AVANCES EN
LEXICOGRAFÍA, TERMINOLOGÍA Y TRADUCCIÓN

Marisela Colín Rodea
Erika Ehnis Duhne
(coordinadoras)

ENALLT ESCUELA NACIONAL DE LENGUAS, LINGÜÍSTICA Y TRADUCCIÓN UNAM

Sésamo, proyecto integral en ingeniería lingüística

GERARDO SIERRA

Introducción

Un término como *ingeniería lingüística* nos abre naturalmente a preguntarnos qué quiere decir, con lo cual podríamos llegar a formularnos una serie de conjeturas al respecto. ¿Se trata de una lingüística con métodos de la ingeniería?, esto es, ¿se refiere a una lingüística cuantitativa o a una lingüística matemática? Ambas nos parecen razonable al considerar que la ingeniería utiliza métodos cuantitativos y, por tanto, basados en las matemáticas. Y aunque hay algo de cierto en todo esto, no es así. Entonces, ¿podríamos pensar en una lingüística de la ingeniería o bien para la ingeniería? Considero que esto lo descartaríamos, pues de manera análoga, y por tanto equivocada, también diríamos que una lingüística de la biología sería una biología lingüística.

Con un poco más de conocimiento y sabiendo que la lingüística cuantitativa o la matemática que tienen principios propios y no tienen nada que ver con la ingeniería, cabe entonces preguntarnos qué haría o sería la ingeniería lingüística. Desde el punto de vista de algunos colegas lingüistas, se trata de herramientas de cómputo para lingüistas, como los procesadores de textos, los bancos de datos o las hojas de cálculo, todos ellos son de gran utilidad, más con los aditamentos y ventajas que ponen a nuestro alcance hoy en día. En ese mismo orden de ideas, están los recursos lingüísticos electrónicos, tales como diccionarios, corpus o acervos bibliográficos. De manera más específica, tenemos el *software* especializado para lingüistas, como un ejemplo, para realizar análisis fonético, para traer las concordancias de una palabra o en algún analizador sintáctico. Temo desilusionar a los colegas esperanzados en que

estos sean los objetivos de dicha área, aunque tampoco lo desmiento del todo, pues de alguna manera la ingeniería lingüística busca desarrollar herramientas de cómputo e incluso *software* especializado para las labores lingüísticas, en tanto que se nutre —y en algunos casos también desarrolla— de diversos recursos lingüísticos electrónicos.

Para entender el área, conviene primero conocer otra muy cercana: la *lingüística computacional*. Podemos entenderla como una interdisciplina entre la lingüística y las ciencias de la computación, mediante la cual se proponen modelos de la lengua en términos formales e inteligibles a las computadoras para la creación de sistemas de cómputo que realicen, en lo posible, las mismas actividades lingüísticas que los seres humanos, esto es, reconocer, entender, interpretar y generar lenguaje humano en todas sus formas, tanto oral y escrito como el de señales o señas. Por su parte, la ingeniería lingüística constituye la parte más aplicada de la lingüística computacional y está dirigida a la obtención de productos para el mercado. Para ello, utiliza una serie de técnicas propias y se basa en un conjunto de recursos lingüísticos que se aplican en el primer caso, por medio de programas de cómputo y que, en el segundo, constituyen una fuente de conocimientos a los que se puede acceder por medio de programas informáticos (*LingLink*).

Entre las múltiples aplicaciones de la ingeniería lingüística, simplemente para dar idea de la diversidad, cabe mencionar, la traducción automática, la recuperación y extracción de información, las interfaces en lenguaje natural, el aprendizaje asistido por computadora, los sistemas de ayuda a la redacción de documentos, las herramientas para el discapacitado, la identificación de sonidos, el reconocimiento tanto de voz como de caracteres escritos, la generación de documentos, el correo electrónico multilingüe y los diccionarios electrónicos.

El grupo de ingeniería lingüística

En septiembre de 1999 recibí la invitación del Secretario Académico del Instituto de Ingeniería de la UNAM para un área aún poco

conocida en México, con la tarea de formar el Grupo de Ingeniería Lingüística (GIL) con un doble objetivo: primero, crear una base de conocimiento relativa y concerniente a esta área de trabajo poco explotada en México; y segundo, formar gente especializada y comprometida con el estudio y desarrollo de toda la gama de oportunidades que ésta ofrece.

Con estos objetivos en mente y gracias al apoyo del Instituto de Ingeniería, así como también con el patrocinio principal del Consejo Nacional de Ciencia y Tecnología y de la propia UNAM, durante este tiempo transcurrido en el GIL hemos realizado diversos proyectos vinculados con el procesamiento de lenguaje natural. Sin embargo, con fines estratégicos, hemos tenido como eje rector un proyecto central de investigación aplicada, sobre el cual giran las diferentes líneas de investigación. Este proyecto central, motor del GIL, surge de un prototipo diseñado y elaborado durante el periodo comprendido de 1992 a 1996 en el Instituto de Ingeniería, con el cual se buscó crear un sistema de búsqueda onomasiológica, esto es, un diccionario que permita la búsqueda de términos a partir de la descripción del concepto mediante el uso de lenguaje natural. El prototipo se corroboró para 33 términos concernientes al área de desastres.

Con base en los intereses de investigación en el ámbito de la ingeniería lingüística hemos tomado el proyecto central con el fin de obtener por un lado, un producto específico aplicado a dominios de especialidad, con lo que cumplimos el compromiso con los patrocinadores y recibir recursos para la formación de personal y adquisición de equipo; por el otro lado, logramos desarrollar técnicas de punta y avanzar en la creación de diferentes líneas de investigación de la ingeniería lingüística, las cuales son necesarias durante las distintas fases del proyecto central, tales como lingüística aplicada —en particular, terminología y lexicografía—, lingüística computacional, ciencias de la computación e informática, bibliotecología y ciencias de la información, por nombrar sólo lagunas. Gracias a la metodología desarrollada en las distintas fases del proyecto ha sido posible elaborar sistemáticamente y en un tiempo razonable diccionarios integrales que permiten tanto la búsqueda semasiológica como la onomasiológica, aplicados a diversas áreas de conocimiento.

Hoy en día, después de ocho años de creación del GIL, hemos realizado varios proyectos de investigación y hemos logrado obtener algunos desarrollos. Toda la información detallada se encuentra nuestra página de Internet: www.iling.unam.mx

El proyecto Sésamo

La denominación del proyecto *Sésamo* proviene de la conocida historia de Alí Babá y los cuarenta ladrones. Recordemos que en este famoso cuento el codicioso Casím entra a la cueva debido a que pronunció con precisión las palabras mágicas, con una docena de mulas preparadas con cofres y cestas vacías para llenarlos de oro, plata y joyas. Sin embargo, seguramente debido a su emoción, olvidó las palabras mágicas para abrir la puerta de la cueva. Recordaba el «ábrete…» pero lo demás, aunque lo tenía en la punta de su lengua, desgraciadamente no había forma de acordarse. Ustedes saben lo que sucedió al pobre Casím que, por no poder salir, los ladrones acabaron colgando su cuerpo cercenado a la entrada de la cueva para evitar la curiosidad de nuevos intrusos. Pero sin seguir más con la historia, regresemos y supongamos qué hubiera pasado si Casím hubiera ido acompañado de un genio —acorde con las historias de *Las mil y una noches*— que le ayudara a encontrar la palabra olvidada tan sólo proporcionándole la descripción del concepto. Ahora supongamos que Casím tuviera la habilidad de recitar la definición proporcionada por la Real Academia de la Lengua Española y, por tanto, le dijera al genio la siguiente descripción: «Planta herbácea, anual, de la familia de las Pedaliáceas».

Coincidirán conmigo en que, difícilmente el genio podría haberle ayudado, quien a pesar de ser un mago era un lego en conocimiento sobre las cosas. Si Casím siguiera con la descripción dada por el diccionario de la RAE, diría entonces: «Semillas amarillentas, muy menudas, oleaginosas y comestibles».

Con esta última el genio podría estar más cerca. Pero con toda seguridad, el genio le habría atinado si Casím dijera algo como: «Cosita que se pone arriba del pan de las hamburguesas», con lo

cual prácticamente todos sabemos que se refiere al ajonjolí o, por su equivalencia al inglés, al sésamo. Con esta nueva historia cabría preguntarnos si existe la posibilidad de contar con un diccionario que nos ayude a encontrar las palabras olvidadas con descripciones similares a esta última, esto es, con base en el conocimiento que tiene la mayoría de la gente y que no necesariamente coincide con el de los académicos. El objetivo del GIL ha sido desarrollar la metodología para obtener diccionarios de búsquedas onomasiológicas a partir de la descripción del concepto en lenguaje natural. Como mencioné, este diccionario constituye el eje central de diversas líneas de investigación, las cuales explicaré y que se basan en las diferentes etapas del desarrollo de dicho diccionario, mostradas en la Figura 1.

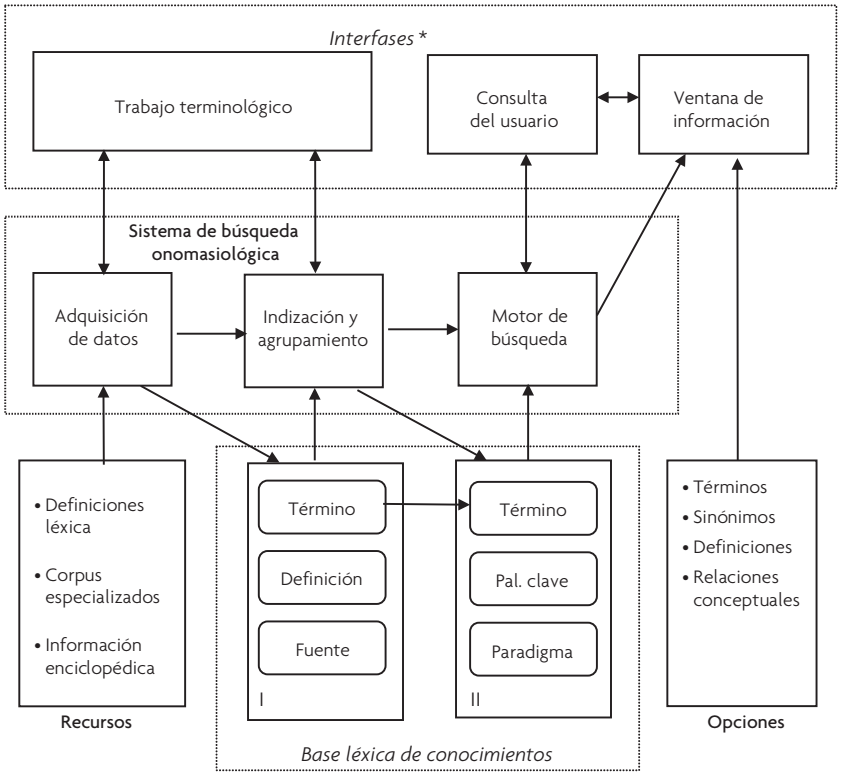

Figura 1. Arquitectura del diccionario onomasiológico.

Recursos léxicos

El punto de partida para cualquier diccionario de especialidad consiste en la obtención de los términos y sus definiciones. Por un lado podemos utilizar, sin duda, los diccionarios existentes en la materia —y en caso de existir, las enciclopedias—, los cuales constituyen el recurso léxico mínimo indispensable. Sin embargo, no siempre podemos encontrar estos recursos, por lo que la otra posibilidad es recurrir a los textos de especialidad, los cuales no sólo utilizan la terminología correspondiente, sino que además la describen. Identificar dichos textos, contar con ellos en un formato manejable para la computadora y procesarlos integrados, forma parte de una línea de investigación denominada *lingüística de corpus*, que da como resultado los corpus lingüísticos, los que constituyen uno de los recursos lingüísticos más utilizados en la ingeniería lingüística.

Corpus lingüísticos

Gracias a una convocatoria de CONACYT en la que se apoyaba la creación de bases de datos propusimos la constitución del Corpus Lingüístico en Ingeniería (CLI). Con esto buscamos elaborar, desarrollar y mantener un corpus lingüístico que contenga textos selectos en el área de ingeniería debidamente codificados y organizados, así como las herramientas de programación adecuadas para poder utilizar el corpus en el desarrollo de diversas investigaciones en el área de lingüística e ingeniería lingüística. Así, las dos metas principales son: primero, contar con el primer corpus lingüístico en el área de ingeniería, el cual si bien fue un corpus relativamente pequeño —cerca del medio millón de palabras—, fue lo suficientemente amplio en las áreas de la ingeniería y debidamente codificado y anotado para su procesamiento automático; la segunda, diseñar las herramientas computacionales necesarias para llevar a cabo distintos tipos de análisis del corpus, tales como concordancias, conteo de palabras y algunas medidas de colocación.

Debido a los avances del proyecto hemos podido realizar diversos análisis del corpus y hemos podido desarrollar otros sistemas

de ingeniería lingüística y parte del diccionario onomasiológico, como el de extracción de términos y de contextos definitorios (que comentaré más adelante).

La experiencia adquirida en el CLI nos motivó a proponer la creación del Corpus Histórico del Español de México (CHEM), patrocinado por la Dirección General de Asuntos del Personal Académico (DGAPA) de la UNAM, ahora nos encontramos desarrollando el Corpus de las Sexualidades en México.

Adquisición de datos

Con el fin de obtener información relevante recabada a partir de textos especializados y diccionarios existentes para formar la base de conocimiento para el desarrollo del diccionario onomasiológico, esta fase se enfoca hacia la búsqueda y extracción tanto de la terminología como de las definiciones o descripciones dadas en dichos textos.

Extracción terminológica

A partir de los textos de especialidad proporcionados por los corpus lingüísticos es posible obtener la terminología correspondiente, la cual será integrada posteriormente al diccionario onomasiológico. Una de las áreas de la ingeniería lingüística la constituye la *terminótica*, que tiene por objetivo crear herramientas y desarrollar metodologías para las distintas actividades del trabajo terminográfico, como la extracción automática de términos. El GIL ha ocupado una parte de sus investigaciones para esta tarea. En un principio nos apoyamos en herramientas comerciales —específicamente *WordSmith Tools*®— que nos permitieron comparar las frecuencias de las palabras en dos textos de diferentes temáticas y con ello, mediante técnicas estadísticas, pudimos extraer los términos correspondientes a nuestra área objeto de estudio.

Conforme fuimos profundizando en esta línea de investigación y gracias a la colaboración, tanto de colegas de la Universidad de

Manchester y de la Universidad de Montreal, así como con el apoyo del macroproyecto Tecnologías para la Universidad y la Computación de la UNAM, desarrollamos nuestro primer sistema extractor de términos para el español, el cual consiste en un método híbrido que utiliza tanto conocimiento lingüístico como métodos estocásticos.

Extracción de contextos definitorios

Con el fin de alimentar la base de datos de conocimientos del diccionario onomasiológico no sólo con los términos del área de especialidad sino también con una cantidad suficientemente rica de definiciones, iniciamos una línea de investigación bajo el patrocinio de CONACYT y con la colaboración inicial de la Universitat Pompeu Fabra. En una primera etapa, bajo el patrocinio de la DGAPA, realizamos un estudio descriptivo de posibles patrones recurrentes para la introducción de nuevos términos y sus definiciones en textos de especialidad, por medio del cual hicimos un análisis de contextos definitorios y de sus elementos constitutivos: términos, definición, patrones verbales definitorios y patrones pragmáticos. Posteriormente, bajo el patrocinio de CONACYT, profundizamos en las relaciones entre el tipo de definición y el patrón verbal definitorio y desarrollamos reglas lingüísticas y computacionales para sistematizar la identificación de patrones recurrentes, con el fin de extraer de manera automática información tanto sobre las unidades léxicas que se utilizan, como de las reglas de su utilización.

Como desarrollo del GIL en el campo de la ingeniería lingüística, creamos un sistema extractor de contextos definitorios para el español a partir de corpus lingüísticos —incluyendo Internet—, clasificándolos en tres tipos de definición (analítica, funcional y extensional) e identificando sus dos principales elementos constitutivos: término y definición. Trabajamos en un desarrollo propio para extraer los tres tipos de definiciones de textos en Internet para un término introducido por el usuario en cualquier área temática.

Bases de datos

La base de conocimientos léxicos constituye una parte fundamental del diccionario onomasiológico. Por ello, desde un principio hemos dedicado un gran esfuerzo no sólo para crear, sino para actualizar e incorporar a la red un banco terminológico para los fines de nuestros proyectos. Así, ésta resulta ser una herramienta útil y de provecho para aquellos a quienes les interese o necesiten obtener información actual y fidedigna sobre la terminología de un área de especialidad. En este sentido, el banco terminológico del GIL es *multipropósito,* ya que está pensado para cumplir con una gran variedad de necesidades y se toma en cuenta tanto a usuarios en general, como a los que están llevando a cabo sus investigaciones dentro del GIL.

Diseñamos la base de datos terminológica para capturar ágilmente la información de diccionarios electrónicos y la obtenida de los textos de especialidad, así como para vaciar los datos necesarios que integran la base de conocimientos léxica del diccionario onomasiológico. Entre los datos capturados se encuentran el término, su área temática, los contextos definitorios y las definiciones dadas por todas y cada una de las fuentes encontradas. Actualmente, el banco terminológico del GIL cuenta con las siguientes bases en operación: física (con 320 términos), lingüística (con 2,473 términos), desastres (con 1,702 términos en español y 850 en inglés), metrología (con 342 términos en inglés), fenómenos destructivos (con 81 términos en inglés), sexualidad (con 598 términos), ingeniería lingüística (con 91 términos) y córpora lingüísticos (con 94 términos).

Con el fin de proporcionar un servicio de acceso multiusuario que sea útil tanto para los integrantes del GIL, así como para los usuarios en general, como ya se describió, se implantó el banco en un servidor accesible en línea. Así, se obtiene un sistema capaz de proporcionar un fácil acceso a la base de datos a través de la Red, confiable y seguro, permitiendo que sólo los usuarios autorizados alimenten la base de datos de manera simultánea, sin restricción de acceso, y que a la vez permita la consulta a toda persona que visite la página *web*.

La experiencia adquirida en el desarrollo de nuestro banco terminológico nos permitió participar en un proyecto de la Facultad de Me-

dicina Veterinaria y Zootecnia de la UNAM con el fin de diseñar y poner en funcionamiento un banco terminológico de esta área en línea.

Indización y agrupamiento

Los datos capturados en el banco terminológico, en particular de términos y definiciones, deben ser procesados para obtener la base de conocimientos léxicos necesarios para el diccionario onomasiológico. Por un lado, está formada por las palabras clave, esto es, aquellas palabras de contenido que resultan significativas en las distintas definiciones para cada uno de los términos, identificando así los términos que corresponden. Por otro lado, también se integra por los paradigmas semánticos, que no son otra cosa que grupos de palabras clave, las cuales, aunque no pertenezcan a un grupo semántico común —acaso ni siquiera a una función gramatical común— pueden, en ciertos contextos, ser utilizadas como pares sinonímicos referenciales al momento de hacer alusión a una situación u objeto específico.

Identificación de palabras clave

Para la identificación de las palabras clave consideramos todas las definiciones que existen. Usamos una lista de paro que determinamos *ex profeso*, la cual contiene palabras funcionales (tales como preposiciones, conjunciones e interjecciones) y algunos verbos, adjetivos y adverbios. Diseñamos un programa especial para quitar las palabras de esta lista de paro y observar las palabras clave que quedaban. Mediante una comparación entre todas las definiciones que corresponden a un término, así como de las definiciones que corresponden a términos de un mismo campo semántico, hemos podido llegar a determinar de una manera más amplia los rasgos semánticos y, por tanto, las palabras clave que corresponden a cada término. Todas estas palabras clave son capturadas en la base de datos e identifican el término con el que están asociadas.

Determinación de paradigmas semánticos

Con el fin de que el proceso de búsqueda onomasiológica sirva a la descripción dada por cualquier usuario, es preciso que la base del conocimiento sea lo suficientemente rica. Con esta finalidad, el GIL trabaja en la obtención de grupos de paradigmas semánticos con una metodología propia creada *ex profeso*, la cual además explota las definiciones capturadas en el banco terminológico.

Una vez determinados los paradigmas semánticos es posible expandir la formulación de búsqueda inicial del usuario, de manera tal que en la base de datos indexada se buscarán los términos que contengan no sólo las palabras clave introducidas por el usuario, sino además, todas aquellas palabras clave que pertenezcan a los paradigmas semánticos correspondientes. Dichos paradigmas son vertidos en una base de datos adecuada a la aplicación del sistema de búsqueda onomasiológica.

Nuestro método de determinación de paradigmas semánticos se basa en el alineamiento de las definiciones que están en el banco terminológico para cada término, de manera que se comparan los cambios que es necesario hacer para que una definición sea igual a la otra. Este método se basa en otro utilizado en la lingüística computacional, denominado *distancia de edición* (Wagner & Fisher, 1974), pero junto con otros algoritmos nos ha permitido encontrar pares semánticos que difícilmente son encontrados en diccionarios de sinónimos y, sin embargo, pueden ser considerados sinónimos en un contexto determinado como el de la descripción de conceptos. Hemos refinado este método desde los puntos de vista lingüístico y computacional, de manera que ha sido una línea de investigación de interés tanto en un área como en la otra.

Diseño del motor de búsqueda

Para cumplir con el objetivo de producir un diccionario integral es necesario considerarlo como un sistema de búsqueda de información, en este caso, terminológica. Se consideraron tres distintas formas en que el usuario puede introducir los datos e interactuar

con la computadora: operadores booleanos, formulación en lenguaje natural y diálogo con la computadora. Con el fin de no restringir al usuario al uso exclusivo de los operadores booleanos o aquéllos impuestos por el diálogo iterativo con la computadora, desarrollamos un motor de búsqueda que permita la descripción del concepto en lenguaje natural, con lo que permitimos que el usuario pueda expresar su búsqueda sin restricciones de ningún tipo y en su particular lenguaje.

El motor de búsqueda toma en consideración todas las palabras de contenido del usuario, así como sus variantes morfológicas, entonces las analiza con el banco de conocimientos que está integrado por las palabras clave y su agrupación en los paradigmas semánticos. Asimismo, se utiliza un método de jerarquización de los resultados a fin de mostrar al usuario los términos más probables a su búsqueda, con lo que se constituye un sistema inteligente de búsqueda.

Diseño de la interfaz del usuario

Toda la metodología expuesta para el diccionario onomasiológico se concretiza en la interfaz del usuario, en donde el usuario interactúa con el programa de cómputo para buscar los términos que corresponden a una descripción dada. Además de esta búsqueda onomasiológica diseñamos la interfaz, para también permitir búsquedas semasiológicas, es decir, para conocer las definiciones de un término dado. Todo esto lo logramos en lo que denominamos ILex (esto es, Interfaz Lexicográfica), la cual es una interfaz diseñada con criterios ergonómicos para contener el diccionario semasiológico y el onomasiológico, pero que a la vez contiene búsquedas inteligentes. Es el caso, por ejemplo, de que en el diccionario semasiológico es posible buscar la definición de un término aún cuando éste sea introducido con faltas de ortografía. ILex también tiene un árbol de dominio que permite ubicar a un término dado dentro del mapa conceptual al área que corresponde, con lo que el usuario puede encontrar otros términos relacionados. Por otra parte, ILex puede proporcionar imágenes, videos y sonidos que complementan la información lexicográfica. Como valor agregado del ILex, se ofrece una

parte lúdica, en donde mediante los juegos de ahorcado y sopa de letras se permite que el usuario refuerce sus conocimientos en la terminología del diccionario.

El diseño del ILex ha sido una línea de investigación que hemos podido desarrollar gracias al patrocinio de DGAPA y a la colaboración de expertos del CCADET, UNAM, en el área de ergonomía e interfaces inteligentes.

Conclusiones

En el presente artículo he descrito varias líneas de la investigación aplicada de la ingeniería lingüística que han sido desarrolladas a partir de la descripción de las etapas de un proyecto principal, que es el *Diccionario onomasiológico* o *Sésamo*. Con esto hemos cumplido con los objetivos planteados con la creación del Grupo de Ingeniería Lingüística, gracias al apoyo del financiamiento del CONACYT y de la Dirección General de Asuntos del Personal Académico de la UNAM. Esto nos ha permitido integrar un grupo interdisciplinario, no sólo formado por sus miembros que consuetudinariamente asisten al GIL, sino también por aquellos que han puesto su granito de arena a este gran proyecto que constituye formar un área interdisciplinaria con alto potencial de desarrollo. A todos ellos un agradecimiento. Entre todos hemos tenido ya varios desarrollos y escrito varias publicaciones que prefiero omitir en este artículo, pues sería difícil escoger una sobre otra y me llevaría a dejar alguna fuera. Sin embargo, la lista de publicaciones y desarrollos los pueden encontrar en la página del GIL: www.iling.unam.mx

Referencias

LingLink, Harnessing the Power of Language, *Anite Systems*, Luxemburgo.

WAGNER R. A. & M. J. Fisher (1974). The string-to-string correction problem. *Journal of the ACM*, Vol. 21(1), p. p. 168-173.