSSIER, y Daphne KOLLER, 2005: "Word-Sense Disambiguation for Machine Translation", *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 771-778.

VIDHU BHALA, Vidyasagard, y Murugappan ABIRAMI, 2014: "Trends in word sense disambiguation", *Artificial Intelligence Review*, volumen 42, número 2, 159-171.

VIGLIOCCO, Gabriella, y David WILSON, 2005: "Semantic Representation", en Gareth GASKELL (ed.), 2007: *The Oxford Handbook of Psycholinguistics*, Oxford: Oxford University Press.

VILLAYANDRE, Milka, 2008: "Lingüística con Corpus", *Estudios Humanísticos Filología*, número 30, 329-349.

WAGNER, Christian, 2006: "Breaking the Knowledge Acquisition Bottleneck Through Conversational Knowledge Management", *Information Resources Management Journal*, volumen 19, número 1, 70-83.

WEAVER, Warren, 1955: "Translation", en William LOCKE y Andrew BOOTH (eds.), 1955: *Machine Translation of Languages*, Nueva York: John Wiley & Sons.

WEINREICH, Uriel, 1964: "Webster's Third: A Critique of its Semantics"*, International Journal of American Linguistic*, número 30, 405-409.

WIDLAK, Magdalena, 2004: *Influence of Word Sense Disambiguation on Text Classification*. Tesis de maestría, Universidad de Ottawa.

WILKS, Yorick, Dan FASS, Cheng-Ming GUO, James MCDONALD, Tomy PLATE y Brian SLATOR, 1998: "Machine Tractable Dictionaries as Tools and Resources for Natural Language Processing", comunicación presentada en The 12th Conference on Computational Linguistics.

WILKS, Yorick, Dan FASS, Cheng-Ming GUO, James MCDONALD, Tomy PLATE y Brian SLATOR, 1989: "A tractable machine dictionary as a resource for computational semantics", en Branimir BORUGAEV y E.J. BRISCOE (eds.), 1989: *Computational lexicography for natural language processing,* Harlow: Longman.

WU, Zhibiao, y Martha PALMER, 1994: "Verbs semantics and lexical selection", comunicación presentada en The 32nd Annual Meeting on Association for Computational Linguistics.

YAROWSKY, David, 1995: "Unsupervised word sense disambiguation rivaling supervised method", comunicación presentada en The 33rd Annual Meeting of the Association for Computational Linguistics.

YAROWSKY, David, 1997: "Homograph Disambiguation in Text-to-Speech Synthesis", en Jan VAN SANTEN, Joseph OLIVE, Richard SPROAT y Julia HIRSCHBERG (eds.), 1997: *Progress in Speech Synthesis*. Nueva York: Springer.

YURAFSKY, Daniel, y James H. MARTIN, 1998: *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Nueva Jersey: Prentice Hall.

ZHANG, Zhongheng, 2016: "Introduction to machine learning: K-nearest neighbors", *Annals of Translational Medicine*, volumen 4, número 11, 218-218.

ZHANG, Yin, Rong JIN, y Zhi-Hua ZHOU, 2010: "Understanding Bag-of-Words Model: A Statistical Framework", *International Journal of Machine Learning and Cybernetics*, número 1, 43-52.

ZHOU, Zili, Yanna WANG, y Junzhong GU, 2008: "New Model of Semantic Similarity Measuring in WordNet", comunicación presentada en The Second International Conference on Future Generation Communication and Networking Symposia.

Anexos

Anexo 1. Minidiccionario para los sentidos de «partido» correspondiente al corpus SENSEVAL-3

```
<lexelt>
<sense>
        <sense id="partido.1" definition="Organización política cuyos miembros comparten la misma ideología" used="yes">
                <example text="el principal partido del país"/>
                <example text="el partido en la oposición"/>
                        <collocation text="partido centrista"/>
                        <collocation text="partido comunista"/>
                        <collocation text="partido conservador"/>
                        <collocation text="partido cristianodemócrata"/>
                        <collocation text="partido de derechas"/>
                        <collocation text="partido de izquierdas"/>
                        <collocation text="partido de la oposición"/>
                        <collocation text="partido demócrata"/>
                        <collocation text="partido ecologista"/>
                        <collocation text="partido estatal"/>
                        <collocation text="partido fascista"/>
                        <collocation text="partido gobernante"/>
                        <collocation text="partido laborista"/>
                        <collocation text="partido local"/>
                        <collocation text="partido mayoritario"/>
                        <collocation text="partido minoritario"/>
                        <collocation text="partido nacional"/>
                        <collocation text="partido nacionalista"/>
                        <collocation text="partido político"/>
                        <collocation text="partido populista"/>
                        <collocation text="partido progresista"/>
                        <collocation text="partido republicano"/>
                        <collocation text="partido socialdemócrata"/>
                        <collocation text="partido socialista"/>
                        <collocation text="partido tradicional"/>
                        <collocation text="partido ultra"/>
                        <collocation text="partido ultraderechista"/>
                        <collocation text="partido ultraortodoxo"/>
                        <collocation text="partido verde"/>
        <synset wordnet="1.5" id="05259394n"/>
</sense>
<sense>
        <sense id="partido.2" definition="Prueba deportiva en la que se enfrentan dos equipos o jugadores" used="yes">
                <example text="partido de baloncesto"/>
                <example text="partido de tenis"/>
                <example text="el mejor partido de la temporada"/>
                        <collocation text="partido amistoso"/>
                        <collocation text="partido de consolación"/>
                        <collocation text="partido de desempate"/>
                        <collocation text="partido de exhibición"/>
                        <collocation text="partido de fútbol"/>
                        <collocation text="partido de homenaje"/>
                        <collocation text="partido de ida"/>
                        <collocation text="partido de tenis"/>
                        <collocation text="partido de vuelta"/>
        <synset wordnet="1.5" id="04780657n"/>
</sense>
</lexelt>
```

Anexo 2: Selección de 120 instancias para la unidad léxica «partido» extraída desde el corpus SENSEVAL-3

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<corpus lang="Spanish">
	<lexelt item="partido.n">
		<instance id="partido.n.1" docsrc="efe_3607_2000/01/07">
			<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
			<cat scheme="IPTC" code="15000000"/>
			<answer instance="partido.n.1" senseid="partido.2"/>
			<context>
			<previous>Con anterioridad, ha militado en San Lorenzo de Almagro, equipo en el que permaneció durante ocho temporadas y con el que llegó a jugar casi doscientos partidos en primera división.</previous>
			<target>El nuevo jugador del Elche fue traspasado al Numancia a principios de temporada, pero no ha entrado en los planos de Andoni Goicoechea, quien tan sólo le ha utilizado en un <head>partido</head> de Copa del Rey.</target>
			<following>Con respecto a la categoría en la que juega el Elche, el nuevo jugador ilicitano, ha dicho que "al igual que la primera división, la segunda es muy competitiva, hay grandes equipos. Hay que pelear mucho para sacar adelante un encuentro y eso es algo que no se me va a poder criticar, puesto que vengo a darlo todo, porque quiero jugar y triunfar en España".</following>
			</context>
		</instance>
		<instance id="partido.n.2" docsrc="efe_4178_2000/01/08">
			<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
			<cat scheme="IPTC" code="15000000"/>
			<answer instance="partido.n.2" senseid="partido.2"/>
			<context>
			<previous> Manzano auguró que Shoji Jo se aclimatará al fútbol español porque tiene "gran rapidez de movimientos" y que lo poco que puede faltarle lo adquirirá con un mínimo periodo de adaptación.</previous>
			<target>La incorporación de Jo ha propiciado que el club vallisoletano haya recibido numerosas peticiones de la camiseta del jugador desde Japón, desde donde se han solicitado también acreditaciones para ver los
			<head>partidos</head> del equipo.</target>
			<following>Unos treinta periodistas nipones estarán presentes el próximo lunes en la presentación del jugador, según confirmaron a Efe fuentes del club</following>
			</context>
		</instance>
		<instance id="partido.n.3" docsrc="efe_6478_2000/01/11">
		<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
		<cat scheme="IPTC" code="15000000"/>
		<answer instance="partido.n.3" senseid="partido.2"/>
			<context>
			<previous>Con el resultado, los "diablos rojos" acumulan cuatro puntos, terminan con cero en el saldo de goles y quedan a la espera de una goleada por tres tantos del Vasco da Gama al Necaxa mexicano para conquistar la plaza por el tercer lugar del torneo al que también opta el Real Madrid español.</previous>
			<target>Sin jugar el <head>partido</head> de la tercera jornada del Grupo B, los mexicanos acumulan cuatro puntos, pero tienen dos goles a su favor.</target>
			<following>El renovado Manchester no tuvo pena de ofrecer su espectáculo a una clientela inicial de 2.000
			aficionados que fue multiplicándose mientras llegaba la hora del duelo central entre el Vasco da Gama y el Necaxa mexicano.</following>
			</context>
		</instance>
		<instance id="partido.n.4" docsrc="efe_12580_2000/01/19">
		<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
		<cat scheme="IPTC" code="15000000"/>
		<answer instance="partido.n.4" senseid="partido.2"/>
			<context>
			<previous>El combinado polaco tiene prev isto ejercitarse esta tarde en el campo de césped artificial de la ciudad deportiva de alicante y, a partir de mañana, entrenarán en sesiones dobles, primero en el estadio Rico Pérez y
			posteriormente en el nuevo campo municipal de Alicante.</previous>
```

<target>El próximo lunes se incorporarán a la expedición tres jugadores que este fin de semana disputan <head>partidos</head> con sus respectivos equipos, en ligas distintas a la polaca, así como otros tantos directivos.</target>
<following>La delegación polonesa ha llegado a España encabezada por el que fuera guardameta de la selección polaca y del Hércules, Jan Tomasewsky. EFE</following>
</context>
</instance>
<instance id="partido.n.5" docsrc="efe_14371_2000/01/21">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.5" senseid="partido.1"/>
<context>
<previous>Al margen de esos dos millones de marcos, el nerviosismo va en aumento ante la cita, el domingo, de la cúpula cristianodemócrata con el informe de los auditores que investigan la contabilidad de esa formación desde 1990.</previous>
<target>Un portavoz de la CDU aseguró hoy que hasta después de la reunión del "presidium" del domingo no va a haber información sobre los resultados de la auditoría encargada por el <head>partido</head>.</target>
<following>Sin embargo, a mediados de semana se filtró ya la noticia de que los investigadores han detectado en las cuentas del partido hasta diez millones de marcos (unos cinco millones de dólares) de los que se desconoce la procedencia.</following>
</context>
</instance>
<instance id="partido.n.6" docsrc="efe_14819_2000/01/21">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONMANO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.6" senseid="partido.2"/>
<context>
<previous>El conjunto malagueño, que llegó hoy a Santander, viajó con las bajas de la extremo Ana Belén López y de la pivote Alma Valiente, mientras que la portera Diana Serrano y la pivote Ana Isabel Forner, ambas con gripe, son dudas para el partido.</previous>
<target>El equipo santanderino, que la pasada temporada jugó la promoción de descenso y se salvó en los últimos <head>partidos</head>, se ha reforzado esta campaña con varias jugadoras extranjeras y aún no conoce la derrota en casa.</target>
<following>El Cajacantabria comenzó la fase por la permanencia con cinco puntos, dos más que el Famadesa, al que también le valdría el empate para continuar con sus aspiraciones y mantenerse en la categoría.</following>
</context>
</instance>
<instance id="partido.n.7" docsrc="efe_18037_2000/01/25">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.7" senseid="partido.2"/>
<context>
<previous>Tras el entrenamiento, el defensa uruguayo Tabaré Silva y el cordobés Juan Carlos resumieron el sentir general y el optimismo que mantiene la plantilla, pese a la delicada situación del equipo.</previous>
<target>"Si tiramos ya la toalla, mejor no seguir, y lo que no podemos hacer es convertirnos en unos perdedores. Queda prácticamente toda la segunda vuelta y aquí no hay nada hecho. Estamos a tres <head>partidos</head> de salvación y por supuesto que lo vamos a intentar", aseveró Juan Carlos, máximo goleador del equipo sevillista, con ocho tantos.</target>
<following>El ex vallisoletano aseguró que la pasada campaña "estaba más complicado el ascenso, ya que en marzo el equipo estaba a diez o doce puntos del cuarto y, al final, en dos meses se consiguió todo".</following>
</context>
</instance>
<instance id="partido.n.8" docsrc="efe_19418_2000/01/27">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.8" senseid="partido.2"/>
<context>
<previous>Aitor Ocio, defensa central del Albacete Balompié, por una rotura fibrilar, es baja segura para el partido que disputará el próximo sábado, a partir de las 19 horas, en El Madrigal ante el conjunto castellonense del Villarreal.</previous>

<target>Ocio, que arrastraba una sobrecarga muscular, cayó lesionado en el transcurso del primer entrenamiento de la semana, el martes, y se suma a la ya conocida con anterioridad del meta Julio Iglesias, que ha sido suspendido un <head>partido</head> por acumulación de amonestaciones.</target>
<following>Pero los problemas no acaban ahí para Julián Rubio, técnico del Albacete, quien mañana, por cierto, cumple años, pues la ausencia de Julio Iglesias será cubierta por Carlos Cano, pero el problema surge con el suplente, pues iba a ser convocado Alejandro, titular del filial, pero se ha lesionado esta semana y estará en el banquillo José Almendros, suplente del Albacete B, de 19 años.</following>
</context>
</instance>
<instance id="partido.n.9" docsrc="efe_19465_2000/01/27">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.9" senseid="partido.1"/>
<context>
<previous>El presidente de Coalición Canaria, Paulino Rivero, advirtió hoy que sería "mucho más difícil" que esta formación diera apoyo a un hipotético Gobierno PSOE-IU que a una mayoría parlamentaria sólo de los socialistas.</previous>
<target>Rivero, en conferencia de prensa, hizo un balance positivo de la colaboración de CC con el Gobierno del PP en la pasada Legislatura y dijo que "todo apunta" a que el entendimiento con ese <head>partido</head> sea el que se produzca a partir del 12 de marzo.</target>
<following>Así, destacó el hecho de que no se hayan producido "graves tensiones, ni grandes dificultades" y que Coalición Canaria ha cumplido "al cien por cien" todos sus objetivos, tanto en defensa de los intereses de Canarias, como en su afán de contribución a la gobernabilidad del Estado.</following>
</context>
</instance>
<instance id="partido.n.10" docsrc="efe_20667_2000/01/28">
<cat scheme="ANPA" code="DEP:DEPORTES,VOLEIBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.10" senseid="partido.2"/>
<context>
<previous>Con todo, matizó que no deben confiarse ante el CV Vigo, al ser "un equipo que, a pesar de ir penúltimo, t iene buenos jugadores y que en su cancha es difícil, al jugar muy arropado por su público".</previous>
<target>Hervás piensa que ganar en Vigo les valdría para afrontar los siguientes compromisos "con más t ranquilidad", ya que en las tres próximas jornadas se enfrentarán al Guaguas Las Palmas, Numancia de Soria y Unicaja Almería, <head>partidos</head> en los que, según el técnico, el Ivesur se jugará la clasificación para la fase por el título".</target>
<following>La jornada decimoquinta en la Superliga se completa con los siguientes encuentros: Universidad Complutense-Escáner Cartagena; Hospitalet-U.Granada; Unicaja Almería-Pepsi Gijón; Numancia-Aguas de Huelva y Guaguas Las Palmas-Sant Pere y Sant Pau de Tarragona</following>
</context>
</instance>
<instance id="partido.n.11" docsrc="efe_21080_2000/01/29">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.11" senseid="partido.2"/>
<context>
<previous>El Málaga CF intentará dar continuidad a la buena racha que lleva fuera de casa y repetir la victoria que logró hace once campañas en el estadio bético, en la 88/89, la última en la máxima categoría del desaparecido CD Málaga, cuando venció por 1-2.</previous>
<target>El conjunto malagueño llega herido, al igual que el Betis por sus dos últimas derrotas, y buscará resarcirse del <head>partido</head> que perdieron contra el Alavés la pasada jornada en La Rosaleda y situarse en una posición cómoda de la tabla.</target>
<following>El equipo blanquiazul ha conseguido hasta ahora cuatro victorias lejos de Málaga, frente al Barcelona (1-2), Celta (2-4), Racing de Santander (2-3) y Espanyol (0-2), por lo que tratará de que el estadio Manuel Ruiz de Lopera sea un lugar propicio para seguir encadenando éxitos que palíen los tropiezos que sufre en su campo.</following>
</context>
</instance>
<instance id="partido.n.12" docsrc="efe_420_2000/02/01">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>

<answer instance="partido.n.12" senseid="partido.2"/>
<context>
<previous></previous>
<target>El Deportivo recibirá mañana al Osasuna en el <head>partido</head> de vuelta de octavos de final de la Copa del Rey con la intención de lavar la mala imagen que en las últimas semanas ha ofrecido en la Liga, mientras que el conjunto de Pamplona tratará de repetir la gesta que consiguió ante el Valencia en la anterior eliminatoria.</target>
<following>Los deportivistas, a pesar de mantenerse como líderes en la competición liguera, han perdido la mayor parte de la renta de la que disfrutaban frente al Barcelona y el Zaragoza. Todos entienden que eliminar al Osasuna es el antídoto para superar la situación.</following>
</context>
</instance>
<instance id="partido.n.13" docsrc="efe_2479_2000/02/03">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONCESTO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.13" senseid="partido.2"/>
<context>
<previous>"Veremos si somos un equipo con la suficiente ambición para movernos entre los puestos del sexto al décimo, pero no más abajo porque entonces nos pondríamos nerviosos", agregó.</previous>
<target>Los insulares llegarán a Gijón sin la presencia de uno de los pilares básicos del equipo, el estadounidense Deon Thomas, que sufrió una elongación muscular torácica en el <head>partido</head> ante el Caja San Fernando en Vitoria. Lo médicos observarán su evolución en 72 horas y diagnosticarán si mejora y puede estar presente en el encuentro ante el León Caja España, dentro de una semana.</target>
<following>Hussein, ante este contratiempo que supone la baja de Thomas, ha incorporado al jugador dominicano Marlon Martínez, pívot del CB Gran Canaria de la Liga EBA, aunque es difícil que el jugador caribeño disfrute de algunos minutos en la cancha.</following>
</context>
</instance>
<instance id="partido.n.14" docsrc="efe_5411_2000/02/07">
<cat scheme="ANPA" code="SOC:SOCIEDAD-SALUD,COMUNICACION"/>
<cat scheme="IPTC" code="07000000"/>
<cat scheme="IPTC" code="14000000"/>
<answer instance="partido.n.14" senseid="partido.1"/>
<context>
<previous>CAMBIO 16: "HEIL HAIDER!"</previous>
<target>La revista "Cambio 16" dedica su portada de esta semana al líder del <head>partido</head> liberal austriaco, Joerg Haider, que "se ha convertido en un escalofrío que recorre Europa. El pacto de los conservadores austriacos con la ultraderecha reabre en Europa el temor de una ideología que debía estar enterrada".</target>
<following>La revista recuerda algunas de las líneas generales de las opiniones de Haider como por ejemplo su idea de que "Austria es un aborto ideológico", así como su admiración por el III Reich, que considera que tenía "una metódica política de empleo".</following>
</context>
</instance>
<instance id="partido.n.15" docsrc="efe_5882_2000/02/07">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.15" senseid="partido.1"/>
<context>
<previous>Frutos dijo que "el acuerdo ha removido las aguas estancadas de la izquierda", criticó la reacción del presidente de la CEOE, José María Cuevas, y, en referencia a Federico Trillo, afirmó que "algunos ya no encuentran otros argumentos que meterse con las barbas de Víctor Ríos".</previous>
<target>Por su parte, Joaquín Almunia aseguró que con el acuerdo firmado entre su <head>partido</head> e IU ganan "los progresistas", y se mostró convencido de que "los únicos perdedores son quienes van a perder las elecciones el 12 de Marzo".</target>
<following>"Ellos lo saben, porque son de derechas, pero no tontos", aseguró el candidato socialista, quien añadió que "por eso están nerviosos, crispados y, con perdón de sus asesores de imagen, descentrados".</following>
</context>
</instance>
<instance id="partido.n.16" docsrc="efe_11006_2000/02/14">
<cat scheme="ANPA" code="DEP:DEPORTES,NATACION"/>
<cat scheme="IPTC" code="15000000"/>

<answer instance="partido.n.16" senseid="partido.2"/>
<context>
<previous></previous>
<target>El Canoe, campeón de la Liga de waterpolo en 1999, se enfrentará al Martiánez en el primer <head>partido</head> de los cuartos de final de la Copa del Rey que se disputará el próximo fin de semana en las instalaciones municipales del polideportivo Prado de Santo Domingo, de la localidad madrileña de Alcorcón.</target>
<following>En esta XIV edición de la Copa del Rey de waterpolo participan los ocho primeros equipos clasificados al término de la primera vuelta del campeonato de Liga Nacional que finalizó ayer.</following>
</context>
</instance>
<instance id="partido.n.17" docsrc="efe_13956_2000/02/17">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.17" senseid="partido.1"/>
<context>
<previous>Por otra parte, Puigcercós se ha mostrado partidario de la financiación "mayoritariamente pública" de los partidos, aunque ha aceptado un máximo del 25% de donativos privados, siempre que se asegure su transparencia e implique desgravaciones fiscales para los donantes.</previous>
<target>"La financiación ilegal de los <head>partidos</head> políticos es un freno a la economía productiva", ha sostenido el candidato, que ha explicado que, durante la reunión, "algún miembro del Círculo de Economía ha estimado en 20.000 millones de pesetas anuales la cantidad destinada a financiación irregular en Cataluña".</target>
<following>En este sentido, el líder republicano ha exigido un mayor control de las adjudicaciones de obras y servicios a empresas privadas y de la contratación pública, así como sanciones más duras para los casos de cobro de comisiones ilegales y financiación irregular</following>
</context>
</instance>
<instance id="partido.n.18" docsrc="efe_18566_2000/02/23">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONCESTO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.18" senseid="partido.2"/>
<context>
<previous>Jamison, de 23 años, no había participado en los últimos cuatro partidos de los Warriors debido a las dolencias en la rodilla izquierda.</previous>
<target>El alero es el máximo anotador de los Warriors con un promedio de 19,6 puntos por <head>partido</head>, además de ser el mejor debajo de los tableros con una media de 8,4 rebotes por encuentro</target>
<following></following>
</context>
</instance>
<instance id="partido.n.19" docsrc="efe_18663_2000/02/23">
<cat scheme="ANPA" code="DEP:DEPORTES,POLIDEPORTIVO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.19" senseid="partido.2"/>
<context>
<previous></previous>
<target>Las manifestaciones de los jugadores el Real Madrid y Barcelona sobre el próximo partido entre ambos equipos y las lesiones que les afectan, el adiós a Antonio Díaz Miguel y el <head>partido</head> que disputará hoy la selección española contra la de Croacia son las noticias que destacan hoy los periódicos deportivos.</target>
<following>El diario "As" recoge un su primera página "un mensaje de Roberto Carlos desde Tailandia", "Líderes cuanto antes. Ya es hora de que empecemos a meter miedo", en cuanto al maratón de partidos que ha disputado el jugador brasileño, afirma: "No se preocupen por mí, ya descansaré en el avión".</following>
</context>
</instance>
<instance id="partido.n.20" docsrc="efe_21281_2000/02/25">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONCESTO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.20" senseid="partido.2"/>
<context>
<previous>Babkov jugará ante el Breogán este fin de semana, a la espera de que Collins le supla la siguiente semana ante el Adecco Estudiantes, si en ese plazo de tiempo la entidad badalonesa da su visto bueno al fichaje del alero.</previous>

<target>El estadounidense es un escolta de 26 años y 1,94 metros de altura, que en la presente temporada ha jugado dos <head>partidos</head> con el conjunto francés del Basket Strasburgo, en los que promedió 25,5 puntos por encuentro.</target>
<following>Collins, formado en la Universidad de Florida State, fue elegido por Phoenix en la segunda ronda del sorteo universitario de 1.997 con el número 37. En la temporada 1997-98 jugó 23 partidos con Angeles Clippers.</following>
</context>
</instance>
<instance id="partido.n.21" docsrc="efe_22653_2000/02/27">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.21" senseid="partido.2"/>
<context>
<previous>Por su parte, el técnico del Rayo, Juande Ramos, no ha puesto ningún reproche a la victoria españolista, aunque cree que está sobredimensionada, porque entiende que el 5-1 "no es la diferencia entre los dos equipos".</previous>
<target>"Ha sido un marcador duro. Después de controlar el <head>partido</head> (0-1), nos hemos relajado y nos han marcado dos goles por errores infantiles", ha subrayado Ramos.</target>
<following>"Con la mínima ventaja nos pensábamos que estaba todo hecho y que íbamos a ganar. Las tres ausencia en defensa nos han hecho mucho daños. Se han notado mucho", ha añadido</following>
</context>
</instance>
<instance id="partido.n.22" docsrc="efe_22816_2000/02/28">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.22" senseid="partido.2"/>
<context>
<previous>La goleada del Universidad de Chile, flamante campeón de la temporada pasada, destacó en la segunda jornada del torneo de apertura del fútbol chileno al vencer al O`Higgins de Rancagua, por 4-0.</previous>
<target>El cuadro universitario, que el pasado miércoles goleó también al Atlético Nacional de Colombia, por 4-0, en <head>partido</head> válido por el Grupo 4 de la Copa Libertadores de América, esta vez repitió la cifra aunque ahora los autores de los tantos fueron los jugadores suplentes.</target>
<following>Los titulares del Universidad de Chile instalados en la gradas del estadio "El Teniente" de Rancagua, vieron como los goles de Marcos González (a los 3 minutos); del argentino Diego Rivarola (m.36), de Sebastián Pardo (m.53), y Edson Monsalve (m.78), dieron cuenta fácil del dueño de casa, que en la primera fecha había sufrido otra derrota frente al Unión Española, por 5-2.</following>
</context>
</instance>
<instance id="partido.n.23" docsrc="efe_23430_2000/02/28">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.23" senseid="partido.2"/>
<context>
<previous>Además, el conjunto cordobés es el segundo de la categoría que menos goles ha recibido a domicilio, con trece en total y sólo superado en esta faceta por el Badajoz, que lleva siete.</previous>
<target>El Córdoba, tras ganar 0-2 a Osasuna en El Sadar, está situado actualmente en el undécimo puesto de la tabla, con 39 puntos, después de haber encadenado una racha de seis <head>partidos</head> sin perder, en los que ha sumado dieciséis de los dieciocho puntos posibles, gracias a cinco triunfos y un empate</target>
<following></following>
</context>
</instance>
<instance id="partido.n.24" docsrc="efe_23527_2000/02/28">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.24" senseid="partido.1"/>
<context>
<previous>Y confió en que "más allá de la rabia y desánimo, haya plena confianza" en que la movilización "permanente y constante en el País Vasco no cesará ya" hasta que finalmente triunfen "los principios sobre las tácticas".</previous>
<target>En el acto también intervinieron el secretario general del PP, Javier Arenas; el presidente del PP de Euskadi, Carlos Iturgaiz, y el presidente del <head>partido</head> en Vizcaya, Leopoldo Barreda.</target>

<following>Unos minutos después de comenzado el acto, se escuchó un repentino estruendo en el interior de la sede que hizo saltar a todos los presentes, hasta que segundos más tarde se comprendió que el sonoro golpe no era más que una de las tribunas dispuestas para los cámaras de televisión que se había desplomado.</following>
</context>
</instance>
<instance id="partido.n.25" docsrc="efe_24214_2000/02/29">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.25" senseid="partido.2"/>
<context>
<previous>Según el desarrollo del partido, también podrán tener su oportunidad Armando, Serrizuela e Ibagaza, entre otros.</previous>
<target>El Mallorca sólo ha perdido un <head>partido</head> en Son Moix -ante el Real Madrid en la primera jornada de Liga- y confía en la fortaleza que está demostrando en su campo para obtener un resultado positivo ante el Mónaco.</target>
<following>Vázquez no podrá contar con el camerunés Samuel Eto`o, porque ya disputó unos minutos de la Liga de Campeones con el Real Madrid ante el Molde noruego, ni con los lesionados Biagini, "Polo" Quinteros y Djokaj</following>
</context>
</instance>
<instance id="partido.n.26" docsrc="efe_24440_2000/02/29">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.26" senseid="partido.2"/>
<context>
<previous>---------</previous>
<target>- Un <head>partido</head> de suspensión por doble amonestación y consiguiente expulsión a Darío Silva (Málaga), Neru (Racing), Fuentes (Real Sociedad) y Guti (Real Madrid).</target>
<following>- Un partido de suspensión por acumulación de amonestaciones a Urrutia (Athletic), De los Santos (Málaga), Bornes, Oli y Karhan (Betis), Héctor (Sevilla), Lauren (Mallorca), Munitis (Racing), Sa Pinto (Real Sociedad), Karanka (Real Madrid), Kluivert y Abelardo (Barcelona), Pacheta y Muñiz (Numancia).</following>
</context>
</instance>
<instance id="partido.n.27" docsrc="efe_1776_2000/03/02">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.27" senseid="partido.2"/>
<context>
<previous>La victoria del Ulker se fraguó después con tres tiros libres, dos anotados por Sarica y uno por Allen, lo cual supone la disputa de un tercer partido de desempate en Barcelona el próximo jueves.</previous>
<target>"Es una ventaja afrontar el tercer <head>partido</head> en nuestra pista y sentir el apoyo del público", dijo Aito, quien se quejó por el juego hecho hoy por su equipo</target>
<following></following>
</context>
</instance>
<instance id="partido.n.28" docsrc="efe_1959_2000/03/02">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONCESTO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.28" senseid="partido.2"/>
<context>
<previous>El PAOK Salónica frenó la excelente marcha del Maccabi en Liga Europea, propició la primera derrota del conjunto israelí del año 2000 en la competición continental tras imponerse por 67-55 y aplazó el desenlace de la eliminatoria para dentro de una semana, cuando ambos equipos resuelvan su enfrentamiento en el pabellón de la Mano de Elías la próxima semana.</previous>
<target>La fuerza con la que el cuadro israelí afrontó el partido fue frenada de inmediato por el conjunto griego, que tras la desventaja inicial (0-5, m.2), se hizo con el mando del <head>partido</head> y del marcador después de imponer su ritmo y aprovechar el excelente momento anotador de Bill Edwards, que anotó diecisiete puntos.</target>
<following>La aportación de Giorgios Balogiannis resultó determinante. Tres triples consecutivos del alero griego terminaron por dar la victoria al cuadro heleno y forzar el desempate para dentro de siete días</following>
</context>
</instance>

<instance id="partido.n.29" docsrc="efe_4242_2000/03/06">
<cat scheme="ANPA" code="POL:POLITICA,ELECCIONES,PRESIDENCIALES"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.29" senseid="partido.1"/>
<context>
<previous>De ellas, la más importante es la de California, en la que se decidirán 162 delegados republicanos y 367 demócratas.</previous>
<target>En ese estado el sistema de votación está abierto a los independientes y a los representantes de cualquier <head>partido</head>, aunque los delegados demócratas y republicanos se definirán sólo por el voto de cada fuerza política.</target>
<following>En California puede, por tanto, darse la situación de que un candidato gane la votación general, pero pierda frente al rival de su mismo partido en cuanto al número de compromisarios.</following>
</context>
</instance>
<instance id="partido.n.30" docsrc="efe_4831_2000/03/06">
<cat scheme="ANPA" code="POL:POLITICA,ELECCIONES"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.30" senseid="partido.1"/>
<context>
<previous></previous>
<target>Dos candidatos del <head>partido</head> Los Verdes de Canarias llegaron hoy a la playa, se quitaron la ropa y se dedicaron a repartir sus programas electorales, en la famosa Playa del Inglés, en el municipio de Valle del Rey de la isla de La Gomera.</target>
<following>Los desnudos candidatos, entre los que se encontraba su cabeza de lista por la provincia de Santa Cruz de Tenerife, Joaquín Galera, llevaron así a las últimas consecuencias la necesidad de su partido de tener un contacto directo con la gente e innovar sobre los tan frecuentes mítines políticos.</following>
</context>
</instance>
<instance id="partido.n.31" docsrc="efe_8761_2000/03/10">
<cat scheme="ANPA" code="POL:POLITICA,ELECCIONES"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.31" senseid="partido.1"/>
<context>
<previous>La presidenta de EA, Begoña Errazti, opinó hoy que los "imperialistas españoles" intentan "eliminar cualquier viso de esperanza y solución" en el País Vasco porque son conscientes de que cuando desaparezca la violencia "nos dedicaremos a construir país y no nos parará absolutamente nadie".</previous>
<target>Errazti, quien intervino esta noche junto al fundador del <head>partido</head>, Carlos Garaikoetxea, y la candidata Begoña Lasagabaster en un mitin de cierre de campaña en San Sebastián, denunció "la estrategia contra las ideas nacionalistas" de los partidos estatales y su interés por crear "confusión" entre nacionalismo y violencia.</target>
<following>"Que no nos utilicen", advirtió la dirigente de EA, quien instó al próximo presidente del Gobierno central a que sepa "responder a las necesidades de paz y normalización política del pueblo vasco con inteligencia" y busque "los intereses de Estado y no los intereses partidistas".</following>
</context>
</instance>
<instance id="partido.n.32" docsrc="efe_9698_2000/03/12">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.32" senseid="partido.2"/>
<context>
<previous>El técnico del RCD Espanyol, Paco Flores, destacó el "magnífico encuentro" de su equipo, el "excelente" trabajo de Raúl Tamudo y los buenos resultados de terceros equipos que se ha combinado con la victoria españolista para alejarse de los puestos de descenso.</previous>
<target>"Todos los resultados nos han beneficiado, pero estoy especialmente contento porque la imagen del equipo ha sido muy buena", dijo Flores, quien no comparte la opinión con Víctor Fernández sobre que la expulsión de Coira resultara determinante en la suerte del <head>partido</head>.</target>
<following>Flores destacó la actuación de Tamudo, autor de dos tantos, y minimizó el hecho de que el Celta no haya contado con ocho jugadores de su plantilla, porque "disponen de una plantilla de gran nivel".</following>
</context>
</instance>
<instance id="partido.n.33" docsrc="efe_10551_2000/03/13">

<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.33" senseid="partido.2"/>
<context>
<previous></previous>
<target>Los jugadores del Atlético de Madrid Carlos Aguilera y el yugoslavo Veljko Paunovic, que sufrieron sendas lesiones en el <head>partido</head> del domingo en Valladolid, serán bajas para el derbi del próximo sábado contra el Real Madrid, al igual que el sancionado Juan Carlos Valerón.</target>
<following>Según informó hoy el médico del equipo, José María Villalón, Aguilera sufre un esguince cervical que le mantendrá alejado de la competición durante dos semanas y que le obligará a guardar reposo en los próximos días con un collarín.</following>
</context>
</instance>
<instance id="partido.n.34" docsrc="efe_11261_2000/03/14">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="ANPA" code="POL:POLITICA,ELECCIONES"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.34" senseid="partido.1"/>
<context>
<previous>Recordó que los partidos constitucionalistas en el País Vasco han conseguido once diputados frente a los ocho de los partidos nacionalistas y consideró que la polarización que, a su juicio, se está dando en la vida política en los últimos tiempos "ha beneficiado" al PNV y al PP.</previous>
<target>Redondo anunció que ante esta situación el PSE-EE remarcará su "proyecto autónomo", sin estar mirando a ambos <head>partidos</head> para establecer su estrategia, y en referencia a los populares hizo un "emplazamiento para que estén a la altura de las circunstancias y abandonen discursos vacíos de contenido".</target>
<following>Hizo referencia a los resultados obtenidos por el PSOE en el conjunto de España y dijo que su primera consecuencia directa, la dimisión del secretario general, Joaquín Almunia, "aunque tal vez no haya sido políticamente correcta, engrandece a la persona y pone al partido en una encrucijada que requiere tranquilidad".</following>
</context>
</instance>
<instance id="partido.n.35" docsrc="efe_12422_2000/03/15">
<cat scheme="ANPA" code="POL:POLITICA,ELECCIONES"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.35" senseid="partido.1"/>
<context>
<previous>En su opinión, debe abrirse una reflexión sobre la situación de "personas contratadas a dedo" para obras de emergencia en la Cañada de Hidum o en cooperativas "que sorprendentemente bajan su rendimiento en campaña electoral porque algunas trabajan para las campañas de determinadas formaciones políticas".</previous>
<target>Denunció que en esos colegios "se vota a ciegas a determinados <head>partidos</head>" y reivindicó el derecho a expresarse en las urnas "en libertad, con capacidad de decisión y no condicionado por un puesto de trabajo".</target>
<following>Velázquez dijo que "aún estamos a tiempo de que en Melilla se pueda votar con mayor grado de libertad" y abogó por formar un nuevo Gobierno "sin ataduras, que no precise del clientelismo político" y que con su gestión y control interno pueda llevar la "madurez" a los citados barrios</following>
</context>
</instance>
<instance id="partido.n.36" docsrc="efe_13566_2000/03/16">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.36" senseid="partido.1"/>
<context>
<previous>A los "barones" del PSOE se les ha encargado ahora la formación de la comisión gestora que deberá dirigir al partido hasta el Congreso de julio próximo, en el que los socialistas elegirán al sustituto de Almunia, quien la misma noche electoral anunció su renuncia tras tres años en el cargo.</previous>
<target>La dimisión de Almunia, que fue aplaudida unánimemente por sus correligionarios como un gesto de honradez, imprescindible además para acometer la renovación del <head>partido</head>, ha desatado una oleada de renuncias en el seno de la dirección.</target>
<following>En los últimos días han renunciado todos los miembros de la Ejecutiva del partido, así como las direcciones locales de Murcia y Burgos, mientras que el líder del PSOE en Cataluña, Narcís Serra, adelantó su inminente retirada del cargo, y hoy mismo dimitió la dirección local de Ourense -"para dar ejemplo"- y el secretario general del PSOE en Aragón, Isidoro Esteban.</following>

```
        </context>
</instance>
<instance id="partido.n.37" docsrc="efe_14153_2000/03/17">
<cat scheme="ANPA" code="POL:POLITICA,EXTERIOR"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.37" senseid="partido.1"/>
        <context>
        <previous>"El problema es la naturaleza del partido FPOE", afirmó Guterres, refiriéndose a la formación de
        ultraderecha presidida hasta hace poco por Joerg Haider.</previous>
        <target>"O bien el Gobierno cambia de <head>partido</head> o el partido cambia de naturaleza", recalcó el primer
        ministro portugués.</target>
        <following>Los catorce socios comunitarios de Austría decidieron el boicot político de Viena en respuesta a la
        alianza de los conservadores austríacos con el FPOE para formar Gobierno, liderado por el canciller Wolfgang
        Schuessel.</following>
        </context>
</instance>
<instance id="partido.n.38" docsrc="efe_14517_2000/03/17">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.38" senseid="partido.2"/>
        <context>
        <previous>Por su parte, Luis Oliver, apoderado del Xerez Deportivo, club del Grupo IV de la Segunda División B
        que utiliza habitualmente el Municipal de Chapín, señaló que esta entidad recibirá con los abrazos abiertos al Sevilla
        FC y a sus seguidores.</previous>
        <target>Oliver señaló que el estadio jerezano es un recinto muy moderno con cerca de veinte mil localidades sentadas
        y que cree que no todos los abonados sevillistas se desplazarán para el <head>partido</head>, por lo que confía en
        que no existan problemas para dar ubicar a los espectadores</target>
        <following></following>
        </context>
</instance>
<instance id="partido.n.39" docsrc="efe_17070_2000/03/21">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.39" senseid="partido.2"/>
        <context>
        <previous></previous>
        <target>Las selecciones nacionales de fútbol sub`23 de México y la mayor de Venezuela se verán las caras este
        miércoles en la ciudad mexicana de Villahermosa, en un <head>partido</head> amistoso que será el último antes de
        que ambas jueguen eliminatorias olímpicas y mundialistas, respectivamente.</target>
        <following>Venezuela empezará la semana próxima su participación en la eliminatoria suramericana para el Mundial
        del 2002 contra Ecuador y el equipo sub`23 de México tiene su mente puesta en las selecciones de Jamaica, Costa
        Rica y Honduras, sus rivales del preolímpico que se jugará en abril próximo en la ciudad mexicana de
        Guadalajara.</following>
        </context>
</instance>
<instance id="partido.n.40" docsrc="efe_18886_2000/03/23">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.40" senseid="partido.2"/>
        <context>
        <previous>Tras el entrenamiento que hoy tuvo la plantilla deportivista en las instalaciones de Acea da Ama, Pauleta
        insistió ante los periodistas en que no hizo "nada para merecer una sanción de dos partidos" y se mostró convencido
        de que eso mismo "lo hubiera hecho un jugador del Real Madrid o del Barcelona no le habrían
        castigado".</previous>
        <target>El jugador portugués expresó no obstante su confianza en que el Deportivo no precisará de sus servicios para
        obtener un resultado satisfactorio en el <head>partido</head> de Liga que el próximo domingo, a partir de las siete y
        media de la tarde, disputará en Riazor ante el Oviedo.</target>
        <following>Respecto a las opciones que su equipo puede tener para conseguir su primer título de Liga, Pauleta dijo:
        "nos lo merecemos y tenemos las condiciones necesarias para conseguirlo".</following>
        </context>
</instance>
```

<instance id="partido.n.41" docsrc="efe_19563_2000/03/24">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.41" senseid="partido.1"/>
<context>
<previous>La presidenta del PDNI, Cristina Almeida, calificó hoy a Manuel Chaves como "un hombre capaz de aunar voluntades y una persona que tiene una representatividad dentro del PSOE gobernando Andalucía, una de las comunidades más amplias".</previous>
<target>"Deben serenarse las cosas en el PSOE para poder reflexionar de cara a la sociedad y seguir trabajando políticamente", indicó Almeida en relación a la decisión del comité federal socialista de crear una gestora encabezada por Chaves, que dirigirá el <head>partido</head> hasta el congreso ordinario de julio. </target>
<following>En declaraciones a Efe en el aeropuerto de Barajas antes de embarcar en un vuelo con destino a La Coruña para inaugurar la casa de la mujer de esta ciudad, la presidenta del PDNI y del Grupo PSOE-Progresistas de la Asamblea de Madrid opinó que en el PSOE debe producirse una reflexión colectiva hacia la sociedad "que debe alcanzar algo más que el cuestionamiento interno".</following>
</context>
</instance>
<instance id="partido.n.42" docsrc="efe_22018_2000/03/27">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.42" senseid="partido.2"/>
<context>
<previous>El reporte médico señaló que Cagua necesitaba 48 horas de descanso para volver a la actividad y eso incidió para convocar a Reasco.</previous>
<target>"Estoy motivado por la convocatoria y muy tranquilo. Espero hacer un buen trabajo para convencer al técnico de que debo estar entre los titulares en el <head>partido</head> frente a Venezuela", expresó hoy Reasco a la prensa local.</target>
<following>Cagua, de 24 años, permanece concentrado con sus compañeros y no descarta la posibilidad de recuperarse y jugar el choque eliminatorio, que se disputará el miércoles en la capital ecuatoriana</following>
</context>
</instance>
<instance id="partido.n.43" docsrc="efe_22755_2000/03/28">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.43" senseid="partido.2"/>
<context>
<previous>"Clubes importantes est aban interesados, pero el Valencia siempre fue mi primera opción. Lo he visto por televisión y me ha impresionado. El Valencia es de los mejores clubes españoles y no podía desaprovechar esta oportunidad", indica al joven punta nórdico, quien pertenece actualmente al Rosenborg de Trondheim.</previous>
<target>Carew, que se encuentra en Bellinzona para disputar un <head>partido</head> amistoso con la selección noruega sub`21 contra Suiza, nació el 5 de septiembre de 1979 y reconoció que ya solamente queda la rúbrica de la documentación.</target>
<following>"No he firmado todavía el contrato, pero todas las condiciones están pactadas", afirmó el noruego, cuyo fichaje le costará al Valencia entre 8,4 y 9,6 millones de dólares, según fuentes cercanas al club noruego.</following>
</context>
</instance>
<instance id="partido.n.44" docsrc="efe_25423_2000/03/31">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.44" senseid="partido.2"/>
<context>
<previous>El jugador de la cantera zaragocista destacó la sorpresa que supone ver al Betis en la zona baja de la clasificación. "Tiene una plantilla muy buena y su situación no es conforme con su potencial deportivo, por lo que serán dos problemas que tengamos que resolver", señaló.</previous>
<target>El defensa del Real Zaragoza no cree que las tensiones que ha vivido el conjunto sevillano, reflejadas ampliamente en los medios de comunicación, vayan a suponer una merma en la capacidad verdiblanca porque saldrá "a por todas" para ganar el <head>partido</head> y sumar tres puntos que alivien su difícil situación en la clasificación</target>
<following></following>
</context>
</instance>

<instance id="partido.n.45" docsrc="efe_157_2000/04/01">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOLSALA"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.45" senseid="partido.2"/>
<context>
<previous>Habrá dos grupos de tres equipos, que serán sorteados próximamente, y los dos primeros de cada uno de ellos disputarán las semifinales.</previous>
<target>El torneo se jugará con las reglas de la FIFA y durante la primera fase se disputarán dos partidos cada día, el 26, 27 y 28 de abril, a las 18.30 y 20.30 horas. Las semifinales están programadas para el domingo 30 de abril a las 12.00 y 17.00, mientras que el <head>partido</head> por el tercer puesto y la final se jugarán el lunes 1 de mayo a las 10.00 y 12.00 horas, respectivamente.</target>
<following>La anécdota del acto de presentación del Campeonato de Europa de Clubes la protagonizó Miguel Angel de Vicente, que al sentarse en la silla que le había sido reservada, en un pequeño estrado, sufrió una espectacular caída y tiró el decorado que había detrás.</following>
</context>
</instance>
<instance id="partido.n.46" docsrc="efe_344_2000/04/01">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONMANO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.46" senseid="partido.2"/>
<context>
<previous>Comentario:</previous>
<target>Caja España Ademar y FC Barcelona empataron un <head>partido</head> que estuvo marcado por la igualdad y en que se dio el mismo resultado que la pasada temporada en el de la liga regular (29-29).</target>
<following>El Ademar, a pesar de las bajas de Entrerríos y Pérez Canca, dio la cara y aguantó a un Barcelona con notables ausencias, ya que la aportación de Juanín, con su velocidad, del central Esquer y del pivote Juancho, que aprovechó la ausencia de Chepkin, permitió al equipo leonés llegar al descanso con un 16-16.</following>
</context>
</instance>
<instance id="partido.n.47" docsrc="efe_1012_2000/04/03">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.47" senseid="partido.2"/>
<context>
<previous>Los hondureños fueron los primeros en amenazar con abrir el marcador en el minuto 35 cuando Eduardo Bennett dejó en el camino al defensa panameño Antony Torres, pero no pudo vencer al portero Ricardo James, que logró desviar un potente disparo del delantero hondureño, en una de las tantas veces que salvó al equipo panameño.</previous>
<target>Después de esta jugada Panamá perdió a su mejor defensa Antony Torres, que tuvo que salir del <head>partido</head> tras sufrir una fractura en dos costillas como resultado de un choque con el delantero hondureño Cárcamo.</target>
<following>A partir de ese momento, Panamá perdió organización en el campo y los hondureños sacaron provecho de esto con algunos contragolpes que pusieron en aprietos a la debilitada defensa panameña, mientras que el sustituto de Torres, Ubaldo Guardia, aunque lo tuvo difícil, hizo un buen trabajo.</following>
</context>
</instance>
<instance id="partido.n.48" docsrc="efe_1560_2000/04/03">
<cat scheme="ANPA" code="DEP:DEPORTES,TENIS"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.48" senseid="partido.2"/>
<context>
<previous>John Newcombe, capitán del equipo, ha informado de que Wayne Arthurs será el sustituto de Philippoussis, aunque es casi seguro que no será él quien juegue uno de los partidos individuales, sino Patrick Rafter.</previous>
<target>Rafter, quien ya se ha recuperado de su lesión en un hombro, dijo que está listo para jugar los <head>partidos</head> individuales.</target>
<following>Newcombe anunciará mañana la escuadra definitiva para enfrentarse a los alemanes</following>
</context>
</instance>
<instance id="partido.n.49" docsrc="efe_3550_2000/04/05">
<cat scheme="ANPA" code="POL:POLITICA,PARLAMENTO"/>

<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.49" senseid="partido.1"/>
<context>
<previous>La decisión de Mauritania fue muy criticada por varios países árabes, especialmente por Irak, país con el que Nuakchot rompió sus relaciones diplomáticas poco después, el 4 de noviembre del año pasado.</previous>
<target>Los <head>partidos</head> mauritanos de oposición también condenaron la decisión del gobierno y convocaron a finales de octubre del año pasado varias manifestaciones de protesta en Nuakchot</target>
<following></following>
</context>
</instance>
<instance id="partido.n.50" docsrc="efe_5253_2000/04/07">
<cat scheme="ANPA" code="POL:POLITICA,GOBIERNO"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.50" senseid="partido.1"/>
<context>
<previous></previous>
<target>La capilla ardiente con los restos mortales del ex presidente tunecino Habib Burguiba quedó instalada hoy en la sede central del <head>partido</head> gubernamental de la Agrupación Constitucional Democrática (RCD), en lo alto de la medina de Túnez.</target>
<following>El féretro llegó a Túnez procedente de Monastir, a 160 kilómetros de la capital del país, donde ayer falleció el que se considera como el "padre" de la independencia tunecina.</following>
</context>
</instance>
<instance id="partido.n.51" docsrc="efe_5719_2000/04/07">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.51" senseid="partido.2"/>
<context>
<previous></previous>
<target>El Hamburgo continúa su persecución sobre el Bayer Leverkusen y el Bayern Múnich, al derrotar hoy por 1-0 al Hansa Rostock, en <head>partido</head> adelantado a la vigésima octava jornada de la Liga alemana de fútbol.</target>
<following>El Hamburgo está ahora a un punto del Bayern y a tres del Leverkusen, si bien éstos todavía tienen que jugar sus respectivos encuentros.</following>
</context>
</instance>
<instance id="partido.n.52" docsrc="efe_5923_2000/04/08">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONCESTO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.52" senseid="partido.2"/>
<context>
<previous></previous>
<target>El Caja San Fernando afronta el <head>partido</head> de la trigésima segunda jornada de la Liga ACB con la intención de conseguir una victoria, lo que le aseguraría estar clasificado a la conclusión de la fase entre los cuatro primeros y disputar las eliminatorias por el título con el factor cancha a su favor.</target>
<following>El equipo de Javier Imbroda se encontrará en Badalona con un Bruguer Joventut que a falta de tres jornadas para la conclusión de la fase ya ha perdido todas las opciones de estar clasificado entre los ocho primeros, por lo que la motivación del conjunto catalán no pasa por sus mejores momentos.</following>
</context>
</instance>
<instance id="partido.n.53" docsrc="efe_7380_2000/04/10">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.53" senseid="partido.2"/>
<context>
<previous>El presidente del Leeds United, Peter Ridsdale, acusó hoy al equipo turco de Galatasaray de "falta de respeto" y "oportunismo" tras el asesinato de dos seguidores del equipo inglés el miércoles en Estambul, la víspera del partido de ida de las semifinales de la Copa de la UEFA.</previous>
<target>Ridsdale ha prohibido la entrada de aficionados del Galatasaray para el <head>partido</head> de vuelta como medida de seguridad, y el equipo turco ha pedido que se celebre en campo neutral para poder contar con el apoyo de su hinchada.</target>

<following>"Dos personas han sido asesinadas. He visto las heridas y no quiero volver a ver nada parecido en mi vida", dijo Ridsdale.</following>
</context>
</instance>
<instance id="partido.n.54" docsrc="efe_7555_2000/04/10">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="ANPA" code="POL:POLITICA,REGIONES-AUTONOMIAS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.54" senseid="partido.1"/>
<context>
<previous>El lehendakari, Juan José Ibarretxe, anunció hoy que comparecerá mañana ante los medios de comunicación para dar a conocer el análisis que, sobre esta situación, efectuará el Consejo de Gobierno en su habitual reunión semanal.</previous>
<target>El lehendakari solicitó el pasado sábado a través de sus colaboradores "calma" a los <head>partidos</head> ante el nuevo escenario político.</target>
<following>Entre las formaciones nacionalistas, la única que hizo declaraciones fue la diputada de Eusko Alkartasuna Begoña Lasagabaster, quien consideró que la "delicada" situación del Gobierno Vasco ha sido originada por la "labor obstruccionista" en el Parlamento autonómico del PP, el PSE y EH.</following>
</context>
</instance>
<instance id="partido.n.55" docsrc="efe_10411_2000/04/13">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.55" senseid="partido.2"/>
<context>
<previous>Julián Rubio, entrenador del Albacete Balompié, con vistas al choque del próximo domingo frente al Córdoba, podría sorprender al dar entrada en el once inicial al veterano Sala a tenor de lo visto en el partido-ensayo realizado hoy en el estadio Carlos Belmonte.</previous>
<target>Sala, en el caso de jugar, retornaría después de no jugar, por decisión técnica, desde el 6 de febrero, en la jornada vigésima quinta, contra el Levante (2-0), partido en el que entró en el minuto 83 en sustitución de Geli y coincidió con su <head>partido</head> 500 como profesional.</target>
<following>Las sorpresas podrían ampliarse, pues de ser fiel a lo hecho hoy el ariete Sabas podría jugar de salida, pero en una posición alejada de la suya natural, ya que como ocurrió en Compostela el técnico manchego probó al delantero madrileño como volante izquierdo.</following>
</context>
</instance>
<instance id="partido.n.56" docsrc="efe_11023_2000/04/14">
<cat scheme="ANPA" code="POL:POLITICA,GOBIERNO"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.56" senseid="partido.1"/>
<context>
<previous>El Gobierno intentará ponerse de acuerdo sobre un candidato común después de que el Parlamento rechazara su propuesta para permitir la reelección de Demirel pero, según círculos políticos, las posibilidades de llegar a un consenso son bastante remotas.</previous>
<target>El Partido de la Madre Patria (ANAP) parece decidido a nombrar a su líder, Mesut Yilmaz, como candidato presidencial, pero sus compañeros de coalición, la Izquierda Democrática (DSP) de Ecevit, y el Partido de Acción Nacional (MHP) del vice primer ministro Devlet Bahceli, se oponen a nombrar a un jefe de <head>partido</head>.</target>
<following>El ministro de Trabajo, Yasar Okuyan, del ANAP, dijo que Yilmaz no ha anunciado todavía su candidatura, y subrayó que "lo que haga Yilmaz dependerá de las decisiones que tomen los líderes de la coalición en la reunión de hoy".</following>
</context>
</instance>
<instance id="partido.n.57" docsrc="efe_14032_2000/04/18">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.57" senseid="partido.2"/>
<context>
<previous>"Defendiendo bien también se ganan los partidos. Pero anticipar cómo se va a desarrollar el encuentro a priori no es fácil. Podríamos controlar el juego y perder. En la ida, por ejemplo, llevamos la iniciativa y no ganamos. Es una incógnita aventurar ahora mismo cómo va a discurrir el partido", señaló Del Bosque.</previous>

<target>Sobre el estímulo extra que van a tener los jugadores por enfrentarse a un Manchester en un estadio que reúne todos los condicionantes ideales para jugar un <head>partido</head> de fútbol, Del Bosque explicó que toda la plantilla está muy motivada.</target>
<following>Y al ser cuestionado sobre si era un día ideal para Raúl, que podría haberse dosificado para estar a tope mañana, Del Bosque fue tajante: "Este partido es un estímulo para Raúl y para todos. Raúl pelea siempre. Se vacía en todos sus compromisos. En Liga y en Europa", apostilló Del Bosque</following>
</context>
</instance>
<instance id="partido.n.58" docsrc="efe_15333_2000/04/20">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.58" senseid="partido.2"/>
<context>
<previous>Un grupo de hinchas del Lanús recibió hoy con una lluvia de flores al portero paraguayo José Luis Chilavert, del Vélez Sarsfield, en la reanudación de un partido de la Liga argentina de primera división que se había suspendido en marzo pasado.</previous>
<target>El simpático recibimiento se produjo momentos antes de comenzar el <head>partido</head> entre el Lanús y el Vélez Sarsfield, de la quinta jornada del torneo Clausura, que había sido suspendido el pasado 12 de marzo cuando se llevaban jugados 35 segundos, a causa de un petardo arrojado por los hinchas locales.</target>
<following>El artefacto explosivo cayó muy cerca de Chilavert, quien fue retirado del campo de juego visiblemente aturdido, lo que obligó al árbitro Sergio Pezzotta a suspender el partido.</following>
</context>
</instance>
<instance id="partido.n.59" docsrc="efe_15430_2000/04/20">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.59" senseid="partido.1"/>
<context>
<previous>El alcalde de San Sebastián defendió la necesidad de superar "los frentes" desde la defensa "del diálogo y la distensión, el respeto escrupuloso de los derechos humanos, tanto desde la exigencia del respeto del derecho a la vida como del fin de la violencia callejera, un cambio en la política penitenciaria, y la aceptación de los principios democráticos del Estado de Derecho y las reglas de juego democráticas".</previous>
<target>Elorza admitió que la contradicción existente entre la línea oficial de su <head>partido</head> y su posición "minoritaria" hace "insostenible" su presencia en la dirección del PSE/EE, al tiempo que dijo echar en falta "un debate en profundidad" para redefinir la estrategia de su formación y que ésta le permita "salir del bocadillo entre el nacionalismo y la derecha española".</target>
<following>A su entender, la opinión pública comparte el principio de que, fuera de toda intimidación, "la voluntad democrática de los vascos libremente expresada será siempre respetada", por lo que "nadie debería excluir" la posibilidad de "una consulta popular".</following>
</context>
</instance>
<instance id="partido.n.60" docsrc="efe_17287_2000/04/24">
<cat scheme="ANPA" code="ECO:ECONOMIA,SECTORES-EMPRESAS,-"/>
<cat scheme="ANPA" code="ECO:ECONOMIA,MACROECONOMIA,FISCALIDAD"/>
<cat scheme="IPTC" code="04000000"/>
<answer instance="partido.n.60" senseid="partido.1"/>
<context>
<previous>Los chaebol emitieron la pasada semana un comunicado conjunto en el que exigían al gobierno que se hiciera a un lado y dejara de entrometerse en las reorganizaciones de estos grupos industriales.</previous>
<target>Este mensaje, aireado pocos días después de que la oposición venciera al <head>partido</head> del presidente Kim Dae Jung en las elecciones parlamentarias del pasado día 13, causó una importante reacción entre los grupos cívicos, que exigen a las autoridades acabar con la excesiva concentración de poder en los chaebol.</target>
<following>Entre los personajes que serán investigados estarán el hijo del presidente del grupo Samsung, el recién nombrado presidente del grupo Hyundai y el hijo-sucesor en la presidencia del grupo SK.</following>
</context>
</instance>
<instance id="partido.n.61" docsrc="efe_18420_2000/04/25">
<cat scheme="ANPA" code="TRI:JUSTICIA-INTERIOR-SUCESOS,JUSTICIA"/>
<cat scheme="IPTC" code="02000000"/>
<cat scheme="IPTC" code="03000000"/>
<answer instance="partido.n.61" senseid="partido.1"/>

<context>
<previous>Pero la pretensión del magistrado español Baltasar Garzón de sentar en el banquillo de los acusados a Pinochet por genocidio, terrorismo y torturas se vio truncada por la decisión del ministro británico del Interior, Jack Straw, quien en consideración al estado de salud del ex gobernante, interrumpió el proceso de extradición.</previous>
<target>Mientras tanto, en Chile se multiplicaron las querellas contra el ex dictador presentadas por los parientes de los desaparecidos y ejecutados, ex presos políticos, sindicalistas, colectivos profesionales y <head>partidos</head> políticos.</target>
<following>A esta lista se sumaron recientemente el primer ciudadano extranjero, el médico ecuatoriano Alfonso García Franco, y el Partido Socialista de Chile, la formación en la que milita el presidente chileno, Ricardo Lagos.</following>
</context>
</instance>
<instance id="partido.n.62" docsrc="efe_19884_2000/04/27">
<cat scheme="ANPA" code="POL:POLITICA,PARLAMENTO"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.62" senseid="partido.1"/>
<context>
<previous>A su juicio, con el modelo vigente "da la impresión, a veces, de que estamos en una almoneda donde se compra y se vende, se regatea y se consiguen dividendos a cambios de apoyos parlamentarios, pero España es algo más que un mercado de la aritmética el gobierno", recalcó.</previous>
<target>Por otra parte, se refirió a su candidatura a la secretaría general del PSOE, sobre lo que señaló que "no deseo ser ningún problema para mi <head>partido</head> y no estaré si creo problemas, pero si puedo ayudar a resolver algún problema estaré en el puesto que sea, el primero o el último, esa es mi disposición".</target>
<following>Agregó que su proyecto lo plantea "en primera persona del plural porque en el PSOE hay que acabar con el personalismo; el PSOE no tiene ombligo y, si lo tuviese, no soy yo", tras lo que afirmó que el objetivo es que en el 2004 haya en la Moncloa un presidente que sea "progresista" porque ocho años del Gobierno de la derecha "se nos van a antojar ocho siglos a la mitad de los españoles y a algunos más".</following>
</context>
</instance>
<instance id="partido.n.63" docsrc="efe_21975_2000/04/29">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONCESTO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.63" senseid="partido.2"/>
<context>
<previous>El pívot Alonzo Mourning, que estuvo la mayoría del segundo tiempo sentado en el banquillo contribuyó con 12 puntos y nueve rebotes y el alero Jamal Mashburn, que no fue factor, logró 11 tantos.</previous>
<target>"Weatherspoon surgió como el jugador decisivo del <head>partido</head> porque fue el que mejor estuvo a la hora de tirar a canasta y es algo que necesitábamos no sólo para este partido sino para el resto de la competición", destacó Riley.</target>
<following>Los Pistons volvieron a demostrar que sin Hill su único jugador competitivo es el alero Jerry Stackhouse, que cada día incrementa más su cotización, y lo demostró al conseguir 25 puntos, que no fueron suficientes para darle a su equipo la posibilidad de mantenerse en la competición.</following>
</context>
</instance>
<instance id="partido.n.64" docsrc="efe_1276_2000/05/03">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONCESTO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.64" senseid="partido.2"/>
<context>
<previous>El escolta Hardaway, en su mejor momento de juego, y Rogers, en su línea de producción constante, aportaron 23 puntos cada uno para dejar a los Suns por primera vez en las semifinales de la Federación Oeste desde 1995 y convertir a los Spurs en el primer equipo campeón de liga que no repite título desde que fueron eliminados los Celtics de Boston en 1986.</previous>
<target>"Me he sentido muy bien y como me dijeron los médicos he jugado en plenitud sin sentir ningún tipo de molestias o problemas físicos", declaró Kidd. "Ha sido muy emocionante el poder ayudar a mis compañeros en el cuarto <head>partido</head> para que estemos en las semifinales".</target>
<following>Los Suns lograron un 46,3 por ciento en los tiros de campo, 63,2 desde la línea de personal y 25,0 en los triples (3 de 12), comparados al 37,0; 61,5 y sin ningún triple, respectivamente, para los Spurs.</following>
</context>
</instance>

<instance id="partido.n.65" docsrc="efe_2450_2000/05/04">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.65" senseid="partido.2"/>
<context>
<previous>Quini, que llegó al CD Toledo la anterior temporada, procedente del Talavera CF, y a instancias del entonces técnico Gregorio Manzano, ahora en el Real Valladolid, ha anotado en la presente campaña solo tres goles, dos de penalti, cuando la campaña fue el máximo goleador del equipo con 16 dianas, cinco de pena máxima.</previous>
<target>En la actual temporada solo ha disputado seis partidos completos, aunque ha participado en otros doce, cinco en la segunda vuelta, para un total de 1.135 minutos, cuando la pasada campaña disputó 15 <head>partidos</head> completos y otros tantos parciales para un total de 2.263 minutos, el séptimo de la plantilla que jugó más minutos.</target>
<following>El año pasado recibió cinco cartulinas amarillas y en la presente campaña ha visto cuatro, dos en un mismo encuentro, ante el Lleida en la segunda vuelta, siendo expulsado en el minuto 40 de partido.</following>
</context>
</instance>
<instance id="partido.n.66" docsrc="efe_4156_2000/05/06">
<cat scheme="ANPA" code="TRI:JUSTICIA-INTERIOR-SUCESOS,TERRORISMO"/>
<cat scheme="IPTC" code="02000000"/>
<cat scheme="IPTC" code="03000000"/>
<cat scheme="IPTC" code="16000000"/>
<answer instance="partido.n.66" senseid="partido.1"/>
<context>
<previous></previous>
<target>El ministro británico para Irlanda del Norte, Peter Mandelson, calificó hoy de "muy positivo" el comunicado del Ejército Republicano Irlandés (IRA), y pidió a los <head>partidos</head> que lo estudien con "seriedad".</target>
<following>Mandelson destacó en Belfast que la nota es "muy significativa" y ofrece la oportunidad de construir en Irlanda del Norte un futuro libre de violencia y armas.</following>
</context>
</instance>
<instance id="partido.n.67" docsrc="efe_5580_2000/05/08">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONMANO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.67" senseid="partido.2"/>
<context>
<previous>Los árbitros suecos Patrick Hakasson y Mats Nilson serán los encargados de dirigir el partido de ida de la final de la Recopa de Europa que disputarán, el sábado 20 de mayo, el Krasnodar y el Milar en la cancha del equipo ruso.</previous>
<target>El primero de los <head>partidos</head> de esta final se disputará a las 15.00 horas (13.00 hora española), mientras que el encuentro de vuelta, que se jugará en el polideportivo municipal de L`Eliana, se disputará el domingo 28 de mayo a las 12.30 horas.</target>
<following>Para este segundo partido la Federación Europea de Balonmano (EHF) también ha dado a conocer los árbitros, que serán los alemanes Frank Lemme y Bernd Ullrich</following>
</context>
</instance>
<instance id="partido.n.68" docsrc="efe_7717_2000/05/10">
<cat scheme="ANPA" code="DEP:DEPORTES,BALONCESTO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.68" senseid="partido.2"/>
<context>
<previous>Pepe Rodríguez, técnico del conjunto melillense, ya tiene decidido el cinco inicial que pondrá en liza ante los menorquines, que será el formado por el base Ron Rutland, el escolta Silva, el alero Paco Martín, el ala pívot David Doblas y el pívot Cedric Moore.</previous>
<target>El entrenador del Melilla considera que su equipo está capacitado para pasar a semifinales, aunque la eliminatoria será "muy dura" a tenor de los encuentros disputados ante el Menorca en la fase regular, que se resolvieron a favor de su equipo en los últimos minutos de cada <head>partido</head>.</target>
<following>Rodríguez aseguró que el Menorca Basquet llegará a Melilla con la moral muy alta tras sus tres victorias en octavos de final ante Los Barrios. "Es un equipo muy compensado, con un juego interior muy fuerte en el que

destaca su pareja de extranjeros, Montgomery y Delany, y por fuera cuenta con grandes artilleros como el veterano Patricio Reynés o el alero Josep Pancreu", subrayó</following>
</context>
</instance>
<instance id="partido.n.69" docsrc="efe_8099_2000/05/11">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.69" senseid="partido.1"/>
<context>
<previous></previous>
<target>El portavoz del PP del País Vasco, Leopoldo Barreda, exigió hoy al PNV que "rediseñe" el proyecto político que aprobó en su Asamblea General, equivalente a un congreso en otro <head>partido</head>, el pasado mes de enero.</target>
<following>En conferencia de prensa celebrada en Bilbao para presentar dos iniciativas legislativas, Barreda dijo que con el PNV no hay "un problema de estrategia, sino del proyecto político de fondo que pactó con ETA, excluyente, por la independencia, y que declaraba enemigos del pueblo vasco a los que no votan nacionalista".</following>
</context>
</instance>
<instance id="partido.n.70" docsrc="efe_11154_2000/05/15">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.70" senseid="partido.2"/>
<context>
<previous>Tantos decisivos para el Lazio en partidos importantes. En los que el equipo estaba sin las ideas claras ante la puerta rival, pero que se le pusieron muy favorables merced a la fuerza del "Cholo" (como se le conoce) que viniendo desde atrás sorprendió a los defensas rivales.</previous>
<target>Al final ha conseguido una nada despreciable cifra de cinco tantos, que le sitúan tras Verón como el goleador más decisivo, pues han traído consigo 12 puntos. Es también el jugador lacial que durante la temporada liguera más veces a entrado durante los <head>partidos</head> para sustituir a compañeros, con 13 veces.</target>
<following>Almeyda, por su parte, pese a perderse algunos partidos por lesión, ha sabido poner el "filtro" a los rivales, además de apoyar en todo instante la labor de organización de sus compañeros de zona. Su fuerza, faltas tácticas y entrega, hicieron "rocoso" en el centro del campo a un equipo ofensivo. Es el que más tarjetas amarillas ha recibido del Lazio (8).</following>
</context>
</instance>
<instance id="partido.n.71" docsrc="efe_11191_2000/05/15">
<cat scheme="ANPA" code="POL:POLITICA,MUNICIPAL"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.71" senseid="partido.1"/>
<context>
<previous>La moción ha prosperado con la abstención del Partido Independiente Valle del Guadiaro (PIVG) y los votos en contra de los siete concejales del PSOE y del único de IU, partidos que gobernaban en minoría desde que el pasado 20 de marzo los tres ediles del PP abandonaran el equipo de Gobierno tras la disolución del GIL.</previous>
<target>La marcha de los tres concejales populares fue argumentada por la dirección provincial del
<head>partido</head> en que el paso, días antes, de los seis concejales del GIL al Grupo Mixto "dejaba sin fundamento" el pacto suscrito entre las tres formaciones para evitar que el Grupo Independiente Liberal accediera a la alcaldía de San Roque, una localidad de unos 22.000 habitantes.</target>
<following>El nuevo alcalde, que recibió el bastón de mando abucheado por medio centenar de vecinos que se agolpaban en el salón de plenos del Ayuntamiento, y ovacionado por otros tantos, achacó la moción de censura a "decisiones políticas y no personales".</following>
</context>
</instance>
<instance id="partido.n.72" docsrc="efe_11488_2000/05/15">
<cat scheme="ANPA" code="POL:POLITICA,GOBIERNO"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.72" senseid="partido.1"/>
<context>
<previous>Humberto de la Calle anunció que el Gobierno tomó la decisión de adelantarse a esta eventual negativa e iniciar un proceso de reflexión y de diálogo "con la mente amplia" sobre el texto de referéndum para mejorarlo.</previous>

<target>En cuanto a los demás elementos del referéndum, agregó que el Ejecutivo tiene la intención de mantener la esencia: temas electorales, lista única, cifra repartidora, voto obligatorio, inhabilitación perpetua, fortalecimiento de la pérdida de investidura, régimen interno de los <head>partidos</head>, transparencia en la financiación de las campañas.</target>
<following>El ministro del Interior reveló que tiene previsto entrevistarse en breve con Serpa y con la también ex candidata presidencial Noemí Sanín, así como con los integrantes del nuevo directorio conservador, con el líder sindical Luis Eduardo Garzón, con los movimientos de tipo social y con los congresistas.</following>
</context>
</instance>
<instance id="partido.n.73" docsrc="efe_15720_2000/05/20">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.73" senseid="partido.2"/>
<context>
<previous>Mañana, sábado, el líder del torneo, el Nacional, jugará a domicilio en Paysandú, con el equipo local en el segundo encuentro adelantado de la jornada.</previous>
<target>El domingo próximo se completará el calendario de la jornada con los <head>partidos</head> Tacuarembó-Bella Vista, Danubio-Maldonado, Frontera- Liverpool, Rocha-Rentistas, Villa Española-Huracán Buceo, Juventud-Peñarol y Racing-Defensor.</target>
<following>- Clasificación:</following>
</context>
</instance>
<instance id="partido.n.74" docsrc="efe_16468_2000/05/21">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.74" senseid="partido.1"/>
<context>
<previous>El secretario general del CDC, Pere Esteve, ha asegurado hoy que no se presentará a la reelección en el cargo, porque cree que el que ocupa debe ser "para el número dos" (del partido) y apostilló que su candidato es Artur Mas.</previous>
<target>El secretario general de CDC, Pere Esteve, ha anunciado hoy a la ejecutiva y al consejo nacional de su partido que no optará a la reelección para que Artur Mas asuma esa responsabilidad y sea el número dos de CDC, tras Jordi Pujol, lo que le situaría como líder del <head>partido</head> en el futuro.</target>
<following>La decisión, que, a su juicio, demuestra la sintonía políticacon Mas ante el congreso del partido previsto para el mes de noviembre, favorece las opciones del portavoz del gobierno catalán y conseller de Economía en la "pugna" que mantiene con el lider de UDC, Josep Antoni Duran i LLeida, por la sucesión de Jordi Pujol en el seno de CiU.</following>
</context>
</instance>
<instance id="partido.n.75" docsrc="efe_22148_2000/05/27">
<cat scheme="ANPA" code="SOC:SOCIEDAD-SALUD,SOCIEDAD"/>
<cat scheme="IPTC" code="07000000"/>
<cat scheme="IPTC" code="14000000"/>
<answer instance="partido.n.75" senseid="partido.1"/>
<context>
<previous>Esta presencia de mujeres se debe, en su opinión, a que la sociedad exige esos cambios y "es muy importante que podamos permitir y, a su vez, exigir que las mujeres no se vean obligadas a hacer un proceso de transformación para ser aceptadas en la sociedad masculina, sino impregnar la estructura política de planteamientos femeninos".</previous>
<target>Para Carmena, las cuotas en los <head>partidos</head> políticos son positivas como sistemas de promoción y de igualdad, aunque en la práctica sea necesario algún tipo de concreción o de corrección.</target>
<following>Las jornadas "Mujer y Sociedad" pretenden reivindicar la dignidad de la condición femenina y eliminar las desigualdades que existen en el ámbito social y laboral, de ahí que en las distintas sesiones de trabajo diversos especialistas traten materias relacionadas con la salud, la educación, la justicia, o la violencia doméstica</following>
</context>
</instance>
<instance id="partido.n.76" docsrc="efe_24454_2000/05/30">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.76" senseid="partido.2"/>
<context>

<previous>Lahore dejó claro que el estadio no se venderá salvo que las generaciones venideras decidan lo contrario y que el consejo de administración, que planteó en la última junta general la firma del convenio de uso con el Estadio Olímpico para jugar algunos partidos en él si no se obtenía liquidez, "nunca ha hablado de venderlo".</previous>
<target>El club abrió la semana pasada un plazo hasta el 9 de junio para que los aficionados puedan aportar dinero mediante un préstamo con garantía real de hasta 3.000 millones de pesetas para afrontar las deudas más urgentes, como aprobó la junta de accionistas para eludir el tener que jugar <head>partidos</head> en el Olímpico desde la próxima campaña.</target>
<following>El vicepresidente del Sevilla consideró que cumplir ese objetivo "va a ser difícil y complicado", ya que se necesita mucho dinero y hasta ahora han sido muy escasas las aportaciones hechas por el sevillismo en la primera semana de suscripción a dicho préstamo.</following>
</context>
</instance>
<instance id="partido.n.77" docsrc="efe_24495_2000/05/30">
<cat scheme="ANPA" code="POL:POLITICA,CONFLICTO"/>
<cat scheme="IPTC" code="16000000"/>
<answer instance="partido.n.77" senseid="partido.1"/>
<context>
<previous>El radical Partido Democrático Unionista del Ulster (DUP), contrario al acuerdo de paz de Viernes Santo del 10 de abril de 1998, decidirá esta noche si se incorpora o no al Gobierno de poder compartido en Irlanda del Norte.</previous>
<target>En declaraciones a la prensa, el diputado del DUP Peter Robinson no ha querido adelantar si su
<head>partido</head> ocupará los dos escaños que le corresponden en el Ejecutivo de la provincia, restablecido la pasada medianoche.</target>
<following>Londres suspendió el Gobierno compartido en Irlanda del Norte el pasado 11 de febrero para evitar que el Partido Unionista del Ulster (UUP, mayoritario entre la comunidad protestante norirlandesa) lo abandonara debido a que el Ejército Republicano Irlandés (IRA) no había empezado a desarmarse.</following>
</context>
</instance>
<instance id="partido.n.78" docsrc="efe_2422_2000/06/04">
<cat scheme="ANPA" code="DEP:DEPORTES,BEISBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.78" senseid="partido.2"/>
<context>
<previous>Después de dos "outs" en la octava, Mike Lowell y Kevin Millar dispararon sendos cuadrangulares para que los Marlins de Florida vencieran por 2-1 a los Azulejos de Toronto.</previous>
<target>El <head>partido</head> se destacó por el duelo de picheo de los abridores Chris Carpenter, de los Azulejos, y Ryan Dempster de los locales.</target>
<following>Carpenter colgó siete ceros antes de ser relevado por Billy Koch (3-1), que retiró a los primeros dos hombres en el octavo para luego servir los ofrecimientos a Lowell (8) y Millar (7) que dio la derrota al equipo canadiense.</following>
</context>
</instance>
<instance id="partido.n.79" docsrc="efe_4140_2000/06/06">
<cat scheme="ANPA" code="TRI:JUSTICIA-INTERIOR-SUCESOS,TERRORISMO"/>
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="02000000"/>
<cat scheme="IPTC" code="03000000"/>
<cat scheme="IPTC" code="11000000"/>
<cat scheme="IPTC" code="16000000"/>
<answer instance="partido.n.79" senseid="partido.1"/>
<context>
<previous></previous>
<target>El vicesecretario general de Unidad Alavesa y ex concejal del Ayuntamiento de Vitoria, Francisco Probanza, ha decidido abandonar Euskadi al figurar en varios documentos requisados a ETA, algunos de ellos recientes, confirmaron hoy fuentes de este <head>partido</head>.</target>
<following>Probanza, concejal en la pasada legislatura, aparecía en la documentación incautada por la Policía al comando "Basurde" de ETA que tenía previsto asesinarle, ya que contaba con información detallada sobre él.</following>
</context>
</instance>
<instance id="partido.n.80" docsrc="efe_5366_2000/06/07">

<cat scheme="ANPA" code="TRI:JUSTICIA-INTERIOR-SUCESOS,JUSTICIA"/>
<cat scheme="IPTC" code="02000000"/>
<cat scheme="IPTC" code="03000000"/>
<answer instance="partido.n.80" senseid="partido.1"/>
<context>
<previous>Pujades ha explicado que el sacerdote gerundense "trabajó durante la huelga gremial previa al golpe de Estado, y no sólo no participó, sino que descontó los días de huelga a los trabajadores que sí la siguieron".</previous>
<target>"Además -ha agregado Pujades-, todo el mundo sabía que era simpatizante de la Unidad Popular, aunque nunca se afilió a un <head>partido</head> político".</target>
<following>Según el relato del biógrafo, Joan Alsina fue a trabajar el martes 19 de septiembre, y hacia las dos de la tarde fue detenido y conducido al internado Barros Arana, que era utilizado como lugar provisional de reclusión.</following>
</context>
</instance>
<instance id="partido.n.81" docsrc="efe_6503_2000/06/08">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.81" senseid="partido.2"/>
<context>
<previous>El seleccionador Erich Ribbeck ha querido tranquilizar a la afición ante los temores de que la defensa germana sea la línea más endeble.</previous>
<target>"Todo lo que sé es que el lunes saltaremos al campo con 11 jugadores", comentó Ribbeck al llegar a Holanda con respecto a su debut ante Rumanía en Lieja en <head>partido</head> correspondiente al grupo A.</target>
<following>Vaals, la pequeña localidad elegida como cuartel general de los alemanes, está enclavada en el triángulo que forman las fronteras entre Bélgica, Holanda y Alemania. </following>
</context>
</instance>
<instance id="partido.n.82" docsrc="efe_9501_2000/06/12">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.82" senseid="partido.2"/>
<context>
<previous>Mondragón tiene contrato con el equipo argentino hasta que finalice el actual torneo Clausura, del que faltan disputarse cinco jornadas.</previous>
<target>El futbolista jugará con el Independiente el viernes ante el Instituto de Córdoba el <head>partido</head> adelantado de la decimocuarta jornada y los dirigentes del club local presumen que podría ser su último encuentro en Argentina antes de incorporarse al Metz.</target>
<following>El técnico del equipo francés, Joel Muller, había impuesto como condición del fichaje de Mondragón que se incorporase a la plantilla antes de finales de este mes, debido a que tiene previsto comenzar la pretemporada el día 23 próximo.</following>
</context>
</instance>
<instance id="partido.n.83" docsrc="efe_10738_2000/06/14">
<cat scheme="ANPA" code="DEP:DEPORTES,POLIDEPORTIVO"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.83" senseid="partido.2"/>
<context>
<previous>3-3. EL ORGULLO RESCATA A YUGOSLAVIA DEL DESASTRE</previous>
<target>Charleroi (Bélgica), (EFE).- El orgullo y el amor propio de Yugoslavia le permitió salvarse del desastre en el duelo balcánico de la Eurocopa (3-3), un <head>partido</head> en toda la extensión de la palabra en el que Eslovenia, debutante en estas lides, acusó su falta de experiencia, y desperdició en seis minutos una ventaja de 3-0. (DH0110)</target>
<following>FUTBOL-BRASIL</following>
</context>
</instance>
<instance id="partido.n.84" docsrc="efe_11939_2000/06/15">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.84" senseid="partido.2"/>
<context>

<previous>El centrocampista del Juventus, con su gran visión de juego y su facilidad para manejar el balón, aspira a reeditar el éxito personal y colectivo que ya tuvo en el Mundial de Francia.</previous>
<target>Junto con estos tres jugadores y el citado Rui Costa, pocos jugadores más han destacado en una Eurocopa con un pobre, en general, nivel de fútbol y en la que algunos de los grandes han causado una pobre impresión en sus primeros <head>partidos</head>, como Holanda (pese a ganar a la República Checa), España, Inglaterra y Alemania.</target>
<following>Si cabe, puede citarse la agresividad y velocidad endiablada del belga de origen africano Emile Mpenza y algunos detalles técnicos del veteranísimo rumano Gica Hagi (35 años).</following>
</context>
</instance>
<instance id="partido.n.85" docsrc="efe_14074_2000/06/17">
<cat scheme="ANPA" code="DEP:DEPORTES,VOLEIBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.85" senseid="partido.2"/>
<context>
<previous></previous>
<target>La selección argentina, en su debut en el preolímpico femenino de voleibol, se mostró como un temible adversario en el <head>partido</head> ante Japón, equipo anfitrión del torneo (1-3), al que hizo sufrir hasta el último set antes de ceder el triunfo.</target>
<following>Las jugadoras que entrena Claudio Cuello no decepcionaron en su primer importante compromiso. Antes nueve mil apasionados seguidores nipones, el conjunto argentino luchó cada punto y, sólo la mayor experiencia internacional de sus adversarias le privó de un mejor resultado.</following>
</context>
</instance>
<instance id="partido.n.86" docsrc="efe_17088_2000/06/21">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.86" senseid="partido.1"/>
<context>
<previous>El alcalde de León y presidente provincial del PP, Mario Amilivia, calificó hoy como "acertada" la decisión de la Junta Directiva Nacional del PP de decretar incompatibles los cargos de alcalde o presidente de Diputación que también sean senadores con el de presidente provincial del partido.</previous>
<target>Amilivia compareció hoy en conferencia de prensa para comentar el resultado de la reunión que mantuvo ayer en Madrid con el secretario general del PP, Javier Arenas, dentro de la ronda de conversaciones que éste mantiene estos días con todos los presidentes provinciales y regionales del <head>partido</head>.</target>
<following>El alcalde de León comunicó personalmente a Javier Arenas que está a "plena disposición" del partido y que no se presentará a la reelección como presidente provincial, tal y como anunció la pasada semana.</following>
</context>
</instance>
<instance id="partido.n.87" docsrc="efe_17683_2000/06/22">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.87" senseid="partido.2"/>
<context>
<previous>"Hemos cumplido como debíamos. Vinimos a salir campeones y lo hicimos. Merecíamos ganar y ganamos", declaró Bianchi en medio del terreno del estadio Morumbi apenas terminó el encuentro.</previous>
<target>El técnico, que ya había ganado la Copa Libertadores en 1994 al frente del Vélez Sarsfield, afirmó que el gran mérito de su equipo hoy fue "haber venido a jugar de igual a igual y sin complejos", tras haber empatado 2-2 el <head>partido</head> de ida, disputado la semana pasada en Buenos Aires.</target>
<following>Bianchi no olvidó la presencia en las tribunas del ex futbolista argentino Diego Armando Maradona y le envió un mensaje desde el césped paulista.</following>
</context>
</instance>
<instance id="partido.n.88" docsrc="efe_20467_2000/06/25">
<cat scheme="ANPA" code="POL:POLITICA,EXTERIOR"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.88" senseid="partido.1"/>
<context>
<previous>El PD apuntaba en sus previsiones más optimistas a 150 escaños con un programa que optaba por una urgente reforma fiscal y apostaba por candidatos jóvenes.</previous>

<target>Otros dos <head>partidos</head> de la oposición que ganaron terreno fueron el Liberal (PL) de Ichiro Ozawa que con su mensaje de "cambio" pasa de 18 a 22 asientos y el Socialista de Takako Doi, que logró un meritorio avance al pasar de 14 a 19 escaños</target>
<following></following>
</context>
</instance>
<instance id="partido.n.89" docsrc="efe_21868_2000/06/27">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.89" senseid="partido.2"/>
<context>
<previous></previous>
<target>Frank Rijkaard, técnico de la selección holandesa de fútbol, advirtió hoy que la semifinal de la Eurocopa contra Italia será muy distinta al <head>partido</head> de cuartos contra Yugoslavia, en el que su equipo venció por 6-1, y señaló que será muy complicado repetir una goleada similar.</target>
<following>Rijkaard quiere frenar la euforia existente en Holanda por el poderío de su próximo rival. "Italia es totalmente diferente a Yugoslavia", destacó el técnico.</following>
</context>
</instance>
<instance id="partido.n.90" docsrc="efe_2485_2000/07/04">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.90" senseid="partido.2"/>
<context>
<previous>Lapuente si contará con el defensa Rafael Márquez, del Mónaco de Francia, quien hoy se incorporó al equipo para estar en los dos primeros partidos de la eliminatoria, contra Panamá el 16 de julio, y contra Trinidad y Tobago, el 22 de julio, ambos como visitante.</previous>
<target>Para Venezuela, que dirige el argentino José Omar Pastoriza, también le servirá el <head>partido</head> ante México más allá de las prisas para organizarlo, porque le permitirá ver a su plantel antes de que se enfrenten a Uruguay, su próximo rival en la eliminatoria mundialista de Suramérica</target>
<following></following>
</context>
</instance>
<instance id="partido.n.91" docsrc="efe_9155_2000/07/12">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.91" senseid="partido.2"/>
<context>
<previous></previous>
<target>El técnico de la selección colombiana de fútbol, Luis Augusto García, no ha podido resolver el dilema sobre quién, entre Jorge Bermúdez del Boca Juniors o Mario Yepes del River Plate de Argentina, se quedará en el banco de suplentes en el <head>partido</head> contra Perú el próximo 19 de julio en Lima.</target>
<following>Para dicho compromiso, el quinto de Colombia en las eliminatorias suramericanas para el mundial del 2002, García ha confirmado como titular en la zaga a Iván Córdoba, del Inter de Italia, quien dijo que no jugará como lateral izquierdo para que el seleccionado cafetero pueda actuar con cuatro defensores sin sacrificar a Bermúdez ni a Yepes.</following>
</context>
</instance>
<instance id="partido.n.92" docsrc="efe_13004_2000/07/17">
<cat scheme="ANPA" code="POL:POLITICA,GOBIERNO"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.92" senseid="partido.1"/>
<context>
<previous>"Preferimos los labradores de futuro a los que alumbran caminos trillados", comentó, en respuesta a las afirmaciones de Chirac, la ministra de Justicia, la socialista Elisabeth Guigou, quien es candidata a la Alcaldía de Aviñón en las Municipales del 2001.</previous>
<target>Por otra parte, Jospin criticó, sin nombrarlo, a la federación parisiense del <head>partido</head> neogaullista Reagrupamiento por la República (RPR), que fundara Chirac hace casi un cuarto de siglo.</target>
<following>Los grandes partidos políticos tienen un papel "indispensable" cuando "no están desgarrados por querellas fratricidas" y cuando son capaces de tener "proyectos", dijo el político socialista, en una clara referencia a las disputas internas del RPP en París</following>

```
		</context>
</instance>
<instance id="partido.n.93" docsrc="efe_13074_2000/07/17">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.93" senseid="partido.2"/>
		<context>
		<previous>El chico, procedente de una familia de clase media uruguaya que apuesta primero a su formación
		académica que a su éxito futbolístico, deslumbró a los representantes del Gremio con su depurado repertorio técnico,
		aliado a un sorprendente olfato goleador, explicó la fuente.</previous>
		<target>Atraído por los comentarios de hinchas del Gremio en Rivera, Verardi viajó ayer, domingo, a la ciudad
		fronteriza para observarle en acción durante un <head>partido</head> del equipo de su colegio, el Sarandí
		Universitario.</target>
		<following>El conjunto de Scorza se impuso por 16-1 y media docena de goles tuvieron el sello del habilidoso
		creador, que juega con el número diez a la espalda y, a pesar de sus modales tímidos y conservadores, es un activo
		peligro desde la izquierda.</following>
		</context>
</instance>
<instance id="partido.n.94" docsrc="efe_17795_2000/07/23">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.94" senseid="partido.1"/>
		<context>
		<previous>Dominada por caras nuevas y con nueve puestos menos que la anterior, en la nueva Ejecutiva se repite tan
		solo un nombre de la elegida en 1997, la diputada Micaela Navarro, mientras que desaparecen los representantes de
		los tradicionales sectores "guerristas" (por Alfonso Guerra) o "renovadores" y casi todas las anteriores figuras del
		PSOE de Felipe González.</previous>
		<target>Solo permanece Manuel Chaves, que tras haber dirigido la gestora que se ha hecho cargo del
		<head>partido</head> durante la crisis abierta tras la dimisión de Joaquín Almunia, ocupará ahora la Presidencia del
		PSOE, vacante desde la muerte del histórico dirigente Ramón Rubial.</target>
		<following>Estrechos colaboradores de Rodríguez Zapatero ocupan los puestos claves de la nueva dirección, como la
		Secretaria de Organización y Acción Electoral, que es para el diputado José Blanco, o la de Política Económica, para
		Jordi Sevilla, uno de los autores del programa de "Nueva Vía".</following>
		</context>
</instance>
<instance id="partido.n.95" docsrc="efe_23523_2000/07/30">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.95" senseid="partido.2"/>
		<context>
		<previous>Bolivia se enfrentará el próximo 16 de agosto a Ecuador en la séptima jornada de la eliminatoria
		mundialista, partido para el que se preparará en la ciudad de La Paz.</previous>
		<target>Castedo explicó que no tienen ninguna intención de cambiar la sede de la eliminatoria mundialista y aseguró
		que los <head>partidos</head> de local se jugarán en el estadio Hernando Siles de esta ciudad, ubicada a más de
		3.600 metros de altitud sobre el nivel del mar.</target>
		<following>Según el cronograma aprobado a principio de temporada, la plantilla boliviana se concentrará en La Paz
		desde el próximo 7 de agosto para disputar el partido con Ecuador el 16 de este mes</following>
		</context>
</instance>
<instance id="partido.n.96" docsrc="efe_24255_2000/07/31">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="ANPA" code="SOC:SOCIEDAD-SALUD,SOLIDARIDAD-DERECHOS"/>
<cat scheme="IPTC" code="07000000"/>
<cat scheme="IPTC" code="11000000"/>
<cat scheme="IPTC" code="14000000"/>
<answer instance="partido.n.96" senseid="partido.1"/>
		<context>
		<previous>Los trabajos de este grupo, que dirige el nuevo portavoz del grupo Parlamentario Socialista, Jesús
		Caldera, desembocarán en un plan integral "cargado de contenido", en el que se "ofrecerá" al Gobierno las líneas
		fundamentales para crear un "gran Pacto de Estado" de inmigración, señalaron a Efe fuentes socialistas.</previous>
```

<target>En este Pacto de Estado deberán estar implicados <head>partidos</head> políticos, gobiernos, sindicatos y agentes sociales y económicos.</target>
<following>Las mismas fuentes insistieron en la necesidad de constituir este Pacto de Estado y de no reducir la inmigración a la reforma de la Ley de Extranjería propuesta por el Gobierno, ya que ésta "no es suficiente" para resolver el problema y es, además, una "estrategia equivocada".</following>
</context>
</instance>
<instance id="partido.n.97" docsrc="efe_852_2000/08/02">
<cat scheme="ANPA" code="POL:POLITICA,ELECCIONES"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.97" senseid="partido.1"/>
<context>
<previous>Para McCain, Bush "es ahora la persona que mejor representa las mejores esperanzas para el futuro" de los Estados Unidos y vaticinó que se convertirá en el próximo presidente de este país en noviembre próximo.</previous>
<target>En el mensaje principal de la sesión de esta noche de la convención presidencial republicana en Filadelfia, McCain advirtió de que su <head>partido</head> debe eludir "el aislacionismo y el proteccionismo".</target>
<following>"No podemos construir murallas contra el éxito global de nuestros intereses y valores. Las murallas son para los cobardes", aseguró McCain, quién llegó a amenazar la candidatura presidencial de Bush durante las primarias republicanas de principios de año.</following>
</context>
</instance>
<instance id="partido.n.98" docsrc="efe_1427_2000/08/02">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.98" senseid="partido.2"/>
<context>
<previous>El directivo dijo que la decisión fue tomada tras comprobar que el pasado sábado, varios integrantes del grupo de hinchas organizados en la "Ultra" agredieron a una familia, y que cerca de un centenar de ellos, golpearon a un seguidor del Alajuelense y cometieron actos de vandalismo y el robo de joyas y dinero.</previous>
<target>El pasado domingo, durante un <head>partido</head> disputado ante Puntarenas, y en el que el Saprissa cayó dos goles a cero, miembros de la "Ultra" protagonizaron una trifulca en el estadio de Puntarenas, a 115 kilómetros de la capital, y que dejó varios heridos y lesionados.</target>
<following>Méndez reconoció que la "Ultra", creada en 1995 como un grupo oficial de apoyo al Club, "se nos salió de las manos, por lo que tenemos que ser radicales en la decisión de eliminarla".</following>
</context>
</instance>
<instance id="partido.n.99" docsrc="efe_2437_2000/08/04">
<cat scheme="ANPA" code="POL:POLITICA,ELECCIONES,PRESIDENCIALES"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.99" senseid="partido.1"/>
<context>
<previous>Anteriormente, los 2.066 delegados a la convención presidencial republicana de Filadelfia proclamaron por unanimidad la candidatura a la Casa Blanca de Bush.</previous>
<target>La proclamación de Bush como candidato presidencial republicano esta noche constituyó una mera formalidad por cuanto anoche había superado el mínimo de 1.043 delegados necesarios para representar al <head>partido</head> en las elecciones de noviembre</target>
<following></following>
</context>
</instance>
<instance id="partido.n.100" docsrc="efe_4750_2000/08/07">
<cat scheme="ANPA" code="POL"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.100" senseid="partido.1"/>
<context>
<previous></previous>
<target>El Movimiento V República (MVR), <head>partido</head> del presidente Hugo Chávez, indicó hoy que ha establecido cinco áreas prioritarias sobre las que deberá trabajar la recién elegida Asamblea Nacional (AN), entre ellas las de seguridad y empleo.</target>

<following>Luis Miquilena, presidente de la Comisión Legislativa Nacional (CLN) y director del MVR, dijo que los otros tres sectores que requieren de leyes con carácter urgente son el de la seguridad social, el de la economía y el de la alimentación.</following>
</context>
</instance>
<instance id="partido.n.101" docsrc="efe_9531_2000/08/14">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.101" senseid="partido.2"/>
<context>
<previous></previous>
<target>El <head>partido</head> entre Ecuador y Bolivia, que se disputará el próximo miércoles en Quito correspondiente a la séptima jornada de las eliminatorias del Mundial de Japón-Corea`2002, está demostrando muy poco interés entre los aficionados según se pudo comprobar hoy en el primer día de venta de entradas, indicó el jefe de la taquilla Marcelo Valencia.</target>
<following>"Poco a sido el interés puesto de manifiesto por los aficionados para adquirir las entradas en el primer día de ventas, pero ojalá que en los dos días que restan se incremente el entusiasmo", dijo Valencia a los periodistas locales.
</following>
</context>
</instance>
<instance id="partido.n.102" docsrc="efe_11154_2000/08/17">
<cat scheme="ANPA" code="SOC:SOCIEDAD-SALUD,SOCIEDAD"/>
<cat scheme="IPTC" code="07000000"/>
<cat scheme="IPTC" code="14000000"/>
<answer instance="partido.n.102" senseid="partido.1"/>
<context>
<previous>La secretaria de Relaciones con las ONG y Movimientos Sociales del PSOE, Leire Pajín, aseguró hoy a Efe que su partido pondrá un gran esfuerzo en trabajar conjuntamente con las organizaciones sociales, pero no se tratará, precisó, de "captar" a las ONG para su causa, sino de confiar en su proyecto".</previous>
<target>Pajín, que cumplirá 24 años en septiembre y es la diputada más joven del Congreso, señaló que las ONG son fruto de los nuevos tiempos y han creado "espacios" donde ni la sociedad ni los <head>partidos</head> "han podido llegar".</target>
<following>En ese sentido, reconoció que "es verdad" que el PSOE "llevaba tiempo distanciado de la sociedad", una dinámica, dijo, que se tiene que romper para "escuchar más a la gente y menos a nosotros".</following>
</context>
</instance>
<instance id="partido.n.103" docsrc="efe_13577_2000/08/21">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.103" senseid="partido.2"/>
<context>
<previous></previous>
<target>El Liverpool visita hoy lunes al Arsenal en busca del liderato provisional en <head>partido</head> adelantado de la segunda jornada de la liga inglesa cuando hace apenas 48 horas que venció al Bradford City por un gol a cero en su estadio de Anfield Road.</target>
<following>El compromiso adquirido con la televisión hace que este encuentro se dispute, cuando ambos equipos apenas han descansado, aunque el resto de conjuntos deberán también afrontar una semana repleta de competición ya que entre el martes y miércoles se completará la segunda jornada y en el fin de semana se jugará la tercera.</following>
</context>
</instance>
<instance id="partido.n.104" docsrc="efe_14947_2000/08/23">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.104" senseid="partido.1"/>
<context>
<previous>El dirigente socialista Jesús Caldera aseguró hoy que, aunque en el PNV "hay movimientos", algunos de sus dirigentes, como su portavoz Joseba Egibar, con sus declaraciones, "viene a ratificar que es un poco el guardián de la ortodoxia".</previous>

<target>Caldera, en declaraciones a la Cadena Ser, se refirió así a las manifestaciones de Egibar, quien dijo ayer que las bases fundamentales para la constitución de un foro de <head>partidos</head> deben ser "el respeto a todos los derechos humanos y a lo que la sociedad vasca pueda decidir libre y democráticamente".</target>
<following>El dirigente socialista señaló que Egibar "intenta tapar las vías de agua que se abren en el barco de Estella, que se está hundiendo".</following>
</context>
</instance>
<instance id="partido.n.105" docsrc="efe_16564_2000/08/25">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.105" senseid="partido.2"/>
<context>
<previous>Además del triunfo, el conjunto barcelonista dio una impresión aceptable, aunque en contadas ocasiones, pues al buen juego desarrollado en la primera media hora, le continuó un periodo no tan vistoso y una segunda parte realmente preocupante, después de que Serra Ferrer llevase a cabo seis cambios.</previous>
<target>En los seis <head>partidos</head> disputados (victorias contra el WHC, 0-14; Quick 1890, 0-10; Arsenal, 1-2; NEC Nimega, 2-5; y PSV, 2-1; y un empate contra el Lazio, 3-3), el Barcelona ha marcado 56 goles, en los que Dani se ha destacado en el máximo goleador, con siete tantos.</target>
<following>Serra Ferrer ha convocado a 19 jugadores para el partido de mañana, entre los que sólo han quedado descartados los lesionados Luis Enrique, Emmanuel Petit y Marc Overmars, además de los descartados por el técnico Jari Litmanen, Winston Bogarde y Samuel Okunowo.</following>
</context>
</instance>
<instance id="partido.n.106" docsrc="efe_18913_2000/08/29">
<cat scheme="ANPA" code="TRI:JUSTICIA-INTERIOR-SUCESOS,TERRORISMO POL:POLITICA,REGIONES-AUTONOMIAS"/>
<cat scheme="IPTC" code="02000000"/>
<cat scheme="IPTC" code="03000000"/>
<cat scheme="IPTC" code="16000000"/>
<answer instance="partido.n.106" senseid="partido.1"/>
<context>
<previous></previous>
<target>El presidente de Castilla y León y del PP regional, Juan José Lucas, insistió hoy en que no se "debe caer en el desánimo" ante el asesinato que terminó con la vida del concejal de su <head>partido</head> en Zumárraga Manuel Indiano, y apostó por la vía del Estado de Derecho frente a las "pistolas".</target>
<following>Lucas, en declaraciones a Efe, señaló que con este atentado "ha muerto una parte de la sociedad española y de la juventud vasca".</following>
</context>
</instance>
<instance id="partido.n.107" docsrc="efe_18966_2000/08/29">
<cat scheme="ANPA" code="TRI:JUSTICIA-INTERIOR-SUCESOS,TERRORISMO"/>
<cat scheme="IPTC" code="02000000"/>
<cat scheme="IPTC" code="03000000"/>
<cat scheme="IPTC" code="16000000"/>
<answer instance="partido.n.107" senseid="partido.1"/>
<context>
<previous>Al hospital de Zumárraga se desplazaron numerosos representantes institucionales y políticos, entre ellos el lehendakari, Juan José Ibarretxe; el portavoz del ejecutivo autónomo, Josu Jon Imaz; el diputado general de Guipúzcoa, Román Sudupe, y el presidente de las Juntas Generales del territorio, Iñaki Alkiza, entre otras autoridades.</previous>
<target>Manuel Indiano concurrió en las últimas elecciones municipales como independiente en las listas del PP y aunque no era el inmediatamente posterior en la lista para sustituir a Faustino Villanueva, quien renunció en febrero, el <head>partido</head> consideró que era el candidato más idóneo, según las mismas fuentes.</target>
<following>Los restos mortales del concejal han sido trasladados al Instituto Anatómico Forense del cementerio de Polloe de San Sebastián, donde se le practicará la autopsia, y desde allí se llevarán al Ayuntamiento de Zumárraga, donde quedará instalada la capilla ardiente.</following>
</context>
</instance>
<instance id="partido.n.108" docsrc="efe_20577_2000/08/31">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>

<answer instance="partido.n.108" senseid="partido.2"/>
<context>
<previous>La liga portuguesa de fútbol se paraliza el próximo fin de semana por la visita el domingo de la selección nacional a Estonia, en encuentro del Grupo 2 europeo de las eliminatorias del Mundial`2002, que se disputará en Corea del Sur y Japón.</previous>
<target>No obstante, sábado se celebrará en la isla de Madeira el <head>partido</head> Marítimo-Gil Vicente, con el que se cerrará la segunda jornada del campeonato y que enfrenta a dos equipos empatados en al tabla con un punto.</target>
<following>El paréntesis en la liga debido al comienzo de la eliminatorias mundialistas debe ser aprovechado por entrenadores para ultimar la preparación de sus conjuntos para la larga temporada, aunque los tres grandes Sporting, Oporto y Benfica no debe servir para mucho estos días de descanso debido sus numerosos internacionales.</following>
</context>
</instance>
<instance id="partido.n.109" docsrc="efe_7201_2000/09/11">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.109" senseid="partido.2"/>
<context>
<previous>Bonet se recuperó de las dos jornadas que llevaba sin anotar y consiguió hoy igualar el récord de goles en un partido del torneo peruano que logró hace dos años el nacional Claudio Pizarro con la camiseta del Alianza Lima, antes de ser transferido al Werder Bremen de Alemania.</previous>
<target>El delantero argentino, figura excluyente de la octava jornada del Clausura, convirtió sus goles en los minutos 1, 45, 47, 55 y 61 del intenso <head>partido</head> que disputaron el Cienciano y el Melgar en el estadio de la ciudad sureña del Cuzco.</target>
<following>El Melgar consiguió disminuir la diferencia con dos goles del centrocampista brasileño Paulo César Cruvinel, a los 49 y 75 minutos de juego.</following>
</context>
</instance>
<instance id="partido.n.110" docsrc="efe_7536_2000/09/11">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.110" senseid="partido.1"/>
<context>
<previous>El secretario general del PSOE, José Luis Rodríguez Zapatero, se mostró hoy preocupado por la evolución de la situación económica y también porque el Gobierno, en este contexto, se muestre "falto de reflejos, de pulso, sin iniciativas", críticas que los socialistas manifestarán a partir de mañana en el Parlamento.</previous>
<target>En una conferencia de prensa ofrecida en Bilbao tras la reunión de la Comisión Ejecutiva Federal del <head>partido</head>, Rodríguez Zapatero criticó al Gobierno por ofrecer a la sociedad sólo mensajes "negativos y pesimistas" en vez de dar respuestas ante el incremento de la inflación, la pérdida de competitividad o el aumento del precio de los carburantes.</target>
<following>La actitud del Gobierno, a su juicio, se basa en "el inmovilismo y el silencio cuando España se juega en decisiones estructurales y de fondo seguramente su capacidad para ser una sociedad y una economía moderna y avanzada" y, además, cuando "muchísimos sectores" económicos empiezan a tener "serios problemas" en su actividad.
</following>
</context>
</instance>
<instance id="partido.n.111" docsrc="efe_9997_2000/09/13">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.111" senseid="partido.2"/>
<context>
<previous>El centrocampista del PSV Eindhoven Theo Lucius salió al campo y revolucionó un partido que se le había puesto muy cuesta arriba al conjunto holandés tras el tempranero gol del Dynamo de Kiev al conseguir el gol del empate y dar el pase del tanto que suponía la victoria del PSV.</previous>
<target>El conjunto ucraniano vio recompensado su dominio inicial en el <head>partido</head> en el minuto seis cuando el delantero uzbeko Maksim Shatskikh puso el 0-1 en el marcador tras rematar un pase de Bialkevich.</target>

<following>El Dynamo anuló al PSV del campo, pero pronto llegó la reacción de los holandeses, que tuvieron su mejor oportunidad en un remate de Yuri Nikiforov, que cabeceó fuera un balón que había quedado en el área tras un disparo fallido del yugoslavo Mateja Kezman.</following>
</context>
</instance>
<instance id="partido.n.112" docsrc="efe_12451_2000/09/16">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.112" senseid="partido.2"/>
<context>
<previous>Desde este momento el Elche siguió a la busca de un gol que le diera la victoria, pero el Sevilla se defendió de manera ordenada y buscando la velocidad tanto de Gallardo como de Fredi y los disparos duros desde fuera del área.
</previous>
<target>Fruto de un disparo lejano que el portero ilicitano César Gálvez no pudo atrapar, el delantero Gallardo puso la puntilla al equipo de Felipe Mesones al recoger el rechace dentro del área pequeña y desnivelar el marcador cuando todo indicaba que el <head>partido</head> iba a terminar en tablas</target>
<following></following>
</context>
</instance>
<instance id="partido.n.113" docsrc="efe_13183_2000/09/17">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.113" senseid="partido.2"/>
<context>
<previous>Suya fue la primera oportunidad del partido a los ochos minutos, pero su disparo salió ligeramente desviado. Ante esto, los rayistas respondieron inmediatamente, tal vez escarmentados del gol de Milosevic un año atrás que les dejó sin liderato. Luis Cembranos en el minuto 15 y Bolo en el 21 llevaron la inquietud a la meta de Juanmi, aunque sus lanzamientos no encontraron el fin buscado.</previous>
<target>La lesión de Luis Cembranos en el minuto 18 en principio parecía un serio contratiempo para los de Juande Ramos. Hoy por hoy es el jugador más en forma y decisivo de este equipo. Sin embargo la realidad fue diferente. Los últimos veinte minutos de la primera parte fueron los mejores de los locales, que gozaron de las mejores ocasiones del <head>partido</head> en las botas de Bolo y Bolic.</target>
<following>Mientras, el Real Zaragoza pareció entrar en un sueño profundo. Tanto Jamelli como Yordi apenas intervenían en el juego y el fútbol de toque que pretende Lillo no aparecía por ninguna parte. El técnico ovetense sorprendió al dejar en el banquillo a Juanele y eso lo notó la punta de ataque.</following>
</context>
</instance>
<instance id="partido.n.114" docsrc="efe_14062_2000/09/18">
<cat scheme="ANPA" code="DEP:DEPORTES,FUTBOL"/>
<cat scheme="IPTC" code="15000000"/>
<answer instance="partido.n.114" senseid="partido.2"/>
<context>
<previous>Después de que el conjunto universitario cumpliese las tres primeras jornadas con el balance de un empate, dos derrotas y sin marcar un tanto, Sequeiros estimó que "la reacción llegará pronto porque ya demostró en la pretemporada que tiene un equipo muy compacto".</previous>
<target>"Yo estoy muy ilusionado con esta oportunidad, que pone fin a los problemas que he tenido en la pretemporada, y me anima muchísimo poder jugar de nuevo un <head>partido</head> oficial", agregó.</target>
<following>Marcos Sequeiros no sabe si podrá debutar el próximo fin de semana ante el Recreativo de Huelva. "Sólo me preocupa entrenar a tope y la decisión final del que deba jugar dependerá exclusivamente al entrenador", señaló.</following>
</context>
</instance>
<instance id="partido.n.115" docsrc="efe_14716_2000/09/19">
<cat scheme="ANPA" code="POL:POLITICA,ELECCIONES"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.115" senseid="partido.1"/>
<context>
<previous>Bandas de agitadores recorren la ciudad clavando en los ojos del rostro de los carteles electorales del principal candidato opositor, Vojislav Kostunica, el emblema diamantino de la OTAN.</previous>

<target>Kostunica, de la coalición de 19 <head>partidos</head> medianos o pequeños reunidos en Oposición Democrática Serbia (DOS), es un nacionalista como Milosevic pero es dialogante con Occidente, promete no intentar imposibles contra la OTAN, e irse antes de cuatro años.</target>
<following>Kostunica critica a fondo a EEUU, condena el ataque de 1999, y niega que Milosevic sea más patriota que él; en 1995, Kostunica acusó a Milosevic de traición y de ser acólito de EEUU tras los acuerdos de Dayton (1995) que trajeron la paz a Bosnia.</following>
</context>
</instance>
<instance id="partido.n.116" docsrc="efe_17901_2000/09/22">
<cat scheme="ANPA" code="TRI:JUSTICIA-INTERIOR-SUCESOS,TERRORISMO"/>
<cat scheme="IPTC" code="02000000"/>
<cat scheme="IPTC" code="03000000"/>
<cat scheme="IPTC" code="16000000"/>
<answer instance="partido.n.116" senseid="partido.1"/>
<context>
<previous>El punto de salida de la concentración será la confluencia de la calle de Aragón y el Paseo de Gracia, y discurrirá hasta la plaza de Cataluña.</previous>
<target>La manifestación estará encabezada por el presidente de la Generalitat, Jordi Pujol, y el alcalde de Barcelona, Joan Clos, y se prevé la asistencia de las principales autoridades catalanas y los líderes de todos los
<head>partidos</head>.</target>
<following>Habrá además una concentración pacífica a las diez y media de la mañana en la plaza de Sant Jaume, también en Barcelona, en la que participarán todas las instituciones, y a las 12.00 manifestaciones silenciosas ante la Diputación de Barcelona y ante el Ayuntamiento de Sant Adriá, convocadas por los sindicatos CCOO y UGT, y también ante varios ayuntamientos catalanes.</following>
</context>
</instance>
<instance id="partido.n.117" docsrc="efe_18965_2000/09/23">
<cat scheme="ANPA" code="POL:POLITICA,CONFLICTO POL:POLITICA,INTERIOR"/>
<cat scheme="IPTC" code="16000000"/>
<answer instance="partido.n.117" senseid="partido.1"/>
<context>
<previous>Los representantes acordaron en la reunión del viernes, además, buscar "el consenso" para facilitar "la más inmediata aprobación" del proyecto de Ley que presentó el Ejecutivo para desactivar el SIN.</previous>
<target>Se acordó, asimismo, que los <head>partidos</head> de oposición estudien el proyecto de modificación constitucional en materia electoral que ha presentado el Ejecutivo para facilitar la realización de nuevas
elecciones generales el próximo año.</target>
<following>El Ejecutivo ha elaborado en los últimos días diversos proyectos de ley y de modificación de la
Constitución que deberán ser aprobados por el Congreso para que se concreten los anuncios de Fujimori.</following>
</context>
</instance>
<instance id="partido.n.118" docsrc="efe_22987_2000/09/27">
<cat scheme="ANPA" code="POL:POLITICA,PARTIDOS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.118" senseid="partido.1"/>
<context>
<previous>"Vamos a hacer frente a todas las responsabilidades, se les va a pagar a todos", aseguró
López.</previous>
<target>Exiplastic, una de las empresas afectadas por las deudas del PRI, vendió al <head>partido</head> un millón de gallardetes de plástico para colocarlos en postes de electricidad en las principales ciudades del país.</target>
<following>El gerente de la compañía, Julio César Rumalan, dijo que el partido únicamente cubrió el 50 por ciento del coste de los gallardetes.</following>
</context>
</instance>
<instance id="partido.n.119" docsrc="efe_655_2000/10/02">
<cat scheme="ANPA" code="POL:POLITICA,REGIONES-AUTONOMIAS"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.119" senseid="partido.1"/>
<context>
<previous>En este sentido, recordó el reciente asesinato del presidente de la patronal guipuzcoana Adegi, José María Korta, así como el "chantaje" y la "coacción inadmisible" que supone el denominado "impuesto revolucionario", ante la cual pidió a los empresarios que "no cedan ante la extorsión".</previous>

<target>"Tenemos la obligación de anteponer la defensa del derecho a la vida y a la libertad a cualquier aspiración o estrategia política", añadió el diputado general, quien destacó la necesidad de que los <head>partidos</head> políticos sean capaces de manifestarse "unidos, sin fisuras ni tensiones estériles, en contra de la violencia y en favor del derecho a la vida y a la libertad de las personas".</target>
<following>Reclamó asimismo a quienes "dan cobertura política" a las acciones terroristas que "se comprometan con la paz", ya que "no es posible ninguna colaboración política mientras no expresen un rechazo expl ícito de la violencia".</following>
</context>
</instance>
<instance id="partido.n.120" docsrc="efe_5414_2000/10/07">
<cat scheme="ANPA" code="POL:POLITICA,ELECCIONES"/>
<cat scheme="IPTC" code="11000000"/>
<answer instance="partido.n.120" senseid="partido.1"/>
<context>
<previous>Estas son las terceras elecciones parlamentarias regulares que se celebran en Eslovenia desde que este país, hoy miembro asociado de la Unión Europea y participante del programa "Asociación por la paz" de la OTAN, se independizó (1991) de la antigua Yugoslavia comunista.</previous>
<target>En 1996 el LDS obtuvo el mayor número de escaños (25), pero se produjo un "empate" entre el bloque de centroizquierda, encabezado por este <head>partido</head>, y el de centroderecha, puesto que los dos lados obtuvieron en mismo número de diputados.</target>
<following>El LDS formó Gobierno entonces con el SLS, que había obtenido 19 escaños, y con DESUS, pero la coalición se desintegró en junio de este año y el ejecutivo de Drnovsek fue reemplazado por una coalición del centroderecha con el primer ministro Bajuk al frente</following>
</context>
</instance>
</lexelt>
</corpus>

Anexo 3: Selección de 120 instancias para la unidad léxica «cabeza»

corpusid|source|context|senseid
278645592|chile_tech|Permanecer frente a el monitor de un computador durante horas generalmente provoca malestares propios de el Síndrome de Visión de el Computador (SVC) como cansancio de los ojos, dolores de cabeza, visión borrosa y otras alteraciones.|HEAD_00
279055336|chile_tech|En el año 2000, Compaq se colocó a la cabeza de la industria con una participación de 30.5 por ciento en ingresos por ventas de servidores Linux.|LEADER_00
279112729|chile_tech|Este Portal, que ha sido definido como un sitio no comercial, espera ponerse en breve a la cabeza de los Portales Chilenos.|LEADER_00
280118145|capital|Muchos se toman la cabeza a el recordar que la Comisión Progresa, que se hizo cargo de los activos de el grupo Cruzat-Larraín en la crisis de los 80, vendió Provida en 40 millones de dólares a el Bankers Trust y hoy vale 247 millones de dólares.|HEAD_00
280160900|capital|Se supone que a esas alturas a el pájaro se le ha acabado el oxígeno y ha desarrollado un músculo en la cabeza que lo obliga a picar y salir.|HEAD_00
280161342|capital|Ahí desapareció toda especie viviente, pero como ella tenía el mérito de esconder la cabeza y vivir en la oscuridad, salió triunfante y siguió viviendo.|HEAD_00
280194258|capital|La competencia, la crisis asiática, la obscenidad de los costos de producción de películas recientes y no en último lugar la contracción latinoamericana están dando dolores de cabeza a el ratón Mickey.|HEAD_00
280208241|capital|Acabo de contratar para la oficina de California a un muchacho que viene de hacer un master en Virginia Wolf, muy inteligente y con una cabeza muy precisa.|INTELLIGENCE_00
280276872|capital|Sin embargo, el mayor remezón que se haya visto en el grupo se produjo el año pasado, cuando las cabezas de sus empresas o bien fueron despedidas o renunciaron.|CHIEF_00
280408245|capital|Quien le hizo la advertencia, Benjamín Claro, en ese entonces cabeza de el estudio, con el tiempo se convertiría no sólo en una suerte de segundo padre, sino que amenazó con abandonar la oficina si no hacían socio a Ricardo Claro.|CHIEF_00
280440655|capital|Sí, a lo mejor vamos a terminar teniendo cada uno una parabólica en la cabeza.|HEAD_00
280441046|capital|De hecho en la Entrevista de el domingo en TVN, Gabriel Valdés lo mencionó entre los jóvenes que a él le gustaría ver a la cabeza de el partido.|LEADER_00
280505757|capital|Mario echa la cabeza para atrás, se ríe, y dice que era un personaje de teatro que alguna vez inventó: y lo dice parodiando el habla de un rabino judío.|HEAD_00
280509145|capital|Mario mueve la cabeza en señal de gratitud.|HEAD_00
280525619|capital|Estuvo a la cabeza de la dirección ejecutiva de la Comisión Nacional de el Medio Ambiente (Conama), justo en la época en que se aprobó el estudio de impacto ambiental de el polémico proyecto Ralco.|CHIEF_00

280597338|capital|Veinte mil cosas. Pero es mucho más de acá dice, tocandose la cabeza.|HEAD_00
280633726|capital|Es la cabeza de mercados emergentes de Templeton, uno de los fondos más grandes de Estados Unidos.|LEADER_00
280690090|capital|En la antigüedad, cuando llegaba un mensajero con buenas noticias le daban un banquete, pero si eran malas ,le cortaban la cabeza.|HEAD_00
307296823|el_centro|Sin embargo, como el fútbol carece de toda lógica, cuando el duelo terminaba (a los 44) el pequeño Martel acertó con un impecable centro a el área peruana justo cuando la cabeza de Mirosevic apuntó a el pórtico de Ibáñez y Chile quedó ganando 1-0.|HEAD_00
307361180|el_centro|El fallecimiento de el alto oficial de ejército quedó a el descubierto en Talca el año 1975, cuando su cadáver apareció en el cerro La Virgen con un disparo en la cabeza.|HEAD_00
307449383|el_centro|En el ensayo futbolístico general apreciado ayer en el recinto maulino, el equipo que denominaremos probable titular para el cotejo ante el equipo de Fernando Carvallo ganó 2-1, con anotaciones de Marco Bautista y Ramón Avila, ambos de cabeza.|HEAD_00
307526145|el_centro|Ahí cayó de un tragaluz mientras hacía unas reparaciones, golpeandose la cabeza fuertemente.|HEAD_00
307569027|el_centro|Sin embargo, después que fue dado de alta, el joven ha continuado con fuertes dolores de cabeza y con bastante malestar en el ojo derecho, donde recibió el golpe.|HEAD_00
307934065|estrella_valparaíso|Mampato tenía que volver alguna vez, y de eso se encargó la gente de Cineanimadores, con Alejandro Rojas y Diego Garretón a la cabeza.|CHIEF_00
307966212|estrella_valparaíso|Finalmente, en los relatos chinos, el malvado dios Kong-Kong derriba con su cabeza una de las columnas que sostienen el cielo, perforando la bóveda celeste y dejando pasar trombas de agua que ahogan toda vida.|HEAD_00
308097354|estrella_valparaíso|Sin embargo, Luciano Tarifeño, quien está a la cabeza de el certamen que parte este domingo con la exhibición de la cinta nacional 'Antonia', no piensa así.|LEADER_00
308118354|estrella_valparaíso|Esta melodía suena en la cabeza de cualquier desocupado de más de 30 años.|HEAD_00
308219334|estrella_valparaíso|A mí me pasa con el conjunto barroco, yo soy el que lleva la cabeza y está insistiendo en los ensayos, a todos les gusta hacer esa música, pero tienen otras obligaciones.|CHIEF_00
308337938|estrella_valparaíso|Durante la época mundialera, estuvo a la cabeza Mauricio Correa, de 'Buenos días a todos'; ahora asumió Bibiano Castelló, quien antes dirigía 'Chile tuday', en Mega.|CHIEF_00
308349688|estrella_valparaíso|Un ítem aparte consiste la evaluación de los ministros, en la que la titular de Relaciones Exteriores, a el igual que en otros sondeos, sigue a la cabeza.|LEADER_00
308365353|estrella_valparaíso|El popular rockero Jon Bon Jovi está a la cabeza de este proyecto que Cristián protagonizaría en enero, en México.|CHIEF_00
308380657|estrella_valparaíso|Pero usted es la cabeza de el grupo, el que prometió estar entre los 8 primeros equipos de el torneo y ganar un tramo.|LEADER_00
316891973|estrella_valparaíso|La palabra la tiene el plantel de jugadores con su entrenador a la cabeza, de responder a los esfuerzos directivos y a los deseos de su hinchada de clasificar a el menos a la Copa Libertadores de América.|CHIEF_00
318111436|gran_valparaíso|Solo movió su cabeza afirmativamente y sintió que sus ojos se le nublaban por las lágrimas.|HEAD_00
319600785|lider_san_antonio|No te vas a estar levantando la falda para pegar una patada en la cabeza o en el pecho si alguien te ataca.|HEAD_00
320117437|lider_san_antonio|A la cabeza está Julio Acevedo, DT de varias escuelas de fútbol de la comuna, quien junto a un grupo de adiestradores dan los primeros pasos de esta entidad y quien por el momento es la cara visible de el grupo.|CHIEF_00
448126667|lun|Otra pequeña de nueve años, Rana Adnan, necesita oxígeno para una herida en el pecho y una contusión en su pulmón con conmoción cerebral, asimismo, presenta una herida en la cabeza, y un trozo de metralla en su brazo izquierdo.|HEAD_00
448629144|lun|El muro es la columna vertebral de ese monstruo sin cabeza que se llama Santiago.|HEAD_00
449599378|lun|A la cabeza figura Sadam, así como sus hijos Udai y Qusay.|LEADER_00
449795664|lun|A sus 28 años, y después de recibir una pedrada en la cabeza, quedó con secuelas irreversibles.|HEAD_00
504560692|el_mercurio|En cualquier caso, todavía no hay ninguna cabeza visible para el grupo, sabido el gusto de Miranda por operar fuera de las luces.|LEADER_00
505152884|el_mercurio|Su constitución frágil y el mayor tamaño de su cabeza en relación con el cuerpo son dos variables que juegan en contra en un impacto.|HEAD_00
505429203|el_mercurio|Colo Colo llega a 16 puntos y se mantiene a la cabeza de el Grupo D, a tres puntos de distancia de la Universidad Católica.|LEADER_00
505841786|el_mercurio|La explosión de dinamita voló la cabeza y ocasionó roturas múltiples en el monumento.|HEAD_00
507328929|el_mercurio|El amor tiene razones que la cabeza ignora.|INTELLIGENCE_00
474601807|el_mercurio|Teñirse el pelo en un tono radicalmente distinto a el natural, o cortarselo después de haberlo llevado largo por años, son cambios drásticos de look, que hacen que el dueño o la dueña de esa cabeza tenga que enfrentarse a el mundo con una nueva imagen.|HEAD_00
474608618|el_mercurio|¿Qué tanto se te ha pasado por la cabeza tener un espacio donde estés cómodo y tal vez cumplir esa expectativa que sabes que se comenta... de ser un cheque a fecha?|INTELLIGENCE_00
632360863|la_tercera|A la cabeza de los expertos se encuentra el técnico electoral de el PS, Francisco Aleuy, quien cumplió similar labor en las primarias de la Concertación, en mayo pasado.|CHIEF_00
632975648|la_tercera|Por ello, el 12 de diciembre debes votar con la cabeza y con el corazón.|INTELLIGENCE_00
633079185|la_tercera|Alvarez Quinteros resultó con una herida cortante en la cabeza y lesiones.|HEAD_00
724822606|austral_valdivia|También planteó una reforma a la salud, pero 'pensando con la cabeza y no con la guata'; la flexibilización de el Estatuto Docente y crear un fondo de inversión social que se financie con la venta de empresas públicas.|INTELLIGENCE_00

725437174|austral_valdivia|A la cabeza de el Ejército, a Izurieta le tocó enfrentar un período marcado por el proceso judicial a el ex Presidente Pinochet y, en el último lapso, por los escasos resultados arrojados por la denominada Mesa de Diálogo, que informó de el paradero de 200 detenidos desaparecidos durante el Régimen Militar (1973-90).|CHIEF_00
726372211|austral_valdivia|Ella sólo es una cabeza sin cuerpo la cual ha sido colocada sobre una bandeja.|HEAD_00
800835972|mostrador|Francisco Tepper estará a la cabeza de la nueva empresa como Director Gerente .|CHIEF_00
800886393|mostrador|El Chino, que fuera número uno de el mundo durante seis semanas en 1998, y cabeza de serie número siete en el torneo japonés dotado con 800 mil dólares en premios, mantiene su excelente racha de resultados y sigue sin ceder una solo manga.|LEADER_00
801625276|mostrador|Vinker repitió su gesto de rechazo y agitó la cabeza negativamente.|HEAD_00
803100521|mostrador|Ella inclinó la cabeza, vencida, subyugada y, luego, pudorosa, cubriendo su desnudez como mejor pudo con los jirones de el vestido, fue a poner se otra vez de rodillas ante el capitán.|HEAD_00
842196090|que_pasa|La cabeza de la empresa de telecomunicaciones más importante de España no dejó que su obsesión descansara.|CHIEF_00
842191652|que_pasa|Por citar algunas, South-Net, controlada por el fondo de inversiones Southern-Cross, tiene como cabeza a Norberto Morita; ClickNest pertenece a el grupo Pescarmona y el conglomerado Exxel Group invierte en InternetCo.|CHIEF_00
842375783|que_pasa|La selección chilena de Paddle Tenis será cabeza de serie número cuatro en el Mundial de la especialidad, que se disputará entre el 25 de junio y el 2 de julio, en Toulouse.|LEADER_00
842560616|que_pasa|Sin pensar en más, encajé la braga en mi cabeza, tapandome así la cara.|HEAD_00
843126300|que_pasa|Hay muchos que aún tienen la cabeza debajo de la arena, como el avestruz.|HEAD_00
331280561|mercurio_valparaíso|'Yo -a el igual que todos- vine a ganar. A Ulihrach tengo que ganarle, creo que lo puedo lograr ya que le gané la última vez que jugamos y espero ganarle otra vez. Tengo claro que todo depende de mí, de la confianza que tenga. Si gané el lunes es porque mi cabeza anduvo bien y espero que se repita'.|INTELLIGENCE_00
331285211|mercurio_valparaíso|Caballero explicó que la suspensión se mantendrá mientras el peligro de contagio exista y resaltó que se están reforzando las medidas de control preventivo en las zonas cordilleras desde la Cuarta a la Décima Región, donde se encuentran unas 250 mil cabezas de ganado en veranadas.|HEAD_00
331291888|mercurio_valparaíso|En otro tema, la cabeza visible en Chile de Europroject, Juan Arimany, explicó que la deuda de la filial Coinco S.A. a la Municipalidad de Viña de el Mar, por concepto de no pago de la renta de concesión, en caso alguno debiera empañar la imagen de solvencia económica de el consorcio para ser capaces de asumir la inversión de la marina, que demandaría US $70 millones.|LEADER_00
331298270|mercurio_valparaíso|Era el primer cabeza de serie de el torneo y absoluto favorito para coronarse en el cemento de el challenger de Salinas (Ecuador, 50 mil dólares).|LEADER_00
331377815|mercurio_valparaíso|Más de 160 mil cabezas de ganado han sido ya sacrificadas y otras 65 mil están sentenciadas.|HEAD_00
331422788|mercurio_valparaíso|'Estábamos muy ansiosos, además tenemos que pensar que este es un equipo relativamente nuevo y que debemos lograr compenetrar nos de mejor forma', explicó el delantero trasandino, quien terminó con un fuerte golpe en su rostro producto de una patada de el sureño Víctor Oyarzún 'el creyó que llegaba con el pie y yo con la cabeza, pero estoy seguro que fue sin mala intención'.|HEAD_00
331440521|mercurio_valparaíso|A la cabeza de la delegación viajaron la secretaria general de la Corporación, Angélica Rubiños y la directora de el Area Educación, Patricia Colarte, quienes fueron invitadas como parte de el proyecto colaboración mutua que mantienen la entidad local y la casa de estudios superiores de Puerto Rico.|CHIEF_00
331441737|mercurio_valparaíso|¿Estaría hidrófobo? El niño Manuel Ponce de doce años de edad, que vive en el cerro de la Mariposa Nº 57, comunicó ayer a el cuartel central de la policía que en circunstancias que pasaba por la avenida de el Brasil, entre la calle Molina y callejón Huito, rosó con un portaviandas que conducía a el Puerto a uno de los caballos de el carretón número 58, lo que fue suficiente para que el animal se diera por ofendido y le diera de coces a el muchacho hasta arrojar lo a el suelo y ocasionar le dos heridas, una en la cabeza y otra en el ojo derecho.|HEAD_00
331441927|mercurio_valparaíso|Este, según todas las apariencias, será sostenido vigorosamente por la asamblea, la cual se propone no usar de más condescendencias en adelante, con el monarquista que está jugando a la república, a la cabeza de ella.|LEADER_00
331499288|mercurio_valparaíso|Froilán Elespe, de 54 años, fue baleado en la cabeza mientras se encontraba en un bar.|HEAD_00
331499410|mercurio_valparaíso|El atentado ocurrió alrededor de las 13.40 GMT, cuando Froilán Elespe, de 54 años, recibió un tiro en la cabeza mientras estaba en un bar a el que acudía todos los días antes de ir a almorzar a su domicilio, informó el gobierno regional vasco.|HEAD_00
331544262|mercurio_valparaíso|El economista dijo no tener proyecto para servir el cargo en TVN ya que en ese momento -hacía sólo media hora que había sido informado- , 'seguía con su cabeza en el Banco Central', institución que abandonará antes de asumir sus nuevas funciones.|INTELLIGENCE_00
331544891|mercurio_valparaíso|De hecho uno de los ejemplares más populares es 'El topo que quería saber quien se había cagado en su cabeza'.|HEAD_00
331566699|mercurio_valparaíso|La cabeza de los wanderinos está en el campeonato local y el amistoso de el domingo ante San Luis, pero sin lugar a dudas que el corazón lo tienen puesto en Lima, ciudad que recibirá el choque entre chilenos y peruanos por las eliminatorias.|INTELLIGENCE_00
331569521|mercurio_valparaíso|La quiebra está bajo control y la empresa continúa expandiendose, sin embargo, a el interior de la firma esperan que la Corte de Apelaciones revoque la medida que les provocó más de un dolor de cabeza.|HEAD_00
331603084|mercurio_valparaíso|Por dentro, Express Tour mantiene la delantera y derrota por cabeza a su compañero de stud, Street Cry, en el Derby de los Emiratos Arabes.|HEAD_00
331603160|mercurio_valparaíso|Se trata de Express Tour, quien por cabeza aventajó a su compañero de corral Street Cry.|HEAD_00
318270324|gran_valparaiso|Se entiende con la cabeza, pero el cuerpo y el corazón tiran para otro lado.|INTELLIGENCE_00
443181187|lun|A el principal denunciante a el interior de EFE, el dirigente sindical Héctor Escobar, le pareció raro que la cabeza de la empresa no conozca lo que ocurre en las máquinas.|CHIEF_00
443194388|lun|'Tengo una admiración por Camilo y su participación suena contundente. Este es un disco hecho para tocar en vivo, eso es lo que tenía en mi cabeza cuando lo hice', cuenta.|INTELLIGENCE_00

443209640|lun|Otra palanca de ráting cuyo poderío no se debe despreciar (sobre todo la competencia) es el sello de garantía que ha patentado el programa de Guillermo Muñoz, ya sea rastreando las borrosas pistas de el asesinato de un campesino en Cochamó o tratando de aclarar las circunstancias que hicieron aparecer la cabeza de un narcotraficante en un puente santiaguino; ya sea recreando la cinematográfica fuga de la cárcel de Valparaíso en 1994 o intentando desentrañar las claves de el imbarajable magnetismo de el gurú peruano de Pelequén.|HEAD_00
443232341|lun|Eyzaguirre 'A el tema de sobresueldos le hemos escondido la cabeza hace 50 años'.|HEAD_00
443232384|lun|El Ministro de Hacienda, Nicolás Eyzaguirre, sostuvo hoy en el Congreso que el tema de los sobresueldos a altos funcionarios públicos es un asunto en el que 'durante 50 años hemos escondido la cabeza, recurriendo se a mecanismos como las transferencias desde empresas públicas, o los sobrecitos'.|HEAD_00
443253945|lun|He enfrentado a grandes y siempre salgo con la cabeza en alto.|HEAD_00
443262897|lun|Sin embargo, Dabrowski no se inmutó, tapó sus oídos y se hizo el sordo ante los insultos, para tapar les él la boca a todos los que pedían su cabeza.|HEAD_00
443263144|lun|Por eso cuando el árbitro Enrique Osses puso fin a los primeros 45 minutos y el marcador seguía en blanco, los pocos creyentes en el Polaco se tomaron la cabeza, presintiendo lo peor para su equipo.|HEAD_00
443264624|lun|Un saque de arco que cruza todo el terreno de juego, pica por sobre las cabezas de Ítalo Díaz y Germán Navea para dejar indefenso a el arquero Carlos Tejas ante la voracidad de Jaime González, quien con la punta de su zapato anotó su primer gol desde que volvió a la UC ('me pareció que González fue con la pierna muy arriba, así que creí que el árbitro cobraría falta', señaló Tejas, el segundo arquero menos batido de el campeonato).|HEAD_00
443266172|lun|Salvo la eliminación de Unión Española (ver página 20), todos los cabezas de serie avanzaron a la siguiente instancia, en la que destaca nítidamente el encuentro entre los dos últimos campeones de el fútbol nacional: Santiago Wanderers y Universidad Católica, cuyo primer duelo se escenificará en el estadio Municipal de Valparaíso.|CHIEF_00
443275553|lun|Todo iba bien hasta que comenzó el encuentro, cuando un grupo de barristas blancos lanzaron piedras a los rojos, quienes en ese momento fueron desplazados por Carabineros hacia el otro sector de las galerías, donde incluso, un hincha cayó y se rompió la cabeza.|HEAD_00
443276758|lun|La calva cabeza que exhibió hace algunos años Pedro Santos en su regreso a la hípica, tras vencer un siempre complicado cáncer, ha ido quedando en el pasado.|HEAD_00
443283282|lun|Otra área en que Maver se hace sentir con fuerza es en los llamados Nutracéuticos, con el té Adelgazul a la cabeza, y en la línea cosméticos, donde la marca fuerte es Hawaiian Tropic.|LEADER_00
443286765|lun|Vestía polera amarilla y calzoncillos negros y su cabeza estaba semienterrada en el barro.|HEAD_00
443320294|lun|Anuncia que en el futuro la ciudadanía conocerá lo que realmente pasó, ya que nadie le quita de la cabeza que dentro de la propia coalición de gobierno le tendieron una trampa.|INTELLIGENCE_00
443333436|lun|Después de soportar cuatro años de trastornos de colon irritable, problemas urinarios y de columna, dolores de cabeza y cervicales, y alteraciones de el sueño, a Susana, de 42 años, secretaria ejecutiva, le diagnosticaron lo que realmente tenía fibromialgia.|HEAD_00
443336692|lun|De visita en nuestro país, este sociólogo, síquico y clarividente dice que no sólo interpreta los colores que se encuentran contenidos en las aureolas que rodean las cabezas de las personas sino que ayuda a limpiar las para curar diversas enfermedades.|HEAD_00
450945499|lun|¿Cómo transcurrió la carrera en la cabeza y el corazón de ambos pilotos?, ¿cómo lograron desentenderse de el dolor, abstraer se a la pérdida y enfrentar el vacío?, son preguntas sin respuesta por ahora.|INTELLIGENCE_00
450952875|lun|Ivo Basay volvió a sacar la voz en una entrevista emitida anoche en el programa Futgol de el Canal 13 y no dejó títere con cabeza.|HEAD_00
450968640|lun|¿Aún con la pelea en la cabeza?|INTELLIGENCE_00
841713264|que_pasa|Si bien se le reconoce una cercanía mayor con los atletas top, algunos critican su administración a la cabeza de el CAR, pues aún no ha logrado que la infraestructura disponible sea utilizada en la medida que corresponde por entrenadores y dirigidos.|CHIEF_00
841718357|que_pasa|Pero la que más divisiones generó en el seno de la coalición de gobierno fue la presentada por el Likud, partido derechista y cabeza de la oposición, a raíz de la inclusión en el programa escolar israelí de cinco poemas escritos por el poeta nacional palestino Majmud Darwish.|LEADER_00
841719217|que_pasa|Y lo que es aún más importante, se perfila como la más probable sucesora de Wolfgang Schäuble a la cabeza de el partido, asunto que se dirimirá a mediados de abril.|LEADER_00
841719399|que_pasa|Pero haber pertenecido a las filas de la CDU en tiempos de Kohl y, a la vez, estar distante de el ex canciller no es fácil 'De los grandes nombres de el partido, ¿quién no fue protegido de Kohl? Fueron 25 años a la cabeza de la colectividad y puede decir se que todos han estado bajo su alero. Tampoco existen dentro de la CDU otros líderes fuertes aparte de Kohl', aseguró a Qué Pasa, desde Berlín, Eusebio Val, analista de el diario La Vanguardia.|LEADER_00
841728344|que_pasa|A pesar de que en estos momentos la permanencia de René Cortázar a la cabeza de TVN se ha transformado en un ítem de orden político, para algunos analistas esta disputa es un reflejo de el crecimiento y logros de la cadena estatal en su permanente batalla con el Canal 13 y de las claves que le han dado el éxito económico la independencia.|CHIEF_00
841737300|que_pasa|Las dos cabezas ejecutivas de los dos más importantes canales de la televisión chilena, Rodrigo Jordan, de Canal 13, y René Cortázar, de TVN, parecen estar a punto de pasar por la guillotina.|CHIEF_00
841744895|que_pasa|El temor es que, con él a la cabeza, el PPD termine siendo un partido opositor a Lagos.|LEADER_00
841745011|que_pasa|Aunque está cubierto de críticas, asoma la cabeza para asegurar que el suyo es uno de los mejores municipios de el planeta.|HEAD_00
841754920|que_pasa|En una serie de entrevistas de el ganador de el Premio Cervantes con testigos de los episodios que trata el libro, el propio secretario de Trujillo le contó los dolores de cabeza que le provocaban los innumerables padres que llevaban a sus hijas a los aposentos de el general para que las desvirgara.|HEAD_00

841765226|que_pasa|Las consecuencias de esta disfunción se traducen en efectos subjetivos (ansiedad, apatía, aburrimiento, depresión, irritabilidad, poca estima), conductuales (drogadicción , pérdida de apetito, consumo excesivo de cigarrillos, risa nerviosa, inquietud), cognoscitivos (incapacidad para tomar decisiones, pérdida de concentración, bloqueo mental) y fisiológicos (aumento de corticoides en la sangre y orina, dolores de cabeza, elevación de los niveles de glucosa sanguíneos, incremento en el ritmo cardíaco, molestias estomacales, sequedad en la boca, dilatación de pupilas, tensión muscular, dificultad para respirar y escalofríos).|HEAD_00
841776320|que_pasa|A pesar de que sus dos primeras medidas de peso, los proyectos de reformas laboral y tributaria, ya le han causado más de algún dolor de cabeza, la justa por la Capital Federal será su primera lid en las urnas cara a cara con el peronismo.|HEAD_00
841776837|que_pasa|A la cabeza de una coalición que agrupa a su partido, Acción por la República, y a Nueva Dirigencia -formación encabezada por el también ex ministro de Menem, Gustavo Beliz- ha logrado posicionarse como el 'líder natural' de la centroderecha argentina; y de paso, como una carta de salvación para el alicaido Partido Justicialista .|LEADER_00
841782758|que_pasa|Este último ejecutivo es hoy el segundo hombre de el Santander en Latinoamérica, cabeza por más de un lustro de Telefónica en la región y mencionado como el hispano más experimentado por estas tierras.|CHIEF_00
841788138|que_pasa|Tres meses para que se terminaran las colas en los consultorios, sentenció, de lo contrario, rodaría la cabeza de la ministra de Salud, Michelle Bachelet.|HEAD_00
841799247|que_pasa|Aquí hay una cabeza, que es el Presidente de la República, y ojalá podamos tener debates muy sustantivos en el gabinete y no intentar que unos ministros manden sobre otros, eso no es conveniente.|LEADER_00
841801620|que_pasa|Aunque tienen claro que será teoría, porque el verdadero modelo está bien guardado en la cabeza de Lagos.|INTELLIGENCE_00
841810268|que_pasa|Estos son los rostros y personalidades de los responsables de el éxito de la teleserie de TVN y de tantos dolores de cabeza de los directivos de Canal 13.|HEAD_00
841824202|que_pasa|Por el contrario, otros equipos adoptan la postura de seguir la carrera y de brindar le apoyo a quien quedó fortuitamente a la cabeza de el grupo.|LEADER_00
841833905|que_pasa|Además, a diferencia de las otras religiones mayoritarias, es la única que tiene una clara jerarquía y una sola cabeza visible.|LEADER_00
841834868|que_pasa|No obstante, con poco más de una semana por delante para los comicios de el 9 de abril, el abanderado de Perú Posible no sólo ha desplazado en las encuestas a sus dos 'compañeros' de la oposición, sino que se ha convertido en un fuerte 'dolor de cabeza' en los pasillos de el Palacio de Pizarro.|HEAD_00

Anexo 4: Selección de 120 instancias para la unidad léxica «cara»

corpusid|source|context|senseid
278905263|chile_tech|Además, presenta un sistema de copiado de transferencia electrostática en seco y un alimentador automático de originales de doble cara en configuración estándar.|SIDE_00
279206286|chile_tech|Cuando se trata de aplicaciones de tipo gráfico, la AcuLáser ofrece la opción de imprimir pequeños mailings y ediciones cortas como revistas internas, promocionales, catálogos, ofertas, etc., tanto en entornos corporativos como en empresas de servicios gráficos, edición o impresión, destacando por la flexibilidad que ofrece en el uso y manejo de el papel que puede imprimir por ambas caras.|SIDE_00
279235845|chile_tech|Fibramold, empresa dedicada a producir doorskins o caras de puertas y puertas, adquirió el software Progression Series de Macola, representado en Chile por Exxis S.A., para automatizar las áreas de Administración, Comercialización y Manufactura.|SIDE_00
279465863|chile_tech|Los usuarios pueden beneficiarse de las diferentes funciones de acabado tales como impresión de folletos y por ambas caras de el papel.|SIDE_00
279466526|chile_tech|Todos los nuevos modelos ofrecen velocidades rápidas de impresión para su clase, tanto en impresión a una cara como a doble faz, produciendo la primera página en muy corto tiempo.|SIDE_00
280160986|capital|Cara de avestruz|FACE_00
280174658|capital|Uno se topa con caras conocidas (el actor E.G. Marshall, la escritora Elena Garro) y, casi siempre , con total desconocidos abogados, jueces, doctores, editores, científicos, alcaldes, profesores.|FACE_00
280182894|capital|Miro las caras.|FACE_00
280194893|capital|El objetivo es que los estadounidenses adquieran videos, ropa y regalos con la cara de Mickey y preparen sus vacaciones en los parques de atracciones a través de internet.|FACE_00
280201593|capital|Sus acciones, sus caras y los veredictos que consiguen para sus defendidos son las razones de que los abogados sean tan mal considerados en Estados Unidos.|FACE_00
280213396|capital|Las caras nuevas son las de Sobral; Alberto Hirmas, cuya familia tiene casi un 3 %; Reinaldo Solari tío de Cúneo y accionista de Falabella y Juan Bilbao, que iría representando a Penta.|FACE_00
280224120|capital|Entre las caras nuevas de El Mercurio destacan Jorge Lesser, quien hace dos meses asumió la dirección de Empresas El Mercurio ; Felipe Lehuedé, que entró en febrero a reemplazar a Manuel Labra, asesor de la presidencia de el diario y de Agustín Edwards en sus negocios personales, y terminó ocupando la gerencia general de el diario , tras la renuncia de Fernando Cisternas.|FACE_00
280505531|capital|Todos con cara de expectación.|FACE_00
280519379|capital|A lo mejor su cara de chico travieso no es para tanto.|FACE_00
289607541|diario_financiero|La superficie exterior de la misma queda siempre hacia arriba y la cara interior bien protegida.|SIDE_00
280947210|diario_financiero|En la otra cara de la moneda, el peor desempeño y el mercado menos cotizado es Colombia.|SIDE_00
281048504|diario_financiero|Se trató de la primera reunión en la cual se vieron las caras quienes, por una u otra razón, tienen compromisos impagos tanto con la entidad en falencia como con varios de los otros capítulos legales o no de el holding de el procesado Eduardo Monasterio.|FACE_00

281048613|diario_financiero|Y aunque hubo algunas caras largas entre los acreedores más pequeños cuando se confirmó que Corfo había verificado su deuda, gracias a una resolución de el ministro Patricio Villarroel lo que le dio derecho a ser acreedor, a la salida no se registraron mayores comentarios sobre el punto.|FACE_00
281232463|diario_financiero|Agregó que ese ha sido el debate que en estos días, la opinión pública ha visto un debate que tiene dos caras. Por una parte en Chile han bajado los impuestos, porque antes se pagaba un 6 % de impuestos por las cosas que se traían de Europa, ahora no se paga (...) y por otra parte, lo que estamos viendo es cómo vamos a buscar esos pesos para invertir los en los más modestos.|SIDE_00
281309446|diario_financiero|Allí los parlamentarios DC liderados por Zaldívar dijeron en la cara a Eyzaguirre que no darían sus votos para apoyar los impuestos específicos.|FACE_00
281429017|diario_financiero|Las caras sonrientes y las palabras de elogio con que el director gerente de el Fondo Monetario Internacional, Horst Koehler, puso fin a su visita a Argentina, el mes pasado, prometían una nueva era de entendimiento entre el gobierno trasandino y el organismo financiero.|FACE_00
281435014|diario_financiero|Anteriormente, uno hacía negocios y operaciones con gente que le veía la cara.|FACE_00
281463959|diario_financiero|Sanin critica que las mujeres, sobre todo las jóvenes, están obligadas a demostrar que son buenas y que no ascienden por su cara bonita.|FACE_00
282363589|diario_financiero|En la otra cara de la moneda, los negociantes dejan para enero la búsqueda de empresas donde invertir y a el llegar esa fecha se preocupan y visitan la página web.|SIDE_00
282426406|diario_financiero|La otra cara de la medalla es que es posible que la economía de los países centrales, especialmente de Estados Unidos y Europa supere la fase semi recesiva de el último año, y ello nos permita recuperar tasas más altas de crecimiento.|SIDE_00
282464748|diario_financiero|Sin embargo, la simple aparición de caras nuevas no es garantía de éxito o seriedad en la gestión futura.|FACE_00
282480870|diario_financiero|El consumidor es muy preocupado de su cara y piel.|FACE_00
282773242|diario_financiero|Sin embargo, a mitad de el año, las caras de los actores de el negocio no eran las más felices.|FACE_00
283006436|diario_financiero|A medida que pasaban los minutos, se podía apreciar el nerviosismo en la cara de los representantes de las administradoras, dado que tras conversar con algunos tenedores de ADR se dieron cuenta que la meta de nombrar a dos directores también estaba en peligro.|FACE_00
283045889|diario_financiero|Esta vez fue la primera que la veía a fines de invierno, con frío y lluvia en Roma y Florencia y un viento helado que casi congelaba las pelucas, caretas y caras blancas que se paseaban ceremoniosamente por la Plaza San Marcos de Venecia.|FACE_00
311746626|estrella_valparaiso|El tripulante se esforzó por mantener su cara fuera de la superficie, pero pronto advirtió que no podía nadar con él hasta el muelle.|FACE_00
318109050|gran_valparaiso|El balde sólo le salía menos que medio y apenas alcanzaba para lavarse la cara y las manos.|FACE_00
318110008|gran_valparaiso|Martita salió corriendo y fue a lavarse la cara, la boca, las manos, el delantal chorreado y todo lo que guardaba ese olor inaguantable.|FACE_00
318124115|gran_valparaiso|La señora Paulita, mirando su cara sucia de lágrimas y las piernas enronchadas por los golpes, movía la cabeza con desaliento e impotencia.|FACE_00
318127008|gran_valparaiso|Se lavó las manos sudorosas y también la cara.|FACE_00
318130371|gran_valparaiso|La Srta. María se preocupó y se fijó que la niña tenía un enorme moretón en la cara.|FACE_00
318155653|gran_valparaiso|A muchas gentes ha bastado un telescopio para espantarse de la cara que vemos de la luna.|SIDE_00
318159980|gran_valparaiso|Consigue los primeros resultados sobre un pergamino, con tipos cúbicos y móviles de madera, unidos mediante un agujero lateral y con una de sus caras en relieve.|SIDE_00
318168567|gran_valparaiso|Y vulgarmente a el que ve lo que no se ve, como suele también decirse, con los ojos de la cara a el que ve de ese modo se le dice que ve visiones; que ve lo que no ve, o, a el menos, lo que solamente él está viendo y no tiene, por tanto, para los demás, realidad alguna.|FACE_00
318168621|gran_valparaiso|Los ojos de la cara no parecen que sean los mismos que los ojos de el alma, a el menos a primera vista.|FACE_00
318168650|gran_valparaiso|Con los ojos de la cara sólo vemos, naturalmente, todo lo que tenemos delante.|FACE_00
318168734|gran_valparaiso|Los poetas, desde Heráclito, no se cansaron de comparar el tiempo metafóricamente con los ríos, y a los ríos se les ve correr con los ojos de la cara.|FACE_00
439881649|lun|Nacido como Jeff Atkins y apenas un veinteañero con cara de ser una década mayor, Ja Rule fue respaldado por el prestigioso sello de rap Def Jam y en 1999 debutó con el disco Venni Vett Vecci, un trabajo de rtimos agresivos que incluía colaboraciones con otros gigantes de el hip-hop, como Jay-Z y DMX.|FACE_00
439891307|lun|Cuando un llamado requiere ser procesado con tacto empieza a balbucear, se le llena la cara de musarañas y descerraja toda clase de consultas inconducentes.|FACE_00
439897374|lun|A el hombre no se le movía ni un músculo de la cara y yo no me imaginaba cómo íbamos a negociar.|FACE_00
439897714|lun|Sin jamás haberse visto las caras se habían convertido en amigos de sangre a través de el chat y eso fue crucial para que el veinteañero trasandino no dudara en postular a Salinero, 30 años y editor de el semanario La Gironda, a el nuevo proyecto que tenía entre manos.|FACE_00
439930985|lun|Uribe negó con la cabeza y señaló Es tabaco. ¿A ver?, dijo Allendes, dando una piteada y echando le una bocanada en la cara a María Gracia.|FACE_00
439935039|lun|Tapandose con una tarjeta que luego retiró, Bosé tomó la cara de la animadora y le dio un beso en la boca.|FACE_00
439950418|lun|Las imágenes de video captadas por el camarógrafo Mitch Crooks desde un hotel cercano muestran a un oficial blanco golpeando brutalmente contra un vehículo la cabeza de un joven negro de 16 años, esposado, mientras otro policía lo golpea en la cara.|FACE_00
439966816|lun|Uno de los que llegó a poner cara de espanto fue Demetrio Marinakis, dirigente de Santiago Morning, después de visitar ayer a el ex timonel de Colo Colo.|FACE_00
541081353|el_mercurio|En dos días más presentarán en sociedad su más reciente y orgulloso proyecto la Fundación Alma Mater (que incluye dos casas de acogida) y, junto a ella, su cara más visible el Club VMA.|SIDE_00

541097430|el_mercurio|Posiblemente, la única cara que regaló algo de alegría a los millonarios fue la de Marcelo Salas, quien volvió a entrenar.|FACE_00
541186563|el_mercurio|Y todo lo grato de el momento se termina con uno o dos insoportables pinchazos en la cara o en la misma boca.|FACE_00
541191306|el_mercurio|Después de una hora de cirugía, el doctor Umaña sale de el pabellón con la satisfacción pintada en la cara.|FACE_00
547445963|el_mercurio|Sin embargo, el fenómeno no será visible en todo el país, pues los únicos que podrán observar la cara oscura de Venus, que aparecerá como una pequeña mancha esférica en la superficie de el Sol, serán los habitantes de Arica, Iquique y localidades de el altiplano nortino.|SIDE_00
478396484|el_mercurio|La cara poniente es ondeada y allí están todas las áreas de servicios de los bomberos.|SIDE_00
478464947|el_mercurio|En la otra cara de la moneda, y pese a no conseguir que las autoridades acogieran sus demandas aun cuando reconoció el error de el gremio de bloquear las arterias de Santiago; el presidente de el Consejo Superior de el Transporte, Manuel Navarrete, repitió el discurso de las últimas semanas en cuanto a que el sector que dirige no está en contra de la modernización que desea impulsar el Gobierno mediante el Plan Maestro de Transporte Urbano; y señaló que de ahora en adelante habrá dos caminos, uno para conversar la licitación de el próximo año, y el otro para seguir denunciando las irregularidades y falencias de el proceso Metrobús.|SIDE_00
498142020|el_mercurio|Su cara de desconcierto me conquistó en medio segundo.|FACE_00
498144729|el_mercurio|Sólo sonreí, y educadamente hice una seña, me di media vuelta y me fui... no podía dejar que vieran mi cara de vergüenza, ni su color.|FACE_00
498254252|el_mercurio|Un engendro extraño vestido con harapos, con la cara deformada por alguna explosión, por alguna enfermedad que no conozco.|FACE_00
498274158|el_mercurio|'No, por dios. Este señor es el Príncipe Andrés!', me dice la misma mujer mirando me con cara de asco.|FACE_00
498277224|el_mercurio|A medida que avanzo empiezo a sentir un olor raro, una fetidez que contrae los músculos de mi cara.|FACE_00
498297572|el_mercurio|Durante la celebración de la liturgia, los presentes vieron llegar a una joven que descubrió la cara de el muerto y lo besó, quedando allí reclinada hasta que en el momento de iniciar se el entierro fueron a apartarla y vieron que se trataba de Isabel de Segura.|FACE_00
608044736|la_tercera|El que volvió a las prácticas es Esteban Valencia Tenía una contractura en el isquiotibial (cara posterior de el muslo) izquierdo.|SIDE_00
608057881|la_tercera|En contraste, Tomás Jocelyn Holt se tomaba la cara e Ignacio Walker mostraba un dejo imborrable de amargura.|FACE_00
608097370|la_tercera|Cuando Vicuña señaló que sólo tendría actores y no caras bonitas, te molestaste mucho.|FACE_00
608146244|la_tercera|Porque lo que uno supo a través de las declaraciones de Rozental es que le dolía la cara lateral exterior de la rodilla, lo que puede deber se a una serie de otros problemas que no tiene nada que ver con la lesión original.|SIDE_00
608146677|la_tercera|Es un ligamento demasiado específico que va de el menisco externo, que pasa por delante de el ligamento cruzado y va a dar a la cara posterior de el cóndilo femoral.|SIDE_00
608213390|la_tercera|Aquel 24 de abril, cuando se inaugure la Junta Nacional extraordinaria, todos se verán las caras en la DC.|FACE_00
608222332|la_tercera|Roja, con tres caras nuevas.|FACE_00
608222572|la_tercera|Así como aseguró que nadie está descartado, porque la lista definitiva sólo la entregaré a fines de abril o la primera semana de mayo, Acosta argumentó la nominaciones de las tres caras nuevas.|FACE_00
697465947|la_tercera|Tocarse la cara frecuentemente.|FACE_00
697466063|la_tercera|Si una persona mira directamente a los ojos mientras comenta un logro reciente, pero a el mismo tiempo no puede dejar de tocarse la cara, eso es una pista que indica que no todo es verdad.|FACE_00
697477598|la_tercera|En la otra cara de la moneda, el dólar se alzó como el gran beneficiado de las turbulencias internacionales.|SIDE_00
697547732|la_tercera|De poca estatura, la cantante lleva en la cara la imagen exacta de su padre y aparece siempre muy maquillada, especialmente en las sombras de los ojos, tal como lucía su madre cuando se casó muy joven con Elvis.|FACE_00
697576952|la_tercera|Salió por la puerta de pasajeros internacionales y no quiso comentar nada, se tapaba la cara cuando le sacaban fotos y trataba de evadir a la prensa, relata el reportero de El Universo.|FACE_00
697584907|la_tercera|Le quebraron un vaso en la cara y quedó todo ensangrentado, con 13 tajos.|FACE_00
697585499|la_tercera|Bajan después de una hora y cerca de las tres de la mañana suben de nuevo, con cara de felicidad, y hasta el otro día.|FACE_00
697590707|la_tercera|A Jennifer Warner le dicen que tiene cara de caballo y que casa en el hipódromo.|FACE_00
691844729|la_tercera|La nueva regulación establece además que las advertencias sobre daño para la salud cubran el 30% de la cara frontal de las cajetillas.|SIDE_00
675291487|la_tercera|Tenía la cara desencajada.|FACE_00
675406702|la_tercera|Durante 53 días no conocerá la cara de sus captores todo contacto se hará por escrito, sobre billetes de reales.|FACE_00
675495864|la_tercera|A las oficinas de la administración de Saville Row llegan caras nuevas, entre ellas está Felipe Bertin, con una línea más joven y la idea de complementar la sastrería.|FACE_00
675506162|la_tercera|'Teníamos sólo un torneo y un amistoso previo. Y jugar con dos mil personas igual afecta . Ahora hay que olvidar se de esto y ganar todo lo que viene. Y ojalá que me toque España en semifinales para romperles la cara', dijo Pablo Jara, el más experimentado de el equipo nacional.|FACE_00
675669404|la_tercera|La cara de Alvaro Fillol cuando entró a el court central de el Club Naval Las Salinas era de preocupación.|FACE_00
675740793|la_tercera|El locutor explicó que cuando ello ocurre sale de el estudio y corre a mojarse la cara y el pelo.|FACE_00
675763337|la_tercera|Habían transcurrido más de 30 horas desde su despegue de Nueva York y, agotado, luchaba por vencer el sueño abofeteandose la cara.|FACE_00
675825570|la_tercera|A el margen de el glamour de ver caras conocidas ingresando a el mundo de el vino, el presidente de Chilevid, Rodrigo Alvarado, plantea que es muy positivo que más empresarios incursionen en el rubro.|FACE_00
675898381|la_tercera|Arranca entusiastas aplausos de la platea y Myriam agradece con cara de sorpresa y se lleva las manos a el pecho.|FACE_00

675907055|la_tercera|La otra cara de la moneda ha sido Roberto Carlos, que, por el contrario, sí podrá jugar de titular.|SIDE_00
675916998|la_tercera|Aprovechando el éxito de figuras jóvenes, 'caras nuevas' postularán a un cargo en la directiva.|FACE_00
676005900|la_tercera|Alicia Pedroso volvió de Cuba muy delgada y da entrevistas en los pasillos con cara muy seria, pero la de peor gusto es Patricia Manterola, la co-animadora que enfrenta el tumulto de la llegada con un jockey Burberry, cartera Vuitton y un buzo deportivo de esos para hacer las compras de el supermercado.|FACE_00
676031776|la_tercera|El equipo de expertos de las compañías japonesas Matsushita Electric Industrial (Panasonic-Technics), Sony ,Toshiba, Hitachi, Pioneer y Sharp , la holandesa Philips, la francesa Thomson Multimedia (RCA) y las surcoreanas LG Electronics y Samsung Electronics, desarrollarán también equipos de grabación y discos DVD con una memoria seis veces mayor a la de los aparatos actuales, con lo que será posible almacenar 30 gigabites en una sola cara o el equivalente a 40 horas de imagen y sonido de la televisión convencional o dos horas de video de alta calidad.|SIDE_00
676047786|la_tercera|De esta manera, Valdivia no buscará un refuerzo interior, por lo que Leonel Méndez (ex Provincial Osorno) será la única cara nueva en el equipo que dirige Marcos Guzmán.|FACE_00
676090225|la_tercera|Además, sostiene Gómez, una cara de la moneda es el alza en el valor de el terreno, pero, por otro lado, se ha abaratado el costo de la urbanización (alcantarillado, electricidad, pavimentación) por el mismo efecto de la mayor cantidad de conjuntos residenciales que han llegado a esas comunas.|SIDE_00
676090862|la_tercera|El domingo se verán las caras.|FACE_00
676130670|la_tercera|Según fuentes gubernamentales de Estados Unidos, la cinta muestra a Pearl hablando con alguien como si fuese una entrevista cuando imprevistamente una persona a la cual no se le ve la cara aparece por detrás, toma bruscamente de la cabeza a el periodista y le corta el cuello con un cuchillo.|FACE_00
676206076|la_tercera|Las nuevas caras cruzadas son el seleccionado paraguayo Jorge Campos y los chilenos Rodrigo Barrera, Eduardo Arancibia, Carlos Verdugo, Carlos Tapia, Nelson Cossio, Fernando Solís, Branco Matijevic y Albert Acevedo.|FACE_00
676206565|la_tercera|El entrenador Víctor Hugo Castañeda cruza su brazo izquierdo, apoya el codo en éste y se toca la cara.|FACE_00
676206922|la_tercera|También lo hizo Castañeda, que dejó de tocarse la cara.|FACE_00
676209933|la_tercera|De rodillas, Nicolás Massú se toma la cara y el ¡vamos! que ya ha patentado en el Buenos Aires Lawn Tennis brota espontáneo de su boca.|FACE_00
676282541|la_tercera|También está la Conejita Playboy, Karla Brandt, que está de portera, forrada en brillos y escotes brillantes estilo segunda mano y pone cara de mala y 'Chavito' es el más famoso que se retuerce en medio de codazos, empujones y roces sin querer porque este lugar está lleno.|FACE_00
676335002|la_tercera|Ni siquiera a lo Rita Hayworth, con vestido brillante de tajo escandaloso, rojo intenso y cara de gata, bailando un número poco ensayado de éxitos de discoteca playera y con los chicos de el ballet elevando la como Rafaella Carrá, sedujo a la gente.|FACE_00
676359913|la_tercera|Silva explicó que es decisión de las personas mostrar o no sus caras.|FACE_00
676389412|la_tercera|La primera vez que se vieron las caras fue el año pasado en Austria, cuando el europeo se impuso por 7-6 y 6-2 en la arcilla de Sainkt Pölten.|FACE_00
676393799|la_tercera|Javiera permanece bajo observación médica en casa de su madre, debido a las lesiones que presenta en su cara y cuerpo, además de varias costillas rotas.|FACE_00
676404254|la_tercera|Fuentes de el departamento de Sanidad de el Gobierno regional vasco informaron que Cabezudo tiene numerosos cortes en la cara y una fractura en un pie de la que podría ser intervenida, mientras que Torres sufre cortes por cristales y problemas en un oído.|FACE_00
676421054|la_tercera|En tal sentido una de las pruebas más esperadas será el slalom en categoría open , donde nuevamente se verán las caras los especialistas Rodrigo Miranda y Rodolfo Guzmán, quienes han protagonizado estrechos duelos en las fechas anteriores.|FACE_00
837739196|que_pasa|Un ejemplo de ello es el gran mural que pintó en la cara poniente de el edificio de Compañía con Bandera, con esa imagen de el caudal de el río Mapocho que se nos viene encima, corriendo sobre un azaroso plano inclinado, mientras nuestros pies de distanciados espectadores descansan sobre una superficie perfectamente horizontal.|SIDE_00
838252388|que_pasa|A pesar de su corta edad y su cara de niña, Josefowicz tiene una larga trayectoria que ha puesto a prueba en escenarios como el Carnegie Hall de Nueva York y el Royal Festival Hall de Londres.|FACE_00
838308347|que_pasa|Desde 1992, su cara ha aparecido en las portadas de Cosmopolitan, Elle, Marie Claire y Glamour, entre otras, y hoy prepara su propio calendario de fotografías en traje de baño.|FACE_00
838455574|que_pasa|Raúl Ruiz, que tempranamente deslindó su obra de cualquier parentesco con otros connacionales, relata que le importaba en esta cinta hacer planos cercanos de los actores, filmando lugares como la frente y la nuca, que dejan a el espectador una sensación incómoda, como si le estuvieran pasando una mano por la cara.|FACE_00
838486077|que_pasa|Cada vuelo, le dijo a sus colegas de el Kremlin, 'es una cachetada en la cara'.|FACE_00
799801196|mostrador|Con cara de niño a los 29 años, sigue vistiendo se con su ropa habitual, a pesar de que la productora estadounidense de Los otros insistiera en comprar le una chaqueta de Miyake y unos zapatos Gucci que no me los voy a poner nunca más.|FACE_00
799813257|mostrador|Cada vez que la sicóloga decía la palabra orgasmo, había un sospechoso corte que incluso provocó quiebres de ejes ella aparecía hablando con la cara hacia la izquierda y luego, sin un paso armónico como corresponde, salía con su rostro hacia la derecha.|FACE_00
799842852|mostrador|Cáceres le dio un manotazo en la cara a el volante Patricio Luna, fue expulsado y dejó a su equipo con un jugador menos cuando quedaba más de una hora de encuentro.|FACE_00
800071592|mostrador|El presidente de el Banco de el Estado, Jaime Estévez, explicó que esta medida tiene dos caras, porque cuando se pagan intereses también se cobra por las cuentas corrientes.|SIDE_00
800077680|mostrador|Busca la distancia con esta tierra, desde la que le llegan esos postes restantes, esas cartas no contestadas, devueltas, la cara opuesta, busca encontrar se en el desencuentro con Chile.|SIDE_00
800078018|mostrador|Para cerrar la idea, volviendo a el epígrafe de Yourcenar, la cara opuesta de una cárcel que puede tener por barrotes las ilusorias fronteras de el mundo entero.|SIDE_00

800129115|mostrador|La preocupación en el entrenador de Agassi se notaba en su cara y Massú se estaba convirtiendo en una gran sorpresa para el público.|FACE_00

Anexo 5: Selección de 120 instancias para la unidad léxica «carta»

corpusid|source|context|senseid
278704965|chile_tech|Los educadores, los profesionales y las pequeñas empresas deben producir materiales colaterales de calidad profesional, desde folletos, cartas, y reportes en tiempo limitado, con un escaso presupuesto y sin experiencia en diseño.|LETTER_00
280068127|chile_tech|Los usuarios pueden jugar, comunicarse y conectarse en línea para juegos interactivos de cartas, tableros, fichas y otros|CARD_00.
278736013|chile_tech|Según la carta de intenciones, firmada en diciembre de 2000, las tres empresas compartirán sus recursos para prestar servicios de gestión de viajes a más de dos millones de clientes.|LETTER_00
278863520|chile_tech|Las tres compañías orientadas a la aeronavegación mundial anunciaron la firma de una carta de intención para desarrollar una iniciativa empresarial global para proveer comunicaciones de banda ancha y servicio de información digital para la aviación comercial.|LETTER_00
278913717|chile_tech|En febrero de el presente año, CasaChile.cl publicó una carta de un lector de Australia que solicitaba ayuda para ubicar a un hermano.|LETTER_00
278931387|chile_tech|Tras la situación, Network Solutions bloqueó la petición de el dominio, enviando una carta a Chile Networks, en la cual informaba haber acogido la petición de Chilenet de detener el uso de los dominios, restandoles 37 días para apelar.|LETTER_00
278955422|chile_tech|Como un paso fundamental en el compromiso de Cisco Systems con la educación de Chile, el Presidente y CEO de la compañía John Chambers y el Presidente de la República de Chile, Ricardo Lagos, firmaron una carta de compromiso a través de la cual Cisco entregará un grupo de herramientas para facilitar la formación y el perfeccionamiento de los profesores chilenos.|LETTER_00
278955447|chile_tech|La carta considera la donación, a través de el Cisco Learning Institute (CLI), de una solución de e.Learning (aprendizaje en red) basada en la Web, que permite la creación de cursos y entrega herramientas de administración para un ambiente personalizado de e.Learning para la disponibilidad de contenidos y la evaluación de sus usuarios.|LETTER_00
278955612|chile_tech|La firma de esta carta foma parte de un constante compromiso de Cisco Systems con la educación, tanto a nivel global, como en Chile.|LETTER_00
279189100|chile_tech|El sistema de código de barra, bajo el concepto de módulos automatizados, permite un control de inventario fidedigno de la trayectoria que siguió la carta desde que se despachó hasta que llegó a destino.|LETTER_00
279189129|chile_tech|Es decir, tanto Envíos, como el cliente, pueden realizar la búsqueda y seguimiento de las cartas en cualquier momento de su recorrido.|LETTER_00
427214016|la_cuarta|La rubia tarotista, que ahora lee las cartas en el programa de Ivette Vergara, debutó en pantalla como la sirena de 'Corazón Partío' (TVN).|CARD_00
427214159|la_cuarta|Hace dos años se casó con un suertudo chileno que no tiene nada que ver con el mundo esotérico. 'Él es sumamente terrestre, pero le tiene mucho respeto a esto', cuenta mientras baraja las cartas en una mesita de su local.|CARD_00
427257926|la_cuarta|la cuarta: El mago tenía una carta bajo la manga|CARD_00
427257948|la_cuarta|El mago tenía una carta bajo la manga|CARD_00
427306563|la_cuarta|Dicha labor recaerá en el Ministerio de Relaciones Exteriores, donde ya se halla la carta rogatoria para su tramitación, 'a fin de que se sirva ordenar las diligencias diplomáticas que sean necesarias', indica el fallo de la Suprema.|LETTER_00
427307697|la_cuarta|Según estas fuentes, el informe sólo sería un recurso inventado por el ex detective para entrar a el negocio de el soplonaje a alto nivel, y ocupar el vacío que dejó en este mercado la dupla formada por Lenin Guardia y Humberto López Candia, que quedaron fuera de combate tras ser encanados por el caso de las cartas bombas.|LETTER_00
427307991|la_cuarta|En el lugar Liliana dejó una carta dirigida a sus hijos.|LETTER_00
427318190|la_cuarta|Caballero tuvo el mejor taco como si le hubiesen dado un taco de manos de Thalía quedó el volante argentino nacionalizado mexicano Gabriel Caballero, quien dijo sentirse en las nubes por su convocatoria para formar parte de la selección cuate, que irá a charrasquear a el Mundial asiático. 'Es un sueño caro; me disponía a ver el Mundial por tele y de repente llegó la convocatoria', dijo Caballero, jugador de los Tuzos de el Pachuca y que en diciembre pasado recibió su carta de ciudadanía azteca.|LETTER_00
427323868|la_cuarta|Lorena Osorio, de 23 años, quien tiene discapacidad física, no pudo aguantar las lágrimas cuando recibió el premio especial a la superación por haberse trasladado desde Requínoa, en la Sexta Región, a Santiago para asistir a el curso. 'Por una carta nos enteramos que existía la fundación y mi hermana que vive en Santiago me recibió en su casa. Ahora estoy feliz por todo lo que he aprendido, pero me hace falta encontrar pega, porque para eso uno se capacita', dijo junto a su querida hermana.|LETTER_00
427331282|la_cuarta|Informe de Contraloría afecta orden público de la nación', reza apocalíptica carta de galenos.|LETTER_00
427331618|la_cuarta|El informe ha servido para armar un escándalo público con grave perjuicio para el país y que por cierto afecta el orden público de la nación, a el amenazar la destrucción de la cohesión interna de el sistema de salud estatal', dice la carta colorienta.|LETTER_00
427343189|la_cuarta|Correos entregó cartas y estelas a ladrillos de ex Peni y también a pelados de el Buin.|LETTER_00
427343227|la_cuarta|Los pelados de el Regimiento Buin aprovecharon de escribir y enviar altiro las cartas a sus progenitoras.|LETTER_00
415118538|la_cuarta|La movida fue en las categorías magia de cerca (juegos con cartas, monedas y artículos pequeños) y de salón o escenario (trucos mentales, ilusionismo y escapismo).|CARD_00
415409167|la_cuarta|Por un error, asimilamos a el rol a 'Magic' y a otros juegos con cartas que sería 'como comparar los juegos de salón con el caballito de bronce', dicen los roleros.|CARD_00
424268414|la_cuarta|Por ejemplo, un juego de cartas o de otro tipo donde siempre exista un ganador y un perdedor.|CARD_00
515594379|el_mercurio|La carta demuestra vigencia, coherencia en la elección de las etiquetas y ofertas para todos los gustos.|LETTER_00

515594405|el_mercurio|El año pasado fue elegida como la mejor carta de vinos de Santiago, por la 'Guía de vinos de Chile' (editada por Paula Comunicaciones).|MENU_00
515594930|el_mercurio|En la categoría 'Mejor carta de vinos' hubo dos personas se abstuvieron.|MENU_00
515596029|el_mercurio|De la carta donde ningún plato de pescado está originalmente concebido para tintos, advierte sugiere indagar en las combinaciones entre el kimiday con salsa de crema y ostiones (6.500) junto a un pinot noir reserva, pues la barrica le aporta la mantequilla necesaria; o la sabrosa tilapia a la oliva (6.300) con malbec, ya que los dos últimos van muy bien.|MENU_00
515596311|el_mercurio|Incluso, ya han llegado a las cartas de los restaurantes de mantel largo con altas dosis de creatividad y sofisticación.|MENU_00
515599184|el_mercurio|Aunque el grueso de la carta se centra en la parrilla, y dentro de ésta en los cortes de vacuno, se ofrecen también en esa forma embutidos, interiores, cordero, cerdo, pollo y pescados, y además unas pocas entradas. innecesarias dado el tamaño de las porciones una decena de platos cocinados con pastas, pescados y por cierto carnes y varias ensaladas.|MENU_00
515600079|el_mercurio|El actual chef, Cristián Rebolledo, ofrece una carta claramente inspirada en esa cocina, pero con aportes personales interesantes.|MENU_00
515601757|el_mercurio|El público tiene dos opciones: el menú didáctico (2.900), que es el que elaboran los alumnos de último año, y los platos de la carta.|MENU_00
515601818|el_mercurio|A la carta hay seis variedades de carnes (como lomo vetado, 2.800 o pechuga de pollo deshuesada, 2.800), cuatro alternativas de pescados (como salmón en salsa de camarones, 3.900 o trucha arlequín, 4.500) y fetuccini a la marinera (3.600) o Alfredo (2.600).|MENU_00
515602697|el_mercurio|La carta ofrece numerosos mariscos y carnes, con variadas salsas y algunas especialidades de estilo cajoun; y los postres, todos hechos en casa incluidos los helados, se presentan en un gran naipe cuyas cartas llevan las respectivas fotografías.|MENU_00
515603090|el_mercurio|A un costado de el Mercado de Providencia nos topamos con este minimalista restaurante que tras su sicodélica entrada nos invita a disfrutar de una decoración moderna, pero acogedora, música electrónica y una contundente carta que combina sabores, especies y aliños a la perfección.|MENU_00
515603389|el_mercurio|Se puede empezar desde las 8:00 AM (incluso domingos) con un buen café y un crujiente croissant hecho en casa, para continuar a lo largo de el día con la inédita y sorprendente selección de sabores de la carta, en cuya preparación se ha cuidado hasta el último detalle.|MENU_00
515603848|el_mercurio|Pero su mayor gracia está en la carta: aquí se ofrecen pescados chilenos que no se encuentran fácilmente, como el rollizo, la cojinova, la palometa, el vilagai y vieja.|MENU_00
515604227|el_mercurio|Este restaurante, ubicado en Vitacura y que se inauguró tan sólo hace tres meses, sorprende por la variedad de platos que su carta contempla.|MENU_00
515604260|el_mercurio|Con una carta extravagante, en la que se encuentran distintos sabores, como por ejemplo, sushi con fruta, envueltos en queso crema, fritos, calientes y enteros, este local hace notar la diferencia.|MENU_00
537395192|el_mercurio|¿Llegarán otros juegos? -pregunta ansioso en la tienda Warner, un gringo de dos metros que ya ha comprado cartas, una lámpara, dos camisetas, una manta, varios cuadernos y, por supuesto, la edición completa de las aventuras de el mago con anteojos.|CARD_00
531294538|el_mercurio|La fiesta de el Nacional, además, considera la inauguración de el Museo de Mitos y Leyendas, con la historia, principales hitos y datos técnicos de el único juego de cartas de estrategia chileno.|CARD_00
582693602|la_segunda|Según el relato que está en manos de la SVS, en la carta fechada el 15 de noviembre, AES cuestiona que si los ADR no emiten una posición sobre los temas a tratar en la junta la aprobación de la alianza, propuesta por TotalFinaElf será el presidente de el directorio quien deberá decidir por ellos.|LETTER_00
582696851|la_segunda|El coronel Gran López agrega que para ello habría que traducir los textos pero no se hace necesario cambiar las cartas que aparecen reseñadas ya que están a modo de ejemplo y se han construido usando normas internacionales.|LETTER_00
582722512|la_segunda|También todas deben tener papelería, y entre mandar una carta fea u otra bien diseñada.|LETTER_00
582739130|la_segunda|En enero de 2000, los abogados Patricio Ossa y Guillermo Hevia enviaron una carta circular a todos los accionistas de Campos Chilenos.|LETTER_00
582739226|la_segunda|Curiosamente, y esto lo pudo comprobar la investigación que realizó el Colegio de Abogados, en la cual se hizo parte el ejecutivo de Pathfinder Félix Bacigalupo, pocos días antes de el envío de dicha carta se había montado en la red de internet una página web denominada 'accionistascamposvirtualeave.net' que luego tomó el nombre de 'defiendeteonline.com'.|LETTER_00
582739546|la_segunda|En este asunto lo que se ha dicho es que hay una intención de incitar a el litigio en circunstancias de que yo ya era parte de un litigio contra Félix Bacigalupo en la Superintendencia de Valores y Seguros y había otros dos juicios iniciados en tramitación. Esto se contrapone a lo que dice el fallo en que pareciera que a raíz de mis cartas se hubieran iniciado los juicios. Eso no es efectivo', explica Ossa.|LETTER_00
582777555|la_segunda|En este sentido, informó que le enviaron dos cartas a Longueira, sin recibir ningún tipo de respuesta.|LETTER_00
582778125|la_segunda|En la oportunidad se entregará una carta dirigida a el intendente Jaime Tohá y a el Presidente de la República, argumentando la necesidad de que se apliquen las sobretasas.|LETTER_00
582788080|la_segunda|IVAX envió una carta a Laboratorios Chile comunicandole que aprobaba el due diligence que realizó sobre la empresa y sus filiales, con lo cual se ratifica que sigue adelante la operación en los términos y precios acordados.|LETTER_00
582790441|la_segunda|Enviaron carta de intenciones a el organismo.|LETTER_00
582790504|la_segunda|Una carta de intenciones le envió el Gobierno brasilero el pasado 14 de junio, por intermedio de el ministro de Finanzas, Pedro Malán, y el presidente de el Banco Central, Arminio Fraga, a el presidente de el Fondo Monetario Internacional Horst Köhler, dando le detalles de las políticas que Brasil pretende implementar en el contexto de su solictud de ayuda a el FMI.|LETTER_00
582790743|la_segunda|Esta carta sigue a otra, fechada el pasado 5 de junio, en que los mismos dos personeros dan a conocer un balance de la situación económica brasileña y de las políticas que se han seguido, claro que dado el escenario que ha enfrentado esta semana Brasil, queda un poco desfasado.|LETTER_00
582798729|la_segunda|Vinos... y una buena carta.|MENU_00

582798945|la_segunda|Claro que para estos negocios por el momento no hay nada nuevo en la carta... por ahora el menú sólo incluye explotar a el máximo los negocios inmobiliario y vitivinícola, que es donde concentran su know how.|MENU_00
558633651|la_segunda|Más de una vez lo hemos dicho: la renovación de la carta en un antiguo restaurante acreditado suele traer más aportes que la inauguración de un local aventurero.|MENU_00
561583478|la_segunda|La carta fue enviada supuestamente por Marcos a un cantante, Angel Luis Lara, alias 'El Ruso', con motivo de la inauguración de un local prozapatista en Madrid, Aguascalientes.|LETTER_00
561913925|la_segunda|A el almuerzo, la primera carta se reemplaza por un menú ejecutivo (3.500).|MENU_00
562052438|la_segunda|Son muchas las cartas que reciben diariamente, como si el público supiera que (a algunos de ellos) esos 15 minutos de fama les cambiaría la vida.|LETTER_00
563595226|la_segunda|Todo casi perfecto, si no fuera porque en un local de esta categoría no es posible aceptar una carta con tantos errores de ortografía.|MENU_00
563565791|la_segunda|Qué hacer: Aprende un refrán, un trabalenguas o un truco con cartas.|CARD_00
563569162|la_segunda|En eso tuvimos un acierto importante que fue el jugar nos una carta con Vanessa Miller como Bárbara.|CARD_00
433260963|lun|Si el retoño comienza a reclamar por el juego de cartas intercambiables, deben tener se a mano, para empezar, 10.500 pesos.|CARD_00
433260554|lun|Los juegos de cartas son una de las novedades para los regalones.|CARD_00
433260436|lun|El señor de los anillos, en tanto, con sus hobbits, elfos, enanos y humanos, ya tiene en tiendas una colección de figuras de acción, cartas coleccionables y juegos de estrategia, en lo que se supone sólo es el comienzo de la arremetida.|CARD_00
442976051|lun|Con ese juego de cartas, sólo faltará Jorge Hevia para tener el póker completo.|CARD_00
446648204|lun|Se inauguró la nueva tienda Salo, donde se pueden encontrar desde juegos de estrategia como 'Mitos y leyendas', las cartas de 'El señor de los anillos', hasta álbumes descontinuados.|CARD_00
456988057|lun|Pero el juego requiere, en otros niveles, mazos de 60 cartas, por lo que hay que comprar sobres adicionales.|CARD_00
52904639 |la_tercera|Esta contenía 60 cartas de personalidades destacadas y detalles sobre la vida en esa ciudad en el año 1901.|LETTER_00
652904955|la_tercera|En su interior se encontraron cerca de 100 cartas, decenas de fotografías y gran cantidad de recuerdos de la época, que en el futuro serán exhibidos en la Biblioteca de el Tutt College.|LETTER_00
652904981|la_tercera|Entre las cartas recuperadas existe, incluso, una de el entonces vicepresidente de Estados Unidos Theodore Roosevelt, que sólo semanas después de escribir la misiva, tuvo que asumir la presidencia de ese país tras el asesinato de el entonces mandatario William McKinley.|LETTER_00
652905166|la_tercera|Además, hay una carta de el entonces presidente de la YWCA, la asociación cristiana de jóvenes, para quien ocupara su lugar el 2001.|LETTER_00
652905215|la_tercera|Esto, porque varias cartas no se encuentran en muy buen estado y serán traspasadas a microfichas por los funcionarios de la biblioteca de la Universidad.|LETTER_00
652923447|la_tercera|Paulatinamente ha ido desapareciendo la costumbre de comunicarse a través de una simple carta.|LETTER_00
652923481|la_tercera|Ya casi nadie manda cartas de tipo personal para comunicar se a distancia.|LETTER_00
652923536|la_tercera|El sociólogo Carlos Catalán señala que quienes cuentan con internet dejaron definitivamente de enviar cartas a través de el correo tradicional.|LETTER_00
652923588|la_tercera|Es un sistema donde se deja de lado la formalidad de las cartas, se utiliza un lenguaje mucho más coloquial y, por lo tanto, la comunicación se hace más fluida y flexible. 'Esto se traduce en una mayor cercanía entre las personas que se están comunicando', explica Catalán.|LETTER_00
652923730|la_tercera|Por su parte, el sociólogo Eduardo Rodríguez, de Feedback Comunicaciones, opina que si bien el e.mail permite tener una comunicación más rápida y, a veces, casi instantánea, es mucho más fría e impersonal. 'Cuando las cartas están escritas a mano y son firmadas por el remitente todo se hace más cercano para el destinatario', dice.|LETTER_00
652923993|la_tercera|Posiblemente, durante la Navidad que recién pasó, gran parte de los chilenos y casi con seguridad los que están conectados constantemente a internet recibió más de un saludo a través de e.mail, reemplazando a las habituales tarjetas y cartas de papel que llegan por medio de el correo tradicional.|LETTER_00
652924013|la_tercera|Este es un ejemplo claro para graficar que las tradicionales cartas personales aquellas que más de una vez adquirieron ribetes míticos en la literatura clásica en boca de poetas u otro personaje legendario han ido perdiendo peso entre las costumbres de la gente, que ha optado por la ayuda de la tecnología moderna que entrega alternativas para comunicarse a distancia en forma simple y óptima.|LETTER_00
652924096|la_tercera|Actualmente, en Chile se movilizan unas 350 millones de piezas postales a el año, lo que significa un promedio cercano a las 23 cartas por persona, una cifra pequeña si se compara con otros países.|LETTER_00
652924122|la_tercera|Por ejemplo, en Canadá, cada persona recibe un promedio de 370 cartas a el año.|LETTER_00
587105813|la_tercera|Su carta de vinos consta de 14 viñas tanto emergentes como tradicionales, cuyos precios fluctúan entre los 3.100 y los 6.900.|MENU_00
587107324|la_tercera|Amplia carta de vinos a un costo promedio de 8.000 la botella.|MENU_00
798832294|mostrador|Algunas sugerencias de la carta son, en las entradas, el Pulpo a el Olivo (3.200), Ocopa Arequipeña (papas a el vapor con salsa de huacatay, maní y queso, 2.200); de fondo, Calamar a el vino tinto (reducción de vino, champiñón y calamares, 4.200), Parihuela de corvina con mariscos (sopa gruesa con cebolla, tomate, cilantro, fondo de corvina y vino blanco, 4.600), Tallarines a la peruana (filete, cebolla, tomate, cebollín, soya, jengibre, 4.400) o Seco de cordero con frejoles a el estilo peruano (4.000).|MENU_00
798832800|mostrador|El viernes 27 y el sábado 28 de el mismo mes se abrirá a todo público con la carta convencional pero con la especial misión de conducir la celebración.|MENU_00
798838565|mostrador|En igual sentido que lo referido con relación a la 'Carpeta nro.1', se observan de las cartas y notas contenidas en este legajo, la constante actividad de las personas sospechadas, siendo que unos son los operadores de las órdenes que imparten los otros, y todos ellos dentro de la

escala jerárquica de la Dirección de Inteligencia Nacional de Chile (DINA), como así su continuadora, la Central Nacional de Inteligencia (CNI).|LETTER_00
798838876|mostrador|A fojas 25.27 de la misma, de fecha 09.056.78 Carta manuscrita: Estimado Luis Felipe (ARANCIBIA). Firma GEORG (WILLICKIE), la que textualmente dice 'Estimado Luis Felipe: Te doy las gracias por tu atenta carta que la recibí esta semana.|LETTER_00
799359726|mostrador|En la carta, Lupe Zevallos, hermana de el fundador de la compañía, Fernando Zevallos, expresa que 'este reinicio no debe ser transitorio, sino definitivo hasta que se produzca un fallo de el más alto nivel de el Poder Judicial chileno'.|LETTER_00
799372480|mostrador|En un contacto hecho ayer domingo con Radio Chilena, el dirigente sindical señaló que 'hemos agotados todas las instancias correspondientes. Se le mandó carta a el Presidente, a el intendente de la cuarta Región y así lo han hecho los colegas de las distintas regiones de el país. Hablamos con diputados y con el presidente de el Sistema de Administrador de Empresas (SAE), Felipe Sandoval.|LETTER_00
799380167|mostrador|González , jefe de el Ministerio Público (MP) , dijo que el funcionario se encontraba de vacaciones desde la semana pasada y que su carta de renuncia la recibió hoy , aunque Zeissig abandonó Guatemala ayer , domingo , rumbo a El Salvador.|LETTER_00
799403364|mostrador|En la medida que se conozcan los resultados de las demás compañías, se verá quienes jugaron bien sus cartas y quienes reprobaron la materia.|CARD_00
799404757|mostrador|De hecho , en una carta fechada el 19 de junio de este año , el Presidente Ricardo Lagos envió un saludo a la comunidad escolar por su 69 aniversario.|LETTER_00
799441385|mostrador|Aunque se ha dedicado a la fotografía publicitaria spa 'hay que vivir' y a las clases, el artista ya ha publicado un libro con una serie sobre la ex cárcel pública de Santiago titulado Despojos de sueños. 'En 1993 cuando desapareció la cárcel de Santiago, me vino la curiosidad en entrar a ella, aunque ya había ingresado alguna vez a una cárcel visitando gente, pero siempre me imaginé cómo eran estando vacías. Esa vez lo que más me robé fueron los muros escritos con la vida de los presos, sus oraciones, sus cartas, sus amores, ¡todo estaba ahí!', contó.|LETTER_00
799444029|mostrador|Las cartas señalan que tu pena va a pasar y para ello es necesario que armonices con el pasado, debes tratar de quedar en paz con todo aquello que te significó dolor.|CARD_00
799480285|mostrador|Carta enviada por Lan Chile a la Fiscalía Nacional Económica (24.07.2001) 'La Comisión Resolutiva hizo suya la proposición de Lan Chile y en estos momentos la compañía está proponiendo a la autoridad, un mecanismo que garantice el cumplimiento de la forma de operar ofrecida', dice el comunicado.|LETTER_00
799499166|mostrador|Mientras ésta se preparaba llegó a poder de Alejandro una carta de Parmenión, en la cual le aconsejaba desconfiar de Filipo, y le acusaba de estar secretamente entregada a Darío.|LETTER_00
799499208|mostrador|Alejandro, sin manifestar emoción alguna, apuró la copa de un trago, entregando simultáneamente a Filipo la carta acusatoria, cuya falsedad quedó inmediatamente demostrada.|LETTER_00
840330599|que_pasa|El ex 'carapintada' quien cumple condena perpetua por haber intentado un golpe contra el gobierno de Menem en 1990 basó sus afirmaciones en una carta de el mandatario argentino publicada por el diario inglés The Sun.|LETTER_00
840342182|que_pasa|El lunes 19, le envió una carta por fax en la que le pedía que ayudara a los senadores chilenos que iban rumbo a Londres a obtener una reunión con personeros de el más alto nivel en el gobierno de Blair.|LETTER_00
840342325|que_pasa|Una sorpresiva carta recibieron algunos colegios públicos y subvencionados de parte de el Ministerio de Educación, en la que se exigía la devolución de un texto de educación sexual que se les había hecho llegar para sus bibliotecas.|LETTER_00
840342981|que_pasa|Las declaraciones de Délano indignaron a Gómez, quien, después de defender a Frei en Valdivia, le envió una carta de rechazo y luego lo invitó a almorzar a un restaurante en Santiago, donde cordialmente se dejaron en claro que ya no eran amigos.|LETTER_00
840346314|que_pasa|Aunque el día anterior, la ex Premier de Gran Bretaña, Margaret Thatcher, había publicado una carta en el diario inglés The Times, en la cual pedía la liberación de Pinochet y reconocía el apoyo chileno en la guerra de las Malvinas, analistas bonaerenses creen que Menem no hizo referencias a el tema para no perturbar las relaciones bilaterales con Chile. 'Se trata de gestos habituales en Menem y que buscan crear un ambiente favorable para solucionar el conflicto pendiente de Campos de Hielo'.|LETTER_00
840354158|que_pasa|José Miguel Insulza v.s Gladys Marín El altercado se produjo el 14 de abril de 1998, cuando el ministro salía por la puerta principal de la Cancillería, en el preciso momento en que llegaba la dirigente comunista a entregar una carta con sus interrogantes sobre la II Cumbre de las Américas y la no invitación de Cuba a la reunión.|LETTER_00
840355264|que_pasa|Algunos informes periodísticos, basados en los archivos descubiertos en Paraguay, hablan de una carta escrita por el general (r) Manuel Contreras a el general paraguayo Guanes Serrano, en la que ofrece el cuartel de la DINA como cuartel general de la operación, para así 'centralizar la información sobre los antecedentes de personas, organizaciones y otras actividades conectadas directa o indirectamente con la subversión'.|LETTER_00
840357771|que_pasa|Por eso, esta semana, varios senadores enviaron cartas a sus colegas de derecha para notificar les que si no concurrían a votar proyectos de ley, ellos no se abstendrían de hacerlo, porque el acuerdo no rige para ausencias por protestas políticas.|LETTER_00
840359983|que_pasa|La fórmula de los propietarios (Rebeca Iver y su hijo, Felipe Castillo, que además maneja el excelente restaurante 'Doña Paula', en el interior de Viña Santa Rita), y que en general comparte la gente que sabe comer, es cambiar de platos muy de tarde en tarde y mantener una calidad constante, una carta estable, una cocina que no va a deparar sorpresas ni sobresaltos a el público exigente y, en general, conservador que allí concurre.|MENU_00
840360321|que_pasa|La repostería de 'Carrousel' ostenta ya una impronta netamente francesa en la muy buena calidad de sus Crepes Suzette o la Mousse Orleans de castaña, en una reiteración de carta precisa, pero impecablemente realizada.|MENU_00
840362808|que_pasa|Mama Fresia había puesto en la maleta sus botas más firmes, así como sus cuadernos y el atado de cartas de amor de Joaquín Andieta.|LETTER_00
840385975|que_pasa|Horas antes de que las 3.500 páginas fueran entregadas a la luz pública, se jugaron su última carta.|CARD_00
840393292|que_pasa|Los síntomas de el cambio ya son observables en Estados Unidos, donde el tráfico de datos a través de Internet sobrepasa las llamadas telefónicas y el correo electrónico ya supera a el envío de cartas tradicionales.|LETTER_00

835717997|que_pasa|Desde su inauguración en 1914, la cadena de restaurantes incluye en su carta el té en hojas con toda la implementación necesaria.|MENU_00
840518035|que_pasa|Desde la inauguración de la carta de vinos y de su servicio por copas, ofrece también de lunes a sábado, en este orden, una especialidad por día como plato fuerte (5.500), ya sea Pollo de campo, Cochinillo, Mariscos, Cordero patagónico, Salmón y Carne de caza.|MENU_00
810316013|primera_linea|A el momento de barajar el naipe, las cartas posibles para marcar el nuevo juego son una revitalización de la agenda valórica progresista, reimpulsar la reforma a la salud -que en el ámbito técnico ha desentrabado los escollos para una autoridad sanitaria- o revivir las enmiendas pendientes a el financiamiento de la educación superior.|CARD_00
810350928|primera_linea|Jesse Jackson recibe carta de el Talibán pero declina mediar.|LETTER_00
810350995|primera_linea|La carta 'no aumenta mi inclinación a ir', dijo, después de informar que reenvió la misiva a el secretario de Estado Colin Powell.|LETTER_00
810351026|primera_linea|Firmada por el mulá Abdul Salam Zaeef, la carta reza 'En la actual situación crítica, se necesitan más prudencia, sagacidad y paciencia para resolver las cuestiones entre Afganistán y Estados Unidos a través de medios pacíficos, y apreciamos su mediación con altos funcionarios gubernamentales de Afganistán'.|LETTER_00

Anexo 6: Matrices de confusión para cada *dataset* del sentido +HEAD_00 de «cabeza»

DATASET 01 +HEAD_00
True Positives: 14
True Negatives: 11
False Positives: 11
False Negatives: 4

True Positive Rate (a.k.a. Recall or Sensitivity): 0.777777777777778
True Negative Rate (a.k.a. Specificity): 0.5
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.56
Negative Predictive Value (a.k.a. Negative Precision): 0.733333333333333
False Positive Rate (a.k.a. Fall-out): 0.5
False Discovery Rate: 0.44

Accuracy: 0.625
Efficiency: 0.638888888888889
Error Rate: 0.375
Euclidean Distance: 0.54715876676645
F-Score: 0.651162790697674
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.285449612859225
Prevalence: 0.45
Standard Error: 0.0765465544619743

DATASET 02 +HEAD_00
True Positives: 15
True Negatives: 11
False Positives: 8
False Negatives: 6

True Positive Rate (a.k.a. Recall or Sensitivity): 0.714285714285714
True Negative Rate (a.k.a. Specificity): 0.578947368421053
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.652173913043478
Negative Predictive Value (a.k.a. Negative Precision): 0.647058823529412
False Positive Rate (a.k.a. Fall-out): 0.421052631578947
False Discovery Rate: 0.347826086956522

Accuracy: 0.65
Efficiency: 0.646616541353384
Error Rate: 0.35
Euclidean Distance: 0.508839829043267
F-Score: 0.681818181818182
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.29621772025328
Prevalence: 0.525

Standard Error: 0.0754155156449918

DATASET 03 +HEAD_00
True Positives: 11
True Negatives: 13
False Positives: 9
False Negatives: 7

True Positive Rate (a.k.a. Recall or Sensitivity): 0.611111111111111
True Negative Rate (a.k.a. Specificity): 0.590909090909091
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.55
Negative Predictive Value (a.k.a. Negative Precision): 0.65
False Positive Rate (a.k.a. Fall-out): 0.409090909090909
False Discovery Rate: 0.45

Accuracy: 0.6
Efficiency: 0.601010101010101
Error Rate: 0.4
Euclidean Distance: 0.564437720038324
F-Score: 0.578947368421053
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.201007563051842
Prevalence: 0.45
Standard Error: 0.0774596669241483

Anexo 7: Matrices de confusión para cada *dataset* del sentido +CHIEF_00 de «cabeza»

DATASET 01 +CHIEF_00
True Positives: 3
True Negatives: 29
False Positives: 1
False Negatives: 7

True Positive Rate (a.k.a. Recall or Sensitivity): 0.3
True Negative Rate (a.k.a. Specificity): 0.966666666666667
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.75
Negative Predictive Value (a.k.a. Negative Precision): 0.805555555555556
False Positive Rate (a.k.a. Fall-out): 0.0333333333333333
False Discovery Rate: 0.25

Accuracy: 0.8
Efficiency: 0.633333333333333
Error Rate: 0.2
Euclidean Distance: 0.700793201387621
F-Score: 0.428571428571428
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.384900179459751
Prevalence: 0.25
Standard Error: 0.0632455532033676

DATASET 02 +CHIEF_00
True Positives: 1
True Negatives: 30
False Positives: 4
False Negatives: 5

True Positive Rate (a.k.a. Recall or Sensitivity): 0.166666666666667
True Negative Rate (a.k.a. Specificity): 0.882352941176471
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.2
Negative Predictive Value (a.k.a. Negative Precision): 0.857142857142857

False Positive Rate (a.k.a. Fall-out): 0.117647058823529
False Discovery Rate: 0.8

Accuracy: 0.775
Efficiency: 0.524509803921569
Error Rate: 0.225
Euclidean Distance: 0.841596860078667
F-Score: 0.181818181818182
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.0529256124024963
Prevalence: 0.15
Standard Error: 0.0660255632312213

DATASET 03 +CHIEF_00
True Positives: 3
True Negatives: 26
False Positives: 3
False Negatives: 8

True Positive Rate (a.k.a. Recall or Sensitivity): 0.272727272727273
True Negative Rate (a.k.a. Specificity): 0.896551724137931
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.5
Negative Predictive Value (a.k.a. Negative Precision): 0.764705882352941
False Positive Rate (a.k.a. Fall-out): 0.103448275862069
False Discovery Rate: 0.5

Accuracy: 0.725
Efficiency: 0.584639498432602
Error Rate: 0.275
Euclidean Distance: 0.734593197364055
F-Score: 0.352941176470588
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.211681709717738
Prevalence: 0.275
Standard Error: 0.0706001062322147

Anexo 8: Matrices de confusión para cada *dataset* del sentido +LEADER_00 de «cabeza»

DATASET 01 +LEADER_00
True Positives: 4
True Negatives: 26
False Positives: 6
False Negatives: 4

True Positive Rate (a.k.a. Recall or Sensitivity): 0.5
True Negative Rate (a.k.a. Specificity): 0.8125
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.4
Negative Predictive Value (a.k.a. Negative Precision): 0.866666666666667
False Positive Rate (a.k.a. Fall-out): 0.1875
False Discovery Rate: 0.6

Accuracy: 0.75
Efficiency: 0.65625
Error Rate: 0.25
Euclidean Distance: 0.534000234082346
F-Score: 0.444444444444444
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.288675134594813
Prevalence: 0.2
Standard Error: 0.0684653196881458

DATASET 02 +LEADER_00
True Positives: 1
True Negatives: 28
False Positives: 4
False Negatives: 7

True Positive Rate (a.k.a. Recall or Sensitivity): 0.125
True Negative Rate (a.k.a. Specificity): 0.875
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.2
Negative Predictive Value (a.k.a. Negative Precision): 0.8
False Positive Rate (a.k.a. Fall-out): 0.125
False Discovery Rate: 0.8

Accuracy: 0.725
Efficiency: 0.5
Error Rate: 0.275
Euclidean Distance: 0.883883476483184
F-Score: 0.153846153846154
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0
Prevalence: 0.2
Standard Error: 0.0706001062322147

DATASET 03 +LEADER_00
True Positives: 4
True Negatives: 27
False Positives: 7
False Negatives: 2

True Positive Rate (a.k.a. Recall or Sensitivity): 0.666666666666667
True Negative Rate (a.k.a. Specificity): 0.794117647058823
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.363636363636364
Negative Predictive Value (a.k.a. Negative Precision): 0.931034482758621
False Positive Rate (a.k.a. Fall-out): 0.205882352941176
False Discovery Rate: 0.636363636363636

Accuracy: 0.775
Efficiency: 0.730392156862745
Error Rate: 0.225
Euclidean Distance: 0.391789043189962
F-Score: 0.470588235294118
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.368482976175321
Prevalence: 0.15
Standard Error: 0.0660255632312213

Anexo 9: Matrices de confusión para cada *dataset* del sentido +INTELLIGENCE_00 de «cabeza»

DATASET 01 +INTELLIGENCE_00
True Positives: 0
True Negatives: 35
False Positives: 1
False Negatives: 4

True Positive Rate (a.k.a. Recall or Sensitivity): 0
True Negative Rate (a.k.a. Specificity): 0.972222222222222
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0
Negative Predictive Value (a.k.a. Negative Precision): 0.897435897435897
False Positive Rate (a.k.a. Fall-out): 0.0277777777777778
False Discovery Rate: 1

Accuracy: 0.875
Efficiency: 0.486111111111111
Error Rate: 0.125
Euclidean Distance: 1.00038572807606
F-Score: NaN
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): -0.0533760512683624
Prevalence: 0.1
Standard Error: 0.0522912516583797

DATASET 02 +INTELLIGENCE_00
True Positives: 2
True Negatives: 30
False Positives: 5
False Negatives: 3

True Positive Rate (a.k.a. Recall or Sensitivity): 0.4
True Negative Rate (a.k.a. Specificity): 0.857142857142857
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.285714285714286
Negative Predictive Value (a.k.a. Negative Precision): 0.909090909090909
False Positive Rate (a.k.a. Fall-out): 0.142857142857143
False Discovery Rate: 0.714285714285714

Accuracy: 0.8
Efficiency: 0.628571428571429
Error Rate: 0.2
Euclidean Distance: 0.616772375569226
F-Score: 0.333333333333333
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.223814129085897
Prevalence: 0.125
Standard Error: 0.0632455532033676

DATASET 03 +INTELLIGENCE_00
True Positives: 0
True Negatives: 32
False Positives: 3
False Negatives: 5

True Positive Rate (a.k.a. Recall or Sensitivity): 0
True Negative Rate (a.k.a. Specificity): 0.914285714285714
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0
Negative Predictive Value (a.k.a. Negative Precision): 0.864864864864865
False Positive Rate (a.k.a. Fall-out): 0.0857142857142857
False Discovery Rate: 1

Accuracy: 0.8
Efficiency: 0.457142857142857
Error Rate: 0.2
Euclidean Distance: 1.00366674687145
F-Score: NaN
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): -0.107624400500126
Prevalence: 0.125
Standard Error: 0.0632455532033676

Anexo 10: Matrices de confusión para cada *dataset* del sentido +FACE_00 de «cara»

DATASET 01 +FACE_00
True Positives: 16
True Negatives: 3
False Positives: 6
False Negatives: 15

True Positive Rate (a.k.a. Recall or Sensitivity): 0.516129032258065
True Negative Rate (a.k.a. Specificity): 0.333333333333333
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.727272727272727
Negative Predictive Value (a.k.a. Negative Precision): 0.166666666666667
False Positive Rate (a.k.a. Fall-out): 0.666666666666667
False Discovery Rate: 0.272727272727273

Accuracy: 0.475
Efficiency: 0.424731182795699
Error Rate: 0.525
Euclidean Distance: 0.823756977432035
F-Score: 0.60377358490566
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): -0.126357084250573
Prevalence: 0.775
Standard Error: 0.0789580584867688

DATASET 02 +FACE_00
True Positives: 23
True Negatives: 2
False Positives: 5
False Negatives: 10

True Positive Rate (a.k.a. Recall or Sensitivity): 0.696969696969697
True Negative Rate (a.k.a. Specificity): 0.285714285714286
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.821428571428571
Negative Predictive Value (a.k.a. Negative Precision): 0.166666666666667
False Positive Rate (a.k.a. Fall-out): 0.714285714285714
False Discovery Rate: 0.178571428571429

Accuracy: 0.625
Efficiency: 0.491341991341991
Error Rate: 0.375
Euclidean Distance: 0.775906854066447
F-Score: 0.754098360655738
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): -0.0143576830751316
Prevalence: 0.825
Standard Error: 0.0765465544619743

DATASET 03 +FACE_00
True Positives: 24
True Negatives: 3
False Positives: 9
False Negatives: 4

True Positive Rate (a.k.a. Recall or Sensitivity): 0.857142857142857
True Negative Rate (a.k.a. Specificity): 0.25
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.727272727272727
Negative Predictive Value (a.k.a. Negative Precision): 0.428571428571429
False Positive Rate (a.k.a. Fall-out): 0.75
False Discovery Rate: 0.272727272727273

Accuracy: 0.675
Efficiency: 0.553571428571429
Error Rate: 0.325
Euclidean Distance: 0.763484225943998
F-Score: 0.786885245901639
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.129219147676184
Prevalence: 0.7
Standard Error: 0.0740565662179931

Anexo 11: Matrices de confusión para cada *dataset* del sentido +SIDE_00 de «cara»

DATASET 01 +SIDE_00
True Positives: 3
True Negatives: 16
False Positives: 15
False Negatives: 6

True Positive Rate (a.k.a. Recall or Sensitivity): 0.333333333333333
True Negative Rate (a.k.a. Specificity): 0.516129032258065
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.166666666666667
Negative Predictive Value (a.k.a. Negative Precision): 0.727272727272727
False Positive Rate (a.k.a. Fall-out): 0.483870967741935
False Discovery Rate: 0.833333333333333

Accuracy: 0.475
Efficiency: 0.424731182795699
Error Rate: 0.525
Euclidean Distance: 0.823756977432035
F-Score: 0.222222222222222
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): -0.126357084250573
Prevalence: 0.225
Standard Error: 0.0789580584867688

DATASET 02 +SIDE_00
True Positives: 2
True Negatives: 23
False Positives: 10
False Negatives: 5

True Positive Rate (a.k.a. Recall or Sensitivity): 0.285714285714286
True Negative Rate (a.k.a. Specificity): 0.696969696969697
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.166666666666667
Negative Predictive Value (a.k.a. Negative Precision): 0.821428571428571
False Positive Rate (a.k.a. Fall-out): 0.303030303030303
False Discovery Rate: 0.833333333333333

Accuracy: 0.625
Efficiency: 0.491341991341991
Error Rate: 0.375
Euclidean Distance: 0.775906854066447
F-Score: 0.210526315789474
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): -0.0143576830751316
Prevalence: 0.175
Standard Error: 0.0765465544619743

DATASET 03 +SIDE_00
True Positives: 3
True Negatives: 24

False Positives: 4
False Negatives: 9

True Positive Rate (a.k.a. Recall or Sensitivity): 0.25
True Negative Rate (a.k.a. Specificity): 0.857142857142857
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.428571428571429
Negative Predictive Value (a.k.a. Negative Precision): 0.727272727272727
False Positive Rate (a.k.a. Fall-out): 0.142857142857143
False Discovery Rate: 0.571428571428571

Accuracy: 0.675
Efficiency: 0.553571428571429
Error Rate: 0.325
Euclidean Distance: 0.763484225943998
F-Score: 0.315789473684211
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.129219147676184
Prevalence: 0.3
Standard Error: 0.0740565662179931

Anexo 12: Matrices de confusión para cada *dataset* del sentido +LETTER_00 de «carta»

DATASET 01 +LETTER_00
True Positives: 14
True Negatives: 11
False Positives: 6
False Negatives: 9

True Positive Rate (a.k.a. Recall or Sensitivity): 0.608695652173913
True Negative Rate (a.k.a. Specificity): 0.647058823529412
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.7
Negative Predictive Value (a.k.a. Negative Precision): 0.55
False Positive Rate (a.k.a. Fall-out): 0.352941176470588
False Discovery Rate: 0.3

Accuracy: 0.625
Efficiency: 0.627877237851662
Error Rate: 0.375
Euclidean Distance: 0.526959739141466
F-Score: 0.651162790697674
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.252860868712087
Prevalence: 0.575
Standard Error: 0.0765465544619743

DATASET 02 +LETTER_00
True Positives: 17
True Negatives: 11
False Positives: 7
False Negatives: 5

True Positive Rate (a.k.a. Recall or Sensitivity): 0.772727272727273
True Negative Rate (a.k.a. Specificity): 0.611111111111111
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.708333333333333
Negative Predictive Value (a.k.a. Negative Precision): 0.6875
False Positive Rate (a.k.a. Fall-out): 0.388888888888889
False Discovery Rate: 0.291666666666667

Accuracy: 0.7
Efficiency: 0.691919191919192
Error Rate: 0.3

Euclidean Distance: 0.450430305888955
F-Score: 0.739130434782609
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.389789721434554
Prevalence: 0.55
Standard Error: 0.0724568837309472

DATASET 03 +LETTER_00
True Positives: 20
True Negatives: 10
False Positives: 4
False Negatives: 6

True Positive Rate (a.k.a. Recall or Sensitivity): 0.769230769230769
True Negative Rate (a.k.a. Specificity): 0.714285714285714
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.833333333333333
Negative Predictive Value (a.k.a. Negative Precision): 0.625
False Positive Rate (a.k.a. Fall-out): 0.285714285714286
False Discovery Rate: 0.166666666666667

Accuracy: 0.75
Efficiency: 0.741758241758242
Error Rate: 0.25
Euclidean Distance: 0.367269779496009
F-Score: 0.8
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.470756541762004
Prevalence: 0.65
Standard Error: 0.0684653196881458

Anexo 13: Matrices de confusión para cada *dataset* del sentido +CARD_00 de «carta»

DATASET 01 +CARD_00
True Positives: 5
True Negatives: 23
False Positives: 9
False Negatives: 3

True Positive Rate (a.k.a. Recall or Sensitivity): 0.625
True Negative Rate (a.k.a. Specificity): 0.71875
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.357142857142857
Negative Predictive Value (a.k.a. Negative Precision): 0.884615384615385
False Positive Rate (a.k.a. Fall-out): 0.28125
False Discovery Rate: 0.642857142857143

Accuracy: 0.7
Efficiency: 0.671875
Error Rate: 0.3
Euclidean Distance: 0.46875
F-Score: 0.454545454545455
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.288278330098528
Prevalence: 0.2
Standard Error: 0.0724568837309472

DATASET 02 +CARD_00
True Positives: 4
True Negatives: 27
False Positives: 6
False Negatives: 3

True Positive Rate (a.k.a. Recall or Sensitivity): 0.571428571428571
True Negative Rate (a.k.a. Specificity): 0.818181818181818
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.4
Negative Predictive Value (a.k.a. Negative Precision): 0.9
False Positive Rate (a.k.a. Fall-out): 0.181818181818182
False Discovery Rate: 0.6

Accuracy: 0.775
Efficiency: 0.694805194805195
Error Rate: 0.225
Euclidean Distance: 0.465544112439868
F-Score: 0.470588235294118
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.341881729378914
Prevalence: 0.175
Standard Error: 0.0660255632312213

DATASET 03 +CARD_00
True Positives: 2
True Negatives: 29
False Positives: 4
False Negatives: 5

True Positive Rate (a.k.a. Recall or Sensitivity): 0.285714285714286
True Negative Rate (a.k.a. Specificity): 0.878787878787879
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.333333333333333
Negative Predictive Value (a.k.a. Negative Precision): 0.852941176470588
False Positive Rate (a.k.a. Fall-out): 0.121212121212121
False Discovery Rate: 0.666666666666667

Accuracy: 0.775
Efficiency: 0.582251082251082
Error Rate: 0.225
Euclidean Distance: 0.724497384371673
F-Score: 0.307692307692308
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.175050164393881
Prevalence: 0.175
Standard Error: 0.0660255632312213

Anexo 14: Matrices de confusión para cada *dataset* del sentido $MENU _00 de «carta»

DATASET 01 $MENU_00
True Positives: 4
True Negatives: 29
False Positives: 2
False Negatives: 5

True Positive Rate (a.k.a. Recall or Sensitivity): 0.444444444444444
True Negative Rate (a.k.a. Specificity): 0.935483870967742
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.666666666666667
Negative Predictive Value (a.k.a. Negative Precision): 0.852941176470588
False Positive Rate (a.k.a. Fall-out): 0.0645161290322581
False Discovery Rate: 0.333333333333333

Accuracy: 0.825
Efficiency: 0.689964157706093
Error Rate: 0.175
Euclidean Distance: 0.559289107898544
F-Score: 0.533333333333333

Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.444312651764607
Prevalence: 0.225
Standard Error: 0.0600780742034896

DATASET 02 $MENU_00
True Positives: 6
True Negatives: 29
False Positives: 0
False Negatives: 5

True Positive Rate (a.k.a. Recall or Sensitivity): 0.545454545454545
True Negative Rate (a.k.a. Specificity): 1
Positive Predictive Value (a.k.a. Precision or Positive Precision): 1
Negative Predictive Value (a.k.a. Negative Precision): 0.852941176470588
False Positive Rate (a.k.a. Fall-out): 0
False Discovery Rate: 0

Accuracy: 0.875
Efficiency: 0.772727272727273
Error Rate: 0.125
Euclidean Distance: 0.454545454545455
F-Score: 0.705882352941176
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.682085509090488
Prevalence: 0.275
Standard Error: 0.0522912516583797

DATASET 03 $MENU_00
True Positives: 5
True Negatives: 28
False Positives: 5
False Negatives: 2

True Positive Rate (a.k.a. Recall or Sensitivity): 0.714285714285714
True Negative Rate (a.k.a. Specificity): 0.848484848484849
Positive Predictive Value (a.k.a. Precision or Positive Precision): 0.5
Negative Predictive Value (a.k.a. Negative Precision): 0.933333333333333
False Positive Rate (a.k.a. Fall-out): 0.151515151515152
False Discovery Rate: 0.5

Accuracy: 0.825
Efficiency: 0.781385281385281
Error Rate: 0.175
Euclidean Distance: 0.323402990400342
F-Score: 0.588235294117647
Matthews Correlation Coefficient (a.k.a. Phi Coefficient): 0.493829164658431
Prevalence: 0.175
Standard Error: 0.0600780742034896