



Lingüística



FACULTAD DE
FILOSOFÍA Y LETRAS

LINGÜÍSTICA COMPUTACIONAL EN MÉXICO

INVESTIGACIÓN Y DESARROLLO

EDICIÓN DE
GERARDO SIERRA MARTÍNEZ Y JAVIER CUÉTARA PRIEDE

PRÓLOGO DE
SERGIO M. ALCOCER MARTÍNEZ CASTRO

LINGÜÍSTICA COMPUTACIONAL EN MÉXICO

INVESTIGACIÓN Y DESARROLLO

Primera edición: agosto de 2015
Edición de Gerardo Sierra y Javier Cuétara
Diseño: Ruth Eunice Pérez Pérez

2015. Instituto de Ingeniería, UNAM.
Facultad de Filosofía y Letras.

ISBN: 978-607-02-6365-0

Reservados todos los derechos. Queda rigurosamente prohibida, sin la autorización previa y por escrito de los titulares del copyright, bajo las sanciones establecidas en las leyes, la reproducción parcial o total de esta obra por cualquier medio o procedimiento.

Versión digital

ÍNDICE

Agradecimientos.....	9
Prefacio.....	11
Prólogo.....	13
Introducción.....	15
Lingüística de corpus: ¿una materia de lingüística o de ingeniería?.....	19
La red electrónica mundial como herramienta de enseñanza-aprendizaje: el diplomado para formación de profesores de español como lengua extranjera.....	26
Escenarios virtuales para la enseñanza de la pronunciación de lengua extranjera.....	32
Una página electrónica para la enseñanza de fonética y fonología en la FFyL de la UNAM.....	35
Comparación diacrónica de perfiles morfológicos: distancias entre los documentos del CHEM y los del CEMC.....	41
Desarrollo de un transcriptor fonético-fonológico para corpus textuales diacrónicos.....	46
Propuesta para un transcriptor automático del español del siglo XVI para el Corpus Histórico del Español de México.....	49
El desarrollo tecnológico y su impacto en los procesos de traducción.....	55
Uso de técnicas y recursos de la lingüística computacional en sistemas de extracción de información.....	59
Métodos para la obtención automática de términos en un área de especialidad.....	68
La estructura de las obligaciones y el “common ground” en diálogos prácticos.....	72
La silabificación en el <i>Corpus DIME</i> como fenómeno fonético de vital importancia para las tecnologías del habla.....	84
Fenómenos de síncopa en el <i>Corpus DIME</i> para su inclusión en el reconocedor de habla DIMEX.....	88
Detección y corrección de asociación sintáctico-semántica basada en la web.....	93
Criterios sintácticos aplicados a un programa de generación de pares semánticos.....	102
Etiquetado de contextos definitorios.....	111

Anáforas y otras relaciones de correferencia en la expansión de contextos definitorios.....	118
Hacia una tipología de definiciones basada en estructuras predicativas para su extracción automática.....	122
Desarrollo de un sintetizador del habla para el Corpus Histórico del Español de México.....	131
Identificación automática del lenguaje hablado sin información fonotáctica.....	137
Corpus paralelo alineado español-inglés de textos literarios.....	144
Control de la estabilidad en el agrupamiento de textos.....	150
Alineamiento de corpus paralelos náhuatl-español con enfoque de extracción léxica.....	156
Índice de abreviaturas de instituciones.....	167

AGRADECIMIENTOS

La edición de estas actas fue posible gracias al esfuerzo de varias personas que, en conjunto, contribuyeron con diversas actividades tales como la grabación, la transcripción, la recolección de diapositivas, la revisión de estilo y el diseño final. A todos ellos, un agradecimiento muy especial.

Rodrigo Alarcón Martínez
César Antonio Aguilar
Georgina Barraza Carbajal
Reyna Cristal Díaz Salgado
Ariadna Carolina Hernández Angulo
Jorge Adrian Lázaro Hernández
Esperanza Montserrat Martínez Herrera
Víctor Germán Mijangos Cruz
Margarita Palacios Sierra
Ruth Eunice Pérez Pérez
José Manuel Posada de la Concha
Jesús Jerónimo Ramírez Galicia
Teresita Adriana Reyes Careaga
Alejandro Rosas González
Octavio Augusto Sánchez Velázquez
Miriam Yuridia Sevilla Román

Asimismo, a las siguientes instituciones que apoyaron en la realización del Coloquio y en la edición de este libro.

Consejo Nacional de Ciencia y Tecnología
Dirección General de Asuntos para el Personal Académico, UNAM
Facultad de Filosofía y Letras, UNAM
Instituto de Ingeniería, UNAM

PREFACIO

En el año 2003, los profesores Gerardo Sierra Martínez y Javier Cuétara Priede se unieron para crear el Coloquio de Lingüística Computacional (CoLiCo) realizado y organizado en la Facultad de Filosofía y Letras (FFL), con la colaboración del Grupo de Ingeniería Lingüística del Instituto de Ingeniería (IINGEN), ambos de la Universidad Nacional Autónoma de México. El CoLiCo es una actividad académica multidisciplinaria en la que confluyen lingüistas, ingenieros, matemáticos y muchos académicos más. El CoLiCo, asimismo, se ofrece como un espacio abierto a toda la comunidad académica: recibe tanto a investigadores y profesores de renombre como a estudiantes en formación.

Otra virtud del Coloquio de Lingüística Computacional es que ha logrado unir a instituciones muy variadas: en él confluyen diversas instancias de la UNAM (además de la FFL y del IINGEN, el Centro de Enseñanza de Lenguas Extranjeras, el Centro de Enseñanza para Extranjeros, la Facultad de Contaduría y Administración, la Facultad de Ingeniería, el Instituto de Investigaciones Filológicas, el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, y muchos más), así como otras instituciones nacionales (Benemérita Universidad Autónoma de Puebla, Instituto Politécnico Nacional, Universidad Autónoma Metropolitana, entre otras) e, incluso, de otros países, como la Universidad Autónoma de Barcelona, La Universidad de Aviñón y la Universidad Carolina de Praga.

El texto que hoy sale a la luz representa las versiones estenográficas de los trabajos presentados en la tercera edición del CoLiCo, llevada a cabo en febrero del 2007. En ese momento, y debido al reconocimiento que el CoLiCo iba cobrando, se mostró conveniente realizar estas memorias. Esto implicó una intensa y larga labor: conseguir las grabaciones, transcribirlas, retocar los textos para darles coherencia gramatical a los discursos orales, etc. Posteriormente —y por razones de muy diversa índole— el texto quedó archivado.

El día de hoy es apropiado retomar este texto por los valores que brinda. Por un lado, aporta un valor histórico, pues da cuenta del estado de la investigación académica en lingüística computacional en México en la primera década del siglo XXI. Por otra parte, el carácter oral de los trabajos presentados muestra la atmósfera y el ambiente en los que se desarrolla el Coloquio; de alguna manera servirá para atraer a más participantes y público a los próximos Coloquios en Lingüística Computacional, al notar los lectores que las presentaciones se exponen de una manera clara y accesible.

En suma, el trabajo que aquí se presenta sale a la luz en el año 2015 porque, cuando se pensaría que los contenidos están superados y que la tecnología ha adelantado casi dos lustros desde que se realizaron las investigaciones, este libro presenta un estado de la cuestión de la lingüística computacional en el mundo académico de nuestro país en los primeros años de este milenio, así como otros valores agregados que no podemos desaprovechar.

PRÓLOGO

He recibido con beneplácito la invitación de los organizadores para escribir el prólogo a las memorias del Tercer Coloquio de Lingüística Computacional (COLICO), celebrado en la Facultad de Filosofía y Letras de la Universidad Nacional Autónoma de México del 26 al 28 de febrero de 2007. Este acontecimiento reunió tanto a especialistas como a estudiantes de diversas instituciones educativas de nuestro país.

La lingüística computacional es un campo muy complejo y vasto, en el que intervienen disciplinas tan diversas como la lingüística, la ingeniería, con raíces de sus disciplinas, como el cómputo, las telecomunicaciones y la electrónica, entre otras. También se podría decir que es un área de conocimiento relativamente nueva y poco explorada en México. Mientras que en los Estados Unidos de América surge en los años cincuenta como un esfuerzo para obtener computadoras capaces de traducir textos automáticamente de lenguas extranjeras, en particular del ruso al inglés, en nuestro país, es hasta los ochenta cuando la lingüística computacional comienza a ser considerada en algunos estudios lingüísticos o de cómputo.

El mayor desarrollo de la lingüística computacional lo podemos ubicar como consecuencia del esfuerzo de distintos equipos de científicos y técnicos de Estados Unidos y de la Unión Soviética, quienes trabajaban en proyectos para los servicios de inteligencia y de las fuerzas armadas. Así, se produjeron importantes avances en áreas que resultaron claves para las tecnologías de procesamiento de lenguaje natural: la teoría de los autómatas, que se originó en los trabajos de Alan Turing y los modelos de teoría de la información, que surgieron del quehacer de Claude Shannon. Más tarde, las investigaciones fueron concentrándose en dos campos: el simbólico y el estocástico.

Dentro de la primera clasificación se encuentra el trabajo desarrollado por Noam Chomsky, lingüista, filósofo, escritor y analista político, quien cambió por completo la perspectiva, los programas y métodos de investigación en el estudio del lenguaje, elevando dicha disciplina a la categoría de ciencia moderna. Para Chomsky, la lingüística es una teoría de la adquisición individual del lenguaje y una explicación de las estructuras y principios más profundos de éste. Postuló la teoría del innatismo a propósito de la adquisición del lenguaje y la autonomía de la gramática sobre otros sistemas cognitivos, así como la existencia de un “órgano del lenguaje” y de una gramática universal. También llevó a cabo la clasificación de lenguajes formales.

La teoría del innatismo sostiene que el humano cuenta con un dispositivo cerebral innato que le permite aprender y utilizar el lenguaje de forma casi instintiva, lo que, aunado a la universalidad de los principios generales abstractos de la gramática y de sus reglas, y de un conjunto finito de términos, crea la posibilidad de que el humano pueda producir un número infinito de frases. Estos descubrimientos aportaron elementos fundamentales para la creación de lenguajes nuevos como puede ser el computacional.

La segunda clasificación, la del campo estocástico, ha sido desarrollada fundamentalmente por los ingenieros electrónicos cuyo trabajo se elabora mediante estadísticas y probabilidades, donde surgió el método de Bayes para el reconocimiento óptico de caracteres. En las décadas subsiguientes el interés se centró en los corpus textuales, especialmente en inglés, al desarrollo de distintos lenguajes de programación con insumos de la lingüística teórica y de distintos programas para el análisis morfológico y sintáctico.

Como es de todos conocido, en la década de los noventa, el Internet revolucionó los sistemas de información de todo el mundo, impactando casi todas las actividades económicas y sociales, especialmente el área

del conocimiento y la educación. Como consecuencia se dio la necesidad de perfeccionar las tecnologías para el procesamiento automático del lenguaje. Los lingüistas computacionales se enfocaron a desarrollar productos informáticos para el análisis automático de la fonética, la fonología, la morfología, la sintaxis y la semántica. Así también adquirió mayor relevancia la generación de lenguaje natural o textos a través de modelar en representaciones semánticas, que a su vez, son procesadas y transformadas en textos en una lengua dada. También trabajan en elaborar sistemas que hacen posible el diálogo entre personas que hablan lenguas diferentes o entre humanos y máquinas. Para ello se requiere del etiquetamiento morfológico, el análisis sintáctico, procesos de interpretación semántica, traducción automática, técnicas de reconocimiento de voz o conversión de texto a voz, recuperación inteligente de información, etcétera.

Casi todas estas aplicaciones fueron abordadas en el Tercer Coloquio de Lingüística Computacional por los ponentes de la UNAM, del IPN, del COLMEX y de la UAM, a través de la conjunción de esfuerzos para tratar temas especializados y de gran actualidad.

La publicación presenta 23 temas dentro de la amplia gama de la lingüística aplicada y de la ingeniería lingüística, que abarcan, entre otros, lingüística de corpus; red electrónica mundial; escenarios virtuales para la enseñanza del español y de lenguas extranjeras; estudios diacrónicos y sincrónicos del español; transcritores fonéticos y automáticos del español; extracción de información relevante; diccionarios semasiológicos y onomasiológicos; etiquetado de contextos; desarrollo de un sintetizador del habla para el corpus histórico del español de México; identificación automática del lenguaje hablado (incluye español y lenguas indígenas).

Esta interdisciplina está presente en la vida diaria de gran parte de la población humana; favorece el trabajo académico, es decir la enseñanza y la investigación, en muy diversas ramas del conocimiento; en el área tecnológica son múltiples sus aportes, sólo por mencionar algunos: la telefonía celular, el diseño de software, el desarrollo de redes; asimismo es vital para empresas bancarias, financieras, aseguradoras y todas aquellas que manejan grandes bases de datos.

Actualmente la Universidad Nacional Autónoma de México desarrolla trabajos en este campo en las facultades de Ingeniería y de Filosofía y Letras a través de la impartición de materias tales como: Lingüística de Corpus, Tecnología del Habla, Procesamiento del Lenguaje Natural, Lingüística Computacional e Introducción a la Ingeniería Lingüística.

Asimismo hay que hacer notar que México cuenta con el Grupo de Ingeniería Lingüística (GIL) que tiene su sede en el Instituto de Ingeniería de la UNAM; el Grupo de Tratamiento del Lenguaje Natural, ubicado en el Instituto Nacional de Astrofísica, Óptica y Electrónica, en Puebla, y el grupo de estudiantes dedicados a esta área en el Departamento de Ciencias de la Computación en el Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas.

Actividades como el Tercer Coloquio de Lingüística Computacional promueven la interdisciplina, la divulgación de temas innovadores y creativos, lo que representa una gran satisfacción y motivo de orgullo para nuestra máxima casa de estudios, ya que colaboran de manera importante para avanzar en el conocimiento de la materia, conjuntando esfuerzos de los profesionales que se dedican a este tema, al tiempo que impulsan la investigación de una manera definitiva, facilitan el debate racional y fundamentado y colaboran con la formación de los profesionales que el país requiere; además de tener ante sí un amplio futuro, tanto en la investigación científica como en el desarrollo tecnológico. Igualmente promueven el acercamiento entre diversas instituciones educativas que comparten temas de interés.

Por todo ello quiero hacer patente mi reconocimiento al doctor Gerardo Sierra Martínez y al maestro Javier Cuétara Priede, quienes con gran compromiso promueven, coordinan y participan de manera activa y entusiasta en el desarrollo de esta área de conocimiento tan útil para el avance científico-tecnológico de nuestro país.

Sergio M. Alcocer Martínez de Castro
Universidad Nacional Autónoma de México
Secretario General

INTRODUCCIÓN

Hoy tengo la misión o el privilegio de introducirlos en el edificio que la ciencia del lenguaje ha empezado a construir desde hace apenas unos 60 años, por lo que deseo describirles sus grandes líneas de investigación, su estado actual, llevarlos a recorrer su pasado —apenas reciente—, o de pronosticarles su futuro...

Ferdinand de Saussure, Conferencia Inaugural de su Curso de Lingüística, en la Universidad de Ginebra

Una de las tareas más difíciles que se llevan a cabo en el mundo de las ciencias, y en particular dentro del margen histórico que nos ha tocado vivir en los últimos años, es otorgarle el atributo de “consolidada” a una nueva área de investigación. Ello equivale a decir que, independientemente de que se esté de acuerdo o no con sus postulados y sus resultados, dicha área cuenta ya con obras básicas, con una comunidad académica sólida, con una metodología específica y, sobre todo, con la posibilidad de crecer y evolucionar a partir de los aportes que vayan generando estudiantes que, con el paso del tiempo, se convertirán en los investigadores y profesores que con su labor sustenten dicha área.

Fiel reflejo de esta dificultad es precisamente lo que Saussure expuso en la inauguración de un seminario de lingüística —en 1891—, quizá el primer curso en su clase que se impartía en la Universidad de Ginebra. Saussure fue justo con sus palabras: con un pasado apenas reciente, y con una gran prospectiva por delante, la lingüística era dentro de las ciencias humanas de su época una de esas nuevas áreas emergentes, la cual tendría que esperar todavía varias décadas tras este discurso (aproximadamente seis), para lograr su punto de consolidación.

Vistas así las cosas, nos resulta ahora casi imposible pensar que la lingüística, en su momento, tuviera una génesis bastante complicada: el mismo Saussure en su exposición se preocupa por tratar de dejar en claro a su audiencia cuáles son los aportes que brinda el estudio del lenguaje a las demás ramas del conocimiento. Para él, la pregunta clave es la siguiente: *¿el lenguaje, o la lengua, puede ser el objeto de estudio —único y exclusivo— de una ciencia?* Así, Saussure se propone con su seminario —insistimos, por entero pionero en el ámbito universitario de su época— responder a esta pregunta.

El fruto de este esfuerzo por responder tal duda es su famoso *Cours de Linguistique Générale*, el cual pasó de ser un conjunto de simples notas de clase, a convertirse en una de las obras fundacionales de la lingüística. No es casual que haya ocurrido esto: con un pleno dominio del rigor y la labor metódica que caracterizan a un científico, Saussure va organizando de forma nítida los orígenes, los objetivos y las metas que debe cubrir la lingüística como una rama de estudio frente a sus alumnos. En todo caso, lo que llama la atención es justo ese origen humilde del *Cours...*, el haber sido engendrado a partir de las exposiciones y discusiones entre un docente brillante y sus alumnos, todos dueños de una mente abierta a los nuevos cambios que se daban en su época.

Si miramos hacia nuestro presente, y particularmente al espacio de nuestra Universidad, podría verse cierta similitud entre la experiencia vivida por Saussure al presentar a sus colegas ginebrinos su seminario pionero, y nuestra experiencia al mostrar ahora las exposiciones, los diálogos y las consideraciones generadas por este 3^{er} Coloquio de *Lingüística Computacional*, celebrado en nuestra Facultad de Filosofía y Letras los días 26, 27 y 28 de febrero de 2007.

La similitud que planteamos, hay que decirlo, no debe de entenderse como una copia exacta de dos momentos históricos (un *déjà vu* con tintes de cliché vanidoso), sino como una extensión de la ardua tarea que hemos descrito al inicio: ¿en qué momento y bajo qué circunstancia podemos dar a una nueva área científica un certificado de aceptación por parte de una comunidad académica? He aquí el *quid* que subyace entre el discurso inaugural de Saussure y la introducción que hacemos hoy de nuestro Coloquio.

En nuestro caso, somos conscientes de que nuestro evento no es fundacional, en el sentido en que este Coloquio no gesta una nueva área científica. Como se sabe, la lingüística computacional y la ingeniería lingüística son dos ramas gemelas de investigación y conocimiento surgidas hace unos 50 años, fruto de esfuerzos conjuntos entre lingüistas, computólogos, filósofos, psicólogos, matemáticos y otros. Todo este esfuerzo llevado en colaboración tiene un objetivo concreto: siguiendo a Alan Turing –uno de los grandes pensadores de nuestra época–, si una máquina fuese capaz de interactuar con un ser humano vía el lenguaje, ¿podremos decir que dicha máquina es inteligente, que el concretar dicha interacción refleja un comportamiento complejo, semejante al de cualquier persona normal?

Sin embargo, este objetivo se ha multiplicado en muchísimos senderos, los cuales abarcan muy distintos objetos de estudio, en proporción similar a los diferentes enfoques que asume cada investigador para abordarlos. Los textos que se reúnen en este libro son un claro ejemplo de tal diversidad: desde la formulación de modelos teóricos e hipótesis agudas –lo que las hace sumamente atractivas–, hasta la exposición de datos y soluciones generadas por métodos híbridos y herramientas computacionales aplicadas al análisis de algún fenómeno lingüístico. Así, hemos tratado de caminar un poco por todas estas sendas, atendiendo a las inquietudes que hemos tenido todos los participantes del Coloquio en torno a la lingüística computacional y la ingeniería lingüística. Asimismo, este intento por recorrer estos múltiples caminos es quizá también un reflejo de nuestra muy personal forma de llegar a este cruce entre computación y lingüística.

En el caso de Javier Cuétara, llegó por la varios senderos: tras la carrera de Lengua y Literaturas Hispánicas en la UNAM, su interés se focalizó en los estudios fonéticos y fonológicos del español, lo que fue perfilando su ingreso a la Maestría de Lingüística Hispánica de nuestra Universidad. Fruto de este periodo de estudios es la tesis titulada *Fonética de la Ciudad de México. Aportaciones desde las tecnologías del habla*. Este trabajo lingüístico, hay que señalarlo, es pionero en los estudios fonéticos aplicados al reconocimiento de habla en nuestro país, ya que plantea un alfabeto fonético computacional para el español de la Ciudad de México compuesto por un inventario de 37 alófonos, los cuales han sido utilizados para crear modelos acústicos capaces de mejorar el desempeño de un sistema de reconocimiento automático del habla. A partir de una labor de promoción y de la impartición de cursos de Lingüística computacional y de Tecnologías del habla, Javier se ha convertido, sin lugar a dudas, en uno de los principales promotores dentro de la Facultad de Filosofía y Letras del desarrollo de investigaciones lingüísticas con un perfil computacional.

Respecto al caso de Gerardo Sierra, su camino ha ido de la ingeniería hacia la lingüística. Sus intereses iniciales lo involucraron con la ingeniería de sistemas, enfocada en concreto hacia el área de desastres. Posteriormente, a partir de la necesidad de establecer una terminología precisa para una nueva área técnica, su trabajo se orientó hacia la lexicografía computacional, particularmente hacia la creación de un diccionario onomasiológico especializado de tipo electrónico, capaz de extraer términos y definiciones desde textos especializados. Este proceso de creación lo llevó a hacer una Maestría en Lingüística Hispánica en la UNAM, y a concretar un Doctorado en Lingüística Computacional en la University of Manchester Institute for Science and

Technology (UMIST). Tras este doctorado, regresó al Instituto de Ingeniería de la UNAM a desarrollar y fortalecer toda una línea de investigación pionera en el país, la ingeniería lingüística, cuyo fruto principal ha sido la conformación de un grupo de investigación, el Grupo de Ingeniería Lingüística (GIL), el cual se ha consolidado a lo largo de casi 10 años como un motor de difusión importante de esta área en nuestra Universidad.

Tras este breve relato de nuestras experiencias personales sobre nuestra aproximación al procesamiento de lenguaje natural, resulta claro que al final no parece tan clara –mucho menos tajante– la frontera que tendría que dividir a lingüistas de ingenieros dentro de esta área. Más allá de buscar esta barrera entre ambas partes, lo que uno puede notar es justo lo contrario: la fusión que por necesidad se ha dado entre ambas formaciones, en aras de encontrar soluciones a problemas tales como el desarrollo de un sistema de reconocimiento de habla (en el caso de Javier), o la creación de un diccionario electrónico especializado (en el caso de Gerardo). De este modo, parece que volvemos de nuevo a nuestra pregunta inicial: dados nuestros perfiles, junto con estas experiencias profesionales, ¿somos capaces de aceptar o no, dentro del ámbito de nuestra Facultad de Filosofía y Letras, la existencia de una nueva rama científica? Al igual que Saussure hace unos cien años, podemos preguntarnos: ¿nuestros colegas serán capaces de aceptar la invitación para conocer las ideas, los trabajos y los logros que se alcanzan dentro del procesamiento de lenguaje natural en español, generados justo por investigadores, profesores y estudiantes miembros de su misma comunidad?

He aquí el punto final de nuestra reflexión. Más allá de las comunes limitaciones que tiene que enfrentar cualquier empresa académica, incluidas la poca o casi nula apreciación que pueda mostrarse entre colegas (no tanto por vanas envidias, sino por un auténtico escepticismo intelectual), lo que deseamos con estas palabras, de la manera más humilde, es brindarle al lector un panorama –o incluso, un paisaje– de las propuestas y trabajos que hoy en día se gestan en nuestra Facultad sobre lingüística computacional e ingeniería lingüística. Lo que el lector podrá reconocer, además de hipótesis, métodos y resultados, es un constante y significativo interés por parte de nuestros expositores y su público por tratar estos temas. Sobra decir que este interés no es algo casual, sino que más bien va acorde con un nuevo signo de los tiempos, con un cambio de paradigma en la lingüística contemporánea. Nuestra Facultad, y en general toda nuestra Universidad, no es ajena a este cambio, y los organizadores de este Coloquio, Javier Cuétara y Gerardo Sierra, creemos que este texto es una prueba fehaciente de este cambio.

Así, amparándonos nuevamente en Saussure, los invitamos a hacer un recorrido por una nueva y fascinante área de investigación y desarrollo, la cual consideramos tendrá un futuro promisorio en las próximas décadas. Confiamos en que lo disfrutarán, tanto como nosotros lo hemos hecho en su momento.

Gerardo Sierra Martínez
Javier Cuétara Priede

Ciudad Universitaria, 6 de julio de 2009

LINGÜÍSTICA DE CORPUS: ¿UNA MATERIA DE LINGÜÍSTICA O DE INGENIERÍA?

GERARDO SIERRA MARTÍNEZ
GIL-IINGEN, UNAM

Gerardo Sierra: El cuestionamiento del título de esta ponencia abre una discusión controversial que vale la pena afrontar, tratándose de la primera plática en el seno de este *Tercer Coloquio de Lingüística Computacional*, en donde vale la pena preguntarse de manera similar si la lingüística computacional es un tema que debiera debatirse en la Facultad de Filosofía y Letras o en la Facultad de Ingeniería. Pero, más que cuestionar sobre este coloquio, al que estamos aquí por interés o por curiosidad, vayamos al punto de esta plática.

Antes de continuar, quiero hacer un paréntesis para aquellos que han seguido de cerca mi trayectoria en diversos foros, pues se preguntarán por qué, en lugar de apoyarme en una presentación vistosa con Power Point en la que intento captar la atención de la audiencia, ahora en esta plática inaugural recurro a la lectura. Por el contrario, a aquellos que no me conocen y que han asistido a grandes conferencias en lingüística, no les parecerá extraño sino que les parecerá más propio este estilo. Así, lo que pretendo desde un inicio es portar la camiseta de lingüista que gracias a esta universidad puedo, con orgullo, mostrar, y con ello quisiera más que verme como un extraño del Instituto de Ingeniería que viene a contaminar, como el colega que he buscado ser, preocupado por el avance de la investigación y la enseñanza en lingüística.

Ahora bien, para entender nuestro tema central, qué es la lingüística de corpus, cuál es su dominio y cuál su ámbito de desarrollo, vale la pena iniciar con la definición misma de lo que significa un corpus. Algunos de mis alumnos aquí presentes no me dejarán mentir en que desde la primera clase del curso lo definimos como: *la recopilación de un conjunto de textos de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos*. Fue toda una clase de dos horas para ahondar en los detalles de esta definición, que por cuestiones de tiempo tendré que omitir, sobre cuáles son estos criterios mínimos y cuáles son estos análisis lingüísticos.

La ventaja principal de los corpus lingüísticos es que éstos se componen a partir de hechos reales, por lo que sus resultados ofrecen una base empírica sólida para el análisis lingüístico. Así, su función central es establecer la relación entre la teoría y los datos, por lo que un corpus equivale a una muestra real a pequeña escala, la cual permite hacer hipótesis pertinentes respecto al funcionamiento de una lengua natural.

Haciendo un poco de memoria, se puede notar que las labores con corpus no son nuevas: Samuel Johnson, desde un enfoque lexicográfico, fue uno de los primeros en hacer trabajos de este tipo para su diccionario desde 1755, logrando agrupar unas 40,000 palabras organizadas de forma alfabética, y vinculadas con sus correspondientes definiciones. En esta obra podemos ver cómo un corpus de palabras se elabora bajo ciertos criterios específicos. Para el siglo XIX, gracias al interés de la filología europea por la evolución histórica y variación dialectal entre lenguas, se llevaron a efecto trabajos de recopilación de palabras sumamente valiosos, tales como el que realizó Joseph Wright para la edición de su *Diccionario de Inglés dialectal*.

De hecho, buena parte de la lingüística de principios del xx consideraba el trabajo de corpus como una herramienta sumamente valiosa, la cual aportaba datos de primera mano respecto de cualquier lengua. Franz Boas, por ejemplo, logró coleccionar un gran número de datos sobre diferentes lenguas amerindias de los Estados Unidos, siguiendo una metodología precisa que delimitó en mucho el análisis estructuralista norteamericano.

En nuestra Máxima Casa de Estudios, varias han sido las veces que se han utilizado los corpus para diversas investigaciones lingüísticas de excelencia, siguiendo la tradición de nuestro querido Lope Blanch de hacer una selección rigurosa de documentos para disponer de distintos registros en su *Estudio Histórico del Español de América*. Tanto los estudios sincrónicos y diacrónicos del español, junto con los estudios del español como lengua materna y de la historiografía lingüística que se vienen desarrollando en el Instituto de Investigaciones Filológicas, en el Centro de Lingüística Hispánica, no en vano llamado “Juan M. Lope Blanch”, y parten del análisis de colecciones de textos, orales y escritos, que constituyen diversos corpus lingüísticos.

Conscientes de su importancia, el maestro Javier Cuétara y un servidor, el primero en cuestiones orales y el segundo en textos, nos hemos comprometido a la tarea de impartir a los futuros lingüistas en esta Facultad los métodos para diseñar y construir corpus, las herramientas y técnicas para su análisis, así como las aplicaciones a diferentes fines lingüísticos. No quisiera comentar las renuencias de algunas autoridades para impartir nuestra cátedra, sino el gran interés por los alumnos en aprender de estos humildes profesores y cambiar su camino de las Letras por el de la Lingüística.

Sin embargo, por todo lo que he mencionado sobre los estudios basados en los datos reales que nos proporcionan los corpus, cabe preguntarse el porqué del título de esta ponencia. Es evidente su importancia y uso en las investigaciones lingüísticas, pero por qué preguntarse si es una materia que debiera impartirse en la Facultad de Ingeniería. No contestaré esto en función de los argumentos fatuos que me dieron sobre el porqué este semestre no se abrió la materia de lingüística de corpus en el Colegio de Letras Hispánicas, pues dudo seriamente que sea una aversión mal ganada ante un prejuicio a las ingenierías o a la computación, o bien por no querer desprenderse del romanticismo a la tradición de anotar corpus en fichas de papel con puño y letra del investigador, para quedar grabado en un Museo de la Filología.

Para explicar la razón del tema inaugural de este coloquio, permítanme cambiar mi camiseta de lingüista, y por tanto la forma de presentación que he seguido, por la camiseta que orgullosamente he llevado durante los últimos ocho años en la UNAM, la de ingeniero lingüista.

En el curso que estamos dando este semestre, actualmente han asistido —si bien no todos están inscritos— diez alumnos del Colegio de Letras Hispánicas de esta Facultad y cuatro alumnos del Posgrado de Lingüística, lo cual es obvio puesto que estamos hablando de que el curso de corpus está orientado para los lingüistas. Sin embargo, con todo, hay tres alumnos de la carrera de Ingeniería en Computación y tres alumnos del Posgrado en Computación.

Aquí habría que hacerse varias preguntas.

La primera pregunta sería: ¿por qué lingüistas y a la vez ingenieros? ¿Existe alguna razón? Bueno, yo supongo que una de las primeras respuestas sería el hecho de que a los ingenieros les interesa la cuestión de Humanidades y, de alguna manera, quieren asistir a un curso de esta naturaleza. Sin embargo, la realidad no es así. No es tanto que algunos ingenieros nos hayan convertido en lingüistas, como sería el caso de Luis Fernando Lara, Raúl Ávila, Alfonso Medina, un servidor y muchos otros más. No. Es, más que nada, porque hay un interés conjunto, y este interés conjunto no lo debemos olvidar.

La siguiente pregunta sería: ¿cuál sería la pertinencia de este curso en la Facultad de Ingeniería y en la Facultad de Filosofía y Letras? Esa es una gran duda que me gustaría en algún momento contestar, y me gustaría que llegáramos a un consenso. Me imagino dos respuestas a esta pregunta. La primera de ellas es la aplicación de los corpus para la lingüística computacional, que es el coloquio en el que estamos en este momento. Si estamos hablando de que la lingüística computacional es un área interdisciplinaria, en donde se tiene por un lado ingenieros en computación y, por otro, lingüistas, entonces es

comprensible, en este sentido, si los corpus van a servir como herramienta, como un recurso léxico para la lingüística computacional, que los ingenieros estén interesados en esta área. De aquí, alguna de las áreas, que inclusive se van a tocar en este coloquio: por ejemplo, la que viene después de la mía, la de enseñanza de lenguas; las cuestiones de tecnologías de voz, de las que también hay varias ponencias en este coloquio; la recuperación y extracción de información, que también tocaremos en algún momento; y de ayuda a la escritura, que es otro de los tantos temas que se tocan en el coloquio. Éstas son simplemente algunas de las áreas que se presentan en este evento. De esta manera, consideramos que es una de las razones por las que habría entonces ingenieros en el curso.

Ahora, la segunda respuesta a la pregunta sería que los corpus son herramientas electrónicas que ayudan a procesar, organizar, administrar y analizar enormes cantidades de documentos. Necesitamos contar con una herramienta de esta naturaleza. Esto es, cuando manejamos grandes cantidades de textos, es necesario, de alguna manera, tener una herramienta que pueda procesar la información que estamos manejando; organizarla y extraer específicamente lo que estamos buscando; analizar justamente la información que nosotros queremos para propósitos lingüísticos. Entonces, dado que es necesario construir este tipo de herramientas, ésa sería otra de las razones por la que los ingenieros estarían involucrados en esta área.

¿Cuál sería la definición de corpus informatizados? Si bien ya he hablado de cuál sería la definición de un corpus lingüístico, vale la pena ahora decir cuál sería la definición de un corpus informatizado: es un conjunto de textos elegidos y anotados con ciertas normas y criterios para el análisis lingüístico, lo cual se parece completamente a la definición anterior, pero con esta salvedad: que se sirve de la tecnología y de las herramientas computacionales para generar resultados más exactos. Y no solamente para generar resultados más exactos, que éste sería uno de los puntos controversiales de un lingüista, sino además de una manera más rápida. Sin duda alguna, la velocidad de procesamiento de una computadora es infinitamente superior a la velocidad de procesamiento que tendría un lingüista buscando en su fichero. Y si hablo de “generar resultados más exactos”, simplemente me gustaría que ustedes contaran manualmente las palabras que existen en un corpus. Piensen ustedes en el *Quijote* de Cervantes. Para contar todas las palabras, y no solamente contarlas, sino, además, hacer concordancias de las distintas palabras o tener medidas de colocación, me gustaría saber si, de alguna manera, lo pueden hacer manualmente. Estoy de acuerdo con que pueden ustedes, manualmente, identificar las distintas partes de la oración de un texto, pero, de la misma manera, con la computadora lo pueden hacer y, de hecho, al igual que sucede con los sistemas de traducción, que no pretenden, como tal, duplicar el trabajo del traductor, sino servir como una herramienta, las computadoras pueden ser una herramienta para el lingüista para anotar, por ejemplo, las partes de la oración de un texto.

Simplemente, para que veamos la importancia de los corpus informatizados, podemos tener el ejemplo de la Real Academia Española. Ésta maneja un corpus de, aproximadamente, cuatrocientos millones de palabras —o de registros, como ellos lo manejan— en dos corpus: el Corpus de Referencia del Español Actual (CREA) y el Corpus Diacrónico del Español (CORDE), donde se encuentran datos desde los orígenes del español hasta unos años antes de nuestra época. En este corpus podemos poner como entrada una palabra, un lema, una palabra truncada. Así, por ejemplo, si nosotros buscamos “*de acuerdo*”, nos dará medidas estadísticas e indicará que existe *de acuerdo*, *de acuerdo con*, *de acuerdo a*, *de acuerdo en* y en ese orden de importancia. Estamos haciendo una búsqueda restringida en México, en el área de ciencias sociales, en 27 documentos, y obtenemos 293 casos. Me gustaría saber si ustedes pueden hacer un conteo de esto, trayendo todas sus

fichas sobre ciencias sociales en México, y buscar todas las veces que aparece *de acuerdo* y cuántas veces aparece *de acuerdo con*, *de acuerdo en*, *de acuerdo a*. Creo que es una labor que difícilmente la puede hacer el ser humano solo.

De igual forma, nos ofrece las concordancias de este *de acuerdo* y, de esa manera, nosotros podemos ver si el *de acuerdo a* es correctamente usado o no lo es. La Real Academia ha llevado a cabo la tarea tan grande de construir un corpus de 400 millones de palabras. Por ello, ustedes podrán darse cuenta la importancia que da la Real Academia al uso de corpus y también cómo hay toda una labor informática detrás para poder desarrollar una página del corpus que además es accesible y completamente gratis. Una de las ventajas que tiene un corpus de esta naturaleza, y a diferencia de cualquier corpus que tengamos en fichas, está, lógicamente, en que se puede compartir fácilmente y extender su aplicación y sus usos a diferentes fines lingüísticos. Y no solamente para aquellos que pensó el investigador original.

Pero, por otro lado, no solamente esto, sino que también nos puede dar las concordancias de este “de acuerdo” y, de esa manera, nosotros podemos ver si el *de acuerdo a* es correctamente usado o no lo es.

De esta manera nace lo que se conoce, hoy en día, como la *lingüística de corpus*, que es la parte de la lingüística en la que se estudian con medios informáticos, de diferentes tipos, grandes masas de datos, porque, evidentemente no vamos a hablar de hacer un análisis del *Primero Sueño* de Sor Juana Inés de la Cruz, que en su momento fue durante el inicio de la computadora y se hicieron dos trabajos en paralelo para mostrar que se podían encontrar las concordancias y el número de palabras del *Primero Sueño*, pero esto ya es historia. Hoy en día, hablamos de grandes masas de datos que serían inabordables de otro modo. Para analizar, por ejemplo, las características lingüísticas de un idioma en un determinado momento de su historia, de cierto tipo de textos, de un conjunto de autores o de un autor determinado, la ventaja del corpus de la Real Academia, en este caso, es que ustedes pueden concretizar su búsqueda únicamente, por ejemplo, para Cervantes y, de esa manera, el proceso se vuelve mucho más rápido.

REAL ACADEMIA ESPAÑOLA

Concordancias (RAE)

Consultar: *de acuerdo, en libros, en CREA, en Ciencias sociales, en MÉXICO*
 Resultado: 293 casos en 27 documentos.

OBTENCIÓN DE EJEMPLOS

Recuperar Concordancias Normal Clasificación: [dropdown]
 Agrupación: [dropdown] Marcas: [dropdown]

Cómo citar el CORPUS

De palabras	%	Casos
de acuerdo	100.00	293

Agrupaciones.

De 3 palabras	%	Casos	De 5 palabras	%	Casos
de acuerdo con	65.52	192	de acuerdo con la cual	1.70	5
de acuerdo a	19.45	57	de acuerdo con la costumbre	1.36	4
de acuerdo al	5.80	17	de acuerdo a las necesidades	1.36	4
de acuerdo en	2.73	8	de acuerdo a sus intereses	1.02	3
de acuerdo. Frente	0.34	1	de acuerdo con las necesidades	1.02	3
de acuerdo, se	0.34	1	de acuerdo con los intereses	1.02	3
de acuerdo y	0.34	1	de acuerdo con la edad	1.02	3
de acuerdo -agree	0.34	1	de acuerdo a los datos	0.68	2
de acuerdo.	0.34	1			

Cuadro 1. Búsqueda de concordancias en la página de la Real Academia Española

Pero, con todo, entonces no sé si siga todavía la interrogante: ¿para qué ingenieros en un curso de esta naturaleza? Sería como decir: ¿se trata, algo así como el agua y el aceite, que tenemos que estar separados y no podemos estar juntos los ingenieros y los lingüistas? ¿Es, acaso, una de las razones? ¿Y que podríamos decir: *¡no voy a acercarme a ellos porque me voy a contaminar!*? ¿O, por el contrario, podríamos decir: *¡podemos estrechar las manos y sumar nuestros esfuerzos con un solo objetivo y, de esa manera, lograr algo que no podríamos de ninguna manera hacer individualmente!*? Creo que ahí hay una gran diferencia y creo que tendríamos que abrirnos, justamente, a esta sinergia para poder hacer de los corpus las herramientas que deben ser hoy en día.

REAL ACADEMIA ESPAÑOLA

Concordancias (RAE)

Consulta: *de acuerdo, en Libros, en CREA, en Ciencias sociales, en MÉTICD*
 Resultado: 293 casos en 27 documentos.

OBTENCIÓN DE EJEMPLOS

Recuperar Concordancias Normal Clasificación: [dropdown]
 Agrupación: [dropdown] Marcas: [dropdown]

Cómo citar el CORPUS **Concordancias.**
 Pantalla: 1 de 12. Siguiente 1 2 3 4 5 6 7 8 9 10 11 12 Ver párrafos

Nº	CONCORDANCIA	AÑO
1	onia de iniciación, el discípulo recibe un mantra de acuerdo a la naturaleza de sus inclinaciones, gust	** 1990
2	ustos y temperamento por una parte y por la otra, de acuerdo a la deidad favorita del aspirante. El ma	** 1990
3	ejecuta el sacrificio con una formalidad rigurosa de acuerdo al código establecido, y brahma, el sacer	** 1990
4	a ramas se han dividido, a su vez, en dos partes, de acuerdo a su contenido: el karma-kanda, formado po	** 1990
5	ro que apela al diverso carácter de las personas, de acuerdo a su inclinación particular hacia el pensa	** 1990
6	o destacar alguna en particular sobre las otras, de acuerdo a su capacidad y las indicaciones de su ma	** 1990
7	por uno mismo, causados y aprobados por otros y, de acuerdo a su intensidad tenue, mediana o excesiva,	** 1990
8	realizarse en función exclusiva de la meditación de acuerdo al sentido formal del Raja-Yoga. Pranayama	** 1990
9	la práctica. 22. Habrá diferencias en los logros de acuerdo al esfuerzo aplicado, suave, mediano o int	** 1990
10	percibiendo. 17. Se conoce o desconoce un objeto de acuerdo al estado de ánimo de la mente. 18. Debido	** 1990
11	criterio de clasificación adoptado por los países de acuerdo al promedio de tiempo de permanencia obser	** 1990
12	n. Turismo autónomo es el que practica el turista de acuerdo con un itinerario que él mismo elabora con	** 1990
13	geográfico o cultural de un lugar. De tal manera, de acuerdo a su constitución misma, podemos distingui	** 1990
14	las temporadas más favorecidas por los turistas, de acuerdo con las condiciones climatológicas más apr	** 1990
15	por lapsos mayores de una semana. De tal manera, de acuerdo con la clasificación adoptada, este hotel	** 1990
16	ale, etc., ya que tiene la opción de seleccionar, de acuerdo con sus preferencias, el mejor sitio para	** 1990
17	nado, procederemos a clasificar estas localidades de acuerdo a los atractivos que presentan, mismos que	** 1990
18	que se manifiesta en múltiples y variadas formas, de acuerdo con la estructura socioeconómica del suiet	** 1990
19	ica. Esta forma acusa características especiales, de acuerdo con el recurso utilizado para su proyecta	** 1990
20	Asimismo, designará al personal de la Secretaría, de acuerdo con el reglamento que al respecto sea apro	** 1990

Cuadro 2. Obtención de concordancias en la página de la Real Academia Española

O bien, podríamos preguntarnos ¿en qué momento vamos a decirles a los alumnos: *hoy no vengan los lingüistas porque nos vamos a dedicar únicamente a cuestiones de cómputo, u hoy, no vengan los de cómputo porque vamos a tratar cuestiones lingüísticas que a ustedes no les interesa?*

Primeramente, en la introducción a la lingüística de corpus, definimos lo que es un corpus lingüístico, vemos los corpus existentes, clasificamos los corpus y vemos que hay una variedad tremenda de ellos. Además vemos cómo el Internet puede, de alguna manera, ser usado como un corpus; si bien desde la primera clase lo definimos, el Internet no es un corpus como tal, porque no cumple con los criterios y normas que debieran seguirse para un buen corpus. Entonces, hasta ahí ¿podríamos decir: *está bien, es adecuado para los lingüistas tomar esta primera clase?* Podríamos decir que sí. Los de cómputo podrán decir: *qué aburrido, a lo mejor nada más el Internet como corpus, pero ¿qué vamos a ver? ¿Vamos a diseñar páginas de Internet?* No, pero trataremos de utilizar muchas herramientas muy interesantes que se van creando hoy en día: las herramientas, técnicas y aplicaciones de la computación en el desarrollo de la investigación lingüística y en las tecnologías del lenguaje.

TEMARIO DEL CURSO
<p>1. INTRODUCCIÓN A LA LINGÜÍSTICA DE CORPUS</p> <ul style="list-style-type: none"> • Definición de corpus lingüístico • Clasificación de corpus • Corpus existentes • Internet como corpus
<p>2. COMPILACIÓN DE CORPUS</p> <ul style="list-style-type: none"> • Corpus textuales • Corpus orales
<p>3. ANOTACIÓN</p> <ul style="list-style-type: none"> • Bases para anotación de corpus • Textual • Morfosintáctica • Sintáctica • Semántica, discursiva y referencial • Fonética y prosódica
<p>4. HERRAMIENTAS Y TÉCNICAS DE ANÁLISIS</p> <ul style="list-style-type: none"> • Conteo de palabras • Concordancias • Colocaciones y medidas de asociación • Análisis fraseológico y oracional • Herramientas de análisis disponibles
<p>5. RESULTADOS DEL PROCESAMIENTO DE CORPUS</p> <ul style="list-style-type: none"> • Aplicaciones a la lingüística sincrónica y diacrónica • Usos para lingüística aplicada: lexicografía, terminología y enseñanza de lenguas • Corpus para la ingeniería lingüística

Cuadro 3. Temario del curso de lingüística de corpus

A partir del segundo tema, sobre compilación de corpus textuales y orales, buscamos realmente que los alumnos vayan compilando. Vemos cómo van a seleccionar los libros, cómo van a seleccionar los textos y, luego, cómo van a realizar la anotación textual, morfosintáctica, sintáctica, semántica, discursiva, referencial, fonética y prosódica. Toda esta anotación ¿mientras sea con papelito, sí lo hacen?, ¿pero si es con la computadora ya no? Yo creo que no.

El curso lo terminamos con las herramientas y técnicas de análisis, desde el conteo de palabras, concordancias y demás. Finalmente los resultados de procesamiento de corpus tanto para la lingüística aplicada como para las tecnologías del lenguaje.

A manera de conclusión, cabe mencionar que, con miras a lograr su óptimo aprovechamiento, actualmente muchos corpus se encuentran en soporte informático, facilitando así el uso de herramientas computacionales para fines lingüísticos, lo que permite trabajar incluso con millones de textos. Esto se traduce en un mayor rendimiento para cualquier investigación. Por esto, resulta necesario aprender a detalle las técnicas y recursos relacionados con los corpus lingüísticos, con miras a dominar su teoría y su práctica. Desde esta óptica, lo que se espera lograr es que los alumnos comprendan la importancia de crear y manejar corpus lingüísticos para sus trabajos de investigación, así como tomar en cuenta los múltiples recursos que hay actualmente para su explotación.

Sin pretender agotar todas las posibilidades que ofrecen actualmente los corpus lingüísticos al estudio del lenguaje, ya sea desde un enfoque netamente lingüístico, o ya sea desde una visión multidisciplinaria como la que toma este curso, cabe considerar algunos aspectos:

- La necesidad de homogeneizar conceptos básicos respecto a la concepción y producción de esta clase de materiales, de modo que pueda fijarse un conjunto de estándares mínimos útiles para llevar a efecto cualquier proyecto de corpus.
- Por lo mismo, es indispensable compartir conocimientos venidos de áreas diferentes (por ejemplo, cómputo, estadística y lingüística), con miras a mejorar el desempeño de los alumnos en el cumplimiento de esta clase de tareas.
- Finalmente, el objetivo de esta clase de cursos debe ser, además de ofrecer conocimientos y métodos útiles para las investigaciones en lingüística de corpus, crear conciencia entre lingüistas sobre la enorme necesidad y las claras ventajas que ofrece desarrollar corpus con recursos electrónicos. Si bien es cierto que no es un requisito inmutable, la evolución y auge de las tecnologías para manejo y flujo de la información hacen necesario adquirir y administrar dichos datos en muy poco tiempo. Un corpus en este caso es una excelente herramienta que permite filtrar datos no relevantes (según los criterios que establezca la persona que lo estructure), y de este modo tener una visión realista sobre un fenómeno o varios fenómenos lingüísticos, la cual pueda ser compartida con la comunidad científica correspondiente.

Con esto, no me resta más que agradecer su atención e invitarlos a adentrarse en el mundo de la lingüística moderna. Hoy, en este siglo XXI en el que podemos aprovechar las bondades de las herramientas, técnicas y aplicaciones de la computación en el desarrollo de la investigación lingüística y en las tecnologías del lenguaje. ■

LA RED ELECTRÓNICA MUNDIAL COMO HERRAMIENTA DE ENSEÑANZA-APRENDIZAJE: EL DIPLOMADO PARA FORMACIÓN DE PROFESORES DE ESPAÑOL COMO LENGUA EXTRANJERA

LAURA GALINDO ISLAS
CEPE, UNAM

César Aguilar: Muchas gracias al Dr. Sierra por su participación. En algún momento, Roman Jakobson decía: “Las ventajas en las reuniones de lingüística es que no son, de algún modo, políticas”. Entonces, siempre hay una apertura al diálogo y siempre se discute desde una perspectiva más científica que personal. Tomado en cuenta esto, la invitación sería que las preguntas las podemos acumular al final y pasar a la siguiente presentación.

En este caso se trata de la maestra Laura Galindo Islas, quien tiene una maestría en lingüística aplicada y actualmente es profesora en el Centro de Enseñanza para Extranjeros (CEPE). Ha publicado una serie de colecciones en la que ella ha participado como autora. En estos momentos, el trabajo que nos presenta tiene el siguiente título: *La red electrónica mundial como herramienta de enseñanza-aprendizaje: el diplomado para formación de profesores de español como lengua extranjera*. Los dejo con la maestra Galindo.

Laura Galindo: Tenemos un tema elaborado para mostrar el diplomado, sobre todo en diferentes instituciones extranjeras, donde no nos es posible asistir. En el cuadro 1 puede apreciarse la página de entrada al diplomado en formación de profesores de español como lengua extranjera.



Cuadro 1. Página del Diplomado para profesores de español como lengua extranjera

Tenemos la fortuna en el CEPE de estar abiertos a las nuevas tecnologías. En la presentación les voy a contar un poco sobre la historia de este diplomado. Se trata de un diplomado en formación de profesores de español como lengua extranjera en línea, existente desde el año 2003.

Tenemos el gusto, además, de contar aquí, en el auditorio, con la presencia de la sexta generación de alumnos que cursan este diplomado de una manera presencial, el cual es un curso bastante similar, cuando menos en cuanto a los contenidos de las materias. La idea es preparar a profesionales en los aspectos lingüísticos, didácticos y de actitud, relacionados con la enseñanza de español como lengua extranjera, porque sabemos que un maestro de lenguas, normalmente, era un personaje que se hacía sobre la marcha, con la experiencia y desarrollando metodologías por su cuenta.

A nivel general, podemos hablar sobre el tipo de sesiones que este diplomado contempla; es semipresencial. Existen materias, como la de *Planeación, observación y práctica en clases*, que serían muy difíciles de llevarse a cabo a través de Internet. Por esa razón, 90% de las materias se lleva a cabo por este medio, mientras 10% es presencial; entonces, los alumnos de cualquier parte del mundo tienen que venir a presentar sus prácticas de clase aquí y a realizar sus observaciones.

Los requisitos para cursar este diplomado son que tengan una licenciatura afín, generalmente, humanidades o ciencias sociales, un compromiso de veinte horas de trabajo a la semana y conocimientos y experiencia en cómputo.

La idea, entonces, es formar especialistas en la enseñanza del español que puedan ofrecer una respuesta innovadora, ética e independiente a las demandas educativas en este campo, mediante la aplicación de conocimientos, habilidades y destrezas adquiridas durante el diplomado.

La metodología que se contempla está basada en la lectura de información sobre el tema y en la ejecución de actividades. Es interesante mencionar que este diplomado tiene todos sus contenidos en línea.

Con respecto a la estructura, tienen las materias en el cuadro 2. Ha variado un poco el número de horas, pero en total son 450 horas que se imparten en dos módulos; el primero sobre el sistema y uso de la lengua, y el segundo sobre formación didáctica.

El cupo máximo de estudiantes es veinticinco. Normalmente los estudiantes llevan un año en cursarlo. Es un diplomado que permite, si hay algún problema específico con los alumnos, que pueda recurrirse al año siguiente. Hay requisitos para obtener un diploma. Se obtiene diploma con constancia.

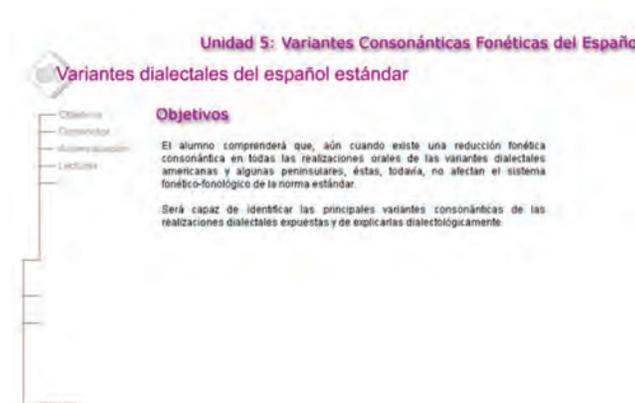
Las materias son diez. Casi todas contemplan un número de cinco, seis o siete unidades. Si entramos a una sección llamada "métodos de enseñanza", allí se explican los objetivos generales y particulares. Por ejemplo, se habla de métodos experimentales, el enfoque oral, se habla un poco de la evolución de la metodología de enseñanza de las lenguas. Hay cierto tipo de actividades. La idea es que los alumnos realicen estas actividades en el tiempo que ellos puedan hacerlo y las envíen a su maestro. Este último tiene la capacidad de responder, para llevar a cabo una retroalimentación.

Podemos ver ahora otra materia: la de variantes dialectales del español estándar. Es una materia que suele ser una de las preferidas porque todo el material está en línea; entonces, si los alumnos tienen interés en ampliar esta información, pueden hacerlo. Suele suceder que son los mismos alumnos los que proporcionan bibliografía actualizada, nueva, muchas veces de Internet, y la comparten.

d iplomado en formación de profesores de español como lengua extranjera en línea		Sistema y uso de la lengua	
		<ul style="list-style-type: none"> Lingüística general 45 horas Fonología y fonética aplicada 30 horas Morfosintaxis (análisis gramatical) 45 horas Aspectos gramaticales problemáticos de la enseñanza del español 60 horas Variantes dialectales del español estándar 45 horas 	TOTAL 225 horas
D.R. © INDA: 03-2006-070813143900-01 contactos créditos puf		Formación Didáctica	
		<ul style="list-style-type: none"> Didáctica (Métodos y técnicas de enseñanza) 45 horas Habilidades lingüísticas y diseño de materiales 60 horas La tecnología en la enseñanza de lenguas 30 horas Evaluación 45 horas Planeación, observación y práctica de clase 45 horas 	TOTAL 225 horas

Cuadro 2. Materias del diplomado

Entremos ahora a unidades, donde podemos ver todo lo relacionado con variantes consonánticas del español (véase cuadro 3). Se refiere a comprender traducciones fonéticas en las realizaciones orales de las variantes dialectales americanas y/o peninsulares.



Cuadro 3. Ejemplo de una unidad del diplomado

Por ejemplo, uno de los ejercicios llevados a cabo consiste en que los alumnos escuchan una grabación de un área andaluza y deben realizar la descripción de lo que oyeron.

Como les comentaba, este diplomado para formación de profesores surgió para satisfacer una demanda: la Secretaría de Relaciones Exteriores, en el año 2003, nos solicitaba maestros del CEPE para dar clases al extranjero, entonces era imposible

atender esta necesidad y, en un esfuerzo conjunto con el gobierno mexicano, pudimos armar este diplomado y capacitar a los maestros. Todos los maestros del CEPE se animaron con este diplomado. Hasta el momento, ha formado a 261 profesores de español.

Su antecedente son los cursos de formación para profesores de español que se impartían durante el verano, pero no había un programa o una actividad sistemática hasta que se generó este diplomado. En 2001, en lugar de este diplomado, teníamos el *Diplomado en enseñanza del español para no hispano hablantes* que después se convirtió en el *Diplomado para la enseñanza del español como lengua adicional* y ese diplomado, que algunos de nuestros compañeros aquí presentes están cursando en su sexta generación, es un antecedente, también del DEFPELE.

Como les comentaba, fue en el 2003 que, con 450 horas en línea, empezaba a funcionar el DFEPELE. Este diplomado para profesores de español forma cuadros docentes especializados en metodología para la enseñanza del español a extranjeros a través de los dos diplomados que ya mencioné, el DEFPELE y el DELA, uno a distancia y uno presencial. Hasta el momento, hemos formado 261 profesores. Es interesante que casi la mitad de esos profesores se formaron en línea; 123 profesores se formaron en línea en cuatro años, y el resto, que es un 53%, se formó en el presencial en seis años. Llevamos, entonces, seis generaciones de profesores que ejercen en los cuatro continentes; 3%, en Asia; 5%, en Europa; 27% en Norteamérica, Canadá y Estados Unidos y, en México, 65%.

El propósito de este diplomado, como ya lo hemos mencionado, es satisfacer esa demanda; el español se ha vuelto un activo fijo en todo el mundo y a todos los hispano hablantes alrededor del mundo, en algún momento, se les solicita que den clases, ya no de literatura, no analizan la poesía española, sino que den clases de español porque es realmente algo que está siendo sumamente necesario, sumamente capitalizable en el mundo de los negocios, en el mundo de las relaciones con América Latina en Europa.

Entonces, este diplomado pretende lograr una formación resultante de conocimientos del sistema lingüístico, así como de teorías metodológicas, didácticas y de evaluación de enseñanza de las lenguas. Específicamente, estos conocimientos, con respecto al sistema lingüístico, intentan describir la lengua en los diferentes niveles: el fonético, el morfosintáctico, el semántico y el pragmático. Por otro lado, también existe la intención de dotar al alumno de conocimientos sobre las diferentes variantes dialectales del español. En cuanto a los procedimientos didácticos, hay un énfasis en la metodología de la enseñanza de las lenguas y en cómo éstas han ido evolucionando. Se trabajan, además, los conceptos y los criterios fundamentales en la formación del proceso de aprendizaje.

GENERACIÓN	PERIODO	ASPIRANTES	INSCRITOS
1a	3 feb 2003	16	16
2a	3 sept 2003	23	19
3a	11 feb 2004	28	14
4a	7 mar 2005	52	25
5ª	17 abril 2006	68	25
6ª	12 feb 2007	58	24
Total	2003 a 2007	245	123

Cuadro 4. Estadísticas del DEFPELE en línea 2003-2007

Como les había comentado, son diez materias, con un componente teórico y práctico, cada una de ellas aplicada a la enseñanza del español a extranjeros.

En el cuadro 4, podemos ver estadísticas que resultan de interés. Tenemos, entonces, la estadística del DEFPELE.

Podemos comparar las del DEFPELE y las del DELA. Como les había comentado, el DEFPELE tiene cuatro años y es interesante que, como ustedes ven, el número de aspirantes, digamos el interés que ha despertado este tipo de diplomado, fue creciendo. Nos detuvimos un poquito en cuanto a descripción el año pasado, ya que a esta generación se le exigió que estuviera titulada y por esa razón bajó un poco el ingreso, pero en general, estas cifras demuestran un crecimiento cada vez mayor. Es interesante ver cuántos aspirantes tenemos y a cuántos podemos atender. También debemos considerar que hay que pasar un examen de admisión y, como les comentaba, hay que estar recibido de una licenciatura. Esto siempre reduce mucho el número de aspirantes.

GENERACIONES	PERIODO	ASPIRANTES	INSCRITOS
2001			19
2002	16 ene - 13 dic 2002	46	21
2003	13 ene - 24 nov 2003	56	19
2004	12 ene - 10 dic 2004		21
2005	10 ene - 9 dic 2005	54	22
2006	16 ene - 13 dic 2006	74	20
2007	17 ene - 10 dic 2007	23	16
Total		253	138

Cuadro 5. Estadísticas DELA presencial 2001-2007

En el cuadro 5 vemos las cifras del DELA, el cual tiene seis años. También en el DELA se ha mostrado un gran interés; el número es apenas un poco mayor, a pesar de que son seis años, dos más que el DEFPELE. Como ven ustedes aquí, el número de aspirantes bajó por la misma razón que ya mencionamos: esperamos a que la gente se titule y luego le damos oportunidad de que se especialice en otras cosas porque si no, es muy difícil. Estábamos formando muchos maestros que no podían realmente trabajar o irse comisionados al extranjero por no tener su título de licenciatura.

Me interesó destacar el índice de conclusión del DEFPELE. Es, más o menos, un 70%. Digo más o menos porque, como ustedes se dan cuenta, estoy tomando en consideración las cifras de una sexta generación que recién comienza. Es hoy doce de febrero; actualmente estamos dando la clase de lingüística teórica y aplicada y la clase de didáctica. En el primer par de materias, hay una deserción, por problemas personales. Yo calculo, siendo optimista, que vamos a tener cuatro deserciones nada más, y esto me dio, más o menos, 70% de maestros que, esperamos, concluyan sus estudios.

Con el DELA, el índice de conclusión es un poco mayor. Esto creo que se entiende; precisamente porque parece ser que los aspirantes, los candidatos a este tipo de diplomado en línea piensan que es muy fácil tomar un curso por Internet.

DEFPELE formó casi a la mitad (47%) del total de profesores en menos tiempo que el DEELA (53%).

Firma de convenios con Universidades de Canadá y EEUU para ofrecer el diplomado en línea en programas conjuntos.

Cuadro 6. Logros del DEFPELE

Todo esto implica una disciplina, implica una capacidad, una fuerza de voluntad, un tipo de participación y de interacción con sus compañeros y con los profesores, y, a veces, no están dispuestos a darlo. Sin embargo, de todas maneras, el índice de conclusión de estudios es alto.

Como conclusión, cabe destacar que este joven diplomado formó casi a la mitad de sus maestros ya profesionistas en menos tiempo que el diplomado presencial, lo que es una ventaja. Por otro lado, el hecho de que ha resultado sumamente atractivo para diferentes universidades en Estados Unidos, en Canadá, en Brasil, el hecho de tener la posibilidad de establecer relaciones, por ejemplo con Brasil, y formar en cascada a muchos profesores especializados, es otro gran logro de este diplomado.

En cuanto a retos, hay algo muy importante con respecto a la figura del tutor en línea. Esta universidad, como muchas otras —creo que los ingenieros nos lo pueden decir muy bien— todavía no está preparada, en cuanto a cuestiones administrativas o jurídicas, para darle al tutor en línea su lugar en cuanto a salario, por ejemplo, en cuanto a la valoración que realmente requiere el tipo de trabajo que se hace. Así que no se firma una tarjeta, no se le cuenta el tiempo dedicado en línea al profesor ni se checa tarjeta ni nada parecido. Hemos enfrentado problemas, por esta falta de definición, problemas en la contratación de nuestros maestros. Entonces, uno de los retos es incrementar, precisamente, nuestra planta docente y lograr que esta modalidad sea atractiva para nuestros profesores, pero para nuestros profesores ya expertos en la enseñanza del español. No tanto para las nuevas generaciones, sino para los que ya saben hacer esto presencialmente, el capacitarse para incursionar en esta modalidad.

También tenemos que incrementar el número de profesores extranjeros que toman este diplomado por la cuestión de la lengua en el examen que tienen que presentar para inscribirse a este diplomado, que es totalmente en español. Queremos también reducir el índice de deserción. Y, con todo esto, lograr la acreditación internacional que, sin duda, nos va a dar el derecho a la firma de convenios que logrará que, en un momento dado, tengamos posgrados conjuntos, diplomados con aplicación doble (muy frecuentes y solicitados en la actualidad), gracias a la globalización. La página de DEFPELE es <http://diplomados.cepe.unam.mx/DEFPELE/>, y los invito con mucho gusto a que se acerquen a averiguar un poco más sobre esto.

César Aguilar: Muchas gracias a la maestra Laura Galindo por su exposición. A mí, en lo personal, me pareció muy interesante. Sobre todo, el hecho de que una institución como el CEPE, y esto creo que ya se ha comentado en otros lugares, se esté involucrando con este tipo de nuevas tecnologías, sobre todo para un fin como lo es la formación de profesores y sí, estoy de acuerdo en lo que ella comentaba: pareciera que, actualmente, la tendencia va más orientada a dejar una postura tradicional del profesor como alguien que emana conocimiento y se lo brinda a todos los demás, para pasar a una relación de tipo apoyo-integración con los estudiantes, vía el uso de nuevos recursos electrónicos. Muy interesante y yo creo que va a haber preguntas al final, también, al respecto. ■

ESCENARIOS VIRTUALES PARA LA ENSEÑANZA DE LA PRONUNCIACIÓN DE LENGUA EXTRANJERA

ROSA ESTHER DELGADILLO
CEPE, UNAM

César Aguilar: A continuación pasamos con la maestra Rosa Esther Delgadillo. Ella tiene una maestría en literatura por parte de la UNAM, es doctora en literatura y actualmente ha terminado la maestría en lingüística aplicada. Ella también es parte del Centro de Enseñanza Para Extranjeros (CEPE). Igual que el caso anterior, ha publicado varios trabajos al respecto, entre ellos, artículos que ha presentado en varios lugares y es responsable del proyecto *IXTLI: el aparato fonador humano en tercera dimensión*. Lo que aquí nos presenta lleva por título: *Escenarios virtuales para la enseñanza de la pronunciación de lengua extranjera*. Los dejo, entonces, con la Dra. Delgadillo y su presentación.

Rosa Esther Delgadillo: Buenos días. Realmente lo que menos importa son los títulos, me interesa rescatar el trabajo con el Dr. Sierra; el trabajo del lingüista y el trabajo del ingeniero. Si recordamos un poco, ¿de dónde viene la palabra ingeniero? Viene de la palabra *ingenio* y, precisamente, fue Descartes quien recuperó ese concepto y lo incluyó en un vasto terreno de donde surge la palabra *ingeniería*, pues realmente *ingeniería* quiere decir ‘construcción’. También un gran lingüista, Chomsky, toma el concepto de *ingenio*; el generativismo de ahí viene; algo que se está generando, que se está construyendo, que se está dando.

Bien. Quiero platicarles un poco de este proyecto: es un proyecto que está en desarrollo y ganó un concurso. Su génesis está determinada, precisamente, por las necesidades que tenemos cuando enseñamos a extranjeros. Ustedes saben que, cuando se enseña otra lengua, nos enfrentamos a muchos problemas, sobre todo la forma en que cada uno de ellos, de nuestros alumnos extranjeros, ha conformado su sistema lingüístico de una manera natural. Hay una transferencia lingüística que se encuentra en cada uno de los hablantes. Y, precisamente, dentro de este campo de la lingüística, de la lengua, como se ha mencionado aquí, tenemos esas cuatro áreas fundamentales: una que son los sonidos, la forma de las palabras, la organización de las palabras y su significado.

El proyecto tiene grandes intereses, principalmente, tres. Uno, que es ayudar a los alumnos extranjeros que asisten al CEPE a que puedan hacer la transferencia lingüística del sistema que traen “guardado” en su cerebro a un nuevo sistema, empezando por los sonidos. El otro es ayudar a los alumnos del curso de formación de profesores, del cual también soy tutora, o asesora (como me quieran llamar, no me importa). Este concepto de trabajar con las nuevas tecnologías nos da la oportunidad de ver todo lo que necesitamos conocer para poder transmitirlo. También ha tenido la oportunidad de impartir la materia de fonética, tanto en forma presencial, como en línea, y me di cuenta de fenómenos que escapan a la literatura. Entonces, con esta idea, fue que me atreví a proponer el proyecto que se llama: *El aparato fonador humano en tercera dimensión*.

Los antecedentes son, un poco, los que les he mencionado. Luego, otro aspecto importante: ¿cómo se configura la lengua a nivel cerebral? Estos hallazgos son recientes. Muchos de nosotros ni siquiera nos imaginamos la importancia que tiene ahora la forma de cómo se está configurando todo un sistema de comunicación en el cerebro. Por ejemplo, en la jerga lingüística se habla de la *interlengua*, esa interlengua nos permite crear un nuevo sistema de comunicación cuando aprendemos otra lengua. Sin embargo, tenemos que ir más allá. Tenemos que hablar de un *intra-lenguaje* y, precisamente, es en la conformación de este intra-lenguaje en donde estoy focalizando este desarrollo del aparato fonador. Vemos que el cerebro funciona a través de una

serie de conexiones neuronales. Esto no lo voy a tocar porque realmente es apasionante, y nos llevaría aquí como tres o cuatro horas.

Sabemos también que el lenguaje tiene bases neurobiológicas y que todo esto está conformado en este sistema. Ya un psicólogo, basado un poco en Chomsky, Eleunler, hablaba de esa estructura interna que cada uno de los hablantes tiene y que nosotros, en el momento en que empezamos a socializar con los demás, empezamos a reactivar. Y, ¿qué es lo primero que reactivamos? Pues, precisamente, son sonidos. Son balbuceos. Esos sonidos no tienen significado aún, pero son sonidos que van conformando ya lo que más adelante va a ser, primero, una forma de comunicación con el entorno, primero, y después con los demás. Todo esto, les decía, no lo voy a trabajar, aunque me parece fascinante, pero es tema de otra ponencia. Empero, considero interesante tomar conciencia de esa función creativa de la que nos hablaba Chomsky. Siempre estamos creando. Estamos jugando con la lengua que tenemos guardada aquí (en la cabeza) y que ya ha dejado huella en nosotros. Tenemos una función de envío y tenemos una función auditiva, tenemos la ruta de los nervios y los órganos vocales. Éste es, un poco, el origen del proyecto que estoy generando.

Vamos a ver la fonética. Si bien, generalmente, hablamos de *fonología*, a mí lo que me interesa destacar es ese conjunto de sonidos que corresponde a todas las lenguas, independientemente de que su representación gráfica sea diferente. Todos tenemos esa conformación de sonidos en la mente. Entonces, vamos a ver cómo se representa en cada lengua, a nivel de sonido. De ahí, podemos tomar en cuenta tres perspectivas. Una, que es, precisamente, la *fonética articuladora*. Fundamental. Si hablamos francés, sabemos que algunos sonidos los articulamos con la garganta; otros, con los dientes; otros, con el paladar. Todas las lenguas tienen esa conformación. Lo que sucede es que la realización en cada lengua es diferente. Por eso es importante tener en cuenta cómo se articulan los sonidos en español. Tenemos, también, la *fonética perceptual*: cómo percibimos los sonidos a partir de la audición. Cada uno de nosotros tiene diferentes formas de percibir auditivamente, puesto que, a nuestro alrededor, hay una gran cantidad de sonidos y solamente los identificamos si los asociamos con aquellos que tenemos guardados en nuestro cerebro. Y, finalmente, la *fonética acústica*. Este aspecto es fundamental para el desarrollo que estoy haciendo, ya que uno de los problemas que hemos tenido, principalmente, es la sincronía entre la imagen de la realización y el sonido.

Ahora bien, el ingeniero lingüista Gerardo Sierra nos mencionaba que ya no podemos vivir aislados. Una de mis principales limitantes fue poder transmitir mi sentir de lingüista a los compañeros que me estaban ayudando con el desarrollo del proyecto. Les decía: “Bueno, es que miren, la lengua tiene que funcionar de tal o cual forma”. Entonces, siempre hacían objetos planos porque necesitaban tener conformada la idea que yo quería transmitirles a las personas, a los ingenieros, que son diseñadores gráficos, pero que no conocen cómo funciona nuestro aparato fonador. Creo que es el momento de trabajar en equipo, de trabajar de una manera multidisciplinaria. No podemos evitarlo. Es algo que ya está aquí y que es una necesidad: trabajar, siempre, en comunidad. La universidad es una, y tenemos que trabajar todos, con ingenieros, con médicos... He tenido que trabajar con neurólogos, con psicólogos, con ingenieros. Es un momento de apertura a lo multidisciplinario.

Bien, entonces, ¿qué son los *ambientes virtuales*? Los *ambientes virtuales* o *entornos* o *mundos* están constituidos por una base de datos gráficos interactivos explorables y visualizables en tiempo real, en forma de imágenes tridimensionales en síntesis, capaces de provocar una sensación de inmersión en dichas imágenes. Se trata de una simulación. Una simulación que nos va a permitir trabajar con todo eso que tenemos

enfrente. Un entorno visual, por lo tanto, es un espacio de síntesis en el que uno tiene la sensación de moverse físicamente con una visión estereoscópica total: una correlación de sensación muscular llamada *propioceptiva*.

Este modelo se generó por el observatorio de visualización IXTLI, el cual, creo que todos los universitarios deberíamos conocer y visitar. Orgullosamente, es la primera sala de esta naturaleza, no solamente en México, sino, también, en Latinoamérica. Considero que vale la pena conocerlo porque, para poder disfrutar de todo esto, necesitamos conocer ese ambiente virtual.

Tenemos tres partes en el proyecto: el modelo, el objeto y la etapa de experimentación. Se ha decidido así debido a la metodología. Todavía es un proceso de construcción, tiene algunas fallas porque algunos documentos se escanearon y cuando se hace esto, cambian palabras, se ponen acentos, etc., pero creo que eso se puede solucionar. Lo importante aquí es ver cómo se ha generado el modelo. Veán ustedes, todo es geometría, es el punto de partida: la geometría. Uno dice: *sí, yo quiero hacer esto*. Sí, pero ¿cómo lo vamos a hacer? Para poder generar un modelo en tercera dimensión, se trabajó con un programa que se llama *MAYA 7*. Se tuvo que generar toda una geometría que tiene características específicas. Cada cuadrito tiene un peso y cada peso implica, también, el que se vea o que no se pueda ver en este observatorio de visualización, porque todo está configurado por una lógica matemática. Es así como se ha ido conformando el modelo. Tuvimos que hacer una investigación científica en la Facultad de Medicina. Investigamos cómo estaba funcionando, tomamos las fotografías, lo dibujamos primero en papel, lo vimos, lo reproducimos en un CD, lo pasamos para que se viera cuál era el movimiento y se empezó con el modelado. Todas estas son las diferentes etapas de este objeto que se llama *Sarahí*, hasta llegar a esta etapa.

En un principio lo teníamos perfecto. Tenía su camiseta de los Pumas y ¡oh, sorpresa!: no corrió. Eso fue porque habíamos puesto 8245 polígonos, era muy pesado. Entonces hubo la necesidad de hacer una reducción de esos polígonos. Ésa fue la primera etapa: el modelado.

De ese modelado se pasa a la programación. Ése es el gran reto: ¿cómo vamos a nombrar cada archivo? Tuvimos que elaborar todo un código que nos fuera familiar a todos los involucrados en el proyecto. Todo esto es programación.

Al final, *Sarahí* tiene una pareja, que es *Renato*. En este caso lo que me interesa es cómo está el corpus, que ése es otro aspecto fundamental que hay que señalar. El corpus consta de 430 archivos que incluyen sonidos, sílabas, palabras, frases y trabalenguas. Este corpus se eligió porque tenemos que notar los diferentes contextos de realización de los sonidos que a mí me interesa rescatar, son los sonidos más problemáticos en todas las lenguas: /r/, /d/, y /t/. Ésos son los problemas fundamentales. Por supuesto, también las vocales y las consonantes.

Bueno, esto es solamente un ejemplo de cómo está organizado el corpus. La ventaja del objeto virtual es que el usuario o los alumnos pueden entrar, si bien deben usar unos lentes para poder penetrar al objeto. Ahora estamos en etapa de experimentación. En esta etapa hemos detectado algunos detalles, como los dientes superiores. Hay que ponerlos en transparencia para que el alumno pueda ubicar perfectamente donde se coloca la lengua. Y nada más poner eso en transparencia implica trabajar otros seis meses porque hay que quitar todo y volverlo a configurar. Pero ya estamos trabajando. Quiero decirles que la experiencia ha sido maravillosa. Los alumnos quedan fascinados, sobre todo porque ya saben donde tienen que colocar la lengua y dónde tienen que articular el sonido. Muchas gracias, esta es *Sarahí* y ojalá tengan la oportunidad de ir a conocerla en el observatorio de visualización IXTLI que está en la Dirección General de Cómputo Académico. Muchas gracias.

César Aguilar: Muchas gracias a la doctora Rosa Esther Delgadillo. Yo creo que ya quedó más o menos claro lo que se puede hacer con esta combinación entre computación y lingüística. Retomo la invitación de ir a visitar a *Sarahí* y, en algún momento, también conocer a su novio. Esa invitación sigue abierta. ■

UNA PÁGINA ELECTRÓNICA PARA LA ENSEÑANZA DE FONÉTICA Y FONOLOGÍA EN LA FFYL DE LA UNAM

JAVIER CUÉTARA PRIEDE/FFYL, UNAM
MARGARITA PALACIOS SIERRA/FFYL, UNAM
JONATHAN MARTÍNEZ PALLARES/FI, UNAM

César Antonio Aguilar: Pasamos a la última presentación de esta mesa. Está a cargo de Jonathan Martínez Pallares, pasante de la carrera de Ingeniería por parte de la Facultad de Ingeniería de la UNAM; y dos viejos conocidos de ustedes, por una parte, la Dra. Margarita Palacios Sierra, de quien voy a hacer algunos comentarios muy breves: tiene una maestría en lingüística por parte de la UNAM, posteriormente hizo el doctorado en la Université de Paris, campus de La Sorbonne. Es una profesora que ha impartido cursos durante muchos años aquí, en la Facultad de Filosofía y Letras, y obviamente tiene una considerable cantidad de publicaciones, además de ser muy querida por todos nosotros. Nos acompaña también el maestro Javier Cuétara Priede, quien cursó la maestría de lingüística hispánica en la UNAM, y ha colaborado con el proyecto DIME que está a cargo del Dr. Luis Pineda y que se está realizando en el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. Tiene intereses en la lingüística computacional y promueve la relación entre lingüística y computación. Sobra señalar que es uno de los organizadores de este evento. Finalmente, la presentación que ellos nos van a dar hoy se llama: *Una página electrónica para la enseñanza de fonética y fonología en la Facultad de Filosofía y Letras de la UNAM*. Si no hay otra cosa que añadir, los dejo con ellos. Adelante.

Margarita Palacios: Lo primero que quiero agradecer es la asistencia que indica persistencia y credibilidad en campos todavía muy pioneros. Yo creo que una de las cosas más difíciles de nuestro momento cultural es aprender a trabajar con la diferencia y éste es, realmente, un espacio interdisciplinario. Me parece maravillosamente bien que se dé. Mil gracias por ello.

Muchos de ustedes, probablemente han tenido ya antecedentes sobre nuestra página. La página nace, realmente, como un material de apoyo, nada más, para el curso de fonética y fonología. ¿Por qué? Porque fonética y fonología tenía un record histórico en el que la gente no pasaba; había conflictos para poderse graduar; después de mucho tiempo la gente seguía arrastrando esta materia y realmente la razón no obedecía a una justificación de complejidad, sino, más que nada, a una sistematicidad en la transcripción; por ello, el maestro Cuétara y yo decidimos hacer este material de apoyo desde hace un tiempo. Antes había realizado una guía para SUA. En fin, había unos antecedentes que, de alguna manera, condensaron en esta primera página. Gracias al trabajo, participación y colaboración, hoy tenemos una página que pronto estará ya completamente renovada. Esta página fue obra casi totalmente, como podrán imaginarse, del maestro Javier Cuétara, más que mía; para mí la máquina era no solamente un reto, sino un golpe brutal a la edad que tengo. Entonces, gracias al apoyo de él, en primer lugar, y en segundo lugar, mucho al reto, pudimos organizar esta primera página casera, que hoy tiene una visión muy diferente, gracias al trabajo en colaboración con el

Instituto de Ingeniería y con la Facultad de Ingeniería. Ésta es, realmente, una valiosa aportación. Considero que los muchachos serán los primeros beneficiados con este trabajo.

Me gustaría decir por qué surge la página. Surge por el valor que tiene la oralidad para la fonética y la fonología. La oralidad es la base de mi relación con el otro. Lo primero que hacemos es emitir sonidos. Lo primero que hacemos es tratar de comunicarnos con el otro; el llanto y la risa son las primeras fuentes de comunicación del ser humano. Luego, la fonética y la fonología, no solamente son la base de la lingüística, sino son, además, la base de la comunicación humana. De ahí la importancia de esta materia dentro del currículum de nuestra carrera.

Después, porque creemos que todos aprenden en la anticipación, esto es: leo y escucho anticipando lo que el otro me va a decir. Esta experiencia la hemos vivido todos. Cuando empezamos a hablar con alguien ya sabemos qué nos va a decir, esto es la anticipación de la comprensión del sentido. Y esto se ejercita, justamente, mientras aprendemos la palabra, que se encuentra en un contexto determinado. Es tan interesante, tan importante esto, que aquí está el principio de la comunicación humana. Cuando un niño empieza a meter dados en una cubeta y dice —uno, dos, tres— ese niño no está contando; para él, ese “tres” es meter dados en la cubeta, porque no está contando, como la madre, que sí cuenta. Al chico le queda tres, eso es lo que él reproduce y ése es el principio de la comunicación humana. Por eso creemos que la fonética y la fonología, empiezan desde allí.

Después, creemos que, efectivamente, la comunicación consiste en una modificación del entorno acústico del oyente realizada por el hablante, como resultado de la cual el oyente concebirá pensamientos semejantes a los del hablante. Cuando yo hablo, espero que la palabra que estoy comunicando sea recibida más o menos de la misma manera por el oyente. La fonética y la fonología, hoy, gracias a los estudios tecnológicos, nos permite medir ese “más o menos”. Señores, esto es verdaderamente histórico en la historia de la humanidad. Es verdaderamente una raya. ¿Por qué? Porque antes sabíamos que la vocal era, más o menos, abierta o cerrada; era más o menos nasal. Cuando alguien me dice: *¡que padre!*, ¿qué me está diciendo, efectivamente? Cuando alguien me dice: *está a todo dar tu vestido o qué padre vestido*, en la segunda ocasión no me está diciendo que el vestido está bonito; me está diciendo: *¡que horror! ¿Cómo te vestiste así?*. Y ¿dónde está esto? No está en el texto. Esto está en el valor de la oralidad, está en el sentido, está en la entonación. De ahí que la página tenga, todavía, un gran reto por delante, que es el trabajo de entonación, sobre el que, en español, se ha hecho muy poco. El verdadero sentido de la comunicación humana está allí.

Bueno, ya me cansé de hablar acerca de hacer nuestro trabajo realidad en una página que se encargará de presentarles el maestro Cuétara, quien ha hecho una labor fenomenal. Ahora dejamos la presentación tanto al maestro Cuétara como a Jonathan. Quisiera destacar la importancia del trabajo interdisciplinario en el sentido de que la formación de Jonathan Martínez es, por supuesto, la de ingeniero. Nuestro puente es Javier Cuétara. La pasamos muy bien Jonathan y yo cuando estamos juntos, pero siempre necesitamos al traductor para entendernos.

Javier Cuétara: Yo dejé la introducción a Margarita y la presentación estaba a cargo de Jonathan. Sólo diré que entre los tres trabajamos muy bien. Es algo que yo he hecho antes; el objetivo es tener la página a disposición de los alumnos del Colegio de Letras Hispánicas, ya muy pronto con una nueva cara.

Jonathan Martínez: Muchas gracias. También ha sido un placer para mí trabajar con ustedes. Me gustaría comenzar por el concepto de lo que son los *sistemas de edición de contenido*.

Generalmente se define un sistema de administración de contenido como aquel sistema que permite la creación y administración de contenidos, principalmente en páginas web, y consiste en una interfaz que controla una o varias bases de datos,

donde se logra el contenido HTML, el cual contiene los detalles del formateo de las tablas, así como las llamadas a la diversidad de funciones.

Cuando se encuentra la función *condem*, los diversos *scripts* que están en el archivo *cms.php* examinan el archivo de contenido *condem* y se construye la tabla de contenidos, así como la página seleccionada, esta construcción de etiquetas se añade a la cadena de la variable *O* anteriormente mencionada a través de expresiones regulares, haciendo que *O* sea una página HTML válida, y pueda ser remediada correctamente por cualquier *browser*.

Dentro de las herramientas electrónicas que utilicé existen básicamente *scripts*, PHP, hojas de estilo y obviamente HTML. Me parece que la ventaja de este sistema es que es bastante sencillo ya que, como les repito, no utilizo bases de datos sino que simplemente a través de una cadena grande y a través de expresiones regulares lo voy acomodando en páginas HTML. Y siempre ha sido mi objetivo, nuestro objetivo, tener en mente a los usuarios, que son humanistas, que no están muy familiarizados con el uso de estas tecnologías, para, básicamente hacerlo lo más fácil posible. A grandes rasgos eso sería todo.

Margarita Palacios: Cuando Jonathan escribe, escribe dos hojas y ya acabó todo, yo necesito doscientas para decir lo que él. Quiero decirles que hasta en ese sentido nuestros lenguajes son diferentes y que esto además es mágico y maravilloso. Cuando uno trabaja en interdisciplinas siempre se aprende. Mi área de trabajo fundamentalmente es el discurso, y he aprendido a trabajar con antropólogos, sociólogos, comunicólogos, por supuesto gente de computación, de ingeniería y esto nos va enseñando o nos va reeducando en el proceso.

Puedo mencionar que se trata de una página un poco aburrida. Al correr del tiempo espero perspectivas muy diferentes, pero incluso a esa página la gente se ha ido acostumbrado a bajar. Es todo un nuevo ejercicio: nosotros estamos acostumbrados a leer de izquierda a derecha y la página no tiene direcciones. *Es horrible que yo tenga que entrar por donde yo quiera*. Esta primera reflexión frente a la página es una posición diferente frente al otro: resulta que nadie me va a decir de dónde a dónde voy, sino que yo tengo que decidir por donde entrar. Perdón, esto es un cambio, no es solamente un cambio físico, y sobre esto quisiera que reflexionemos todos un poco en este coloquio, porque estamos precisamente frente a un cambio de vértice cultural y no solamente ante un trabajo interdisciplinario. No sé si Javier quiera decir algo.

Javier Cuétara: Sí, quiero decir algo que me parece muy importante. Este ejercicio es una herramienta de trabajo; no hay ni una sola materia que tenga una página similar en otra escuela fuera de Letras; incluso en otras universidades o en otras facultades dentro de esta propia universidad. Es algo normal en nuestra facultad, es algo de la vida diaria aquí. Por lo tanto, sí es una aportación, aunque hay algunos intentos por aquí y por allá. Yo mismo tengo otras materias con otras páginas. Creo que somos los únicos, no sé si haya otra y esté equivocado, pero no creo. Ese es otro de los valores que puede tener este ejercicio que a la vista de muchos sería tímido, sencillo, y lo es, pero es bonita y es una herramienta de trabajo.

César Aguilar: Terminamos entonces con este ejercicio, como lo plantearon. Yo sí quiero retomar esa parte con respecto a lo que se puede comparar con otras facultades y sí creo que en la Facultad de Filosofía y Letras sería interesante ese tipo de interacción de recursos electrónicos; y como lo comentaba Javier, creo

que los podemos contar con los dedos de la mano y hasta nos sobran. Entonces, retomemos la idea del coloquio, sobre todo el énfasis a colaborar en una u otra parte. Creo que llega el momento de las preguntas. Continuemos con el diálogo, invito a los ponentes a pasar al frente. Los dejo a su disposición para lo que quieran comentar, preguntar, señalar, bienvenido, yo creo que los ponentes en ese sentido no tendrán ningún reparo, así que por favor.

SECCIÓN DE PREGUNTAS

Pregunta 1: Me gustaría saber qué vislumbran ustedes como posibles dinámicas de cooperación entre computólogos, ingenieros o personas que trabajen con computadoras y gente de humanidades: lingüistas, letristas.

Margarita Palacios: Personalmente quisiera decirte algo: la página ha sido muy útil, y será útil. Ustedes nos hicieron el favor de ayudarnos con dos tesis, la primera fue un análisis de distribución semántica, con el programa de distribución semántica, en el que logramos a través de un trabajo de investigación saber qué valor tenía para la familia, para niños, adolescentes, en escuelas públicas y escuelas privadas, y los resultados fueron interesantes. Vale la pena mencionarlo. El COLMEX aportó en una canción de Barney; puede ser muy estúpido lo que estoy diciendo, pero es tan importante en la formación de niños. Probablemente ustedes la han oído: “yo soy feliz con mi familia así” y luego el niño, o el que canta empieza “mi familia tiene mamá, papá, dos hermanos, un perrito”, una familia tradicional “yo soy feliz con mi familia así”, después dice “yo tengo una amiga que tiene a su mamá y a papá, pero su papá no vive ahí, yo soy feliz con mi familia así, yo tengo dos papás, yo soy feliz con mi familia así” y empieza a pasar una pluriconstitución de la nueva familia. Esto estuvo sacado de los principios de esa tesis, de la que, sin el programa de computación, no se hubiera podido hacer. Y el segundo ejemplo es el de una persona que hizo una tesis sobre el estado de Oaxaca, de un pueblito muy centrado, donde hicimos un análisis de investigación, justamente para tratar de ver cuál es la actividad dominante de este pueblito. Efectivamente, no era la tierra, ni la cerámica ni ningún trabajo de tipo manual; lo que trabajaba toda la comunidad era la preparación de dos grandes fiestas, todo lo que apareció fueron palabras referentes a fiestas, tamales, comida, porque el pueblo es básicamente de inmigrantes. Entonces, cuando llega el dinero que mandan todos para la fiesta, todo mundo trabaja para eso; quiero decir que aquí hay un trabajo para poder aplicar en los corpus los programas de cómputo y poder identificar campos semánticos, desplazamientos, cambios de la lengua, no solamente semánticos, sino cambios estructurales de la lengua. Hace un momento el ingeniero Gerardo Sierra, lingüista, nos preguntaba ¿qué está pasando con la lengua?, ¿hacia dónde vamos?, ¿cómo se está modificando? Concretamente, en el terreno de la fonética hay todo un trabajo titánico por hacer en el campo de la información. Por eso, al término de este coloquio tendremos el curso del profesor Butragueño sobre entonación, que es el área más renombrada y más ignorada de la producción oral de la lengua hablada, y es dónde está el sentido. Así que yo veo no un campo, veo el futuro de la lingüística en trabajos interdisciplinarios.

Gerardo Sierra: Como alguna vez tú me hiciste reflexionar, estamos hablando de dos disciplinas: la lingüística y la computación, dos disciplinas que en algún momento se juntaban y se hacen trabajos multidisciplinarios. Sin embargo, me mostraste que estamos hablando de dos transdisciplinas: la lingüística como tal es una transdisciplina porque, realmente, en cualquier disciplina estamos hablando de lenguaje y necesitamos el lenguaje para comunicarnos; para informar usamos el lenguaje. Por otro lado, está la computación, otra gran transdisciplina que se utiliza para cualquier área, para biología, para química, geografía, para lingüística, para la computación misma. Entonces, ¿qué hacer cuando tenemos la congruencia de estas dos grandes transdisciplinas para formar a su vez una interdisciplina? Creo, por lo tanto, que es importante esta vinculación entre la lingüística y la

computación, y lo estamos viendo en esta primera mesa de recursos electrónicos para la enseñanza. Y es nada más una de las partes. Finalmente, ¿de qué manera puede participar activamente la lingüística con la ingeniería? Creo que van a tener que ir juntas, de la mano de aquí para adelante.

Rosa Esther Delgadillo: Yo solamente quiero subrayar que hay un cambio de paradigma, ya los esquemas de hace 20 o 30 años son obsoletos, inclusive la misma literatura se tiene que ver ya con otros ojos. Esta realidad virtual de la que yo les he hablado puede desarrollar programas en donde podemos ver las diferentes fases narrativas que nos permiten darle sentido a la obra. Ahí lo tenemos y estamos inmersos en ese mundo: la Internet, los celulares, las *palm*s, todos estos objetos, ¿qué usan? Usan códigos y muchos de ellos son códigos lingüísticos. Estaban solicitando, de una universidad en Barcelona, el trabajo de un fonetista para hacer comparaciones de diferentes variedades del español, y la plaza quedó vacía porque ahora sí tenían al fonetista pero no tenían el complemento. Entonces, creo que hay un trabajo extraordinario que se puede hacer aquí, en esta universidad, de la cual, insisto, debemos sentirnos orgullosos: no es la Facultad de Filosofía y Letras, no es el Centro de Enseñanza para Extranjeros, no es la Facultad y el Instituto de Ingeniería, es la Universidad Nacional Autónoma de México, con todo lo que implica su desarrollo hacia fuera. Por ello es una de las universidades más importantes ahora de América Latina. Creo que debemos sentirnos orgullosos y empezar a cambiar esos paradigmas, esas formas de ver el mundo, esa forma de interactuar también con los otros. Gracias.

Pregunta 2: Quiero felicitarlos; creo que es un trabajo extraordinario lo que han estado haciendo; me encanta sobre todo este trabajo interdisciplinario en el que combinan la medicina con la computación, con la lingüística, con lo que sea, y muy particularmente quiero felicitar al ingeniero. La verdad es que más o menos entiendo el trabajo que llevan y, como mencionaba la maestra Palacios, son lenguajes distintos. El hecho de programar una página es un gran problema, y que un lingüista, un humanista lo entienda es una tarea difícil. Me parece que vale la pena destacar el trabajo de un ingeniero en el desarrollo de la página, todo el trabajo previo a la proyección en la red. Los demás trabajos son magníficos, la verdad es que estoy muy contenta de que se desarrollen, y creo que esto es apenas un comienzo; considero que aún hay mucho por hacer en este campo y, en la medida de lo posible, vamos a tratar de apoyarlos, al menos cuenten conmigo, muchas felicidades y gracias.

Pregunta 3: Yo quería comentar sobre la efectividad de las páginas. Tomé el curso de introducción a la lingüística con el profesor Javier y en el segundo semestre se hizo una catarsis entre los alumnos porque ya no teníamos que entregar la tarea en hojas. Antes nos pedía definiciones, nos pedía corpus; entonces llegábamos con los papeles, entregábamos y de repente: *no encontré tal palabra, pásamela*. La entregábamos, y de repente con todo el aviso de un *coco wash* que nos iba a hacer. El profesor nos dijo: “va a haber formas distintas de trabajar en este semestre, va a haber un foro, o podemos hacer esto y esto”, y nos pintó tan hermoso el foro y nos avisó que nos iba a obligar a hacerlo, que participamos en ello. Era complicado, porque no estamos acostumbrados a esas cosas, somos humanistas, la computadora se usa para redactar y ya. Incluso cuando empezó recién la página, hubo problemas con el acceso, por lo que, a veces, llegábamos sin hacer la tarea. Empezó, entonces, a hacer un foro en donde teníamos que leer la tarea de otras personas para hacer

la nuestra, fue retroalimentarnos también, porque cuando entregábamos un trabajo no era para él. Cuando entregábamos la tarea era para el grupo. Realmente el foro funciona, porque no sólo participas con el profesor, ves la tarea del de al lado para mejorarla o complementarla y eso realmente te hace aprender.

Al final, también hubo otro “coco wash” en el trabajo final: nos dio seis textos de diez personas diferentes, todos eran de un tema en común, un tema abierto. Entonces, se cayó en el error de no leer los demás trabajos; recuerdo que llegó y dijo “no están leyendo los trabajos, me estoy dando cuenta porque están repitiendo la información”. Es más sencillo trabajar así porque, a partir de ver que diez se equivocan, yo por los errores que llevan ellos. Puedo decir que realmente funcionan esas páginas, y es muy bueno que existan ya en la facultad, puesto que a veces nos aislamos, decimos: “somos humanistas”, “con el libro ya estoy aprendiendo”. Este es un escalón más que tenemos que subir para trabajar realmente, y, sí, las páginas funcionan para el estudio de lingüística.

Javier Cuétara: Te agradezco que lo comentaras, te agradezco que me recordaras esa experiencia del 2006. Fue un gran trabajo para todos y todos estuvimos obligados a trabajar y a disfrutar. Hay algo que quiero mencionar y que es muy importante: la modalidad. Sí, hay una página electrónica, es un recurso y apoyo; en ese caso, en el curso de introducción a la lingüística, yo tenía que consensar si ésa era la manera de evaluar. El grupo la eligió y fue una muy buena experiencia.

Margarita Palacios: Quisiera hacer una pequeña aclaración: cuando hablamos de lingüística, en análisis del discurso, nosotros hemos tratado de romper la barrera entre textos orales y textos escritos porque el *talk show* no es un texto oral, los *talk shows* y los *reality shows* y todo esto no son textos orales, son textos escritos, dichos oralmente. Por ejemplo, muchas veces una conferencia simplemente es un texto escrito dicho oralmente. De alguna manera el profesor Sierra, cuando hizo su exposición, jugó con las dos formas. Entonces, no hay un límite determinado entre lengua oral y lengua escrita, toda lengua oral tiene estructuras escritas y que toda lengua escrita tiene formas de oralidad. Basta recordar cuando a Rulfo le preguntaron si él hablaba como escribía, él dijo *no, yo escribo como hablo* y éste es el ejemplo más fehaciente de que no hay límites determinados. Sin embargo, para el trabajo sistematizado, hemos tomado lengua escrita como aquella que está producida como grafía, lengua oral como aquella que es un *dictum*, es decir, por transmisión. Valdría la pena la aclaración: de alguna manera, no podemos tomar el corpus a menos que un investigador quiera abarcar lo que es oral desde este otro punto de vista del discurso, o esté de acuerdo en tomar lo grafo como grafo. Es más, por qué no se realiza un mini coloquio o alguna mesa para ver cómo lo vamos a trabajar, porque ahí hay un problema de disciplina. ■

COMPARACIÓN DIACRÓNICA DE PERFILES MORFOLÓGICOS: DISTANCIAS ENTRE LOS DOCUMENTOS DEL CHEM Y LOS DEL CEMC

ALFONSO MEDINA URREA
GIL-IINGEN, UNAM

Alfonso Medina: Buenas tardes, primero que nada voy a esbozar los objetivos. ¿Qué es lo que quiero hacer con esta ponencia? Sencillamente se trata de proponer un esquema para medir qué tanto ha cambiado el español en el nivel morfológico desde el siglo XVI, y el objetivo —ya que estamos en un coloquio de lingüística computacional, si bien no estamos tratando la lingüística computacional en su sentido estricto de formulismos y de inteligencia artificial, sino más bien en el sentido del área más general del lenguaje en computación de investigación lingüística apoyada con computadoras— es ver cómo esto es posible gracias a los desarrollos que hemos hecho en tecnologías del lenguaje e ingeniería lingüística, etcétera.

Para esto primero quiero presentar los corpus que utilizo en el análisis cuantitativo y, en especial, el *Corpus Histórico del Español de México* (CHEM), ya que las dos ponencias que vienen están relacionadas con éste: primero la de Carlos Fonseca y luego la de Teresita Adriana, que van a hablar sobre asuntos de fonología y fonética. Después voy a hablar de algunos antecedentes del experimento que llevamos a cabo para esta ponencia, mostraré algunos catálogos de afijos extraídos para cada ciclo y compararé estos perfiles morfológicos. Después propondré una manera de medir éstos, las diferencias entre estos perfiles.

Empezaré hablando del CHEM (el *Corpus Histórico del Español en México*). Es un proyecto de DGAPA que está en el tercer año; nos ha servido para tener una página de Internet donde ya se pueden hacer búsquedas muy sencillas, pero muy valiosas, a partir de los pocos documentos que hemos reunido y que esperamos crezcan. El CHEM fue posible, en primer lugar, gracias a experiencias previas que teníamos en el Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería (II), concretamente, el desarrollo de un Corpus Lingüístico en Ingeniería (CLI), para el cual también existe una humilde página web, y del cual fue responsable el Dr. Gerardo Sierra. Esta experiencia sirvió para construir el corpus.

Además, la disponibilidad de costosos materiales de filólogos y lingüistas. En el caso del corpus del CHEM son los documentos que Concepción Company, del Instituto de Investigaciones Filológicas, proporcionó para que pudiéramos incorporar al CHEM como primer conjunto de corpus.

El objetivo general fue construir un corpus diacrónico que pueda ser utilizado en Internet. Los documentos típicos que están aquí son de la Nueva España y del México independiente. Las metas específicas, primero que nada para garantizar la representatividad del corpus, es que aunque no tengamos los documentos que todavía están por incluirse allí, tenemos que saber qué hay en esos siglos que deba después incluirse, de tal manera que si hay algo prominente en ese siglo pueda incorporarse después, y hay que tomar en cuenta los géneros literarios, los géneros temáticos, los tipos de textos, no olvidar a los autores prominentes, etcétera.

Otra meta específica es salir y buscar, pedir a los filólogos para que nos presten su trabajo y lo podamos incluir allí, sobre todo si reúnen ciertos criterios de transcripción para que más o menos tengamos documentos unificados en cuanto a formatos, etc. Finalmente, como meta, diseñar y desarrollar herramientas específicas para explorar y analizar el corpus.

El generador de concordancias ya existe; en la página de Internet puede entrar cualquiera de ustedes, hacer exploraciones y ver las concordancias de palabras y búsquedas específicas que quieran hacer. También, y relacionado con el experimento que explicaré después, pondremos a disposición segmentadores morfológicos y lematizadores para el corpus, que serán de utilidad en otros corpus que sirven también para desarrollar analizadores y etiquetadores morfosintácticos y para desarrollos como transductores ortográficos fonológicos que puedan servir como datos para sintetizadores. De eso se tratan las dos próximas ponencias.

En pocas palabras, uno de nuestros criterios principales es poder lograr una representatividad, aunque sea relativa, buscando variedad y equilibrio de géneros, documentos y registros; los primeros fueron los de Concepción Company, pero estamos procesando otros documentos de los siglos XVI, XVIII y XX que serán incorporados próximamente y que enriquecerán y ayudarán a aumentar la representatividad del corpus en cuanto a los siglos en cuestión.

Por otra parte, el otro corpus que utilicé para este experimento es el compilado en El Colegio de México para obtener la base del *Diccionario Español Usual en México*. Es una muestra de cerca de mil documentos y transcripciones del siglo XX y desde un principio fue diseñado para que hubiera una variedad y equilibrio de géneros, que es exactamente lo que queremos.

Pasando al del experimento, diré que entre los antecedentes de éste podríamos mencionar la glotocronología de Maurice Swadesh, que estuvo varios años en México y que, en pocas palabras, diseñó un método inspirado en el fechado de carbono del material orgánico. Presumía un índice de cambio lingüístico por milenio fijo, que fue distinto en otras versiones de la glotocronología, debido a que es obvio que no hay un índice de cambio lingüístico fijo.

Su método se basaba en examinar la base léxica de entre cien y doscientas palabras de las lenguas que comparaba; esta base léxica consistía en aquellas palabras dedicadas a las partes del cuerpo, cuerpos celestes, uno y dos, los pronombres personales, etc. La idea era comparar los porcentajes de cognados que se compartían en estas lenguas, los cognados son las palabras que tienen un mismo origen. La idea es que, mientras mayor sea ese porcentaje de cognados, más reciente en el tiempo se presume que se supera su separación como lenguas independientes.

Abarcaré un poco más sobre los cognados. Como decía, son palabras con un origen común. Encontramos cognados entre lenguas como *posible* y *possible* los dos tienen un origen común, uno de español y uno de inglés. *Starved* y *Starben* de inglés y alemán, tienen el mismo origen aunque diferentes significados, *starved* es 'morirse de hambre' y *starben* es 'morir', o palabras como *exquisito* que en español significa 'refinado' o 'muy bueno' mientras que en portugués es 'algo muy extraño e incluso desagradable'. También puede haber cognados en una misma lengua, tenemos formas con un mismo origen, como *delicado* y *delgado*, o en inglés *shirt* y *skirt*.

Podemos hablar de cognados morfológicos. Vemos que el español y el portugués comparten varios clíticos pronominales que se adhieren a los verbos y se puede decir que tienen un mismo origen en latín. También, podemos ver que la segunda persona del plural es un morfema que parece ser utilizado únicamente en discursos muy serios o paródicos, porque los hablantes comunes del español en México no lo utilizamos sino como una vaga segunda persona, mientras que en España sigue viva la idea de una segunda persona muy productiva y con el estatus de flexión.

Hablaré ahora específicamente de los perfiles morfológicos y propongo que se trate de conjuntos de personas prominentes de una lengua específica. Cómo determinar cuáles son los conjuntos prominentes es el método del que hablaré brevemente, pero mientras, diré que los antecedentes del método tienen que ver con Zellig Harris, con Claude E. Shannon y con Josse de

Kock, que trabajaron con métodos para segmentar palabras y determinar cuáles son los morfemas de cada una en diversos idiomas y, principalmente, en lengua escrita.

Zellig Harris utilizó diversas lenguas del mundo, Josse de Kock el francés y el español, y a Shannon se le conoce por su trabajo en teoría de la información; él no es lingüista, pero tuvo influencia de la lingüística de los años 50 antes de llegar a Chomsky.

La idea del manejo computacional de los corpus electrónicos nos permite entonces definir catálogos de afijos y grupos afijales; eso hemos hecho previamente en el grupo aplicándose a corpus del chuj (una lengua maya), del tarahumara, del checo; lo hemos y lo vamos a seguir aplicando a diversas lenguas para ver qué tan universal, entre comillas, puede ser esto.

La idea es medir, para cada grupo, afijos o grupos afijales, lo que puede caracterizarse como su *afijalidad*. Hablando del experimento en cuestión, diré que tomé tres estados de lengua de los siglos XVI y XVIII y para eso utilicé el *Corpus Histórico del Español en México* (CHEM) con documentos de Concepción Company, pero otros también, (todavía no están en red), y para el siglo XX el *Corpus del Español Mexicano Contemporáneo* (CEMC).

Hicimos una extracción automática de los sufijos y grupos sufijales. En muy pocas palabras, el método tiene que ver con cálculo de entropías de Shannon, con medidas estructurales de economía y de cuadros. Lo importante es que se generan catálogos de afijos que se pueden ordenar por qué tan afijales son, y en esencia corresponden a qué tan prominentes son en ese corpus. Lo importante es que se generan automáticamente para cada estado de lengua y tomamos los cognados morfológicos de estos tres estados comunes a los tres siglos para compararlos entre sí. Fueron 282.

En el cuadro 1 se encuentra el catálogo que mencioné, primero generado para el siglo XVI. En la segunda columna tenemos los afijos o grupos sufijales, luego tenemos toda una serie de estadísticas y al final una columna que sencillamente es un índice de qué tan afijal es ese afijo o grupo de afijos en determinado corpus. De manera similar, para el siglo XVIII tenemos una tabla general. No necesariamente tienen el mismo orden lo afijos, ya que va a variar de corpus a corpus; no necesariamente tienen la misma fijalidad de un siglo a otro. Finalmente para el siglo XX se presenta la misma idea.

	CHEM XVI	CHEM XVIII	CEMC XX
-o	0.8040	0.8127	0.8222
-ito	0.5108	0.5496	0.6093
-ería	0.3662	0.4776	0.5364
-eis	0.6117	0.2747	0.0000

Cuadro 1. Distancias en afijalidad

Ya para acercarme a las distancias en la afijalidad para comparar los siglos, se promediaron los valores absolutos de las diferencias en la afijalidad de cada cognado morfológico. En este cuadro se presentan cuatro ejemplos, donde tenemos el sufijo, un sufijo polisémico, en los tres siglos, tiene un valor de afijalidad entre 0 y 1, muy afijal; es decir, 0.8, y parece ser muy estable.

-ito, por ejemplo, que suele ser de interés porque es una derivación, pero es tan frecuente que parece flexión, también tiene algo estable. Sin embargo, en el siglo xx parece subir un poco. Y similarmente -eis, por ejemplo, en el siglo xx, no surge; no es que no ocurra la forma, estadísticamente no puede contar como un morfema del español y al examinar ese corpus se ve que en esos contextos, este morfema ocurre.

Para explicar cómo calculé las diferencias, sencillamente hice una resta y saqué el valor absoluto de estos valores entre los siglos, de tal manera que resté cada valor de afijalidad para conseguir una diferencia; posteriormente, la diferencia para cada afixo fue promediada, y así se obtiene el promedio de diferencias de afixos en cuanto a su afijalidad.

En el cuadro 2 tenemos resumida la distancia morfológica entre los siglos, tenemos una fracción entre 0 y 1 y vemos que más bien son números pequeños. Eso quiere decir que habrá que afinar esto porque, ¿qué tan pequeño es pequeño? Lo importante es que se trata de la misma lengua, después de todo y, si consideramos que el 99% de los genes que tenemos los compartimos con los chimpancés, que son otra especie. Entonces también entre lenguas la diferencia debería ser mínima. Si aquí comparo la relación de un siglo con el otro, vemos que el del siglo xvi al xviii tenemos un 0.5, mientras que la distancia del siglo xvi al xx tenemos un 0.06. Por otra parte del siglo xviii al xx tenemos la distancia menor, lo que quiere decir que entre estos siglos, los más similares serían el xviii y el xx, y si ha habido un quiebre entre los dialectos que ha habido entre estos siglos, tendría que haber sido en algún momento antes del xviii donde podríamos decir que ha nacido el español en México. Lo que corrobora algunas intuiciones de filólogos que han dicho que el xviii está más cercano al xix.

	CHEM XVI	CHEM XVIII	CEMC XX
CHEM XVI	0.000000	0.054705	0.061739
CHEM XVIII	0.054705	0.000000	0.046545
CEMC XX	0.061739	0.046545	0.000000

Cuadro 2. Distancia morfológica entre siglos

En términos de porcentaje de similitud morfológica podemos decir que, utilizando la misma tabla, los más similares, de nuevo en un 95% serían el siglo xx con el siglo xviii, mientras los menos similares serían el xx con el xvi, mientras que el xviii se aproxima más al xx.

Conclusiones: Acabo de mostrar un esquema para medir cuánto ha cambiado el español en México, en el nivel morfológico desde el siglo xvi. Se ve que quedaron más asociados el xviii con el xx. Se generaron datos cuantitativos puros sobre cambios en un nivel de lenguaje; es decir, estamos más allá de las intuiciones, no se trata simplemente de conjeturas.

De igual forma, falta mucho por hacer, definir *qué tanto es tantito*; estas diferencias tan pequeñas son importantes: o son que la lengua todavía sea lo mismo, o podría uno decir que hay que desarrollar métodos para decir qué tan significativa es, estadísticamente, significativa entre los siglos, y hay que tomar en cuenta otras unidades de la lengua, como los clíticos pronominales, otros modificadores muy frecuentes, que también a ellos se les puede calcular medidas de entropía y de economía, que pueden servir para determinar su *cliticidad*, por ejemplo, y eso puede compararse entre los siglos. Sería interesante también aplicar esto entre los dialectos del español, no solamente los geográficos. Lo obvio sería comparar el español de otros países en América con el de México, con el de España; pero también dialectos sociales, podría verse una diferencia interesante y para

ver hacia afuera la relación con otras lenguas, como el portugués, que se parece tanto al español y donde seguramente hay muchos cognados morfológicos. Eso es todo.

SECCIÓN DE PREGUNTAS

Pregunta 1: Supongo que se está manejando algún umbral para poder decir cuáles sí y cuáles no y tal vez eso sea por la cantidad de información. ¿Cuántos se obtuvieron en cada siglo?

Alfonso Medina: Bueno se tomó la mejor segmentación de todas las palabras de cada corpus y eso dio un número finito de formas para cada siglo. De esas formas sólo 282 fueron compartidas en todos los siglos y esas fueron las que se tomaron. ■

DESARROLLO DE UN TRANSCRIPTOR FONÉTICO-FONOLÓGICO PARA CORPUS TEXTUALES DIACRÓNICOS

CARLOS FONSECA
FI / GIL-IINGEN, UNAM

Carlos Fonseca: Yo formo parte del Grupo de Ingeniería Lingüística (GIL). Este grupo está en el Instituto de Ingeniería (II); mi trabajo está pensado para el CHEM, el *Corpus Histórico del Español en México*, y se trata de un transcriptor fonético-fonológico para corpus textuales diacrónicos. A pesar de que el desarrollo es para este proyecto, se tiene pensado que pueda utilizarse en algún otro corpus de esta génesis.

Como vamos a ver a lo largo de la presentación, hay que, de alguna manera, pegarle a nuestro trabajo unos módulos o unos segmentos de programación adicionales para que podamos obtener lo que queremos.

Primero, vamos a tener una pequeña introducción, mencionaremos algunos antecedentes, trabajo previo a este desarrollo, vamos a tocar, obviamente, el tema del corpus, pero centrándome en la estructura de los archivos del corpus. Además de esto, y lo que es fundamental para este desarrollo, hablaremos del procesamiento de estos archivos, para poder hacer una descripción general de nuestro transcriptor fonético-fonológico.

Se está trabajando conjuntamente, en este caso, los ingenieros en computación con lingüistas. ¿En dónde está ahora nuestro trabajo y qué es lo que nos falta?

Primero que nada, quiero mencionar que es una labor de dos áreas, dos grupos: uno, el desarrollo de sistemas computacionales, el otro es la fonética y la fonología. ¿Qué sucede cuando ambos se conjuntan? Se dan muchos desarrollos, muchos programas de cómputo, varios de ellos incluso se comercializan. Podemos mencionar, por ejemplo, los sintetizadores de voz, esos programas que hacen que la computadora hable; de igual forma, podemos hablar sobre los que sirven para identificar la voz, programas de dictado, etcétera.

El desarrollo que nos interesa es el de la transcripción automática. La transcripción automática se utiliza como una herramienta para llegar a estas aplicaciones que ya mencioné, dado que para que la computadora pueda hablar, necesita primero algunos símbolos especiales en el texto y con base en esos símbolos es que se puede generar la señal sonora.

Como antecedentes directos, encontramos el trabajo del maestro Javier Cuétara: el llamado *TranscríbEMex*, desarrollado por Cuétara y Villaseñor en el 2004. El *Transcriptor Fonético Automático para el Español de México* es su nombre completo, cuya referencia principal se encuentra en el trabajo de Cuétara del 2004: *Fonética de la ciudad de México*, su tesis de maestría. *TranscríbEMex* está desarrollado en un lenguaje de programación llamado *Perl*, y está basado en módulos; además, utiliza como alfabeto fonético *Mexbet*, también un estudio del maestro Cuétara.

Pero ¿a qué nos referimos con los módulos? Si nosotros vemos *TranscríbEMex* en nuestra computadora, notamos una serie de archivos, que tienen incluso un ícono que puede parecer extraño; esto es lo que conforma *TranscríbEMex*. Esta serie de archivos son módulos, como ya lo mencioné, en otras palabras, son un conjunto de instrucciones que hacen una cosa en particular. Por eso no pudimos utilizar *TranscríbEMex* para aplicarlo al CHEM.

¿Por qué? Porque las reglas de transcripción ya vienen dentro de este código. Si nosotros abriéramos, por ejemplo, el módulo que transcribe de grafías a fonemas, vamos a ver que allí tenemos una serie de condicionantes que son propiamente las reglas

de transcripción. Más adelante, en la descripción de este desarrollo, vamos a ver que nosotros tomamos de diferente manera, aplicamos de diferente forma estas reglas. Lo que hace *TranscribEMex* es que tiene un archivo de entrada y crea varias salidas, de hecho son más de tres, pero las que nos interesan a nosotros son las de sílabas. Genera una segmentación silábica, la transcripción fonética y la transcripción fonológica. Basta mencionar que *TranscribEMex* fue diseñado para aplicarse en otro proyecto totalmente diferente.

Esa también es otra condicionante para aplicar o no el programa al CHEM. En la pantalla gráfica que nos presenta, encontramos la entrada de texto que escribimos, le damos *clíc* en *transforma* o damos *enter*, y obtenemos las transcripciones y además la segmentación silábica.

La estructura de los archivos del CHEM es la siguiente: como ya el Dr. Medina mencionó, tenemos varios siglos. Imaginemos que nuestros archivos están guardados en carpetas, cada carpeta contiene archivos de un solo siglo, ¿y qué es lo que contiene cada carpeta? Cada una contiene una serie de archivos, si se fijan el ícono es un poco extraño, tal vez habrán pensado que sean archivos .txt; no lo son, vamos a ver por qué. Cada archivo, como vemos aquí, es un documento en XML.

XML es un lenguaje de etiquetado; por sus siglas significa *Lenguaje de etiquetado extensible*, es algo similar a HTML o el código en el que están hechas las páginas de Internet (hay una discusión sobre qué es más poderoso, si HTML o XML). Entonces, para cada documento XML, si lo abrimos, lo que vamos a ver es una serie de lo que llamamos *etiquetas*. No voy a explicar la estructura, lo que es importante resaltar es que toda nuestra información va a estar entre dos marcas, una de inicio y una de final. Lo que nosotros tenemos que hacer para poder procesar los archivos del CHEM es lo siguiente: tenemos un primer programa de pre-proceso, nuestro proceso central que es en este caso el transcriptor y un post-proceso. ¿Qué es lo que hace el pre-proceso? Primero vamos a buscar la etiqueta que identifica el año, que es <YEAR>, y vamos a tomar la información que está entre estas dos etiquetas. Con esta información, lo que hacemos es identificar de qué siglo es el documento; una vez que se identificó, vamos a extraer tres archivos: el archivo *excepciones*, el archivo de *reglas fonéticas* y el archivo de *reglas fonológicas*.

Aquí está la diferencia con *TranscribEMex*: nosotros no tenemos las reglas dentro del código del programa, lo que estamos haciendo es que tenemos, por así decirlo, seis pares de reglas para cada carpeta, y, dependiendo de a qué siglo pertenezca nuestro documento XML, extraemos o abrimos esos archivo de reglas. El archivo de excepciones es mucho más grande que el de las reglas, porque allí estamos metiendo todas aquellas grafías o palabras completas a las que no se puede aplicar directamente una de las reglas de los archivos o que presentan alguna ambigüedad.

Una vez que tomamos nuestros tres archivos, lo siguiente es identificar cada una de las cadenas de entrada. Por ejemplo, después de una etiqueta <CUERPO_DOC> vienen nuestras cadenas de entrada. Lo que hace el pre-proceso es que, para cadena de entrada, llama al transcriptor y realiza esta transcripción al post-proceso. Si nosotros corriéramos directamente *TranscribEMex* sobre este archivo XML, lo que haría sería transcribir desde la primer grafía que tenemos que es el pico paréntesis de nuestra estructura XML. Evidentemente, eso no lo queremos transcribir.

Ya que tenemos la cadena de entrada, nuestros dos archivos de reglas y las excepciones, las pasamos al proceso, a nuestro transcriptor, allí hay ciertas transformaciones. Lo que va a sacar son cadenas de fonemas. Estas cadenas de fonemas se van a pasar al post-proceso, y lo que el post-proceso va a hacer es insertar en el documento una cadena con una etiqueta de XML que explique que se trata de la transcripción. Se trata

de la misma cadena de grafías, pero ahora enseguida de esa cadena de grafías tenemos la cadena de fonemas. ¿Esto para qué nos va a servir? Una de las cosas es que esta transcripción, en un futuro, va a servir para alimentar un sintetizador de voz y podremos tener un acercamiento de, por ejemplo, cómo se hablaba en estos siglos.

Estoy haciendo el desarrollo del *Transcriptor Fonético-Fonológico para Corpus Digitales*, éste es el título de tesis de licenciatura en la que estoy trabajando para obtener el grado de ingeniero en computación en la Facultad de Ingeniería. Lo que estamos nosotros tomando es un conocimiento que se llama *ingeniería de software*, el cual nos dicta una serie de pasos para desarrollar programas. Yo soy de la idea de que cualquier persona puede programar, independientemente de que si es ingeniero o lingüista, o la carrera que sea, la diferencia está en cómo programar. Muchos de nosotros programamos de manera empírica, es decir, nos sentamos como tales con un libro al lado y empezamos a hacer nuestros algoritmos y programar, si acaso escribimos antes nuestros algoritmos, lo que sea. Entonces lo que trata de hacer el ingeniero en software es crear pasos, una metodología para desarrollar un programa, desde una idea básica hasta el código y el programa funcional.

Lo primero que hacemos es definir nuestro problema, nuestros alcances, qué cosas del problema sí se pueden solucionar y qué no, después analizamos nuestros requisitos, qué es lo que nos está pidiendo el cliente, cuáles de sus peticiones sí podemos satisfacer con nuestro desarrollo. Después con base en este análisis, en la definición del problema diseñamos el sistema a través de diagramas; son simple y sencillamente diagramas. Después de esto, con base en el diseño, nos vamos a programar, nos sentamos con nuestro buen libro y, ya que tenemos nuestro programa hecho, lo probamos.

Aquí quiero mencionar que oficialmente ya llevo un año haciendo la tesis, y la razón es que aunque esté en ingeniería de software, tenemos estos requisitos de desarrollo, pero lo que hemos observado en la literatura y en la práctica es que realmente el desarrollo sigue descansando en la experiencia del programador y del desarrollador, en esta nube abstracta de conocimiento, de abstraer el problema hasta obtener el programa, lo que hemos estado haciendo es una metodología de desarrollo de software en donde nosotros, justamente, primero analizamos el problema sin pensar en qué lenguaje lo vamos a programar. Posteriormente, abordamos la cuestión de los diagramas, vamos a describir nuestro programa ya de una manera más formal, más computacional, después vamos a programar en sí nuestro sistema y terminar haciendo las pruebas.

Hacemos un trabajo conjunto, ¿con quién? Tomando otra vez nuestros dos conjuntos desde el principio, lo que hacemos es que, para obtener nuestro sistema transcriptor, quien se dedica a fonética y fonología hace un análisis de corpus, genera las reglas que forman parte de este análisis y diseño. Nosotros las formalizamos, programamos, hacemos pruebas, pero también quien nos dio las reglas nos va a ayudar a estas pruebas.

Desgraciadamente se me ha terminado el tiempo, por lo que no explicaré de qué manera interactuarían para sacar las transcripciones; sin embargo, puedo decir que el módulo verificado lo que hace primero es verificar si nuestra cadena de entrada está dentro de las excepciones, de nuestro archivo de excepciones, y si no está allí, entonces arranca con las transcripciones. En nuestro trabajo futuro, lo que queremos hacer es obtener el módulo de transcripción fonética, pero además de tenerlo, que esté disponible para los alfabetos AFI y RFE, y para esto necesitamos una interfaz para visualizar la información. Esto sería todo. ■

PROPUESTA PARA UN TRANSCRIPTOR AUTOMÁTICO DEL ESPAÑOL DEL SIGLO XVI PARA EL CORPUS HISTÓRICO DEL ESPAÑOL DE MÉXICO

TERESITA ADRIANA REYES CAREAGA
FFYL / GIL-IINGEN, UNAM

Raúl del Moral: Por lo pronto, presentaría yo a Teresita Adriana Reyes Careaga, alumna de la Facultad de Filosofía y Letras, quien nos presenta *Una propuesta para un transcriptor automático del español del siglo xvi para el Corpus Histórico del Español de México*.

Teresita Reyes: Buenas tardes. En el programa mi presentación se llama *Propuesta para un transcriptor automático*. Yo no voy a hacer el transcriptor, sino Carlos Fonseca; lo que yo haría sería precisamente darle las reglas fonéticas y fonológicas para que se desarrolle el transcriptor; así, mi presentación se llama *Propuesta fonético-fonológica para un transcriptor automático del español del siglo xvi para el CHEM*. Este trabajo será la investigación de mi tesis de licenciatura. El objetivo es obtener las correlaciones grafía-fonema del español de México del siglo xvi para la implementación en este transcriptor automático; se logrará mediante el estudio de los documentos del siglo xvi del CHEM, se concluirá con una serie de reglas fonético-fonológicas que sean lo suficientemente amplias para transcribir cualquier texto del siglo xvi que se incorpore al CHEM.

Algo que me gustaría resaltar es que ésta es una investigación de alguna manera lúdica, como dice mi tutor; muy imaginativa, ya que no tenemos datos reales de cómo se hablaba en el siglo xvi, pero me gustaría mucho destacar el carácter serio y exhaustivo con que se está haciendo, ya que hay que ir buscando correlaciones grafía-fonema en más de cien mil palabras que tienen los dos documentos del siglo xvi. Son estos dos los que hasta el momento están trabajándose en esta investigación: los *Documentos lingüísticos de la Nueva España. Altiplano central*, edición de Concepción Company, integran 78 documentos, y también *El habla de Diego de Ordaz*, de Lope Blanch, que son 9 cartas. Entre estos dos tenemos alrededor de cien mil palabras.

Se están estudiando los contextos de manera manual: palabra por palabra. El CHEM está en línea en la dirección www.iling.unam.mx/chem, tenemos, además, otros documentos que se van a integrar posteriormente a la investigación. Se trata de un trabajo paleográfico de María Buelna y Martha Guzmán de la UAM Azcapotzalco; tiene 20 documentos y son los *Procesos inquisitoriales contra indígenas que realizó Fray Juan de Zumarraga en la Nueva España*.

Ahora, la primera parte de la investigación consistió en la comparación de los sistemas fonológicos antes y después del siglo xvi. Para lo correspondiente a antes del siglo xvi, hubo una proliferación de fonemas sibilantes y los autores que he consultado, como Rafael Lapesa, Moreno de Alba, la propia Concepción Company, coinciden en que hay una falta de sistematización en las grafías que representan estos fonemas. También se presenta el mismo problema en la aspiración. Antes del siglo xvi, se produjo una reducción en el sistema

de sibilantes para el español en general, pero es más o menos en este periodo cuando empieza a haber un reajuste: es un periodo de cambios.

El cuadro 1 presenta la propuesta presentada por Concepción Company como sistema fonológico del siglo XVII, basado en el estudio de las cartas de un panadero, incluidas en los DLNE.

b		[d]				g	
p		t				k	
	f		[s]	(š)	y	x	(h)
m			[n]	ñ			
					l	/	[r]
						r	

Cuadro 1. Sistema fonológico del siglo XVII (Concepción Company)

Las mismas cartas están incluidas en sus *Documentos lingüísticos de la Nueva España* y forman, además, parte del CHEM. Con base en estos estudios se hizo el sistema fonológico del siglo XVI para México (véase cuadro 2).

En el cuadro 2 podemos ver las sibilantes que existen; se presenta un fonema alveolar que sería /s/ y un prepalatal /š/.

Tenemos algunas complicaciones para la investigación que son: la calidad de los amanuenses, quienes tenían diferente nivel cultural y educativo, y diversidad de origen. En este periodo muy poca gente sabía leer y escribir, y por esto se recurría a los amanuenses; estos van a reflejar, en la escritura, su nivel cultural y educativo, su origen, y también, de alguna manera, todos los aspectos de aquel que le estaba dictando.

Tenemos el problema del género textual en los *Documentos lingüísticos de la Nueva España*: son documentos de carácter oficial. Por otro lado, los escritos de Diego de Ordaz son cartas personales. Entonces, tenemos que ver si por el género textual algunas palabras estaban ya cristalizadas y su grafía no representa precisamente la pronunciación correcta, o si todavía tenían vigencia algunos fonemas. Nos encontramos aquí, por ejemplo, la diversidad de grafías de *licenciado*, que se usa mucho en los documentos. *Licenciado*, tal como lo escribimos hoy en día, tiene una frecuencia de treinta apariciones, esto solamente en los *Documentos lingüísticos de la Nueva España*. Hallamos, igualmente, *liçenciado* graficado con ç, *ljçenciado* graficado con una *i* larga, con doble ç, con vacilación entre una y otra, y en la última una con *j* y ç.

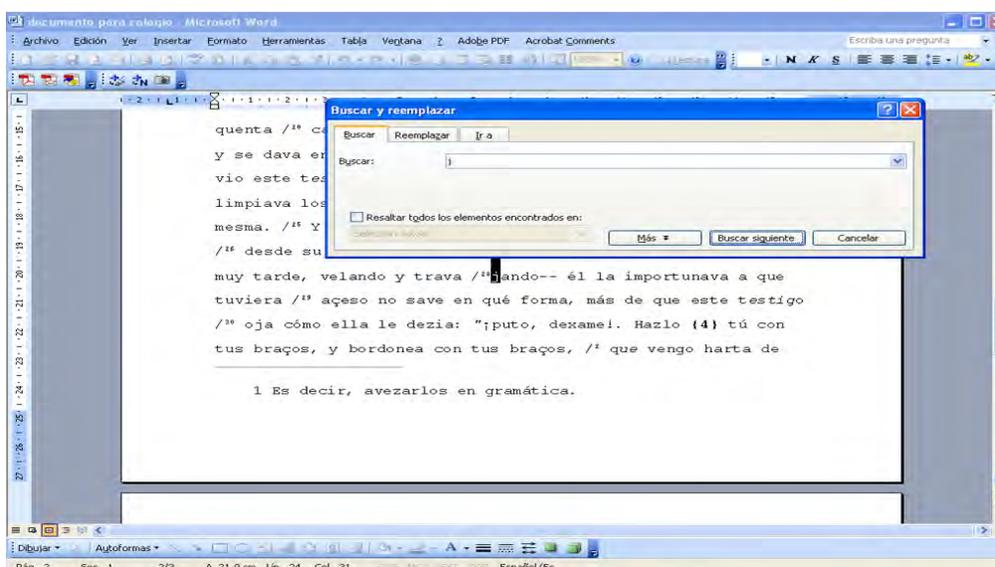
La diversidad de grafías es, entonces, la siguiente: Licenciado con 30, Liçenciado con 22, Ljçenciado con 5, Liçençiado, igualmente, con 5, Licençiado con 3 y Ljçenciado con 1.

p		t		k	
b		d		g	
			ç		
	f	s	ŝ	x	h
			y		
m		n	ñ		
		l			

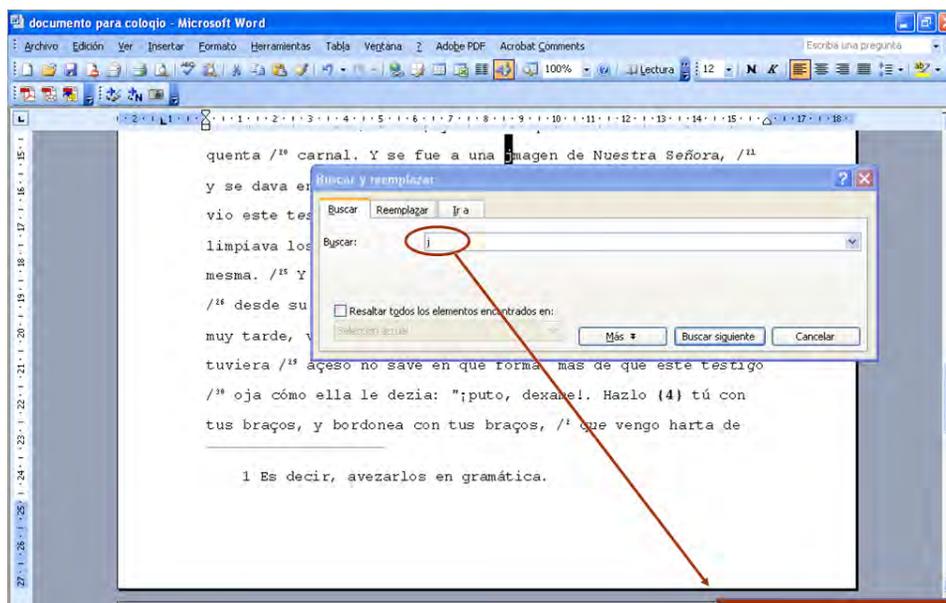
Cuadro 2. Sistema fonológico del siglo XVI en México

El problema, también, con esta diversidad de grafías, sería que se refleja la falta de normatividad para este periodo: no había gramáticas que normaran el uso de la lengua o había muy pocas y no tenían la difusión ideal. Algo muy obvio es la falta de datos reales: no hay grabaciones del siglo XVI.

El método que estoy usando para esta investigación es ir buscando palabra por palabra. Si bien los documentos los tenemos en formato electrónico, hay que hacer una búsqueda manual palabra por palabra. En el cuadro 4 podemos ver un ejemplo; se trata de una búsqueda de *j*, como sonaba esta grafía: si tenía un valor vocálico o un valor consonántico. En este caso, es pertinente presentar todas las grafías que tienen un valor vocálico para /i/.



Cuadro 3. Ejemplo del método utilizado en búsqueda de lexemas



Cuadro 4. Ejemplo con la grafía "j" con sonido /i/

Estos son: "y", "j" e "i". Un ejemplo podría ser la palabra *imagen*, escrito con una *j* inicial, sin embargo tiene un valor vocálico. También la grafía "y" puede equivaler a /y/, como en *yerba* y *yugo*, y "j" a /x/, como en *perjudiciales*. Por lo tanto, debemos crear una regla que permita saber en qué contexto una y o una *j* suenan como vocal, y en cuáles contextos suenan como consonante. La búsqueda manual en los documentos del corpus permite conocer estos contextos y finalmente crear la regla.

Como dijo también Carlos Fonseca, hay excepciones siempre, y de acuerdo a las necesidades del proyecto vamos a elaborar una lista de excepciones para cada grafía.

Las reglas que se están elaborando son las siguientes: una *i* suena /i/ con un valor vocálico en todos los casos; una *y* va a sonar como /i/ con valor vocálico entre dos espacios vacíos: sería el primer ejemplo en el caso de *cuenta* y *relación*. Estos ya son ejemplos que se obtienen del corpus: va a sonar con un valor vocálico entre una vocal y una consonante, como en el caso de *reynos* y *oyr*; cuando es inicial de palabra y le sigue una consonante: en el caso de *yria*, *yndios*; y cuando no importa qué le preceda, una vocal o una consonante, y es final de palabra: en el caso de *asy*, *oy*, *sy*; después, cuando está precedida por una consonante, no importa cual, que es por ejemplo *yan* del verbo *ir*. La *y* va a sonar con un valor vocálico aunque no esté dentro de ninguna de estas reglas, cosa que se ve por el contexto. Tomemos como ejemplo la oración "*yan a conservarse por muchos años*".

En este caso, también, la regla que se obtuvo para la *y* con un valor consonántico es que una *y* seguida de vocal va a sonar como /y/. Veíamos en los ejemplos anteriores el caso de *yugo*, *yo*, *ya*. Si no tenemos esta lista de excepciones, el transcriptor va a hacer una transcripción con un valor consonántico que sería el incorrecto.

También tenemos que la *j* representa el fonema /i/, vocal, entre una consonante y una vocal como *opinión*, *compañja*; cuando es inicio de palabra y le sigue una consonante en el caso de *jmpedor*, *jventario*, cuando está entre dos consonantes, en el caso de *qujsa*, *sjno*. Aquí hay que hacer una observación, que hay una vocal gráfica antes: esa *q* más una *u* va a tener un valor consonántico, eso es lo que se está tomando en cuenta. No importa qué le preceda y que sea final de palabra como *muj*, *nj*, *esrbj*; la lista que se obtuvo fue *oja* del verbo *oír*, igual debe verse por contexto ya que una *j* entre vocales debería sonar como

consonante y se puede confundir con *hoja* en caso de que aparezca en alguna parte del corpus, el ejemplo es: *este testigo oja cómo ella le dezía.*

En este momento de la investigación tenemos ya las reglas para las vocales y algunas otras grafías que aparentemente no tienen problemas, como los que había mencionado de las sibilantes y la aspiración. Asumimos para todas las vocales y algunas consonantes la /p/, /t/, /k/, /b/, /d/, /g/, /ç/, /y/, /m/, /n/, /ɲ/, /l/, /r/ y /r̄/.

Los fonemas que van a representar problemas son: la aspiración, en el caso de /f/ y /h/: no sabemos si se conserva o se aspira en palabras como *fazer*, *fecho*. Efectivamente, todas las del verbo *hacer* alternan la forma: unas tienen *f* y otras tienen *h*, y *fasta*.

También la producción de la *x* presenta conflictos. De alguna manera, sabemos que *dixo* se pronuncia como /x/, sería bastante obvio decirlo, pero no podemos afirmar que sea en un contexto intervocálico ya que tenemos casos como *próximo*, también en un contexto intervocálico, donde suena como /ks/. Hay que ver si se va a hacer en lista o se va a crear una regla y cómo sería esa regla.

Para las sibilantes, una /s/ o una /s̄/ en palabras como *braços*, *toçino*, *çédula*, no sabemos si se pronunciaba /brá-sos/, como ahora, o /brá-s̄os/ como el fonema medieval.

Tenemos un ejemplo de un texto perteneciente al siglo XVI del CHEM, el año es 1576:

1576, ciudad de México.

A.G.N.: Inquisición 117-2, ff.3r-4r.

Denuncia de Pero Díaz contra Tomé Nuñez, candelero, por decir éste que “tener acceso carnal con su mujer como y cuando quisiera no era pecado”. La mujer lo acusa de homosexual.

En la çudad de México, dos dias del mes de noviembre de mill y quinientos y setenta y seis años, antel señor inquisidor, licenciado Avalos, en su audiencia de la mañana, pareçio sin ser llamado, y juró en forma de derecho de dezir verdad, un hombre que dixo llamarse Pero Díaz, natural de Sevilla, çurrador, vecino desta çudad, a san Pablo, de hedad de quarenta y tres años. Y dixo que por descargo de su conçiencia viene a dezir y manifestar que podra aver cinco años questando en la çudad de los Angeles, y biviendo en ella este testigo y Tomé Nuñez, candelero, y Luisa Gallegos, su muger, en cuya casa este testigo estava entonçes, un dia, que serian las nueve de la mañana, la dicha Luisa Gallegos vino del monasterio de sancto Domingo de confesarse, queste testigo la llevó y truxo de braço y vio como se confesó alii con un fraile dominico de alii. Y llevandola a su casa desde el monasterio, la susodicha yva llorando. Y preguntandole este testigo que por qué llorava, la susodiclia le respondió que llorava su mala ventura; que ella lo diría a su marido porque le convenia. Y llegados a casa, ella dixo al dicho Tomé Nuñez, su marido, delante deste testigo: “Tomé Nuñez, mira que me da mi confesor por gran pecado esto que hazeis comigo”.

Sin declarar qué, más de que su confesor le mandava que antes consintiese la miceria, que dexaele hazer tal. Y el dicho Tomé Nuñez le respondió: “Anda, calla; que hazello con mi muger de qualquier suerte no es pecado. ¿Qué saven los friles dominicos? Que les basto yo a dalle catedra y vezalles gramatica”. Y ella le rogo mucho que le hiziese plazer, por amor de Dios, de dezille quién era su confesor para yrse a confesar con él. Y él le respondió “Tereso de Brenes”, reyendose.

Y porfiándole ella en ello, le bolvió él a dezir: “Tereso de Bolaños”, riyéndose. Y otro domingo adelante, levantándose este testigo de dormir la siesta, halló al dicho Tomé Nuñez, que estava limpiando los pechos a la dicha su muger con una turca verde della, y luego se baxa atacándose las calças por el escalera abaxo. Y la dicha Luisa de Gallegos, acuitándose y llorando, dixo a este testigo: “mira este mal christiano, que donde le da la gana de tener açeso conmigo, él haze”. Y le enseñó las manos suzias del acto, y le dixo que en ellas avia tenido quenta carnal. Y se fue a una jimagen de Nuestra. Señora, y se dava en los pechos. Y en las mesmas manos conoçio y vio este testigo la mesma simiente del açeso. Y que le limpiava los pechos donde ella dezia que avia faltado de la mesma. Y despues desto, dos o tres noches, oyó este tesfigo desde su cama cómo —quando ella se yva acostar que hera muy tarde, velando y trabajando— él la importunava a que tuviera açeso no save en qué forma, más de que este testigo oja como ella le dezia: “¡puto, dexame! Hazlo tú con tus braços, y bordonea con tus braços, que vengo harta de travaxar”. No oyó más este testigo, ni savido otra cosa, más de que sobre esto don Rodrigo Maldonado, acalde mayor de la Puebla, hizo informacion, no save ante que scrivano, pensando que hera putto, como ella se lo avia llamado, y creyendo que la avia acometido por detras. Y él estuvo preso, y despues lo soltaron. Y ella se quexava de que lo soltasen. E que se remite al proçeso donde parecerá lo que sobre ello oió.

Notamos la diversidad de grafías que se presentan. Hay muchas ç, por ejemplo, letras geminadas, tal es el caso de *mil*. Hay muchas palabras que se repiten como *dixo* con *x*; no sabemos si ya está cristalizado o el fonema todavía tiene alguna vigencia. Se pretende que el transcriptor con base en estas reglas haga una transcripción fonológica. Hasta aquí va la investigación. Se terminará ya con una idea clara del comportamiento de las sibilantes y la aspiración pero no deja de ser un trabajo muy imaginativo. Eso sería todo, gracias.

Raúl del Moral: Agradecemos a Teresita Reyes y abrimos un espacio para comentarios.

SECCIÓN DE PREGUNTAS

Alberto Barrón: Mencionaste que hay algunas letras que presenta ambigüedad como en el caso de *hoja* y *oja*. Entonces, ¿el transcriptor solamente hace un análisis sintáctico, o también un análisis semántico para ver el contexto, que era lo que mencionabas sobre la diferencia del sonido?

Teresita Reyes: Yo creo que el transcriptor se va a guiar solamente por las reglas, no tiene ningún carácter semántico; precisamente por eso se crea la lista de excepciones.

Alberto Barrón: ¿Va a dirigirse directamente a las listas, si encuentra algo, y lo va a sustituir?

Carlos Fonseca: Sí. Lo primero que va a hacer es verificar si la cadena de entradas está dentro de las excepciones y si está, va a transcribir directamente la regla que tiene en ese archivo, se va a aceptar el movimiento de transcripción y seguiría con todo lo demás.

Raúl del Moral: Hasta aquí con las preguntas. Muchas gracias. ■

EL DESARROLLO TECNOLÓGICO Y SU IMPACTO EN LOS PROCESOS DE TRADUCCIÓN

ANTONIO REYES PÉREZ
GIL-IINGEN, UNAM

Gerardo Sierra: La plática de hoy es sobre extracción de información. La mesa estaba bien definida para que se pudieran dar tres pláticas muy importantes sobre este tema; sin embargo, por cuestiones de horario, tuvimos que hacer un ajuste. Por lo tanto, la primera plática, de Antonio Reyes, no es lo más cercano a extracción de información, no es lo más adecuado; sin embargo, es parte de todo lo que es el evento de Lingüística Computacional.

La plática que nos presentará Antonio Reyes se titula: *El desarrollo tecnológico y su impacto en los procesos de traducción*. Como antecedentes: Antonio Reyes es maestro en Lingüística Hispánica, estudió la carrera de Letras en esta Facultad, ingresó al Grupo de Ingeniería Lingüística (GIL) hace como siete años más o menos; fue, de hecho, el segundo integrante de la Facultad de Filosofía y Letras en el GIL (el primer integrante está haciendo el doctorado en Barcelona) Antonio no se quiere quedar atrás; dado que no hay un buen posgrado aquí sobre lingüística computacional, ha decidido hacerlo en la Universidad Politécnica de Valencia. Estamos de acuerdo con él completamente. Comencemos con su ponencia.

Antonio Reyes: Buenas tardes a todos y gracias por su asistencia. Como ya bien dijo Gerardo, el tema que les voy a platicar no tiene mucha relación con lo que es el trabajo en extracción de información; sin embargo, espero que les sea de utilidad lo que les voy a dar en esta plática.

Básicamente, voy a exponer dos cosas: la primera, una reflexión acerca de cómo la tecnología nos ha ayudado día con día en todos nuestros ámbitos profesionales; prueba de ello es que hay máquinas y *softwares* para todas las especialidades que continuamente están reformándose. Vamos a ver cómo esta tecnología, este impacto, llega hasta la traducción, al proceso de traducción automática, de traducción asistida o traducción manual. El segundo punto es ver cómo la tecnología del lenguaje puede mejorar y utilizar estos procesos de traducción.

El orden que voy a seguir para la presentación es, primero, darles una introducción muy breve de dónde parte el trabajo y hacia dónde va, un tema de discusión, es decir, ver por qué la traducción no es un proceso trivial y por qué requiere de una atención especial; hablaré sobre la lingüística, cómo la lengua tiene un impacto muy importante en todo proceso de traducción y cómo hay que analizarla para hacer un proceso más o menos exitoso en cuanto a los signos se refiere. Hablaremos de la tecnología, cómo ésta tiene impacto en estos procesos y cómo se podría mejorar. Y por último unas consideraciones finales.

Comenzando con lo que es el primer tema, hay que mencionar que la lengua es un sistema vivo, dinámico, que está en constante movimiento; por lo tanto, es muy difícil que pueda ser aprendido por un sistema informático, hay una gran cantidad de fenómenos de referencia intrínseca a la lengua que hace que sea muy

complejo el hecho de poder analizarla, y su contraparte, llevarla a un sistema informático para entender un idioma, por supuesto no significa sólo comprender la gramática de esta lengua sino toda una serie de procesos y fenómenos que encontramos en ésta.

Esto se vuelve bastante complejo. ¿Por qué? Para un proceso de traducción se debe estar consciente que la lengua, que la materia prima de trabajo, es algo muy importante; si no se está consciente de ello, difícilmente se podrá tener acceso a un trabajo más o menos decente en cuanto a traducción se refiere; no basta únicamente con traducir literalmente, sino que hay que implementar la riqueza, la fidelidad de todo el discurso argumentativo. Hay que estar conscientes de que la lengua como un sistema de estudio es muy importante para la traducción. Como último punto, la tecnología: cómo se mezcla desde un principio, cómo ha estado abarcando casi todos nuestros años profesionales y cómo esta tecnología puede ayudarnos a tener un mejor proceso de traducción.

Dentro del tema de lo que es la traducción, antes que nada, tratemos de ver que traducir no es un proceso trivial.

La Real Academia Española define traducción como 'expresar en una lengua lo que está escrito o se ha expresado antes en otra'; también: 'Referido a algo que está en determinada lengua, expresarlo en otra'. ¿A qué se refiere el traducir? Básicamente es llevar una lengua fuente a una lengua meta que está expresada en la lengua fuente; pero no es un proceso trivial, y no es un proceso trivial no porque yo lo diga u otra gente lo diga, sino porque hay gama, una materia prima que está dentro del proceso de traducción.

En este sentido, este tema lingüístico, como tal, es el que regula cómo se está comportando la lengua y qué proceso debe tener el traductor. Tenemos el sistema lingüístico, que se encuentra en constante cambio. Hay fenómenos de distintos niveles, por ejemplo, en el latín *lacte*, que se traduce en español como *leche*, puede haber fenómenos fonéticos, fonológicos; en *elder*, traducido como *older* en inglés moderno, hay fenómenos que tienen que ver con la semántica, por especialización de significados. No es algo que sea tan cuadrado, no es algo que esté ya definido *a priori*; es algo que va creciendo día con día.

Tenemos, como mencionaba hace un instante, factores internos y externos de la misma lengua: desde los niveles de análisis fonéticos, morfológicos, hasta los pragmáticos y discursivos, que de una u otra forma están revolucionando el funcionamiento de la lengua.

Encontramos también aspectos sociales, culturales, económicos, que a fin de cuentas se traducen en todo lo que se está corrigiendo y lo que tenemos a través de nuestra lengua. Es muy importante tomarlos en cuenta y, como si fuera poco, hallamos, además, un aspecto que sale de la gramática: la comunicación. En todo proceso comunicativo hay una serie de fenómenos bastante complejos de analizar y, que a la hora de traducir o de llevarlos a una lengua meta, resultan demasiado complicados: diferencias, implicaciones, análisis, creencias, etc. Son fenómenos que están regulando la lengua. Por lo tanto, tratar de plasmar esta fluidez, esta riqueza discursiva en una lengua meta resulta muy difícil.

Por tanto, ¿qué podemos decir? La lengua no puede ser caracterizada únicamente bajo criterios gramaticales, no podemos tener un proceso de traducción basándonos únicamente en lo que aprendemos en la clase o en nuestro libro gramatical. Hay una serie de procesos que tratan de involucrar al traductor en lo que son las lenguas que está trabajando. Por lo tanto, la lengua se debe de entender como un sistema vivo que tiene referencias tanto internas como externas y que está en constante movimiento. Claro está que el movimiento no es arbitrario, sino que hay una cierta regularización que constantemente está moviéndose. Todo esto hace que el proceso de traducción, que el trabajo del mismo traductor se vuelva muy complejo. En este sentido, ¿qué ofrece la tecnología?, ¿cómo ayuda al traductor a tratar de minimizar y simplificar su tarea? No va a resolvérsela, porque ése no es el punto, sino va a tratar de darle herramientas que faciliten su trabajo.

En este punto la tecnología ofrece bastantes cosas: en primera instancia está el uso de la computadora: tener una máquina que facilite el trabajo de alguna u otra forma está ayudando. Debería de simplificar el proceso. Tenemos sistemas informáticos que hacen que la interacción entre un humano y una computadora sea mucho más eficiente con *softwares* especializados. En este caso hay gran cantidad de programas para traductores que podrían servir para su trabajo; otro punto importante es Internet; debemos cuestionarnos si es válido o si no lo es, si es pertinente, qué tan veraz es la información; sin embargo, es una herramienta más que se encuentra a disposición.

¿Todo esto en qué se traduce? En ventajas y desventajas. Se supone que el uso de la computadora, el uso de sistemas, debería agilizar el trabajo manual y utilizar los tiempos. Creo que todos los que hemos tenido contacto con las máquinas sabemos que esto no es tan real: a final de cuentas, si yo como estudiante de lingüística trato de interactuar con la máquina —con un *software* especializado—, me va a costar algún trabajo, por lo que las ventajas pueden traducirse en una desventaja, debido a que no se está consciente de todo lo que hay que manejar para interactuar con sistemas.

Otra ventaja es tener un *software* especializado para un proceso de traducción, existen bastantes programas que pueden facilitar mi trabajo, si no me basta con uno o dos que satisfagan mis necesidades.

Existen bastantes desventajas, sólo mencionaré algunas: por lo general es un trabajo semiautomático que requiere tanto la supervisión como la revisión y en muchos casos la necesidad de un experto para avalar o tirar lo que la máquina está diciendo en forma automática. Hay una información desactualizada, si estoy trabajando con un *software* de hace dos años. A nivel lingüístico, tenemos que no necesariamente esa información corresponde con la información lingüística que manejan hoy en día, lo que implica un problema: no hay actualización de información; pocas veces, al mismo tiempo, el *software* toma en cuenta la lingüística. La lingüística como tal es una disciplina que puede ayudar bastante desde su óptica, desde su perspectiva a los desarrollos tecnológicos.

Tenemos el caso de Internet. Creo que todos estamos conscientes de que es una fuente muy importante de información; sin embargo, debemos cuestionarnos pertinentemente su uso en un proceso de traducción y también de lo veraz de la información que estamos obteniendo. Yo, como traductor, quizá no tengo la misma frescura si busco en un libro de gramática la información que requiero; sin embargo, en Internet, por ejemplo, puede haber foros en línea, correos, *chat*, que de cierta forma me están refiriendo la actualidad de la lengua, y es una ventaja que está en mí y en mi desarrollo tratar de aprovechar o no. Y claro, siempre viendo qué tan veraz y qué tan confiable es esta información.

En este sentido lo que está haciendo la tecnología no deja de tocar la lengua. Surgen lo que se llaman las tecnologías del lenguaje. ¿Qué es lo que buscan esas tecnologías? Tratar de traducir toda la información, todo el conocimiento que existe *per se* en la lengua, llevarlo a sistemas informáticos que pueden representar un mero aprovechamiento en cuanto a la lengua se refiere.

Los que usan esa tecnología pueden darse cuenta de todo lo importante, tanto para el trabajo lingüístico como para el trabajo literario, el trabajo del traductor, el intérprete, etc. Hay que tomarlos en cuenta dentro del proceso de traducción. Algunas tecnologías que pueden aplicarse directamente a este desarrollo profesional son, por ejemplo, los cortes electrónicos. El hecho de que yo pueda ver cómo una palabra dentro de un contexto especializado o dentro de un contexto coloquial puede ir variando su significado, es algo que podría encontrar, de una forma más eficiente y veraz, en un corpus electrónico. El hecho de buscar información por

su categoría gramatical también me facilita el trabajo, mi desenvolvimiento, el uso de diccionarios, y no únicamente de diccionarios de lengua, sino diccionarios especializados, es también otra herramienta que las tecnologías del lenguaje prestan al traductor (tesaurus, bancos terminológicos, etcétera).

Un caso muy particular del trabajo del traductor es enfrentarse a trabajos ya muy especializados, como lenguajes científicos y técnicos. El hecho de que existan bancos terminológicos que puedan proporcionar la información que se requiere y mostrar cómo se va desarrollando esta lengua en el proceso de un contexto es muy valioso. El uso de ontologías, por ejemplo, que tratan de representar cómo están ubicados los contextos que presentamos a través de la lengua, también, es muy valioso para el traductor. Hay más ontologías que pueden servir para simplificar el trabajo, entender una lengua y poder hacer una mejor traducción.

Mi invitación es para todos los lingüistas: traten de entender que la lengua no es un estadio aparte del resto de la comunidad con la sociedad; el uso de las tecnologías, por lo tanto, puede ser muy útil.

Para concluir, considero que todos estamos conscientes de que toda lengua implica un mundo de conocimiento y que en la medida en que un traductor esté consciente de toda esta información podrá desenvolverse de una mejor forma; su traducción se llevará a cabo de mejor manera, ya sea terminológica o coloquial. Visto desde una perspectiva que involucre a la lengua como una herramienta que le dé información acerca de lo que está marcando será muy importante. El uso de herramientas informáticas va a simplificar el desenvolvimiento profesional, aunque no es suficiente.

En la medida en que estas herramientas de la tecnología del lenguaje puedan involucrar todo proceso que se analiza dentro la lengua, va a beneficiar el desarrollo y el perfeccionamiento de software y a las mismas tecnologías del lenguaje. No nos resuelve, empero, un trabajo, sino que lo sistematiza y lo simplifica para que todo trabajo, no sólo el del traductor, sino de cualquier profesional que tenga que ver con la tecnología, se vea beneficiado y simplificado. Con eso termino mi plática. Gracias por su asistencia.

Gerardo Sierra: Muchas gracias. Debemos tomar en cuenta que muchas veces el trabajo de los que estudian letras y lingüística es la revisión y corrección de estilo o, bien, la traducción. En ese sentido, las herramientas que se plantean sobre las tecnologías del lenguaje no sólo son útiles para lo que es traducción, sino inclusive para lo que es revisión de estilo. El punto es saber utilizar esas herramientas y saber que existen. No sé si haya alguna pregunta para el maestro Antonio Reyes.

SECCIÓN DE PREGUNTAS

Pregunta 1: ¿El proyecto está abarcando una lengua en particular?

Antonio Reyes: No; de hecho, esto se puede aplicar a cualquiera. Toda lengua es un mundo,;entonces no puede ser centrado nada más al español, al inglés. Cada una tiene ciertas particularidades, por lo que esto se puede aplicar para cualquier lenguaje. La tecnología como tal es aplicable a la lengua que estés utilizando.

Gerardo Sierra: Muchas gracias. ■

USO DE TÉCNICAS Y RECURSOS DE LA LINGÜÍSTICA COMPUTACIONAL EN SISTEMAS DE EXTRACCIÓN DE INFORMACIÓN

CARLOS MÉNDEZ/GIL-IINGEN, UNAM
JUAN HERNÁNDEZ DE ANDA/FCA, UNAM
CECILIA LUZÁN/FCA, UNAM
LIZBETH MAGAÑA/FCA, UNAM
JOSÉ LUIS MARTÍNEZ MARTÍNEZ/FCA, UNAM

Gerardo Sierra: Vamos con la segunda plática de esta tercera mesa, que es *Uso de técnicas y recursos de la lingüística computacional en sistemas de extracción de información*. Esta ponencia es de todo un grupo asistido por Carlos Francisco Méndez Cruz del GIL. Él proviene de la Facultad de Contaduría y Administración de la carrera de informática, cuyos conocimientos decidió utilizarlos para la lingüística. Empezó estudiando la carrera de letras, pero después vio que era un camino largo y prefirió entrar directamente a la maestría; actualmente la está cursando, después de haber llevado todos los requisitos correspondientes de la lingüística. Como profesor de la Facultad de Contaduría y Administración ha dado muchos cursos y además ha dirigido tesis. Entre las tesis propuso un tema en el cual se incluye el que estamos presentando. Los alumnos que exponen son Juan Carlos Hernández de Anda, Carmen Cecilia Luzán Hernández, Lizbeth Mireya Magaña López y José Luis Martínez Martínez Reyes, todos ellos de la Facultad de Contaduría y Administración de la carrera de informática.

Carlos Méndez Cruz: Buenas tardes a todos. Brevemente quiero aprovechar para agradecer a Javier Cuétara y al Dr. Sierra, como organizadores, la oportunidad que me dan a mí y a mis chicos para que tengan este espacio para mostrar su trabajo de investigación. Lo que vamos a mostrar como ya bien dijo el Dr. Gerardo es una tesis que desarrollaron para su titulación, la cual no hubiéramos podido terminar sin el apoyo del Grupo de Ingeniería Lingüística, porque nos proporcionó el material bibliográfico; también tuvimos el apoyo del Dr. Diego Isidoro del Instituto Politécnico Nacional. Aprovecho aquí para agradecerles a los chicos nuevamente que hayan aceptado el reto de entrar a un área nueva, porque en realidad en la informática nos dedicamos a los desarrollos de sistemas de información, a dar soluciones tecnológicas; sin embargo, jamás están ligadas a un procesamiento de lenguaje o a este tipo de áreas que tienen que ver con la inteligencia artificial o simplemente natural.

Les agradezco que hayan tomado el reto que llevaron a un buen término como ya también mencionaron. Los presento: Juan Carlos Hernández, él trabaja en el área de sistemas en la Facultad de Derecho donde también da algunas clases; Cecilia González Hernández, que labora en una consultoría en el Tropylogy, que está trabajando para Nacional Financiera, y Lizbeth Mireya Magaña, José Luis Martínez Martínez, que se encuentra en el área de sistemas de Sanborn's. Todos se han titulado con este trabajo el pasado 20 de octubre del 2006, son recién egresados. No me resta más que darles paso para que ellos presenten su aportación.

Lizbeth Magaña: Buenas tardes, mi nombre es Lizbeth Magaña y para iniciar les voy hablar sobre los sistemas de extracción de información. Los Sistemas de Extracción de Información procesan grandes cantidades de datos conformados por un conjunto enorme de textos, llamado corpus, con el objetivo de extraer y organizar únicamente los datos que cumplan con un perfil deseado, es decir, la información útil o de interés. Para lograrlo utilizan herramientas lingüísticas con las que se hace un análisis morfológico, léxico y semántico a dicho corpus.

Como antecedentes a los Sistemas de Extracción de Información, encontramos que estos surgen principalmente a raíz de las MUC (*Message Understanding Conferences*) patrocinadas por la DARPA (*Defense Advanced Research Projects Agency*). Las MUC se desarrollaron entre los años 1987 y 1998 realizándose siete conferencias, en las cuales se proporcionaba el desarrollo de los Sistemas de Extracción de Información en dominios de aplicación y el uso de corpus etiquetados. En estas conferencias se les proporcionaba a todos los investigadores un corpus ya etiquetado y un dominio de aplicación; los sistemas desarrollados competían por dar los mayores resultados. Los investigadores eran de universidades particulares y de esta institución.

Como ejemplos de la MUC tenemos la MUC 3 y la MUC 4; la tarea era extraer y almacenar características específicas de incidentes terroristas. El dominio de aplicación fueron noticias y artículos sobre este tipo de ataques. El corpus se recopiló principalmente de noticias de periódicos y revistas.

El objetivo de la tesis fue desarrollar un Sistema de Extracción de Información al cual llamamos *Sistema de Extracción de Información de Nuevo Software*, que toma archivos de un texto no estructurado y los analiza para obtener un lanzamiento de *software*.

El lanzamiento de *software* es el evento donde una compañía lanza o pone a la venta un nuevo programa (*software*) y la información a extraer es el nombre de la compañía, el nombre del *software* y la versión del mismo.

- Un ejemplo sería, en este caso, la compañía Microsoft, el *software* sería *Messenger* y la versión sería 8.

Ahora, con el objetivo de de exponer más claramente los problemas a los que nos enfrentamos, expongo diversos ejemplos:

- “Microsoft lanza *Messenger* 8...”
- “Microsoft ha anunciado el lanzamiento de la nueva versión 8 del *Messenger*...”
- “Microsoft Inc. mejora su plataforma de aplicaciones con el lanzamiento de SQL SERVER 2005, VISUAL STUDIO 2005 y la próxima versión del *Messenger*...”

En el primero de estos ejemplos, aparentemente sencillo, Microsoft es la primera palabra y es el nombre de la compañía, después sigue un verbo en tercera persona, inmediatamente después sigue el nombre del producto seguido por la versión. El segundo ejemplo es un poco más complicado, puesto que encontramos el mismo Microsoft, sin embargo, el verbo es compuesto, éste es *ha + -ado*; además, tenemos información irrelevante: “*lanzamiento de la nueva versión*”; encontramos primero la versión y después el nombre del producto. El tercero es ligeramente más complicado: en él tenemos a Microsoft, que es una compañía formada por dos palabras, tenemos también algo de información irrelevante, encontramos más de un producto y estos están formados por más palabras. Ahora, ¿cómo nos enfrentamos a estos problemas a lo largo de la tesis? Utilizamos herramientas de lingüística computacional que les explicará Cecilia Luzán.

Cecilia Luzán: Buenas tardes, mi nombre es Cecilia. Yo les voy a explicar las técnicas que utilizamos para desarrollar nuestro Sistema de Extracción de Información. Les voy a decir lo que es la *tokenización*. Se trata de dividir un texto en varias unidades

mínimas o pequeñas unidades, en este caso nosotros *tokenizamos* por palabras gráficas, que es cualquier palabra que nosotros pensamos, como el artículo *la*, una palabra, un nombre, y por oraciones gráficas; las oraciones gráficas las dividimos en mayúscula inicial o de *punto a punto* como las mencionamos aquí. Otra técnica que utilizamos fue la *lematización*, que consiste en el proceso de asociar variantes de una misma palabra a una forma del diccionario; por ejemplo, las palabras *vas*, *fui*, *iré* pueden ser asociadas con la forma del diccionario *ir*; es decir, lo que encontramos es el lema de cada una de ellas. Las palabras, por ejemplo, *caminaré*, *caminaron* y *camino* pueden estar asociadas a una forma del diccionario, la cual es *caminar*.

Para hacer la *lematización* en nuestro sistema, utilizamos un léxico, el cual contenía todas las variantes de una palabra, por ejemplo ésta: *abalancé*, *abalances*, *abalance*, y la asociamos con una entrada del diccionario la cual es *abalar*, esto es, su lema.

De igual forma, utilizamos etiquetado POST o *Parts of Speech Tagging*; entendemos etiquetado como el marcado de lo que deseamos distinguir, en este caso, lo que quisimos resaltar fueron las partes de la oración. ¿Cómo etiquetamos nuestro corpus? Asignando a cada palabra su categoría gramatical. ¿Qué entendemos por asignar a cada palabra su categoría gramatical? Identificar si se trataba de un artículo, de un sustantivo, de un verbo, y a la vez asignarle los elementos en que se componían cada uno de estos elementos gramaticales. Verbigracia, buscábamos si un verbo estaba en tercera persona, si era singular o plural. Nuestro sistema utilizó un lexicón, el cual incluía la categoría gramatical. Utilizamos, igualmente, el POST que ya había mencionado. Las características de cada una de las palabras nos las indica nuestro etiquetado, que es, por ejemplo, “V” cuando se trata de un verbo, “I” de indicativo, “S” pasado, “1” primera persona, “S” de singular, aunque algunas de las características que encontramos en este estándar no aplicaban en ninguna palabra.

En un corpus etiquetado (véase cuadro 1) se puede observar la palabra, su categoría gramatical y aparte su lengua. *Entity Named Recognition* es el reconocimiento de entidades con nombre, el cual consiste en identificar entidades en los textos que pueden ser cualquier cosa, por ejemplo lagos, ciudades, organizaciones, fechas, etc. En nuestro caso, identificamos empresas de software mediante un conjunto de compañías, el cual nosotros mismos recolectamos. Por mi parte es todo, muchas gracias.

José Luis Martínez: Buenas tardes, vamos a hablar sobre los recursos utilizados. Para empezar, el corpus. Un corpus es un conjunto de textos reales; es decir, se pueden encontrar en la vida diaria, pueden tratar de un área específica, esto es, de un dominio; cumplen con ciertas características y forman parte de un mismo idioma. Además cuentan con su corte informático, pudiendo llegar a tener miles de palabras.

Nuestro corpus para el desarrollo consistió en 240 noticias de internet. Primero lo empezamos con 120, treinta cada quien. Tomamos noticias de internet que creíamos que eran lanzamientos de software; pero como los resultados no eran muy claros, decidimos bajar otras 120 noticias para tener las 240. Todas se trataban de lanzamientos de *software*; fue bajo este esquema que hicimos nuestra aplicación. Finalmente, obtuvimos un corpus de prueba, que consistió de 70 noticias.

Microsoft_E1_E ha_VIIP3S0-haber **lanzado_V0P00SM-**
lanzar la_TDFS0-la beta_NCFS000-beta_NCFP000-beta
del_SPCMS-del futuro_AGMS000-futuro_NCMS000-
futuro Messenger_X1 sin_SPS00-sin la_TDFS0-la
habitual_AGIS000-habitual_AGIS000-habitual
fanfarria_NCFS000-fanfarria y_CC00-y
publicidad_NCFS000-publicidad que_CS00-
que_PE3CN00-que acostumbra_VOIP3S0-
acostumbrar_V0N0000-rodar a_SPS00-a los_NCMS000-los
productos_NCMP000-producto estrella_VOIP3S0-
estrellar_V0R02S0-estrellar_NCFS000-
estrella_NP00000-estrella

Cuadro 1. Ejemplo de un texto etiquetado

De estas 70, diez eran completamente ajenas a lanzamientos de *software*, como podía ser espectáculos, deportes, política, etc. De esas 70, veinte eran de tecnologías de información, pero no eran lanzamientos de *software*, eran tecnologías de información como videojuegos u otros. De esas, también 10 eran lanzamientos de cualquier producto que no sea *software*: lanzamientos de películas, de un nuevo *shampoo*, entre otros. 30 eran lanzamientos de *software*. Estas 30 son las que nos interesaban.

Según José G. Moreno de Alba, un *lexicón* es un repositorio de información léxica, donde a cada unidad léxica se asocia un conjunto de información que incluye su categoría sintáctica, su interpretación semántica a nivel léxico y diversas propiedades morfológicas, sintácticas y semánticas. Se utilizó el lexicón del Dr. Grigory Sidorov del Instituto Politécnico Nacional.

En el lexicón encontramos desde la letra “A” hasta la “Z”, por lo que hay una gran cantidad de palabras dentro de éste. Tenemos la palabra, su categoría gramatical y su lema asociado, o su palabra de diccionario asociada. También contamos con un “*Gazetteer*”. Se denominó *Gazetteer* a un listado de compañías de *software*, es decir, ENR.

Nosotros tenemos nuestro corpus; a éste lo vamos a *tokenizar*, a dividir por palabras. Se le va a hacer un etiquetado POST, *parts of speech*, esto es, su categoría gramatical y, por último, su lematización.

¿Qué pasa cuando una palabra no está en el lexicón? Por ejemplo, Microsoft no la vamos a encontrar en él, para eso nos sirvió el *Gazetteer*. Nosotros lo hicimos con base en cientos de compañías de *software* que bajamos de internet o que pudimos suponer que eran este tipo de compañías.

Nuestro método de extracción fue mediante *expresiones regulares*. Una expresión regular es un conjunto de patrones que sirve para representar una cadena de caracteres, por lo que provee una caracterización finita de un conjunto infinito de dichos patrones.

Con respecto a una expresión regular, yo puedo encontrar un producto si tengo un pronombre personal, o un artículo, o un determinante; si además de eso, le sigue la palabra “nueva” o “nuevo”, y a eso le sigue un conjunto de palabras.

Después de esta cantidad de palabras recibo un nombre propio, seguido por el verbo ya sea *ha lanzado* o *ha creado* o *ha liberado*, *anunciado* o *publicado*, y, por último, encontramos un conjunto más de palabras y un nombre propio. Ésa puede ser mi expresión regular.

¿Qué es lo que podría caer en una expresión regular? Con el patrón anterior podría hacer concordancia un texto como: *una nueva liberación de software se dio a conocer por la compañía Microsoft, que ha lanzado ahora Messenger Plus*. Esto me indica

que posiblemente yo puedo encontrar el producto que se va a lanzar en esta expresión regular. En breve les explicarán el proceso que llevamos a cabo.

Juan Hernández de Anda: Ahora vamos a explicar un poco acerca de cómo funciona el sistema y las soluciones que utilizamos. Lo primero, es que necesitamos identificar el corpus. Este corpus puede ser dado a partir de la recopilación manual por parte del usuario, quien lo guardó en una carpeta; o también se le puede dar un sitio de Internet del cual el sistema automáticamente puede recolectar el corpus. La entrada puede ser ya sea el texto plano o HTML. Ya armado el corpus, empieza a procesar el sistema cada una de las noticias o cada uno de los documentos que tenemos. Lo primero que hacemos es una *tokenización* por palabra, es decir dividimos todo nuestro documento por palabra y por oración; después de que ya se ha dividido, lo etiquetamos con el etiquetador POST (partes de la oración); le ponemos algunas cosas para identificar si es verbo, si es sujeto, el número, si es masculino, si es femenino, y es a través de las etiquetas que vamos identificando cada una de las palabras.

Posteriormente, realizamos un proceso de reconocimiento, que tiene que ver con el *Gazetteer* de compañías. Nosotros tenemos una tabla donde tenemos un listado de compañías, y si el nombre de la compañía encaja con alguna de las palabras que viene en el documento, entonces, se etiqueta automáticamente como una compañía de *software*. El resultado de este proceso nos arroja una salida: archivos de texto etiquetados, localizados por oración y por palabra, y etiquetados, además, semánticamente. Las empresas que fueron identificadas se etiquetaron distintamente.

A continuación, volvemos a procesar el documento; esta vez identificamos las *keywords* de lanzamiento o *keywords* de TI. *Keywords* de lanzamiento son palabras que nos van a identificar un lanzamiento como: *nuevo*, *propiamente*, *lanzar*, *saldrá*, *anunciará*. Palabras que indican que es una presentación de algo. En fin, éstas se contabilizan.

Las *keywords* de TI, palabras que son de tecnología: *computación*, *software*, *informática*; palabras que nos hablen acerca de que el documento es propiamente de esa materia: *informática*.

A base de modelos de oración, tratamos de identificar lo siguiente: las compañías, los productos y las versiones. Por ejemplo, *Microsoft lanzó*, palabra que empieza con mayúscula más el verbo *lanzar*, *Microsoft lanzó*, *lanzará*, por eso es que encontramos el lema o el infinitivo raíz, para que les sea más sencillo poder entender.

En este punto, ya tenemos en nuestra base de datos las compañías que se identificaron, los productos, las versiones; y esto lo guardamos por cada una de las oraciones donde se van encontrando. Pasamos a armar *templates*, apuntar por párrafo, por oración, si se encontró en ese párrafo alguna versión de algún producto, propiamente el producto o alguna compañía. De esta forma armamos cierto número de *templates* que salgan por noticia. Esta cantidad de *templates* va a ser filtrada: el primer filtro es que necesitamos encontrar una *keyword* que nos hable de que es un lanzamiento de alguna situación y de eso, palabras que nos hablen de tecnología. Porque si no cumple con este estándar, lo descartamos. Después los siguientes filtros que hacemos es elegir los *templates* que tengan mayor cantidad de información; podemos hallar un *template* que tenga *compañía*, *producto*, *versión*; otro que nada más tenga *producto*, *versión*; entonces nos quedamos con el que tiene los tres elementos y eliminamos el que tiene dos.

Si después de este filtro tenemos todavía más de un *template* por compañía, lo que hacemos es que elegimos el *template* que está más cerca del origen del documento, debido a que también, con base en nuestra

investigación del corpus, pudimos notar que la información más relevante viene al principio de la noticia. Inmediatamente, vamos a tener un *template* por noticia, aquellas que fueron seleccionadas a través de estos filtros como lanzamientos de algún producto tecnológico.

Finalmente, se le presenta al usuario los resultados que ya acabamos de obtener. Este proceso, esencialmente, es lo mismo que haríamos manualmente nosotros: meternos a un sitio de Internet, estar leyendo alguna cantidad de noticias, y, manualmente, quedarnos únicamente con las que hablen de algún lanzamiento de tecnología. Alguna vez, un científico en computación hablaba que la programación es simplemente enseñar al sistema lo que manualmente hacemos. Nosotros intentamos, con base en la lingüística, desarrollar un sistema como el que nos tocó desarrollar. Lo que realizamos manualmente fue pasarlo al software para poder dar un resultado final.

En una de las pruebas que hicimos fue que de 30 noticias se recuperaron 26; y de las 26, nueve noticias eran correctas. Teníamos un corpus de 30 noticias de lanzamientos de tecnología del cual se extrajeron 26. Y de estas 26, únicamente 19 eran correctas; lo que nos dio un *precision* de 73% y un *recall* de 86%.

Nuestras conclusiones son las siguientes: Usamos tres de los cinco niveles del lenguaje: el morfológico, el sintáctico y el semántico. No pudimos abarcar más por la complejidad que ya implica el nivel pragmático; y el fonológico no es propiamente la materia, ya que estamos trabajando sólo con texto; no nos metemos con sonido en ningún momento.

Para el reconocimiento de entidades (*entity recognition*), además del uso de *Gazetteers*, se utilizaron *keywords* y expresiones regulares.

La constitución y análisis manual de un corpus fue de gran utilidad para el desarrollo del sistema, ya que, como ya les explicaba, todos los filtros que realizamos los hicimos a base de experiencia, de estar etiquetando manualmente y hacer el proceso manual.

Métodos computacionales más complejos darían mejores resultados. Hay una gama más extensa de técnicas que se pueden aplicar, de recursos que se pueden utilizar para poder obtener mejores resultados; pero en nuestra tesis, utilizamos los que ya les explicamos.

En algunos años, los Sistemas de Extracción de Información serán herramientas básicas en la búsqueda por Internet. Eso es una realidad, pero la tendencia es que los buscadores sean más fuertes, más completos hasta que en alguna ocasión podamos hacerle una pregunta en lenguaje natural y nos pueda hacer una búsqueda tal y como la haríamos manualmente.

Los sistemas de extracción de información son áreas poco investigadas en nuestro país. También las empresas poseen muy poco conocimiento de su existencia y, por lo tanto, de los usos que se pueden hacer de ellos. Al tratar de consultar bibliografía, la mayor parte estaba en inglés o era de origen europeo, propiamente de España. No existe bibliografía escrita en el país.

Esta tecnología puede ser susceptible de utilizarse en cualquier organización, cualquier empresa, desde algún área económica, áreas de gobierno, cualquier institución que utilice grandes cantidades de información, necesita que ésta sea analizada para poder extraer lo que ellos necesitan para poderse utilizar en sus términos.

La extracción de información puede presentar una ventaja a las organizaciones al economizar tiempo. Si uno tiene un buen sistema, lo que manualmente alguien se tardaría días, semanas en hacer, la máquina lo podría procesar en un tiempo comparativamente mínimo de horas, de tiempo. Eso sería todo.

Gerardo Sierra: Creo que ha sido muy interesante ver cómo la tecnología del lenguaje está participando en un tema como éste. Porque trabajar con las industrias del lenguaje realmente no es sencillo. Como nos han mostrado, buscar cuestiones de software fue todo un trabajo entre todo un equipo de cinco personas, donde cada quien tenía bien identificadas sus tareas. Llegaron

a buenos resultados, por supuesto, pero hay mucho más que se puede hacer. Justamente hay que abrirnos hacia este campo y ver todas las posibilidades. No sé si haya alguna pregunta.

SECCIÓN DE PREGUNTAS

Pregunta 1: Tengo una pregunta sobre la expresión regular. Dijeron algo de cualquier número de palabras; mi pregunta es, ¿no tienen un límite?

José Luis Martínez: Sí, tenemos un límite. Por ejemplo, aquí encontramos un conjunto de palabras hasta que encuentre un nombre propio. Con un nombre propio me refiero a una palabra que inicie con letra mayúscula; o, de lo contrario, que termine con un punto, por ejemplo. Ese sería su límite.

Pregunta 2: ¿Lo primero que ocurra?

Carlos Méndez: Lo que pasa es que sí se notó que la desventaja de usar expresiones era esta situación. Como lo comentaron, se hace un corte por enunciado, entonces este tipo de expresiones se aplica a los enunciados, por lo que si el enunciado no entra, se descarta. Por eso es que se logra un *ranking* de enunciados. No es hacia todo el texto, porque si fuera hacia todo el texto podríamos encontrar una palabra al principio y otra al final, como tú lo decías.

Pregunta 3: Yo tengo una pregunta, la bibliografía que encontraron, ¿de cuándo era?, ¿qué tan nueva era?

José Luis Martínez: Bueno la bibliografía, básicamente, la encontramos en libros, la más actualizada era de los 90. Si tú quieres encontrar algo del 2000 en adelante, la mayor parte está en Internet, sobre todo, en coloquios, en tesis, en otras investigaciones.

Pregunta 4: ¿Y puedes consultarlo entonces?

José Luis Martínez: Sí.

Pregunta 5: El conjunto, ¿tenían alguna ventaja en especial?

Carlos Méndez: En realidad sí tienen una gran ventaja. Digamos que una de las propuestas de la tesis precisamente consiste en un sistema de información, porque consideramos un escenario como el siguiente: un gerente quiere estar actualizado en los nuevos lanzamientos de software, porque tiene que estar enterado, un gerente de informática, un gerente de tecnología de la información; tendría, entonces, que leer todas las noticias de todos los periódicos o estar leyendo mucho; quisimos brindarle la oportunidad de tener un sistema de información; pero ese sistema de información tendría la ventaja competitiva de usar un

procesamiento con recursos y técnicas lingüísticas. ¿Por qué? Pues porque, a partir del análisis que hicieron del corpus original, del corpus que ellos etiquetaron a mano, se dieron cuenta de cuál fue la lista de *keywords* que podían utilizarse, cuáles eran los rasgos lingüísticos que estos tenían, cuál era la estructura sintáctica que los lanzamientos de *software* llevan o conllevan; a lo mejor parece obvio, van a decir “lanzó” pero cuando ellos hicieron su análisis encontraron muchas otras palabras que no necesariamente eran “lanzó” y también hablaban de un lanzamiento de *software*. Creo que la ventaja competitiva estuvo en esto: en hacer un análisis lingüístico; esa era la dirección de la tesis; y porque si lo hiciéramos meramente computacional, creo que sólo daríamos con búsqueda de patrones de cadenas. Por eso comentábamos que hay métodos más sofisticados que alcanzan niveles de precisión mejores: árboles de decisión o entrenamientos de programas, sin embargo, entre la investigación que hicieron, uno de los métodos eran las expresiones regulares; ése fue el que decidimos tomar, pero conscientes de que hay métodos más complejos.

Pregunta 6: ¿De casualidad pensaron lanzarlo al mercado?

Carlos Méndez: Sí, ellos lo pensaron. Considero que es un sistema que se puede vender, así como está en ese momento, porque ahorran tiempo al usuario.

Pregunta 7: ¿Por qué no usan el nombre propio? Entiendo que son limitadas las empresas que trabajan este tipo de recursos ¿por qué no sustituir ese nombre propio, que el sistema identifique una mayúscula y reducir las posibilidades a los nombres de las empresas sobre las que se pretende localizar las noticias?

Juan Hernández de Anda: De hecho se utiliza una variante un poco más extensa; ésas las dimos solamente como ejemplo. Si es de montar la constitución de la palabra, porque no solamente necesitamos una mayúscula al principio, que estén puras mayúsculas, que contenga alguna letra distinta o un número, utilizamos un rango un poco mayor; aunque finalmente decidimos que la primera sea mayúscula, o que todas sean mayúsculas, que contenga un número.

Alberto Barrón: Una razón para no hacer eso es que lo cierras, solamente el sistema funcionaría para esos nombres, si lo quieres usar para otra cosa, ya no lo podrías hacer.

Pregunta 8: A eso me refiero, si queremos que sea sólo para computación es más preciso el nombre de las empresas, porque viendo el prototipo podría ser: “ha lanzado una línea de bicicletas” y no nos interesa saber sobre las novedades en las bicicletas o en los deportes. Yo entiendo que está enfocado a sistemas de *software*.

Juan Hernández de Anda: Eso es cortarlo, más que por lo que usted me comenta, por la discriminación que ya hacíamos de que necesitamos encontrar por lo menos nueve palabras con algo de tecnología, porque, si quitáramos ese candado, también podría decir “L’oreal lanzó” o, como usted decía “Benotto lanzó la bicicleta”. Si quitáramos ese candado podría ser un poco más amplio, aunque también, para tener mejores resultados, tendríamos que hacer *Gazetteers* de compañías de bicicletas, del ambiente que fuéramos a trabajar.

Carlos Méndez: Con el análisis que hicieron, también se dieron cuenta de que no es tan cerrado, como tú dices. Porque había empresas que en nuestra vida habíamos escuchado; como dice Alberto, si nos hubiéramos cerrado al lexicón, hubiéramos

dejado afuera posibles empresas que no conocíamos. Y si bien, como tú dices, la tecnología está cerrada a unas empresas, también por otro lado, estas empresas salen muy rápido porque la tecnología también avanza mucho y una nueva empresa se forma y lanza un nuevo producto o lanza una nueva versión, etcétera. ■

MÉTODOS PARA LA OBTENCIÓN AUTOMÁTICA DE TÉRMINOS EN UN ÁREA DE ESPECIALIDAD

ALBERTO BARRÓN
IIMAS / GIL-IINGEN, UNAM

Gerardo Sierra: Ahora nos va a hablar Alberto Barrón, quien ingresó al Instituto de Ingeniería para realizar su tesis de maestría. Él acaba de terminar la Maestría en Ciencias de la Computación en el IIMAS. Él nos va a hablar de *Métodos para la obtención automática de términos en un área de especialidad*.

Alberto Barrón: Voy a platicar algunas técnicas que se han utilizado en la extracción de términos y en particular la que nosotros hemos implementado, esto es, C-value y NC-value. Vamos a dar algunos ejemplos para que ustedes entiendan el proceso y algunas conclusiones. Este proyecto es sólo una parte de uno más grande que se llama *Extracción de conceptos en textos de especialidad a través del reconocimiento de patrones lingüísticos y metalingüísticos* que dirige Gerardo Sierra; él fue quien ideó en principio la adaptación y la creación de un extractor de términos en español. Del lado de la lingüística, este extractor se puede aplicar en la creación de diccionarios como apoyo, debido a que, como lo menciona el proyecto central, de lo que se trata es de detectar los distintos contextos, que son los bloques de texto en los que un término es definido. Mi objetivo es detectar aquellos términos que estén definidos, del lado computacional, es útil en los buscadores que no se basen solamente en la búsqueda de cadenas estáticas sino que realicen búsquedas más semánticas; también es importante que en la creación de documentos no se utilicen sólo palabras, sino los términos claves que participan en ellas. Se han desarrollado muchos extractores de términos, la mayoría de ellos son para el inglés y el francés.

Los primeros que se crearon estaban basados únicamente en reglas lingüísticas: de lo que se trataba era de dictar un texto, como ya había dicho Juan Carlos, identificar las partes de la oración y simplemente adecuarse a los patrones que se sabían de antemano. Por ejemplo, en inglés, el patrón más común es adjetivo-nombre como en *personal computer*.

Algunos ejemplos de extractores son *Lexter*, que fue hecho para el francés y *Heid*, para el alemán. Las técnicas estadísticas se basan en una de las primicias de esta ciencia, es decir, que un suceso que ocurre frecuentemente es importante. Entonces, si una cadena ocurre muchas veces en un documento, debe ser importante para el documento y si este documento es especializado, lo más probable es que sea un término. Lo que hay que eliminar aquí son las palabras funcionales como las preposiciones, artículos, que son las más frecuentes en un documento, pero fuera de ellas, verbos, sustantivos y adjetivos, si ocurren frecuentemente en un documento ya son candidatos para ser términos. La ventaja de este enfoque, que es puramente estadístico, es que no importa el idioma que se esté trabajando, simplemente se trata de buscar palabras, así que pueden ser palabras del francés, alemán o inglés, y no hay que hacer grandes adaptaciones al algoritmo. Finalmente, vienen las técnicas sencillas en las cuales básicamente hay que combinar las dos anteriores.

Uno de los ejemplos fue desarrollado en Canadá. Lo que hace es que, si los términos ocurren en documentos especializados, se espera que en los que no lo son no ocurran, simplemente comparado un documento general con un documento especializado y las palabras diferentes entre estos dos documentos son los candidatos a términos y estos se usan como semillas para detectar más candidatos; éste funciona para francés, inglés, español e italiano. *TerMine* fue desarrollado en Inglaterra, y originalmente su objetivo era detectar los términos o candidatos a términos que aparecían en documentos del área

de lingüística. Es precisamente el *TerMine* en el que basamos este trabajo. El algoritmo a utilizar en *TerMine* para la extracción de términos se llama NC-value y es un método híbrido desarrollado primero para el inglés y para el área de medicina. Se divide en dos etapas, la primera parte es C-value; lo primero que se hace a un documento es identificar las partes de la oración, después la detección de candidatos de términos que son aquellos sintagmas que cumplan con ciertos patrones; la eliminación de candidatos por medio de una lista de paro, y la etapa estadística es simplemente ordenar estos candidatos con base en la frecuencia y longitud de los mismos sintagmas candidatos.

Esto es un texto de ejemplo del área de computación, utilizamos dos corpus para este desarrollo, uno es el *Corpus Lingüístico de Ingeniería* del Dr. Gerardo Sierra y el otro es el corpus de informática del español del Dr. Patrick Drouin y la Dra. Marie-Claude L'Homme, de la Universidad de Montreal.

Utilicé un etiquetado que se llama *TreeTagger*, desarrollado en Alemania; la ventaja de este etiquetado es que hace la averiguación. Esto se basa en técnicas de inteligencia artificial, una etapa de decisión y una etapa de entrenamiento. Si ustedes les dan un etiquetado que ya está validado por un experto, el sistema es capaz de aprender cuáles son las combinaciones de palabras para decidir qué es lo que va a trasladar cada uno. Como ya les mencionó Carlos, se trata de identificar cuál es un verbo, un adverbio, un nombre común. También existe uno que se llama *Freeling*, desarrollado en la Universidad de Cataluña y también detecta entidades nombradas, pero tiene un desempeño inferior al *TreeTagger* para etiquetar las partes de la oración.

Los patrones que detectamos para español estaban enfocados en el área de computación. Fue muy fácil, se dio más ocurrencia de un nombre común, un nombre propio o una palabra extranjera. Estos patrones fueron definidos porque los hice con base en el etiquetado que se hace con *TreeTagger*; por ejemplo *software*, aunque sea una palabra extranjera del español. Otro de los patrones característicos en español es *nombre común más adjetivo* y en ocasiones seguido de una preposición que normalmente es *de* y otro nombre común, como en este caso en el sistema operativo, un ejemplo sería *sistema central de procesamiento* o CPU. También otra de las reglas que es muy productiva es *nombre común más preposición* o *nombre común más nombre propio*, como podría ser alguna marca o, por ejemplo, *memoria RAM*, y una unidad de medida, por ejemplo, *banda ancha*. Otras reglas que no son tan productivas son *verbos*; el problema aquí es que no podemos identificar qué verbos son términos y cuáles no lo son. En el caso anterior de las reglas que les mencionaba, si encontramos un patrón con dos nombres repetidos, ya es muy probable que sean términos.

Posteriormente, basta definir si son del área tratada o no, pero en el caso de los verbos, si hay un verbo *ser, distribuir, tratar, trabajar*, lo que sea, no hay manera de ver si es cercano al área o no; otra regla que no es tan productiva es un *nombre común* y un *acrónimo* o únicamente el acrónimo, la relación significa 0 o 1 ocurrencia de esta categoría, en el caso del *IP* es un protocolo en internet. Otra que también es muy rara es *nombre común* o *preposición*, además de una combinación nombre con un adjetivo o adjetivo nombre común, éstas también detectan muchos términos. El problema es que traen mucha más basura que términos, es por eso que en el primer prototipo sólo se consideran estas construcciones.

Lo primero es detectar aquellos sintagmas que cumplen con los patrones que ya les mencioné, como *fábrica* que es un nombre común, *disco duro* que es un nombre común más un adjetivo, *sistema operativo*, el mismo patrón, *formatear* que sí es un término en el área de computación pero no está modificando a *computadora*. Ya que el programa detectó estos candidatos, lo que se hace con esta lista es pasarlos por una *lista de paro*. La *lista de paro* es un conjunto de palabras sin nombres comunes y adjetivos que no se espera

que aparezcan en computación; en documentos de otra área de conocimiento simplemente hay que crear la *lista de paro*. En computación hay un ejemplo de las palabras que aparecen en la *lista de paro*, ya vimos que éstas no aparecen en los términos de computación, por lo tanto, cualquier candidato que la contenga se elimina.

El problema de la *lista de paro* es que es dependiente del área de conocimiento. Ésta es para el área de computación; si lo queremos para otra cosa, hay que hacerla nuevamente y eso se hace simplemente con un conteo de palabras: las más frecuentes son las que se incluyen.

Esto fue la etapa lingüística, después viene la etapa estadística. En la etapa estadística lo que sigue es dar la frecuencia total de sintagmas en el corpus. Un evento que ocurre muchas veces debe ser importante. Otra variante es la frecuencia total de actividades del sintagma, pero como parte de sintagmas más largos. Por ejemplo, en las áreas de especialidad es común que a la *red neuronal artificial* se le cambie el nombre simplemente a *red neuronal* cuando el término real es *red neuronal artificial*; por lo tanto, si en un texto aparecen ambos, el resultado es negativo para el más corto, porque se puede asumir que es una simplificación de otro término; este factor es negativo. Otro factor es el número de los candidatos ya seleccionados que sean de mayor longitud, esto implicaría dependencia; un ejemplo sería *red de computadoras*, *red de computadoras inalámbricas* y *red de computadoras de largo alcance*, los tres son términos, pero como *red de computadoras* es el término más pequeño y cubre a los otros dos, se hace dependiente de ellos y se puede asumir que los dos más largos son variedades del término más corto, por lo tanto benefician al más pequeño. Finalmente, entre más largo sea el término, será más probable que se repita muchas veces.

Para la frecuencia de aparición, simplemente se hace el cálculo; por ejemplo, *sistema operativo* tiene un C-value, un potencial de ser un verdadero término, de cuatro, lo cual no les dice mucho; lo mismo ocurre para *sistema*, tomando en cuenta los factores que ya he comentado y su C-value. Ya he mencionado que había un sistema que únicamente conservaba la frecuencia de aparición del sintagma; si tomamos eso en cuenta, ése sería el orden de la probabilidad de que estos candidatos sean términos; el más probable de ser un término es *disco duro* y el menos probable es *partición*, pero si calculamos en C-value, se puede ver que *sistema* bajó hasta cuarto lugar de la lista de candidatos, y ¿por qué bajó? Porque *sistema* aparece dentro de *sistema operativo*, y si nuestro corpus solamente fueran estos cuatro renglones, se podría asumir que ésta es una simplificación del verdadero término. Es por eso que la salida de C-value es mejor cuando se ve en un gran volumen de texto, ya que se nota mejor el cambio.

La siguiente parte es el algoritmo NC-value. De lo que se trata aquí es que los términos no aparecen con un número arbitrario de palabras, sino que se genera una ventana en la que aparecen palabras de contexto. Por ejemplo, un candidato a término es *USB*; vamos a tomar una parte de 25 palabras, las palabras funcionales se eliminan, y *pequeño* y *dispositivo* son palabras de contexto; es muy probable que *pequeño dispositivo* se refiera a *USB*, por lo tanto estas dos son palabras de contexto y le van a ayudar en su potencial de ser término, es el mismo caso de este candidato a término, *dispositivo de almacenamiento*; si nosotros vemos en una ventana cinco palabras antes —los términos tienden a aparecer juntos y son palabras del mismo tema— un *USB* y *pequeño* sería palabra de contexto, *utiliza* no nos va a dar mucho pero *memoria RAM* sí.

El mismo caso es de *memoria flash* y la *capacidad de almacenamiento* que sería candidato; las palabras de contexto le van a ayudar. Las palabras de contexto son aquellas palabras que ocurren en la vecindad de un candidato a término que son sustantivos o adjetivos, o incluso verbos. Es muy común que en el contexto de candidatos como *disco duro* aparezcan palabras como *formatear*, *guardar* o *gigabyte*. Y con *sistema operativo* aparecen muy frecuentemente palabras como *instalar*, *grabar*; eso es lo que se utiliza para reordenar la lista, y de nuevo se le asigna un peso y a este candidato se le calcula otro potencial. No se observa mucha diferencia entre C-value y NC-value, aunque si decrecen, el orden sigue siendo el mismo, el objetivo aquí es que el orden deje los verdaderos términos, que estos queden en el tope de la vista para que el experto o la persona que vaya a ver esto no esté observando toda la salida, para que ahorre tiempo.

Hay que tomar en cuenta que esta lista puede ser de cientos o miles de candidatos; en un experimento que hicimos con cien mil palabras, por ejemplo, *usuario* fue el primer candidato; si lo ordenamos tanto por frecuencia como por aparición en C-value y NC-value, ocupa el primer lugar. Pero ustedes observan que *estación de trabajo*, que es un término de computación, si solamente lo ordenáramos por frecuencia, ocuparía el lugar 69 y con C-value llegó hasta el siete. Sí resulta ser mejor este método de ordenación. En el candidato *problema* es al revés, *problema* no es el verdadero término; si observamos sólo la frecuencia que es 119, es candidato, pero ya con estos dos factores bajó un poco, y *memoria flash* fue como el caso de *estación de trabajo* debido al cálculo, donde se consideraron su longitud y su frecuencia en otros candidatos. Al final de cuentas, el valor de C-value es mejor.

La salida del sistema es un documento XML, hay colores que marcan los candidatos a términos, los rojos son los que es más probable que sean términos, sin embargo, tiene errores. Por ejemplo, USB no lo detectó por un error de etiquetado, porque cuando pasamos el etiquetado de partes de la oración se consideró como un adverbio. Como les digo ésta es una de las salidas que da; otra es un documento HTML, que es simplemente una tabla con todos los candidatos ordenados; se trata de un documento separado por punto y comas.

Las conclusiones son: vimos que C-value puede ser una buena opción para la abstracción de términos en español, simplemente lo que se hace es adaptar las reglas lingüísticas y generar una lista de partes; sin embargo, hay errores en la salida, la precisión es más o menos de 60%. En el caso de texto libre, todavía nos falta hacer la evaluación en lo que sería la aplicación de este sistema que es la detección de términos o candidatos a términos dentro de contextos definitorios. Sólo como un dato, una persona se tardó más o menos tres horas en obtener la terminología de un documento de 2385 palabras, lo que sí es que hay que ocupar un poco más de programación, un poco más algoritmos, porque, para procesar un documento de 140 mil, el sistema actual resulta tardado. Es todo, gracias.

Gerardo Sierra: Muchísimas gracias. Hemos acabado la sesión número tres. Nos vemos mañana. Gracias. ■

LA ESTRUCTURA DE LAS OBLIGACIONES Y EL “COMMON GROUND” EN DIÁLOGOS PRÁCTICOS

LUIS PINEDA/IIMAS, UNAM
VARINIA ESTRADA/FFYL, UNAM
SERGIO CORIA/IIMAS, UNAM

Javier Cuétara: Vamos a empezar con la cuarta mesa: “Tecnologías del habla”. Habrá dos sobre este tema; ésta es la primera. Los tres trabajos que se presentarán ahora mismo se insertan dentro del proyecto DIME, que se desarrolla en el Instituto de Investigación de Matemáticas Aplicadas y Sistemas.

El primer trabajo hablará sobre *La estructura de las obligaciones y el “common ground” en diálogos prácticos*, es un trabajo en conjunto con Luis Pineda que es el coordinador del grupo y del proyecto DIME del departamento de Ciencias de la Computación del IIMAS, por Sergio Coria, el doctorando, también miembro del mismo grupo, y por Varinia Estrada, que es estudiante egresada del Colegio de Letras Hispánicas; tenemos un trabajo en conjunto entre lingüistas, computólogos, matemáticos, doctores en filosofía. Empezamos con este trabajo.

Varinia Estrada: Como dijo Javier Cuétara, este trabajo está dentro del proyecto DIME. El proyecto DIME, *Diálogos Inteligentes Multimodales en Español*, es coordinado por el doctor Luis Pineda. Comenzó en 1998 en el Instituto de Investigaciones Matemáticas Aplicadas a Sistemas en el departamento de Ciencias de la Computación.

El principal objetivo del proyecto DIME es la creación de Sistemas conversacionales para el español hablado en dominios especializados y el objetivo particular de este trabajo, que yo presento, es la creación de una teoría de los actos del habla que tenga una base empírica y la creación, también, de una metodología de anotación de actos del habla que sean correspondientes a la teoría.

La base empírica del proyecto DIME es el *Corpus DIME*, un corpus multimodal, donde tenemos 26 diálogos prácticos, en habla y video. La interacción es humano-humano, el dominio es el diseño de cocinas y se utilizó la interfaz CAD, que es un programa de diseño. El cuadro 1 muestra el programa de diseño.

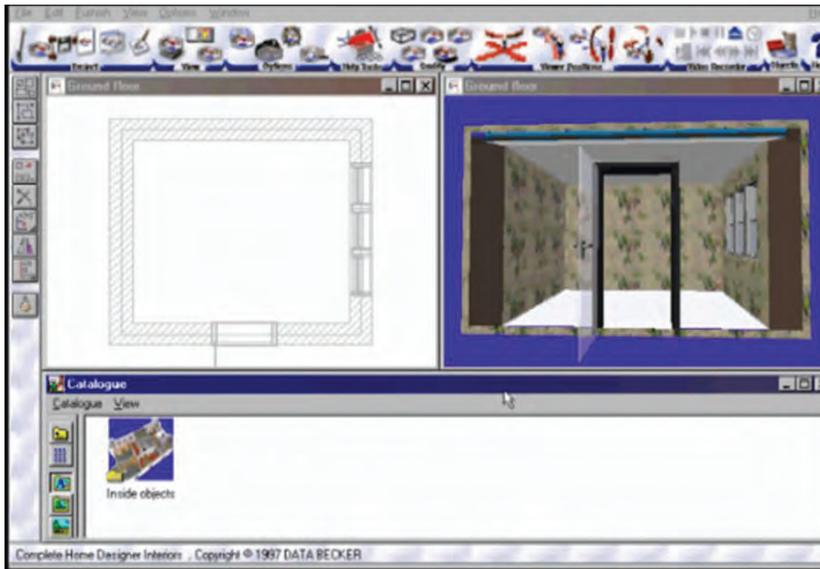
Como se ve en el programa, existen dos vistas distintas de un solo lugar, es la cocina que se diseña; un ejemplo de diálogo a este respecto sería el siguiente:

“¿Quieres que traiga o mueva un objeto a la cocina?”

“Sí quiero...”

Hay un usuario en el sistema y es quien le va diciendo al sistema qué es lo que tiene que hacer y conjuntamente van desarrollando la tarea. Éste es un ejemplo de lo que ocurre en los 26 diálogos del *Corpus DIME*.

El corpus, primero, es transcrito ortográficamente y etiquetado en enunciados; actualmente se etiqueta en todos estos niveles lingüísticos: alófonos, sílabas fonéticas, palabras, —de hecho estos niveles tienen que ver con los dos trabajos que se presentan más adelante—, también se etiquetan tonos con el modelo ITSINT, modalidad oracional que es una afirmación, una interrogativa y demás, partes de la oración, reparaciones del habla, y también etiquetamos actos del habla con el esquema DIME-DAMASL.



Cuadro 1. Imagen del Corpus DIME

Quiero hablar de este esquema, el esquema DIME-DAMSL para etiquetar actos del habla. DIME-DAMSL está basado en DAMSL que es *Dialogue Act Markup in Several Layers* que fue creado por Allen & Core. Este esquema tiene cuatro niveles, dimensiones de análisis para cada enunciado: el estado comunicativo, nivel de información, funciones hacia adelante, funciones hacia atrás.

El primero tiene que ver con asignar a la elocución si tuvo un valor comunicativo o no dentro de la conversación; si no lo tuvo, se etiqueta como abandonada, ininteligible o monólogo; si no tiene esa etiqueta, es que sí tuvo una aportación, entonces se puede etiquetar.

DIME-DAMSL utiliza los cuatro niveles de análisis de DAMSL y además agrega un set de actos que son acciones gráficas, porque nuestro corpus es multimodal; entonces, el usuario pide que se muevan objetos, que se quiten o que se pongan; todas esas acciones gráficas que están aquí, poner o quitar objetos, también vehículos, cómo quitar o cambiar de lugar, o un objeto. Nosotros además estamos agregando dos puntos, ésta es nuestra verdadera aportación: “los límites de transacción” y “obligaciones y *common ground*”.

Voy a hablar de estos dos puntos, la teoría en la que está fundamentado todo el esquema de etiquetación. Primero que nada, asumimos que los diálogos prácticos son conversaciones en las que conjuntamente se intenta resolver una tarea en un dominio específico. Esto se basa en la hipótesis de Allen; para empezar, la competencia es más sencilla que en la conversación general, y la hipótesis de la independencia del dominio tiene que ver con que la interpretación del lenguaje y el manejo del diálogo son independientes de la tarea concreta que se resuelve.

De igual forma, postulamos que los diálogos prácticos son secuencias de transacciones; la transacción es, de hecho, una unidad que nos permite hacer un análisis de la conversación, una unidad más completa, en la que hay dos fases: en una se especifica la intención, qué es lo que el usuario quiere, y en la segunda se satisface esa intención y el sistema realiza lo que el usuario solicitó.

Entonces tenemos esto: primero el usuario expresa, el sistema interpreta y pueden llevarse muchas elocuciones sin poder determinar exactamente qué es lo que el usuario quiere. En la satisfacción, el sistema va

a intentar satisfacer lo que quiso aquí el usuario y también pueden entrar en una posibilidad de volver a explicar, y así hasta que se logre el acuerdo; después de esto, la transacción estaría terminada.

Creemos que las transacciones construyen el perfil de los actos del habla y los actos del habla los definimos como acciones e intenciones que son expresadas en los enunciados. Entonces tenemos tres tipos de actos del habla: solicitudes de información, directivas de acción, respuestas, compromisos, afirmaciones, ofertas, etcétera.

En el esquema DIME-DAMSL tenemos definido nuestro inventario de tipos de actos del habla.

También encontramos que hay actos del habla que sirven para especificar las intenciones dentro de las transacciones, y actos del habla que ayudan a establecer durante la conversación el acuerdo mutuo.

Dentro de los actos del habla, los que especifican o satisfacen intenciones forman parte de la estructura de las obligaciones; mientras que los que establecen el acuerdo mutuo son parte de la estructura del “common ground”.

Las que especifican intenciones, las que pueden formar parte de obligaciones son directivas de acción, compromisos, donde se compromete a realizar algo, solicitudes de información, solicitudes que ya fueron aceptadas y demás; y las que satisfacen intenciones; por lo menos en nuestro corpus, son acciones gráficas —si yo quiero que mueva algo hay una disponibilidad para satisfacer lo que yo quería—, las respuestas, afirmaciones y demás. Estos serían los actos del habla dentro de la estructura de las obligaciones.

Dentro del “common ground” hay dos niveles distintos: uno que es el *acuerdo*, en donde se lleva a cabo si acepta, rechaza y en donde se puede poner también la conversación en espera a que algo más ocurra; y del otro lado, el otro nivel es el del *entendimiento* en donde están todos los *acknowledgements*, *backchannels*, las repeticiones; todos estos tienen que ver con cómo es que los que conversan establecen que se están entendiendo o no las señales del entendimiento.

Por lo general, el “common ground” se mantiene implícito; parece que con el simple acto de la conversación se da por hecho que nos estamos entendiendo; aun así, hay formas explícitas, y las explícitas por lo general refuerzan el “common ground” o hacen explícito que hubo una falla en él, y lo que hemos visto es que todas las fallas deben poderse resolver; de hecho, deben ser resueltas para poder continuar con la tarea que se estableció.

Ambas estructuras de las obligaciones del “common ground” van a hablar sobre convenciones sociales de la conducta conversacional, los típicos ejemplos son: si Ana pregunta, si uno pregunta, el otro responde, por lo menos te dice “no te voy a responder”; si Ana ofrece algo, Beto acepta o rechaza la cosa, y ambas estructuras no sólo están basadas en convenciones sociales fuertes sino que además son dependientes del contenido conceptual que se expresa. Esto es porque ambas estructuras son estructuras conversacionales; y eso es su verdadera misión.

Esto fue como un poco la teoría, pero nosotros tenemos que poder representar de alguna manera a la hora de realizar el análisis del corpus. Entonces vemos las relaciones que existen en los actos del habla como relaciones de cargos y abonos; y cada acto del habla contribuye como cargo o como abono dentro de la conversación en alguna de las estructuras, sea de las obligaciones o del “common ground”.

Para que una transacción esté terminada debe estar balanceada entre todos los cargos que vienen de su abono en ambas estructuras. Un ejemplo de transacción bien portada puede ser el siguiente:

ENUNCIADO	ACTOS DE HABLA
1. u: después <sil> me puedes poner <sil> el extractor de aire encima de la <sil> de la estufa	dir-acción
2. s: okey	acepta, compromiso
3. s: <acción-graf>	mueve-obj
4. ¿así está bien?	sol-inf
5. u: sí así está bien	respuesta, acepta

Cuadro 2. Actos del habla

Tenemos los enunciados, uno, dos, tres y cuatro; el primero es una directiva de acción; el segundo es una aceptación, de compromiso, porque va a realizar lo que le pidieron, la tercera realiza la acción gráfica; el cuarto es la recepción de información; y el quinto es una respuesta y una reacción. Pero lo que nosotros queremos es poder hacer un análisis un poco más completo. Por lo tanto tenemos los enunciados, las obligaciones con sus cargos y sus abonos, el "common ground" con sus dos niveles de acuerdo/entendimiento para cada uno y al final los actos del habla.

El etiquetado se da de la siguiente forma: la directiva de acción genera siempre un cargo en ambos lados. Las obligaciones, que tienen que estar de acuerdo a lo que se le pidió en el acuerdo, porque primero que nada el sistema tiene que aceptar lo que va a realizar. Entonces el "okey" representa que él acepta el acuerdo y un cargo, porque está comprometiéndose a realizarlo.

A la hora de mover el objeto, se abonan los dos cargos en las obligaciones, pero por otro lado en el acuerdo se cargan; como el estado de las cosas y el espacio virtual ha cambiado, el usuario tiene que aceptar los cambios, lo que se hizo. En "¿así está bien?" hay una solicitud de información donde debe haber una confirmación. En "sí, así está bien" está la respuesta para que acepte la acción que se realiza.

Ésta sería una transacción balanceada. El problema es que las cosas no son tan sencillas como parecen. Por ejemplo, si se le pide que ponga algo en la pared del fondo, y el sistema no sabe cuál es la pared del fondo, se tiene que entrar a resolver cuál es la pared del fondo y demás. Se pueden complicar muchísimo las transacciones, tanto la especificación como la satisfacción, ya que pueden seguir múltiples caminos.

Una transacción puede empezar con una oferta, con una aceptación de información, con una directiva del usuario, con una oferta del sistema, o, bien, puede empezar de muchas maneras distintas y además seguir caminos distintos.

En la satisfacción también puede el usuario no estar de acuerdo, por lo que se tiene que volver a especificar hasta que se logre la satisfacción, y, de igual forma, en cualquier momento de la conversación puede fallar el "common ground", puede haber algo que el otro no sabe qué es, por lo que se tenga que entrar a una situación específica en la que se tiene que resolver el problema para poder seguir con la tarea. Los límites de las transacciones a veces son un poco borrosos; no sabemos si hay transacciones embebidas. Un ejemplo sería el siguiente:

– Quiero que pongas la estufa, *pero antes* quiero que muevas el refrigerador para acá y luego eso para allá y todo esto para acá.

enunciado	obligaciones		"common ground"				Actos del habla
	ch	cr	acuerdo		entendimiento		
			ch	cr	ch	cr	
1. u: después <sil> me puedes poner <sil> el extractor de aire encima de la <sil> de la estufa	①		①				dir-accion
2. s: okey	②			①			compromiso, acepta
3. s: <mueve-obj>	② ①		③				mueve-obj
4. así está bien?	④		④				sol-info
5. u: sí así está bien	④			③ ④			respuesta. acepta

Cuadro 3. Estructuras diferentes

TURNO	enunciado	obligaciones		"common ground"				actos del habla	
		C	A	acuerdo		entendimiento		obligaciones	common ground
				C	A	C	A		
1s	¿quieres que traiga algún mueble a la cocina?			①					oferta
2u	sí	①			①			oferta	acepta
3	necesito una estufa	③		③				dir-acción	
4s	un segundo				③				acepta
5	éstos son los modelos de estufas que contamos estufas sencillas y estufas con alacenas laterales			⑤					opción- abierta
6u	mm <sil> voy a seleccionar esta estufa			⑥	⑤				acepta, afirma
7s	okey				⑥				acepta
8u	eh por favor lo necesito <sil> en la pared del fondo			⑧					afirma
9s	¿cuál es la pared del fondo?	⑨						sol-inf	espera
10u	a ver aquí	⑨		⑩				resp	afirma apt-zona
11s	¿ahí?	⑪				⑪		sol-inf	espera repetición
12u	sí	⑪				⑪		resp	acepta
13S	un segundo	⑬			⑩ ⑧			compromiso	acepta
14s	<agrega-objeto>	⑬ ③ ①		⑭					accion-graf
15	¿así está bien?	⑮		⑮				sol-inf	
16u	sí de momento sí	⑮			⑭ ⑮			resp	acepta

Cuadro 4. Ejemplo de estructuras

Se trataría, entonces, de una transacción con transacciones medidas, o una secuencia de transacciones más pequeñas; nosotros creemos que son transacciones embebidas. A veces el acuerdo se da de manera implícita. No es algo tan fácil.

Pero con todo y esta dificultad, hemos encontrado que hay recurrencias en lo que sucede, en las relaciones de los actos del habla y estas regularidades las hemos postulado como reglas en ambas estructuras conversacionales.

CARGO	ABONO	EN PARTICIPANTE
sol-info	respuesta	otro
dir-acción	acción	otro
compromiso	acción	mismo
oferta (aceptada)	acción	mismo

Cuadro 5. Estructura de las obligaciones

En la estructura de las obligaciones, encontramos los cargos, los abonos, mientras que en participante, encontramos quién tiene que abonarlo. Una solicitud de información siempre se va a abonar por respuesta por el otro participante; una directiva de acción por una acción por el otro actante; un compromiso con una acción que realiza el mismo que estableció el compromiso; y la oferta, que ha sido ya aceptada, se va a abonar con una acción, también, por el mismo que hizo la oferta.

En la estructura "common ground" también se establecen estas reglas; en el acuerdo hay una solicitud de información del tipo sí/no, puesto que necesita una respuesta. Una directiva de acción igual necesita ser aceptada o rechazada —una oferta, una acción abierta, una afirmación— de alguna manera tiene que ser aceptada o rechazada.

En el nivel del entendimiento unas señales se vienen a cumplir en el *acknowledgements*, en el *backchannels*. Cuando hay una señal de entendimiento o un perdón, se abona con el usuario o con el hablante, que es la persona que no ha entendido y se tiene que corregir.

Nuestras formas de anotación son dos, por lo menos para estos actos del habla: uno es para límites de transacción y relaciones de cargos y abonos; y en el otro etiquetamos los actos del habla para el esquema DIME-DAMSL.

El cuadro 6 muestra los límites de transacción, aquí están "inicio-fin", obligaciones, "common ground", e igual DIME-DAMSL.

1	A	B	C	D	E	COMMON GROUND				DIME-DAMSL		
						OBLIGACIONES	ACUERDO		ENTENDIMIENTO		DIME-DAMSL	
							CARGOS	ABONOS	CARGOS	ABONOS	CARGOS	ABONOS
2	ELOCUCIÓN	INICIO (I) O FIN (F) DE TRANS.	SEGM (0 / 1)									
3												
22	utt19: ahí está bien ?			19	18,11,1	19					accion-graf, sol-inf	accion-graf
23	utt20: u: hijole <sil> tal vez está un poco <sil> este separado de la estufa		21	20	19	20	19				resp, dir-accion	accepta-parte, visual, dir-accion
24	utt21: podrías juntarlo un poco más ?		21	20	19	20	19				resp, dir-accion	accepta-parte, visual, dir-accion
25	utt22: s: okey			22			20				compromiso	accepta
26	utt23: ahí está bien ?			23	22,20	23					accion-graf, sol-inf	accion-graf
27	utt24: u: bueno par-parece que está		24		23						resp	quiza, visual
28	utt25: corresponde al al dibujo		24		23						resp	quiza, visual
29	utt26: am aunque se ve un poco ahí		26									afirm, pt-zona
30	utt27: se alcanza a traslucir el la la lo que es la pared		26									afirm, pt-zona
31	utt28: pero no sé si es <sil> esté correcto eso		26									afirm, pt-zona
32	utt29: hijole <sil> <ruido>											exclamación
33	utt30: s: <no-vocal> perdón											sne
34	utt31: u: no		31				23					rechaza
35	utt32: creo que me equivoqué		31				23					rechaza
36	utt33: porque ten[go] tenemos que mover <sil> los <sil> <ruido> los objetos que acabas de bueno el objeto que acabada de colocar este		33	33		33					dir-accion	dir-accion
37	utt34: hay que desplazarlo hasta topar con la con la pared		33	33		33					dir-accion	dir-accion
38	utt35: eh igual con la <sil> con la estufa		33	33		33					dir-accion	dir-accion
39	utt36: okey necesito que quede libre el espacio debajo de la <sil> de la ventana		33	33		33					dir-accion	dir-accion
40	utt37: s: okey						33					accepta
41	utt38: un segundo			38								compromiso
42	utt39: ahí está bien ?			39	38,33	39					accion-graf, sol-inf	accion-graf

Cuadro 6. Límites de transacción y relaciones de cargos y abonos

El cuadro 7 muestra dónde etiquetamos los actos del habla:

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
2	ELOCUCIÓN	SEGMENTO	INICIO (I) O FIN (F) DE TRANSA CC. (1)	edo.com	edo.com	edo.com	nlv.inf	nlv.inf	nlv.inf	nlv.inf	dec	dec	dec	sol.inf	sol.inf	sol.inf	info.acc.fut
3				asistexp	mooc	abandonada	talea	admin-tarea	admin-com	otro	afirm	realfirm	otro	prep.s/n	prep.pro	prep.imp	nlv.accion
4	utt1: s: quieres que desplace o traiga algún objeto a la cocina ?		1				1		1								
5	utt2: u: mm bueno <ruido> <sil> por el momento						1										
6	utt3: ay es que no no recuerdo bien que era eso <ruido>	3					1									1	
7	utt4: <ruido> qué objeto es el que está	3					1									1	
8	utt5: qué objeto es éste	3					1	1								1	
9	utt6: ya no recuerdo cómo						1										
10	utt7: s: éste ?									1						1	
11	utt8: u: sí									1							
12	utt9: s: es <sil> una alacena inferior <sil> un gabinete						1				1						
13	utt10: u: okey						1										
14	utt11: entonces <sil> por el momento quisiera mover este esta alacena al lado de <sil> de la estufa <sil> que está aquí						1										
15	utt12: s: <no-vocal> <sil> * okey *									1							
16	utt13: u: + pero no sé						1										
17	utt14: s: quieres que mueva <sil> este objeto <sil> en esta posición ?									1					1		
18	utt15: u: mm bueno am <sil> tal cual está en la posición no									1							
19	utt16: tendría <sil> tendríamos que girarlo							1				1					
20	utt17: s: okey							1									
21	utt18: un segundo							1									
22	utt19: ahí está bien ?							1	1						1		
23	utt20: u: hijole <sil> tal vez está un poco		21														

Cuadro 7. Actos del habla DIME-DAMSL

Están a diferentes niveles: estado comunicativo, niveles de información, transacciones hacia delante y transacciones hacia atrás y a cada una de las locuciones se le asigna el valor del acto del habla que están teniendo.

Primero tuvimos que pasar por una especie de proceso de entrenamiento y etiquetación, por lo menos para saber si lo que estamos haciendo en cuanto a modelos es consistente conforme a nosotros mismos; después de etiquetar, obtuvimos muy buenos “kappas”. Con “kappas”, a grandes rasgos, quiero decir que los que etiquetamos estamos de acuerdo con lo que hace el modelo.

DIÁLOGO	CARGOS/ABONOS	DIME-DAMSL	NO. UTTS	NO. TRANSACCIONES
d01	✓	✓	116	6
d02	✓	✓	196	12
d03	✓	✓	168	9
d05	✓	x	118	14
d06	✓	✓	371	34
d10	✓	✓	100	9
d11	✓	✓	285	30
d12	✓	✓	117	9
d13	✓	✓	191	16
d14	✓	✓	137	10
d15	✓	x	90	8
d17	✓	x	237	22
d18	✓	1/2	216	8
d19	✓	✓	105	11
d20	✓	1/2	179	18
d21	✓	x	69	7
d22	✓	x	181	16
d23	✓	x	81	8
d25	✓	x	116	9
d26	✓	x	210	13
TOTAL	20	11	3283	269

Cuadro 8. Avances de etiquetación

Obtenemos unos resultados preliminares en los que el 80% de los actos del habla aparecen, ya que son los actos del habla que hemos documentado en la transacción como la definimos, tanto para la especificación como para la satisfacción. De alguna manera los métodos prácticos no son muy cooperativos, los actos del habla que más ocurren es el “acepta”.

El cuadro 8 presenta los avances de etiquetación:

Todavía no terminamos de etiquetar todos los diálogos, nos faltan unos cuantos. Eso es todo. Gracias.

Javier Cuétara: Ha sido muy interesante ver el proceso, cómo este trabajo ha ido madurando y evolucionando. Creo que merece que le hagamos algunas preguntas. ¿Alguien tiene alguna duda?

SECCIÓN DE PREGUNTAS

Pregunta 1: Dices que tu trabajo se concentra en la interacción humano-humano y después hablas de “usuario y sistema; ¿es lo mismo?

Varinia Estrada: No. En el corpus la interacción es humano-humano, pero el objetivo es modelar los diálogos prácticos para que el sistema sea un sistema y no un humano.

Pregunta 2: En este contexto la comunicación es más sencilla, ¿por qué?

Varinia Estrada: No exactamente. Dije que la competencia es más sencilla, pero, mejor dicho, es más fácil de modular. Que dos quieren hacer algo conjuntamente no va a ocurrir; es por eso que digo que son muy sencillos, lo que es más sencillo es la modelación.

Pregunta 3: Entonces sería más claro decir que es más sencilla con respecto a la conversación general. Y la otra es el estado de comunicación...

Varinia Estrada: El estado comunicativo.

Pregunta 4: No entendí eso. La idea de los diálogos prácticos es: si tú vas a comprar una hamburguesa, lo más común es que te la vendan ¿no es así?

Varinia: Y no vas a hablar de cualquier cosa, sino que la pides.

Pregunta 5: Es decir, hay un dominio altamente especificado, en el que el cliente es muy cooperativo, el vocabulario es muy restringido, etcétera, y son prácticas que hace la gente cuando se necesita. Entonces la idea está en que cuando haya sistemas van a estar destinados a este tipo de tareas; no va a haber sistemas que hablen en cualquier contexto como un ser humano.

Varinia Estrada: Hasta con un ser humano en los contextos no hablas de cualquier cosa, va a ser muy restringido; por eso se habla de temas sencillos de la conversación general, porque con una persona puedes hablar de lo que sea, lo que sucede en la conversación es tan complejo que uno también tiene que aceptar sus limitaciones y saber que no se puede modelar cualquier

sistema de conversación. Con respecto a lo del estado comunicativo, esta etiqueta es la que vamos a tener al principio, pues sólo etiquetas cuando la elocución no tuvo una aportación en la comunicación; un monólogo, ya que no tiene una aportación en un diálogo, porque se está hablando solo, se están diciendo cosas a uno mismo o en función de decir algo en una conversación.

Pregunta 6: Respecto al modelo que tomas, en el que sí aportaron dos aspectos, ¿qué efectos tiene en los actos del habla?

Varinia Estrada: Pusimos los límites de transacción: cómo se balancea la conversación en las dos funciones conversacionales. Y es a partir de los actos del habla.

Pregunta 7: En términos prácticos, si se basan en un nivel ya hecho, entonces las aportaciones que han realizado ¿cómo se ven? ¿En qué se reflejan?

Varinia Estrada: Se reflejan en la capacidad de moderar diálogos prácticos; de hecho, una forma de medirlo es, por ejemplo, en el acuerdo que se establece en los resultados que obtenemos, porque de hecho si pones a la gente a etiquetar y le explicas cómo hacerlo, la gente etiqueta cosas muy diversas, aunque tengan instrucciones comunes porque ya a la hora de ver un fenómeno práctico, resulta que ya no es así, entonces una forma de medir qué tan buena es una teoría, es qué tanto tú puedes entrenar gente, generar expertos para que tus teorías se reflejen; en ese sentido nosotros hemos logrado buenos resultados.

También el objetivo de la etiquetación es generar un sistema conversacional; digamos que se va a programar un manejador de diálogo que tenga que ver con todas estas cuestiones de la estructura de las conversaciones y más importante que la estructura de la conversación, es modelar la conversación. Nosotros decimos que se puede hacer independientemente del contenido conceptual. Es a partir también de esa estructura y quien sepa el programa lo va a hacer muy bien, considerando no solamente los actos del habla sino todo esto.

Todo esto es para un modelo computacional; en principio, este protocolo es de un sistema para que una máquina lo haga, si no lleva esa dirección entonces no es muy interesante para nosotros, ya que se trata de otro objetivo, no el que nosotros tenemos.

Pregunta 8: Las "kappas" las están usando nada más para comparar personas o ya tienen un programa. Quiero decir que la "kappa" es utilizada para ver qué tanto coinciden las personas. ¿Tienen un programa que etiquete y las comparan con todas, o solamente personas?

Varinia Estrada: No, etiquetamos personas, varias etiquetamos el mismo diálogo. Sergio es el experto, él saca las "kappas", pero la etiquetación es por varios humanos, por nosotros mismos en varias rondas de los mismos diálogos.

De hecho, los formatos de etiquetación están hechos en el sentido de que automáticamente es el dato de entrada en las "kappas", porque realmente desde un punto de vista práctico sobre todo cuando hay mucha fuente de información sí responde un programa para computarlo de manera eficiente.

Sergio Coria: El método “kappa” permite medir qué tanto están de acuerdo los etiquetadores por encima del límite del azar, es decir, si la etiquetación se hiciera de modo aleatorio, simplemente por suerte se podría contar con una cierta coincidencia en las etiquetaciones, pero cuando se quiere medir la consistencia de un modelo, de un esquema de etiquetación, es necesario medir esa consistencia por encima del límite de la suerte, del azar; entonces el método “kappa” hace un ajuste en estas coincidencias en la etiquetación tomando en cuenta la cantidad de etiquetas disponibles, es decir, la cantidad que podemos elegir y el número de casos que estamos etiquetando. Y como decía Varinia, esa etiquetación que estamos haciendo ha sido manual, pero al tener un modelo consistente podemos llegar a tener un etiquetador que lo haga de manera automática.

De hecho podríamos hacer un *speech* orientado a los actos del habla, pero se trata de algo muy complejo.

Pregunta 9: Precisamente a eso va mi pregunta, porque yo traté de evaluar mi etiquetador con “kappa”, puse a varias personas a etiquetar a mano y no logré hacerlo, por eso mi pregunta de si nada más eran personas o había un programa.

Varinia Estrada: Las “kappas” se utilizan en la teoría general en relaciones persona-persona, persona-máquina o máquina-máquina; es posible usarse para todos los casos, la cosa es que es complicada entre aparatos, pero realmente Sergio ha sido de gran colaboración porque ha trabajado muchísimo en cómo se da este proceso.

Javier Cuétara: ¿Alguien más quiere hacer otra pregunta?

Pregunta 10: Sólo tengo un comentario. Es correcto si el discurso que estás usando en la interacción comunicativa está limitado al hecho mismo del desarrollo, incluso hay una gran cantidad de datos comunicativos que no son estrictamente lingüísticos, que se dan por contextualización, por movimientos, por aseveraciones que no son textuales. A mí me parece bien y creo que el objetivo de esto es llegar a sacar reglas de actuación lingüística en contextos determinados, hasta ahí me parece correcto. Yo como lingüista seguro que aspiro a mucho más que a eso, esa es la parte interesante para poder así generalizar algo; pero en el momento en que esto se convierta en factores múltiples, el diagrama que acabas de presentar en una conversación no restringida es imposible, es imposible porque hay una parte que tú mencionaste que se lo impide, que por supuesto en su intención comunicativa. Pero es un buen inicio.

Varinia Estrada: Hay varias cosas. En este trabajo, lo que estamos postulando desde el punto de vista lingüístico es que la estructura del “common ground” de asimilaciones es una estructura lingüística a la par de la estructura sintáctica o semántica, que se construye con los actos del habla, y eso lo hace completamente lingüístico en lo general, y es cierto, claro que postular desde una estructura lingüística a nivel pragmático se vuelve complicado; pero de todas maneras, el hecho de que nosotros modelemos esto desde el punto de vista computacional, no quiere decir que los seres humanos usemos este tipo de estructuras cuando estamos hablando en una conversación normal. De hecho, de la misma manera que construimos una construcción sintáctica, podemos preguntarnos: ¿la estructura sintáctica es real? ¿Alguien la ha visto?

Pregunta 11: No. Al contrario, la estructura siempre está, lo único que es útil para la estructura es un marco de referencia porque la lengua, finalmente, como estructura, siempre es una teoría; la única manera en que la lengua realmente se comporta es en el uso, es la única forma, es decir, para mí como lingüista lo otro siempre será un magnífico marco de referencia pero nunca será la lengua; y eso es lo que durante muchísimos años hizo la gramática tradicional, hoy la considero un punto de partida, pero definitivamente, no consideraría un trabajo serio, de lingüística actual, sin tomar la lengua en uso.

Varinia Estrada: Esto es lengua en uso; sin embargo, nosotros estamos postulando estructuras de la lengua en uso, y, aparte, desde el punto de vista computacional, la estructura no es un artefacto de la teoría, no lo es; en la teoría computacional la estructura se construye en la máquina; esto es, se construye una estructura de datos, un objeto informacional, real, que está ahí presente, que además tiene una propiedad muy importante, que tal vez no sea necesario en estudios de lingüística clásica, que esas estructuras son causales porque los programas tienen que obedecer a la conducta conversacional; entonces esa estructura tiene que ser causal en el sentido en que permite interpretar el lenguaje y producirlo. Sí, la estructura es algo muy real, es el objeto que se va construyendo informacionalmente a lo largo del proceso de interpretación; y si bien así no nos interesa, la lingüística computacional por un lado pretende aplicar la tecnología computacional para apoyar el uso del lenguaje de manera amplia, en cualquier tipo de situaciones lingüísticas que podamos usar, que va desde el tratamiento de textos hasta la conversación humana. Pero por otro lado también tenemos interés de que esas teorías tengan una realidad desde un punto de vista científico, no sé si ustedes consideren la lingüística como una ciencia o no.

Pregunta 12: Es una ciencia.

Luis Pineda: Yo no estoy seguro que sea una estructura real de un artefacto mental.

Javier Cuétara: Tampoco podemos perder de vista desde dónde se marca esta propuesta, son diálogos prácticos en un contexto restringido y dentro de un proyecto de tecnologías del habla. El tema, como se dan cuenta, da para mucha discusión. ¿Algo más?

Luis Pineda: Quería platicarles que Varinia presentó esta plática en el seminario *The Speech pictures* del departamento de Ciencias de la Comunicación en la Universidad de Rochester, donde se reúnen todos los investigadores de ese departamento para presentar resultados sólidos de investigación, y fue muy bien recibida. Varinia estuvo allá para hacerlo y realmente estamos muy orgullosos del progreso. ■

LA SILABICACIÓN EN EL CORPUS *DIME* COMO FENÓMENO FONÉTICO DE VITAL IMPORTANCIA PARA LAS TECNOLOGÍAS DEL HABLA

ALEJANDRA ESPINOZA CRUZ
FFyL / IIMAS, UNAM

Javier Cuétara: El siguiente trabajo lo presenta Alejandra Espinoza, que es egresada también del Colegio de Letras Hispánicas y becaria del proyecto DIME; precisamente, este es un trabajo sobre su tesis de licenciatura, y ella nos presentará *La silabificación en el Corpus DIME como fenómeno fonético de vital importancia para las tecnologías del habla*.

Alejandra Espinoza: Buenas tardes. Es importante aclarar que este trabajo está en curso. Voy a hablar un poco del proyecto DIME, en donde se enmarca todo este trabajo. Su principal objetivo es el desarrollo de una teoría de la estructura conversacional en la aplicación a la creación de un sistema de administración de diálogos. El proyecto comenzó a desarrollarse en 1998-1999 en el IIMAS de la UNAM; es un proyecto multidisciplinario que está conformado entre otros recursos por dos corpus: el *Corpus DIME* y el *Corpus DIME-DAMS*.

En el primer corpus, se encuentran los mismos recursos lingüísticos y está constituido por 26 diálogos del habla espontánea, grabados con el modelo de cocinas. Es importante decir que estos recursos lingüísticos son aplicados para las tecnologías del habla que Llísterri ha dividido en dos: *síntesis del habla* y *reconocimiento de habla*. La *síntesis de habla* es la generación automática de habla a partir de una representación simbólica; en el *reconocimiento de habla*, la conversación es una representación simbólica.

El corpus que se utilizó para esta investigación fueron cinco diálogos que se etiquetaron o se transcribieron en tres niveles segmentables: el nivel T54, que es el nivel alofónico, el TP, que es el nivel ortográfico de palabras y el T, sílaba fonética. En total, de los cinco diálogos etiquetados, se obtuvieron 974 oraciones, de las cuales se descartaron 577 por ser muy ruidosas, presentar mucho *ok*, un monosílabo *sí/no* y obviamente oraciones que no presentan un fenómeno. El corpus útil con el que se trabajó es de 397 oraciones.

Es importante decir que el trabajo es empírico. La herramienta que se utilizó fue el *Speech Viewer*. El primer nivel es el nivel de palabra, el segundo nivel es el nivel de sílaba fonética no fonológica y el último nivel es el alofónico. Este trabajo se hizo con las 397 oraciones de los cinco diálogos del *Corpus DIME*.

La hipótesis de la que se parte en esta investigación es que la división silábica fonética difiere de la fonológica; en realidad, el sistema nos dice que tendríamos que dividir silábicamente como dice aquí [sju·dád de mé·xi·ko], pero lo que en realidad ocurre es [sju·dá·de·mé·xi·ko]. Esta sílaba queda libre: [dá] y la siguiente consonante forma sílaba con la siguiente letra: [de]. El análisis empírico de una muestra de habla puede arrojar información sobre los contextos fonéticos que provocan estos fenómenos, los cuales pueden ser formalizados en la definición de reglas fonéticas.

Uno de los objetivos principales con base en el estudio de este corpus es crear las reglas que permitan la sistematización del fenómeno, es decir, la silabificación para, posteriormente, aplicarla no sólo en síntesis del habla sino en reconocimiento del habla y comprobar qué tanto difiere la silabificación fonética de la sonora.

De acuerdo con el análisis que se hizo de este corpus base, encontramos las siguientes categorías: la formación de sílaba abierta; un ejemplo puede ser el siguiente: nosotros tenemos [ko-lo-kár és-ta] y lo que realmente ocurre es [ko-lo-ka-rés-ta], esto es la sílaba abierta [ka], como ya lo mencioné, es dejar la sílaba libre, que no esté trabada por una consonante y es una tendencia del español de México; igual aquí: [las a-la-sé-nas], lo que en realidad tenemos es [la-sa-la-sé-nas] otra vez la sílaba abierta. Ése es uno de los fenómenos que se encontraron.

La siguiente es la homologación de elementos idénticos, vocálicos o consonánticos. Vocálicos con [e] y [a] y consonánticos con [s] y [b]; el ejemplo está aquí: en [és-te es-tán-te], estos dos segmentos vocálicos idénticos se reducen a uno solo y es [es-tes-tán-te]; [las sí-yas], tenemos [la-sí-yas]; estos segmentos son idénticos y se reducen en un solo sonido.

Y las sinalefas, en este caso es [la es-tú-fa] y lo que ocurre es [laes-tú-fa]. Estos son los tres fenómenos que se encontraron con base en el estudio empírico del corpus.

Vamos con el primero, que es la sílaba abierta. Quilis la define de esta manera: la sílaba que termina en vocal; es decir, en el mismo núcleo silábico, recibe la denominación de *abierta*. El español muestra una clara tendencia a la sílaba abierta, por ejemplo [tjé-nes al-gún]. Este es un ejemplo real en *tienes algún*; lo que el sistema nos indica es [tjé-nes al-gún], y lo que realmente ocurre es [tjé-ne-sal-gún].

Y los resultados de la sílaba abierta son favorables y de todos los contextos encontrados los más significativos fueron los constituidos por sílaba abierta [sa], [ne], [re] y [se]. Por ejemplo: [ú-nos ar-má-rjos] aquí está [ú-no-sar-má-rjos]; *en esta pared* en vez de [en és-ta pa-réd] es [e-nés-ta pa-réd], [mo-bér en] igual [mo-bé-ren] y [tjé-nes en] igual [tjé-ne-sen], esa es la abierta, y se hicieron fichas para analizar los datos. Tenemos una columna corresponde al fenómeno, la regla, las apariciones totales en el corpus y las apariciones que tuvo el fenómeno, la transcripción alofónica, el T54, la transcripción ortográfica, el número del diálogo en el que se encontró y si lo emitió el usuario o el sistema.

El total de apariciones fueron 42 fichas. Con el fenómeno en el corpus útil fueron 30, y el porcentaje del fenómeno de sílaba abierta fue 71%.

En cuanto al fenómeno de homologación tenemos que es el fenómeno de la fusión de dos sonidos idénticos en contacto y es importante decir que no respeta el límite de palabra. Un ejemplo de vocálico es [de e-le-xír] y estos dos segmentos quedan [de-le-xír]; y consonánticos, [las sí-yas] queda [la-sí-yas]. Este es un ejemplo también de homologación: [es-te es-tán-te ke-da a-í] igual [es-tes-tán] en una sola sílaba.

La homologación también fue favorable. La homologación consonántica con [s] es del 100%, ocurre 100% en el corpus; segmentos vocálicos con [a], 32%, y segmentos vocálicos con [e], 29%. Este es un ejemplo de homologación consonántica de [s], que apareció veinticinco veces y siempre ocurre el fenómeno de homologación consonántica.

La sinalefa: la última vocal de una palabra se une con la vocal inicial de la palabra siguiente; [fre-ga-dé-ro a-kí] y lo que tenemos en una sola sílaba es [fre-ga-de-roa-kí]; [xí-ra en], que da [xí-raen], no respeta el límite de palabra. Los resultados de la sinalefa: [ea] 37%, [ao] 30% y [ae] 11%.

Con base en el análisis de todos estos fenómenos y tratando de hacer una sistematización de todo esto para poder crear reglas que puedan ser implementadas en las tecnologías del habla, se formuló lo siguiente: la primera es formación de sílaba abierta cuando una consonante se encuentra entre dos vocales, la consonante se agrupa o forma sílaba con la vocal siguiente, no respeta límite de palabra.

La segunda regla: reducción de segmentos idénticos vocálicos cuando la última sílaba de una palabra termina en una vocal y la primera sílaba de la siguiente es la misma vocal; éstas se reducen a una sola vocal, y la vocal resultante forma sílaba con la primera palabra. La homologación en consonánticos se da cuando la última sílaba de una palabra termina en consonante y la primera sílaba de la siguiente es la misma consonante; ésta se reduce a una sola y la consonante resultante forma sílaba con la segunda palabra.

La sinalefa se presenta cuando una sílaba termina en vocal y la siguiente sílaba comienza en otra vocal; éstas se agrupan para formar una sola sílaba con la primera, el caso es [la es·tú·fa] da [laes·tú·fa].

Esto sólo es un poco de teoría; según los porcentajes que Guerra nos presenta acerca de la frecuencia de los tipos de la estructura silábica, la sílaba abierta es la que mayor porcentaje tiene en la lengua hablada y nosotros comprobamos esto, ya que en el corpus la sílaba abierta sacó 71%. Gracias.

Javier Cuétara: Muchas gracias. Se ha trabajado mucho en esto y la investigación está en proceso, podemos hacerle algunas preguntas.

SECCIÓN DE PREGUNTAS

Luis Pineda: Ante todo felicitarte, yo creo que desde el principio mostraste tu deseo de superar cuanta adversidad se te enfrentara, mereces un aplauso. Ahora, cuando veo todos tus resultados, mientras estaba yo observando me acordé de Navarro Tomás; él encontró todo esto sin la utilización de herramientas computacionales. Esto es preocupante porque significaría que después de este heroico trabajo no haces aportación alguna. Entonces mi primera pregunta es: ¿no encontraste, por ejemplo, entre vocales homófonas, un alargamiento? ¿Se trata solamente de la reducción a una sola vocal?

Alejandra Espinoza: De hecho, creo que la aportación principal es la aplicación a las tecnologías del habla.

Luis Pineda: Es comprobar lo mismo que Navarro Tomás con otros sistemas. Lo que yo estoy tratando de pensar es que otro sistema, indiscutiblemente más fino y más exacto que el oído humano, nos daría algo, por eso viene la pregunta de si efectivamente a la hora que mides tu etiqueta es igual a la de una vocal simple.

Alejandra Espinoza: En primer lugar, no se trata de refutar a Navarro Tomás, de hecho demuestra que los resultados son consistentes, pero por otro lado las estadísticas no son iguales porque Navarro Tomás con todas las cosas que tenía, y su buen oído, tiene resultados diferentes que los de nosotros con el corpus.

Luis Pineda: Navarro Tomás no hace estadísticas en esa época, pero suponíamos que íbamos a encontrar algunos alargamientos; de hecho, en clase lo decimos. Eso ya es una aportación. Sabemos que hay que decir que en la técnica verdadera se reduce a una vocal.

Alejandra Espinoza: Eso tiene que ver más con el habla cuidada.

Javier Cuétara: Hay que recordar que, como dijo Alejandra, esto no es una aportación. Es comprobar empíricamente, con un corpus con una aplicación en un ámbito específico, lo que corresponde a una sistematicidad de la lengua española, el contraste

en un sistema fonológico y en uno fonético. Lo interesante es que Alejandra ha estado estudiando los rasgos fundamentales: cuándo se cumple y cuándo no. Lo que hace es proponer una serie de reglas para aplicarlas a las tecnologías del habla.

Luis Pineda: Claro, es un poco lo que hizo de alguna manera Navarro Tomás. Lo que yo quería destacar es cómo desde un punto de vista lingüístico sí hay una aportación, sí hay un avance: de vocal-vocal no se presenta ningún alargamiento, y lo puedes decir ya casi como una regla, eso es una aportación y a mí lo único que me encantaría encontrar, que seguramente encontraste en los antecedentes de tu trabajo, es que de alguna manera estamos confirmando y afinando un trabajo tan viejo como el de Navarro Tomás, ni siquiera Quilis.

Javier Cuétara: Bueno, ¿quién sería Quilis sin Navarro Tomás?

Luis Pineda: Claro. Eso es lo que me parece muy valioso.

Javier Cuétara: El *Manual de pronunciación* es de 1918.

Alejandra Espinoza: Una cosa que nos pareció interesante —primero que nada— es que la etiquetación de corpus, la no coincidente con la sílaba esperada fuera de contexto y la sílaba fonética era realmente muy notable, y es algo que nos sorprendió realmente. Ahora, por otro lado, había muy poca literatura al respecto de cómo se comporta la sílaba en la lengua hablada, la literatura que se puede encontrar es muy limitada y de hecho confirmaciones empíricas, porque todo eso se hace para experimentar. Lo que se empezó es un trabajo empírico que realmente necesitaba comprobación. Por otro lado, también se hizo un trabajo de computación, que les interesará saber, que es cómo hacer reglas para su aplicación en las tecnologías del habla.

Javier Cuétara: Correcto. Muchas gracias, Alejandra. ■

FENÓMENOS DE SÍNCOPA EN EL CORPUS *DIME* PARA SU INCLUSIÓN EN EL RECONOCEDOR DE HABLA DIMEX

ANA CEBALLOS
FFyL / IIMAS, UNAM

Javier Cuétara: Otra propuesta interesante es la de Ana Ceballos; dentro del mismo marco, en el mismo proyecto, con conceptos similares, presenta otro tipo de cambio fonético que son fenómenos de síncopa según su análisis en el *Corpus DIME* para considerar su inclusión en un reconocedor de habla.

Ana Ceballos: Buenas tardes. Yo voy a presentar los fenómenos de síncopa. Esta investigación es parte del proyecto DIME, que es habla espontánea. La base del reconocimiento de habla es la pronunciación, que es un listado léxico de palabras que es posible que reconozca un sistema como el que ya se ha presentado.

Es importante mencionar que en el habla que encontramos ocurren muchísimas modificaciones y, por eso, un diccionario de pronunciación debe tener para cada vocablo tantas realizaciones como sea posible, lo cual hace muchísimo más exacto el reconocimiento de voz, ya que, como hemos visto en muchos estudios, cuando estamos interactuando con la computadora, se dicen muchas cosas mal, hablan muy rápido, hay ruido. Entonces, es importante que todas las variaciones fonéticas que pudieran ocurrir, dependiendo de un tipo de hablante, las tuviera un diccionario de pronunciación y fuera infalible. Ya hay varios diccionarios en este proyecto, el que se va a enriquecer es el del corpus creado por Pérez Pavón en el 2006 que utilizó como base el *Corpus DIMEX 100*, que es de habla controlada.

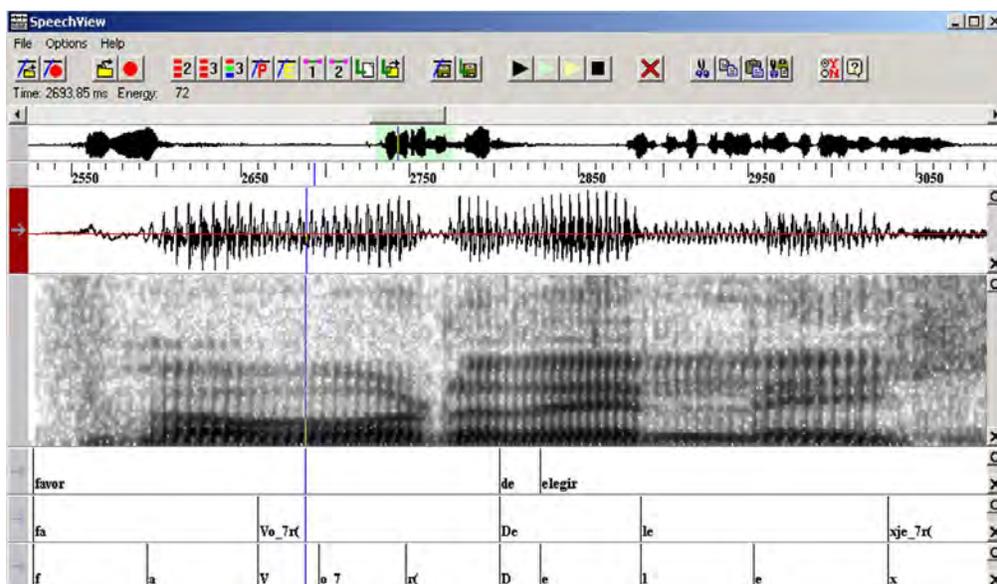
Del corpus base fueron seleccionados cinco diálogos que contenían entre 92 y 210 elocuciones; esto fue transcrito en los niveles de alófonos, palabras y sílabas.

En el cuadro 1, encontramos la herramienta con la cual se obtuvieron los datos. Con esta herramienta se trabaja con un oscilograma, un espectrograma y con el audio; también tiene una curva entonativa, es decir, que realmente no es como pura percepción auditiva, sino que se pueden buscar las pérdidas en estas dos partes de la herramienta para no equivocarnos, porque muchas veces lo que hacemos nosotros es reconstruir lo que se quiso decir, lo que provoca que pasen muchos fenómenos desapercibidos.

El objetivo es darle más entradas y formas de pronunciación al diccionario de pronunciación, enriquecerlo, y que sea más exacto que el diccionario que ya se tiene.

La hipótesis es que este análisis empírico va a mostrarnos muchísimos fenómenos, pero vamos a tomar los que sean sistemáticos para poder crear reglas contextuales que puedan ser utilizadas en este diccionario de pronunciación. En el corpus final en el que trabajé hubo que descartar muchísimas oraciones ya contaminadas que están mal segmentadas, o que no son relevantes, como monosílabos, para fenómenos en general, y el resultado fueron 828 oraciones de las que se descartaron 157; quedaron 671 oraciones y de ellas se obtuvieron 451 ejemplos de fenómenos.

Se encontraron muchos fenómenos, los más comunes fueron elisiones, aperturas y cierres vocálicos, pero la elisión fue la que arrojó un mayor porcentaje. Entonces se decidió que ése era el hecho en el que nos enfocaríamos. Para los datos no tomamos en cuenta la silabificación que se está trabajando en la otra investigación.



Cuadro 1. Segmentación

Para la parte de elisiones se tomó en cuenta la posición de las pérdidas; la mayoría ocurría en inicio absoluto o después de pausa, igual que las apócope, ocurrían al final o ante pausa, y las síncopas al interior de palabra.

Hay que tener en cuenta que el vocabulario es muy específico porque el dominio del corpus es específico y que el tratamiento podría ser fonético o léxico. Para tomar como regla un fenómeno tenía que ocurrir por lo menos cinco veces y ser pronunciado por varios hablantes.

En todos los casos de elisión, la apócope y la aféresis resultaron muy poco frecuentes y los de síncopa ocurrieron en el 71% de los casos de elisión, por lo que decidí enfocarme a la síncopa para mostrar reglas. De la aféresis lo más importante fue la pérdida de la e inicial ante consonantes alveolares; por ejemplo, antes se dio en un 20%, tenemos el ejemplo de *espera* por [spé-ra]. La apócope fue también muy poco frecuente y se encontró en vocales postónicas, esto sólo como un dato.

Hubo muchos contextos que presentaron síncopa; los catalogué en cuatro categorías: las síncopas de sílaba, las síncopas de consonantes intervocálicas, las síncopas de consonantes en posición implosiva y las de vocales en diptongos. La síncopa de sílaba tuvo una frecuencia de 32% y ocurrió sólo en algunas palabras; por ejemplo para la sílaba *-des* en *puedes* se pronunciaba [pués]; luego, en un 52% de los casos, para la sílaba *-res* en *quieres* pasó a ser [quiés] y para la sílaba *-ses-* como se ve en varias conjugaciones del verbo *necesitar*, *necesito* se decía [nesíto], lo que ocurrió en un 20% de los casos. Esto significa que el tratamiento de la síncopa de sílaba va a ser léxico para estas palabras y que hay que crearles nuevas realizaciones en este diccionario.

Ahora, las consonantes intervocálicas —un fenómeno muy trabajado— coincidieron en muchos casos. Encontramos, entre, las intervocálicas incluidas, las consonantes que participan con líquidas, como el elemento *bl*. El conjunto *br* no ocurría, pero *bl* sí.

La *b* simple intervocálica se perdió en 20% de los casos. *Favor* se decía [faór], en las intervocálicas había un proceso de relajamiento, pérdida total, o, bien, una sonorización. En el segmento *bl* también se perdió la

b y en vez de *problema* quedaba [prolema]. En el 18% de los casos y en tres de los seis hablantes para la *r* intervocálica no hubo nada, fue muy poca la frecuencia de cambio; la *r* se mantiene casi siempre en los contextos, no tiende a perderse, ni en intervocálica, ni en posición implosiva. Y que más porcentaje aportaron fueron los segmentos *tr*, *fr* y *dr* intervocálico.

La posición implosiva es la que más porcentajes aporta; suelen perderse más las consonantes en esta posición. También la teoría maneja que se pierden fácilmente las *e* intervocálicas y eso no ocurrió, por lo menos en esta muestra, pero sí con las implosivas. La implosiva se perdió en un 20% de los casos; en el segmento *ns* trabado paso a ser sólo *s* y entonces en *tridimensional* tenemos [trídimensionál]; para la *n* final o ante pausa ocurrió en el 11.5% de los casos, también es un porcentaje bajo; para la *VnIV* ocurrió en un 12% de los casos y esto ocurría en híbridos de palabras como en *son las* pasaba a ser [sólas]. También vemos que son consonantes alveolares, así que son fenómenos de asibilación. Para la *r*, como decía, es casi nula su pérdida, la *r* se mantiene muy bien. La *s* en posición implosiva sólo se perdió en esta situación, pasó por procesos de sonorización, igual que las velares y líquidas, se perdió ante la *l*, en el 44% de los casos, que es de los porcentajes más altos que obtuve, en *todas las sillas* daba [todala síllas]. La que presentó porcentaje más alto fue la *k* ante *s*, sin importar si seguía consonante o vocal, tuvo el 60% de los casos y también en muy pocas palabras, pero se perdía en casi todos los hablantes más de dos veces. En la palabra *extractor* se pronuncia [estrátór] y en la palabra *exacto* se pronuncia [esáto]. La pérdida implosiva de *d* se dio en 22% de los casos, pero sólo en posición final donde *pared* pasó a ser [paré].

Para la síncope de vocales, con un porcentaje más amplio el diptongo *ie* tuvo el 17% de los casos en palabras muy distintas y en seis de los siete hablantes, lo cual también nos habla de que es un tema importante. El diptongo *ie* pasa a ser *i* o *e* tónica en igualdad de circunstancias, en *bien* podríamos tener [bín] o [bén], esto ocurre en palabras muy frecuentes en el corpus como también son frecuentes en la lengua.

A nivel de léxico, la palabra *okey* representa una pérdida de *i* y queda [oké]; la palabra *quieres* tiene las realizaciones [kíres], [kíre], [kíes]; y *refrigerador* puede perder todas sus intervocálicas.

Como conclusiones generales, encontramos que son muchos los fenómenos que ocurren, pero para crear reglas es necesario que los fenómenos sean sistemáticos y que los porcentajes realmente ameriten la creación de reglas. Estos reconocedores tienden a enfocarse en situaciones conversacionales concretas, pero no todos los hablantes son iguales, así que es necesario que haya gran cantidad de realizaciones para eso y, por tanto, tenemos que cuidar que los fenómenos no sean exclusivos de una persona o de los errores que puede tener un hablante que hable muy rápido o mal, y también que sean exclusivos de una palabra; es decir, que solamente si tenemos, como ya vimos, un [kíes] en un determinado contexto y de una determinada palabra, no vamos a crear una regla para aplicarla a todas las demás.

La frecuencia de las palabras en la lengua es importante, y en el corpus las palabras que más ocurrían son las palabras que presentaron más fenómenos, ya que son las que están en uso constante en la lengua y los porcentajes más significativos son los que deben ser considerados para las reglas fonéticas.

Las reglas creadas al final, tomadas de los porcentajes más relevantes, son: para *ns* hay una pérdida de *n* ante vocal ($n \rightarrow 0 / _sV$); para *s* ante *l* o *r* va a haber una pérdida también ($s \rightarrow 0 / \{ _l, _r \}$); va a haber pérdida de *k* ante *s* trabada implosiva ($k \rightarrow 0 / _s$); tal vez crearemos alguna regla para apartar la *r* en el segmento *fr* para una posible realización $VfrV \rightarrow VfV$; la *b* intervocálica tiende a desaparecer ($VbV \rightarrow VV$).

El tratamiento que se hace del léxico puede ser aplicado a más contextos, pero en este caso son más específicas de este dominio y de este proyecto. Por ejemplo, las palabras: *puedes*, *quieres*, *okey*, *favor* y *pared*, que son de alta frecuencia y de varias realizaciones, deben ser incluidas para que el reconocimiento sea más eficiente.

Eso es todo y muchas gracias.

Javier Cuétara: Podemos hacer algunas preguntas.

SECCIÓN DE PREGUNTAS

Pregunta 1: Mi duda es qué tanto está afectando el hecho de pasar las frecuencias por el micrófono, porque yo entiendo los fenómenos de las síncopas como el de los veracruzanos que se comen las s. Mi pregunta es si tienen en cuenta la influencia del micrófono con la captura de los datos.

Ana Ceballos: Mencioné que en un proceso de comunicación humano-humano, en cualquier ambiente, siempre hay un proceso de reconstrucción; tal vez esté haciendo gran cantidad pérdidas y síncopas de cualquier tipo, pero yo participo en un código, entonces yo voy a ir interpretando, a menos que sean muy obvias como las pérdidas de los costeños.

Javier Cuétara: Ana, sabes que para nosotros es muy evidente que un costeño aspira o pierde ciertas consonantes, pero para ellos no.

Pregunta 2: Pero las síncopas recuerdan muy bien la historia del español.

Ana Ceballos: Además, así ha evolucionado la lengua.

Javier Cuétara: Quiero decirles que es verdaderamente increíble darse cuenta de que hacemos más cambios de los que nos percatamos, pero puedes tener cierta razón en que la calidad del audio quizás no es tan buena por las condiciones, pero, de alguna manera, se está simulando la situación que puede ocurrir entre un usuario y un sistema.

Pregunta 3: Para darte el ejemplo, yo intentaba pasar el lenguaje de inglés. He presentado pruebas en inglés y las cintas siempre dan dolor de cabeza. Cuando a mí me hicieron un examen oral, para mí era más fácil escuchar de mi interlocutor; una cinta siempre me costaba mucho trabajo. Yo supongo que sí hay cierta constante, porque por ejemplo hablar de vocales es algo simple, ya que todas tienen más o menos un timbre, pero ya detectar sonidos como /k/ que son de golpeo, creo que puede ser muy difícil, ver dónde cambió la frecuencia o si están entre dos frecuencias; casi tendrías que meterte a analizar la onda o que tus micrófonos fueran muy especiales. En todo caso lo que puedo comentar es que se debe poner más cuidado en el audio, ya que no es lo mismo que escuchen un concierto en la radio, o en una cámara, puesto que cambian mucho las frecuencias.

Ana Ceballos: Mi intuición es que realmente el ruido del micrófono no es un factor importante. Si hubiera algún inconveniente más bien sería por ruido en la grabación.

Pregunta 4: Cuando dices que hay desaparición de la n , ¿no observaste ningún índice de nasalización?

Ana Ceballos: Sólo tomé las pérdidas absolutas, pero todos los fenómenos pasaron también por los procesos de debilitamiento.

Javier Cuétara: Algo que me llama muchísimo la atención, también, es la pérdida de la vocal e en inicio absoluto de un enunciado. ¿Algún otro comentario para Ana?

Pregunta 5: Este tipo de trabajos está enfocado a un área muy específica para enriquecer y agrandar el diccionario de pronunciación y, a partir de esto, podemos llegar a postulados fonéticos como a los que llegó ella.

Ana Ceballos: Lingüísticamente hubo muchos fenómenos que sí respetaron la teoría, como estos casos de la pérdida de e, d en posición intervocálica.

Javier Cuétara: Muchas gracias Ana, ahora vamos a cerrar la sesión. Mañana tendremos la quinta mesa de sintaxis y semántica a las 10:30 en el salón de actos y una segunda mesa sobre tecnologías del habla. Les agradezco su presencia. ■

DETECCIÓN Y CORRECCIÓN DE ASOCIACIÓN SINTÁCTICO-SEMÁNTICA BASADA EN LA WEB

SOFÍA GALICIA HARO
FACULTAD DE CIENCIAS, UNAM

Jeannette Reynoso: Vamos a dar inicio a esta segunda jornada del Tercer Coloquio de Lingüística Computacional en la UNAM. Esta mesa es bastante extensa y presenta cinco trabajos dedicados al área de semántica y sintaxis. Nuestro primer trabajo va a ser presentado por la doctora Sofía Natalia Galicia Haro, doctora por el Instituto Politécnico Nacional, especialista en adquisición de conocimiento léxico y análisis sintáctico. Actualmente la doctora Galicia está con nosotros en esta universidad en la Facultad de Ciencias. Ella nos va a presentar un trabajo titulado *Detección y corrección de asociación sintáctico-semántica basada en la web*.

Sofía Galicia: El título de este trabajo es *Detección y corrección de asociación sintáctico-semántica basada en la web* y la idea es presentarles aquí dos métodos que hemos trabajado utilizando precisamente la web como un gran corpus. Entonces, les voy a hablar de la web como un corpus y precisamente de esos dos métodos: uno es detección y corrección de una palabra anómala en un contexto, y otra es desambiguación de uso de frases preposicionales. El uso de colecciones grandes de texto es muy común en la lexicografía para construir diccionarios; tenemos, por ejemplo, el *Diccionario del español usual de México*, cuyo corpus se construyó por los años setenta.

Pero en el procesamiento de lenguaje natural por computadora también es común utilizar grandes textos y desde los años sesenta está, por ejemplo, el *Brown Corpus* o el *British National Corpus*; la idea en ellos es que tienen una página en la que albergan ciertas herramientas que permiten el uso de la web como un corpus.

El problema de estas colecciones de textos es que, estadísticamente, son insuficientes para encontrar los resultados que queremos en el procesamiento de lenguaje natural por computadora, y por ejemplo, para resultados confiables no lo encontramos en textos de periódicos o recopilaciones. Entonces, una solución ha sido precisamente usar Internet como el corpus más grande que se haya utilizado, y bueno, conocemos varios buscadores; verbigracia, *Google*.

Utilizar la web como un corpus tiene un enorme valor potencial como un recurso lingüístico; la podemos utilizar para muy diferentes tareas. Es la más grande recopilación, se actualiza y amplía constantemente, aunque, claro, el problema también es que las páginas que aparezcan hoy probablemente ya no aparezcan mañana, pero cada día aumentan esas páginas. Hay una amplia cobertura de dominio de todos los temas y creo que la mayor ventaja para los que queremos utilizarlos es que no tiene ningún costo. Si nos referimos a corpus que han sido compilados para construir diccionarios, no siempre el acceso es sencillo. A estos muchas veces no se puede tener acceso, en cambio el Internet lo podemos usar todos.

El cuadro 1 presenta una recopilación bastante fidedigna en el año 2001 y corresponde a un coloquio que se hizo precisamente sobre considerar la web como un corpus; los números aparecen en cantidad de

palabras. Si nosotros lo vemos en los últimos renglones para el español son 2 658 000 palabras; para el francés, en el año 2001, más de 3 800 000 palabras; para el inglés, veinte veces más que para el francés en esa época. En 2005 decían que el crecimiento era de 7.3 millones de páginas por día y en enero del 2005 decían que había once mil quinientos millones de páginas; la cantidad de textos que se tienen allí es muy grande.

Welsh	14,993,000
Lithuaniam	35,426,000
Basque	55,340,000
Latin	55,943,000
Turkish	187,356,000
Catalan	203,592,000
Finnish	326,379,000
Czech	520,181,000
Norwegian	609,934,000
Portuguese	1,333,664,000
Italian	1,845,026,000
Spanish	2,658,631,000
French	3,836,874,000

Cuadro 1. Web AltaVista en 2001

Un problema de la web para utilizarla como un corpus son los errores que tiene, ya que cualquier persona tiene acceso a ella. Podemos mencionar, precisamente, algunos errores: “pienso de que” aparece 19 700 veces, pero “pienso que” se encuentra más de un 1 400 000; “atravezar” 34 500, “atravesar” más de 1 200 000 veces. Existen problemas, está mal escrito; lo que podemos ver es que la cantidad de veces en que aparece la forma correcta es mucho más grande que las incorrectas. Eso daría una mayor confiabilidad.



Cuadro 2. Ejemplo de búsqueda

Otra cosa que ha pasado con los buscadores es que incorporan herramientas; en este caso, poner un asterisco en Google permite obtener contextos. En el cuadro 2 estoy presentando tres ejemplos: lo que está mostrando en negritas son las frases fijas “a fin de cuentas”. El asterisco permite traer desde una palabra, un signo de puntuación, hasta frases. Es una ventaja para los que queremos utilizar la web como un corpus.

Se ha utilizado este tipo de recursos desde el 2000-2001. Se han empleado para extraer patrones lingüísticos, traducción automática, aprendizaje de ontologías y sobre todo en búsqueda de respuestas.

En este caso, lo que quiero presentar son dos trabajos: el primero, para detección y corrección de una palabra anómala; la idea en general es que una palabra es incorrecta en un contexto porque rompe los vínculos sintáctico-semánticos; un ejemplo sería “pasear por el centro histórico”, y la palabra errónea, en lugar de *histórico*, “pasear por el centro *histérico*”. Entonces, podemos utilizar la web para corregir ese error. Otro trabajo es la desambiguación de uso de frases preposicionales del tipo preposición-pronominal-preposición, que son más o menos fijas: “a fin de”, “con motivo de”, y vemos, por ejemplo, “a fin de año”. Si utilizamos “a fin de” en ese contexto de “a fin de año haremos una fiesta”, tiene que ver con “a fin” y tiene que ver con “año”. Sin embargo, en la siguiente, “a fin de evitar el deterioro”, “a fin de” lo podemos sustituir por una preposición “para evitar el deterioro de”, y en el último caso es el inicio de una frase fija “a fin de cuentas, utilizó todas las herramientas”. Hablaré sobre estos dos métodos.

El primero, que es detección y corrección de una palabra anómala, tiene que ver con los malapropismos. Se trata de un error, donde una palabra se reemplaza por otra similar en sonido pero de diferente semántica. Cuando aparecen estas palabras se destruyen las relaciones semánticas y sintácticas de las colocaciones. Las colocaciones son esas palabras que están relacionadas sintáctica y semánticamente: un ejemplo es “mañana soleada”, “mañana” es la palabra rectora, “soleada” la dependiente; “mechón de canas”, “pateando puertas”. Pero cuando hay errores en algunas de estas palabras, ya no se forman esas colocaciones con las palabras vecinas, y se podrían presentar: “mañana sopeada”, “lechón de canas”, “pateando muertas”; es la misma categoría gramatical, pero hubo un error en esas palabras.

Para corregir esas palabras anómalas podemos buscar cuáles son los parónimos. Los parónimos son palabras similares por su etimología o solamente por su forma o sonido, y en este trabajo lo que consideramos es que las palabras aparecían erróneas pero solamente en una letra, entonces un ejemplo: “histórico” cambió a “histérico”, “mesar” por “besar”, claro que en esas colocaciones se mantiene el tipo sintáctico, la categoría gramatical y entonces, por ejemplo, en verbos: “aportar” podría cambiar a “acordar”, “acostar”, “abortar”; adjetivos, si fuera “tenso” podrían cambiar a “denso”, “terso”. Nos estamos limitando a esos errores, donde solamente una letra se cambió. Los sustantivos de modos podrían cambiar a cualquiera de esos; si tenemos ese recurso de cuáles son los parónimos para una palabra, lo que queremos hacer es ver cuáles son los que tienen mayor relación para obtener la corrección. Y lo que hicimos fue utilizar Internet; teniendo claro ese recurso de cuáles son las palabras parónimas para una, y encontramos este conjunto experimental. En la primera parte de la frase, la corrección que es errónea dice “mañana sopeada” y “mañana” la podemos cambiar por “macana”.

Los parónimos que encontramos en diccionarios fueron “soleada”, “topeada”, “copeada”, sus estadísticas en Internet son las siguientes: cero no apareció “mañana sopeada” en internet, “mañana” como palabra apareció en más de 3 000 000 de páginas y “sopeada”, en 99; podemos utilizar esos valores para obtener datos estadísticos, algo parecido a información mutua, y determinar cuál es la correcta. Aquí por ejemplo

vemos que el tercer renglón, que es el de “mañana soleada”, obtiene 1 710 páginas en Internet. El siguiente caso, el 2 es el de “hora liso”, “ora”, en sus parónimos, podemos encontrar “ara, era, osa, ova”. Y para “liso”, el “luso”, “laso”, “leso”; y aunque encontramos que aparece “hora liso” siete veces, encontramos que el tercer renglón “era liso” es el que obtiene mayor cantidad de páginas en Internet. Podemos obtener esas páginas, necesitamos recursos como los parónimos, y obtenemos estadísticas; podemos determinar cuál es la corrección para esa colocación.

1)	1.2	“mañana sopeada”	0, mañana: 3180000, sopeada:99
	1	“macana sopeada”	0, macana:173000
	2!!	“mañana soleada”	3710, soleada:48100
	2	“mañana topeada”	0, topeada:117
	2	“mañana copeada”	0, copeada:15
	2	“mañana hopeada”	0, hopeada:4
	2	“mañana jopeada”	0, jopeada:26
2)	6.1	“ora liso”	7, ora: 13100000, liso:382000
	1	“ara liso”	0, ara:4160000
	1!!	“era liso”	922, era: 37700000
	1	“osa liso”	30, osa: 3810000
	1	“ova liso”	0, ova: 1880000
	2	“ora luso”	1, luso: 579000
	2	“ora laso”	1, laso: 144000
	2	“ora leso”	2, leso: 247000
3)	3.1	“rey vago”	10, rey: 6330000, vago: 552000
	1	“bey vago”	0, bey: 1670000
	1	“ley vago”	1, ley: 11800000
	1	“reo vago”	7, reo: 1360000
	2	“rey lago”	198, lago: 4280000
	2!!	“rey mago”	8320, mago: 705000
	2	“rey pago”	88, pago: 5160000
	2	“rey vaho”	0, vaho: 28800
	2	“rey vaso”	4, vaso: 882000
	2	“rey vado”	2, vado: 659000

Cuadro 3. Conjunto experimental

En el caso de desambiguación de frases preposicionales, lo que hallamos es que esta otra frase preposicional “al pie de” a veces tiene un sentido composicional. Frase idiomática: “al pie de las plantas”; y parte de una frase fija: “tomada al pie de la letra”. Hay tres usos para esa frase. Y en el procesamiento del lenguaje natural por computadora necesitamos decirle a la máquina cuáles son esos usos.

Lo que hacemos es considerar las propiedades lingüísticas de esas frases preposicionales idiomáticas; “a fin de” puede sustituirse por “para”, a veces. Vemos que una es la modificación restringida: esas frases no se pueden modificar sin cambiar su significado. Hay un ejemplo: “Por el gran temor a su estruendosa magia”, ya no tiene el significado “por temor a = para evitar” como en el siguiente caso “tengo la libertad bajo fianza por temor a una posible duda”, no es “por temor” el significado, sino “para evitar una posible duda”; la modificación restringida nos da una idea de estas frases preposicionales fijas.

Tampoco se pueden sustituir los sustantivos en estas frases preposicionales idiomáticas; “se tomará la decisión de si está a tiempo de comenzar la rehabilitación”, si aún puede comenzar y no puede sustituirse por “a periodo de” o a “época de”, es porque se rompe la relación sintáctico-semántica, además de que puede ser parte de una frase fija. Lo que hicimos fue buscar en Internet cuál es esa modificación, si es restringida, como lo suponemos para esa frase “con motivo de”, y no tan restringida para una frase preposicional regular. En el cuadro 4 están los casos que aparecieron, donde se utilizó asterisco. Vemos que la frase se mantiene “con motivo de” en 34 casos, que hay una modificación solamente en cuatro, y en 17, aparecen como asterisco, trae varias posibilidades, realmente está dividiendo esa frase.

Está bien que con motivo de fin de año todos tomen vacaciones pero ¿todos a la vez?

Búsqueda en Google: “con *motivo* de fin”.

NADA	...que se celebran en fechas connotadas, generalmente relacionadas con celebraciones históricas o con motivo de las fiestas de fin de año...	34
MODIF	Con el motivo adicional de fin de año, esperamos desde ya su asistencia y participación Jorge Castro, Jorge Raventos, Pascual Albanese 27/11/2006	4
FRASE	...Orizaba, Tampico, etcétera, etcétera, han estado cambiando, con este último motivo, telegramas para ponerse de acuerdo a fin de llevar una manifestación...	17

Cuadro 4. Modificación restringida

Está bien que con motivo de fin de año todos tomen vacaciones pero ¿todos a la vez?

Búsqueda en Google: “de *fin* de año”.

MODIF	(Cada/este un) la primera quincena del mes de diciembre; de cada fin de año b Participar de todos los beneficios que pueda otorgar UTP, ya sea colectiva o...	38
FRASE	chico busca trabajo para fin de año chico busca trabajo de camarero para fin de año. Publicar anuncio gratis...	59
NADA	Esta cuestión de procedimiento habitual se relaciona con las reuniones de mitad de año y de fin de año de Comité, y pasó a ser una cuestión de...	3

Cuadro 5. Modificación amplia

La idea de que hay unos casos en que no se modifica, que muy poco se modifica y otros que se separa nos puede dar una idea de lo que es la frase preposicional en ese contexto. Obtuvimos datos de una cantidad de frases preposicionales idiomáticas; encontramos que las que se modifican tienen un valor mínimo. Esto nos sirve de ayuda.

Vemos que la modificación para una frase regular en donde tiene que ver el sentido composicional es más amplia, fueron 38 casos con modificación. Otra frase que no se refiere realmente a la frase preposicional, sino que tiene que ser otra cosa muy diferente y en donde se mantuvo “de fin de”. Entonces, vemos que hay muchos más modificadores y eso también nos puede ayudar para determinar cuál es la frase preposicional.

En el caso de probar si se puede sustituir en sustantivo, lo que hicimos fue tomar un diccionario de sinónimos y sustituir cada uno de los sinónimos y ver si encontrábamos en Internet si existía esa frase en ese contexto. Se encontró, entonces, lo siguiente:

De fin de año

- “motivo de SINONYM de año”:

Final – 2 pages

Terminación – 3 pages

- “de SINONYM de año”:

Final - 157000 pages

Cabo - 121 pages

Remate - 12 pages

Terminación - 4 pages

Propósito - 132 pages

Meta - 4 pages

Aquí mostramos una frase regular, donde sí son sustituibles y obtenemos diferente número de páginas. En cambio, para las frases preposicionales idiomáticas “con motivo de” encontramos muy pocas o no encontramos en el contexto de la frase:

Con motivo de

- “con SINONYM de fin de año”

“con pretexto de fin de año” - 1 page

“con pretexto de fin” - 2 pages

A fin de

- “a SINONYM de incrementar de manera considerable”

- “a SINONYM de incrementar”

Para ver si una frase es una frase fija obtuvimos también cuáles son las cantidades en las que aparecen en Internet y fue en relación a una medida estadística similar a información mutua que obtuvimos los valores (véase cuadro 6); “al pie de la letra” y “al pie del cañón” son los que obtuvieron un mayor valor. Lo que hacemos es buscar esas relaciones sintáctico-semánticas en grupos de cadenas, valiéndonos de las páginas en Internet.

FRASE	# PAG	CONTEXTO	# PAG	SCI
al pie de la sede	31	la sede	9900000	-1.44
al pie de la Suburban	0	la Suburban	940	0
al pie de un cactus	9	un cactus	51000	0.57
al pie de una cruz	205	una cruz	964000	2.96
al pie de unas colinas	90	unas colinas	9930	5.07
al pie del Castillo	9930	Castillo	15000000	6.58
al pie de la torre	13500	la torre	7210000	7.55
al pie del Monumento	14500	Monumento	4230000	8.04
al pie de la montaña	49500	la montaña	4480000	10.06
al pie del cañón	140000	cañón	2290000	12.98
al pie de la letra	767000	la letra	10200000	14.33

Relaciones sintáctico-semánticas entre:

P-NP-P y contexto

NP P₂ + contexto

Cuadro 6. Partes de una frase fija

Uno de los problemas en el caso de ir examinando qué es lo que trajo el buscador en lugar del asterisco, es que necesitamos otro programa que vaya revisando. Podemos utilizar Internet para buscar estas cantidades, utilizar métodos estadísticos. El resultado: para frases idiomáticas fue el 100%; para frases regulares, el 99% de precisión; aunque existen algunos problemas en las primeras.

TIPO	PRECISIÓN	RECALL	# DETECTADOS	# CORREC MANUAL	# DETECTADOS CORRECTOS
EXP _{PNP}	56	-	31	18	18
IDIOM	100	80	4	5	4
REG _{PN} P	99	82	99	120	98

Presición: #P-NP-P detectados correctos / #P-NP-P detectados
Recall: #P-NP-P detectados correctos / #P-NP-P manualmente

Cuadro 7. Resultado

Finalmente, lo que quiero expresarles es que necesitamos mucha cantidad de ejemplos de lenguaje para los métodos en el procesamiento del lenguaje natural. Los corpus son muy útiles pero nos hace falta mucho más cantidad de esos textos para métodos por computadora, porque se basa en estadísticas; la web es muy útil pero necesitamos más opciones de acceso en esos buscadores. Muchas gracias.

Jeannette Reynoso: Le damos las gracias a la doctora Galicia Haro, y, debido a que esta mesa de trabajo incluye varias presentaciones, un total de cinco, vamos a dedicar al final de cada una de las presentaciones unos minutos para dialogar con nuestros ponentes, esto nos va a permitir sin lugar a dudas, tener la información mucho más cercana y tener una dinámica mucho más fácil para dialogar con ellos. La doctora Galicia tiene unos minutos para contestar preguntas, dudas, comentarios.

SECCIÓN DE PREGUNTAS

Pregunta 1: ¿Para hacer la búsqueda necesitan algún programa o el sistema puede hacerlo?

Sofía Galicia: Hay diferentes lenguajes, pero yo, por ejemplo, utilicé *Pearl* que lo que hace es que tiene posibilidades de accesos automáticos. El problema es la cantidad que nos permite, estamos hablando de mil, a veces mil quinientos accesos; parecen muchos, pero cuando se trata de obtener estos valores resultan pocos, sin embargo, con ese programa se puede hacer automático.

Pregunta 1: Entonces, ¿no han usado Gate todavía?

Sofía Galicia: Bueno, *Gate* tiene herramientas que permiten determinar los nombres y su objetivo es la extracción de información, pero no se basa tanto en la Web.

Pregunta 2: Por lo que entendí, utilizan el buscador de *Google*, pero mi pregunta es ¿el número de visitas que tiene cada página no afecta en estos estudios? Porque obviamente se nos dan las páginas más visitadas, ¿qué pasaría si ese ranking fuera diferente?

Sofía Galicia: Sí, precisamente, ése es un gran problema. Precisamente en lo que mostré hay más de un millón de páginas; entonces lo que hacemos automáticamente es buscar, claro no en todos porque es de muchísimo tiempo, pero hemos hecho unos estimados; por ejemplo, revisar diez páginas completas automáticamente y ver qué proporciones hay o, si eso no está funcionando, vamos a veinte páginas y revisamos también automáticamente esas páginas para tratar de obtener proporciones. La ventaja que yo le veo a la web es que incorpora herramientas lingüísticas; uno le pide “administración” y salen “administraciones” y variantes en la palabra, lo que lo hace más rico en el caso de fenómenos lingüísticos; realmente tenemos esa limitante, no podemos buscar en todas. Nosotros lo que hacemos es buscar por cantidades de páginas.

Jannette Reynoso: ¿Alguien más?

Pregunta 3: Como comentario breve, es interesante lo que presentaste. Se me ocurre consultar, por ejemplo, tratar de ver qué tanta irregularidad hay en ciertas combinaciones; bueno, es cierto, la *Gramática de la lengua española* ya lo ha dicho y lo vuelve a repetir y hasta ahora pareciera que no hay mayor ciencia en escribir frases nominales, lo curioso es que cuando confrontas ese tipo de datos obtenidos de patrones de lenguaje natural, escribir reglas para ese tipo de combinaciones es un reto. La gramática dice esto, pero se están generando este otro tipo de construcciones. Me parece un buen punto de análisis.

Por otro lado, por ejemplo, para un estudio de variación lingüística, también sería interesante ver qué tanto entran en competencia cierto tipo de frases para ciertos contextos, quizá se pueda encontrar distintos tipos de español, ver qué tanto están entrando en competencia unas con otras. Supongo que tu investigación no va sobre esa parte, pero son datos que se arrojan y creo que ése es otro buen trabajo que se puede hacer.

Sofía Galicia: Sí. Respecto a su pregunta, también, nosotros nos limitamos a español, pero como *Google* y muchos buscadores no consideran los acentos, ñ y eso, hay veces que trae páginas del portugués; sí, entonces hay que ponerle más cosas al programa para que vaya eliminando.

Jeannette Reynoso: Me voy a permitir justo hacer un comentario porque tu trabajo implica que se abre una posibilidad para uno de los problemas más fuertes en el análisis variacionista, que es tener corpus adecuados. Sí, muchos lingüistas en la actualidad sienten todavía una reminiscencia para trabajar, por ejemplo, las secuencias de *chateo*, porque les parece que están excesivamente marcadas por errores, sin embargo, se está abriendo a otra de variación lingüística; también el corpus es un problema para los que queremos hacer comparación, análisis, etcétera. Hay una parte que me interesó muchísimo: tu punto de referencia son los diccionarios y ése es un grave problema, porque, incluso para los lingüistas tradicionales, tomar como punto de referencia diccionarios tradicionales o con metodologías que ahora están muy cuestionadas es un problema. Habría tal vez la posibilidad, seguramente en un futuro, de ampliar esto a los estudios de variación y que tus puntos de referencia fueran de otro tipo.

Sofía Galicia: Sí.

Jeannette Reynoso: ¿Alguien más?

Pregunta 4: Me queda un poco la inquietud en cuanto a si podemos llamar corpus a Internet porque en realidad el concepto que yo tengo de corpus es un documento que está estable. Tal vez valga la pena llamarle o darle un nombre adecuado a esa fuente de datos, tal vez el nombre más apropiado no sería corpus porque finalmente está variando y si volvemos a hacer otro estudio mañana, los datos cambiarían. Me parece que abre otra línea de investigación de comportamiento del lenguaje; tal vez sería bueno pensar que se le podría dar otro nombre más apropiado a esa fuente de información.

Sofía Galicia: Sí, lo que pasa es que, a lo mejor, como yo estoy más en la parte de lingüística computacional, lo manejamos diferente. Nosotros vemos corpus sincrónico, diacrónico, balanceado; lo que todo mundo quisiera, sería un corpus balanceado, pero desafortunadamente para trabajo en lenguaje natural por computadora necesitamos una cantidad de gestos, por eso decimos para los de lingüística computacional, la web es nuestro corpus. Quizá sería apropiado ponerle algún adjetivo que lo marcara.

Janette Reynoso: Sí, yo me voy a permitir hacer un comentario sobre esa línea: creo que el concepto de “fuente de datos” va a depender mucho del objetivo final que se plantee en un estudio de lengua, porque incluso la movilidad, las diferencias permiten para las últimas tendencias de variación que sea un corpus mucho más fidedigno, mucho más cercano al habla común; al contrario, para algunas metodologías es mucho más atractivo que un corpus fijo y bien delimitado.

Pregunta 5: La web como tal no es un corpus, pero puede ser considerado como tal, estrictamente utilizando la definición moderna sobre lo que sería un corpus, no la sigue, pero la realidad es que es muy utilizado, hay muchas herramientas que sí se están desarrollando en distintos países. Por ejemplo, en donde se pueden traer las concordancias y edificando cada uno de los países de donde viene esa concordancia. Existe otro tipo de herramientas donde se pueden extraer las concordancias únicamente ordenadas como lo presenta el corpus de la *Real Academia Española*, una tras otra. En fin, hay muchas herramientas que se están utilizando. Hay que considerar si bien no lo es estrictamente, puede ser considerado y es una fuente de datos que vale la pena utilizar. El objetivo es también abrirse a nuevas herramientas, puesto que *Google* permite acciones muy limitadas, aunque es muy práctico, por supuesto, pero todavía falta más por hacer.

Sofía Galicia: Sí, precisamente ése es el punto, toda la información que está en Internet es muy útil, lo podemos usar, yo lo considero como mi corpus, pero hacen mucha más falta herramientas, porque hay que hacer mucha labor de computación para que lo pudiera utilizar gente que no tienen interés en hacer esa parte. Hacen falta herramientas.

Jeannette Reynoso: Muy bien, pues le damos las gracias a la doctora Galicia Haro. ■

CRITERIOS SINTÁCTICOS APLICADOS A UN PROGRAMA DE GENERACIÓN DE PARES SEMÁNTICOS

SONIA MORETT ÁLVAREZ
FFyL / GIL-IINGEN, UNAM

Jeanette Reynoso: A continuación Sonia Morett, egresada de la licenciatura en Lengua y Literaturas Hispánicas de esta Facultad y becaria de investigación en El Colegio de México, nos presentará la ponencia *Criterios sintácticos aplicados a un programa de generación de pares semánticos*, un tema donde se interrelacionan la sintaxis y la semántica para ser aplicadas en procesos de recuperación de información. Bienvenida, Sonia.

Sonia Morett: Muchas gracias y buenos días a todos. El trabajo que aquí presento forma parte de la investigación que realicé para mi tesis de licenciatura y tiene por objetivo elaborar propuestas para un mejor desempeño de los sistemas recuperadores de información desde la sintaxis, a partir del análisis del funcionamiento de un programa computacional denominado *clustering*.

La investigación se enmarca en el proyecto central del Grupo de Ingeniería Lingüística del que muy probablemente ya han oído hablar, el *Diccionario Electrónico de Búsqueda Onomasiológica* (DEBO). Las características principales de este diccionario están contenidas en su propio nombre: *electrónico* y *onomasiológico*. Con *electrónico* queremos decir simplemente que la colección de voces y sus definiciones se presenta en un formato digital (y no en papel) y que la información se recupera mediante un sistema computacional; con *onomasiológico* nos referimos al modo de búsqueda desde un punto de vista semántico, que parte de la descripción del concepto para llegar al término y no al revés (como ocurre con los diccionarios semasiológicos). Se trata, en este caso, de que el usuario pueda introducir en lenguaje natural su propia definición de un concepto y el sistema le proporcione el término que desconoce o no recuerda.

Ejemplo:

Búsqueda semasiológica: Termómetro (información conocida)

→ 1. m. Fís. Instrumento que sirve para medir la temperatura (incógnita).

Búsqueda onomasiológica: Cosa para medir la temperatura (información conocida)

→ Termómetro (incógnita).

En nuestros días, los diccionarios onomasiológicos y electrónicos están estrechamente relacionados, pues una base de datos computarizada formada por términos y sus correspondientes definiciones (independientemente de su propósito original) admite dos formas de búsqueda: la semasiológica, en donde el usuario teclea un término para que el sistema recupere el o los significados vinculados a éste, y la onomasiológica, en donde se inserta una definición con la intención de obtener el término que le corresponde. Es así que en la era digital los diccionarios onomasiológicos en soporte electrónico, gracias a su flexibilidad en las búsquedas, adquieren mayor relevancia sobre sus pares limitados a las posibilidades del estático libro impreso.

Sin embargo, la mayor posibilidad de éxito en las búsquedas en los diccionarios onomasiológicos no se produce automáticamente al ser llevados a formato digital. Para que pueda lograrse que el usuario de un diccionario inserte su propia

definición y ésta lo lleve con éxito al término deseado, es necesario un trabajo de preparación del sistema y de preprocesamiento de las definiciones, donde los estudios lingüísticos tienen grandes cosas que aportar. Uno de los mecanismos que han demostrado ser eficientes en este sentido es la consideración de *paradigmas semánticos*.

Entendemos por *paradigma semántico* un conjunto de palabras que pueden ser mutuamente sustituibles en el contexto de las definiciones sin que se vea alterado el significado del término al que corresponden. Un ejemplo podría ser el siguiente:

- Capacidad de un sistema material para **producir** trabajo, con las propiedades de la conservación y la inconvertibilidad. (*Océano Uno Colos Diccionario Enciclopédico*).
- Es la capacidad de **efectuar** trabajo. (*Física Weber*).
- Capacidad de un sistemas físico para **realizar** trabajo. (*Encarta 98*).
- Propiedad de un sistema: su capacidad para **hacer** un trabajo. (*Dictionary of Physics*).

Las cuatro definiciones corresponden al término *energía* y fueron extraídas del banco terminológico de Física del que dispone el Grupo de Ingeniería Lingüística (GIL) de la UNAM. Como puede observarse, *producir*, *efectuar*, *realizar* y *hacer* conforman un paradigma semántico, pues en el contexto de estas definiciones aceptan cualquier posibilidad de intercambio sin que se produzcan modificaciones de significado. Así por ejemplo, dentro del DEBO, un usuario podría decir que quiere saber cómo se denomina:

- La capacidad de producir trabajo,
- La capacidad de realizar trabajo o
- La capacidad de hacer trabajo.

El sistema lo debería remitir a *energía*, aunque en el banco terminológico que conforma la base de datos de sus búsquedas no hayan sido registradas esas definiciones sino *la capacidad de efectuar trabajo*. Así, con la base de una definición obtendríamos cuatro posibles definiciones que conducirían a la obtención exitosa del término buscado. La misma operación de sustitución podría aplicarse a cada una de las definiciones que aportaron miembros al paradigma. Incluso, de este ejemplo podríamos extraer otro paradigma semántico {*capacidad*, *propiedad*}, que, en este caso, es un *par semántico*.

Como puede observarse, la ampliación de los criterios de búsqueda en un diccionario onomasiológico a la consideración de paradigmas semánticos multiplica las probabilidades de éxito del potencial usuario para acceder al término deseado. En este caso, a partir de cuatro definiciones conseguiríamos 32 posibilidades.

Con la intención de obtener automáticamente estos paradigmas o agrupamientos semánticos (y con ello expandir las posibilidades de las búsquedas onomasiológicas) se utilizó el método *clustering*. Este método es un sistema de recuperación de información que trabaja con criterios estadísticos.

En una primera etapa, el *clustering* alinea dos definiciones de un mismo término; el procedimiento se repite hasta agotar todas las posibles combinaciones de pares de definiciones para cada término. (Cuando digo todas las posibles combinaciones de pares de definiciones me refiero a las definiciones que contiene el banco terminológico, y el número de estas es variable para cada término). Ya con los alineamientos, se localizan los *pares iguales*, los *pares nulos* y los *pares correspondientes*.

Veamos el siguiente cuadro:

DEF.1	Energía	Capacidad	de	un	sistema	material	para	producir	trabajo	con	las	...
DEF.2	Energía	Capacidad	de	un	sistema	físico	para	realizar	trabajo			
TIPO	par igual	par correspondiente	par igual	par correspondiente	par igual	par nulo	par nulo	pares nulos				

PARES IGUALES: Parejas de palabras pertenecientes a dos definiciones distintas que comparten posición y lema (no necesariamente flexiones).

PARES NULOS: Se consideran así las palabras que dentro de un alineamiento no tienen una paralela en la otra definición o, dicho de otro modo, su par es una palabra vacía.

PARES CORRESPONDIENTES: Parejas de palabras diferentes que ocupan una posición paralela dentro del alineamiento.

Cuadro 1. Tipos de pares

Posteriormente, se determina si los pares correspondientes se promueven a *pares vinculados* mediante el cálculo de un coeficiente de similitud llamado LCC (*longest colocation couple*).

El valor asignado por este coeficiente de similitud a cada uno de los pares de palabras del alineamiento del cuadro anterior se muestra en el siguiente cuadro:

DEF.1	Energía	Capacidad	de	un	sistema	material	para	producir	trabajo	con	las	...
DEF.2	Energía	Capacidad	de	un	sistema	físico	para	realizar	trabajo			
TIPO	par igual	par correspondiente	par igual	par correspondiente	par igual	par nulo	par nulo	pares nulos				
LCC	0	0	0	0	0	7	0	3	0	0	0	0
PAR VINCULADO	no	no	no	no	no	si	no	no	no	no	No	no

Cuadro 2. Ejemplo 1

Los pares vinculados son generados por el sistema; se trata de los candidatos a pares semánticos que una revisión humana de los alineamientos identificaría. Los pares semánticos son, como ya les decía anteriormente, parejas de palabras que en el contexto de los alineamientos pueden intercambiarse sin que se altere el significado de ambas definiciones.

Ni los pares iguales ni los pares nulos son considerados para el cálculo de LCC, puesto que ellos no pueden ser promovidos a pares vinculados y conformar pares.

A los pares correspondientes, los que sí pueden promoverse a vinculados, el sistema les asigna un valor de LCC, es decir, de similitud. En el ejemplo anterior (cuadro 2) es de 7 para el par {material, físico} y de 3 para el par {producir, realizar}. Este número coincide con la suma de los pares iguales registrados inmediatamente antes y después del par correspondiente en evaluación, más una unidad que le corresponde al mismo par. Se considera que un LCC equivalente a 5, es lo suficientemente bueno para ser promovido a vinculado.

Como el objetivo del programa es extraer automáticamente pares de palabras que coincidan con los pares semánticos que una persona podría reconocer, se trata de hacer que el funcionamiento del sistema se acerque lo más posible al razonamiento humano, y para esto no podemos valernos exclusivamente de métodos estadísticos.

Si nos fijamos bien, la tesis que subyace al funcionamiento del *clustering* es que los pares semánticos afloran en contextos idénticos (palabras iguales a la izquierda y a la derecha). Esto tiene un correlato en gramática: las similitudes de sentido se infieren del parecido en las relaciones sintácticas que establecen dos palabras.

Por lo tanto, la información gramatical prioritaria para el funcionamiento de esta fase del programa es la sintáctica. Con esta idea enfoqué mi tesis de licenciatura, que consistió en analizar sintácticamente los alineamientos arrojados por el *clustering* para un corpus conformado por las definiciones del área de Física del banco terminológico mencionado y, con esa información, realizar propuestas de modificaciones al sistema.

A continuación describiré algunas de estas modificaciones que planteo en mi trabajo y que inciden en el tratamiento de perífrasis gramaticales, nexos, adjetivos y adverbios contenidos en las definiciones.

Hay que tomar en cuenta que la instrumentación de cada una de las siguientes propuestas requiere que las definiciones hayan recibido previamente un etiquetado de categorías gramaticales. En lingüística computacional, esto quiere decir que a cada palabra se le asocia una marca con la información de la categoría gramatical a la que pertenece. Existen programas que realizan esto de manera automática, por lo que no sería complicado incorporar al motor del *clustering* un etiquetador.

Perífrasis gramaticales.

Sin más preámbulos, comienzo por las perífrasis gramaticales. Recordemos que una perífrasis es una construcción gramatical fija en la lengua que refiere a un concepto único. Entonces, las perífrasis serían funcionalmente palabras antes que frases. Mi propuesta es que el programa las considere así, ya que actualmente el *clustering* agrupa por separado las palabras que forman parte de una perífrasis.

Para este trabajo, consideraré tres tipos de perífrasis: locuciones, términos compuestos y perífrasis verbales.

- **Locuciones**

Las locuciones son estructuras que semántica y sintácticamente forman una unidad pero aparecen separadas en la escritura. En este ejemplo tienen ustedes una locución prepositiva:

DEF.1	E	p	g	...	y	que		está	a	una	altura	h	por	encima	de	un	determi- nado	nivel
DEF.2	E	p	g	...	Si	un	objeto	está	a	una	altura	h	sobre	el				nivel
TIPO DE PAR	I	I	I	C	C	C	N	I	I	I	I	I	C	C	N	N	N	I
LCC				1	1	1	0	0	0	0	0	6	1					
PAR VINCULADO		no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

Cuadro 3. Ejemplo 2

Por encima de desempeña la misma función gramatical que *sobre*, y por ser sus significados equivalentes, resultan intercambiables en el contexto de los alineamientos. Sin embargo, el *clustering*, tal como está programado actualmente, jamás arrojaría este par que el ojo humano sí identifica como semántico, pues el sistema establece parejas de palabras gráficas exclusivamente.

Para subsanar esta deficiencia y poder obtener pares más precisos se propone que las perífrasis sean tratadas como palabras, pero susceptibles de ser alineadas sólo con palabras con las que compartan función gramatical: prepositiva, conjuntiva.

- **Términos compuestos.**

Los términos compuestos constituyen unidades léxicas con un significado unitario y las relaciones sintácticas que pueden establecer son las mismas que cualquier palabra:

DEF.1	Difracción	Desviación	de	los	rayos	luminosos	cuando	pasan	por	los	bordes	...
DEF.2	Difracción	Es	la	dispersión	de	la	luz	en	una	región	situada	...
TIPO DE PAR	I	C	C	C	C	C	C	C	C	C	C	C
LCC	0	2	1	1	1	1	1	1	1	1	1	2
PAR VINCULADO	no	no	no	no	no	no	no	no	no	no	no	no

Cuadro 4. Ejemplo 3

La propuesta, en este sentido, es localizar y etiquetar como una unidad léxica los términos compuestos insertos en las definiciones. En el ejemplo que tenemos aquí, si *rayos luminosos* recibiera el tratamiento de una palabra, estaría en condiciones de alinearse con *luz*. La obtención de estos términos podría realizarse de manera automática en un corto plazo. Actualmente, el Grupo de Ingeniería Lingüística desarrolla un programa para la obtención de términos compuestos o multipalabra en textos de carácter científico o técnico para el español.

- **Perífrasis verbales**

Por perífrasis verbal vamos a entender toda reunión de formas verbales que, en conjunto, refieren a un solo proceso o estado, pues si nos ciñéramos a las definiciones que aportan las gramáticas, dejaríamos fuera muchas que no se ajustan al prototipo.

La importancia de darle un tratamiento especial a estas expresiones verbales se deriva de la alta incidencia de apariciones que detecté dentro de mi corpus, lo que se corresponde con su abundancia en la lengua española.

Aquí he seleccionado un ejemplo:

DEF.1	Péndulo	cuerpo	degrave	que	puede	oscilar	suspendido	de				un	punto	
DEF.2	Péndulo	cuerpo	rígido	que	cuelga	de	un	hilo	delgado	sujeto	a	un	soporte	fijo
TIPO DE PAR	I	I	C	I	C	C	C	C	N	N	N	I	C	N
LCC	0	0	4	0	2	1	1	1	0	0	0	0	2	0
PAR VINCULADO	no	no	no	no	no	no	no	no	no	no	no	no		

Cuadro 5. Ejemplo 4

La primera definición del alineamiento tiene por núcleo verbal una combinación de dos perífrasis básicas (verbo conjugado + infinitivo + participio), que en conjunto remiten a un proceso equiparable al expresado por el verbo de la segunda, aunque entre ambas expresiones existe una diferencia semántica: en la segunda definición la forma verbal *cuelga* señala que el acontecimiento existe, mientras que en la primera *puede oscilar suspendido* manifiesta la posibilidad de que se dé esa situación.

La presencia de las perífrasis dentro de los alineamientos desencadena que aparezcan como pares correspondientes palabras con diferente categoría gramatical: {oscilar, de}, {suspendido, un}. Recordemos que el programa establece agrupamientos a partir de palabras gráficas.

Si consideramos verbo a toda estructura gramatical que funcionalmente actúa como tal, tendremos alineamientos más precisos, pues no sólo se obtendrán pares formados por un verbo y una perífrasis del tipo {cuelga, puede oscilar suspendido}, sino que esto repercutirá en una mejor organización en pares de todo el alineamiento. De esta manera podría agruparse el par {punto, hilo}. Esta consecuencia es válida para todos los tipos de perífrasis que les he presentado hoy, y es más, es válida para todas las modificaciones sugeridas al programa: como es lógico, un ordenamiento más preciso de determinado tipo de pares desencadenará un reordenamiento general de los alineamientos, y esto es favorable para la obtención de pares semánticos más precisos.

Nexos

El nexo es una categoría funcional en la que se reúnen los elementos gramaticales que sirven para enlazar a otros. Las clases de palabras que funcionalmente actúan como nexos son las conjunciones, las preposiciones y también los verbos copulativos. No hablaré de las conjunciones, pues dentro de ellas existen varias categorías diferentes que he considerado para efectos del *clustering* y requeriría más tiempo abordarlas.

- **Preposiciones**

A parte de las locuciones con función prepositiva (como vimos anteriormente), en mi trabajo consideré las preposiciones que son regidas por un verbo.

DEF.1	Cinemática	Parte	de	la	mecánica	que	trata	del	movimiento	...
DEF.2	Cinemática	Parte	de	la	mecánica	que	estudia	el	movimiento	...
TIPO DE PAR	I	I	I	I	I	I	C	C	I	C
LCC	0	0	0	0	0	0	7	2	0	2
PAR VINCULADO	no	no	no	no	no	no	no	no	no	no

Cuadro 6. Ejemplo 5

En el alineamiento se considera el par correspondiente {trata, estudia} con un LCC de 7, como decía, un muy buen índice de similitud y, sin embargo, el sistema no lo considera vinculado. ¿Por qué? Porque no cumple con otra condición necesaria que es la condición de frontera. Esto es, que un par correspondiente, además de registrar un LCC igual o superior a 5, debe tener, por lo menos, un par igual a su izquierda y otro a su derecha para ser promovido a vinculado.

Los contextos sintácticos donde aparecen las palabras que integran este par correspondiente son muy similares. El verbo de la oración que expresa la primera definición rige un suplemento (o complemento de régimen de verbo prepositivo) y el verbo estudia un implemento (o complemento directo). Puesto que suplemento e implemento desempeñan la misma función transitiva, considero que la preposición del primero debería ser

tratada como un incremento verbal. Entonces el par correspondiente ya no sería {trata, estudia}, sino {trata de, estudia} y el reagrupamiento de los pares del alineamiento se haría como se muestra a continuación:

DEF.1	Cinemática	Parte	de	la	mecánica	que	trata de	el	movimiento	...
DEF.2	Cinemática	Parte	de	la	mecánica	que	estudia	el	movimiento	...
TIPO DE PAR	I	I	I	I	I	I	C	I	I	C
LCC	0	0	0	0	0	0	9	0	0	2
PAR VINCULADO	no	no	no	no	no	no	no	no	no	no

Cuadro 7. Ejemplo 6

Para poder realizar este cambio, descompose la partícula *del* en sus constituyentes originales con la finalidad de que preposición y artículo pudieran integrarse a pares independientes en el contexto de los alineamientos. La separación de las partículas que son contracción de formas lingüísticas con diferente categoría gramatical es una propuesta general para la etapa de preprocesamiento de las definiciones que contemplo en la tesis.

Si *trata de* y *estudia* formaran un par, el siguiente par correspondiente sería un par igual {el, el}, por lo que se cumpliría la condición de frontera, y el par que manualmente hemos identificado como semántico sería promovido a vinculado con un LCC de 9. El coeficiente de similitud tan alto y el sentido común nos dicen que el nuevo par muestra un parecido muy cercano. El hecho de aparecer en un contexto idéntico deja demostrado que ambos verbos (y el incremento de uno de ellos) pueden ser intercambiados sin que se produzcan alteraciones de sentido.

El mecanismo para incorporar esta modificación al motor de *clustering* es muy sencillo. Se trata de proporcionar al sistema una lista de verbos que rigen preposición y éste automáticamente haga los cambios pertinentes.

- **Verbos copulativos.**

Los verbos copulativos tienen un significado tan amplio que requieren de un atributo que los especifique. Una estructura de sujeto + atributo donde no se observa verbo presupone la existencia de uno de este tipo que se ha elidido y esto es bastante frecuente. Es común que, cumpliendo la función de nexos, quede como rastro una coma.

Observen el siguiente alineamiento:

DEF.1	Móvil	Cuerpo	que	está	en	movimiento
DEF.2	Móvil	Cuerpo			en	movimiento
TIPO DE PAR	I	I	N	N	I	I
LCC	0	0	0	0	0	0
PAR VINCULADO	no	no	no	no	no	no

Cuadro 8. Ejemplo 7

Las dos definiciones expresan exactamente la misma idea, la única diferencia entre ellas radica en que en la primera el sujeto y el atributo aparecen vinculados mediante un verbo copulativo que en la segunda definición no se muestra.

Mi aportación, en este sentido, es que los verbos copulativos (y en su caso, el nexos que los precede) que se alinean con conjuntos vacíos, se consideren pares *semi nulos*. Un par *semi nulo* es un par nulo formado por una palabra funcional y un espacio vacío en la cadena. Este tipo de pares, que no había explicado con anterioridad, surgieron de una revisión anterior al *clustering* y su aplicación tiene el propósito de impedir que los pares nulos corten las cadenas discursivas de los entornos en que aparecen para efectos del cómputo de LCC.

Adjetivos.

En lo que respecta a los adjetivos, estos han sido considerados en un sentido amplio: un adjetivo es toda unidad gramatical (independientemente de su clase sintagmática) que desempeña la función de modificador nominal.

En el siguiente alineamiento aparecen diferentes formas de modificadores nominales:

DEF.1	Primera	ley	de	Newton	Todo	cuerpo	continúa	en	su	estado	de	reposo	o		...
DEF.2	Primera	ley	de	Newton	Todo	cuerpo	continúa	en	su	estado	de	reposo	o	de	...
TIPO DE PAR	I	I	I	I	I	I	I	I	I	I	I	I	I	N	I
LCC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PAR VINCULADO	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no

Cuadro 9. Ejemplo 8

DEF.1	...	movimiento	rectilíneo	uniforme				a	menos	que	...
DEF.2	...	movimiento		uniforme	en	línea	recta	a	menos	que	...
TIPO DE PAR		I	N	I	N	N	N	I	I	I	I
LCC		0	0	0	0	0	0	0	0	0	0
PAR VINCULADO		no	no	no	no	no	no	no	no	no	no

Cuadro 10. Ejemplo 9

DEF.1	...	sea	impelido	a	cambiar	dicho	estado	por	fuerzas	que	actúen	sobre	él,
DEF.2	...	sea	obligado	a	cambiar	este	estado	por	fuerzas	impresas		sobre	él.
TIPO DE PAR		I	C	I	I	C	I	I	I	C	N	I	I
LCC		0	7	0	0	6	0	0	0	4	0	0	0
PAR VINCULADO		no	sí	no	no	sí	no	no	no	no	no	no	no

Cuadro 11. Ejemplo 10

En este alineamiento, el adjetivo *rectilíneo* cumple con respecto a *movimiento* la misma función especificativa que la frase *en línea recta*. Por otra parte, en la oración *que actúen sobre él* el nexos *que* y el verbo conjugado se pueden sustituir por el participio en función adjetiva *impresas* debido a que cumplen la misma función de modificador nominal.

Con la intención de dar un tratamiento homogéneo a las diferentes formas de adjetivación que presentan las definiciones se sugiere aplicar un *chunking* a las definiciones contenidas en nuestro banco terminológico, para así localizar frases y determinar con qué categoría gramatical se corresponden. Como consecuencia de

esto, se podría habilitar ya el sistema para alinear adjetivos, frases adjetivas y frases preposicionales en función de complemento adnominal cuando fuera pertinente.

Una vez con las etiquetas de categorías gramaticales para las frases, se podría ampliar este tratamiento a los adverbios. A manera de ejemplo puede verse en el siguiente alineamiento *libremente* y *con libertad* son dos modificadores verbales que bien podrían integrar un par semántico:

DEF.1	Péndulo	Cuerpo	que		...	puede	oscilar	libremente		alrededor	...
DEF.2	Péndulo	Cuerpo	rígido	que	...		oscilar	con	libertad	alrededor	...
TIPO DE PAR	I	I	C	N	N	N	I	C	N	I	
LCC	0	0	3	0	0	0	0	2	0	0	
PAR VINCULADO	no	no	no	no	no	no	no	no	no	no	

Cuadro 12. Ejemplo 11

A manera de conclusión de esta exposición puedo decirles lo que se logró con el trabajo al que aquí he hecho referencia:

- 1) Se delimitaron los aspectos sintácticos que deben tomarse en consideración para posteriores modificaciones al sistema.
- 2) Se confirmó la hipótesis de que programar el *clustering* para reconocer rasgos fundamentales de la gramática española mejora el proceso de recuperación de pares semánticos.
- 3) En el Grupo de Ingeniería Lingüística nos dimos cuenta de que la mayoría de las iniciativas aportadas puede extenderse a solucionar problemas relativos a la extracción de información relevante en otras bases de datos léxicos.

Muchas gracias por su asistencia y su atención. Ojalá que esta exposición haya resultado de su interés.

Jeanette Reynoso: Muchas gracias, Sonia, por compartirnos este interesante trabajo en el Coloquio. ■

ETIQUETADO DE CONTEXTOS DEFINITORIOS

BERTHA LECUMBERRI
FFyL, UNAM

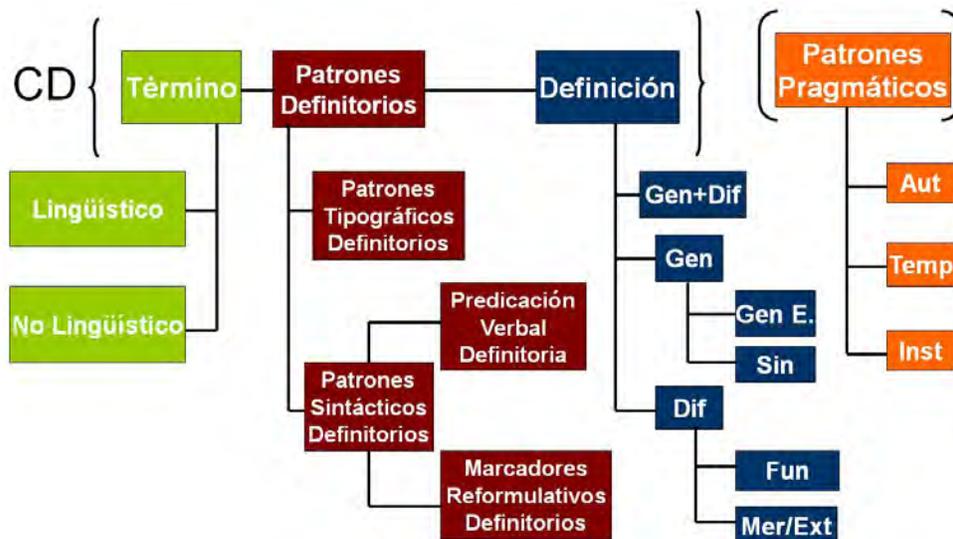
Jeanett Reynoso: Vayamos al tercer trabajo de esta mesa. Lo presenta Bertha Lecumberri, también parte del Grupo de Ingeniería Lingüística (GIL), actualmente estudiante de la licenciatura de Lengua y Literaturas Hispánicas. Su trabajo se titula *Etiquetado de contextos definitorios*.

Bertha Lecumberri: Actualmente, en el Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería (IINGEN) se está desarrollando el proyecto *Extracción de conceptos en textos de especialidad a través del reconocimiento de patrones lingüísticos y metalingüísticos*; el objetivo es obtener contextos definitorios de terminología especializada para apoyar la construcción de diccionarios electrónicos onomasiológicos y semasiológicos, la elaboración de bancos terminológicos, el diseño de ontologías, además de agilizar la búsqueda automática de términos y definiciones.

Nuestro objeto de estudio son los contextos definitorios (CD). Un contexto definitorio es un fragmento textual cuyos principales elementos son un término y una definición, además de otras unidades que ligan estos dos componentes. Pueden ser de tipo: *X define Y como Z*, *X es Y*, *X se compone de Y^o... Yⁿ*, *X funciona en Y* y *X se llama también Y*.

Un ejemplo del primer tipo –*X define Y como Z*– es: *Lafourcae (1980) define el perfil profesional como una especificación de habilidades, rasgos y disposiciones que orientan la construcción del plan de estudios y asuntos que definen el que hacer de los miembros de cierta profesión*.

Para el estudio de los contextos definitorios, se han dividido en las siguientes partes:



Cuadro 1. División de los Contextos Definitorios

Término, patrones definitorios y definición, además de los patrones pragmáticos. Un término puede ser lingüístico, formado por palabras, y no lingüístico, esto es, cuando está formado por fórmulas, las cuales son muy comunes en el área de ingeniería.

Los patrones definitorios representan el nexo que une al término con su definición. Pueden ser patrones tipográficos definitorios como dos puntos, comillas, paréntesis, etc.; patrones sintácticos definitorios, esto es, una predicación verbal definitoria; o bien marcadores reformulativos definitorios, los cuales son construcciones que explican el lenguaje mismo, y de alguna manera están modificando el contexto.

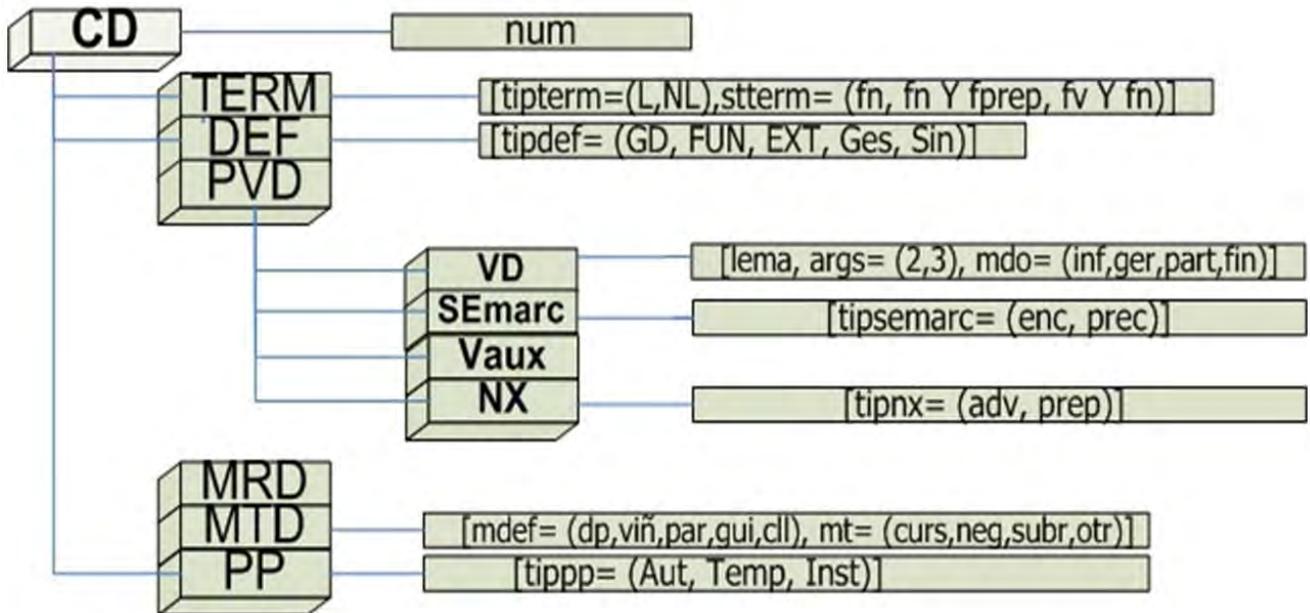
Existen tres tipos de definiciones:

- *Genus + differentia*: el género próximo al que pertenece el término y la *differentia* es lo que lo separa de los demás conceptos que pertenecen a este género.
- *Genus*: puede ser *genus* exclusivo, o sinónimica. Un *genus* exclusivo se refiere a cuando sólo aparece el género al que pertenece el término en cuestión.
- *Differentia*: puede ser funcional: dice en qué funciona, para qué sirve, o en qué se usa; y metonímica o extensional que es la enumeración de las partes que componen al término.

Los patrones pragmáticos explican el uso que se le está dando al término dentro del contexto; pueden ser de autoría, temporales o instruccionales, éste último se refiere a la luz de qué corriente o teoría se está definiendo el concepto.

A partir de esa clasificación se desarrollaron las siguientes etiquetas:

Etiquetas



Cuadro 2. Etiquetas utilizadas en los CDs

Del lado izquierdo están las etiquetas en mayúsculas y del lado derecho los atributos que marcan las diferencias que pueden existir en cada una de las partes del contexto.

CD: engloba todo el contexto, su atributo es un número, esto es para tener numerados todos los contextos definitorios en el corpus.

TERM: el término tiene dos atributos: tipo de término, el cual puede ser lingüístico o no lingüístico, y la estructura del término que puede ser frase nominal; frase nominal y frase prepositiva, frase verbal, y frase nominal.

DEF: La etiqueta de definición cuenta con un atributo de tipo de definición, puede ser *genus + differentia*, funcional, extensional, *genus* exclusivo y sinonímica.

El patrón verbal definitorio (**PVD**), tiene cuatro etiquetas:

Verbo definitorio (**VD**): tiene tres atributos: lema, que es la forma en infinitivo del verbo que está definiendo; número de argumentos, que tiene que ver con la valencia verbal; y el modo, que puede ser infinitivo, gerundio, participio o finito, que se refiere a las formas conjugadas.

SEmarc: se refiere a la presencia del clítico *se*, como atributo se pone si es proclítico o enclítico.

Verbo auxiliar (Vaux), que no tiene atributos.

Nexo (NX): cuyo atributo es el tipo de nexo, el cual puede ser adverbio o preposición.

MRD: Los marcadores reformulativos definitorios no tienen atributo.

MTD: Los marcadores tipográficos definitorios se dividen en dos tipos: marcadores definitorios, que pueden ser dos puntos, paréntesis, viñetas, guiones y comillas; y los marcadores tipográficos que son cursivas, negritas, subrayado u otros. La diferencia es que las primeras sirven para introducir una definición o marcar un nexo entre el término y la definición y las segundas son para resaltar de algún modo los términos o las definiciones y así facilitar que la computadora los localice.

PP: Los patrones pragmáticos pueden ser de autoría, temporales o instruccionales.

En el cuadro 3 se pueden ver las etiquetas con sus atributos en lenguaje XML:

BOTÓN	ETIQUETA
1. CD	<CD num="" > </CD>
2. TERM	<TERM tipterm="" stterm="" > </TERM>
3. DEF	<DEF tipdef="" > </DEF>
4. PVD	<PVD> </PVD>
5.VD	<VD lema="" args="" mdo="" > </VD>
6.SEmarc	<SEmarc tipsemarc="" > </SEmarc>
7.Vaux	<Vaux> </Vaux>
8.NX	<NX tipnx="" > </NX>
9.MRD	<MRD> </MRD>
10.MTD	<MTD mdef="" mt="" > </MTD>

Cuadro 3. Atributos de las etiquetas

Para la extracción de CD's se utilizó el Corpus Lingüístico de Ingeniería (CLI), el cual cuenta con 81 documentos especializados en el área de ingeniería y tiene aproximadamente 300 000 palabras.

El primer paso consistió en rastrear las ocurrencias de lemas de verbos definitorios, lo cual da como resultado una lista de posibles candidatos a CD. Posteriormente, se hace una selección manual de los buenos candidatos a CD para ser etiquetados.

Para facilitar el proceso de etiquetado, se implementó un sistema semiautomático por medio de macros en *Microsoft Word*, las cuales son instrucciones cronológicas usadas para economizar tareas. Así, el etiquetador no tiene que escribir manualmente cada una de las etiquetas, simplemente selecciona el texto que quiere etiquetar y al hacer *clic* en el botón de la etiqueta deseada, ésta aparece automáticamente.

Algunas consideraciones que se deben de tomar en cuenta durante el etiquetado, con el fin de facilitarle la lectura de las etiquetas a la computadora, son:

1. Todas las etiquetas deben de ir dentro de la etiqueta CD.
2. Las etiquetas VD, SEmarc, Vaux y NX deben ir dentro de la etiqueta PVD.
3. Todas las etiquetas deben aparecer en cada CD aunque algunas queden vacías.

Un ejemplo de etiquetado:

```
<CD num= "1"> <PP tippp= "Inst"> La Teoría del Buque </PP> estudia el
<TERM tipterm= "L" stterm= "fn"> barco </TERM> <PVD> <VD lema=
"considerar" args= "3" mdo= "part"> considerado </VD> <NX tipnx= "adv">
como </NX> <SEmarc tipsemarc= ""> </SEmarc> <Vaux> </Vaux> </PVD>
<DEF tipdef= "GD"> un flotador que se mueve en un líquido </DEF> . <MRD>
</MRD> <MTD mdef= "" mt= ""> </MTD> </CD>
```

Cuadro 4. Ejemplo de texto etiquetado

Primera etiqueta de CD, con su atributo numérico.

Aparece un patrón pragmático instruccional, esto es, el término se está definiendo a la luz de la *Teoría del Buque*.

El término es *barco*, el cual es lingüístico, y su estructura es frase nominal, como se puede ver el artículo no va dentro de la etiqueta.

El patrón verbal definitorio es *considerado como*.

El verbo definitorio es *considerado* cuyo lema es *considerar*, cuenta con tres argumentos y su modo es participio.

El nexos que une al verbo definitorio con la definición es *como* el cual es un adverbio.

La definición, *un flotador que se mueve en un líquido*, es *genus + differentia*, donde *un flotador* es el *genus* y *que se mueve en un líquido* es la *differentia*.

Por último aparecen las etiquetas que quedaron vacías pero que deben de incluirse: SEmarc, Vaux, MRD y MTD.

Véase el cuadro 5, donde se presenta dos ejemplos de los problemas que surgen a partir del etiquetado.

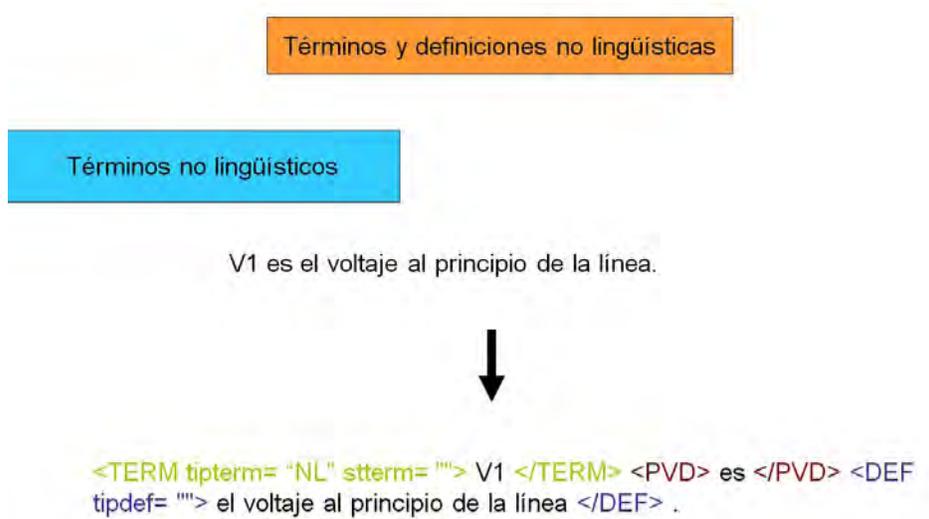
En el ejemplo se puede apreciar un contexto que presenta un término no lingüístico:

Como se puede ver, los términos no lingüísticos no presentan ningún problema, esto está previsto en los atributos, por lo que en el atributo de tipo de término se establece que es un término no lingüístico (NL).

Por otro lado se encuentran las definiciones no lingüísticas como la siguiente:

“El coeficiente de acomodamiento es $F = nk(1-P)(3)$ ”

Teóricamente una definición no puede ser no lingüística pero aquí parece que lo es, por lo que aquí se presentan dos posibles soluciones a este tipo de contextos, comunes en el área de ingeniería.



Cuadro 5. Términos y definiciones no lingüísticas

En la primera, *coeficiente de acomodamiento* es el término y $F = nk(1-P)^3$ la definición. Se propone un etiquetado con una definición no lingüística. La segunda solución entiende que la definición es la parte lingüística del contexto.

Las preguntas que es necesario responder para etiquetar correctamente este tipo de contextos son:

- ¿Es esta estructura una definición o una descripción?
- En caso de ser un CD, ¿ $F = nk(1-P)^3$ es el término o la definición?
- ¿Sería *Donde $n = 1$ es el número de capas, $k = 1.31$ es el coeficiente empírico de la capa, y $P = 0.25$ es la porosidad de la capa de coraza* parte de la definición o bien una descripción de la misma?

Para la extracción de los CD se establecieron dos criterios:

1. Rastrear los lemas de los verbos definatorios, como ya se había señalado anteriormente.
2. Recuperar el contexto que contenga dicho verbo contenido entre dos puntos.

Usando únicamente estos dos criterios provoca que la computadora arroje contextos incompletos como los siguientes:

1. El proceso de interrupción se puede escribir brevemente como sigue: 1.
2. Éste consta de un banco de capacitores sumergidos en aceite en un recipiente de porcelana y conectados en serie para aumentar la resistencia de la línea de alto voltaje.

En el ejemplo 1 falta la definición, mientras que en el ejemplo 2 “Éste” es un pronombre que sustituye al término, el cual fue mencionado anteriormente en el texto, por lo que habría que expandir el contexto para encontrar el término del presente contexto.

Actualmente los contextos incompletos se amplían manualmente, por lo que es necesario desarrollar una herramienta que prevea este tipo de situaciones y arroje los contextos ampliados automáticamente.

Jeanett Reynoso: Gracias Bertha. Evidentemente tu investigación está planteando muchísimos de los problemas que no sólo le pertenecen al área ahora de ingeniería lingüística sino problemas tradicionales en el análisis semántico, entonces, es realmente un nuevo ámbito de aplicación. ¿Alguien está interesado en dialogar con Bertha? Seguramente muchos de ustedes tienen preguntas y comentarios.

SECCIÓN DE PREGUNTAS

Margarita Palacios: Cuando marcaste una de tus últimas preguntas seguramente tienes ya un paradigma para definirlos. Dices: ¿es esto una descripción o una definición? ¿Qué estás usando para determinarlo?

Bertha Lecumberri: Ése es el problema, todavía no tenemos una solución para este planteamiento.

Margarita Palacios: Tengo un esquema cruzado que comparto. Para mí la estructura de descripción es la estructura del género de la definición, ¿cuál es la estructura de la descripción? La estructura de la descripción dice que todo lo que describimos se hace a partir del *ser* y el *hacer*, o sea, un micrófono es lo que sea, es y tiene. Son los dos verbos que yo usaría y luego me voy al *hacer*: *sirve para* y que la unión de estos dos me da una definición. Creo que está cruzado, no está totalmente claro. Seguimos en la misma duda.

César Aguilar: Me adelanto un poco a la presentación que nos toca. Digamos que la cuestión sería una división meramente operativa en el sentido de que partimos de un planteamiento más o menos estándar en lexicología y terminología en el sentido de que definir una definición se puede ver como una relación entre *genus* y *differentia* y de algún modo lo que estamos rescatando es la definición típica aristotélica y en última instancia, a lo mejor, planteamos que subyace una estructura de entidades, cópulas y atributos de lo mismo que tú comentaste, *ser* y *hacer*; de algún modo, lo que hacen es dar una lista de atributos X y ya con eso se arma una definición canónica.

El problema sería más bien, por ejemplo, si consideramos una definición, no se puede referir al ámbito del lenguaje natural y se puede decir que gramaticalmente ocupa palabras, construcciones, frases y demás. Entonces, de algún modo, todavía podemos rescatar que la definición es una estructura lingüística. El problema es que cuando aparece este tipo de definiciones, supongo que la gente con mayor formación en matemáticas, ingenierías y demás, en ese sentido, es que cuando pasas al nivel de descripción, por ejemplo, de fórmulas, de silogismos o de representaciones más o menos abstractas, y decir que representa al valor 3.1416 según el arco tal, lingüísticamente tenemos problemas, porque no podríamos decir que son numerales o que es una vocal en función o representativa de símbolo.

La cuestión es señalar que todo ese tipo de cosas lo ponemos en el cajón de sastre de descripciones o de formulismos y en algún momento vamos a ver si eso puede funcionar operativamente como una definición o no, por lo menos ahora lo que tendríamos visto es que a lo mejor un término sí puede funcionar, ahora hay que ver con las definiciones o, en todo caso, cuando encuentras un patrón de este tipo con estas características gramaticales tómallo como definición y cuando encuentres este otro ponlo aparte y luego vemos qué es eso.

Valeria Benítez: Quiero hacer un reconocimiento al trabajo de Bertha, ya que ha sido el trabajo más difícil en el proyecto de contextos definitorios: el etiquetado, porque no había nadie que fuera específicamente etiquetador, puesto que todos estábamos en los pequeños temas, y en los sujetos específicos del tema. Ella llegó a hacer este trabajo de etiquetado y a partir de su colaboración se han descubierto cosas, porque justo esto que le decimos: mételo en el cajón de sastre, nos hace descubrir problemas que se tienen que resolver forzosamente para poder hacer extracciones automáticas de CD's mucho más precisas.

César Aguilar: Retomando lo que dice Valeria, la idea es hacer una hipótesis a partir de datos concretos y hacer la invitación: si gustan colaborar con el proyecto con la parte de etiquetado, es la parte de trabajo en bruto; sin embargo, sin ese trabajo no salen las grandes ideas y los grandes temas ni tampoco este tipo de problemas que a la hora de pensarlos o sentarse a reflexionarlos uno se pregunta ¿qué pasa aquí?

Es un trabajo muy interesante y una buena forma de meterse a la ingeniería lingüística, por una parte práctica enteramente lingüística de tener el dato y por el otro lado la etapa de resolución de problemas.

Javier Cuétara: Es una pregunta muy puntual. Al principio de tu presentación hiciste un esquema de los contextos definitorios, ¿de dónde viene?, ¿tú la diseñaste?

Bertha Lecumberri: No, ese proyecto ya estaba cuando yo llegué; es trabajo que ya llevaba mucho tiempo en manos de Rodrigo Alarcón, César Aguilar, Gerardo Sierra y del grupo en general.

Javier Cuétara: Entonces, ¿ésta es una creación del Grupo de Ingeniería Lingüística para su aplicación?

Valeria Benítez: De hecho, el concepto de CD es una creación del grupo y poco a poco se han ido distinguiendo las partes de los CD de acuerdo con las definiciones. Si en un principio no se tomaban patrones pragmáticos, ahora se toman, ya que es un elemento recurrente en las definiciones. Entonces, estaba ese mapa que ha ido mejorando poco a poco.

Javier Cuétara: Sí, el grupo lo tiene y lo va ampliando.

Bertha Lecumberri: De hecho, las etiquetas tampoco son definitivas, pueden ir cambiando dependiendo de las necesidades que los mismos contextos vayan presentando.

Valeria Benítez: Esperemos que ahora sí queden para siempre.

Jeanett Reynoso: Bueno, le damos las gracias. ■

ANÁFORAS Y OTRAS RELACIONES DE CORREFERENCIA EN LA EXPANSIÓN DE CONTEXTOS DEFINITORIOS

VALERIA BENÍTEZ R.
FFyL / GIL-IINGEN, UNAM

Jeannette Reynoso: A continuación Valeria Benítez nos presenta un tema relacionado con la ponencia anterior, es parte de la investigación de sus tesis de licenciatura en Lengua y Literatura Hispánicas. Es interesante constatar que en investigaciones como ésta, tanto la lingüística como la informática convergen para dar resultados y aplicaciones muy interesantes. Adelante Valeria.

Valeria Benítez: Buenos días, mi nombre es Valeria. Lo que a continuación voy a presentar es parte de la investigación de mi tesis de licenciatura que tiene por nombre *Anáforas y otras relaciones de correferencia en la expansión de contextos definitorios*, la cual se desarrolla en el marco del proyecto *Extracción de conceptos en textos de especialidad a través del reconocimiento de patrones lingüísticos y metalingüísticos*.

El objetivo principal de dicho proyecto es la extracción automática de estructuras discursivas denominadas Contextos Definitorios (CDs), es decir, fragmentos textuales en los que hay un *término* y su correspondiente *definición*; podemos ver un ejemplo:

La “componente horizontal” es la suma de la fuerza resultante de la distribución de presiones y la fuerza de fricción.

Los CDs se extraen de manera automática haciendo búsquedas de patrones verbales definitorios, es decir, aquellos verbos que ligan al término con su definición. Vemos ente comillas el *término*, subrayado el *patrón verbal definitorio*, lo demás es la *definición*. Los patrones verbales que se emplean como patrón de búsqueda son en especial verbos tales como *caracterizar, comprender, concebir, conocer, considerar, consistir, constar, definir, denominar, describir, entender, identificar, llamar, permitir, servir, usar, utilizar* y *ser*.

Los fragmentos textuales con los que se trabaja se obtienen de corpus especializados del tipo tesis, artículos de divulgación, resúmenes, entre otros; los corpus principales para el proyecto son el *Corpus Lingüístico de Ingeniería (CLI)* y el *Corpus técnico del IULA*.

Cabe señalar que se trata del corpus principal de mi trabajo ya que éste es uno de los que se han generado en el GIL, y por ello podemos acudir a él sin restricciones. La mayoría de los CDs son obtenidos de este documento, aunque además se trabaja simultáneamente con el *Corpus Técnico del IULA*, con el objetivo de aprovechar los adelantos que se han logrado con ese corpus; no olvidemos que se trata de un proyecto en el cual hay distintos temas en desarrollo.

Es importante señalar que el proceso y metodología para la extracción de contextos corresponde a una etapa anterior a mi investigación; de hecho, el tema de mi tesis surgió de algunos problemas detectados en los resultados hasta ahora obtenidos en la extracción automática de CDs. En este sentido, se ha observado que la extensión de las definiciones no corresponde a un mismo patrón y entonces los contextos despliegan distintas estructuras.

Es importante decir que las definiciones están “dispersas” en el discurso, es decir, se trata de fragmentos textuales que aparecen espontáneamente en los textos especializados. De tal manera que al aplicar la metodología para la extracción automática, la herramienta que se ha usado hasta ahora arroja diferentes resultados, principalmente candidatos a CDs completos, pero también incompletos; podemos ver los correspondientes ejemplos:

Un “interruptor” es un dispositivo cuya función es interrumpir y restablecer la continuidad en un circuito eléctrico. (CLI, CD_356).

“Este modelo” se conoce como modelo potencial porque con él el campo de velocidades se puede considerar como el gradiente de una función potencial, que además cumple la ecuación de Laplace en todo dominio. (CLI, CD_30).

Los CDs incompletos han motivado mi investigación, ya que hasta ahora no hay un criterio para determinar la extensión de los CDs en la extracción automática y existen estructuras y mecanismos discursivos diversos que inciden en la extensión de los contextos. La problemática general deriva de la necesidad de describir y estudiar ciertos recursos lingüísticos, en especial fenómenos discursivos que garantizan que un texto pueda ser correctamente interpretado. En este sentido, comúnmente se habla de *cohesión* y *coherencia* textuales, dos propiedades de los textos en general y en particular de los CDs. La *coherencia* se manifiesta en la unidad total entre las partes del texto y se refiere al sentido completo de un discurso, de tal forma que incide en el proceso total de la intención comunicativa, es decir, en el significado global. Por otro lado, la *cohesión* se percibe en el conjunto de funciones lingüísticas que indican relaciones dentro de un texto (*recurrencia, paráfrasis, sustitución, elipsis, marcadores discursivos*, etc.), o sea, mecanismos lingüísticos que sirven para articular las partes del discurso.

Creemos que en gran medida un CD puede funcionar como una unidad discursiva coherente y cohesionada porque sus partes están articuladas de forma que se completa la intención comunicativa, se proporciona la definición de un término. Según lo que se ha venido diciendo, un CD tiene que tener todos los elementos que lo conforman: término, predicación verbal y definición, pero en el proceso de extracción automática se obtienen candidatos incompletos, en los cuales falta el término o la definición; como el patrón de búsqueda es el patrón verbal, sólo se consideran algunas palabras antes y después de dicho patrón.

Se ha observado que ciertos mecanismos discursivos generan candidatos incompletos, al parecer las *relaciones de correferencias, anáforas, elipsis, paráfrasis* y *marcadores discursivos* son los más frecuentes. Mi trabajo particularmente se enfoca a las relaciones de correferencia y anáforas que desencadenan candidatos incompletos; esto se debe a que en los primeros acercamientos al tema se encontraron recurrentemente estructuras tales como sintagmas nominales y pronombres que no tienen antecedente en el fragmento dado por la extracción automática.

Mi investigación se encuentra en una etapa inicial, se trabajará con un total de 250 CDs, 150 del CLI y 100 del Corpus Técnico del IULA. El corpus de mi trabajo fue elegido al azar, es decir, que los 250 CDs se obtuvieron de un repositorio sin atender a ningún criterio, por ejemplo, tipo de predicación verbal, tipo de

candidato, tipo de definición, etc. Dicho repositorio se ha venido construyendo en el GIL y Bertha ya nos introdujo un poco en el tipo de trabajo de etiquetado que se lleva a cabo.

Los objetivos de mi trabajo son:

1. Localizar, clasificar y etiquetar relaciones de correferencia dentro de los CDs y fuera de ellos cuando el término establece relaciones de correferencia con unidades lingüísticas “externas” al CD.
2. Proponer una tipología de anáforas y relaciones de conferencia implicadas en los CDs.
3. Establecer patrones lingüísticos recurrentes en el uso de correferencias y anáforas que permitan llevar a cabo búsquedas automáticas más precisas.

En este sentido partimos de ciertos presupuestos que han sido determinados a partir de la problemática general. Creemos que a través del reconocimiento y análisis de las relaciones de correferencia y anáforas establecidas entre el término definido y otras entidades nominales y pronominales será posible rastrear los límites de un CD. Además, al delimitar la extensión de un CD, la cohesión y coherencia textuales de éste no deberán diluirse a pesar de que se trata de fragmentos discursivos extraídos de un texto completo. En este sentido, si consideramos ciertos mecanismos referenciales como parte de la estructura de los CDs, será posible plantear patrones útiles en la extracción automática.

Si atendemos un poco al ámbito del problema, más en un sentido teórico, será necesario acudir a ciertos conceptos, tales como *coherencia*, *cohesión*, *correferencia*, *anáfora*, *pronombre*, entre los más importantes. Todas estas nociones están implicadas en el tema de la referencia discursiva ¿Qué es *referencia*? Bueno, se trata de un concepto muy importante para la lingüística, pero, para fines de nuestra investigación, basta con precisar que *referencia*, en el ámbito del discurso, es el fenómeno en el cual las unidades lingüísticas establecen relación con otras unidades lingüísticas y constituyen entonces *relaciones referenciales* intrínsecas al acto comunicativo.

Tradicionalmente, la anáfora ha sido tratada como uno de los principales mecanismos de referencia discursiva; ésta nos permite hacer una referencia abreviada a alguna entidad o entidades, con la confianza de que el receptor del discurso sea capaz de desabreviar la referencia y, por consiguiente, determinar la entidad a la que se alude. De tal manera que la anáfora establece una relación de referencia entre una forma lingüística (expresión anafórica) y un *antecedente*.

Cabe señalar que los pronombres se identifican como la anáfora prototípica. Por otro lado, el concepto *correferencia* ha sido considerado como un tipo de relación que se da entre dos entidades del discurso, se trata de la relación de simetría entre dos expresiones referenciales concretas que en cierto contexto de uso comparten una misma identidad, a pesar del sentido que tengan por sí mismas o en otros contextos. La relación de simetría se da porque dos o más expresiones, concretamente sintagmas o unidades nominales, apuntan al mismo referente extralingüístico.

Los conceptos de anáfora y correferencia son muy próximos; en el primer caso se trata de la relación de un elemento anafórico que reintroduce en el discurso una entidad previamente mencionada, antecedente; en el segundo caso se trata de la relación de dos entidades discursivas que señalan al mismo objeto del mundo. Para fines prácticos, se ha determinado que los pronombres, sin contenido léxico, son el elemento principal de una relación anafórica en la cual éstos acuden forzosamente a un elemento del discurso para completar el significado, mientras que los sintagmas nominales, que poseen contenido léxico, están más próximos de una correferencia en la cual dos o más entidades discursivas remiten al mismo referente extralingüístico.

No cabe duda que para describir los mecanismos referenciales, en este caso relaciones de anáfora y de correferencia, que desencadenan candidatos incompletos en la extracción automática, será preciso establecer un marco teórico y determinar qué vamos a entender por anáfora y correferencia, ya que nos percatamos con este breve panorama de que se trata de un tema

muy complejo que requiere acotar conceptos para describir los fenómenos referenciales que inciden en los resultados de la extracción automática.

Otro de los objetivos de describir los fenómenos que dan pie a los candidatos incompletos es extender las etiquetas de las que nos habló Bertha en la presentación anterior, ya que como vimos, hasta ahora hay etiquetas XML para las partes del CD pero no se consideran las relaciones de referencia de las que hemos venido hablando. Es así que de los 250 CDs elegidos al azar de un repositorio que, como dije, ya existe, se separarán los *Candidatos completos* de los *Candidatos incompletos*. En ambos tipos se rastrearán las relaciones de correferencia y anáforas, y en caso de que sea necesario, se acudirá a los documentos originales para localizar antecedentes o recuperar la relación referencial incompleta. Los candidatos completos se expandirán para poder incorporarlos al corpus de contextos definitorios que también se lleva a cabo en el GIL y posteriormente se integrarán las etiquetas para relaciones de anáfora y correferencia.

Mi trabajo se encuentra en una etapa inicial y sin presentar resultados todavía hemos podido hacer observaciones preliminares.

El ámbito del problema es el siguiente:

Relaciones fóricas en el texto.

- Referencias exofóricas (exáforas) – Denotan entidades ajenas al discurso.
- Referencias endofóricas (endáforas) – Refieren entidades expresadas dentro del discurso, es decir, en un contexto discursivo (anáfora/catáfora).
 - Anáfora – hace referencia a un antecedente ya mencionado, se orienta hacia el texto anterior.
 - Catáfora – Refiere a una entidad que será mencionada en el discurso, se orienta hacia el texto posterior.

Se encuentra aún en proceso lo siguiente:

- Contabilización de resultados de “buenos candidatos” de los tres corpus.
- Contabilización de resultados de “candidatos anómalos” en los tres corpus.
- Tipología de anáforas y de expresiones correferenciales más comunes en los CDs.
- Definir patrones regulares de anáforas y correferencias en “candidatos anómalos”.
- Definir patrones regulares de anáforas y correferencias en “buenos candidatos”.

Es así que en esta presentación he tratado de exponer el tema que ha dado pie a mi tesis de licenciatura, con el fin de mostrar la problemática concreta de los primeros resultados obtenidos de la identificación y extracción automática de CDs. Espero haber logrado un acercamiento a mi trabajo. Sin más que agregar, muchas gracias. ■

HACIA UNA TIPOLOGÍA DE DEFINICIONES BASADA EN ESTRUCTURAS PREDICATIVAS PARA SU EXTRACCIÓN AUTOMÁTICA

CÉSAR ANTONIO AGUILAR
GIL-IINGEN, UNAM
ITZIA EREDI BACA
FFyL / GIL-IINGEN, UNAM

Jeanette Reynoso: Vamos a dar inicio al último trabajo de esta mesa. Es un proyecto en el cual participa Itzia Eredi Baca Ibarra, quien actualmente es pasante de la licenciatura en Filosofía, forma parte también del Grupo de Ingeniería Lingüística (GIL) y está desarrollando un proyecto de tesis titulado: *Análisis de definiciones canónicas. Estudio del modelo aristotélico de definiciones*. Ella participa junto con el doctorando César Antonio Aguilar, que también forma parte del Grupo de Ingeniería Lingüística (GIL), y nos presentan el trabajo *Hacia una tipología de definiciones basada en estructuras predicativas para su extracción automática*.

César Aguilar: Buenas tardes. Vamos a tratar de ir un poco rápido. Básicamente lo que les vamos a presentar aquí el trabajo que estamos realizando en dos proyectos, dentro del proyecto de investigación CONACyT de *Extracción de Contextos Definitorios* patrocinado y, específicamente, lo que nos interesa aquí es abordar definiciones. La forma como nos organizamos Itzia y yo es: yo hago la presentación y, a Itzia, le dejo las preguntas al final. Lo que quieran platicar con ella, está disponible.

Podemos empezar.

El área de extracción conceptual es propiamente una de las líneas de trabajo que existe dentro de la Ingeniería Lingüística, de la cual, ayer, el Dr. Gerardo Sierra ya comentó algunas cosas; por ejemplo, el trabajo de corpus.

En este caso, la idea sería que si ya tengo un conjunto de fuentes de información, sea corpus, o sea Internet, como lo presentó la Dra. Sofía Galicia Haro, también tengo revistas, documentos en papel, que voy a digitalizar, etc. Tengo un universo de información. Por otro lado, tengo un conjunto de herramientas computacionales que facilitan realmente el trabajo de una forma increíble donde, lo que yo hago es buscar un conjunto de datos; en este caso, de patrones, ya sean sintácticos, fonológicos, léxicos, etc. Lo que en última instancia busco son patrones y la constitución de ese tipo de patrones que me ayuda a buscar la información que requiero. Finalmente, obtengo esa información y con ese conocimiento que acabo de ganar puedo crear distintas cosas; por ejemplo: puedo hacer diccionarios, puedo hacer bancos terminológicos, puedo hacer nuevos documentos como revistas, glosarios, terminologías y un largo etcétera de cosas que se pueden sacar. Esto es básicamente una definición muy corta de lo que sería Extracción de Información. Ese es el marco donde se ubica toda esa investigación sobre extraer conceptos.

¿Qué se hace aquí? o ¿cómo se trabaja con esto? La cuestión es que tenemos la fusión de dos métodos de trabajo de algún modo. Por una parte, utilizamos métodos estadísticos, más que para hacer un simple conteo de patrones, o decir, realmente el verbo *definir* o el pronombre *el cual* me sirve para obtener todo este tipo de estructuras. Lo que tratamos de hacer mejor es ver qué tanto una herramienta de cómputo está sacando realmente la información que yo requiero o qué tanto el sistema está mal y tengo que volver a reprogramarlo para que sea óptimo en ese sentido.

Esas estructuras cuando se trabajan en lenguaje natural siguen ciertas reglas; cierto estilo de construcción de información. Pensando como lingüista, esas reglas no son gratuitas, y si puedo describir esas estructuras a partir de un conjunto

de condiciones tales que me permitan suponer que cuando aparece un determinante y luego le sigue un sustantivo, puede convertirse en una anáfora, se convierte en un término, se convierte en una unidad de conocimiento o cualquier otra cosa que le quieran llamar, pero eso está constituido por dos elementos o dos unidades lingüísticas.

Hacer una combinación entre métodos estadísticos y este tipo de reglas o de formulaciones para tratar de predecir cómo están constituidos este tipo de patrones es como constituimos métodos híbridos y, de algún modo, eso es lo que nos ayuda a hacer esos procesos de extracción de información.

Ahora, esto es simplemente para reforzar la idea o la noción de contexto definitorio. Se trata de un fragmento discursivo que contiene un término y una definición, que está compuesto por varias entidades. Esto es lo que nos interesa ver.

Tenemos predicaciones verbales definitorias que son tipos de frases verbales que lo único que están haciendo, como bien señalaron Bertha y Valeria, es unir un término con una definición; lo interesante es que aquí, como frases verbales, de algún modo también subyace la idea de que siguen las reglas, más o menos normales como cualquier frase verbal (entiéndase el verbo más sus posibles argumentos o constituyentes).

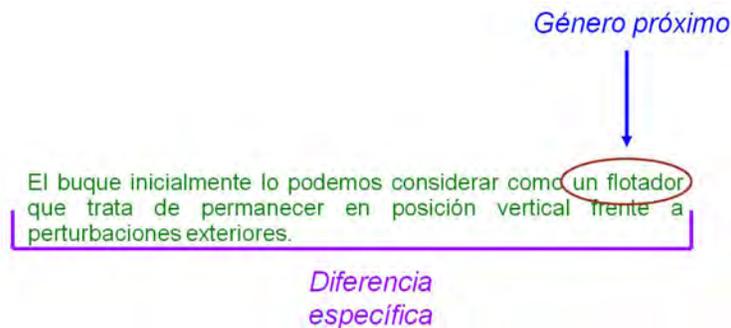
Por tanto, una definición se puede explicar como parte del constituyente de un verbo definitorio. La relación que hay precisamente entre esta predicación verbal y la definición que le sigue después, de algún modo, puede ser analizada y puede ser descrita a partir de reglas sintácticas que, en este caso, tiene un cierto tinte formal, no están cerradas a que se siga con otro modelo lingüístico.

En este caso, los invito a que atiendan a un modelo, el que ustedes quieran: funcional, normativo, descriptivo, tipológico, lexicográfico y todas las combinaciones habidas y por haber. Pueden también analizar este tipo de estructuras y proponer una variante. En este caso, lo enfocamos más que nada a un punto de vista formal, con un modelo de gramática rección y ligamiento, pensando que estamos trabajando con sistemas de cómputo y buena parte de los sistemas de cómputo tienen buena relación con gramáticas formales.

Retomando el punto, lo que a Itzia y a mí nos interesa, por una parte, en el caso de mi investigación, es ver cómo se comportan esas predicaciones verbales y, en el caso de Itzia, si yo tengo un tipo de definición y resulta que esa definición va ligada con cierto tipo de patrón verbal, ¿qué clase de definiciones hay allí?

Para ver qué es una definición, nos podemos remontar unos dos mil años atrás, para ver el caso precisamente de Aristóteles. Más allá de verlo como que la filosofía ya ha establecido que Aristóteles es el padre de las definiciones, lo que nos interesa es ver que la propuesta de Aristóteles de algún modo todavía describe el fenómeno de forma más o menos completa. Fue uno de los primeros, quizá, que atendieron precisamente la relación que habría entre estructuras sintácticas y el tipo de proceso de conceptualización que se hace para presentar estas cosas que se llaman definiciones; obviamente, hay una teoría del conocimiento. Básicamente a Aristóteles lo que le interesó saber es cómo, de algún modo, los seres humanos conocen las cosas o cómo atribuyen ciertas ideas para decir que “esto es una botella”. No nos vamos a meter ahora con eso, simplemente es una descripción. La idea de Aristóteles es que yo tengo un conjunto de estructuras lingüísticas que en este caso podemos llamar *silogismos* o *juicios* y que siguen un proceso de operación.

Al final, lo que interesa es que con esos elementos lingüísticos yo tengo *conclusiones* o *demostraciones*; o codifican conclusiones o demostraciones y eso es conocimiento.



Cuadro 1. Definición aristotélica o analítica

Ahora, este tipo de proceso está relacionado con un conjunto de categorías o con un conjunto de atributos que obviamente Aristóteles propuso, los que conocía en su tiempo o los que le servían para su teoría del conocimiento, pero es amplio y no está cerrado. Cuando hablamos de que la botella es un instrumento que sirve para portar agua, lo que estamos haciendo es poner una entidad y relacionarla con un conjunto de atributos tales como *sirve para hacer tal cosa*. En el cuadro 2, está la lista de las categorías que plantea Aristóteles, que es más o menos lo que proponía en ese momento.

CATEGORÍA	INTERPRETACIÓN DE LA CATEGORÍA	EJEMPLOS
Esencia	"¿Qué es X?"	Hombre Sócrates
Cantidad	"¿Cuánto es X?"	4 metros 2 kilos
Calidad (Qualía)	"¿Qué rasgos tiene X?"	Blanco Soluble
Relación	"¿Con qué se liga X?"	Superior Inferior
Locación	"¿Dónde está X?"	En la escuela En el mercado
Tiempo	"¿Cuándo es X?"	Ayer Hoy
Posición	"¿Cómo está situado X?"	Horizontal Paralelo
Posesión	"¿Qué es propiedad de X?"	Tiene patas Está armado
Acción	"¿Qué hace X?"	Corta Quema
Pasión (Pasivo)	X es afectado por...	Cortado Quemado

Cuadro 2. ¿Qué es una definición?

Desde Aristóteles a la fecha, sobre este tipo de explicaciones, sobre este tipo de problemas se ha regado mucha tinta y básicamente para la lexicografía y la terminología, que son las áreas que nos interesan. Lo que se resalta, por ejemplo, en trabajos como los de Sager, que es un lexicólogo y terminólogo, o un lingüista computacional, es que sí hay una coincidencia muy fuerte entre estructuras lingüísticas y conceptualización; entonces, lo que podríamos suponer es que hay cierto tipo de patrones en la lengua natural que nos permiten codificar información sobre algo y que más o menos esas estructuras mantienen cierta regularidad.

Lo que planteaba Aristóteles en su momento era utilizar relaciones de sujetos-predicados o predicaciones. Retomando un poco esto —un aporte real de un lingüista mexicano Luis Fernando Lara—, estas estructuras se pueden nominar, por ejemplo ecuaciones sémicas y, en todo caso, lo que estaríamos pensando es que el lenguaje natural tiene la capacidad de presentar conceptos; una forma de hacer esto es utilizar predicaciones que corresponderían a preguntas tales como *¿qué es esto? Esto es una botella* y una botella es un objeto que sirve para portar líquidos.

Nos preguntamos si las definiciones presentan realmente conocimientos inamovibles y, por otro lado, cómo suponer que los conocimientos son movibles, que van variando; qué los hace regular, presentar y utilizar ciertas estructuras lingüísticas, dado que si no podemos plantear una regularidad en la organización del conocimiento, es decir, que haya ideas unitarias o estructuras conceptuales unitarias debido a que varía dependiendo del área de trabajo, de cuestiones, de civilizaciones, de cultura, de tiempos, etc. Lo que sí podemos señalar es que hay una cierta regularización en patrones para representarlos.

La definición aristotélica básicamente está compuesta de dos partes: una parte que se denomina *género próximo* que, como ya se lo había explicado, es simplemente un descriptor que abarca un conjunto de objetos que pertenece a un mismo conjunto o universo de cosas y, la *diferencia*, que es simplemente la característica de los atributos que hacen particular ese objeto en ese universo de cosas. Ésa es la definición. A nivel de formalismo lógico lo que estaríamos planteando es que esto es representable y ciertas estructuras que más o menos podemos formalizar en ese sentido son, por ejemplo, a partir de lógica de predicados. Retomando un concepto tradicional, lo que tendríamos sería una relación entre un sujeto y un conjunto de atributos o de predicaciones y, en ese caso, un nexo o una cópula que operativamente lo único que está señalando es que sirve para ligar al atributo con la entidad o con el sujeto.

Podemos manejarlo como cuantificadores universales existenciales y tendríamos dos tipos de definiciones diferentes, pero estructuralmente estamos partiendo de una relación sujeto predicado. Decir que todas las computadoras tienen tales características y sólo si cumplen con esas características las voy a definir como computadoras, o decir que algunos lenguajes de programación ocupan tales cosas en un universo de otros lenguajes de programación, de algún modo, es plantear dos formas de conceptualizar. El punto es que ambas ocupan una estructura o se puede ver que la relación sujeto-predicado, estructura predicativa, subyace dentro de ellas.

Dentro del Grupo de Ingeniería Lingüística, en trabajos anteriores, planteamos una posible tipología de definiciones en donde partimos del modelo canónico *género próximo* y *diferencia específica* o si lo prefieren llamar *definición analítica* o *definición aristotélica*. Podemos tener cuatro tipos de definiciones. Por un lado, si nada más nos quedáramos con la estructura que representa el género, tendríamos, posiblemente, una definición de tipo sinonímica en el sentido de que tal cosa es equivalente a otra. Entonces, la botella es equivalente a un recipiente o, por el otro lado, si yo dijera que una botella es un recipiente y solamente me quedara ahí, lo único que está señalando es una relación de superordinado o de hiperónimo, y decir que la botella pertenece a un conjunto más grande, ¿qué diferencia a la botella de todo ese conjunto que son recipientes? No se puede afirmar con exactitud, pero ahí está la definición y, por el otro lado, si yo solamente tomara en cuenta la diferencia específica y quitara de algún modo ese *genus*, lo que me queda son definiciones en donde, simplemente, puedo describir para qué sirve la botella o puedo decir que la botella está compuesta por una tapa,

el contenedor, la forma, etc. Y, en todo caso, estoy haciendo una definición en donde describo las partes que lo componen ya sea por extensión o por relaciones de meronimia; eso es básicamente lo que estamos proponiendo aquí.

A continuación presentamos los ejemplos del corpus:

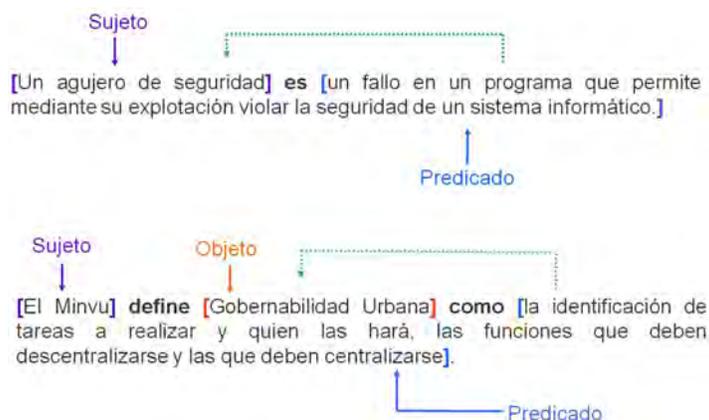
El término “sembrar partículas” [Término] se refiere a [PVD] introducir en el flujo de observación un trazador que es transportado por el mismo, el cual no lo altera visiblemente y es observado fácilmente [Definición].

Cuando por un circuito circulan corrientes de secuencia positiva, la impedancia del circuito, se denomina [PVD] “impedancia” [Término] a la corriente de secuencia positiva [Definición].

El “relevador auxiliar” [Término 1] y “mástiles” [Término 2] son [PVD] los sistemas de protección [Definición].

La “Terminal de Contenedores” [Término] cuenta con [PVD] dos muelles de atraque, el muelle del Bufadero y el muelle del Dique del Este [Definición].

Esto también tiene un equivalente sintáctico; una forma de tratar de describir cómo se constituyen esos patrones del lenguaje natural desde un punto de vista de gramática generativa o de gramática de rección y ligamientos; es decir, las estructuras predicativas se componen de una unidad que funciona como núcleo de la predicación, en este caso, un verbo o alguna entidad que pueda cumplir esa función, como dos puntos, paréntesis, etc., y regularmente habrá un especificador o un sujeto que puede ser, en este caso, el objeto que se va a definir; el término, o puede ser también alguien que de algún modo a nivel semántico sea el que realice el proceso de conceptualización, un ejemplo podría ser: “Lafourcae define tal cosa como...”. Eso vendría a ser el sujeto que realiza la acción de definir. Por el otro lado, si esto ocurre o si aparece este agente, entonces la posición que anteriormente tenía el término como sujeto de predicación puede pasar a ser un objeto; ejemplo: “Chosmky define la gramática generativa como una gramática de X características”. Ahí es en donde en la posición de objeto tradicional se encuentra gramática generativa, y, finalmente, lo que traería como un predicado sería justo la definición como el conjunto de atributos que vamos a introducir. Tienen estructuras de frases, pueden ser descriptibles.

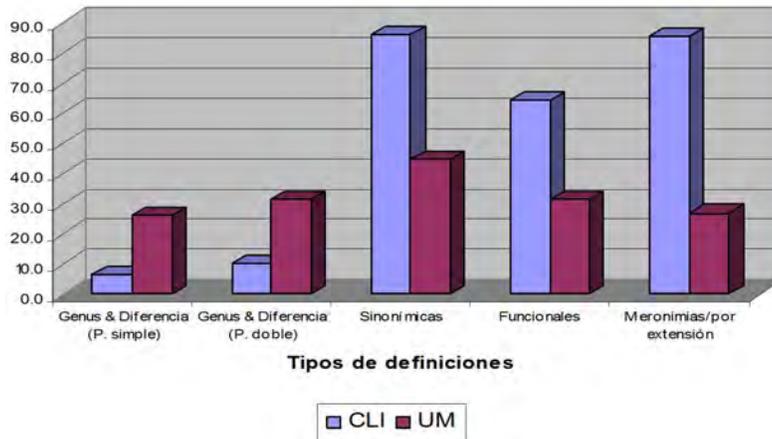


Cuadro 3. Predicación simple y secundaria

No se trata de una lista cerrada, simplemente lo que tratamos de señalar aquí es que ese tipo de verbos combina muy bien con las estructuras predicativas, sea que tengamos una estructura de predicación simple o de predicación secundaria.

Ahora, ¿realmente esto ayuda a sacar definiciones o podemos extraer automáticamente definiciones o Contextos Definitorios? Lo que aprovechamos aquí, en este experimento, es comparar dos corpus: uno de informática y otro de ingeniería, donde estamos tratando de ver qué tanto con estos patrones yo puedo obtener realmente definiciones según la tipología que les presente y más la idea que veíamos es que no va tan mal, está más o menos entre un 20% o 40% de posibilidades, aunque hay mejores candidatos que otros. En ciertos contextos o dada la especificidad o la capacidad multiuso que tendrían ciertas unidades, por ejemplo, definiciones analíticas con un verbo copulativo como *ser*, que es muy productivo para todo el lenguaje natural. Estrictos censos no arrojan muy buenas definiciones porque la generalidad amplía el margen de búsqueda; sin embargo, verbos como *concebir* o *llamarse* o como *servir para*, entre otros, sí son muy restrictivos, nos pueden dar mejores candidatos y obtener definiciones funcionales o extensionales, o de algún otro tipo.

ANÁLISIS COMPARATIVO DEL CORPUS CLI vs. UM



Cuadro 4. Comparación de frecuencias

Lo que estamos haciendo aquí es continuar con el reconocimiento de patrones; lo que tendríamos entonces es que, por ejemplo, para el caso de las definiciones analíticas, tendríamos que ciertas estructuras como, por ejemplo, el *género* y la *diferencia*, pueden ser representados con frase nominal más oración completamente o de relativo con X características y es simplemente dar como que la extensión, considerando que si la predicación verbal de algún modo es la que introduce la definición, la expansión de esa predicación verbal con su correspondiente definición está estructurada en una cadena sintáctica.

En el cuadro 5, se trata del caso de una predicación simple con la cópula *ser*.

Con el verbo *permitir* (cuadro 6), tipológicamente, hay verbos que tienen cierta estructura o predicaciones verbales; mejor dicho, tienen, con su definición, cierta estructura y otras varias.

Para, terminar y aprovechando la ocasión en un espacio como lo es la Facultad de Filosofía y Letras, podríamos pensar que hay cierto patrón canónico para describir qué es una definición. Identificando ese patrón canónico podríamos establecer que, quizá, con una matriz de rasgos discursivos sintácticos y todas sus combinaciones seríamos capaces de señalar una definición desde un punto de vista lingüístico; se admiten variaciones y conforme me vaya acercando o alejando de la variación, puedo plantear una cadena de prototipos.

VERBO SER (PREDICACIÓN SIMPLE)

DEFINICIÓN	GÉNERO	DIFERENCIA	TOTAL
ANALÍTICA	Frase Nominal=Determinante, Cuantificador, Demostrativo + Nombre + (Frase Adjetiva o Frase Prepositiva)	Pronombre relativo (que, la cual, el cual, ya que, por el que, etc.) + Oración.	31
ANALÍTICA	Frase Nominal=Determinante, Cuantificador, Demostrativo + Nombre + (Frase Adjetiva o Frase Prepositiva)	Frase Prepositiva=Preposición (para, de, en, entre o con) + Nombre+Frase Prepositiva, Frase adjetiva o Oración.	71
ANALÍTICA	Frase Nominal=Determinante, Cuantificador, Demostrativo + Nombre + (Frase Adjetiva o Frase Prepositiva)	Frase Adverbial=Adverbio + Nombre + Frase Prepositiva, Frase Adjetiva o Oración.	7

Datos obtenidos del *Corpus de Informática en español* (CIE) de la Universidad de Montreal

Cuadro 5. Algunos experimentos (verbo **ser**)

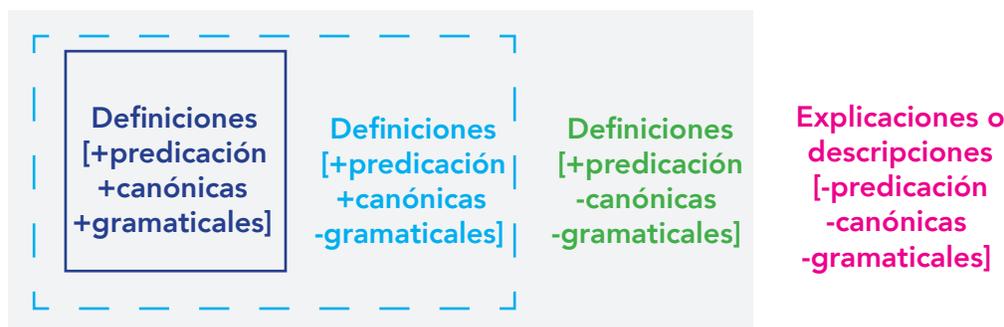
VERBO PERMITIR (PREDICACIÓN SIMPLE)

DEFINICIÓN	DIFERENCIA	TOTAL
FUNCIONAL	Verbo infinitivo + Oración	69
FUNCIONAL	Frase Prepositiva = Preposición (para, de, en, entre o con) + Nombre + Frase Prepositiva, Frase Adjetiva o Oración	24
FUNCIONAL	Frase Nominal = Nombre + Frase Prepositiva, Frase Adjetiva o Oración	14
FUNCIONAL	Frase Adverbial = Adverbio + Nombre + Frase Prepositiva, Frase Adjetiva o Oración	2

Datos obtenidos del corpus CIE de la Universidad de Montreal

Cuadro 6. Algunos experimentos (verbo **permitir**)

CONCEPTOS



Cuadro 7. Representación de conceptos

Al final de ese prototipo de definición canónica textual, como aparece en el diccionario Larousse o una tesis de lo que ustedes quieran, hay otras cosas, que son números, fórmulas, signos, etc., que son complicadas de admitir como definiciones en un sentido lingüístico, pero no dejan de ser, por ejemplo, explicaciones o descripciones de conceptos, y retomar un poco la idea de Aristóteles. De algún modo, siguen cierta estructura más o menos formal, una estructura de predicación, quizás otras estructuras posibles, pero sí hay un cierto *continuum* en ese sentido. Podríamos hablar de ciertas estructuras que se apegan más a un modelo lingüístico y otras estructuras conceptuales que no son lingüísticas y, sin embargo, están presentes en los conceptos. Eso es todo y muchas gracias.

Jeanette Reynoso: Podemos iniciar esta última ronda de comentarios.

SECCIÓN DE PREGUNTAS

Margarita Palacios: Estoy fascinada con su trabajo. Fascinada en el sentido de todas las perspectivas que nos están abriendo. Fundamentalmente, creo que a la gente que hacemos análisis del discurso. ¿Fundamentalmente, por qué? Porque tu último cuadro, elocuentemente demuestra que entre menos predicación y menos canónico y menos gramatical hay menor conocimiento. Después de eso, una duda, sería esto: cuando hablaron de explicaciones, supongo que de la que parten es de las oraciones explicativas de relativo, ¿por ahí empiezan?

César Águilar: Sí.

Margarita Palacios: ¿Ahí han encontrado adjetivas, pero no de relativo, sino con preposición y verbo?

Cesar Águilar: Sí; de hecho, es simplemente por tratar de cumplir un medio de predicación a partir de frases. Si continuamos con la idea de rección, estricto censo, lo que tendríamos que hablar es de frase adverbial, frase complementante, frase prepositiva, etc. Y sigue la secuencia si no aparece frase complementante con

pronombre relativo como lo presentaba Valeria. Tu siguiente opción puede ser frase prepositiva o frase adjetiva con la introducción del adjetivo tal cual, después del verbo. La distinción podría ser por frecuencia estadística, por patrón canónico, podría ser la construcción completa este verbo: frase nominal, frase complementante y, no sé, alguna otra posibilidad que esté compitiendo y abra las posibilidades de que haya la continuidad de una cadena sintáctica con distintas variantes. Esa sería la idea, pero sí se arrojan datos que corroboran lo que estabas preguntando.

Margarita Palacios: En ese sentido, cuando se dice que a mayor conocimiento de mundo hay más predicación, más canonización y más gramaticalización, y a menor conocimiento nos tendríamos que ir al espacio de explicación o descripciones, donde hay menos predicación, menos canonización, menos gramaticalización. ¿Esto se cumple en algo de lo que ves?

César Águilar: Yo creo que dependería. Si este tipo de estructuras o cadenas sintácticas en 100% de los casos van a ser completamente definiciones; verbigracia, de tipo analíticas o de otro tipo. Por ejemplo, en un caso como *Bill Gates considera que Microsoft es una empresa que le produce mucho dinero*, por más estructura predicativa que tengas, sería que algún lexicógrafo o algún terminólogo la viera como una definición *per se*. Sin embargo, quizá lo que podría suceder ahí es que pasaría a la categoría de explicación de concepto o de algún otro tipo de cuestiones; no sé, yo creo que se podría ver como ciertas excepciones a la regla.

Si te sale cierta combinación adjetiva, no la tomes en cuenta. Sin embargo, la idea sería que en algún momento, por lo menos en la propuesta que se puede plantear a nivel sintáctico, o en la propuesta que podría hacer Itzia, desde un punto de vista lógico, es que pareciera que si hay una estructura más o menos formalizable con respecto a cómo dar una definición en lenguaje natural y, una vez descrito, se podrían facilitar estas áreas de detección de conceptos o de detección automática. Sí, a lo mejor una cierta clasificación difusa y resaltando el área de difuso: ¿qué tanto es una cosa definición con más predicación, y que tanto con menos no lo es? Es como preparar un pastel, qué tanto de royal le pongo y qué tanto de harina le quito. Sería el énfasis de lo difuso y en la capacidad de mover categorías.

Pregunta 2: ¿Estos proyectos ya se hicieron en otros países y con qué tanto éxito?

César Águilar: Sí, sobre todo para Procesamiento de Lenguaje Natural.

Jeanette Reynoso: Eso es todo. Vamos a tomar un descanso antes de pasar a la siguiente mesa.

DESARROLLO DE UN SINTETIZADOR DEL HABLA PARA EL CORPUS HISTÓRICO DEL ESPAÑOL DE MÉXICO

AMARANTO DÁVILA JÁUREGUI/FI, UNAM
ABEL HERRERA CAMACHO/FI, UNAM
ALFONSO MEDINA URREA/GIL-IINGEN, UNAM
ADOLFO HERNÁNDEZ HUERTA/FI, UNAM

Dr. Alfonso Medina Urra: Buenas tardes, vamos a dar inicio a la primer plática de esta mesa, la cual va a ser presentada por Amaranto Dávila Jáuregui y Adolfo Hernández Huerta, y se titula: *Desarrollo de un sintetizador del habla para el Corpus Histórico del Español de México*.

Amaranto Dávila: El objetivo principal de este trabajo es desarrollar un sistema adecuado para síntesis de voz con la finalidad de proveer una base de datos de audio para el sitio web del CHEM (Corpus Histórico del Español de México), el cual es una colección de documentos que representan el español de México desde el siglo XVI al XIX. Este corpus continúa en desarrollo y, esencialmente, es una fuente documental con información sociolingüística para investigación científica.

El sistema de síntesis está basado en un método de difonos TD-PSOLA desarrollado en la UNAM. Además, están incluidos los sufijos y las palabras gramaticales encontradas dentro del corpus. Dado que en esta primera etapa restringiremos el sistema al siglo XIX (el español de México ha variado muy poco desde entonces), incluiremos los sufijos más prominentes, aproximadamente 300, y alrededor de 300 de las palabras más frecuentes del español contemporáneo con la finalidad de obtener más naturalidad en el habla, sin aumentar demasiado la base de datos.

La síntesis de voz puede ser producida mediante la concatenación de unidades grabadas y seleccionadas de un solo hablante de la base de datos. En los segmentos del sistema concatenados de la grabación ya hecha se juntan cadenas para formar palabras, frases, etc. Este método provee más naturalidad en la generación de habla, pero también causa fallas de audición. Es usual suavizar la salida de la onda sonora en el punto de concatenación con la finalidad de ganar naturalidad.

Actualmente desarrollamos un sistema de tipo texto-habla para el español basado en difonos, sufijos y concatenación de palabras. Este sistema consta de dos partes: una que traduce texto a fonemas y encuentra la vocal acentuada, y otra que toma la entrada fonética y genera una simple salida en audio.

Hemos añadido un proceso PSOLA para mejorar la naturalidad, aunque genera cierta degradación de la señal. Los resultados reportan un habla natural con pequeñas desventajas, comparada a la de otros sistemas internacionales más complejos para el español.

Uno de los mayores problemas con los sistemas de concatenación es cómo tratar con los límites entre segmentos. Está claro que minimizando el número de ocurrencias de los límites se mejora la calidad del habla reduciendo el número de fronteras involucradas y, por supuesto, usando unidades más largas. El punto es: entre más larga sea, mayor es el número de fronteras y los detalles en éstas.

Nuestro corpus está construido básicamente por difonos. También incluye las palabras más frecuentes y los sufijos más prominentes del español hablado en México. Es sabido que el máximo número de difonos para un lenguaje determinado es el cuadrado de su número de fonemas. Sin embargo, de acuerdo a la fonotáctica de un lenguaje, algunos difonos, de hecho, pueden no ocurrir. Tenemos que tratar con 24 fonemas y 350 difonos. De acuerdo con la investigación llevada a cabo en esta universidad para el CHEM y el Corpus del Español Contemporáneo (desarrollado en El Colegio de México), un conjunto de alrededor de 500 palabras gráficas existe con cierta prominencia gramatical, medido por el alto nivel de entropía de sus contextos (una media más fina de prominencia que la mera frecuencia). Estas son principalmente palabras funcionales como preposiciones, artículos, pronombres, determinantes, conjunciones, pero también palabras de contenido muy frecuentes como nombres específicos, adjetivos, adverbios y verbos en infinitivo.

Como acabamos de señalar, usaremos difonos, sufijos y palabras. Los difonos son nuestras unidades básicas de síntesis para simplificar la complejidad del sistema; estos segmentos serán obtenidos de frases grabadas conservando las características suprasegmentales implicadas.

Usaremos difonos como nuestra unidad básica porque ellos, idealmente, modelan bien las uniones. A pesar de la reserva de un gran número de permutaciones de las unidades, esta elección es mucho más conveniente que usar sílabas porque necesitarían un número mayor de variaciones.

Una primera suposición puede sugerir que entre más larga sea la medida de la unidad, menos problemas nos darán los errores por las grandes cantidades de contenido semántico que capturará cada unidad. Si este es el caso, los errores al unir unidades pequeñas, como los fonos, serían los más críticos para la percepción. Ésta es la razón por la que preferimos los difonos como unidad básica sobre los fonos —la estructura del modelo fue diseñada para reducir la captación de errores, que pueden ocurrir en uniones coarticuladas entre segmentos—.

Los sufijos pueden ser concebidos como grupos muy frecuentes de letras añadidos al final de las raíces, bases e incluso palabras completas. Típicamente, los sufijos y las secuencias de estos pueden flexionar palabras (*casual, casuales; compro, compra, compran, compramos, compró, compraría, compraba, compraban, comprándoselas*, etc.) o crear unas nuevas (*casual, casualidad, casualmente*). Así, los sufijos modifican y extienden el significado y la categoría gramatical de una palabra. Unidades de este tipo, particularmente más largas de dos, son ideales para su inclusión en el sistema porque el número de puntos de concatenación dentro de la palabra decrece cuando son combinados con difonos. En otras palabras, como los problemas de coarticulación se daban entre unidades individuales, los sufijos de más de tres fonemas redujeron considerablemente estas dificultades. La clave para un exitoso conjunto de sufijos relevantes es encontrarlos dentro del corpus propuesto en el lenguaje señalado (en contraposición a usar libros de texto como base de datos), el cual puede privilegiar, por ejemplo, al español peninsular, el cual exhibe flexiones de sufijos muy productivas, a saber, como *-éis, -áis*, lo cual es raro en todos los otros dialectos del español contemporáneo. Así, como nosotros elegimos sufijos que son muy comunes en el lenguaje señalado, creemos que, finalmente, darán más naturalidad al habla sintáctica.

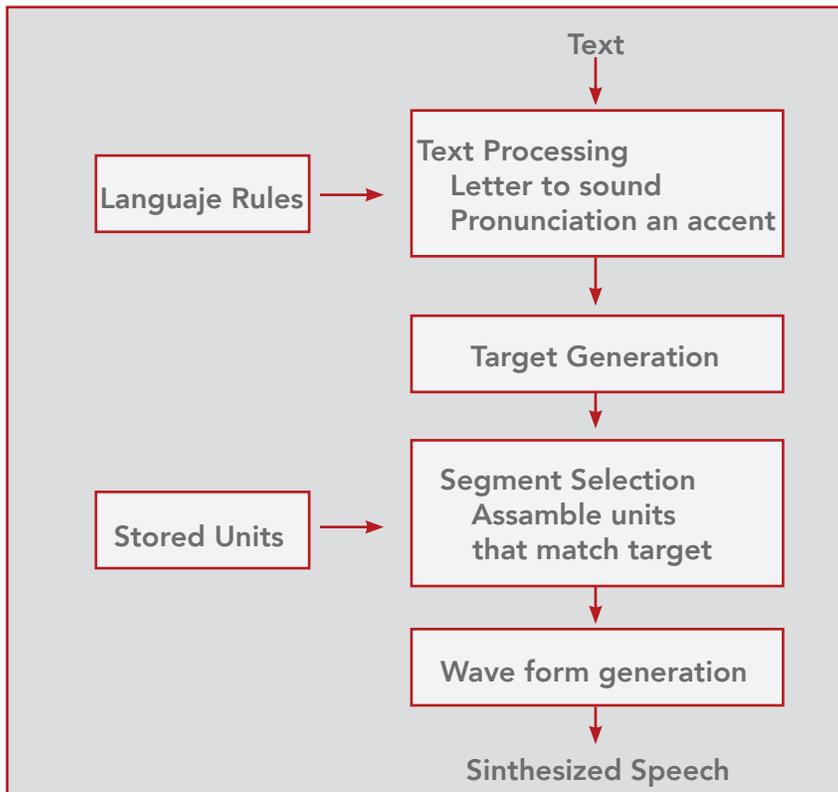
Un argumento similar puede ser propuesto para justificar la inclusión de las palabras gráficas más prominentes en el lenguaje según un corpus. Un subconjunto de las quinientas muestras mencionadas arriba fue elegido para la grabación.

Finalmente, para determinar cuáles difonos grabar, los documentos del CHEM para el siglo XIX fueron analizados automáticamente.

La calidad de los sintetizadores de concatenación depende en alto grado de la calidad de las unidades del habla grabadas. El corpus será grabado en formato WAV *sampling* con velocidad de 22,050 KHz y con 16-bit. Las sesiones de grabación se harán en un estudio profesional; en este primer escenario, el hablante será masculino.

La tarea de etiquetado consiste en analizar las ondas sonoras y los espectrogramas, así como hacer anotaciones a las ondas sonoras del habla grabada, con la finalidad de extraer información sobre las unidades de las frases. Los niveles de prosodia son también necesarios para darnos información sobre el tono y los acentos. Las etiquetas de las palabras consisten en marcadores de tiempo en el principio y final de las palabras. Las etiquetas de tono son representaciones simbólicas de la melodía de las palabras. Este trabajo es usualmente realizado por herramientas de etiquetado automático del habla, debido al tamaño de la base de datos. Para un etiquetado fonético, los reconocedores de habla son usados en un modo de alineación forzada, donde el reconocedor encuentra los límites entre los segmentos. Las herramientas de etiquetado prosódico automático trabajan desde un conjunto de características motivadas lingüísticamente (como las duraciones y las escalas de frecuencia fundamental), más algunas características binarias encontradas en el lexicon (como palabras de acento final contra palabras de acento inicial).

Las unidades de grabación fueron marcadas manualmente debido a que un desarrollo hubiera requerido más tiempo.



Cuadro 1. Diagrama del blog del sintetizador TTS

El sistema tiene la estructura que se muestra en el cuadro 1. Tiene cuatro módulos: procesamiento de texto, generación de objetivos, selección de segmentos y generación de ondas sonoras.

El procesamiento de texto consiste en una representación fonética de una entrada de texto (en formato .txt / .doc); para ser sintetizado, se debe mantener toda la puntuación y las marcas de acentos para preservar

pistas de entonación. También, en texto adicional, será introducida la conversión de datos y números a una cadena fonológica. Los símbolos no relacionados al significado del texto, como los números de páginas y otras anotaciones en el documento original, serán omitidos en archivo de salida.

El primer paso es trasladar la entrada, cualquiera que esta sea, a texto plano, un proceso conocido como normalización de texto. En esta etapa, todos los números y abreviaturas deben ser escritos como palabras completas a través del uso de tablas de referencia y todos los formatos de texto deben ser removidos.

LETRA	SONIDO	LETRA	SONIDO	LETRA	SONIDO
A	/a/	I	/i/	O	/o/
b,v	/b/	J	/x/	P	/p/
D	/d/	K	/k/	s,z	/s/
E	/e/	M	/m/	T	/t/
F	/f/	N	/n/	u,ü	/u/
H	Mute	Ñ	/ñ/		

Cuadro 2. Letra que representa directamente cada fonema

El proceso está hecho de dos maneras: la lectura puede ser también a través del uso de *look-up tables* o puede ser obtenida a través de un conjunto de reglas de pronunciación.

El español tiene distintas letras para hacer asociaciones de sonido y la mayoría de las letras tiene una relación con su correspondiente sonido (véase cuadro 2). Algunos casos necesitan reglas especiales de pronunciación (véase cuadro 3). La mayor parte de este proceso puede cumplirse a través del uso de reglas.

La ortografía española que se encuentra dentro de los textos de los siglos más tempranos del CHEM exhibe importantes variaciones. Por ejemplo, la palabra *indio* tiene formas distintas como: *yndio*, *jndio*, *yndjo*, etc. Sin embargo, la normalización de estas formas no implicó un esfuerzo para el siglo XIX, cuando las representaciones ortográficas comenzaron a ser más estables. También, desde que el sistema fonético del español mexicano permaneció igual desde aquel siglo, el proceso de transcripción es fácil de comprender, al menos para esta primera etapa del proyecto.

El formato de salida es .txt, así que la transcripción no usa el Alfabeto Fonético Internacional (IPA), sino el código ASCII. Para nuestros propósitos, no importa si el IPA no es utilizado.

En español hay muchas palabras que son préstamos, como las palabras tomadas de lenguas indígenas. Hemos tratado eso usando diccionarios y tablas de referencia. Esto es especialmente verdadero en México donde hay un uso extenso de palabras tomadas del náhuatl y otras lenguas nativas.

Actualmente hemos obtenido características prosódicas del texto usando solamente las marcas de puntuación; las sílabas tónicas en español pueden o no estar marcadas gráficamente, lo cual tiene que ver con su posición en la palabra, de acuerdo con las siguientes reglas:

1. Si la palabra no tiene un acento gráfico y termina en -n, -s o vocal, la vocal con acento está situada antes de la última sílaba.
2. Si la palabra no tiene acento gráfico y no termina con -n, -s o vocal, el acento está en la última sílaba.

Dadas estas reglas, si una palabra no tiene tilde, deberíamos ser capaces de hayar la sílaba tónica de la palabra. Nosotros usamos un pequeño autómata para encontrar las dos últimas sílabas de una palabra.

LETRA	REGLA
C	/k/ si va seguida de "a", "o" o "u" /s/ si va seguida de "e" o "i" Si va seguida de "h" en 'ch' se considera un solo fonema
Ch	/C/
G	/g/ si va seguida de "a", "o", "ü", "u" /x/ si va seguida de "e", "i" Si va seguida de "u" se necesita una siguiente letra: Si es "ue" o "ui" se omitirá la "u"
L	/l/ excepto cuando va seguida de 'l' Si va seguida de "l" en 'll' se considera un solo fonema
Li	/l/ si es la última letra en la palabra /L/ cualquier otro caso
Q	/k/, si va seguida de 'ue' o 'ui' se omitirá la 'u'
R	/r/ si va entre vocales /R/ otros contextos Si va seguida de "r" en "rr" se considera un solo fonema
Rr	/R/
X	Se transcribe a 2 sonidos: /k/ /s/
W	Usada solamente en préstamos

Cuadro 3. Reglas que requieren un análisis futuro

Un módulo de generación de objetivos predice las oraciones de frases y el acento de las palabras y, de esto, se generan objetivos.

Adolfo Hernández: En la selección de segmentos, este módulo selecciona las mejores unidades de acuerdo con los objetivos. Cuando hemos completado el análisis de textos para las palabras y tiene una secuencia de difonos, sufijos o palabras y metas prosódicas, tomamos la secuencia apropiada de unidades de la base de datos de difonos. Después, necesitamos concatenar los difonos juntos y ajustar la prosodia (entonación y duración) de la unidad de secuencia para combinar los requerimientos prosódicos del habla.

En el sistema previo, hicimos pruebas para unir segmentos directamente sin más procesamientos. La salida en forma de habla además de ser muy monótona tiene un alto grado de inteligibilidad porque no hubo procesamiento que indujera distorsión de la señal, la cual es un verdadero problema en la mayoría de los sistemas libres y, también, en sistemas comerciales.

A través de la concatenación directa de las unidades esperamos obtener una buena calidad de síntesis de habla debido al uso de unidades más largas. No obstante, para incrementar la calidad de la síntesis de voz —esto es, que las transiciones entre las unidades no sean percibidas— será necesario hacer un procesamiento sobre el resultado por medio del algoritmo TD-PSOLA.

Este algoritmo es usado porque puede ser aplicado directamente a la señal de audio, sin la necesidad de una extracción paramétrica, como en el caso de LPC y otros algoritmos similares usados. Para trabajar con este algoritmo, necesitamos agregar una fase de extracción de marcas entonativas para la creación de la base de datos. Este paso se hace *offline*, por lo que no lleva penalización en tiempo real.

Para la extracción de marca entonativa hemos usado un programa dinámico basado en el algoritmo presentado por Vladimir Goncharoff y Patrick Gries. Este algoritmo fue diseñado para ser fácil de entender y no sacar prácticamente ningún error de extracción. Otro punto a favor es que el código fuente de este algoritmo se distribuye gratuitamente.

Durante el tiempo real, la señal sonora es dividida en ventanas de los segmentos con la frecuencia fundamental centrada. La longitud de estas ventanas debe ser más larga que el periodo de frecuencia fundamental pero proporcional a éste. Para realizar modificaciones estos segmentos se alinean a una nueva posición y se enlazan. Los valores de normalización son calculados a partir de las ventanas para eliminar modificaciones de energía.

Puede también ser necesario duplicar o eliminar segmentos para mantener la duración de la señal en diferentes tonos, o para acomodar la modificación del tiempo o la señal simultáneamente para modificaciones de frecuencia fundamental. Para minimizar discontinuidades en la concatenación de puntos usamos este algoritmo para modificar la frecuencia de cada segmento. Los segmentos están sincronizados, así que su inicio y fin corresponden a una marca de frecuencia y su magnitud es igualada.

Aunque este proceso es incapaz de eliminar todas las discontinuidades, éstas son magníficamente minimizadas, pero al costo de una pequeña distorsión comparada con otros sistemas basados en OLA, con un proceso más lento y con un proceso de carga también más pequeño, para alcanzar una velocidad aceptable en equipos más lentos.

Finalmente, esperamos un habla natural mejor que aquellos sistemas que usan exclusivamente difonos, pero con algunas desventajas en la pérdida de claridad comparado a los mejores sistemas internacionales. De cualquier modo, este sistema usa una pequeña cantidad de memoria comparado con aquellos sistemas, los cuales no están diseñados para el español mexicano. Estamos aún trabajando para mejorar la claridad sin perder naturalidad.

Alfonso Medina: Muchas gracias por su presentación. Si no hay preguntas vamos a pasar a la siguiente mesa.

IDENTIFICACIÓN AUTOMÁTICA DEL LENGUAJE HABLADO SIN INFORMACIÓN FONOTÁCTICA

ANA LILIA REYES HERRERA
INAOE

Alfonso Medina: Ahora nos va a hablar Ana Lilia Reyes Herrera, que está haciendo su doctorado en el INAOE y su ponencia se titula: *Identificación automática del lenguaje hablado sin información fonotáctica*.

Ana Lilia Reyes: Gracias, estoy aquí en colaboración con el Dr. Luis Villaseñor Pineda. En el contenido de mi presentación voy a hacer una introducción: hay dos enfoques para resolver ese problema. Vamos a identificar más o menos cada uno de ellos, cómo se solucionaron y los resultados que se obtuvieron.

¿Qué es la identificación del lenguaje hablado? Es identificar el idioma sin importar el hablante, ni lo que se está diciendo.

Humanamente, lo hacemos cuando ya conocemos un idioma: es, por ejemplo, cuando decimos *esto es alemán, esto es el chino*, dependiendo de cómo se lo escuche, o si tengo algún conocimiento del idioma, de tal forma que lo relaciono. Esto se hace computacionalmente. La pregunta es: ¿para qué me sirve identificar el lenguaje? Existe gran cantidad de aplicaciones sobre esto. Primero, como un pre-procesamiento de un sistema: supongamos que hay un proyecto grande, que ustedes están en México y quieren hablar a Japón, ustedes no saben hablar japonés, en este caso, hablan primero, hablan telefónicamente y la persona en Japón escucha en japonés lo que están hablando en español y ustedes, a su vez, escuchan en español lo que él está diciendo en japonés.

Para hacer eso necesitamos un reconocedor de habla, saber exactamente qué se está diciendo. Actualmente hay muy pocos, y los que están en uso son muy limitados. Un ejemplo podría ser la *secretaria*. Digamos que hay un doctor que habla, y de la voz se está pasando al texto, son sistemas muy limitados. Para este caso, lo que nosotros proponemos es primeramente que identifique el idioma y posteriormente que haga las traducciones, lo que sería la primera parte del proceso.

En la segunda, las compañías de teléfonos han tenido muchos identificadores del lenguaje; ustedes hablan, identifican qué idioma es e inmediatamente los pasan al traductor correcto.

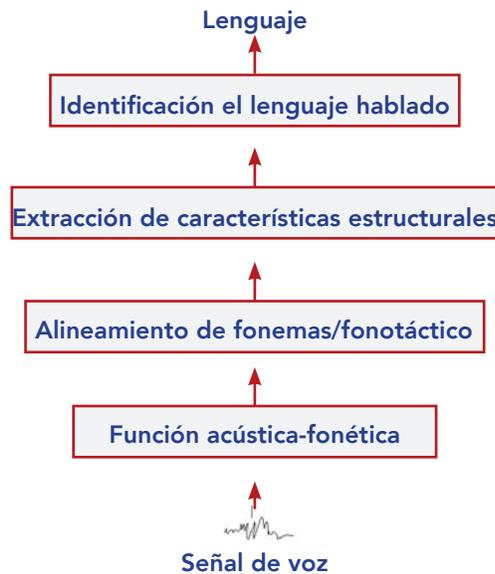
En EU está el 911 para emergencias; en este caso está comprobado que la gente cuando tiene graves problemas o está en una situación de emergencia recurre a su lengua natal, en este caso, cuando ustedes están en un caso de emergencia recurren a su idioma. A veces, se pierden minutos valiosos en tratar de identificar en qué idioma se da la conversación, y si empleamos un identificador de idioma en ese momento le diría a la operadora qué lenguaje se está hablando; se pasa, entonces, al traductor correcto y se soluciona el problema.

Existen actualmente migrantes monolingües en México y en EU, e indígenas que no hablan más que su idioma. Podemos entrar a la página *¿Qué lengua hablan?* Pongamos por ejemplo que hay un indígena

monolingüe con un problema legal o médico, y en ese momento dado se sienta con una persona que sabe usar el sistema; esta persona no tiene que leer todo el instructivo para saber qué hacer, simplemente tiene que hacer oír a la persona grabaciones. Hay 69 lenguas indígenas documentadas, y solamente existen 39 en este archivo. La persona tiene que escuchar la grabación. En un caso, verbigracia, la grabación podría decir *dame la mano*, y si la persona no te extiende la mano, quiere decir que no te entendió; si tuviéramos un identificador de lenguas indígenas en ese momento, ustedes sabrían qué idioma es.

Lo importante es tener un identificador de idiomas y lo último propuesto es el monitoreo. Realmente EU ha invertido mucho en esto para lo que es el terrorismo; esto es, ustedes hablan por un teléfono que está intervenido y, entonces, el gobierno sabe de qué país son ustedes.

¿Cómo resolvemos el problema? Existen dos formas: el primer enfoque dice que hay que tener la señal de voz digitalizada para diversos procesos, de ahí ustedes tienen que segmentar la señal de voz; posteriormente, se siguen los siguientes pasos: primero hay que segmentar la señal en fonemas para saber cada palabra, generalmente se sabe que cuando hablamos lo hacemos en fonemas, ya de ahí se van construyendo los lexemas y la sintaxis; después, pueden decir qué idioma es. Todos esos procesos se hacen para hacer la identificación del lenguaje.



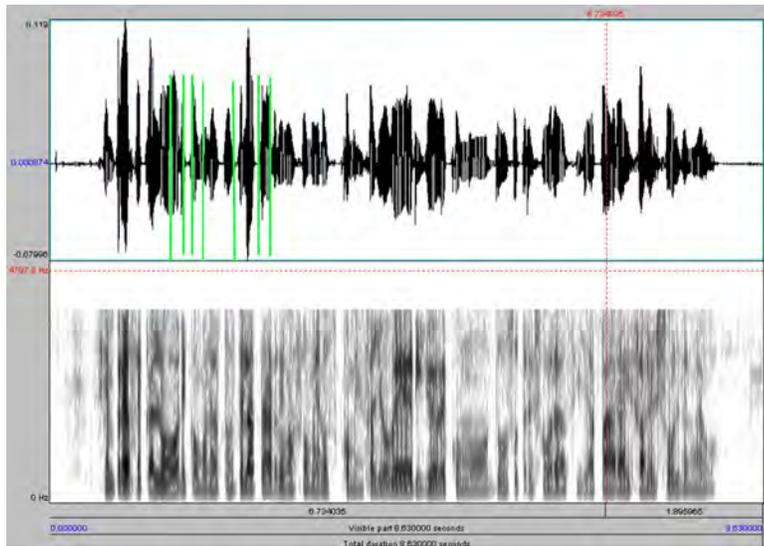
Cuadro 1. Sistemas con representación fonética

En el cuadro 2 se muestra el espectrograma. Las flechas verdes son la segmentación de fonemas, en este caso lo que necesitamos mucho para este sistema es una buena segmentación; si no se segmenta bien la voz, se producen conflictos. Posteriormente, si no tienen el corpus completo de la voz del lenguaje, también hay problemas y obviamente hay que etiquetar todo esto; hay mucho trabajo dedicado a la identificación del lenguaje centrado en fonemas o fonotáctico.

Una de las desventajas es que es altamente dependiente del lenguaje; si ustedes quieren agregar un nuevo lenguaje, tienen que llevar a cabo todo este proceso de recopilación, etiquetado, segmentación de fonemas, modelo del lenguaje. Es un trabajo muy grande el que hay que hacer.

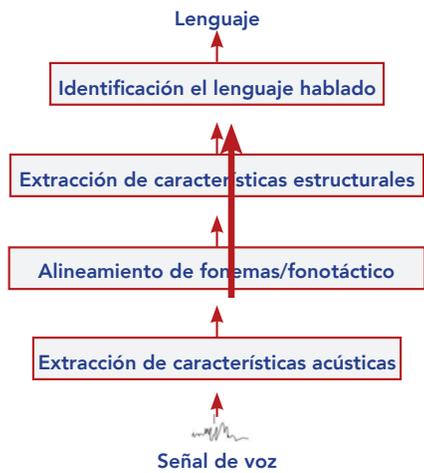
Lo más importante de todo esto es que si queremos aplicarlo para las lenguas que no tienen transcripción fonética, de algunas lenguas indígenas, algunos dialectos africanos, hindúes, etc., no nos va a servir; por eso es que me aboqué a hacer la identificación del lenguaje sin la representación fonética.

El segundo enfoque es hacerlo sin representación fonética. Ustedes se preguntarán: ¿cómo le vamos a hacer? Lo que podemos hacer es explotar directamente la señal acústica para la Interpretación de Lenguaje Hablado (ILH), obteniendo características perceptuales, tales como la señal de voz, la prosodia, el ritmo, la entonación, entre otros.



Cuadro 2. Espectrograma

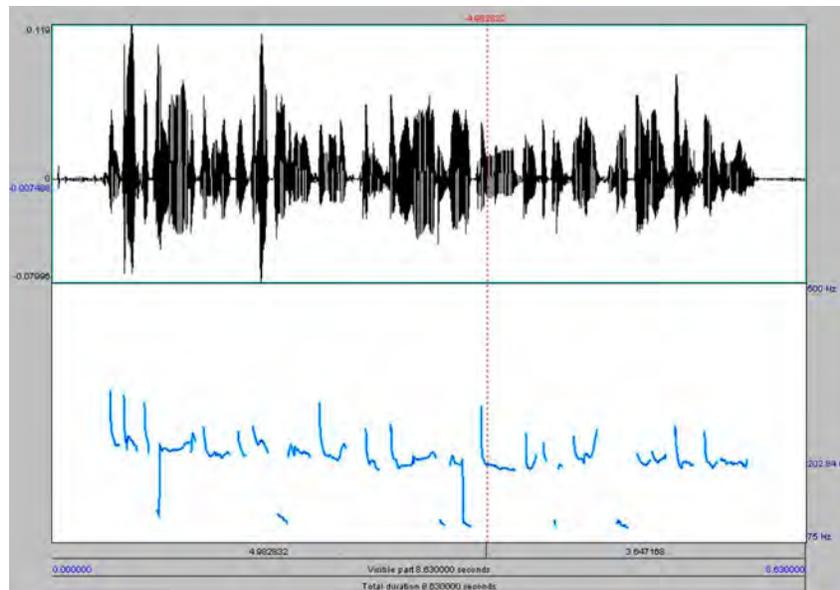
En este caso, lo que propongo es eliminar la representación fonética, y pasar directamente a la extracción de la señal acústica y la identificación del idioma. Ya se han hecho estudios sobre el ritmo, de cómo podemos sacar el ritmo de la señal.



Cuadro 3. Sistemas sin representación fonética

Ya se había hablado de que la frecuencia fundamental y la señal de voz significan la prosodia; hay muchos trabajos sobre esto.

Esto es el estado del arte; en resumen, existen bastantes estudios sobre la parte de lo fonotáctico, es lo que se ha tratado actualmente. Los dos primeros son Casseiro y Torres; han trabajado en ello con buenos resultados en el reconocimiento del habla. Cummins, Samouelian y Rouas son los que han estado trabajando sin la expresión fonotáctica. A grandes rasgos, son los idiomas que ellos reconocen, la señal de voz, los métodos que usaron, la frecuencia fundamental.



Cuadro 4. Frecuencia fundamental (*pitch*)

Por otra parte, Rouas utilizó algo que Ramos desarrolló hace tiempo, lo cual es identificar los idiomas por su ritmo, los clasificaba en tres grandes grupos: el *silabetime*, *stresstime* y el *moratime*. El *silabetime* es como el español y el francés, en donde encontramos un tiempo para las sílabas. El *stresstime* es propio del inglés y el alemán que acentúan la sílaba. Finalmente, el *moratime* que es como una nueva definición que se creó para el japonés. Estos son los tres grandes grupos, no puedo dar más detalles sobre esto.

Lo importante de esto es pasar de la señal acústica a la representación fonética. Para esto, utilicé *coeficientes ceptrales Mel*, que es un término de electrónica, y digitaliza la señal en coeficientes ceptrales donde se divide la señal en varias frecuencias: la frecuencia fundamental, frecuencia primaria y frecuencia secundaria.

Regularmente el estado del arte para lo que es reconocimiento de habla utiliza nada más doce coeficientes centrales, porque a partir del treceavo no aporta más información para el reconocimiento de habla, es decir, para saber qué es lo que me estás diciendo. Yo lo aumenté a dieciséis para tener mayor información sobre las frecuencias secundarias.

Mi método es el siguiente: ustedes tienen la señal de voz, y lo que propongo es segmentación al igual que en el reconocimiento del habla, es decir, en fonemas, en 20 milisegundos, pero si en 20 milisegundos obtengo la identificación del hablante, puedo conseguir información.

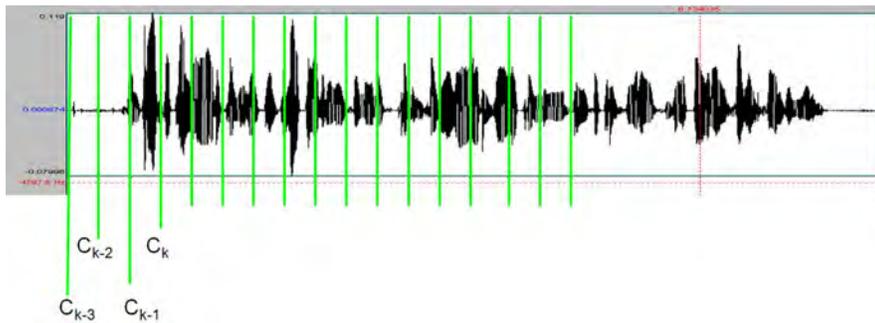
En mi método lo que hago es exactamente sacar los deltas, lo que les he explicado, uno central de cada uno de ellos, sus promedios y sus máximos y mínimos. Con esto obtengo 192 atributos de toda una muestra de señal de voz, porque,

regularmente, para identificar, estamos moviendo la señal de voz, por ejemplo, a diez segundos, a cincuenta segundos, a treinta segundos. Para no mover tanto la señal, convierto cualquier muestra a 192 atributos.

Actualmente, no existe un *multiclase*, es decir, un sistema donde ustedes hablen y estén todos los sistemas grabados y que identifique el idioma. Se va haciendo por pares; lo que nosotros hacemos regularmente es tomar un lenguaje, por ejemplo, náhuatl, junto con el zoque de Oaxaca. También hacemos la *transformada*, posteriormente, la obtención y la de dimensionalidad, después obtengo la identificación del idioma.

Para poder presentarme a nivel internacional, tuve que mostrar el sistema OGI. Es un sistema que tiene 22 lenguajes —de los cuales he utilizado nueve— que son de hablantes distintos, enseguida, tomo cincuenta hablantes diferentes. Cada persona habla lo que se le ocurre en ese momento, no es una grabación determinada y no es regional; de hecho, para el español, tuve personas que hablaban de Colombia, de Venezuela, español de España y español de México.

CALCULO DE PROMEDIOS Y DELTAS



$$\Delta 1C_k = C_k - C_{k-1}$$

$$\Delta 2C_k = C_k - C_{k-2}$$

$$\Delta 3C_k = C_k - C_{k-3}$$

Promedios de $\Delta 1$, $\Delta 2$ y $\Delta 3$

Promedio, máximo y mínimo de C_k

Máximo y Mínimo de $\Delta 1C_k$, $\Delta 2C_k$ y $\Delta 3C_k$

Cuadro 5. Procesamiento usando 16 MFCC

Las pruebas que hice para las diferentes lenguas —para probar si esto funcionaba— fueron del náhuatl y del zoque de Oaxaca, comparándolos con el español, e hice para los diferentes rangos de las muestras de voz (que fueron siete) treinta y cincuenta segundos para el OGI, y para lenguas indígenas tres, siete y diez.

Podemos tener algunas grabaciones, por ejemplo, la de náhuatl donde la persona que las escuchó me dice que hay algún poema, que, por ejemplo, hablaban de algunas de las partes de sus regiones. Entonces, no está predeterminado a un idioma, a algo que está escrito o fijo, a un formato.

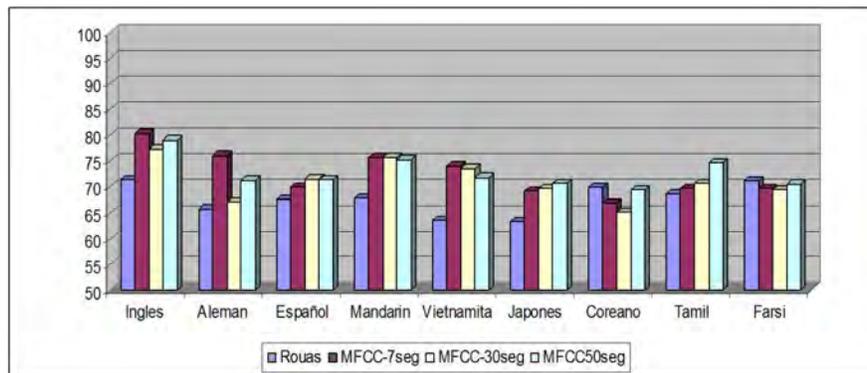
En el cuadro 6 presento mis resultados, me estoy comparando con Rouas, ya que fue de los últimos que obtuvo buenos resultados. Entre paréntesis se pueden apreciar los de él y en negritas están los míos. En este caso, tuve una tasa de reconocimiento más alta que la de Rouas. Cummis empleó la frecuencia fundamental, mientras que Rouas obtuvo la distinción entre unidades vocálicas y consonánticas. Y lo hice por pares de lenguajes.

En este mismo cuadro, muestro por idioma, y con los coeficientes y las señales de voz, los resultados; la primera columna pertenece a Cummis. En la siguiente me estoy comparando con Rouas, que realmente tuvo buenos resultados, pero los estoy superando; mi tasa de identificación del idioma es alta. Todavía no llego a los fonotáticos, pero ésa es la idea, llegar en un momento dado a los fonotáticos y eliminar esta área que es la de identificación del lenguaje hablado sin el conocimiento de la segmentación fonética.

Estos resultados corresponden al náhuatl. Hice la prueba, debido a que, realmente, era muy interesante saber qué se podía hacer para estas lenguas y sí, efectivamente, obtuve una tasa muy buena; entre el náhuatl y el zoque de Oaxaca existe un ritmo que puede distinguirse con facilidad. Por ejemplo, el náhuatl es hablado con un ritmo como /kj, kj, kj/ cosa que no se da en el zoque de Oaxaca. Si tenemos estas diferencias en el ritmo, podemos identificar los idiomas. En fin, estos fueron mis resultados con las lenguas indígenas.

Las conclusiones: lo que ya había mencionado, se presentó mi sistema independiente del proceso fonotático. En los resultados me comparé con Cummis y con Rouas, y algo que se notó es que entre el inglés y el alemán los resultados son muy buenos y eso es lo que se había mencionado, que el inglés y el alemán son *stresstime*, o sea, son muy parecidos en su ritmo. Y agregando esas frecuencias bajas es cuando se obtienen mejores resultados. Se probó que usando varios clasificadores los resultados eran los mismos y, entre más grande era la muestra de la señal de voz, los resultados tienden a decaer y a degradarse.

DELTAS Y PROMEDIOS



Cuadro 6. Avances: conclusiones generales

Lo más importante de todo esto es que se sentaron las bases para la identificación de las lenguas indígenas de México. Claro que para continuar con esto todavía me falta probar con las demás lenguas, encontrar corpus. El corpus no lo pude localizar por ningún lado, me lo tuvo que prestar la ciudad de Texas porque sólo allí contaban con él. Me faltaría tener un corpus. De hecho, es uno de los proyectos que mi asesor quiere hacer: un corpus de lenguas indígenas para poder trabajar esto.

Como trabajo a futuro, se pretende tratar de combinar lo de Rouas con lo mío. He trabajado con estudios de Willes, otro tipo de procesamiento de señal de voz, y realmente obtuve muy buenos resultados. Hasta ahora vamos muy bien con esa parte. Sin embargo, falta un corpus para lenguas indígenas, y poder establecer un sistema. Mi sistema no está bien definido, aún falta un poco de trabajo. Bueno eso es todo.

Alfonso Medina: Podemos dar paso a las preguntas. ¿Alguna pregunta?

SECCIÓN DE PREGUNTAS

Javier Cuétara: Me encantó tu trabajo. Todos nos hemos enterado de indígenas que han sido enjuiciados sin poderse defender, ni tener un traductor; algunos de ellos ni siquiera hablan español. Estamos ahora tratando de acordarnos de dónde hay bancos de datos.

Alfonso Medina: En el Instituto de Investigaciones Antropológicas hay una fonoteca y es un buen recurso; también hay otras fonotecas en el INAH, también existe una radio indígena.

Javier Cuétara: Sí. Ojalá que este trabajo continúe.

Ana Lilia Reyes: Sí, ojalá rinda frutos.

Alfonso Medina: Con esto despedimos esta mesa.

CORPUS PARALELO ALINEADO ESPAÑOL-INGLÉS DE TEXTOS LITERARIOS

GRIGORI SIDOROV/CIC, IPN

Margarita Palacios: Vamos a dar inicio con la ponencia *Corpus paralelo alineado español-inglés de textos literarios*, la cual será presentada por Grigori Sidorov, quien viene del Instituto Politécnico Nacional.

Grigori Sidorov: Estimados colegas, en esta plática vamos a ver los problemas relacionados con los corpus paralelos alineados. Lo voy a presentar utilizando como ejemplo un corpus paralelo español-inglés desarrollado por nosotros. Cabe mencionar que este corpus consiste en textos literarios.

Antes que nada, me gustaría mencionar diferentes tipos de alineación. Normalmente se habla de alineación a nivel de párrafos, nivel de oraciones y nivel de palabras.

Claro está que la diferencia entre esos tipos de alineaciones se basa en las unidades estructurales que se usan. El estado del arte actual permite la alineación a nivel de párrafos y oraciones con un porcentaje bastante alto, mientras que a nivel de las palabras todavía existen bastantes problemas.

Las áreas de aplicaciones de los textos paralelos alineados son innumerables en el campo de la lingüística computacional. Por ejemplo, se usan en compilación automática de los diccionarios, enseñanza de idiomas y extracción de información lingüística de cualquier tipo.

La motivación más directa de este trabajo es contestar la siguiente pregunta: ¿son métodos léxicos de alineación aplicables para textos literarios? Ya que los métodos léxicos se han aplicado a los textos técnicos o jurídicos, donde las traducciones deben ser muy exactas, en el caso de los textos literarios las traducciones pueden ser bastante libres.

Otro enfoque nuestro en la investigación era realizar la comparación de varios algoritmos de alineación usando la similitud léxica.

Por fin, estamos dando cuenta de la necesidad de desarrollo de los corpus alineados para el español.

Como un ejemplo de qué tan libre puede ser una traducción, presentamos dos de éstas, que se realizaron al idioma inglés del mismo texto español. Como se ve en el cuadro 1 las traducciones son bastante distintas aunque transmiten la misma idea general.

TRADUCCIÓN REAL	TRADUCCIÓN LITERAL
<p>I shrugged and smiled and nodded. But I could barely control the emotion that shook me from head to toe. I figured that she must have been upset by the thought that in that blurry face in the photo was written the inevitable, corrupted, degraded future of her own old age. "What do you have to fear from time, little girl, if you have so much of it?" I wondered, as I withdrew to give her space to compose herself and regain her customary serenity.</p>	<p>I made her see that I did not give importance to the situation. But (and I speak about this time) I could barely control the emotion that whipped me from head to toe. "Why should you be afraid of time, if you have a lot of time?" I told her in my inside, thinking that her abashment is due to the abrupt anguish that was caused by the mere idea that, anyway, at the photograph, in that blurry face, there was written something inevitable, the corrupted and degraded future, which my inquisitional eyes (and my theories about physical inheritance) could foresee for her. Something like a shame of a sin not committed yet or a fault assumed without any reason, deep down, the infantile copy of a too immature conscience. "Why are you afraid of time, little girl" I repeated inside, while I was withdrawing to allow her to compose herself and regain her customary serenity.</p>

Cuadro 1. Alineación de dos textos traducidos

Métodos de alineación

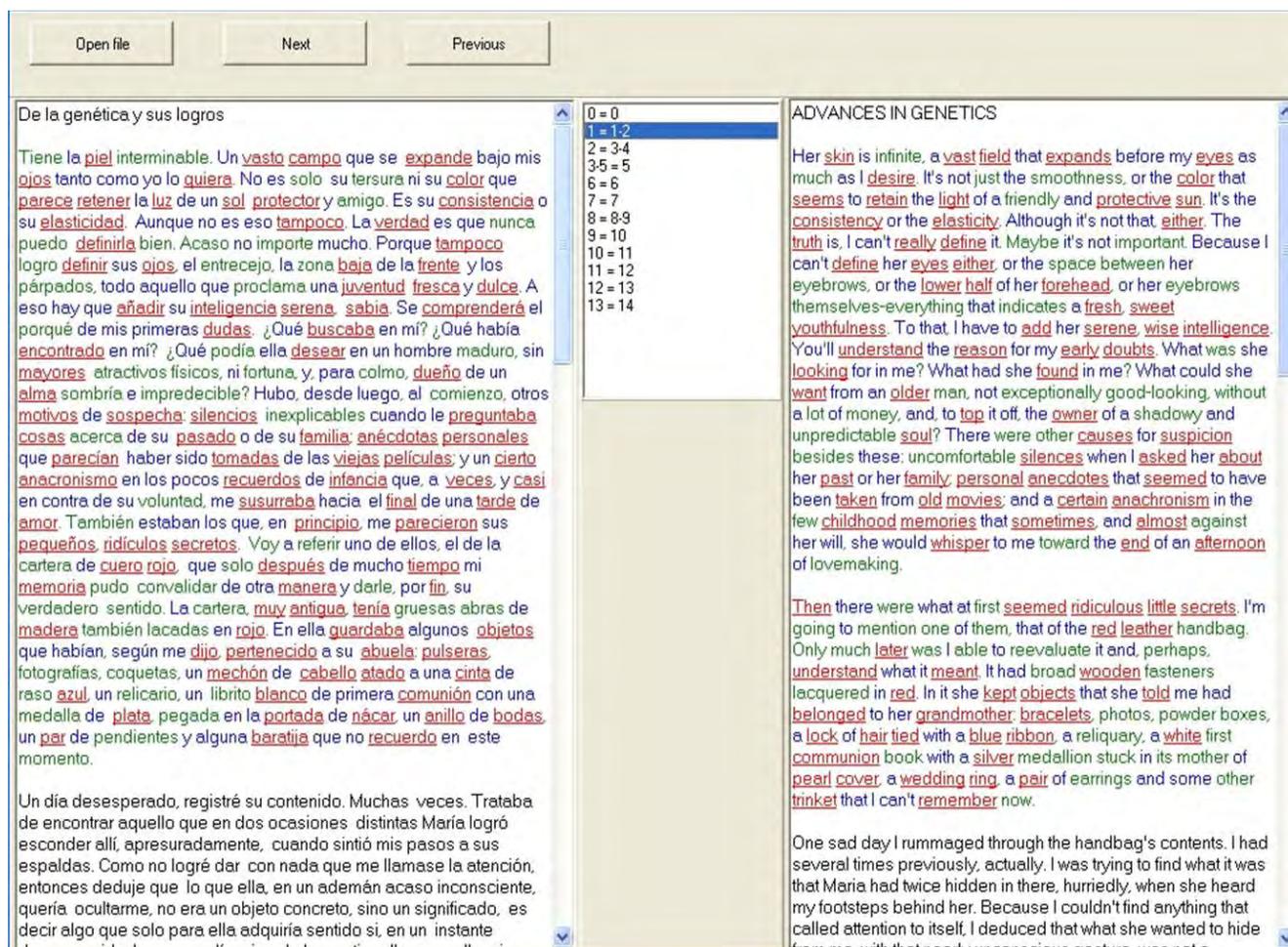
Ahora pasamos a la descripción breve de los métodos de alineación. Hay dos maneras de clasificar los algoritmos correspondientes: dependiendo del algoritmo o dependiendo del método del cálculo de la similitud entre los elementos estructurales.

Según el algoritmo utilizado, los métodos se clasifican en métodos heurísticos y métodos basados en la optimización (por ejemplo, la programación dinámica).

Según el método de cálculo de la similitud, los métodos se clasifican en: léxicos, estadísticos, combinados.

Por ejemplo, el bien conocido algoritmo de Gale y Church es un método estadístico basado en una técnica de la optimización. En nuestros estudios hemos implementado un método heurístico y un método basado en la optimización. En este caso, hemos realizado los experimentos en el nivel de párrafos, aunque obviamente las mismas técnicas son aplicables a nivel de oraciones. Cabe mencionar, que a nivel de palabras es necesario usar otros métodos aunque basados en las mismas ideas.

En el cuadro 2, se presenta el resultado de funcionamiento del sistema de la alineación. Las palabras subrayadas tienen las correspondencias directas en los textos según el diccionario que usamos. Se nota que un párrafo en el español tiene dos párrafos correspondientes en el inglés.



Cuadro 2. Ejemplo de la alineación a nivel de párrafos

Alineación heurística

El método heurístico se basa en unas reglas de detección de correspondencias estructurales. El algoritmo se basa en los patrones: 1-1, 1-2, 1-3, 3-1, 2-1; lo que quiere decir que un párrafo puede tener como correspondencia uno o dos o tres párrafos en otro idioma, y al revés, un párrafo en el otro idioma puede tener como correspondencia uno o dos o tres párrafos en el idioma dado. Es obvio que este algoritmo trabaja solamente con los tres párrafos consecuentes como máximo, determina la similitud entre los patrones mencionados y escoge la correspondencia con la similitud máxima.

Para calcular la similitud se buscan las correspondencias léxicas en los componentes estructurales (párrafos). Se calcula el valor utilizando el coeficiente de Dice con penalización por diferencia en el tamaño.

Otra consideración importante en este método es la utilización de los puntos ancla, es decir, se determinan algunos puntos donde la correspondencia es obvia en los textos, por ejemplo, fechas o nombres propios largos, etc. El algoritmo lo toma en consideración al hacer el proceso de alineación.

Alineación con optimización

Utilizamos como método de alineación la programación dinámica. Se busca la ruta óptima entre todas las posibles rutas de alinear los elementos estructurales, tal como se muestra en el cuadro 3.

		Language B					
		0	1	2	j	...	N_B
Language A	0	0	∞	∞	∞	...	∞
	1	∞	0.1	0.3	0.4	0.6	0.8
	2	∞	0.3	0.5	0.5	0.7	0.7
	3	∞	0.4	0.7	0.7	0.8	0.9
	i	...	0.4	0.6	a		
	N_A	∞					?

Cuadro 3. Algoritmo de la programación dinámica

Existen varias medidas de similitud. Por ejemplo, para el cálculo de la similitud se puede utilizar la función que depende de la posición relativa, específicamente:

Distancia (TA , TB)

$$= |\text{start}(TA) - \text{start}(TB)| + |\text{end}(TA) - \text{end}(TB)|$$

Donde $\text{start}(TA)$ es la posición relativa (medida en porcentaje) de la primera palabra en TA . Aquí TA y TB son los elementos estructurales de los textos en los lenguajes correspondientes.

En el caso del algoritmo de Gale y Church se miden las correspondencias en longitudes de los párrafos.

En nuestro caso la similitud depende del número de elementos léxicos que se interceptan, es decir, la palabra existe en un texto y su traducción existe en otro. Más precisamente; calculamos las palabras que quedaron sin la traducción, porque de esta manera tomamos en cuenta la longitud del texto; es decir, si algunas palabras se tradujeron, pero muchas otras no; la similitud no debe ser muy alta.

Evaluación preliminar del método heurístico

Hemos evaluado ambos métodos. Para la evaluación del método heurístico se utilizó un fragmento de *Drácula*. Se vieron 50 patrones, se evaluaron los resultados de manera manual, la precisión obtenida fue igual a 94%, siendo los resultados como sigue:

PATRÓN	TOTAL OCURRENCIAS	ERRORES
1-1	27	0
1-2	8	2
1-3	7	0
3-1	6	0
2-1	2	1

Cuadro 4. Resultados

Sin embargo, hay que mencionar que los datos fueron bastante limpios en este ejemplo. En una aplicación masiva de este algoritmo, al cometer un error, éste se pierde si no encuentra rápidamente un ancla. Las anclas no aparecen en los textos muy frecuentemente. Entonces es recomendable o bien tener muchas anclas, o basarse en el algoritmo con la optimización.

Evaluación de alineación con optimización

Aplicamos el algoritmo basado en la optimización desarrollado al texto *De la genética y sus logros*. Hemos evaluado los resultados en términos de la precisión y la especificidad presentados en la tabla siguiente. La correspondencia múltiple se presenta cuando varios elementos pueden tener como correspondencia elementos en el otro texto.

MEDIDA	Correspondencia múltiple		Correspondencia simple	
	precisión, %	especificidad, %	precisión, %	especificidad, %
PROPUESTA	89	85	88	90
BASE	65	28	43	54

Cuadro 5. Correspondencias múltiples y simples

AUTHOR	ENGLISH TITLE	PAR.	SPANISH TITLE	PAR.
Carroll, Lewis	Alice's adventures in wonderland	905	Alicia en el país de las maravillas	1,148
Carroll, Lewis	Through the looking glass	1,190	Alicia a través del espejo	1,230
Conan Doyle, Arthur	The adventures of Sherlock Holmes	2,260	Las aventuras de Sherlock Holmes	2,550
James, Henry	The turn of the screw	820	Otra vuelta de tuerca	1,141
Kipling, Rudyard	The jungle book	1,219	El libro de la selva	1,428
Shelley, Mary	Frankenstein	787	Frankenstein	835
Stoker, Bram	Dracula	2,276	Drácula	2,430
Ubídia, Abdón	Advances in genetics ²	116	De la genética y sus logros	109
Verne, Jules	Five weeks in a ballon	2,068	Cinco semanas en globo	2,860
Verne, Jules	From the earth to the moon	894	De la tierra a la luna	1,235
Verne, Jules	Michael Strogoff	2,464	Miguel Strogoff	3,059
Verne, Jules	Twenty thousand leagues under the sea ³	3,702	Veinte mil leguas de viaje submarino	3,515

Cuadro 6. Corpus

Corpus desarrollado

Aparte de probar los algoritmos de la alineación hemos desarrollado un corpus paralelo español-inglés de los textos literarios. Los títulos incluidos en el corpus se presentan en el cuadro 6.

Las columnas “Par” contienen el número de párrafos correspondientes.

El tamaño del corpus es alrededor de 11,5 MB. Los parámetros del corpus son los siguientes:

	INGLÉS	ESPAÑOL
Palabras total	848,040	844,156
Palabras únicas	25,877	43,176
Párrafos	18,701	21,540

Cuadro 7. Parámetros para la construcción del corpus

Conclusiones

Concluyendo nuestro trabajo podemos mencionar lo siguiente:

- Hemos presentado varios algoritmos de alineación.
- Discutimos sobre varios métodos de cálculo de similitud.
- Vimos la comparación de los resultados para los textos literarios que son “difíciles” para los métodos de la alineación.
- Presentamos un corpus de los textos literarios y sus parámetros.

Trabajo futuro

Como trabajo futuro vale la pena mencionar los siguientes puntos:

- Debe hacerse el análisis de los errores en casos de los diferentes algoritmos.
- El método del cálculo de la similitud debe ser más exacto, por ejemplo, hay que evitar calcular traducciones varias veces.
- Es necesario realizar el cálculo de similitud tomando en cuenta el orden de palabras en los párrafos.
- Es deseable utilizar esquemas con asignación de pesos tipo TF-IDF en lugar de remover palabras frecuentes.
- Se planea la aplicación de los métodos mencionados a alineación de las oraciones y palabras.
- Por fin, es necesario intentar utilizar los resultados de alineación para el enriquecimiento de los diccionarios bilingües.

CONTROL DE LA ESTABILIDAD EN EL AGRUPAMIENTO DE TEXTOS

HÉCTOR JIMÉNEZ SALAZAR/FCBIT, UAM
DAVID PINTO AVENDAÑO DSIC/UPV

Héctor Jiménez: El problema de agrupamiento de textos es muy importante por el tipo de las aplicaciones que actualmente tienen demanda. Específicamente, se trata de formar grupos de textos a partir de una colección de textos dada, sin ninguna información adicional; por ejemplo, información como: vocabularios, otros grupos de textos previamente formados, diccionarios, etc. Hay muchos enfoques para realizar agrupamiento. El nuestro fue, primeramente, representar cada texto por un conjunto de términos que en él aparece, posteriormente, se aplica un método de agrupamiento. Es importante decir que, con el propósito de evaluación de los grupos obtenidos, tomamos una colección previamente agrupada en forma manual. Así, sin considerar el grupo al que pertenece cada texto, se agrupa y después se evalúan los grupos comparando con el agrupamiento manual.

Concretamente, nuestro interés está en saber con cuáles términos se representan los textos para obtener la mejor evaluación de los grupos obtenidos. La estrategia que seguimos fue ordenar los términos del vocabulario de la colección de textos según un criterio de importancia y tomar un porcentaje de los términos más importantes, con estos se tendrá cada texto representado y procedemos a realizar el agrupamiento.

No es sencillo elegir un buen porcentaje, puesto que, con un porcentaje bajo de términos, pueden quedar fuera términos que permiten establecer similitudes entre los documentos para ser reconocidos como parte de un grupo. También, un porcentaje alto de términos normalmente incluirá términos que introducen ruido en la representación y, por tanto, grupos mal formados.

Estos métodos son métodos no supervisados. Al estar eligiendo los términos que van a presentar a cada texto, nosotros esperamos que, conforme vayamos añadiendo términos, el agrupamiento sea mejor o se quede ya en un valor de precisión, ya no va a aumentar; sin embargo, después del 10% o 20% de los términos que conforman los textos ya no obtenemos una mejor clasificación; empero, del primer agrupamiento podemos ir obteniendo un buen conjunto y después al aumentar términos, el agrupamiento empieza a causar problemas, empieza descomponer ese agrupamiento.

Lo que tratamos de hacer es controlar este hecho; que al aumentar términos no se descomponga el agrupamiento que ya habíamos hecho antes; ése es el problema que hemos detectado y que es importante resolver. Panorámicamente se pretende hablar de algunos métodos de selección de términos, que es la base para representar los textos, y, después de hacer el agrupamiento, vamos también a ver cuál va ser la propuesta que hacemos para controlar la inestabilidad, y después vamos a hablar del experimento que nos da una idea de qué tan buena es la propuesta y finalmente las conclusiones.

Voy a mencionar de manera muy panorámica cada uno de los métodos. Este primer método de selección de términos para representar los textos lo que hace es asignar una puntuación a cada término y esa puntuación se obtiene a partir del número de documentos que hacen uso del término; entre más documentos hagan uso de un término, mayor va ser la puntuación del término. Una vez que tenemos todos los términos con su puntuación, se ordenan en forma descendente, esto es, tenemos primero los de mayor puntuación hasta los de menor puntuación y se toman porcentajes, por ejemplo, el 10% de estos términos y con éstos se representan todos los textos; una vez que ya están representados se agrupan y se mide qué tan bien está hecha la agrupación. Este es un método muy sencillo, bastante estable y es efectivo.

Otro de los métodos que hemos usado es que podemos traducir como fuerza de enlace; aquí se pretende que un término pueda estar participando o influyendo en que dos documentos sean similares; de tal manera que si un término tiene esta propiedad, va a asumir una importancia mayor que otros que no contribuyan de alguna manera a enlazar varios textos; podríamos decir que el anterior es inestable y, desafortunadamente, es costoso computacionalmente, ya que hay que calcular muchas similitudes para poder definir lo que requiere la fórmula.

Pasamos a este otro método que es llamado *Punto de transición*. Este método de selección de términos se basa en la idea de que los términos de frecuencia media tienen un alto contenido semántico; entonces, hay que localizar las frecuencias medias en los términos de un texto y con esos términos que tienen frecuencia media se representan los textos. El método es bastante sencillo y altamente efectivo, pero es también muy inestable; justamente eso es lo que nos ha motivado a hacer este análisis que ahora estoy presentando, hablo de estos detalles para poder encontrar las frecuencias medias. Con esto terminaríamos los métodos de selección.

Referiremos ahora la idea que hemos formulado. Vamos a ver los resultados y comentaremos más sobre qué otras ideas tenemos para controlar la inestabilidad de los métodos. Nuestra idea es poder establecer un identificador que nos diga qué tan juntos están los textos, quiero decir, que se parecen mucho entre sí. Esta idea nos lleva al concepto de *densidad* que se refiere a procesamiento de textos y está basado en lo que se llama el *centroide*.

El centroide, como sabemos, es un concepto biométrico que se refiere a un punto central o representativo de un conjunto de puntos y este centroide puede calcularse mediante el promedio de diferentes cifras que son numéricas; pero cuando estamos trabajando con términos, el concepto de centroide cambia y, finalmente lo llevamos a números, porque lo que hacemos es utilizar las frecuencias de los términos. Con la idea de obtener un texto centroide de una colección, elegimos los términos más representativos y estos son los términos de frecuencia media. Si ya tenemos el centroide, podemos hablar de la densidad, la densidad de una colección, porque algo que no había mencionado antes es que los métodos de selección nos van a ofrecer diferentes maneras de representar los textos: como ya lo dije, por ejemplo, yo tengo ordenados mis términos de acuerdo con una puntuación y el 10%; con ese 10% represento todos los textos, pero también puedo elegir el 20% y representar todos los textos. En otras palabras, por cada porcentaje de términos que elija que me da un método de selección, tendré una colección diferente, o, dicho de una manera más adecuada, una representación diferente de la colección; vamos a verlas como colecciones a todas estas.

La densidad se refiere a una colección. Es decir, si yo tengo una manera de representar los textos mediante un método y un porcentaje, entonces, definimos la densidad de esa colección como la suma de las similitudes del centroide con todos los textos. La similitud que estamos manejando aquí es una muy conocida, se trata del *índice de Jaccard* o, que algunos llaman, *distancia semántica*. Hemos visto que funciona más o menos bien para este tipo de problemas.

Esbozamos el planteamiento que hacemos de nuestro trabajo. Nuestra idea es que si la densidad es alta, entonces esto nos estará diciendo que habrá mayor parecido entre textos y, a su vez, esto quiere decir que es más fácil formar cada grupo. Eso es un enfoque. Más adelante, voy mencionar las desventajas, y eso nos daría un alto desempeño del agrupamiento. La densidad me puede indicar si la selección que he hecho es una buena selección, es una buena forma de representar los textos. Simplemente nos olvidamos de la

inestabilidad de los métodos, solamente calculamos la densidad y la densidad me indica cuando es una buena selección, buena representación; si es buena representación, hago el agrupamiento. Ésta es la hipótesis en sí.

Referiremos qué está pasando con cada una de las formas en que representamos los textos de la colección al ir tomando diferentes porcentajes. Naturalmente el 20% de los términos incluye al 10%; si el método está funcionando bien, puede tener términos que permitan mejorar el agrupamiento y, cuando el agrupamiento deja de ser bueno, podemos ver un descenso en la densidad; así que la manera de obtener una buena selección es buscando un máximo local. Pasemos entonces a ver qué sucedió con esta hipótesis que acabo de mencionar.

El cuadro 1 es una colección de resúmenes de textos científicos, más específicamente, de física de altas energías, son casi 3000 textos y una cantidad promedio de 46 términos por texto. Es una colección un poco difícil, aquí hay nueve clases, estas nueve clases están muy desbalanceadas; hay una clase que tiene muchos, hay otras que tienen un número mediano y hay otras que tienen muy pocos. Por un lado, esto lo hace difícil, pero también no es muy representativa para los problemas de agrupamiento. Lo que se hizo en el experimento fue justamente preprocesar cada uno de los textos usando lexemas, quedándonos con las palabras cerradas: solamente sustantivos, adjetivos, verbos, etcétera.

SE HA EMPLEADO LA COLECCIÓN <i>HEP-EX</i> . UNA COLECCIÓN QUE REÚNE RESÚMENES DE ARTÍCULOS CIENTÍFICOS DE UN DOMINIO ESPECÍFICO ^a . SUS CARACTERÍSTICAS SON:	
Tamaño en bytes	962,802
Número de clases	9
Número de textos	2,922
Número total de términos	135,969
Tamaño de vocabulario	6,150
Promedio de términos por texto	46.53

^aHigh Energy Physics, compilada por CERN.

Cuadro 1. Colección de prueba

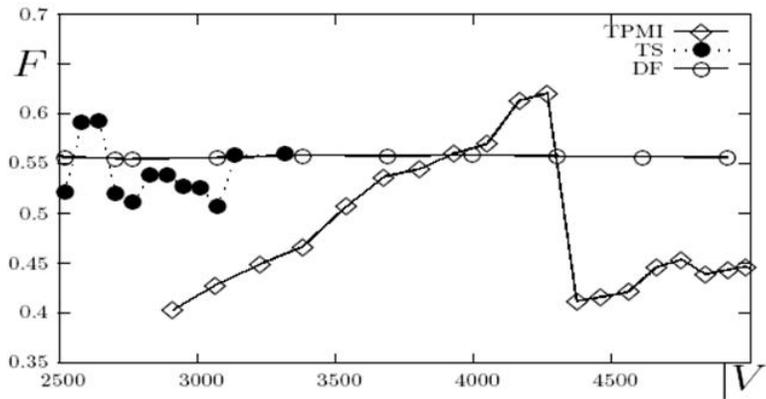
Aplicamos los métodos de selección para varios porcentajes. De igual forma, hicimos el agrupamiento con un algoritmo del vecino más cercano; una variante del vecino más próximo y, finalmente, medimos la calidad de este agrupamiento con estas medidas estándares que son la precisión y la evocación, reuniéndolas en el índice F1 que nos dice qué tan buena es.

El cuadro 2 presenta los resultados. Vemos que los puntos negros son el medio de fuerza de enlace. Como se puede observar, es muy inestable; en el eje horizontal tenemos el tamaño del vocabulario y estamos dando varios porcentajes de términos seleccionados y tenemos que, al aumentar términos con el método de fuerza de enlace, baja en un momento dado y luego vuelve a subir. Está bajando, ésa es la inestabilidad al aumentar términos con el método de fuerza de enlace: baja en un momento dado y luego vuelve a subir. Está bajando, ésa es la inestabilidad de fuerza de enlace. Por otro lado, el método de frecuencia interdocumento es muy estable: son los puntos huecos, que están arriba de .55, su valor de la calidad del agrupamiento. Notamos también que el método basado en el punto de transición, que aparentemente es estable, llega un momento

en que da un giro bastante radical: baja y después vuelve a subir. Ahí está también la inestabilidad de ese método que es el de punto de transición.

Resultados

Índice F para 3 métodos de selección.



Cuadro 2. Resultados

CORRESPONDENCIA ENTRE EL DESEMPEÑO (F) Y LA DENSIDAD (ρ)						
%	17	18	19	20	21	22
F	0.4756	0.5824	0.5988	0.6171	0.4861	0.5074
ρ	8.81	9.16	9.40	9.77	9.72	9.95

Cuadro 3. Densidad contra F

Elegí solamente un franja (véase cuadro 3). El comportamiento nos dice que hay un máximo local justamente en el máximo F; el valor del desempeño del agrupamiento fue para un porcentaje de veinte, punto de transición, que nos da un índice de .61, y es donde se presenta un máximo local; como pueden ver, el 20% corresponde a una densidad de 9.7 y al lado izquierdo y derecho de 9.77 están los valores menores. Eso quiere decir que es un máximo local, o sea, me está diciendo exactamente de dónde se obtiene en mejor agrupamiento.

Las conclusiones: aprovechemos para mencionar que es solamente un experimento y hay muchas cosas que hacer; puesto que nuestra hipótesis la estamos viendo cómo la densidad, nos dice que los textos están muy cercanos, lo que permite agrupar los que son muy similares; pero ¿qué pasa cuando hay textos muy cercanos que no son muy similares? Debe haber un problema. Se supone que los ordena para que no suceda esto, pero esa idea tiene que ser comprobada con otros métodos y también explorar qué sucede

conjuntamente con estos otros términos que al agregar en lugar de darnos una similitud buena nos dan una similitud mala, es decir, que se agrupen textos de diferentes clases en un mismo grupo, eso es todavía una cuestión que debemos continuar y seguir trabajando. Eso es todo, muchas gracias.

Margarita Palacios: Agradecemos mucho la participación del Dr. Jiménez. Considero que su trabajo ya es más que una propuesta, que ya es un proyecto y que desde mi punto de vista y aplicado al terreno específico de la lingüística, hay grandes posibilidades, no solamente lo que estás llamando agrupamiento sino justamente en la dispersión. Me parece muy interesante el proyecto y seguramente habrá muchas preguntas dentro del público.

SECCIÓN DE PREGUNTAS

Gerardo Sierra: Tú trabajaste sobre resúmenes, y los resúmenes normalmente tienen la misma extensión, es de esperarse que estén formados por dos o tres párrafos a lo mucho, ¿qué pasa cuando se tiene una colección de distintos tamaños —pensemos correos electrónicos, noticias de periódicos que pueden ser o muy cortas o muy grandes— qué tan fácil sería hacerlo?

Héctor Jiménez: Como dije, los resúmenes son bastantes cortos y, aún así, no contribuyen a elevar la variedad de la longitud; esta cantidad de términos por documento no suele ser la que más influye en el agrupamiento, sino más bien la frecuencia de los términos, por ejemplo, si los textos son extremadamente cortos —los ejemplo más interesante serían contextos de una palabra polisémica—. Éstos pueden ser una oración o varias oraciones, a lo mejor una antes y otra después, éstos ya nos pueden llevar a textos extremadamente cortos y todos son muy cortos. El problema ahí es que carece de la frecuencia de los términos porque no podemos decir cuáles son importantes y cuáles no lo son. Más que a la variedad entre los diferentes textos es más bien en cuanto a que sea un tamaño reducido de todos. Si tengo un máximo de treinta palabras, me enfrento a un problema más amplio.

Pero, así como éste, hay otros muchos factores, como el equilibrio entre las clases, hay clases que tienen muchos documentos y otras que tienen pocos, esto se ve positivo por problemas normales, no podemos decir que cada uno que tenga cien textos.

Pregunta 1: ¿Cuántas clases obtuviste? ¿Son clases que tú definiste?

Héctor Jiménez: Sí. Este método de vecinos más cercanos define los grupos y, obviamente, uno no le dice cuántos; el mismo método nos dice cuántos van a ser; normalmente da más de los que son; por ejemplo, aquí obteníamos dieciséis grupos, a veces más. Cuando se degradaba el agrupamiento teníamos veinte o más, es decir, un indicador de un buen agrupamiento es que sean pocos grupos. En otro experimento hicimos participar el número de grupos para medir la calidad de la selección, lo que funcionó pertinentemente.

Grigori Sidorov: ¿Se puede elegir qué tan fina va a ser la clasificación? Es decir, tú hablas de noticias, yo hablo de deportes, de política y, digamos, de política externa, interna, ¿depende de qué nivel estamos hablando?

Héctor Jiménez: Sí, hay métodos que, por ejemplo, dado un grupo, puede volver a reagruparse en diferentes subgrupos. No estamos tratando con ese problema de hacer un endograma, una jerarquía que nos diga: toda la colección se divide en tres grandes grupos y a su vez en otros subgrupos hasta hacer algo muy fino. Por ahora, solamente el método es de grupos no jerárquicos, simplemente de grupos generales que componen la colección.

Margarita Palacios: Daremos inicio a lo que será la última presentación de este coloquio.

ALINEAMIENTO DE CORPUS PARALELOS NÁHUATL-ESPAÑOL CON ENFOQUE DE EXTRACCIÓN LÉXICA

SERGIO PAEZ/IIMAS, UNAM
GABRIELA BAYONA/CELE, UNAM

Gabriela Bayona: Buenas tardes, nuestro trabajo es la investigación que estamos realizando como tesis de maestría con el doctor Gerardo Sierra Martínez, que es nuestro asesor. El título del trabajo es *Alineamiento de corpus paralelos náhuatl-español con enfoque en extracción léxica*.

A manera de introducción, como han visto en el coloquio, la unión entre ingeniería y lingüística ha dado muchos frutos de diversas clases en los últimos veinte años, sobre todo, desde el campo de la ingeniería en computación. Lograr automatizar procesos de terminología y léxico es una necesidad de nuestra época, dada la inmensa cantidad de información que tenemos y la urgencia de su disponibilidad que obviamente sobrepasa la capacidad y velocidad humana, es decir, necesitamos computadoras para hacer esto. Sin embargo, la mayoría de las herramientas lingüísticas trabaja con lenguas dominantes como francés, alemán, inglés; son muy pocos los trabajos para lenguas indígenas, como el náhuatl.

Desde la lingüística, aunque existen diversos diccionarios náhuatl-español, ninguno de estos trabajos parte de un corpus paralelo informatizado. Este trabajo consistió en reunir un corpus paralelo informatizado náhuatl-español para su procesamiento; se crearon diversos programas que permitieron el alineamiento y la extracción del léxico.

Teníamos la intención original de trabajar con una lengua indígena mexicana viva, pero nos topamos con dos problemas: el primero es que las herramientas de automatización requieren un corpus alfabetizado de textos en formato digital, cosa que muy pocas o ninguna de las lenguas indígenas mexicanas vivas posee; y el segundo problema es de índole político, y es que las lenguas mexicanas indígenas vivas todavía están en proceso de estandarización y alfabetización; aunque existen diversas propuestas, surgidas desde fuera de las comunidades de hablantes, surgidas de tipólogos, de evangelizadores, de autoridades educativas, ninguna cuenta con el respaldo masivo de las mismas, pues cada lengua viva está representada por una diversidad de dialectos y siempre que se utiliza como norma uno de ellos, se hacen patentes los conflictos políticos de las propias etnias y de su relación con el suelo mexicano. Por estas razones decidimos utilizar el náhuatl clásico para nuestro trabajo.

Nuestro corpus consta de 188,611 palabras en náhuatl clásico y 210,587 palabras en español actual de México. Se denomina náhuatl clásico a la lengua hablada por los mexicas, aztecas de México, Tenochtitlán, del centro de la ciudad de México, en la época del imperio azteca durante la conquista y los primeros años de la colonia del imperio español. Es importante señalar esto porque el náhuatl clásico convivió con otras variantes del náhuatl; actualmente, el náhuatl tiene diferentes dialectos, en otras palabras, el náhuatl clásico está muerto.

El náhuatl clásico fue escrito antes de la llegada de los europeos con un sistema parcialmente ideográfico; estos códices y documentos que contienen ejemplos del sistema de escritura han sido reservados. Con la llegada de los españoles y el fin del imperio azteca, los frailes españoles elaboraron un alfabeto basado en las letras latino-españolas; esa ortografía sirvió para componer una diversidad de manuscritos en pergaminos y folios que se conservan en los archivos, museos, bibliotecas de distintas partes del mundo, en particular en Europa, Estados Unidos y México.

Nuestro corpus está compuesto por cinco textos. El primero de ellos es *Anales de Tlatelolco*, que fue escrito hacia 1620; se conserva en dos manuscritos en el Fondo Mexicano de la Biblioteca Nacional de Francia, donde está catalogado con el número 22 y 22bis. El segundo texto es *La leyenda de los soles*, conocido también como Códice de Chimalpopoca; fue escrito entre 1558 y 1561, por fechas del propio manuscrito, y es anónimo, aunque se le ha atribuido al mexica tlatelolca Martín Jovita; el manuscrito original fue extraviado por el director, entonces secretario del Instituto Nacional de Antropología e Historia, Salvador Toscano, pero se conserva una edición facsímil en el Archivo Histórico de la Biblioteca Nacional de Antropología e Historia. El tercer texto es el Diario de Domingo Chimalpaí, que fue escrito entre 1589 y 1615 por Domingo Francisco de San Antonio Muños Chimalpaí Cuauhtehuanintzin, descendiente de nobles tlatelolcas; se conserva como el manuscrito mexicano 220 de la Biblioteca Nacional de Francia, aunque en la Biblioteca Nacional de Antropología e Historia está un comienzo del Diario. También de Chimalpaí tenemos dos textos más, que son: *Las ocho relaciones* y el *Mural de Culhuacán*, escritos entre 1607 y 1637; se conservan como el archivo 74 de la Biblioteca Nacional de Francia y los folios 1ARA16V del volumen 256-B de la colección antigua del Archivo Histórico de la Biblioteca Nacional de Antropología e Historia.

La paleografía y la traducción de los textos fueron realizadas por el maestro Rafael Itena del Instituto Nacional de Antropología e Historia, quien amablemente también nos dio los archivos electrónicos y digitales para hacer el corpus. Cabe aclarar que ni Sergio Páez ni yo hablamos náhuatl.

Sergio Páez: Los objetivos en este proyecto fueron básicamente tres: construcción del corpus informático náhuatl-español, alineamiento de oraciones y la correspondencia de palabras.

La parte de la construcción del corpus básicamente se realizó en tres pasos. El primero fue la preparación de los textos para ser etiquetados, ya que los documentos tenían notas al pie de página y también algunas marcas del traductor. Una vez con los textos limpios, con únicamente lenguaje náhuatl, etiquetamos con lenguaje XML (Extensible Markup Language) para marcar párrafos, oraciones, títulos, subtítulos. Vamos a ver más adelante las etiquetas que hicimos para hacer el marcado y ya teniendo los textos etiquetados proseguimos a su alineamiento.

La preparación de los textos para ser marcados consistió en eliminar las notas al pie de página, eliminar las anotaciones del autor. Ese trabajo se hizo con la herramienta *Microsoft Word*. Este etiquetado en XML se realizó con las opciones de reemplazo de *Microsoft Word*; de hecho, nuestras marcas para analizar oración es el punto y seguido, y el punto final para párrafos. Los títulos y subtítulos en general se pueden identificar fácilmente, aunque tuvimos que agregar subtítulos porque los textos no contaban con ellos.

Una vez que tenemos los textos etiquetados, únicamente con las marcas de documento, título, subtítulo, párrafo y oración, se pasaron a formato XML. Se tuvo que hacer un proceso: primero pasarlos de XTL (Externalization Template Library) en texto plano, posteriormente, abrirlo en .txt y grabarlos con formato UTF8 con terminación XML para que un buscador fuera capaz de reconocerlos. Cuando el documento en un buscador no se puede abrir quiere decir que una etiqueta está mal: hay que regresarse al texto original, ver qué parte del etiquetado es la que tiene error. Se debe corregir ese error en el texto original, volver a hacer todo el procedimiento original: pasarlo a .txt, luego salvarlo en UTF8 e intentar abrirlo con el *browser*. Finalmente, cuando lo abre completamente, se trata de un texto bien etiquetado.

Las etiquetas que escogimos son las siguientes:

- <doc> para documento.
- <t> para título.
- <st> para subtítulo.
- <p> para párrafo.
- <o> para oración.

Cabe mencionar que el objetivo de XML es que tengamos mayor atención sobre el contenido del texto y no sobre el formato que tiene éste.

El cuadro 1 muestra un ejemplo de etiquetado. La etiqueta “documento” cuenta con “número”, y contiene “título”, “subtítulo”, “párrafo” y “oración”. En este caso ya están numerados tanto los párrafos y las oraciones en un texto que ya se procesó, además tuvimos que hacer un procedimiento de numerado de párrafos y oraciones. La parte del numerado es, hasta cierto punto, el trabajo de alineado.

```
<doc num="1">
  <t1>ANALES DE TLATELOLCO</t1>
  <st>Aquí se referirá cómo vinieron, y cómo comenzó el tlatocáyotl.</st>
  <t2>I. LOS GOBERNANTES DE TLATELOLCO</t2>
- <p num="1">
  <o num="1">Cuando llegaron a Chapoltépec, los mexicas no venían
  separados, todavía estaban juntos.</o>
  <o num="2">Aquí se nombran todos los principales: Poyáhuítl, Memella,
  Xolman, Michiníztac, Cemacachiquíhuítl, Aátlatl, Xomímitl, Atlacuáhuítl,
  Ténoch, Ocelopan, Acacitli, Océlotl, Chalchiuhtlatónac, Ayocuan, Xocóyol y
  Tláquetz.</o>
  <o num="3">Y su caudillo, llamado Tozcuécuex, todavía los gobernó durante
  20 años en Chapoltépec.</o>
</p>
- <p num="2">
  <o num="1">Al morir Tozcuécuex se asentó Huitzililhuitzin; y éste llevaba 23
  años gobernando cuando los mexicas fueron despojados en el año 1 Tochtli
  1298, aunque algunos lograron salvarse.</o>
  <o num="2">Luego se metieron a vivir entre los de Colhuacan, cuando
  Eztlocelopan fue a suplicarles; y quienes lo enviaron fueron Xomímitl,
  Michiníztac, Ténoch e Iztacchiauhtótotl.</o>
  <o num="3">Mientras anduvieron por Nextícpac, por Tecuictollan, por
```

Cuadro 1. Ejemplo de etiquetado en un *browser*

El alineamiento es un proceso complejo: desarrollamos algunos programas. El primer programa numera los párrafos, el algoritmo de Gale y Chuch que usamos es el alineamiento de oración; toma como base que los textos deben de tener el mismo número de párrafos. Para nosotros era importante saber cuántos párrafos contenía cada texto, por eso se desarrolló este programa. Una vez que teníamos nuestro texto etiquetado, corríamos el programa que se encargaba de enumerar cada párrafo. En este punto nos ayudaron mucho las etiquetas porque las de inicio de párrafo nos indican que viene uno nuevo; insertamos la cadena <doc num. + (El número)>, la etiqueta originalmente ya tenía su cierre, mayor que; entonces se inserta esa cadena con numeración conforme se va encontrando la etiqueta. De esa forma podemos irnos al fin del documento y ver el número de párrafos que tiene. Es el mismo proceso lo que hacemos con el documento de otro lenguaje. En este caso, primero en español, luego en náhuatl y verificamos que el número de párrafos sea el mismo número. Si el número de párrafos era el mismo,

no había problema. La probabilidad para que los textos estuvieran alineados a nivel de párrafos era alta; de hecho, éstos ya se encuentran alineados; no tuvimos problemas en eso.

Posteriormente, utilizamos un *Wewor*: un programa que hace algo parecido al anterior, pero en este caso a nivel de oraciones. Tuvimos que tomar una decisión: si numerar todas las oraciones de corrido o numerar las oraciones desde el inicio hasta cada párrafo; finalmente decidimos inicializar la cuenta en cada párrafo. Esto nos iba a ayudar a tener una dirección para cada oración; de hecho, los documentos nos están sirviendo como base de datos, es decir, no necesitamos almacenar información, sino que el mismo documento nos sirve como tal, y si queremos localizar una oración, lo podemos hacer fácilmente a través de las etiquetas y las oraciones de cada una. Su dirección sería número de párrafos y número de oración y de esa forma yo puedo tener dirección para cada oración.

Se enumeran las oraciones en los dos textos; tenemos, empero, el siguiente problema: cuando un párrafo tiene el mismo número de oraciones, podemos esperar que esté alineado. Si dos párrafos tienen diferente número de oraciones, quiere decir que hay un problema, por lo tanto, desarrollamos un programa que encuentre esa diferencia, para no estar revisando todo el documento, viendo cada párrafo y encontrar dónde hay diferencias, porque hay documentos que tienen setenta y cuatro párrafos, en los cuales realizar este tipo de trabajo tomaría demasiado tiempo. Desarrollamos ese programa que va contabilizando si los párrafos tienen el mismo número de oraciones. Si encuentra dos párrafos que tienen diferente número, en ese momento se detiene indicándonos dónde hay un problema; de allí, nos vamos a los textos originales y revisamos qué está pasando.

El formato electrónico del texto a veces tenía unos errores; por ejemplo, le faltaba un fin de párrafo en alguna parte o podía ser un error tipográfico al momento de pasarlo a formato electrónico; otro error puede ser, en efecto, que el traductor lo trasladó como cinco oraciones, por lo que tenemos que tomar decisiones difíciles. Verbigracia, si tenemos un problema en donde tres oraciones o dos oraciones se tradujeron como una oración, entonces tenemos que tomar decisiones. Yo, en lo particular, aparte de mantener la información —en los casos donde se le pudo mantener— lo que hice fue cambiar un punto y seguido por coma; a veces, una oración puede contener varias oraciones en problemas de punto y coma, una oración como punto y aparte, sin embargo, allí se puede tomar la decisión de eliminar esas oraciones; la verdad es que la mayor parte del documento tiene oraciones 1 a 1, no nos afecta estadísticamente tomar decisiones. Para mí lo importante es que la información esté allí, que pueda regresar otra vez a su posición de 2 a 2, por ejemplo, o pueda mantenerse así 3 a 3 cambiando algún punto y coma por punto y seguido o viceversa.

El algoritmo de alineamiento, un trabajo bastante interactivo, es el siguiente: primero, como les había comentado, empezamos numerando los párrafos en el lenguaje A, después los párrafos en el lenguaje B. Aquí cabe mencionar que en un principio empecé a numerar los párrafos y oraciones; luego tenía problemas separados, por lo que me regresaba al principio, incluso hubo ocasiones donde perdían todas las modificaciones que se habían hecho a los documentos numerados en párrafos y todo el trabajo se perdía.

Finalmente, lo que se llevó de mejor manera fue que cada cambio debería estar marcado en el documento original etiquetado en *Word* y tener que procesar todo, otra vez, para que, en un futuro, tuviéramos la fuente pura y fuéramos capaces de regenerar el documento numerando párrafos y oraciones. En el algoritmo se trabajan primero los párrafos, se enumeran y se verifica que tengan el mismo número de éstos; cuando lo tienen, terminamos este ciclo. Entonces verificamos si tienen el mismo número de oraciones, corremos el

programa para que encuentre diferencias. Si encuentra diferencias, de nuevo tomamos decisiones y regresamos a repetir todo el proceso anterior hasta que finalmente los dos documentos tengan lo mismo en cada párrafo, el mismo número de oraciones. Podemos decir que los dos documentos están alineados.

Aquí hay que alinear los documentos en español con los documentos en náhuatl. El cuadro 2 presenta un diagrama de flujo de este proceso, que se lleva a cabo cada vez que haya un cambio: hay que volver a reenumerar párrafos y las oraciones; para todo el proceso hay que pasarlo a .txt y después pasarlo a formato UTF-8 (8-bit Unicode Transformation Format) para tener un documento XML; así obtenemos el documentos final etiquetado con los párrafos y las oraciones numeradas.



Cuadro 2. Diagrama de flujo que presenta el proceso cada vez que hay un cambio

Al contar con los documentos alineados, podemos empezar a trabajar con el alineamiento de palabras; en un principio, comenzamos con la fórmula de información mutua, es decir

$$"I(x:y) = \log_2 \frac{\text{Prob}(x,y)}{\text{Prob}(x) \text{Prob}(y)}", \text{ que se maneja con base en probabilidades.}$$

Tenemos tres probabilidades. La probabilidad de “x”, donde contamos con un número de oraciones en el texto. La probabilidad de “x” es el número de oraciones, donde aparece la palabra “x” sobre el número total de las oraciones del texto, lo mismo la probabilidad de “y”; y la probabilidad de “x:y” es la probabilidad de que las dos palabras concurren en una oración, esto nos da un valor que se llama *información mutua*.

Implementamos los programas para que calcularan la información mutua sobre dos candidatos y los resultados que obtuvimos es que la palabra que nosotros veíamos que era la más probable para ser la correspondencia a la palabra que buscamos no era la que nos daba la información mutua más alta y empezamos a investigar y encontramos que era debido a un problema que se conoce como de “House-Chambre”, que, de hecho, también Gate y Choice encontraron.

En un texto determinado se quiere calcular información mutua, para lo que se cuenta con “House y Chambre” y con “House y Comun”; ya que en inglés la palabra “Chambre Comun”, la cámara de los comunes, es equivalente a la *House of Lords*, de los ingleses; se está haciendo un cálculo entre House y Chambre, el valor de 31950 es el valor donde aparecen juntas. 12004 es donde House aparece únicamente 4793 es para Chambre y donde no aparecen 848333.

	chambre		
house	31,950	12,004	I= 4.1
	4,793	848,330	
	comunes		
house	4,974	38,980	I= 4.2
	441	852,682	

Cuadro 3. Problema "House-Chambre"

Obsérvese el cuadro 3. En comunes vemos que aparecen simultáneamente 4,974, en House aparecen 38,980, comunes aparecen 441 veces y no aparece ninguna de las dos en 852,682. Se calcula la información mutua y se obtiene la más alta con comunes que con Chambre; aquí ocurre algo mal. Básicamente lo que sucede en el fenómeno es que con información mutua, siempre que aparezca la palabra en House, no importa que aparezca en comunes, su información mutua reporta índices muy altos. En cambio, vamos a suponer que comunes aparece 5 veces, pero las 5 veces que aparece, aparece con House, la información mutua siempre va a dar el máximo; ellos comentan que el problema es que no están tomando la información de la diagonal, por lo que diseñan una nueva propuesta: utilizan otro índice llamado ϕ^2 , que sí aprovecha la diagonal y es una medida de asociación, que reporta mejores resultados, reflejando con más frecuencia que una palabra está concurrendo con otra, lo que nos va a dar un índice alto, mientras que una que concurre menos, nos da un índice bajo.

Esta fórmula es un poco compleja. Vamos a simbolizar lo siguiente:

N = número total de regiones

a = $\text{frec}(A \text{ y } B)$

b = $\text{frec}(A) - \text{frec}(AB)$

c = $\text{frec}(B) - \text{frec}(AB)$

d = $N - a - b - c$

Finalmente, la fórmula de ϕ^2 es ésta:

$$\phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Es una fórmula bastante compleja, pero refleja bastante bien los resultados. En éstos, escogimos la primera oración con la que trabajamos y vamos a analizar la palabra "año" que tiene mayor frecuencia en el documento. Estuvimos trabajando con el primero que es por índice alfabético *Anales de Tlatelolco*, por eso

lo seleccionamos; contiene más o menos setenta y cuatro párrafos. Buscamos la primera ocurrencia de la palabra “año”, que aparece en determinado párrafo, en determinada oración, y de allí la tomamos. El texto fue el siguiente:

Al morir Tozcuécue se asentó Huitsilihuitzin; y éste llevaba 23 años gobernando cuando los mexicas fueron despojados en el año 1 Tochtli 1298, aunque algunos lograron salvarse.

Yn omic nima ualmotlai y Uitziliuitzin; e yuh XXIII xihuitl tepachohua ynic mmoyalo que mexica ypan Cen Tochtlixiuitl, au yn oc quezqui mocauhque.

En este caso tenemos un problema de carga de trabajo para la computadora, puesto que las relaciones serían matriciales; encontramos todas las palabras en náhuatl, todas las palabras en español; habría que calcular la ϕ^2 de todas contra todas y sacar la máxima para encontrar las correspondencias. Sin embargo, ese trabajo sería bastante pesado para la computadora tanto en tiempo de cómputo como en espacio de memoria; habría, entonces, que hacer una selección de posibles candidatos. Una forma es escoger una oración y calcular la ϕ^2 por cada una de las palabras de esa oración y ver qué resultados obteníamos con esas palabras; sabemos que aquí muy probablemente está la palabra, pero no sabemos cuál es, y queremos saber cuál de esas palabras es la que tiene la ϕ^2 más alta, que nos va a dar un índice más probable.

Aquí vale la pena comentar que en nuestro algoritmo de contabilización de palabras, lo que se hace es que una vez que encuentra la cadena “año”, la contabiliza como una ocurrencia. Es independiente de estas oraciones. En este caso, el programa revisa esta oración y cuando encuentra la subcadena “año” en esta oración contabiliza como una ocurrencia de la subcadena; hasta cierto punto no nos afectó porque en realidad “año” y “años” es de alguna forma el mismo lema, pero es un detalle que vale la pena aclarar en cuanto a la contabilización. También estamos haciendo cálculos gruesos, burdos, ya que existen oraciones en las que aparece “año” dos veces; nosotros no estamos haciendo contabilizaciones dobles, estamos haciendo contabilizaciones simples y “años” es una palabra particular que corresponde a su singular, a su lema “año”.

Gabriela Bayona: Cabe aclarar que en esta traducción, en particular, cuando se hizo la paleografía, se cambió la puntuación del manuscrito original en náhuatl y se adecuó a la puntuación con la que iba a quedar en español, por eso tenemos alineamiento 1 a 1 casi siempre; los alineamientos que tuvimos que arreglar fueron muy pocos. En este par de oraciones en náhuatl y español podríamos tener puntos anclas, como son los números; en náhuatl se encuentra 23 en número romano y tenemos el año 1 Tochtli y ese sustantivo, creo yo, “Tochtli”, aparece también en náhuatl; entonces tenemos esas dos anclas que a simple vista podrían servirnos. Baste hacer la aclaración de que el programa no calcula con puntos ancla.

Sergio Páez: Contamos con una corrida de nuestro programa con ϕ^2 y la lista de palabras contra las que estamos calculando (véase cuadro 4). Se trata de la lista de palabras de la oración en náhuatl, donde encontramos el cálculo en ϕ^2 . También vemos las ocurrencias de la palabra en lenguaje A, que es “año”, luego las ocurrencias de las palabras en lenguaje B y tenemos las concurrencias en la cuarta columna.

En la columna de ϕ^2 hayamos el índice más alto que puedan encontrar. “Xihuitl” tiene la información mutua más alta, con 0.51; la diferencia de información mutua siempre tiene un intervalo de 0 a 1 porque en información mutua puede dar valores muchos mayores que 1, dependiendo del número de oraciones que se tengan y del número de ocurrencias de las palabras puede dar índices mayores que uno.

PALABRA EN LA LENGUA A: AÑO
VECTOR DE Φ^2

Pal. Leng. B	Fi Cuadrada	OcA	OcB	OcAB
yn	0.055520104	298	634	165
omic	0.011666523	298	34	20
nima	0.041955249	298	195	29
ualmotlali	0.009086255	298	23	14
y	0.078543030	298	828	243
uitziliuitzin	0.002244619	298	1	1
e	0.003979520	298	877	287
yuh	0.003863225	298	64	28
xihuitl	0.009008603	298	4	4
tepachohua	0.002244619	298	1	1
ynic	0.000071971	298	109	35
namoyaloque	0.002244619	298	1	1
mexica	0.002147410	298	113	31
ypan	0.002387951	298	36	16
cen	0.002186142	298	47	20
tochtli	0.126285692	298	53	53
xiuitl	0.516289800	298	188	186
au	0.039884206	298	529	134
oc	0.011333449	298	615	225
quezqui	0.000827745	298	6	1
mocauhque	0.000284648	298	2	1

Cuadro 4. Listas de palabras de la oración en náhuatl

Vemos que el número de ocurrencias de la palabra “año” es 298. Para el número de ocurrencia de la palabra “Xihuitl”, en este caso, el algoritmo encuentra cadenas. Por lo tanto, si aparece la cadena “Xihuitl”, la contabiliza. Esto nos va a ayudar, porque el náhuatl es una lengua aglutinante, puede utilizar “Xihuitl” dentro de la palabra y el algoritmo la va a detectar y la va a contabilizar como una ocurrencia. Hay que tener en cuenta cómo está trabajando el algoritmo de búsqueda de palabra dentro de la oración; se puede cambiar a que sea estrictamente la palabra. En este caso lo que encuentra es la subcadena, la contabiliza y está contabilizando 188 ocurrencias de la cadena “Xihuitl”; las concurrencias son 186. Estas dos palabras son muy útiles; tienen características muy buenas para ésta, muy ideales; no todas se comportan como esta palabra.

“Tochtli” apareció 53 veces en la oración con “año”. Aquí, por ejemplo, en la información mutua, seguramente “tochtli” nos hubiera dado una información mutua más alta, de hecho, la mayor, porque cuando aparece, siempre aparece en la oración “año” y sin embargo ϕ^2 nos está diciendo que no; esto es, se observa que la diferencia es relativamente alta de .51 a .12: el algoritmo ϕ^2 es bastante bueno. Las demás son bastantes

bajas; hay que tomar en cuenta que básicamente estamos trabajando con la hipótesis de que las palabras que ocurran más que una oración tienen mayor probabilidad de ser las correspondencias. En general sería todo.

Ahora vamos a consultar en el diccionario la palabra “xihuitl” y vemos que la primera de sus traducciones es “año”; también observamos —y esto es interesante— que también aparece “xihuitl” con “h”; esto es un punto importante porque la estructura no está estandarizada en los documentos, en ocasiones el autor escribió “xihuitl” con h y “xiuitl” sin h. Por pronunciación la debería de llevar, pero la omitió por práctica.

La palabra “xihuitl” aparece 4 veces en oraciones donde se encuentra año, pero, como no tiene tanta frecuencia, su ϕ^2 es más baja, 0.09, y su información mutua de nuevo es la máxima.

Así concluimos la presentación. Esto nos abre muchas ideas para hacer más programas para procesar esta información, sin embargo, por el momento es un *stop* para presentar los resultados que son bastante frescos; la ϕ^2 apenas la obtuvimos la semana pasada y queríamos impresionar a nuestro asesor al darle esta noticia. Con esto concluimos nuestra plática y esperamos que les haya gustado.

Margarita Palacios: Agradecemos a Sergio y a Gabriela su intervención, y creo que se equivocaron de lugar porque en el Aula Magna está ahora el seminario *A sus cincuenta años de cultura náhuatl*. Así que felicito nuevamente su trabajo; yo creo que es una gran aportación. ¿Alguien quiere abrir un diálogo?, si fueran tan amables.

SECCIÓN DE PREGUNTAS

Gerardo Sierra: En efecto —lo comentábamos con Margarita y contigo la semana pasada— vemos muy lejos la modificación que le ibas a hacer en Church a ϕ^2 , pues ver los resultados ahora es sorprendente, y sí se ve muy diferente a lo que tradicionalmente se ha usado de información mutua. Justamente queremos ver que no necesariamente lo que siempre se utiliza es lo mejor y que justamente hay en algún lugar del mundo alguien que está trabajando el tema.

Estábamos viendo que era “año”. En la tabla estás comparando “año” con “xihuitl”, con “h”, pero realmente no era “año”, sino “años”. En ese sentido la pregunta sería si 23 años corresponde a “xihuitl”. Lo que tú estás diciendo es que “xihuitl” con “h” las 4 veces corresponde con “año” entonces, puede ser una de dos: todas las veces que hay “año” casualmente también hay “años”. ¿En todos los párrafos donde hay “años” más bien hay “año”, o ustedes hicieron lematización de “años”?

Gabriela Bayona: La correspondencia sería con “xihuitl” no con “xiuitl”, porque está el 23 junto y es una forma gráfica de frecuencia baja, es decir, sólo aparece 4 veces en el corpus. La forma gráfica dominante es “xiuitl sin h”, lo que hace el programa es que contabiliza la frecuencia total del corpus a pesar de que físicamente se puede pensar que es la correspondencia física.

Sergio Páez: ¿Podrías repetir tu pregunta?

Gerardo Sierra: Sí. Lo que pasa es que “xihuitl” con “h” ocurre 4 veces y casualmente esas 4 veces corresponde con “año”, pero en este caso “xihuitl” correspondía con años.

Encontramos “xihuitl” con “h” y “xiuitl” sin “h” en el mismo párrafo, justamente cuando aparece “año” aparece “xihuitl” con “h” y algunas veces sin “h”, entonces, en este caso en particular debió de haber coincidido con “año”.

Sergio Páez: Vamos a separar: una cosa es el cálculo estadístico en todo el documento y otra cosa son estas dos oraciones, y cuando se contabilizaron estas 2 oraciones, el programa empezó a contabilizar buscando en español y encontró la subcadena “año”, donde paró, y sumó una más a su cuenta de año. Cuando contabilizó contra “xihuitl” con “h”, empezó a buscar “xihuitl” con “h” y le sumó una de esas 4. De esta forma, las oraciones contribuyeron para el cálculo de la ϕ^2 cuando se requirió; en el caso de “xiuitl”, también buscó “año” y después llegó hasta aquí, por lo que contabilizó el “xiuitl” como una ocurrencia. Estos encuentros sirvieron para el cálculo de las ϕ^2 , lo cual nos lleva a que lo único para que nos sirve esta oración es para sacar candidatos.

De hecho, no hemos acabado este ejemplo; debimos de haber buscado una oración donde nada más apareciera “año” una vez y “xihuitl” para hacer la explicación un poco más simple. Pero esto sirvió para ver cómo se está contabilizando la ϕ^2 . En este momento lo que nos está importando son los candidatos.

Alfonso Medina: Bueno, es que esto de la lematización no quedó muy claro ¿no hicieron lematización? Es muy complejo porque son 2 lenguas: no sólo se trata de lematizar español, sino también náhuatl. ¿Qué tal si ocurre la palabra “añoranza”? Entonces “añoranza” seguramente no va a aparecer como “xihuitl” y la lematización sería recomendable aunque fuera muy sencilla, porque además comentaste que “xihuitl” puede aparecer combinada, como 20 años, y la lematización en náhuatl podría ayudarte a colocar tanto a “xihuitl” con “h” como “xiuitl”, sin “h”, juntas en un mismo lema. Sería algo manual.

Sergio Páez: En un principio fue nuestro primer acercamiento; obviamente, empiezan a salir los comentarios. No es tan fácil encontrar una palabra, buscar una palabra tiene sus bemoles. Decir si la palabra está o no está no es nada más buscar la subcadena. Esto no es tan sencillo y se obtiene del análisis de los resultados.

Javier Cuétara: Podría seguir hablando de “año” y eso, pero a mí me gustó mucho tu trabajo y todo lo que tenga que ver no sólo con la relación entre la lingüística o la ingeniería o la computación; me parece un mérito mayor que trabajemos la lenguas vernáculas. Ojalá no sea sólo un trabajo de tesis, esperemos que prospere y se incorporen muchos más. Claro, están los problemas de los corpus textuales: ¿dónde están?, ¿cómo obtenerlos? Hago este comentario porque atiende a tus últimas palabras que fueron: “Ojalá que les haya gustado”. En efecto, a mí me gustó y estoy verdaderamente seguro de que a todos los que estamos aquí, también. Felicidades. Que tengan muchos resultados, para que nos lo sigan compartiendo.

Margarita Palacios: Me alegra muchísimo haber estado moderando esta mesa.

Gerardo Sierra: Para clausurar quisiera decir unas pocas palabras a esta concurrencia. Este es el tercer Coloquio que hemos hecho, y ha superado a los anteriores. Lo superó en el número de ponencias, en el número de asistencia, se nos acabaron los folletos; en algún momento, me dijeron para qué quieres más de cien si eso es demasiado y se nos fueron volando en el primer día, en el segundo día se hubieran ido los otros cien. En las ponencias, que fueron por una parte de la UNAM, estuvieron el Instituto de Ingeniería, el IIMAS, que siempre hemos presentado, se integró en el segundo Coloquio la participación de la Facultad de Ciencias y

en este Coloquio se incluyó además la Facultad de Ingeniería. Por otro lado, la participación de nuestro colegas hermanos, el CIC, que siempre ha sido constante. La ponencia que nos faltó, la de VALIDE, era de El Colegio de México, que también es otro de los integrantes más en este Coloquio. Los participantes cada vez somos más, lo cual hay que celebrar.

No me queda más que agradecerles y seguirles entusiasmando para que sigamos adelante en estas tareas y como vemos hay mucha colaboración entre los que estamos, ya que no podemos trabajar aislados ni los lingüistas, ni los ingenieros en computación. Por otro lado, esto tiene que ser también interinstitucional, como ya lo hemos mencionado. Damos por terminado. Muchísimas gracias.

ÍNDICE DE ABREVIATURAS DE INSTITUCIONES

CEPE	Centro de Estudios para Extranjeros
CIC	Centro de Investigación en Computación
CONACyT	Consejo Nacional de Ciencia y Tecnología
FCA	Facultad de Contaduría y Administración
FFyL	Facultad de Filosofía y Letras
FI	Facultad de Ingeniería
GIL	Grupo de Ingeniería Lingüística
IIMAS	Instituto de Investigación en Matemáticas Aplicadas y en Sistemas
IINGEN	Instituto de Ingeniería
IPN	Instituto Politécnico Nacional
IULA	Instituto Universitario de Lingüística Aplicada
UAM	Universidad Autónoma Metropolitana
UNAM	Universidad Nacional Autónoma de México
UPV	Universidad Politécnica de Valencia



Gerardo Sierra Martínez, doctor en Lingüística Computacional por la University of Manchester, Institute of Science and Technology (UMIST), Inglaterra. Jefe del Grupo de Ingeniería Lingüística (GIL) del Instituto de Ingeniería en la UNAM. Ha promovido e impulsado el área tanto a nivel docencia como investigación y desarrollo, en áreas como lexicografía computacional, terminológica, recuperación y extracción de información, minería de textos y corpus lingüísticos.

Actualmente es investigador titular B, pertenece al Sistema Nacional de Investigadores (nivel II), es evaluador de proyectos Conacyt y miembro de varios comités científicos y editoriales.

Ha impartido cursos conjuntos en la UNAM para las Facultades de Ingeniería y de Filosofía y Letras, así como en los Posgrados de Lingüística, de Bibliotecología y de Ciencias de la Computación. Ha logrado que en el Programa de Estudios de la Facultad de Ingeniería se abra un módulo de especialidad sobre Tecnologías del Lenguaje en la Carrera de Ingeniería de Computación.



Javier Cuétara Priede, maestro en Lingüística Hispánica y licenciado en Lengua y Literaturas Hispánicas por la UNAM, es profesor e investigador adscrito a la Facultad de Filosofía y Letras y al Centro de Enseñanza para Extranjeros de la UNAM, donde imparte diversas materias de licenciatura, especialización y diplomado, con un espectro variado, en áreas tan diversas como: tecnologías del habla, lingüística computacional, fonética instrumental, enseñanza de la pronunciación del español a extranjeros, dialectología hispánica y fonética diacrónica.

Ha pertenecido a diversos comités académicos y artísticos en la UNAM, el Instituto Nacional para la Evaluación de la Educación (INEE), el Centro Nacional de Evaluación para la Educación Superior (CENEVAL) y el Instituto Nacional de Bellas Artes (INBA), donde ha destacado también como promotor cultural.

Colabora intensamente con el Grupo de Ingeniería Lingüística y el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), ambos de la UNAM. Fue director del Centro de Enseñanza para Extranjeros-Taxco (CEPE-Taxco), donde fomentó las áreas académicas, artísticas, culturales y ecológicas. Asimismo, fue coordinador de la carrera de Lenguas y Literaturas Hispánicas.

Esta obra se terminó de imprimir en el mes de agosto de 2015, en el Instituto de Ingeniería, Circuito Escolar, Ciudad Universitaria, CP 04510, México DF.

La edición, impresa en papel Bond consta de 75 ejemplares.



www.iingen.unam.mx



www.iling.unam.mx