universidad de león

Departamento de Filología Hispánica y Clásica

APROXIMACIÓN A LA LINGÜÍSTICA COMPUTACIONAL

Milka Villayandre Llamazares

León, 2010

Universidad de León

INFORME DEL DIRECTOR DE LA TESIS

(Art. 11.3 del R.D. 56/2005)

El Dr. D. **Salvador Gutiérrez Ordóñez** y el Dr. D. **Manuel Iglesias Bango** como Directores[1] de la Tesis Doctoral titulada "**Aproximación a la Lingüística Computacional**" realizada por D.ª **Milka Villayandre Llamazares** en el Departamento de **Filología Hispánica y Clásica**, informa favorablemente el depósito de la misma, dado que reúne las condiciones necesarias para su defensa.

Lo que firmo, para dar cumplimiento al art. 11.3 del R.D. 56/2005, en León a ___ de _____ de _____.

[1] Si la Tesis está dirigida por más de un Director tienen que constar los datos de cada uno y han de firmar todos ellos.

Universidad de León

ADMISIÓN A TRÁMITE DEL DEPARTAMENTO
(Art. 11.3 del R.D. 56/2005 y

Norma 7ª de las Complementarias de la ULE)

El Departamento de **Filología Hispánica y Clásica** en su reunión celebrada el día ___ de _____ de _____ ha acordado dar su conformidad a la admisión a trámite de lectura de la Tesis Doctoral titulada "**Aproximación a la Lingüística Computacional**", dirigida por el Dr. D. **Salvador Gutiérrez Ordóñez** y el Dr. D. **Manuel Iglesias Bango**, elaborada por D.ª **Milka Villayandre Llamazares** y cuyo título en inglés es el siguiente "**Computational Linguistics: an approach**".

Lo que firmo, para dar cumplimiento al art. 11.3 del R.D. 56/2005, en León a _____ de _____ de _____.

El Secretario,

Fdo.: _____

Vº Bº

El Director del Departamento,

Fdo.: _____

ÍNDICE DE CONTENIDOS

No puedo combinar unos caracteres

dhcmrlchtdj

que la divina Biblioteca no haya previsto y que en alguna de sus lenguas secretas no encierren un terrible sentido. Nadie puede articular una sílaba que no esté llena de ternuras y de temores; que no sea en alguno de esos lenguajes el nombre poderoso de un dios. Hablar es incurrir en tautologías. Esta epístola inútil y palabrera ya existe en uno de los treinta volúmenes de los cinco anaqueles de uno de los incontables hexágonos—y también su refutación. (Un número *n* de lenguajes posibles usa el mismo vocabulario; en algunos, el símbolo *biblioteca* admite la correcta definición *ubicuo y perdurable sistema de galerías hexagonales,* pero *biblioteca* es *pan* o *pirámide* o cualquier otra cosa, y las siete palabras que la definen tienen otro valor. Tú, que me lees, ¿estás seguro de entender mi lenguaje?).

(Jorge Luis Borges, *La biblioteca de Babel*)

0. INTRODUCCIÓN

0. INTRODUCCIÓN

Esta tesis surge con el objetivo de intentar dar respuesta a preguntas básicas planteadas a raíz de una primera toma de contacto con el campo de la Lingüística Computacional (en adelante LC):

1) ¿Qué implica tratar el lenguaje con ordenadores?
2) ¿Cuál es su finalidad?
3) ¿Qué requisitos previos exige?
4) ¿Cómo se lleva a cabo tal tarea?
5) ¿Por qué surgió este acercamiento al tratamiento del lenguaje?

Caben dos perspectivas de aproximación al objeto de estudio: la lingüística y la informática. Como no podía ser de otra forma, el acercamiento que realizamos a esta disciplina parte de nuestra formación lingüística, aunque, no obstante, se harán referencias a la perspectiva de la Informática cuando así lo requiera el tema.

Por otra parte, nos interesa sobre todo su estatus en España, pero sin dejar de lado los referentes internacionales que han sentado las bases teóricas sobre las que se sustenta la LC, porque, como bien dice G. Rojo en el prólogo a un libro de reciente aparición, las publicaciones sobre LC aún son escasas en nuestro país: "No disponemos todavía de bibliografía suficiente en español sobre temas de Lingüística Informática y Computacional, sobre Traducción Automática o Generación del Lenguaje" (Lavid 2005:23).

Precisamente, esa fue la situación que me encontré hace ya algunos años cuando por primera vez me acerqué a este campo de

investigación, en torno a 1996. El material disponible era muy escaso, pero en un par de años, entre 1998 y 2000 aparecieron en el mercado varios manuales. Poco antes, a principios de los 90, se había creado un programa de Tecnología Lingüística en el Área de Industrias de la Lengua de la Sociedad Estatal Quinto Centenario, se editan recursos lingüísticos como el Archivo Digital de Manuscritos y Textos Españoles (ADMYTE), asistimos a las primeras etapas de los corpus académicos, CREA y CORDE, o la puesta en marcha del Seminario de Industrias de la Lengua de la Fundación Duques de Soria (*vid.* Llisterri y Almiñana 1998), por mencionar algunas de las iniciativas más destacadas.

No es de extrañar que este campo suscitara el interés de algunos investigadores, al conjugar dos ingredientes tan atractivos como el lenguaje y una herramienta relativamente nueva, los ordenadores. Como dice M. Bates (1994:239), el lenguaje es tan importante en nuestras vidas que su uso fluido es casi sinónimo de inteligencia[1].

El lenguaje ha sido el centro de atención de diversas disciplinas científicas: filosofía, lógica, psicología, biología, antropología y, por supuesto, lingüística. Cada una ha aportado sus métodos y teorías para su descripción. Desde el momento en que se utiliza el ordenador para el estudio de una conducta en general y el lenguaje en particular, se puede hablar de un nuevo paradigma de investigación (Winograd 1972:ix):

> *This book is part of a newly developing paradigm for looking at human behaviour, which has grown up from working with computers. When faced with highly complex and organised behaviour like language, we ask 'What kind of process could be going on to produce that behaviour?' Computers and computer language give us a formal metaphor, within which we can model the processes and test the implications of our theories.*

[1] "Language is so fundamental to humans, and so ubiquituous, that fluent use of it is often considered almost synonymous with intelligence".

> *Our models are of necessity incomplete. It is not yet clear what connections they have with the processes going on in the human mind. Yet they give us a clear framework for thinking about what it is we do when we understand and respond to natural language*.

En este nuevo paradigma los ordenadores proporcionan una metáfora formal para modelar y probar las teorías que tratan de dar cuenta del funcionamiento del lenguaje, de cómo somos capaces de entender, qué procesos subyacen a la conducta lingüística. Como afirma T. Winograd (*vid. supra*), esos modelos formales para describir el funcionamiento del lenguaje no tienen por qué reproducir exactamente el funcionamiento de la mente humana. Basta con que nos faciliten una plataforma para acercarnos a la comprensión de un fenómeno tan complejo a la vez que tan cotidiano como es el lenguaje.

Lógicamente, la tarea entraña no pocas dificultades, debido a la complejidad inherente a las propias lenguas naturales, tal y como lo expresan Edwards y Kingscott (1997:16-18), cuando dicen que dominar el lenguaje natural es una de las tareas más difíciles que se le puede pedir a un ordenador; jugar al ajedrez como un gran maestro resulta, en comparación, relativamente simple. El lenguaje humano es con frecuencia alusivo y ambiguo. Alusivo, porque las palabras pueden incorporar referencias a múltiples niveles. Las personas, al leer o escuchar, nos hemos acostumbrado a detectar indicios y pistas. Pero los ordenadores no tienen ese sexto sentido. La ambigüedad es, quizás, un problema aun mayor. Muchas frases, hasta el 40% en ciertos tipos de textos, pueden resultar ambiguas para un ordenador, incluso aunque tengan sentido para un traductor humano, porque este tiene su conocimiento “extratextual”.

Esta "dificultad" inherente al lenguaje se convierte en un obstáculo cuando intentamos definir un campo de estudio en el que este juega un papel central. En palabras de H. Cunningham: "Attempting to define something as dynamic and multi-faceted as a research field concerned with human language is a difficult task; given ten researchers there will likely be ten definitions, all using similar terminology" (1999:1).

Este problema se complica cuando entran en juego dos o más lenguas, cuanto más alejadas, más acentuadas las diferencias y por tanto las formas de mirar el mundo; cuando hay que conocer los puntos de vista, los conocimientos, las técnicas y estrategias que desde otras disciplinas se han propuesto, ya que "la lengua no solo ha sido objeto de interés para los lingüistas, junto a ellos están los filósofos, psicolingüistas e ingenieros" (Moreno *et al.* 1999:1); cuando desde la propia ciencia del lenguaje no existe acuerdo sobre unidades de trabajo, teorías explicativas de los fenómenos lingüísticos; o, simplemente, cuando nuestras limitaciones humanas no nos permiten llegar a comprender en última instancia cómo funciona nuestro cerebro y, por tanto, cómo se procesa el lenguaje cuando hablamos, cuando escuchamos, cuando escribimos o leemos.

No obstante, pese a las dificultades y desconocimiento de los que partíamos, nos pareció atractivo por lo menos esbozar una pequeña y modesta aproximación a lo que podía dar de sí este nuevo campo de trabajo, intentar estructurar, en la medida de nuestras posibilidades y disponibilidad de tiempo, los contenidos surgidos de las investigaciones en forma de dos cursos[2] y, finalmente, compartir a

[2] Distribuidos en las asignaturas de la licenciatura de Lingüística "Lingüística Computacional" y "Lingüística Computacional II" impartidas en la Universidad de León.

través de Internet[3] esas conclusiones parciales a las que pudiéramos llegar, por si resultaran de interés para alguien más[4].

Si se han logrado o no las metas propuestas, no me corresponde a mí juzgarlo, aunque en el apartado final de "Conclusiones" expondré las ventajas e inconvenientes encontrados durante el desarrollo del presente proyecto, apoyándome en la propia experiencia y en las aportaciones, sugerencias y comentarios de los diferentes usuarios.

[3] URL: http://www3.unileon.es/dp/dfh/Milka/Milka.htm, en concreto: http://www3.unileon.es/dp/dfh/Milka/LC.htm y http://www3.unileon.es/dp/dfh/Milka/LCII.htm

[4] Se incluyen como anexos algunos de los materiales utilizados, a modo de ejemplo.

1. QUÉ ES LA LINGÜÍSTICA COMPUTACIONAL

1. QUÉ ES LA LINGÜÍSTICA COMPUTACIONAL

De acuerdo con las consideraciones previas y con vistas a la meta que nos proponemos, parece de rigor, cuando nos acercamos a un campo que desconocemos, saber en qué consiste el mismo, qué rasgos lo caracterizan, cuál es su objeto de estudio, qué objetivos se plantea alcanzar y qué métodos de trabajo emplea para lograrlo. Una ciencia no es tal si no delimita previamente estas cuestiones y otras conexas[5].

La respuesta a estas preguntas nos permitirá efectuar una primera aproximación al concepto de Lingüística Computacional y a otros relacionados con el fin de intentar una acotación inicial de la materia que nos ocupa.

El punto de partida para llevar a cabo esta toma de contacto lo constituye una serie de definiciones del término "lingüística computacional" extraídas de diccionarios generales y especializados, manuales, artículos sobre el tema, etc. De su consulta hemos colegido algunas reflexiones sobre la LC que pasamos a comentar a continuación.

---

[5] *Vid.* MOURE (2002), quien destaca, entre los rasgos que deben estar presentes en toda investigación que aspire a ser considerada ciencia, los siguientes (*ibid.*:15):

- Contar con el aval de una tradición de estudios e investigadores.
- Delimitar un objeto real que sea descriptible mediante una serie de leyes.
- Adecuar las ideas a los hechos, es decir, ser verdad.
- Buscar conocimiento, ser crítica.
- Emplear un método que trabaje con hipótesis y datos cuya verificación o falsación permita el progreso.
- Tener como objetivo la sistematización de los conocimientos y la formulación de leyes, no la obtención de productos.

A lo largo de nuestro trabajo y, en especial, en el apartado dedicado a las conclusiones, veremos en qué medida la LC cumple estos requisitos y puede, por tanto, ser calificada como ciencia o, por el contrario, no los satisface y es más apropiado, en consecuencia, otorgarle el estatus de tecnología.

1.1. El estatus de la LC en el conjunto de las ciencias

Lo primero que observamos es que no existe unanimidad a la hora de ubicar la LC en el conjunto de las ciencias. Así, junto a definiciones que la sitúan claramente en el terreno de la Lingüística en general[6] y de la Lingüística Teórica[7] o de la Lingüística Aplicada[8] en particular, otras se decantan por el de la Informática, en especial por el de una de sus subdisciplinas, la Inteligencia Artificial[9], o por otra de sus áreas, el

[6] Así lo hace, p. ej., D. CRYSTAL en su *Diccionario de lingüística y fonética* (2000 [1980]:345):

> Lingüística computacional: Rama de la lingüística en la que se emplean técnicas y conceptos computacionales para la elucidación de problemas lingüísticos y fonéticos. Se han desarrollado varias áreas de investigación entre las que se incluyen el procesamiento del lenguaje natural, la síntesis del habla, el reconocimiento del habla, la traducción automática, la creación de concordancias, la evaluación de las gramáticas y muchas otras áreas en las que se requieren cálculos y análisis estadísticos (p. ej. en los estudios de textos literarios).

[7] J. GÓMEZ GUINOVART (1998:135), en uno de los varios trabajos que ha dedicado a la delimitación de la LC, "Fundamentos de Lingüística Computacional: bases teóricas, líneas de investigación y aplicaciones", estima que "desde el punto de vista de su vinculación a la lingüística, la lingüística computacional puede ser considerada una subdisciplina de la *lingüística teórica*, en tanto que uno de sus objetivos es la elaboración de modelos formales (e implementables informáticamente) del lenguaje humano".

[8] Es la postura que encontramos en el *Encyclopedic Dictionary of Applied Linguistics* de K. JOHNSON y H. JOHNSON (1998:81-82):

> *Computational linguistics can be seen as a branch of applied linguistics, dealing with computer processing of human language. Automatic translation between natural languages, text processing and communication between people and computers are among its central concerns. Speech recognition and understanding and speech synthesis allow people to communicate with computers using spoken language. Computational grammars with top-down and bottom-up processing capabilities have been developed in this connection. Computer-assisted language learning programmes are among numerous applications of the new technology. Computerized corpora of written and spoken texts facilitate research on usage using concordances.*

[9] Es el caso de P.-K. HALVORSEN (1991 [1988]:252) en el apartado dedicado a "Las aplicaciones informáticas de la teoría lingüística" en el vol. II del *Panorama de la Lingüística Moderna de la Universidad de Cambridge*, compilado por F. J. NEWMEYER:

Procesamiento del Lenguaje Natural[10]; sin faltar aquellas que optan por el espacio conformado por la intersección de las anteriores[11], es decir,

---

> La lingüística computacional está considerada como una rama de la inteligencia artificial (IA). Como todos los campos dentro de la IA, se ocupa de la investigación y sistematización de una capacidad cognitiva. En el caso de la lingüística computacional, el objetivo central es la capacidad lingüística. Sin embargo, su preocupación no es necesariamente construir un modelo *psicológicamente realista* del comportamiento lingüístico humano. Su objetivo es identificar y caracterizar las clases de procesos y los tipos de conocimiento que están implicados en la habilidad de comunicar y asimilar información por medio del lenguaje natural, sin tomar en consideración su *status* psicológico. Una de las contribuciones de la lingüística computacional consiste en un conjunto de técnicas que capacitan al conocimiento lingüístico para guiar y constreñir el procesamiento lingüístico realizado por un sistema de procesamiento del lenguaje natural.

O también de J. VIDAL y J. BUSQUETS (1996:393-394) en el capítulo dedicado a la "Lingüística computacional" dentro del manual *Elementos de lingüística* editado por C. MARTÍN VIDE:

> La LC es una rama de la inteligencia artificial (en adelante IA). Si bien las opiniones entre los especialistas divergen, se asume que el principal objetivo de la LC es la investigación y sistematización de la capacidad lingüística entendida como una capacidad cognitiva fundamental. Sucintamente, la LC se orienta hacia el estudio del conocimiento lingüístico obtenido a partir de la aplicación de un conjunto de formalismos y técnicas de representación. Con ello se pretende el procesamiento del LN [lenguaje natural] mediante un ordenador. […] No se trata de elaborar modelos que posean realidad psicológica, sino más bien de construir modelos que simulen los tipos de conocimiento y los procesos que intervienen en la habilidad de transmitir e interpretar información a través del LN. En otras palabras, simular un conocimiento inteligente. Desde este punto de vista, se atribuye una cierta 'racionalidad' a la computadora, aunque es, por supuesto, estrecha y artificial.

J. GÓMEZ GUINOVART (1998:135), por su parte, en el mismo artículo mencionado con anterioridad (*vid. supra*), considera que "desde el punto de vista de su vinculación a la informática, y también por motivos históricos, la lingüística computacional suele ser considerada como una subdisciplina de la *inteligencia artificial*".

En este sentido, la propia REAL ACADEMIA ESPAÑOLA (*DRAE*-01), en un artículo recientemente modificado y como avance de la 23ª edición de su diccionario, ha incluido el término "lingüística computacional" dentro de la entrada "lingüística" y, a la hora de definirlo, parece inclinarse por su vinculación a la Informática y la Inteligencia Artificial: "1. f. *Inform*. Aplicación de los métodos de la inteligencia artificial al tratamiento de cuestiones lingüísticas".

[10] Es el caso de la definición que proporciona W. ARMS (2000 [1999] en su libro *Digital Libraries*: "Computational linguistics: The branch of natural language processing that deals with grammar and linguistics".

[11] Así la define J. KLAVANS (1997:665) en el capítulo dedicado a la LC en el manual de Lingüística editado por W. O'GRADY, M. DOBROVOLSKY y F. KATAMBA: "Computational linguistics is a relatively new discipline that lies in the intersection of the fields of linguistics and computer science. It is but one of many new hybrid

de la Lingüística y de la Informática, o por un ámbito interdisciplinario o multidisciplinar, en conexión con diversos campos científicos, sobre todo con la Ciencia Cognitiva[12], de la que algunos autores, desde esta perspectiva más general, la hacen depender en última instancia.

---

disciplines involving computers that require computational expertise as well as a background in another field".

[12] Es lo que hacen Mª A. MARTÍ e I. CASTELLÓN (2000) en la "Introducción" a su manual *Lingüística Computacional*:

> *La Lingüística Computacional és una àrea de coneixement interdisciplinari on conflueixen la Lingüística Teòrica i Aplicada, la Informàtica, la Intel·ligència Artificial i la Ciència Cognitiva. Encara que amb aquest terme es fa referència sovint a tota mena de processos informàtics que s'apliquen sobre dades lingüístiques, l'objectiu últim de la Lingüística Computacional és la modelització del comportament lingüístic del parlant i de l'oient, és a dir la construcció de programes informàtics que simulin els processos que tenen lloc en els individus quan ens comuniquem.*

O más adelante (MARTÍ y CASTELLÓN *ibid.*:1), cuando dicen: "La Lingüística Computacional (a partir d'ara LC) és una nova disciplina que ha sorgit de la col·laboració entre la Lingüística, la Informàtica i altres àrees de coneixement com la Intel·ligència Artificial (des d'ara IA) i la Ciència Cognitiva".

Las definiciones que encontramos en la enciclopedia de contenido libre WIKIPEDIA inciden en el carácter inter y multidisciplinar de la LC:

> La lingüística computacional es un campo multidisciplinar de la lingüística y la informática que utiliza la informática para estudiar y tratar el lenguaje humano. Para lograrlo, intenta modelar de forma lógica el lenguaje natural desde un punto de vista computacional. Dicho modelado no se centra en ninguna de las áreas de la lingüística en particular, sino que es un campo interdisciplinar, en el que participan lingüistas, informáticos especializados en inteligencia artificial, psicólogos cognoscitivos y expertos en lógica, entre otros [URL: http://es.wikipedia.org/wiki/Portada].

> *Computational linguistics is an interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective. This modeling is not limited to any particular field of linguistics. Traditionally, computational linguistics was usually performed by computer scientists who had specialized in the application of computers to the processing of a natural language. Recent research has shown that human language is much more complex than previously thought, so computational linguists often work as members of interdisciplinary teams, including linguists (specifically trained in linguistics), language experts (persons with some level of ability in the languages relevant to a given project), and computer scientists. Computational linguistics draws upon the involvement of linguists, computer scientists, experts in artificial intelligence, cognitive psychologists, mathematicians, and logicians, amongst others* [URL: http://en.wikipedia.org/wiki/Computational_linguistics].

Como se puede inferir de las definiciones que hasta ahora hemos enumerado, no es descabellado concluir que la LC, en su afán por comprender el lenguaje y simularlo en una computadora –objetivo más ambicioso al que aspira–, aúna los intereses de la Lingüística y de la Informática. Es decir, necesita los conocimientos que ambas le suministran, pero no solo estos. Si bien estas dos disciplinas son los pilares sobre los que se sustenta, es importante destacar el hecho de que la LC se mueve en un marco interdisciplinar, por lo que también acude a la Psicología, las Matemáticas, la Lógica, la Ciencia Cognitiva, etc. en

---

H. USZKOREIT (1996, 2000), en un texto introductorio, sitúa la LC en un espacio intermedio entre la Lingüística y la Informática, aunque destacando sus conexiones con la Ciencia Cognitiva y la Inteligencia Artificial:

> *Computational linguistics (CL) is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty. It belongs to the cognitive sciences and overlaps with the field of artificial intelligence (AI), a branch of computer science aiming at computational models of human cognition. Computational linguistics has applied and theoretical components.*

J. GÓMEZ GUINOVART también insiste en varias ocasiones –en el artículo ya referido (*vid. supra*) y en el capítulo sobre "Lingüística computacional" del *Manual de Ciencias da Linguaxe* editado por F. RAMALLO, G. REI-DOVAL y X. P. RODRÍGUEZ YÁÑEZ– en la interdisciplinariedad como rasgo característico de la LC:

> La lingüística computacional (o lingüística informática) es un campo científico interdisciplinar relativamente reciente –cerca de cincuenta años de investigación y desarrollo– cuyo objetivo radica en incorporar en los ordenadores la habilidad en el manejo del lenguaje humano (GÓMEZ GUINOVART 1998:135).
>
> *A* lingüística computacional *constitúe un eido científico interdisciplinario vinculado á lingüística e á informática, e encamiñado a incorporar nos ordenadores a habilidade no manexo da linguaxe natural humana e a facilita-lo tratamento informatizado das linguas e do seu estudio* (GÓMEZ GUINOVART 2000a:1).

J. LAVID (2005:73), en su libro *Lenguaje y nuevas tecnologías*, se expresa en términos parecidos: "La Lingüística Computacional es un área interdisciplinaria entre la Lingüística y la Informática que se ocupa de la construcción de sistemas informáticos capaces de procesar el lenguaje humano". Y, más adelante (*ibid.*:76):

> […] la Lingüística Computacional es un área interdisciplinaria que se crea y desarrolla gracias a las contribuciones de diferentes disciplinas. En este sentido, la LC forma parte de las Ciencias Cognitivas y se solapa en sus objetivos con los del campo de la Inteligencia Artificial (IA), una rama de la Informática cuyo objetivo es la simulación de modelos computacionales de la cognición humana.

busca de soluciones alternativas. Sin embargo, existen matices que la diferencian de todas ellas y que, por lo tanto, justifican su estatus independiente como saber científico.

De forma esquemática, podríamos resumir de la siguiente manera las distintas posturas evidenciadas en las definiciones que hemos tomado como punto de partida:

1.1.1. La LC como parte de la Lingüística

Desde el momento en que el lenguaje[13] aparece implicado de una o de otra manera en el quehacer de la LC, la vinculación de esta con la Lingüística es incuestionable. Ambas disciplinas se ocupan de investigar los mecanismos que posibilitan la comunicación entre las personas por medio del lenguaje, aunque en el caso de la LC con la ayuda que le proporcionan los ordenadores, que es su rasgo característico[14]. Resulta evidente, así pues, que el lenguaje, como capacidad general del ser humano, y las lenguas naturales, como producto de dicha capacidad, son el objeto de estudio de la LC. Por lo tanto, comparte con la Lingüística el interés por descubrir y describir

[13] En forma de "elucidación de problemas lingüísticos y fonéticos" (*vid. supra* CRYSTAL 2000 [1980]), "elaboración de modelos formales del lenguaje humano" (*vid. supra* GÓMEZ GUINOVART 1998), "*computer processing of human language*" (*vid. supra* JOHNSON Y JOHNSON 1998), "procesos y los tipos de conocimiento que están implicados en la habilidad de comunicar y asimilar información por medio del lenguaje natural" (*vid. supra* HALVORSEN 1991 [1988]), "investigación y sistematización de la capacidad lingüística entendida como una capacidad cognitiva fundamental" (*vid. supra* VIDAL Y BUSQUETS 1996), "aplicación de los métodos de la inteligencia artificial al tratamiento de cuestiones lingüísticas" (*vid. supra* RAE 2001), "modelització del comportament lingüístic del parlant i de l'oient" (*vid. supra* MARTÍ Y CASTELLÓN 2000), "modeling of natural language from a computational perspective" (*vid. supra* WIKIPEDIA), "construcción de sistemas informáticos capaces de procesar el lenguaje humano" (*vid. supra* LAVID 2005), etc.

[14] El simple hecho de emplear medios informáticos en la investigación lingüística no convierte esta en LC, como veremos más adelante (*vid.* 1.4. Principales líneas de investigación), pero de momento basta para establecer la distinción entre Lingüística y LC.

cómo funciona el lenguaje, cómo podemos comunicarnos las personas a través de él, qué elementos y procesos intervienen cuando actuamos como emisores y cuáles cuando lo hacemos como receptores... Y difiere de la Lingüística, entre otros aspectos, en las herramientas que emplea para llevar a cabo sus investigaciones.

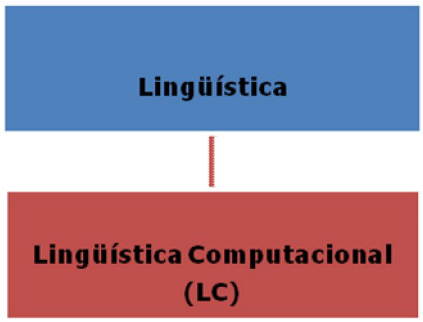


Ilustración 1. La LC como parte de la Lingüística.

Ahora bien, la vinculación puede establecerse, además de con la Lingüística directamente, en tanto que ciencia del lenguaje, a través de la Lingüística Teórica o de la Lingüística Aplicada. En el primero de los casos, se pone el énfasis en la vertiente más "científica" de la LC, aquella que tiene como objetivo elaborar modelos formales sobre el lenguaje o probar y evaluar las teorías que le suministra la Lingüística Teórica o que ella misma diseña, sin importar de forma inmediata las repercusiones prácticas que se puedan derivar. En el segundo caso, el peso recae sobre la orientación más "tecnológica" de la LC, aquella cuya finalidad es, sin menospreciar la fundamentación teórica, desarrollar aplicaciones concretas en las que el lenguaje humano desempeñe un papel central y que tengan trascendencia en la sociedad, tales como la traducción automática, el procesamiento de información textual, la comunicación entre personas y ordenadores, la síntesis y el

reconocimiento del habla, el aprendizaje/enseñanza de lenguas asistido por ordenador, etc., entre las más destacadas.

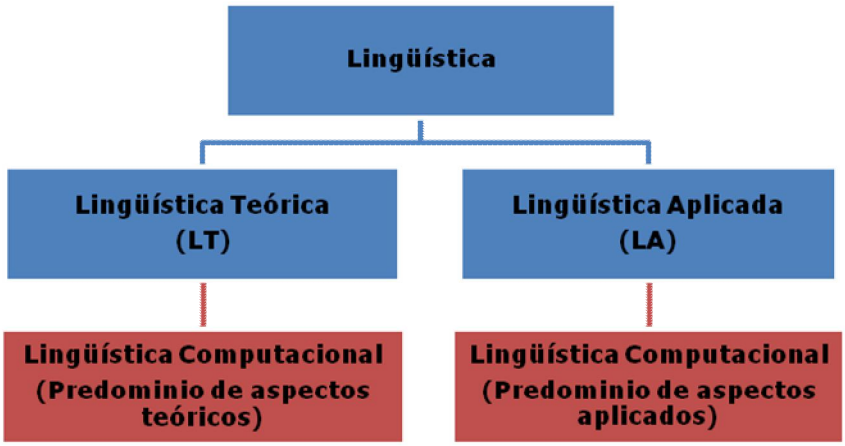


Ilustración 2. La LC como rama de la LT y de la LA.

1.1.2. La LC como rama de la Informática

Si el lenguaje es el vínculo de la LC con la Lingüística, el empleo de los ordenadores como herramienta fundamental de trabajo conecta la LC con la Informática. No se trata solo de estudiar el lenguaje y las lenguas, sino de hacerlo con la ayuda que suponen hoy en día los ordenadores[15]. Además, hay que tener en cuenta que la LC, en su empeño por dominar el lenguaje, no se limita a la mera indagación –labor que por sí sola no la diferenciaría de la Lingüística–, sino que comprende una meta más

[15] USZKOREIT (1996, 2000), más que de ayuda, habla de necesidad. En su opinión, la complejidad que están alcanzando en la actualidad los formalismos gramaticales exige el recurso a programas informáticos para poder manipularlos adecuadamente:

> *Theoretical CL takes up issues in theoretical linguistics and cognitive science. It deals with formal theories about the linguistic knowledge that a human needs for generating and understanding language. Today these theories have reached a degree of complexity that can only be managed by employing computers. Computational linguists develop formal models simulating aspects of the human language faculty and implement them as computer programmes. These programmes constitute the basis for the evaluation and further development of the theories.*

ambiciosa: reproducir una capacidad cognitiva, la lingüística en este caso, en programas informáticos, con algún fin práctico.

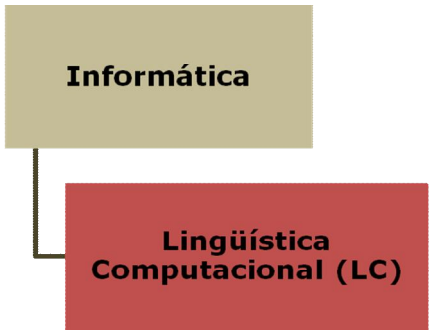


Ilustración 3. La LC como parte de la Informática.

Para alcanzar esta pretensión también le resultan imprescindibles las aportaciones de la Informática, en especial de una de sus subdisciplinas, la Inteligencia Artificial (en adelante IA), que precisamente estudia todas las conductas inteligentes del ser humano, entre las que ocupa un lugar destacado el lenguaje. De esta forma, tanto la Informática como la IA proporcionan a la LC técnicas, estrategias, formalismos de representación y otras herramientas que puedan contribuir, desde una orientación eminentemente aplicada, a ese objetivo de lograr ordenadores capaces de "hablar".

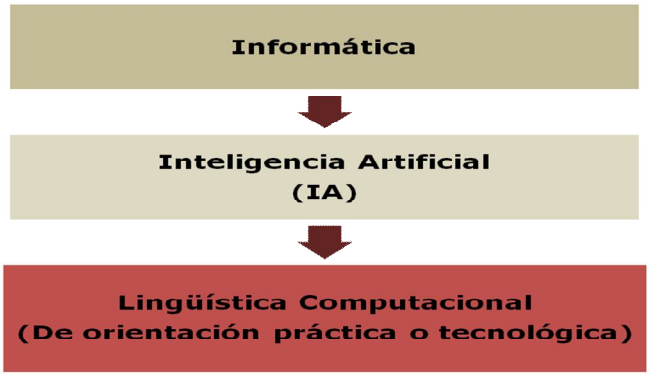


Ilustración 4. La LC como rama de la IA.

Desde la perspectiva de la Informática y de forma paralela a como sucede en el caso de la Lingüística, hay que destacar también que no es extraño hacer depender la LC, no directamente de la Inteligencia Artificial, sino de la parte de la IA que se ocupa específicamente del lenguaje humano, el Procesamiento del Lenguaje Natural (en adelante PLN), subdisciplina que se caracteriza en general por presentar una orientación práctica y por estar centrada en el tratamiento de la lengua escrita.

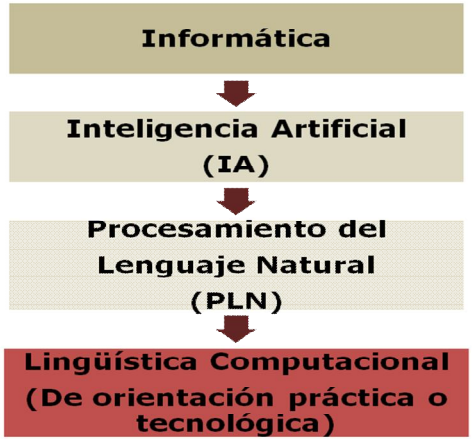


Ilustración 5. La LC como PLN.

Además, como comentaremos más adelante, en muchos casos la LC se llega a identificar con el PLN, hasta el punto de que ambos términos se toman como sinónimos, ya que, por lo demás, sus intereses coinciden plenamente: el primero, más habitual en el ámbito de la Lingüística y el segundo, en el de la Informática. Asimismo, la posición de la LC respecto a la Inteligencia Artificial puede interpretarse también como un solapamiento o coincidencia de objetivos, más que como una dependencia[16].

[16] *Vid. supra* definiciones de USZKOREIT (1996, 2000) y LAVID (2005).

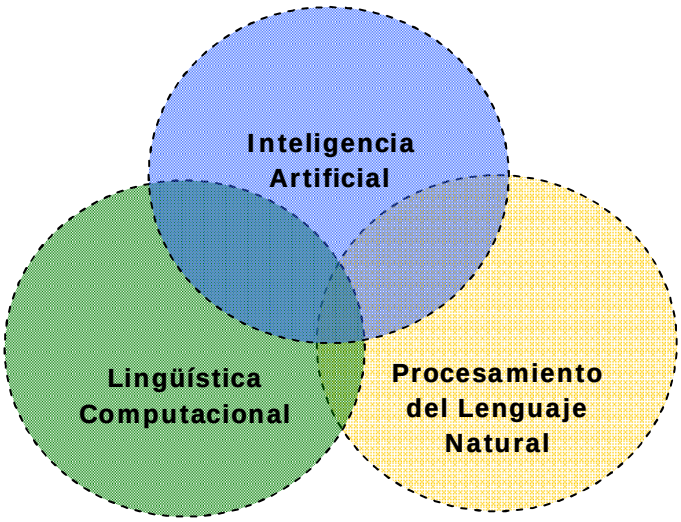


Ilustración 6. La relación de la LC con el PLN y la IA.

1.1.3. La LC entre la Lingüística y la Informática

Como se desprende de lo dicho hasta aquí, ni la Lingüística ni la Informática por sí solas permiten definir la LC, sino que es la suma de las aportaciones de ambas la que ha propiciado el surgir de este nuevo ámbito del saber. Además, no se puede pasar por alto que esta oscilación entre Lingüística e Informática se manifiesta ya en la propia denominación del campo, dado que tanto el lenguaje ("Lingüística") como los ordenadores ("Computacional") están implicados en él, por lo que su intersección parece conformar el ámbito específico de trabajo de la LC.

Así pues, Lingüística e Informática se suman e integran en un enfoque mixto que sirve como punto de partida a la LC en su pretensión de emular en un programa informático la capacidad lingüística humana en su totalidad en tanto que capacidad cognitiva

básica[17]. En aras de lograr esta meta, la LC centra gran parte de sus esfuerzos en la elaboración de formalismos o modelos formales, en lo que sería un acercamiento básicamente teórico y, por tanto, lingüístico. Pero no basta con disponer de descripciones del lenguaje en sus diferentes niveles (fonológico, morfológico, sintáctico, etc.), sino que estas, además, han de ajustarse a los requisitos que imponen las herramientas de trabajo que emplea la LC, los ordenadores, ya que de otra manera dichas descripciones no podrían ser implementadas en un programa informático, de ahí la necesidad de que la descripción lingüística sea "formal", es decir, esté formulada mediante reglas claras, precisas y despojadas de toda ambigüedad, de manera similar a como sucede en Matemáticas o en Lógica.

Por otra parte, la LC, a la hora de abordar su objeto de estudio, no siempre lo hace desde una perspectiva teórica o científica, sino que también está interesada en buscar la aplicación de esos conocimientos al logro de productos finales que tengan una finalidad práctica concreta, lo que entronca con las motivaciones de la Informática y la Inteligencia Artificial en torno al lenguaje. No obstante, ambas partes, teoría y práctica, Lingüística e Informática, son necesarias por igual en el camino hacia el objetivo último que comparten lingüistas e informáticos a propósito del lenguaje, que no es otro que comprender su funcionamiento, aunque en el caso concreto de la LC, como paso previo para desarrollar ordenadores capaces de utilizar el lenguaje igual que las personas.

[17] Aunque, a veces, este ambicioso objetivo se ve limitado a abordar aspectos parciales del lenguaje –p. ej., la morfología– en función de la aplicación concreta que se quiera dar a las investigaciones (como incorporar un conjugador verbal en un diccionario en línea), ya que en muchos casos las aplicaciones no requieren un tratamiento integral, sino solo considerar un fenómeno lingüístico determinado (en el ejemplo, la conjugación verbal).

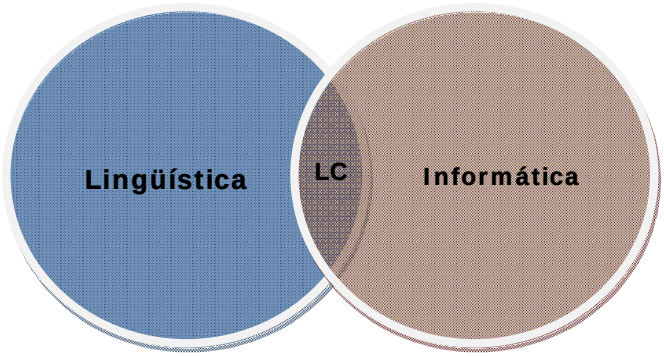


Ilustración 7. La LC entre la Lingüística y la Informática.

En palabras de R. Hausser (2001:1):

> *The goal of computational linguistics is to reproduce the natural transmission of information by modeling the speaker's production and the hearer's interpretation on a suitable type of computer. This amounts to the construction of autonomous cognitive machines (robots) which can communicate freely in natural language.*

1.1.4. La LC en conexión con otras ciencias

Por último, para finalizar este apartado, nos queda por señalar que, precisamente, este carácter "híbrido" que acabamos de comentar, sumado a la complejidad inherente al lenguaje, que se concibe como una parte fundamental del sistema cognitivo humano, es el que, para algunos autores (*cf.* p. ej. Uszkoreit 1996, 2000), sitúa la LC en un marco de confluencia más amplio que el de la mera intersección de Lingüística e Informática: el que le proporciona en la actualidad la Ciencia Cognitiva[18] que, tomando como punto de referencia el objetivo común

[18] O Ciencias Cognitivas, dada la multitud de disciplinas que abarca la etiqueta, como podemos observar en la siguiente definición tomada de WIKIPEDIA:

de estudiar la mente humana[19], aglutina áreas del saber tan diversas como la Lingüística, la Psicología, la Neurociencia, la Antropología, la Filosofía y la propia Inteligencia Artificial, entre otras.

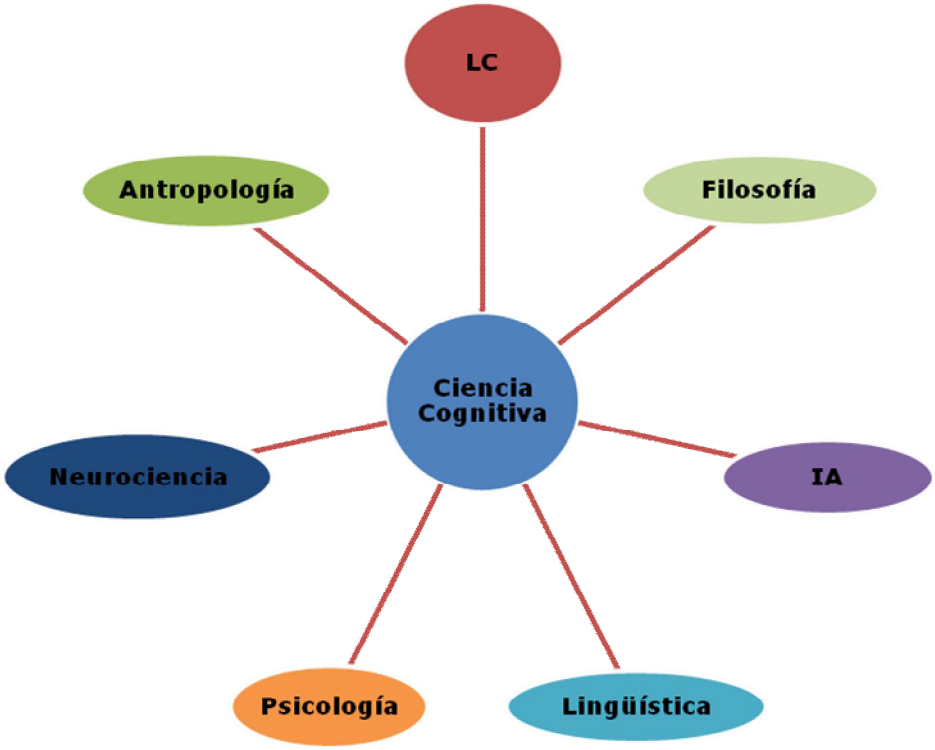


Ilustración 8. La LC en conexión con otras disciplinas.

Se denomina ciencia cognitiva al estudio científico de la mente humana. Su enfoque y su área de investigación es marcadamente multidisciplinar, fruto de la confluencia entre la lingüística, la psicología cognitiva, la neurociencia, la filosofía (en particular la filosofía de la ciencia y la filosofía de la mente) y la inteligencia artificial, por todo lo cual a menudo se designa en plural como *ciencias cognitivas*.

[19] Así lo proclama la "Cognitive Science Society" en su presentación:

> *The Cognitive Science Society, Inc. brings together researchers from many fields who hold a common goal: understanding the nature of the human mind. The Society promotes scientific interchange among researchers in disciplines comprising the field of Cognitive Science, including Artificial Intelligence, Linguistics, Anthropology, Psychology, Neuroscience, Philosophy, and Education.*
>
> [URL: http://cognitivesciencesociety.org/index.html]

Por otro lado, como veremos más adelante, Ciencia Cognitiva y LC comparten, además de intereses, unas mismas raíces.

En concreto, el punto de unión fundamental entre LC y Ciencia Cognitiva lo constituye la metáfora del cerebro como un ordenador capaz de manipular símbolos y de ejecutar complejos procesos basados en el conocimiento almacenado en su interior[20]. De hecho, esta imagen ha sido adoptada como uno de los axiomas centrales de la Ciencia Cognitiva en su intento por caracterizar la inteligencia humana. Dada la imposibilidad de acceder directamente a ella para su estudio empírico, recurre a los ordenadores como herramienta de experimentación que le permite su modelado mediante el diseño de programas informáticos y bases de conocimiento a medida, según las especificaciones de los investigadores. De esta forma, la Ciencia Cognitiva pretende encontrar una explicación para los procesos responsables de toda conducta inteligente y, en especial, la lingüística, lo que conlleva explorar y entender los tipos de conocimiento que subyacen a dicha conducta así como la forma en que están organizados en la mente. Al estar involucrado el cerebro, las aportaciones de la Neurociencia y de la Psicología resultan imprescindibles para comprender su funcionamiento y estructura, pero también las de la Psicolingüística por lo que concierne específicamente al lenguaje, o de disciplinas como la Antropología o la Filosofía, que aportan reflexiones generales de interés sobre los debates éticos que tales cuestiones suscitan.

En este contexto, hay que destacar que el bagaje que la Lingüística proporciona a la LC resulta insuficiente o inadecuado en muchos casos, de ahí la necesidad de acudir a los conceptos, métodos, etc. desarrollados en otros ámbitos de trabajo que de una u otra manera tocan el lenguaje en alguna de sus facetas. Entre ellos sobresale la Inteligencia Artificial, área en la que todo lo concerniente a la organización del conocimiento es un tema central de investigación, de

[20] *Cf.* WINOGRAD (1983), en su obra *Language as a Cognitive Process*.

ahí las estrechas conexiones, o solapamientos[21], que se establecen entre ambas: a la Inteligencia Artificial le interesa la interacción del conocimiento propiamente lingüístico con el conocimiento de naturaleza más general, compartido con otras tareas cognitivas.

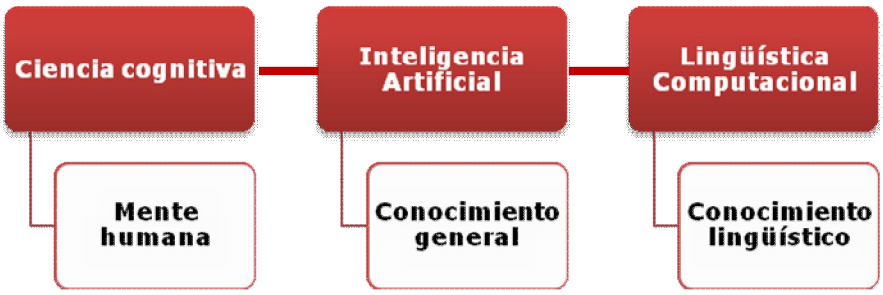


Ilustración 9. La LC como Ciencia Cognitiva.

Así pues, este terreno multidisciplinar en el que se mueve la LC se deriva directamente de la envergadura de la tarea a la que se enfrenta e implica un gran número y variedad de conocimientos, lo que justifica el mencionado carácter interdisciplinar de la LC, destacado en las diferentes definiciones. Además de las aportaciones de la Lingüística y de la Informática, que serían las que le proporcionan las bases sobre las que se cimienta, la LC se beneficia de los descubrimientos procedentes de múltiples áreas, desde la Lógica a la Psicología, pasando por las Matemáticas, la Psicolingüística, la Inteligencia Artificial, etc. en el marco general de lo que se viene denominando Ciencia(s) Cognitiva(s). Es más, según A. K. Joshi (2002 [1999]:745), "al poner en contacto los campos de la ciencia computacional y la lingüística –que están íntimamente relacionados–, la LC desempeña un papel central en la ciencia cognitiva".

[21] *Cf.* USZKOREIT (1996, 2000) y LAVID (2005).

Por lo tanto, la situación que nos encontramos es que el estudio de los contenidos a los que se refiere la etiqueta “Lingüística Computacional” es posible efectuarlo desde diferentes perspectivas, que resumiremos en dos: la de la Lingüística y la de la Informática, con la Ciencia Cognitiva como telón de fondo. Como es lógico, cada una lo aborda con presupuestos teóricos, objetivos, métodos y herramientas muy distintos.

1.2. Objetivos de la LC: LC Teórica y LC Aplicada

Como se ha podido observar en el apartado anterior, la LC aglutina los intereses de la Lingüística y de la Informática, hecho patente en la propia denominación del campo, dado que tanto el lenguaje (Lingüística) como los ordenadores (Computacional) están implicados en ella. Para algunos autores, este hecho sitúa la LC en un marco de intersección más amplio que el de la propia Lingüística o la Informática, el que le proporciona la Ciencia Cognitiva.

No obstante, puesto que tanto el lenguaje como los ordenadores están implicados en la definición de LC, se suelen distinguir dos posibles acercamientos a este ámbito de confluencia:

1) Acercamiento de la Lingüística
2) Acercamiento de la Informática

Desde la *perspectiva de la Lingüística*, se considera que:

- La LC es una rama de la Lingüística Aplicada.
- Su objeto es el estudio del lenguaje en sus diferentes niveles y procesos (fonética, morfología, sintaxis, semántica, pragmática, etc.) desde una perspectiva global que los integre, ya que su objetivo último es emular la conducta lingüística en su totalidad.
- Lo que la diferencia de la Lingüística Teórica es que la LC se sirve de formalismos y técnicas computacionales.

Desde la *perspectiva de la Informática*, se considera que:

- La LC es una rama de la Inteligencia Artificial.
- Su objeto es simular la conducta lingüística humana en cuanto capacidad cognitiva básica, aunque no necesariamente teniendo en cuenta su fundamentación psicológica. Es decir, lo importante es desarrollar programas informáticos capaces de "hablar": reconocer, comprender y producir enunciados, imiten o no la forma en que funciona nuestro cerebro cuando ejercitamos la capacidad lingüística.
- Se integra en un proyecto más ambicioso, el de simular la inteligencia humana en general.

Esta doble vertiente, lingüística e informática, se observa en los dos objetivos o motivaciones con los que se puede abordar el trabajo en LC, objetivos teóricos y objetivos aplicados, que han dado lugar a que se establezca una distinción paralela entre:

1) LC Teórica, más vinculada a la Lingüística.
2) LC Aplicada, más relacionada con la Informática y la IA.

Lingüística Computacional (LC)	
LC Teórica	LC Aplicada
Objetivos teóricos	Objetivos aplicados
Perspectiva de la Lingüística	Perspectiva de la Informática

Tabla 1. LC Teórica vs. LC Aplicada.

Los objetivos teóricos, también llamados "científicos", son independientes de cualquier aplicación y constituyen el ámbito de trabajo de la LC Teórica.

Según R. Grishman (1991 [1986]:16-17), se concretan en:

- Probar las gramáticas que propone la Lingüística Teórica.
- Investigar los procesos psicológicos que intervienen en la producción y comprensión del lenguaje dentro del marco general de la Ciencia Cognitiva.
- Estudiar la forma de representar el conocimiento general o del mundo.

Los objetivos aplicados, también llamados "tecnológicos" o "aplicaciones orientadas a la ingeniería", tienen que ver con sistemas prácticos o programas informáticos específicos y constituyen el ámbito de trabajo de la LC Aplicada.

Según R. Grishman (1991 [1986]:15-16), las tres aplicaciones principales de la LC son:

- Traducción automática.
- Recuperación de información.
- Interfaces hombre-máquina.

Cuando nos referimos a la *Lingüística Computacional Teórica*, estamos ante lo que se entiende por LC en sentido estricto o LC por antonomasia. Esta toma sus temas de trabajo de la Lingüística Teórica y de la Ciencia Cognitiva. Las aportaciones de la Psicología Cognitiva, en especial de la Psicolingüística, también son de especial relevancia, lo que se ha traducido en el surgimiento de una nueva ciencia, la Psicolingüística Computacional. El objetivo de esta vertiente de la LC es proporcionar una explicación del funcionamiento del lenguaje en sus

diferentes niveles: fonético, morfológico, sintáctico, semántico, pragmático, etc.

Este objetivo general, según X. Gómez Guinovart (2000a:223), se concreta en:

- La elaboración de teorías o modelos lingüísticos generales que cumplan dos requisitos:
  - Ser formales.
  - Ser adecuados para su implementación en un programa informático.
- La descripción de fenómenos lingüísticos concretos en el marco de las teorías o modelos anteriores.
- La comprobación automatizada de la consistencia de una teoría lingüística.

Por su parte, la *Lingüística Computacional Aplicada* es una vertiente de la LC que posee una clara orientación tecnológica, lo que ha provocado que hoy en día con frecuencia se aluda a ella con nombres como *ingeniería lingüística* o *tecnología del lenguaje humano*. Se centra en los aspectos prácticos que se puedan derivar de la simulación de la conducta lingüística con medios informáticos y su objetivo es crear productos informáticos que incorporen algún componente en el que intervenga el lenguaje, oral o escrito. Uno de sus principales retos es mejorar la comunicación entre personas y ordenadores mediante el uso del lenguaje. En concreto, consiste en métodos, técnicas, herramientas y aplicaciones en las que el lenguaje desempeña un papel central.

Según X. Gómez Guinovart (2000a:223-224), las principales aplicaciones, que este autor agrupa en cuatro categorías, son:

- Programas para la comprensión y generación de enunciados: consulta a bases de datos, sistemas de diálogo, etc.
- Programas relacionados con las tecnologías del habla: dictado automático, conversión de texto en voz, etc.
- Herramientas para el procesamiento documental: correctores ortográficos y estilísticos, programas para la generación automática de resúmenes, sistemas de extracción y recuperación de información textual, etc.
- Herramientas para el procesamiento plurilingüe: programas para la enseñanza de lenguas asistida por ordenador o para la creación de ejercicios, programas de ayuda a la traducción, etc.

1.3. Lingüística Computacional y Lingüística: el carácter aplicado

Quisiéramos ahora detenernos un momento a reflexionar sobre las implicaciones que, desde la Lingüística, tiene la aplicación de los ordenadores al estudio del lenguaje.

Desde que a principios del siglo XX F. de Saussure, en su *Curso de lingüística general* (1916), estableciera las bases de la llamada "Lingüística científica", esta ha experimentado una evolución y un crecimiento espectaculares. No obstante, es lógico que, en esos momentos fundacionales, la recién inaugurada ciencia del lenguaje dirigiera su atención hacia sí misma[22], hacia el estudio de la estructura interna de las lenguas, defendiendo a toda costa su "inmanencia", contemplada como un aval de cientificidad.

Sin embargo, una vez consolidada esta Lingüística Teórica, pronto, en torno a los años sesenta, empezaron a surgir otros puntos de vista que desbordaban esos límites esbozados por F. de Saussure y proclamaban la necesidad de aproximaciones mixtas, que conjugaran ideas propias con otras ajenas en principio al mundo de la Lingüística. La explicación hay que buscarla en que "la realidad primera que muestran los hechos lingüísticos en sus manifestaciones es su diversidad" (Fernández 1999:21), su complejidad. Por lo tanto, cuando se pasa de contemplar el objeto de estudio como una abstracción inmutable a considerarlo algo concreto y dinámico, sujeto a diferentes tipos de variaciones, el lenguaje revela su carácter polifacético y

[22] Es la denominada "Microlingüística", "Lingüística interna" o "Lingüística del código", que se ocuparía de la Fonética y Fonología, la Morfología, la Sintaxis y la Semántica, disciplinas que constituirían el "núcleo de la Lingüística" (*cf.* ROJO 1986:53 y ss.) o lo que M. FERNÁNDEZ (*cf.* 1986, 1996, 1999) llama "divisiones de la Lingüística", centradas en la dimensión simbólica del lenguaje y que consideran los fenómenos lingüísticos en sí mismos, sin atender a factores externos de tipo social, cultural, biológico... Además, son las que han sido objeto de una mayor atención y cuentan con una larga tradición de estudios que las sustentan.

multidimensional y, en consecuencia, se resiste a encajar en las categorías preestablecidas y rígidas de la lingüística de corte estructural. Se pone entonces de relieve que en su caracterización no solo se debe atender a la forma y a la sustancia de los signos lingüísticos en el marco del sistema concreto de una lengua, sino que también se deben incorporar "saberes periféricos" a ese núcleo de la Lingüística (*cf.* Moure 2002:48), tales como los aspectos psicológicos subyacentes al lenguaje, los procesos neuronales implicados, su dimensión social y cultural, las consecuencias derivadas de su uso por parte de los hablantes, etc[23]. Así pues, para dar cuenta de todas estas caras de su objeto, la Lingüística se vio obligada a abrir sus puertas a otros ámbitos del saber –Psicología, Neurología, Sociología, Antropología, Filosofía del lenguaje–, como única vía para poder ofrecer descripciones y explicaciones más exhaustivas. De esta forma, se consolidan toda una serie de disciplinas lingüísticas, marcadas por la integración en sus investigaciones de perspectivas y enfoques procedentes de otros campos científicos, y también una nueva visión del quehacer lingüístico, aquella que a grandes rasgos comprende el funcionalismo (*cf.* Moure 2002:104 y ss.). Por otra parte, el término *ciencias del lenguaje*, de moda en los últimos años (*cf.* Payrató 1998:25), viene precisamente a poner el acento en el intercambio de conocimientos que se da en la actualidad entre la Lingüística y estas relativamente nuevas disciplinas.

[23] Estas otras dimensiones del lenguaje se suelen recoger bajo etiquetas como "Macrolingüística", "Lingüística externa" o, en términos de M. FERNÁNDEZ (*cf.* 1986, 1996, 1999), "ramas de la Lingüística", que incluirían los entonces nuevos campos de la Psicolingüística, la Neurolingüística, la Sociolingüística, la Antropología lingüística o la Pragmática, respectivamente. G. ROJO (1986:51 y ss.) las engloba bajo el rótulo de "disciplinas no nucleares", puesto que se ubican en un círculo más externo en relación con el central o "núcleo" (*vid.* nota anterior). Hay que señalar que G. ROJO establece diferentes círculos en torno al nuclear, según el grado de alejamiento o de vinculación que mantienen las distintas disciplinas no nucleares con la Lingüística. Las más próximas serían la Psicolingüística, la Etnolingüística y la Sociología del lenguaje, por actuar como puentes entre la Lingüística y otras ciencias de tipo cultural, mientras que la Neurolingüística, en su esquema, ocupa una posición más distanciada, por acercarse al lenguaje desde la perspectiva propia de otra ciencia.

Este hecho obedece a la imposibilidad de describir y explicar el lenguaje desde un único punto de vista: es preciso reconocer la diversidad interna del campo.

Por otra parte, casi al mismo tiempo –o incluso antes[24]– que el horizonte de los conocimientos en torno al lenguaje se ampliara de esta manera, se empezaron a vislumbrar también nuevos caminos en otra dirección, la de las aplicaciones de la Lingüística. La meta ya no es solo observar, describir y establecer generalizaciones –lo que conformaría la labor de la lingüística que ha sido calificada como "teórica", "básica" o "pura"–, sino que el lingüista debe ir más allá y aportar soluciones a las situaciones que le va planteando la sociedad, es decir, la investigación debe tener una finalidad práctica. Como dice M. Fernández (1999:34):

> […] no solo las motivaciones de curiosidad ante los hechos han empujado el crecimiento de la Lingüística, sino que el campo disciplinar –en su progresión y debido al conocimiento logrado– ha permitido delimitar nuevos problemas y ha facilitado la aproximación a circunstancias peculiares, ya no solo con objeto de describirlas sino porque ha de resolverlas. Este panorama es especialmente notable en el ámbito –ya asentado– de la Lingüística aplicada.

[24] Así, p. ej. K. JOHNSON Y H. JOHNSON (1998:9), en su *Encyclopedic Dictionary of Applied Linguistics*, retrotraen el surgimiento de la Lingüística Aplicada a finales de los cuarenta y principios de los cincuenta del pasado siglo XX en centros de Estados Unidos y del Reino Unido, aunque la fecha oficial se suele hacer corresponder con el *Coloquio Internacional de Lingüística Aplicada* celebrado en la Universidad de Nancy, Francia, en 1964, primer encuentro científico consagrado específicamente al nuevo conjunto de saberes y en el que, además, se acordó la fundación de la *Association Internationale de Linguistique Appliquée* (AILA). *Vid.* URL: http://www.aila.info/index.htm. En España, esta institucionalización debería esperar casi veinte años, hasta 1982, año en que se crea la *Asociación Española de Lingüística Aplicada* (AESLA). *Vid.* URL: http://www.aesla.uji.es/

Si bien en un principio la Lingüística Aplicada nace[25] estrechamente vinculada a la enseñanza y aprendizaje de segundas lenguas y de lenguas extranjeras, área con la que se llega a identificar, sobre todo en la tradición anglosajona, a partir de los ochenta experimenta un rápido desarrollo que da lugar a la sucesiva inclusión en su seno de nuevas materias –aunque reservando un lugar preponderante a ese núcleo inicial–, hasta el punto de constituirse en un dominio científico con entidad propia, independiente en cierta medida de la Lingüística Teórica[26] y caracterizado por la variedad de sus intereses, que giran siempre en torno a problemas reales. Surge así toda una serie de disciplinas aplicadas en Lingüística, con objetivos muy concretos en torno al lenguaje, que incluyen desde cuestiones relacionadas con el aprendizaje y enseñanza de lenguas hasta temas de normalización o de desarrollo de políticas lingüísticas, pasando por ámbitos como el de la traducción, la elaboración de diccionarios, el tratamiento automatizado de textos o las patologías lingüísticas. En palabras de T. Moure (2002:130):

> Se trata de una proyección, recientemente desarrollada, de los estudios lingüísticos y que abarca aquellas investigaciones donde los conocimientos lingüísticos se ponen al servicio de un objetivo práctico, de modo que pretenden resolver un problema material e inmediato. Los trabajos en didáctica de lenguas, teoría de la traducción, lingüística clínica, planificación lingüística y lingüística computacional entran en este dominio.

[25] En este sentido hay que tener en cuenta, como dice L. PAYRATÓ (1998:20), que "en definitiva, más que del nacimiento repentino de una disciplina (o subdisciplina, en relación con la 'gran' lingüística), deberíamos hablar de *continuum* de confluencias en tradiciones diferentes, que representan una *dimensión aplicada* de la lingüística". Este mismo autor (*ibid.*:19) alude a cómo ya en 1925, en el primer volumen de la revista *Language*, se hace mención a una dimensión aplicada de la lingüística.

[26] Esta ha contemplado con cierto desprecio y distancia todo lo relacionado con la vertiente aplicada (*cf.* MOURE Y LLISTERRI 1996:212), por lo que no hay que extrañarse del tiempo transcurrido hasta la consolidación de la Lingüística Aplicada.

Así pues, es precisamente en este marco de la Lingüística Aplicada[27] donde la LC encuentra su razón de ser en el terreno de los estudios lingüísticos[28], junto a otra serie de subdisciplinas derivadas cada una de problemas materiales específicos (*cf.* Fernández 1996:22 y ss.).

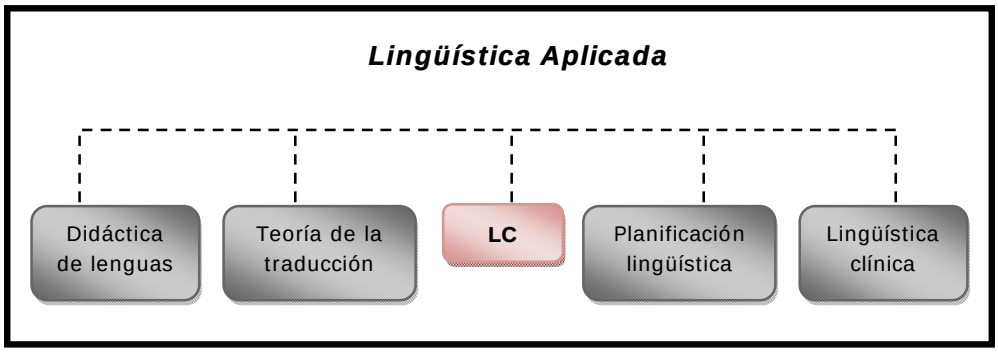


Ilustración 10. La LC en el marco de la Lingüística Aplicada.

Es más, la presencia de la LC en este nuevo ámbito disciplinario se remonta a la propia institucionalización de la Lingüística Aplicada, en 1964, ya que el encuentro celebrado en la Universidad de Nancy (*vid.* nota 24), que se toma como representativo del inicio oficial de la Lingüística Aplicada, estuvo promovido por el "Groupe de traduction automatique" de dicha universidad francesa (*cf.* Slama-Cazacu 1984:44). La traducción automática, área de investigación destacada dentro de la LC[29] que consiste en el empleo de "sistemas informáticos que llevan a cabo traducciones de una lengua a otra, con o sin intervención humana" (Hutchins y Somers 1995:27), fue uno de los campos pioneros en los que

[27] En la concepción de G. ROJO (1986), la Lingüística Aplicada –que define como "una especie de tecnología lingüística" (*ibid.*:51), ya que busca la forma de aprovechar en términos prácticos los conocimientos obtenidos del estudio de las lenguas– se sitúa entre las disciplinas no nucleares, en un círculo intermedio entre las disciplinas puente y la Neurolingüística.

[28] Lógicamente, hasta que no existió la tecnología necesaria fue impensable plantear la existencia de la LC como un área del saber distinta.

[29] *Vid.* p. ej. las definiciones ya mencionadas de D. CRYSTAL (*vid.* nota 6) y K. JOHNSON y H. JOHNSON (*vid.* nota 8).

se centraron los primeros trabajos en LC propiamente dicha (*vid.* más adelante). No obstante, tuvo que pasar un par de décadas hasta que los avances tecnológicos –y científicos– proporcionaron nuevas y mejores posibilidades para el tratamiento de las lenguas con medios informáticos, y estas posibilidades se concretaron en productos útiles para la sociedad. Así lo resume M. Fernández (1999:36):

> El despegue tecnológico de los últimos años, asociado con el propio devenir metodológico en el campo de la Lingüística, ha provocado la atención al procesamiento artificial de las lenguas, al tratamiento informático de ingentes bases de datos lingüísticos, o a los medios automáticos de traducción (ámbito de la Lingüística computacional).

Por lo tanto, la LC es una disciplina aplicada[30]. Como tal, participa de las mismas características que definen la Lingüística Aplicada y que son comunes a todas las áreas que esta comprende, aunque también se define por un objeto, una metodología y unos objetivos específicos, que le permiten erigirse en subdisciplina independiente dentro del marco de aquella.

Para T. Slama-Cazacu (1984:22 y ss.; 96 y ss.), la especificidad de la Lingüística Aplicada frente a la Lingüística Teórica reside en: i) su orientación o finalidad práctica, ii) su necesidad de una base teórica y

[30] Así de explícitamente lo proclamaba A. M. GARRIDO MORAGA (1984:213) en su artículo "La lingüística y los ordenadores. Consideraciones sobre lingüística mecanizada", cuando afirmaba que "La LM [lingüística mecanizada] hay que situarla en una de las varias direcciones de la Lingüística Aplicada y en un contexto ideológico que podemos calificar de neopositivista tal como se ha desarrollado en los últimos tiempos". Más recientemente A. MORENO SANDOVAL (1998:30), en su manual *Lingüística Computacional*, también se manifiesta con contundencia a este respecto: "La Lingüística Computacional es una disciplina aplicada. Entre sus usos principales figuran, entre otros, la traducción automática, los [*sic*] interfaces hombre-máquina, la recuperación y extracción de información y los correctores sintácticos y estilísticos".

iii) su interdisciplinariedad. Ll. Payrató (1998:24) sintetiza este programa así:

> En definitiva, pues, la lingüística aplicada puede concebirse como una orientación o dimensión de la investigación lingüística, propia de todos los campos de estudio incluidos en las ciencias del lenguaje, que, partiendo de marcos (teóricos) interdisciplinarios, persigue como objetivo la resolución de problemas (prácticos) derivados de la praxis lingüística, del uso lingüístico en que se concreta la capacidad humana del lenguaje.

1.3.1. Orientación práctica

En cuanto a la primera de las características señaladas, la Lingüística Aplicada presenta una clara orientación hacia una finalidad práctica, ya que guarda estrecha relación con situaciones concretas de la vida en las que interviene el lenguaje. Estas imponen la selección de los hechos que han de ser objeto de estudio, así como el objetivo que ha de perseguir la investigación. Por este motivo, se pueden reconocer tantas áreas dentro de la Lingüística Aplicada como problemas materiales se identifiquen, de ahí la heterogeneidad propia del campo. No obstante, esta tendencia obedece a un principio que parece gobernar la ciencia hoy en día, el de la especialización de los conocimientos en función de los nuevos aspectos de la realidad que se van descubriendo –en su mayoría, se pueden reducir a "problemas de comunicación" que surgen en la sociedad actual (*cf.* Payrató 1998:27)– y que requieren un tratamiento por parte del lingüista, del que se espera que encuentre una "solución" que redunde en el beneficio de la comunidad de que forma parte. No se trata, por tanto, de obtener nuevos conocimientos sin más meta que

lograr una mejor comprensión del lenguaje y de las lenguas, como hace la Lingüística Teórica, sino que el logro de esos conocimientos está condicionado por su aplicación: los conocimientos han de tener una finalidad práctica (*cf.* Fernández 1996:20-21).

En este sentido, "la Lingüística Computacional se encuadra en el grupo de las disciplinas aplicadas porque proyecta determinados conocimientos sobre el funcionamiento de las lenguas a la resolución de problemas concretos" (Moure y Llisterri 1996:210).

Los avances tecnológicos ocurridos durante la segunda mitad del siglo XX han marcado el nacimiento de la denominada "sociedad de la información". Al triple eje conformado por lenguaje, sociedad e información (*cf.* Llisterri 1999; Martí y Llisterri 2001), los tres pilares sobre los que se sustenta el origen mismo de las diferentes civilizaciones –pues se asocia el desarrollo de la capacidad lingüística con la aparición de la vida grupal, como un medio para transmitir los conocimientos de unos individuos a otros y de una generación a otra–, han venido a sumarse, como elemento definidor del nuevo modelo social, los ordenadores y las nuevas tecnologías asociadas a ellos, que se han erigido en una herramienta básica.

Pronto se hizo evidente la conexión que se podía establecer entre los ordenadores y las lenguas naturales, pues estas son el instrumento que solemos emplear para intercambiar con más eficacia información. Por este motivo, "no es pues extraño que, desde que los ordenadores llegaron a alcanzar un grado suficiente de complejidad, surgiera un interés en tratar el lenguaje, en tanto que portador de información, de un modo automático" (Llisterri 1999:2). Es decir, el nuevo reto que la Informática planteaba a la sociedad ha sido respondido por el surgimiento de la Lingüística Computacional, que ha proporcionado los útiles necesarios para poder manejar ingentes cantidades de

información expresada en una lengua natural. Así, disponemos de programas informáticos que nos permiten introducir, buscar, manipular, etc. información. Para R. Grishman (1991 [1986]), la traducción automática, la recuperación y extracción de información y las interfaces hombre-máquina son los campos que más esfuerzos han concentrado en LC.

Además, la faceta práctica se manifiesta en otro orden de cosas: el ahorro de tiempo y dinero, así como la eficacia que las herramientas informáticas aplicadas al lenguaje aportan en determinadas tareas.

1.3.2. Base teórica

Por otra parte, la Lingüística Aplicada no carece de unos planteamientos teóricos previos ni se limita a la mera aplicación de conocimientos sin más, a "aplicar teorías lingüísticas a un dominio práctico" (Slama-Cazacu 1984:14; Payrató 1998:18), sino que se sustenta en una base teórica propia, conformada por principios generales que se caracterizan por tomar siempre la realidad como punto de referencia. Por este motivo, no le sirve cualquier teoría, sino que para ser válida esta debe considerar la lengua como un fenómeno concreto y dinámico. Estos aspectos teóricos se conjugan perfectamente con los aplicados, ya que elabora sus propios modelos o, como mínimo, reelabora las teorías lingüísticas existentes (*cf.* Slama-Cazacu 1981:15), de acuerdo con sus necesidades, lo que, a su vez, puede revertir en la investigación teórica dentro de la propia Lingüística. El objetivo de estos constructos teóricos no es el de la investigación en sí, sino el de proporcionar una respuesta a determinadas demandas de la sociedad.

La LC no es ajena a esta preocupación. R. Grishman (1991 [1986]:16) hace referencia a la forma como combina "objetivos científicos" o teóricos con aquellos de una orientación más tecnológica o práctica. Desde esa perspectiva "científica", a los conocimientos que le aportan la Lingüística y otras disciplinas (Inteligencia Artificial, Informática, Lógica, Matemáticas, Psicología, Procesamiento de Señales, etc.), se suman formalismos y conceptos propios que la propia LC desarrolla, sobre todo en el tratamiento de la sintaxis y para la descripción del léxico, según las situaciones concretas a las que tenga que atender (*cf.* Moure y Llisterri 1996:152). De estos, en una relación de reciprocidad, se beneficia también la Lingüística Teórica, pues hay casos en los que "la Lingüística computacional está de hecho pesando en los desarrollos teóricos de la Gramática, de la Lexicología e incluso de la Pragmática, en donde las formalizaciones vuelven a estar de moda" (Fernández 1996:28). Es más, el peso de la vertiente teórica es tal que algunos autores (por ejemplo Cunninghan 1999) consideran la LC una disciplina básicamente teórica, aspecto que la diferenciaría de la línea de investigación representada por el Procesamiento del Lenguaje Natural o PLN (*vid. infra*).

A propósito de los fundamentos teóricos, con frecuencia se ha achacado a la LC que sus desarrollos científicos son más bien localistas (*cf.* Moure y Llisterri 1996:211), centrados en dominios restringidos del lenguaje o planteados ad hoc para solventar una cuestión determinada.

Como bien señalan T. Moure y J. Llisterri (1996:211), no hay que perder de vista el hecho de que:

> si la investigación en este terreno no ha avanzado con mayor rapidez no es a causa de limitaciones de tipo teórico, sino porque nuestros conocimientos sobre el lenguaje son todavía más pobres de lo que suponemos. Las aplicaciones computacionales dependen, hoy más que nunca, de una teoría lingüística que las avale y les proporcione el apoyo formal imprescindible para la gestión de sus datos.

Por otra parte, las teorías lingüísticas desarrolladas desde la LC se caracterizan por la necesidad de una formalización, requisito que no cumplen ni buscan muchas de las teorías propuestas desde la Lingüística Teórica.

1.3.3. Interdisciplinariedad

Se puede decir que esta es una constante de la Lingüística en la actualidad: participación de diferentes ciencias y saberes tanto lingüísticos como no lingüísticos en la caracterización del objeto de estudio, del lenguaje, lo que se ha traducido en un cambio en el perfil del lingüista, quien:

> ha dejado de ser un profesional especializado en un conocimiento que solo interesa a sus colegas (el modelo típico de la ciencia) para pasar a construir y elaborar teorías de amplio alcance, con repercusiones en su campo y en campos ajenos, con proyecciones aplicadas y dimensiones teóricas, con interés divulgativo y general (el modelo típico de la filosofía) (Moure 2002:90-91).

Y es que la Lingüística Aplicada y, por tanto la LC, se caracteriza por su interdisciplinariedad, ya que es un dominio científico que utiliza los conocimientos de la Lingüística pero también de otras disciplinas con las que interseccionа. Este acercamiento está motivado por la "necesidad" que existe hoy en día de contar con medios complejos para el estudio de un fenómeno también complejo como es el lenguaje. En lo que a la LC se refiere, "el conocimiento sobre el lenguaje se integra [...] junto con elementos procedentes de otras disciplinas, tanto en la teoría de la lingüística computacional como en la concepción de sistemas y herramientas y en la creación de recursos lingüísticos" (Moure y Llisterri 1996:152). Es decir, la colaboración se produce no solo en cuestiones de índole teórica, sino que también comparte técnicas, herramientas y métodos con otras ciencias:

> la integración de conocimientos procedentes de la inteligencia artificial, de la informática, de los programas cognitivos, etc. resulta fundamental para el trazado y la determinación del ámbito de la Lingüística computacional, en cuyo marco se hace en mayor medida imprescindible la consideración unitaria y conjunta de aspectos que de alguna forma son reinterpretados desde la órbita del área (Fernández 1996:26).

El resultado es que con ello:

> se está consolidando un campo de trabajo caracterizado principalmente por la interdisciplinariedad y por ofrecer la posibilidad de convertir las teorías en realidades, materializadas en productos que, en última instancia, tienen como objeto ayudar a las personas en aquellas tareas en las que el lenguaje juega un papel preponderante (Moure y Llisterri 1996:153).

Ahora bien, la LC, además de estas características comunes a todas las subdisciplinas aplicadas[31], ha de poseer algún matiz diferenciador. En su caso, la parcela de la realidad de la que se ocupa es aquella formada por:

> el conjunto de aspectos, factores, procesos, elementos, etc. que intervienen en la computación del lenguaje; de lo que se trata es de elaborar modelos y técnicas que permitan procesar las lenguas naturales en lenguaje máquina, con objeto de hacer posible no solo el reconocimiento sino también la generación y producción desde la misma computadora (Fernández 1996:25-26).

Y, en definitiva, se define por su propio objeto ("elaborar teorías y procedimientos para conseguir el tratamiento automático de las lenguas"), metodología (híbrida, "a medio camino entre informática y lingüística") y finalidad: "obtener productos tecnológicos relacionados con las industrias de la lengua" (Moure y Llisterri 1996:209).

[31] M. FERNÁNDEZ (1996:25) justifica perfectamente el carácter aplicado de la LC:

> Finalmente, también la *Lingüística computacional* y la *Planificación lingüística* son ámbitos de la *Lingüística aplicada* admitidos por su entidad sobre la base de sus objetos de estudio y sus propósitos resolutivos respecto a determinados problemas materiales. En los dos casos, además, se plantean necesidades de integración de aspectos y factores multidisciplinares; y, naturalmente, existen en ambos terrenos desarrollos teóricos evaluados por su grado de aplicabilidad y alcance sobre los problemas materiales.

1.4. Principales líneas de investigación

Como detallaremos a continuación y como ya hemos apuntado en apartados previos, el campo de la LC, desde sus inicios, ha contado con numerosas vertientes o líneas de investigación, teóricas y aplicadas, "que tienen en común la integración de conceptos y procedimientos informáticos en el tratamiento del lenguaje y del habla. De aquí el término *Lingüística informática* que otros autores utilizan" (Moure y Llisterri 1996:151-152).

Dentro de un contexto general dominado por una creciente necesidad de adquirir, procesar y transmitir información es donde surgen los trabajos de lo que se llamó *Lingüística Computacional* en torno a la década comprendida entre 1940 y 1950 en Estados Unidos, centrados, por una parte, en los cómputos de apariciones así como en la elaboración de índices y concordancias; y por otra, en la traducción automática. En este sentido, se concibe la LC como un campo muy amplio:

> Bajo la denominación de Lingüística computacional es posible agrupar un conjunto relativamente heterogéneo de teorías, métodos, herramientas, aplicaciones y productos que tienen en común la consideración de la lengua como un objeto susceptible de ser tratado mediante procedimientos informáticos (Moure y Llisterri 1996:147),

en el que se integra cualquier tarea lingüística para la cual se utilicen medios informáticos. Sin embargo, esta tradición (en lo referente a los cómputos de frecuencias, etc.; no en lo que concierne a la traducción automática) se ha venido a identificar más tarde con lo que se conoce como "literary and linguistic computing" y ya no se considera LC en

sentido estricto. Se trata de meros cálculos estadísticos que puede realizar cualquier procesador de textos actual.

Sin embargo, ahora, el término de LC se usa de forma más restringida, como sinónimo o al lado de *Procesamiento del Lenguaje Natural*, área que se ocupa de la modelización o emulación con medios informáticos de la conducta lingüística en toda su complejidad y, por lo tanto, está integrada en la *Inteligencia Artificial*, ciencia o subdisciplina de la Informática que persigue la construcción de sistemas computacionales inteligentes que simulen toda conducta cognitiva humana en general, la lingüística entre ellas[32], por lo que también entra en contacto con la Psicolingüística:

> *Computational linguistics: 1. (formerly, and still occasionally) A very broad label covering virtually any activity involving computers and natural language, such as machine translation of natural-language texts, computer searching of texts or the preparation of concordances for literary works by computer. Now usually called 'literary and linguistic computing'. 2. (more usually today) A synonym for natural-language processing"* (Trask 1993:53).

Es decir, se presenta como un área de intersección con numerosas disciplinas, con las que comparte parte de su objeto. Hay que destacar que este marco de interdisciplinariedad en el que se mueve es característico también de la ciencia cognitiva con la que se solapa en ocasiones. Superado el optimismo triunfalista de los primeros tiempos,

---

32

> *Computational linguistics is best viewed as a branch of artificial intelligence (AI). As all fields within AI, it is concerned with the investigation and modeling of a cognitive capacity. In the case of computational linguistics it is the language capacity that is in focus. (...) The goal is rather to identify and characterize the classes of processes and the types of knowledge which are implied by the ability to communicate and assimilate information using natural language regardless of their psychological status. One of the contributions of computational linguistics is a set of techniques which make it possible for linguistic knowledge to guide and constrain the linguistic processing performed in a natural language system* (HALVORSEN 1991 [1988]:202-203).

lo mismo que ocurrirá con la IA, se ha llegado a una postura más realista y práctica, que ha transcendido hasta alcanzar el mercado comercial y el terreno industrial.

Pues bien, como ya se ha señalado en los apartados previos, la de *Lingüística Computacional* no es la única denominación que se utiliza para aludir al campo que nos ocupa, sino que es habitual referirse a él también como *Procesamiento del Lenguaje Natural*, *Lingüística Informática*, *Ingeniería Lingüística*, *Tecnología del Lenguaje Humano*, etc.

Esta oscilación terminológica se debe, por una parte, a que la tarea de simular la conducta lingüística con medios informáticos ha sido abordada por diferentes ciencias y desde distintas perspectivas y, por otra parte, a la propia evolución de la LC.

El resultado es una serie de líneas de investigación que comparten el interés por el lenguaje y por su tratamiento computacional, pero que difieren en la forma de llevarlo a cabo.

Las principales líneas de investigación que se señalan en la bibliografía al respecto son:

- *Procesamiento del Lenguaje Natural (PLN).*
- *Inteligencia Artificial (IA).*
- *Lingüística Informática (LI).*
- *Industrias de la Lengua, Ingeniería Lingüística y Tecnologías del Lenguaje (Humano) o de la Lengua.*

Otras líneas de investigación son:

- *Tecnologías del habla*
- *Lingüística de corpus*

1.4.1. Procesamiento del Lenguaje Natural (PLN)

Este término, traducción del inglés "Natural Language Processing" (NLP), alterna con el de LC para referirse a la línea de investigación básica dentro del campo de intersección entre el lenguaje y los ordenadores. Es más, en la actualidad LC y PLN se tienden a identificar, por lo que ambos términos se pueden considerar sinónimos: "Computational linguistics: [...] 2. (more usually today) A synonym for natural-language processing" (Trask 1993:53).

Igual que la LC, el PLN considera los ordenadores como un instrumento adecuado para la descripción y explicación de las diferentes facetas o niveles del lenguaje: fonética y fonología, morfología, sintaxis, semántica, pragmática, análisis del discurso, etc.

El objetivo general, también común con la LC, es diseñar programas o sistemas informáticos que simulen la conducta lingüística humana en todas o en alguna de sus facetas, programas que sean capaces de utilizar lenguajes naturales: "[...] Lingüística Computacional y Procesamiento del Lenguaje Natural tratan de lo mismo: del desarrollo de programas de ordenador que simulan la capacidad lingüística humana" (Moreno Sandoval 1998:14). O en palabras de J. Allen (1995:1):

> *The goal of this research is to create computational models of language in enough detail that you could write computer programs to perform various tasks involving natural language. The ultimate goal is to be able to specify models that approach human performance in the linguistic tasks of reading, writing, hearing, and speaking.*

Para llevar a cabo este objetivo, el PLN diseña técnicas para representar mediante un metalenguaje los datos lingüísticos referidos a un determinado nivel del lenguaje, ya sea para el análisis o para la generación. Así entiende M.ª F. Verdejo (1995:39-40) el PLN:

> Entendemos por procesamiento del lenguaje natural (PLN) el estudio del mismo con el fin de crear modelos computacionales capaces de utilizarlo. Esta definición abarca una problemática muy amplia: desde la construcción de simples editores de texto en los que el lenguaje se considera como una cadena de caracteres, hasta interfaces para sistemas informáticos complejos capaces de ayudar a un usuario a plantear un problema dialogando en lenguaje natural.

Para L. Moreno *et al.* (1999:13) y L. Moreno y A. Molina (1999:65), "el Procesamiento del Lenguaje Natural (PLN) es una parte esencial de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos que faciliten la interrelación hombre/máquina y permitan una comunicación mucho más fluida y menos rígida que los lenguajes formales".

Como podemos observar, F. Verdejo se centra en la creación de modelos, mientras que L. Moreno en la efectividad de sistemas o interfaces que faciliten la comunicación hombre-máquina. En cualquier caso, "todo sistema de PLN intenta simular un comportamiento lingüístico humano" (Moreno *et al.* 1999:13).

Así pues, igual que se desprendía de las definiciones de LC, la creación de modelos del lenguaje y el hincapié en la comunicación hombre-máquina también parecen ser los pilares centrales del PLN, orientación que aparece estrechamente vinculada con otra rama de la Informática, la Inteligencia Artificial (IA), que se encarga de "codificar

en un programa informático facultades cognitivas" (Moreno Sandoval 1998:14), entre ellas el lenguaje.

La doble faceta teórica y aplicada del PLN también se puede rastrear en otro clásico, J. Allen (1995:1-3), quien distingue en el PLN motivaciones científicas y motivaciones aplicadas, de modo paralelo a como R. Grishman (1991 [1986]) había hecho a propósito de la LC. Para J. Allen (1995:1), el objetivo del PLN es "crear modelos computacionales del lenguaje lo suficientemente detallados que permitan escribir programas informáticos que realicen las diferentes tareas donde interviene el lenguaje natural". Dichos modelos los considera útiles "both for scientific purposes –for exploring the nature of linguistic communication– and for practical purposes –for enabling effective human-machine communication" (*ibid.*).

Los aspectos teóricos y aplicados, siguiendo a H. Tennant (1981:2) y J. Allen (1995:1-3), se pueden resumir así:

PLN TEÓRICO	***PLN APLICADO***
❑ Motivación científica.	❑ Motivación práctica o tecnológica.
❑ Se ocupa de explorar la naturaleza de la comunicación lingüística.	❑ Persigue posibilitar una comunicación hombre-máquina efectiva.
❑ Pretende mejorar el conocimiento que tenemos sobre el funcionamiento del lenguaje y de la mente humana.	❑ Busca la utilidad de no tener que necesitar un lenguaje artificial para comunicarse con el ordenador.
❑ Se encarga de desarrollar modelos computacionales que simulan la conducta lingüística.	❑ Consiste en aplicaciones concretas y en técnicas para representar la información lingüística mediante un metalenguaje.
❑ Estos modelos deben cumplir dos requisitos: i) ser adecuados para su tratamiento informático y ii) ser eficientes computacionalmente.	❑ Trata de que los modelos computacionales funcionen, aunque no reflejen la forma en que las personas procesamos el lenguaje.

Tabla 2. PLN Teórico vs. PLN Aplicado.

Desde el punto de vista teórico, lo destacable es que uno de los ámbitos principales es el desarrollo de gramáticas, analizadores sintácticos o *parsers* y léxicos computacionales, casos todos ellos en los que existe, por lo tanto, un verdadero "tratamiento" del lenguaje o análisis lingüístico.

No obstante, hay que destacar que para algunos autores el PLN se centra específicamente en los objetivos aplicados, en los que confluye con la vertiente aplicada de la LC o LC Aplicada (*vid.* Gómez Guinovart 1998, 1999 y 2000a). Por ejemplo, es la postura de M.ª A. Martí e I. Castellón (2000:4):

> *S'entén per processament del llenguatge natural (des d'ara PLN) el desenvolupament de formalismes de representació de les dades lingüístiques i de llenguatges de programació eficients per tractar-les. Es considera que el PLN s'ocupa dels aspectes tècnics de la LC, ja que tracta de trobar solució als problemes que planteja la comprensió del llenguatge natural en el marc d'aplicacions concretes, com la traducció automàtica, la indexació automàtica de textos, la interacció home-màquina en llenguatge natural, la confecció de resums, l'extracció i la recuperació d'informació, etc.*

O, en otro momento, M.ª A. Martí (2003:10) en el mismo sentido:

> Otra línea de investigación la tenemos en el Procesamiento del Lenguaje Natural, que se centra en los aspectos más aplicados de la LC, ya que trata de buscar soluciones a los problemas que plantea la comprensión del lenguaje natural en el marco de sistemas concretos.

Las aplicaciones del PLN se suelen clasificar en dos grandes grupos (Allen 1995:4-5; Moreno *et al.* 1999:17-19; Moreno y Molina 1999:67-69):

1) Aplicaciones basadas en diálogos: comunicación hombre-máquina escrita u oral	***2) Aplicaciones basadas en el tratamiento masivo de información textual: procesamiento de texto escrito***

• Sistemas de acceso a bases de datos.
• Sistemas de acceso a otros dominios.
• Sistemas de diálogo inteligente.
	• Para la creación de textos:
	  • Correctores ortográficos, sintácticos y de estilo.
	  • Diccionarios de sinónimos y tesauros.
	• Para el procesamiento de textos:
	  • Filtrado de documentos.
	  • Clasificación de documentos
	  • Indexación de documentos.
	  • Generación automática de resúmenes.
	  • Traducción automática.

Tabla 3. Aplicaciones del PLN.

Como se observa, el paralelismo entre las definiciones previas de LC y las de PLN aportadas ahora no puede ser mayor, paralelismo que se mantiene incluso en el terreno de las aplicaciones concretas. Por ejemplo, M.ª F. Verdejo (1995:70-80) y M.ª F. Verdejo y J. Gonzalo (1998:32-35) citan las interfaces, el tratamiento de textos, las ayudas a la escritura y la traducción automática, y señalan que las interfaces constituirían el objetivo prioritario, al incluir prácticamente todos los niveles de procesamiento del lenguaje.

Por este motivo, A. Moreno Sandoval identifica LC y PLN como dos disciplinas coincidentes por compartir el mismo objeto de estudio, el lenguaje y su simulación a través de un ordenador: “Por tanto, Lingüística Computacional y Procesamiento del Lenguaje Natural

tratan de lo mismo: del desarrollo de programas de ordenador que simulan la capacidad lingüística humana" (Moreno Sandoval 1998:14)[33].

Las diferencias entre LC y PLN se establecen considerando argumentos como los siguientes (*vid.* Gómez Guinovart 1998, 1999, 2000a; Cunningham 1999):

a) Ámbito de procedencia

- La LC se sitúa en la esfera de la Lingüística y de la Ciencia Cognitiva.
- El PLN está más vinculado a la Informática y a la Inteligencia Artificial, como dejan claro L. Moreno Boronat *et al.* (1999:13) y L. Moreno y A. Molina (1999:65):

> El Procesamiento del Lenguaje Natural (PLN) es una parte esencial de la Inteligencia Artificial que investiga y formula mecanismos computacionalmente efectivos que faciliten la interrelación hombre / máquina y permitan una comunicación mucho más fluida y menos rígida que los lenguajes formales.

b) Motivación

- En la LC predominan los aspectos teóricos o científicos.
- En el PLN, los aplicados o tecnológicos.

[33] Equivalencia que ya hemos visto registrada por TRASK (1993:53): "Computational linguistics: (...) 2. (more usually today) A synonym for natural-language processing".

En palabras de H. Cunningham (1999:4): “To summarise: CL is a part of the science of language that uses computers as investigative tools; NLP is part of the science of computation whose subject matter is computer systems that process human language”.

El siguiente cuadro sintetiza la relación entre LC y PLN:

LC		***PLN***	
Término más usual desde la perspectiva de la Lingüística		Término más común desde la perspectiva de la Informática	
LC Teórica	LC Aplicada	PLN Teórico	PLN Aplicado
Predominio de aspectos teóricos		Predominio de aspectos aplicados	

Tabla 4. Relación entre LC y PLN.

1.4.2. Inteligencia Artificial (IA)

Tanto la LC como el PLN se conciben como una rama de la IA. La Inteligencia Artificial (en inglés “Artificial Intelligence”, AI) es una de las subdisciplinas de la Informática. Su objetivo es diseñar ingenios artificiales que simulen el comportamiento inteligente humano: “Subdisciplina de la informática, encargada de codificar en un programa informático facultades cognitivas” (Moreno Sandoval 1998:14); “Es la ciencia que trata de que las máquinas hagan la clase de cosas que hace la mente humana” (Gregory 1995 [1987]:609).

En consecuencia, el lenguaje será objeto de estudio de la IA en tanto que facultad cognitiva básica. De hecho, constituye uno de los bloques centrales en que la IA suele estructurar el comportamiento inteligente humano, de ahí el empeño en su dominio desde los inicios de la disciplina, con vistas a obtener la comunicación hombre-máquina en lenguaje natural y no en un lenguaje formal y artificial, ya que las

lenguas son la forma más natural y eficiente de que disponemos para comunicarnos, sea entre nosotros mismos o con los ordenadores.

Para llevar a cabo este objetivo, la IA debe ocuparse de la descripción rigurosa de las estructuras del lenguaje así como de los conocimientos generales que poseemos las personas. Esto es así porque codificar en un programa informático la capacidad cognitiva del lenguaje implica que previamente se posee un conocimiento de cómo funciona este, conocimiento que es posible representar, hacer explícito de manera formal, despojado de toda posible ambigüedad.

Desde la perspectiva de la IA, el lenguaje se concibe como una parte de un todo, el sistema cognitivo humano. Este tratamiento del lenguaje es característico de la IA, pues la LC se ocupa del lenguaje en sí mismo, sin necesidad de integrarlo en un sistema más general.

Simular la conducta inteligente humana ha sido desde la más remota antigüedad una de las mayores aspiraciones de los filósofos, pensadores y científicos[34]. Dicha aspiración se ha concretado en el plano de la ficción y de la realidad en el diseño y construcción de autómatas o humanoides que plagan la mitología, la literatura y la ciencia, como los que ayudaban a Hefestos en su fragua, el Golem, Frankenstein o todo tipo de artilugios mecánicos, eso sin tener en cuenta todos los robots que pueblan la historia más reciente del cine, una de cuyas más recientes aportaciones ha sido precisamente la figura de David, el niño-robot protagonista de *AI*[35], la película de Steven Spielberg. David es el primer "meca" ("mecánico") programado para tener sentimientos.

Pues bien, esta meta de la ciencia y de la ficción se convirtió hace ya más de medio siglo, coincidiendo con la aparición del ordenador o, más

[34] *Vid*. BORRAJO *et al.* (1997 [1993]:23-32) y también MCCORDUCK (1991 [1979]), para una descripción más detallada de los intentos del hombre de construir seres mecánicos semejantes a él.

[35] Siglas del inglés "Artificial Intelligence", Inteligencia Artificial.

bien, gracias al ordenador, en el objeto de una nueva disciplina científica, cuya denominación es precisamente la de *Inteligencia Artificial*. Se inscribe en un marco interdisciplinar general, el que le proporciona la Ciencia Cognitiva, donde convergen campos como la Informática, la Lingüística, la Psicología, la Filosofía, las Matemáticas, la Biología, la Neurología, etc. Aunque también conoce una orientación más aplicada o tecnológica, puesto que algunas de esas investigaciones se concretan en productos comerciales o industriales.

Desde los primeros momentos, el tratamiento del lenguaje dentro del nuevo paradigma proporcionado por el ordenador se convierte en uno de sus pilares fundamentales, ya que se entiende que el lenguaje es una de las capacidades cognitivas básicas y, al mismo tiempo, más compleja, por lo que su dominio supondría un gran avance en la comprensión del funcionamiento de la mente humana. Tal y como ha demostrado la Pragmática, el lenguaje no se reduce a la simple combinación de signos según unas reglas, sino que está estrechamente unido a nuestras capacidades cognitivas generales, a nuestra memoria, a nuestras experiencias pasadas, a nuestras expectativas futuras, al conocimiento mutuo entre los hablantes, etc. De ahí su estatus como conducta inteligente.

Pero, ¿qué es la IA? Veámoslo a través de algunas definiciones. A. Moreno Sandoval (1998:14) define la IA como la subdisciplina de la informática encargada de "codificar en un programa informático facultades cognitivas". Según el *Diccionario Oxford de la mente* (Gregory 1995:60), "es la ciencia que trata de que las máquinas hagan la clase de cosas que hace la mente humana". Esas cosas son, entre otras, el mantener una conversación, contestar preguntas, etc... Y las máquinas son los ordenadores, encargados de ejecutar los programas informáticos que codifican dichas capacidades cognitivas. Para M. Meya (1980:135),

"la Inteligencia Artificial es una ciencia interdisciplinaria que tiene por objeto investigar el funcionamiento de la inteligencia humana, para aplicar luego estos modelos teóricos a una máquina que deberá ser capaz de reflejarlos". Y, por último, para H. A. Simon (1995:95), uno de los pioneros del campo:

> *AI deals with some of the phenomena surrounding computers, hence it is a part of computer science. It is also a part of psychology and cognitive science. It deals, in particular, with the phenomena that appear when computers perform tasks that, if performed by people, would be regarded as requiring intelligence –thinking.*

Así pues, la IA se presenta como una ciencia interdisciplinar, vinculada por una parte a la Informática, en tanto que pretende que las máquinas, es decir, los ordenadores actúen como personas, lo que implica codificar las capacidades cognitivas humanas en programas informáticos; y, por otra parte, el hecho de investigar la inteligencia humana la vincula a la Psicología y a la Ciencia Cognitiva.

Ahora bien, ¿qué es la "inteligencia" para la IA? ¿Qué tiene que hacer una máquina para que se considere "inteligente"?

El intento más serio de definir lo que se entiende por inteligencia es el conocido como "test de Turing", ideado por el matemático inglés del mismo nombre, A. Turing, en 1950. La prueba propuesta consiste en que una persona, situada en una habitación separada, formula preguntas a través de una terminal de ordenador a otra persona y a un ordenador. A partir de las respuestas proporcionadas por estos, debe determinar quién es quién. Si no puede decidirlo, se considera que esa máquina ha pasado el test de Turing y que, por lo tanto, es inteligente, definida la inteligencia de esta forma. Según sus previsiones, referidas a finales del siglo XX, las posibilidades de que un ordenador con 20 Gb

de memoria pudiera confundir a un interrogador humano después de 5 minutos eran inferiores al 30%.

Ahora bien, ¿por qué este empeño en simular en un ordenador los comportamientos inteligentes del ser humano?

Para H. A. Simon (1995:96), las razones por las que se intentan simular en un ordenador los comportamientos inteligentes humanos son tres:

1) Comprender la inteligencia en general a través del diseño e implementación de programas informáticos que la muestren y, a partir de ahí, construir una teoría sobre los sistemas inteligentes.
2) Comprender la mente humana a través del diseño de programas que muestren inteligencia utilizando los mismos procesos que emplean las personas para llevar a cabo esas mismas tareas.
3) Construir sistemas expertos, programas informáticos capaces de suplir o complementar la inteligencia humana en determinadas áreas o tareas.

A. M. Ramsay (1991:28-29) resume estos puntos en dos, que para él son las motivaciones básicas de la IA y que conducen a la concepción de la IA bien como una *ingeniería* bien como una *ciencia*:

a) *IA como ingeniería*: *acercamiento de la ingeniería* o *ingeniería del conocimiento*. Se basa en la utilidad que supone disponer de ordenadores "inteligentes". El objetivo de este acercamiento, eminentemente práctico y comercial, se concreta en resolver problemas reales usando la inteligencia artificial, entendida como un conjunto de ideas sobre la representación del conocimiento y la forma de emplearlo en la construcción de sistemas.

b) *IA como ciencia*: *acercamiento de la ciencia cognitiva*. Se basa en la posibilidad que brindan los ordenadores de investigar cómo la mente humana realiza esas tareas que requieren inteligencia. Desde esta perspectiva, de carácter teórico o de investigación básica, la IA se ocupa de discernir cuáles de esas ideas sobre la forma de representar el conocimiento, etc. aportan alguna explicación de la inteligencia o de alguna de las conductas inteligentes del hombre. En definitiva, trata de comprender la inteligencia en general y la humana en particular. Para ello, se construyen sistemas que imitan la estructura y/o el funcionamiento del cerebro humano.

A estas dos orientaciones fundamentales, la de la ingeniería y la de la ciencia, les corresponden, respectivamente, dos enfoques:

a) El *simbolismo*, que entiende la inteligencia como procesamiento simbólico de la información, es decir, la capacidad para manipular símbolos de acuerdo con unas reglas.

b) El *conexionismo*, que contempla la inteligencia como un todo complejo producto de las conexiones que se establecen entre sus partes. Rechaza, por lo tanto, la idea de que la inteligencia se reduzca a la manipulación de símbolos y reglas. Habla de una red formada por elementos simples y autónomos denominados "neuronas", por analogía con las células cerebrales, que se relacionan unos con otros para representar el conocimiento.

Así pues, igual que sucedía con la LC y el PLN, en la IA se diferencian dos tipos de objetivos: teóricos y aplicados que, en lo que al lenguaje se refiere, se identifican respectivamente con la LC y el PLN

(*vid.* p. ej. Tennant 1981:2; Ramsay 1991:28-29; Fernández y Sáez Vacas 1995:220):

a) Objetivos teóricos: punto de vista de la ciencia cognitiva.

- Simular la capacidad cognitiva del lenguaje en un sistema informático puede contribuir a aumentar nuestro conocimiento sobre el funcionamiento del cerebro humano y del lenguaje, lo que se corresponde con los objetivos teóricos o científicos de la LC-PLN.
- Utiliza los ordenadores y los programas informáticos como un banco de pruebas, como una herramienta para indagar sobre la naturaleza de la inteligencia, con independencia de la utilidad que pueda redundar de ello.
- Aborda los mismos temas de los que se ocupa la Lingüística Teórica, aunque atendiendo prioritariamente a la precisión y formalidad de las teorías.
- A veces se utiliza el término *Lingüística Computacional* para referirse a esta orientación más teórica.

b) Objetivos aplicados: acercamiento de la ingeniería.

- Simular la capacidad cognitiva del lenguaje en un sistema informático es ventajoso por la utilidad que supone disponer de ordenadores con esta habilidad, lo que entronca con los objetivos aplicados de la LC-PLN.
- Ante todo busca la utilidad.
- A este acercamiento corresponden los primeros trabajos relacionados con la LC-PLN, centrados en la traducción automática y, luego, en las interfaces.

- A veces el término *Procesamiento del Lenguaje Natural* se utiliza para referirse únicamente a esta orientación más aplicada.

Por supuesto, el lenguaje no es la única capacidad cognitiva de la que se ocupa la IA. Algunas áreas destacadas de la IA son:

- Comprensión y generación del lenguaje natural tanto en su vertiente hablada como escrita. Esta área tiene su origen en las aportaciones de N. Chomsky, de la Psicolingüística y de la Psicología cognitiva.
- Visión artificial o reconocimiento de patrones o formas, es decir, el desarrollo de estrategias que permiten a un ente interpretar las imágenes que capta del medio.
- Robótica: se ocupa de los mecanismos de control que permiten a un artilugio mecánico moverse en un medio físico y manipular elementos también físicos con cierto grado de autonomía.
- Emulación del razonamiento simbólico o simulación de la inteligencia a nivel funcional. Se trata del núcleo básico de la disciplina, ya que fue el que le dio origen (trabajos de Turing, Newell y Simon). Se ocupa del pensamiento lógico-matemático, de la capacidad de inferencia y de razonamiento, así como de la solución de problemas, juegos (damas, ajedrez, tres en línea) y la demostración automática de teoremas.
- Simulación del funcionamiento neuronal o simulación de la inteligencia a nivel físico. Esta corriente remonta a 1943, cuando McCulloch y Pitts idearon una “neurona electrónica”. El supuesto básico es que quizá la forma de emular la conducta inteligente humana pase por mimetizar la propia estructura física del cerebro humano.

- Sistemas expertos. Surgen en los años 70 como consecuencia de la evolución de la IA desde la búsqueda de soluciones a problemas generales, independientemente del campo considerado, hacia el estudio de los mecanismos empleados por un experto humano para resolver un problema en tiempo real en un campo muy concreto. P. ej. existen sistemas expertos en el diagnóstico médico, en genética molecular, en geología, etc.

Las aplicaciones de la IA se dejan sentir en áreas como la agricultura o la industria, en la que robots controlados mediante ordenador pueden realizar tareas que entrañan algún riesgo; en medicina, ya sea en el diagnóstico o en la intervención; en el hogar inteligente, en la enseñanza, en la aeronáutica, en la astronomía o las matemáticas, en el mundo de los negocios, aplicaciones militares, juegos, etc.

En lo que al lenguaje se refiere, es el PLN –ciencia o tecnología– el que se erige como un área básica de la IA, puesto que el tratamiento del lenguaje forma parte de un objetivo global, ya que se considera como una parte –fundamental– de un sistema más general, un sistema inteligente. De forma gráfica:

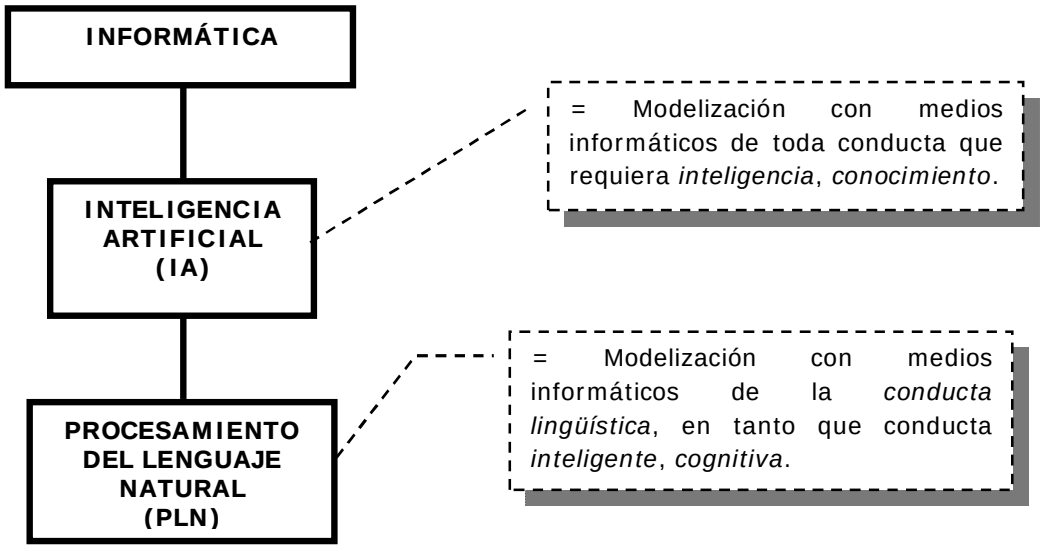


Ilustración 11. El PLN en el marco de la IA.

1.4.3. Lingüística Informática (LI)

Este término, del inglés "linguistic computing", constituye una línea de investigación muy amplia, ya que abarca cualquier uso de los ordenadores en relación con el lenguaje. También se denomina *informática aplicada a la lingüística* (*vid.* Gómez Guinovart 1998, 1999, 2000a).

Se trata de una de las líneas de investigación pioneras del campo. Utiliza los ordenadores como un instrumento más de trabajo en Lingüística. En los primeros tiempos de la LC y todavía a veces en la actualidad se identifica con la LC:

> *Computational linguistics: 1. (formerly, and still occasionally) A very broad label covering virtually any activity involving computers and natural language, such as machine translation of natural-language texts, computer searching of texts or the preparation of concordances for literary works by computer. Now usually called 'literary and linguistic computing'* (Trask 1993:53).

Concibe los ordenadores y los programas informáticos como herramientas eficaces para abordar tareas mecánicas y tediosas (contar, clasificar, buscar y ordenar la información, etc.) por la rapidez, exactitud y economía que introducen. Se concreta en todo tipo de programas y herramientas informáticas que puedan servir de ayuda en los estudios relacionados con la lengua y la literatura:

> Finalmente, tenemos en la Lingüística Informática, orientada al desarrollo de programas de apoyo a los estudios filológicos, lexicográficos, lingüísticos, humanísticos, etc., otra disciplina que ha contribuido al desarrollo de las tecnologías de la lengua. Estos programas tratan los textos como secuencias de caracteres, independientemente de la lengua en la que están escritos, y extraen de los mismos datos de tipo estadístico, concordancias, colocaciones, etc. (Martí 2003:11).

A diferencia de la LC y del PLN, por lo general realiza un tratamiento superficial de la información lingüística (cómputos, cálculos), sin que exista un verdadero análisis:

> *El terme Lingüística Informàtica fa referència als programes orientats a donar suport als estudis filològics, lexicogràfics, lingüístics, humanístics, etc. Aquests programes no cal que continguin coneixement lingüístic i consideren els textos com a seqüències de caràcters, siguin aquests lingüístics, numérics o de qualsevol altre tipus, independentment de la llengua que es tracta. El seu objectiu és l'extracció d'informació superficial dels textos en termes d'un determinat criteri, normalment de tipus quantitatiu o bé basat en el reconeixement de determinades seqüències en el text* (Martí y Castellón 2000:6).

Las áreas en las que mayor repercusión tiene son: lexicografía, sociolingüística, lingüística histórica, estilometría, lingüística estadística, lingüística de corpus, edición de textos, enseñanza y aprendizaje de lenguas, etc.

Los programas más representativos son los que sirven para extraer listas de frecuencias y los programas de concordancias (palabras en contexto), a los que hay que añadir los programas para la enseñanza de cualquiera de las áreas de la lingüística (*vid.* Lawler y Dry 1998).

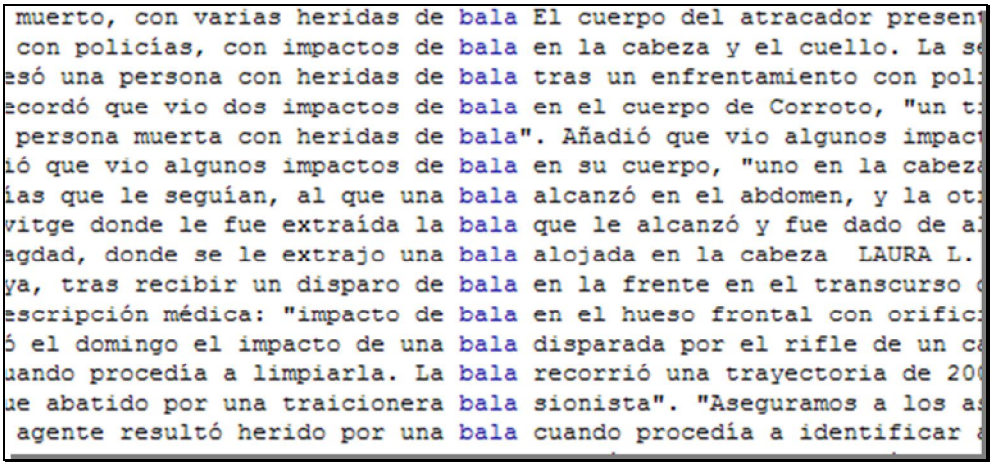

muerto, con varias heridas de bala El cuerpo del atracador presen
con policías, con impactos de bala en la cabeza y el cuello. La s
esó una persona con heridas de bala tras un enfrentamiento con pol
ecordó que vio dos impactos de bala en el cuerpo de Corroto, "un t
persona muerta con heridas de bala". Añadió que vio algunos impac
ió que vio algunos impactos de bala en su cuerpo, "uno en la cabez
ías que le seguían, al que una bala alcanzó en el abdomen, y la ot
vitge donde le fue extraída la bala que le alcanzó y fue dado de a
agdad, donde se le extrajo una bala alojada en la cabeza LAURA L.
ya, tras recibir un disparo de bala en la frente en el transcurso
escripción médica: "impacto de bala en el hueso frontal con orifi
ó el domingo el impacto de una bala disparada por el rifle de un c
uando procedía a limpiarla. La bala recorrió una trayectoria de 20
ue abatido por una traicionera bala sionista". "Aseguramos a los a
agente resultó herido por una bala cuando procedía a identificar

Ilustración 12. Ejemplo de concordancias para la palabra "bala" extraídas del CREA, Corpus de Referencia del Español Actual, Real Academia Española[36].

En la actualidad destaca especialmente la incidencia de lo que se ha dado en llamar "tecnologías de la información y de la comunicación" (TIC) (Internet, correo electrónico, etc.)[37].

[36] Las concordancias hacen posible inferir las relaciones que se establecen entre las palabras en virtud de sus distintas propiedades. En el ejemplo, se aprecian colocaciones del tipo: *impacto de bala*, *herida de bala*, *extraer una bala*…

[37] Visítese el sitio web Lab.Lingua, Laboratorio de Lingüística Informática de la Universidad de Alicante. URL: http://www.ua.es/dfelg/lablingua/

1.4.4. Industrias de la lengua, ingeniería lingüística y tecnologías del lenguaje (humano) o de la lengua

Industrias de la lengua, *ingeniería lingüística* y *tecnologías del lenguaje (humano)* o *de la lengua* son términos que se han ido poniendo de moda sucesivamente desde los años ochenta y principios de los noventa. Por lo tanto, representan las líneas de investigación más recientes en LC, resultado de la propia evolución de la disciplina. En su origen están, según J. Vidal y J. Busquets (1996:441), el crecimiento de las actividades que tienen como base la transferencia de información, y también toda una serie de programas gubernamentales, sobre todo en Japón y en la Comunidad Europea. Como señala el folleto *Lenguaje y tecnología. De la torre de Babel a la aldea global* (1997:12), estamos insertos en la era o sociedad de la información, en la que esta se considera un valor comercial y, como tal, susceptible de explotación industrial. Por otra parte, dado que los mercados tienden a ampliarse y que la información con frecuencia viene expresada a través de la *lengua,* el surgimiento de las llamadas *industrias de la lengua* y de la *ingeniería lingüística* no es más que una imposición de esos mercados de cara a facilitar el comercio internacional.

1.4.4.1. Industrias de la lengua

Según el *Centre de Referència en Enginyeria Lingüística* de la *Generalitat*, el surgimiento de las industrias de la lengua se debe a la confluencia de cuatro tipos de factores:

a) Factores socioculturales: las características de la sociedad moderna.

b) Factores científicos y tecnológicos: el avance de la investigación en todos los ámbitos aplicados y, en particular, en el de las tecnologías de la información.

c) Factores políticos: la defensa oficial de las lenguas.

d) Factores económicos: la generalización de los mercados y la importancia creciente de un "mercado lingüístico".

Parece que el término *industrias de la lengua* se utilizó por primera vez en francés ("les industries de la langue") a principios de los ochenta y, desde entonces, su uso se ha extendido rápidamente para referirse a todo tipo de actividades comerciales, profesionales y organizaciones relacionadas con la lengua (*cf.* Edwards y Kingscott 1997:262). En esta extensión ha tenido mucho que ver el uso del término en ámbitos políticos y planes de investigación europeos, tal y como señalan J. Vidal Villalba y J. Busquets Rigat (1996:434).

En concreto, el término hace referencia a un ámbito amplio que abarca "una serie de actividades comerciales en las que el tratamiento del lenguaje por personas o por máquinas, o por una combinación de unas y otras, forma una parte integrante del producto o servicio" (*Lenguaje y tecnología. De la torre de Babel a la aldea global* 1997:12), idea que también recogen T. Moure y J. Llisterri (1996:149-150):

> Finalmente, cabe considerar las llamadas industrias de la lengua, denominación usada a menudo junto con la de ingeniería lingüística para referirse a las aplicaciones del procesamiento del lenguaje natural y del habla en el desarrollo de productos comerciales, destinados a usuarios finales, que incluyen una parte importante de conocimientos sobre la lengua.

Estos productos, en el ámbito del texto, cumplen funciones relacionadas con la redacción, corrección, gestión y traducción de documentos: correctores de diverso tipo, diccionarios en soporte electrónico, programas de traducción automática o asistida, sistemas de consulta a bases de datos, sistemas de recuperación de información, de resumen automático, de enseñanza de lenguas, etc.

Por lo que se refiere al ámbito del habla, se utilizan para el reconocimiento y la síntesis del habla, la identificación de locutores o de lenguas, sistemas de diálogo, traducción del habla, sistemas de dictado, etc.

Así pues, se trata de una línea de investigación muy amplia, aunque el aspecto más destacado es que persigue la obtención de productos comerciales relacionados con la lengua. J. A. Edwards y A. G. Kingscott (1997:13 y ss.) clasifican estas actividades industriales en torno a la lengua en:

a) Monolingües: todas aquellas que giran en torno a una única lengua.

b) Bilingües: todas aquellas que trabajan con dos lenguas.

c) "Translingües": todas aquellas relacionadas con la traducción.

d) Multilingües: todas aquellas que trabajan con más de dos lenguas.

1.4.4.2. Ingeniería lingüística

El término *ingeniería lingüística* apareció por primera vez en un panel de COLING (*International Conference on Computational Linguistics*), en el congreso celebrado en 1988 (*cf.* Cunningham 1999:6), y de ahí se

extendió por Europa, fundamentalmente por su utilización en una sección del Programa Telemático de la Comisión Europea que lleva el mismo nombre: *Language Engineering*[38]. Es definida como:

> *Language Engineering is the discipline or act of engineering software systems that perform tasks involving processing human language. Both the construction process and its outputs are measurable and predictable. The literature of the field relates to both application of relevant scientific results and a body of practice* (Cunningham 1999:5).

El objetivo de la ingeniería lingüística es proporcionar "medios de ampliar y mejorar la utilización de la lengua para hacer de ella una herramienta más eficaz" (*Ingeniería lingüística. Cómo aprovechar la fuerza del lenguaje*, pág. 1). Para ello:

a) Parte de un cuerpo de conocimientos teóricos que le proporcionan las ciencias del lenguaje y otras en las que el lenguaje es parte de su objeto: "es la aplicación de los conocimientos sobre la lengua al desarrollo de sistemas informáticos para que puedan reconocer, comprender, interpretar y generar el lenguaje humano en todas sus formas" (*Ingeniería lingüística. Cómo aprovechar la fuerza del lenguaje*, pág. 5).

b) Con ellos elabora recursos lingüísticos que posteriormente explotará con la aplicación de técnicas informáticas:

[38] Véanse, para más detalles, LLISTERRI y GARRIDO (1998), así como los folletos *Ingeniería lingüística...* y *Lenguaje y tecnología*.

> En la práctica, la ingeniería lingüística consiste en una serie de técnicas y recursos que se aplican, en el primer caso, por medio de programas informáticos y que, en el segundo, constituyen una fuente de conocimientos a los que se puede acceder por medio de estos mismos programas (*Ingeniería lingüística. Cómo aprovechar la fuerza del lenguaje*, pág. 5).

Se suele englobar dentro del marco general de las industrias de la lengua (*cf.* Llisterri y Garrido 1998:299), al ser un requisito previo para que existan estas. En sentido estricto es una línea de investigación más específica que las industrias de la lengua, aunque en realidad se solapa con dicho término.

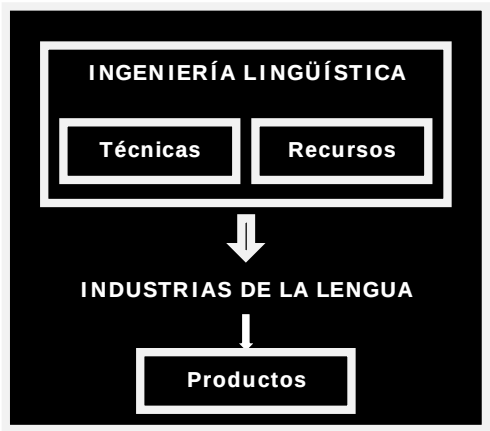


Ilustración 13. La ingeniería lingüística y las industrias de la lengua.

Según M.ª A. Martí e I. Castellón (2000:7), la *ingeniería lingüística* viene a marcar un punto de inflexión en el desarrollo de la LC, al representar la interacción de esta ciencia con la sociedad, a la que brinda productos aplicados: sistemas funcionales que la LC de corte teórico no proporciona y que la sociedad demanda (*cf.* Boguraev, Garigliano y Tait 1995):

> *Linguistic Engineering (LE) is an engineering endeavour, which is to combine scientific and technological knowledge in a number of relevant domains […]. LE can be seen as a rather pragmatic approach to computerised language processing, given the current inadequacies of theoretical CL [Computational Linguistics]* (European Commission, Linguistic Research and Engineering in the Framework Programme 1991 *apud* Boguraev, Garigliano y Tait 1995:1).

En esta orientación aplicada y comercial, coincide plenamente con las industrias de la lengua y se opone a la LC Teórica: "Language engineers make things work without knowing why, whereas computational linguists know why their systems don't work" (Cunningham 1999:5).

1.4.4.3. Tecnologías del lenguaje

Este es el término más habitual hoy en día, utilizado como sinónimo de *ingeniería lingüística*.

El objetivo último de las tecnologías lingüísticas es lograr la comunicación con los ordenadores mediante un lenguaje natural y el acceso a la información:

> Por tecnologías de la lengua o ingeniería lingüística se entiende [*sic*] los programas que procesan el lenguaje humano con los siguientes objetivos: mejorar la comunicación en todas sus modalidades y facilitar el acceso a la información por encima de las barreras que impone la distancia, el uso de lenguas distintas o el modo en que tiene lugar la comunicación, ya sea hablado o escrito. […] Se trata, en último término, de aplicar los conocimientos sobre la lengua al desarrollo de sistemas informáticos, con el fin de que puedan reconocer, comprender, interpretar y generar lenguaje humano en todas sus formas (Martí 2003:9).

Lo distintivo de esta línea de investigación, igual que las anteriores, es su clara orientación aplicada (aunque no carece de base teórica) y, sobre todo, comercial: se busca que las aplicaciones se concreten en productos de mercado (*cf.* Moure y Llisterri 1996, Vidal y Busquets 1996, Gómez Guinovart 1998 y 2000a), aunque parece comprender una meta algo más amplia, reflejada en el término *tecnología del lenguaje humano*[39], más próxima a la del Procesamiento del Lenguaje Natural o la Lingüística Computacional propiamente dicha:

> *The field of human language technology covers a broad range of activities with the eventual goal of enabling people to communicate with machines using natural communication skills. Research and development activities include the coding, recognition, interpretation, translation, and generation of language* (Cole *et al.* 1996).

Los conceptos clave, según M.ª A. Martí (2003:1), son: potencial económico e impacto social.

En el origen de todas estas últimas líneas de investigación están, según J. Vidal y J. Busquets (1996:441) y M.ª A. Martí (2003:9-10), factores como los siguientes:

- El crecimiento de las actividades que tienen como base la transferencia de información.

[39] "Language technologies are information technologies that are specialized for dealing with the most complex information medium in our world: human language. Therefore these technologies are also often subsumed under the term Human Language Technology" (USZKOREIT 1996, 2000).

- Toda una serie de programas gubernamentales e institucionales, sobre todo en Japón y en la Unión Europea, que las han fomentado.
- Las nuevas posibilidades industriales y tecnológicas.
- La mejora de la capacidad de los ordenadores.
- Lenguajes de programación más adecuados.
- La propia evolución de la disciplina.
- La investigación básica realizada en LC y PLN que ha propiciado el desarrollo de aplicaciones reales y no de laboratorio.

El resultado de la suma de todos estos factores es una nueva realidad que se ha venido a denominar *sociedad de la información*, marcada por la globalización, el multilingüismo y las posibilidades que ofrecen las nuevas tecnologías para la información y la comunicación.

El lenguaje tiene un papel central en esta sociedad, de ahí que se haya convertido en objeto de explotación comercial.

Las tecnologías lingüísticas no serían posibles sin el desarrollo previo de recursos lingüísticos y técnicas de análisis, necesarios para la obtención del producto comercial final.

Los *productos* pueden estar relacionados con la lengua escrita o con la lengua hablada.

a) *Lengua escrita*: se trata de productos que cumplen funciones que tienen que ver con la redacción, corrección, gestión y traducción de documentos.

- Correctores de diverso tipo
- Diccionarios en soporte electrónico
- Programas de traducción automática o asistida
- Sistemas de consulta a bases de datos
- Sistemas de recuperación de información
- Sistemas de resumen automático
- Programas de enseñanza de lenguas

b) *Lengua hablada*: se trata de programas que se utilizan para el reconocimiento y síntesis del habla, la identificación de locutores o de lenguas, sistemas de diálogo, traducción del habla, sistemas de dictado, etc.

c) *Recursos lingüísticos*: gramáticas computacionales, corpus y bancos de datos terminológicos.

d) *Técnicas de análisis*: programas informáticos para la introducción, procesamiento y generación de texto escrito y lengua oral[40].

[40] Para más información, es recomendable visitar el sitio web de la *Oficina del Español en la Sociedad de la Información* (OESI) sobre tecnologías lingüísticas: URL: http://oesi.cervantes.es/oesi/tls.jsp

1.4.5. Otras líneas de investigación

De forma breve, mencionaremos otras dos líneas de investigación (*cf.* Moure y Llisterri 1996):

1.4.5.1. Las tecnologías del habla

Aunque en principio la LC-PLN también abarca el tratamiento de la lengua hablada, los trabajos de este último terreno hoy en día han alcanzado tal desarrollo que con frecuencia se recogen bajo la denominación de *tecnologías del habla*. Desde el punto de vista histórico también tiene su justificación esta división, pues los trabajos sobre el tratamiento del habla se iniciaron en el ámbito de la ingeniería de telecomunicaciones, y no en el de la Informática (PLN) o Lingüística (LC). En conjunto, se trata de un campo cuyos intereses se centran principalmente en:

- conversión de textos escritos en su equivalente oral: *síntesis del habla*
- transformación del habla en texto: *reconocimiento del habla*
- traducción automática de conversaciones
- identificación o verificación de hablantes en servicios telefónicos
- sistemas de diálogo oral entra personas y máquinas

1.4.5.2. La lingüística de corpus

Se ocupa de la constitución y explotación de grandes muestras de uso real de la lengua, "corpus", tanto en su vertiente textual como oral. Se puede concebir como un ámbito transversal a todas las líneas de investigación anteriores, ya que les proporciona el *input* necesario, bien para realizar las investigaciones teóricas pertinentes o para desarrollar aplicaciones sustentadas sobre material auténtico.

El avance de los estudios basados en corpus ha sido espectacular desde las décadas de los ochenta y los noventa, propiciado en gran medida por los avances tecnológicos: la capacidad de almacenamiento de los ordenadores, la velocidad de los procesadores, la aplicación de técnicas estadísticas para la manipulación de los datos, etc. De hecho, su importancia hoy en día es tal que puede servir como parámetro para medir el grado de desarrollo computacional de una lengua. Y este lugar de privilegio no se reduce al ámbito de la LC, sino que se extiende a todo tipo de estudio sobre el lenguaje realizado en Lingüística o Filología, donde para muchos se ha erigido en un recurso imprescindible.

Todas estas líneas de investigaciones apuntadas desde una perspectiva teórica, en la práctica tienden a confluir. T. Moure y J. Llisterri (1996) señalan una serie de elementos comunes a todas ellas:

1) Integración cada vez mayor entre el PLN y las tecnologías del habla, apreciables en proyectos europeos (*European Network for the Integration of Language and Speech*, ELSNET[41]), metodologías empleadas (uso de técnicas estadísticas, propias del habla, en el

[41] URL: http://www.elsnet.org/

tratamiento del texto), trabajos en sistemas de traducción automática oral, que tienen que dar cuenta tanto de la lengua hablada como de la escrita, publicaciones (*Computer Speech and Language*), congresos (*International Conference on Spoken Language Processing*, ICSLP), etc.

2) Fronteras difusas entre el PLN y las tecnologías del habla y las industrias de la lengua: la separación teórica entre técnicas (PLN) y aplicaciones comerciales (industrias de la lengua) no se corresponde a veces con la práctica, ya que hay empresas que integran la I + D, en las que trabajan de forma conjunta ingenieros o informáticos y lingüistas.

3) Necesidad de *recursos lingüísticos*, tales como recopilaciones de textos escritos, de grabaciones de lengua hablada, diccionarios, terminologías especializadas o gramáticas, todos ellos en formato digital que posibilite su acceso y tratamiento informáticos. En este sentido, se ha creado la *European Language Resources Association* (ELRA)[42], que tiene su contrapartida en el *Linguistic Data Consortium* (LDC)[43] en América.

[42] URL: http://www.elra.info/
[43] URL: http://www.ldc.upenn.edu/

1.5. Evolución histórica

Caracterizadas *grosso modo* algunas de las principales tendencias o líneas de trabajo que se pueden encontrar bajo la etiqueta de *Lingüística Computacional*, pasamos a hacer un sucinto recorrido por algunos de los principales hitos en la historia de la LC.

Con independencia de la terminología (*Lingüística Computacional*, *Procesamiento del Lenguaje Natural*, *Lingüística Informática*, *Industrias de la lengua*, *Ingeniería Lingüística*, *Tecnología Lingüística*, etc.) o de la perspectiva adoptadas (lingüística o informática, teórica o aplicada), se suelen distinguir cuatro o cinco grandes etapas, según los autores, en el desarrollo histórico del campo de la LC.

Evidentemente, cuando la revisión histórica se realiza desde la perspectiva de la Lingüística, se presta especial atención a los avances científicos del momento mientras que, cuando la perspectiva es la de la Informática, el centro de atención lo constituyen las aplicaciones y sistemas desarrollados en cada fase. Aquí conjugaremos ambos acercamientos para dar una visión general de las tendencias predominantes en cada etapa, con especial énfasis en los "balbuceos" que precedieron a la actual consolidación de este ámbito del saber.

1.5.1. Los orígenes

Se puede decir que todo "comienza" a finales de los años cuarenta, en centros de investigación de EE.UU., Inglaterra, la entonces URSS y Francia, solo unos años después de la invención del ordenador[44] (1946)

[44] Aunque ENIAC es el que se ha llevado la fama al pasar a la historia como el primer ordenador digital de propósito general, realmente se trata de un invento que

al término de la II Guerra Mundial. Es entonces cuando va a surgir una nueva disciplina lingüística, la *Lingüística Computacional* (LC), con el objetivo de estudiar el lenguaje desde la nueva perspectiva que le va a brindar aquel. Desde ese momento inicial hasta mediados de los sesenta y principios de los setenta, cuando alcanza su consolidación, una serie de factores externos e internos a la propia Lingüística van a ir configurando este nuevo campo de investigación que centra nuestro interés.

Entre los acontecimientos acaecidos en ese período que llevaron al nacimiento de la LC y que determinaron su ulterior evolución, lo primero que hay que señalar es que, cualquiera que sea el punto de vista adoptado, lo que resulta evidente es que no se puede hablar de LC sin ordenadores. Sin embargo, estos no se habían pensado para tratar el lenguaje. El desarrollo del ordenador digital estuvo marcado por una motivación bélica. Durante la Segunda Guerra Mundial matemáticos e ingenieros prominentes en sus campos –p. ej. W. Weaver[45] o A. Turing– fueron contratados por sus respectivos gobiernos –el estadounidense y el británico– para que construyeran máquinas capaces de efectuar las complejas operaciones que se precisaban para descifrar mensajes clave o realizar cálculos balísticos. Por lo tanto, estas grandes calculadoras se diseñaron específicamente para trabajar con números, no con palabras.

---

se desarrolló de forma paralela en centros de Inglaterra –donde en 1943 empezó a operar Colossus, diseñado con el propósito específico de descifrar el código en que estaban encriptadas las comunicaciones alemanas- y Estados Unidos sobre todo, pero también de Alemania. P. McCORDUCK (1991:65-66) hace referencia a las solicitudes de patentes por parte del ingeniero K. Zuse en Alemania para sostener que a él se debe el primer ordenador de este tipo, aunque a causa de la "suerte" alemana en la guerra su trabajo quedó olvidado y su máquina destruida (*cf.* COPELAND 2000 para una breve historia de los primeros ordenadores).

[45] Este planteará de forma explícita la idea de utilizar los ordenadores para la traducción de una lengua a otra.

A. Turing, matemático inglés (1912-1954), fue quien sentó las bases conceptuales de la moderna Informática, pero además apuntó algunas ideas sobre las que luego se fundamentará la Inteligencia Artificial[46] como ciencia, en torno a 1956: los ordenadores como sistemas inteligentes[47] capaces de simular cualquier conducta cognitiva humana. En este sentido, en un artículo de 1947 (publicado en 1969), "Intelligent Machinery", especula con la posibilidad de construir máquinas capaces de "pensar", aunque considera más útil por el momento emplearlas para tareas más triviales, en tanto en cuanto le parecen más fáciles de llevar a la práctica y también más provechosas, como jugar al ajedrez, al tres en raya, aprender idiomas, traducir, descifrar mensajes en clave o hacer matemáticas, lo que constituye un resumen perfecto de los que serán los principales temas de investigación en IA una década más tarde:

> Aquí estamos principalmente interesados en el sistema nervioso. Podríamos construir modelos eléctricos bastante exactos para copiar el comportamiento de los nervios, pero esto no parece tener mucha utilidad. [...] Los circuitos eléctricos que se utilizan en las máquinas computadoras parecen tener la propiedad esencial de los nervios. Son capaces de transmitir información de un lugar a otro, así como almacenarla (*apud* McCorduck 1991:71).

No es extraño que A. Turing se expresara en estos términos, pues muchos eran los que por aquel entonces pensaban lo mismo. El enorme desarrollo experimentado por las Matemáticas como sistema formal las

[46] Y que tendrán una gran trascendencia en el futuro desarrollo de la Lingüística Computacional.

[47] De hecho, a él debemos el conocido como "Test de Turing", una prueba objetiva para determinar si se puede considerar "inteligente" la conducta de un programa informático, idea expuesta en su famoso artículo "Computing Machinery and Intelligence" (1950), traducido al español como "¿Puede pensar una máquina?".

había convertido en el lenguaje universal que ansiaba la ciencia, de ahí que vaya a ser el punto de referencia para cualquier actividad que se precie de científica, la Lingüística incluida. Así, con las Matemáticas se va a intentar explicar el comportamiento humano y el de los ordenadores. Este cometido lo asumirá en primer lugar la Cibernética, un nuevo tipo de ingeniería propuesto por primera vez en 1943 por A. Rosenblueth, J. Bigelow y N. Wiener y confirmado en 1948 con la obra *Cybernetics* de Wiener. Se define como la "ciencia de la comunicación y el control en y entre los animales, hombre incluido, y las máquinas"[48] (Borrajo *et al.* 1997:30) y su meta era "encontrar un conjunto de principios sencillos que explicaran las actividades de la mente humana" (*ibid.*). La analogía entre los ordenadores, el ejemplo más prototípico de máquina cibernética, y el cerebro humano era obvia "ya que en los dos casos se trataba de sistemas que facilitaban la entrada de las informaciones, su retención en la memoria, su tratamiento, su transmisión por los canales de la comunicación, así como su entrega en la salida" (Černý 2000:323).

De forma casi paralela, el matemático e ingeniero C. Shannon y el también matemático estadounidense W. Weaver (1894-1978) formulan en 1949 la influyente teoría de la información y de la comunicación. Entre otras cuestiones, destacan que el proceso de descodificación de la información se basa en gran medida en conocimientos de tipo probabilístico: en función de lo que se ha descifrado se calcula la probabilidad de la información siguiente (*cf.* Černý 2000:286). Tanto los ordenadores como los cerebros son sistemas capaces de procesar información expresada en forma de símbolos, numéricos en el caso de

[48] N. Wiener, igual que Weaver o Turing, participó en la guerra. Llevó a cabo trabajos relacionados con el "control" de proyectiles y llegó a construir "la primera bomba volante, o sea, el arma que era capaz de alcanzar un blanco a larga distancia, ya que su trayectoria era controlada continuamente [...] y, según fuera preciso, corregida todavía durante el vuelo" (ČERNÝ 2000:323).

los primeros, y no numéricos en el de los segundos. Sin embargo, C. Shannon pronto pensó en la aplicación de los ordenadores para manipular símbolos no numéricos. En 1950, en su artículo "A Chess Playing Machine", señaló que "las nuevas máquinas podían no solo llevar a cabo cálculos numéricos, sino que eran tan generales y flexibles que podían 'adaptarse para trabajar simbólicamente con elementos que representaran palabras, proposiciones u otras entidades conceptuales'" (McCorduck 1991:112-113).

De esta forma, "el *ordenador de propósito general* se hizo necesario porque las teorías de los procesos mentales se habían vuelto demasiado complejas y estaban evolucionando demasiado rápidamente para poder ser practicadas en máquinas ordinarias" (Minsky *apud* McCorduck 1991:336).

Todas estas teorías, íntimamente relacionadas entre sí, y sus principales representantes, van a coincidir en el verano de 1956 en el Dartmouth College (Hanover, New Hampshire) en un ciclo de conferencias, promovido por J. McCarthy, sobre las nuevas posibilidades que para la investigación ofrecía el ordenador. El resultado será la fundación de la IA sobre tres pilares fundamentales (*cf.* McCorduck 1991:105 y 1993:95):

a) El pensamiento podía simularse fuera del cerebro humano.

b) Podía explicarse en términos científicos, de una manera formal.

c) El mejor instrumento de laboratorio para llevar a cabo esta tarea era el ordenador digital.

Como dice M. Minsky, uno de los presentes en el encuentro:

> Paradójicamente, la ciencia de la computadora, que fue inspirada por una exigencia de enormes cálculos cuantitativos, por su inesperada capacidad de describir procesos señaló, en su lugar, el nacimiento de una ciencia de lo cualitativo. Esta capacidad de manejar lo complejo, dinámico y cualitativo dio lugar a intentos de modelizar el pensamiento humano (*apud* Borrajo *et al.* 1997:31).

Por lo tanto, si era posible de alguna manera simular el pensamiento, la inteligencia humana en términos lógico-matemáticos en un ordenador, si el cerebro humano y los circuitos eléctricos compartían ciertas similitudes de funcionamiento como procesadores de información y manipuladores de símbolos, el siguiente paso era intentar reproducir la capacidad lingüística, una de las conductas cognitivas más complejas pero al mismo tiempo más humanas, en un ordenador. En este momento nace la LC (o PLN), aunque tal denominación todavía tardará una década en acuñarse[49]. En concreto, va a ser la traducción automática[50] la primera aplicación de los ordenadores en relación con el lenguaje que suscite el interés de los investigadores[51].

---

[49] "El procesamiento automático del lenguaje natural (PLN) surge en los años cincuenta, y su historia se entrelaza con las investigaciones que sobre el lenguaje se llevan a cabo en lingüística formal, psicología cognitiva, lógica y ya dentro de la informática la teoría de los lenguajes, la compilación y sobre todo la inteligencia artificial" (VERDEJO y GONZALO 1998:29).

[50] "The use of computers for the translation of natural languages was probably the first application of the newly invented electronic computers to non-numerical tasks, and it was certainly the first application in what was later to be known as computational linguistics. It was also one of the first areas of research in the field of artificial intelligence" (HUTCHINS 2000a:1).

[51] Las conclusiones que se extraen sobre los trabajos en traducción automática son, en consecuencia, representativas de la concepción que se tenía del tratamiento del lenguaje con medios informáticos en este momento, al ser la aplicación pionera.

F. J. Newmeyer (1986:1) describe el ambiente que reinaba entre los lingüistas en EE.UU. en la década de los 50 como de optimismo, debido a los logros teóricos alcanzados, especialmente relevantes para nuestros propósitos los referidos a la formalización del lenguaje (Bloch y Harris), que situaban a la Lingüística a la misma altura científica de la Física, de las Matemáticas o de la Mecánica. Prueba del nivel que se había alcanzado era que consideraban que sus teorías estaban listas para ser informatizadas:

> *Many linguists felt that the procedures had been so well worked out that computers could take over the drudgery of linguistic analysis. The time was near at hand when all one would have to do would be to punch the data into the computer and out would come the grammar!*
>
> *There was also a feeling that computers could solve another traditional linguistic problem –translation* (Newmeyer 1986:2).

Los avances de la lingüística formal en la línea que a mediados de los cincuenta inaugurará N. Chomsky van a contribuir a este clima de euforia. Por una parte, N. Chomsky establece las bases de la llamada *lingüística algebraica*, el estudio de las gramáticas formales dentro de las Matemáticas, al elaborar una tipología de las diferentes clases de gramáticas formales atendiendo a su poder generativo: la conocida como *jerarquía de Chomsky* (*cf.* Moreno Sandoval 1998:48). Este hecho será decisivo en el ámbito de la Informática –en la actualidad la teoría de los autómatas y los lenguajes formales es un capítulo que no puede faltar en ningún manual de la materia– y pronto algunos investigadores intentarán llevar a la práctica computacional esas ideas. Pero también va a tener repercusiones en la Lingüística, ya que dicha clasificación le sirvió a Chomsky para proponer en 1957 las gramáticas generativo-

transformacionales como las más adecuadas de la jerarquía para la descripción de las lenguas naturales. En consecuencia, la sintaxis va a ser el nivel lingüístico que los primeros lingüistas computacionales intentarán someter a las exigencias del ordenador, aunque implementar una gramática generativo-transformacional, con la estructura profunda y la estructura superficial, así como las reglas de transformación que relacionan ambos niveles, se revelará una tarea harto complicada, lo que favorecerá el desarrollo de otros formalismos gramaticales divergentes del paradigma "chomskiano", del tipo de las gramáticas categoriales o las gramáticas de estados finitos, mucho más eficientes en términos computacionales.

Con estas bases, se inicia la primera generación de traducción automática y, con ella, llegan las investigaciones inaugurales de la LC, es decir, los primeros trabajos computacionales sobre el lenguaje. Es frecuente en la bibliografía del momento encontrar también sustantivos como "optimismo", "euforia", "entusiasmo", etc. para describir los ánimos que alentaban a estos primeros trabajadores en traducción automática, la mayoría de ellos ajenos al campo de la Lingüística –ingenieros, matemáticos…– y, por lo tanto, desconocedores de las complejidades inherentes al lenguaje en general y, en consecuencia, de cara a su tratamiento informático. Pero, deslumbrados por las nuevas posibilidades que les ofrecía el ordenador y empujados por las generosas sumas de dinero que, sobre todo, el gobierno norteamericano y las agencias militares y de inteligencia estaban invirtiendo[52], veían posible y próximo en el tiempo el logro de sistemas capaces de llevar a

[52] Hay que tener en cuenta el contexto en el que nos encontramos, finalizada la Segunda Guerra Mundial y con dos países, Estados Unidos y la URSS, dispuestos a iniciar la carrera espacial, por lo que la traducción automática se veía como un medio de acceder de forma rápida y económica a los avances científicos del otro.

cabo "traducciones totalmente automáticas de alta calidad"[53], de todo tipo de textos, los literarios incluidos, e incluso traducciones orales, no solo escritas (*cf.* Hutchins 2000a:3), es decir, traducciones perfectas en las que el hombre no tendría necesidad de intervenir, algo que incluso con la tecnología actual se considera inviable. Y es que creían que el proceso de traducción iba a ser una tarea fácil, a la que se podían aplicar las mismas técnicas criptográficas de base estadística que tanto éxito habían tenido durante la guerra para el desciframiento de mensajes encriptados[54]: todo era cuestión de hallar el código, la clave, que además se pensaba que era universal, que permitiera el trasvase lingüístico. En este sentido se manifestaba en 1949 W. Weaver, uno de los pioneros del campo junto a A. D. Booth:

> *It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code". If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?* (*apud* Hutchins 1999).

Este matemático estadounidense había trabajado como criptógrafo durante la guerra aplicando las matemáticas y la probabilidad para descifrar mensajes del bando alemán. En ese mismo período se había familiarizado con el uso de los ordenadores como potentes máquinas de cálculo (*cf.* Hutchins 1999). Y recordemos que había colaborado con C.

---

[53] Traducción de lo que en inglés se conoce como FAHQT, "Fully Automatic High Quality Translation".

[54] P. ej. el matemático inglés A. Turing había trabajado durante la guerra en un proyecto secreto denominado "Ultra" que tenía que ver con el desciframiento del código alemán a partir de una máquina cifradora, "Enigma", que había sido capturada a los alemanes. La clave se encontró gracias a los esfuerzos de Turing y a una potente máquina electromagnética, clara antecesora de los ordenadores digitales, que permitía implementar cálculos numéricos muy complejos (*cf.* McCorduck 1991:§3).

Shannon en la elaboración de la teoría de la información. En 1947 escribió a N. Wiener, el fundador de la Cibernética, comentándole la posibilidad de emplear los ordenadores para la traducción. Sin embargo, este no sintió interés por la idea, por lo que buscó otro interlocutor, el cristalógrafo inglés A. D. Booth, quien sí que estaba interesado (*cf.* Bennett 1995:445). En 1949[55], para muchos la fecha oficial de inicio de la LC, Weaver elaboró un memorándum, "Translation", en el que especulaba sobre la posibilidad de utilizar el recién inventado ordenador digital para traducir documentos de forma automática, llamando la atención sobre cuatro puntos:

i) la necesidad de tener en cuenta el contexto[56];

ii) los elementos lógicos del lenguaje, comunes a las lenguas, en conexión con la aplicación de las matemáticas y la lógica para formalizar los procesos mentales;

iii) la aplicación de métodos criptográficos independientes de la lengua basados en la estadística; y

iv) los universales lingüísticos, que ejemplifica con el siguiente símil, un poco extenso en la cita, pero revelador de la mentalidad del momento:

[55] La idea de la traducción automática ya había sido planteada en Rusia en la década de los 30, aunque no tuvo demasiada difusión exterior (*cf.* HUTCHINS 2000a:5).

[56]

> *If one examines the words in a book, one at a time through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of words. […] But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning...* (*apud* HUTCHINS 1999).

> *Think, by analogy, of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another, they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed. But, when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers.*
>
> *Thus it may be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to descend, from each language, down to the common base of human communication –the real but as yet undiscovered universal language– and then re-emerge by whatever particular route is convenient* (Weaver 1949 *apud* Hutchins 1999).

Poco después de este memorándum, se celebra la primera conferencia sobre traducción automática organizada por el lógico Y. Bar-Hillel en el MIT (1952) y dos años más tarde, en 1954, se crea la revista *Mechanical Translation* y se realiza, en el centro de IBM en Georgetown, la primera demostración pública de un sistema que traducía del ruso al inglés (*cf.* Kay 2003). El interés suscitado atrae la deseada financiación y provoca que en otros países se creen grupos de investigación en torno a la traducción automática.

Puesto que no estamos ante un campo de trabajo uniforme, ya que los investigadores procedían de la Ingeniería Eléctrica, de la Física, de la Lingüística, de la Filosofía…, sus intereses a propósito de la traducción automática eran diferentes: en el caso de los ingenieros y matemáticos prevalecía una orientación práctica, se esperaba obtener a corto plazo sistemas de traducción automática útiles, basados en métodos estadísticos –lo que se llamó "fuerza bruta" (*cf.* Hutchins 1995:433)–; en

el caso de lingüistas y lógicos, el acercamiento era teórico, se buscaba conseguir a largo plazo sistemas bien fundamentados lingüísticamente – "perfeccionismo" (*cf.* Hutchins *ibid.*)[57].

Asimismo, los puntos de partida también divergían: los primeros se insertaban en la tradición de la Cibernética y la teoría de la información (Wiener, Weaver, Shannon) desarrollada a mediados de los cuarenta; los segundos seguían la tradición de lógicos, filósofos del lenguaje y las figuras más destacadas de la lingüística estructural del momento, europea y americana, desde Saussure hasta Z. Harris primero, y poco después la lingüística generativo-transformacional de N. Chomsky, aunque la influencia de este último se dejará sentir sobre todo a partir de los sesenta. Sin embargo, en este primer momento, se desconfía de las teorías lingüísticas en general, y se prefieren métodos empíricos, basados en la teoría de la información, y el enfoque directo o traducción palabra a palabra, que requiere un tratamiento lingüístico mínimo: básicamente se toman las palabras de la lengua fuente, se consultan en un diccionario bilingüe y se genera la traducción en la lengua meta, a la que, como mucho, se le aplican algunas reglas de reordenamiento sintáctico. Es decir, el peso de la traducción recae en la elaboración de buenos diccionarios bilingües. No obstante, algunos grupos empiezan a investigar otros acercamientos a la traducción, como la interlengua y la transferencia[58].

Esta importancia de la traducción automática se va a confirmar con la fundación en 1962 de la *Association for Machine Translation and Computational Linguistics* y el lanzamiento en 1965 de la revista

[57] Esta doble orientación estará presente desde este momento en el terreno de la LC, del PLN y de la IA. *Cf.* R. GRISHMAN (1991 [1986]) respecto a la LC; H. TENNANT (1981:2) y J. ALLEN (1995:1) distinguen objetivos científicos y objetivos prácticos a propósito del PLN; A. RAMSAY (1991:28-29) diferencia, dentro de la IA, el punto de vista de la ciencia cognitiva y el acercamiento de la ingeniería.

[58] *Cf.* HUTCHINS y SOMERS (1995) para una descripción de los fundamentos de cada método.

Mechanical Translation and Computational Linguistics, antecedentes de la *Association for Computational Linguistics* (1968) y la revista *Computational Linguistics* respectivamente.

Pero desde el principio no faltaron los escépticos o, si se quiere, los que como Y. Bar-Hillel –quien ha pasado a la historia como el primer investigador contratado a tiempo completo para dedicarse a la traducción automática– tenían una visión más realista de los problemas que implicaba el tratamiento computacional del lenguaje, especialmente cuando pronto iban a surgir cuestiones a las que ni la propia Lingüística del momento daba cabida, lo que daría en llamarse la "barrera semántica" (*cf.* Hutchins 2000b). Para Y. Bar-Hillel, era imposible efectuar traducciones totalmente automáticas, sin intervención humana. La presencia del hombre era imprescindible, bien previamente a que se iniciara el proceso (*pre-edición*) o con posterioridad (*post-edición*). Solo con esta colaboración se podrán obtener resultados (*cf.* Hutchins 2000b).

Los grupos de orientación teórica habían intentado abordar el nivel sintáctico: "A number of approaches to syntax were examined, including use of Chomsky's formal grammar by the MIT group and the stratificational approach proposed by Sydney Lamb" (Bennett 1995:447). A partir de los sesenta este será el acercamiento predominante, aunque los problemas relacionados con el significado y con el conocimiento del mundo, obviados en la teoría lingüística, se harán cada vez más patentes: "We will only have adequate mechanical translations when the machine can 'understand' what it is translating" (Yngve *apud* Hutchins 1995:434).

La aparición de la obra de N. Chomsky *Aspects of the theory of syntax* en 1965, si bien suponía la consolidación en Lingüística del modelo generativo-transformacional, también reafirmaba la extensión de otra idea: que las gramáticas de estructura de frase no eran aptas para

describir las lenguas naturales, mientras que las gramáticas transformacionales sí (*cf.* Roeck 1995:154). Esta idea chocaba con los experimentos que los propios colegas de Chomsky en el MIT estaban llevando a cabo en relación con la traducción automática, como era el caso de Bar-Hillel (1951-1953) primero y V. Yngve (1953-1965) después.

Tras probar el modelo generativo-transformacional sin éxito –"The failure of Chomsky's theory in this first test of it was a disappointment to us and it was not for lack of trying or any ineptitude on our part" (Yngve 2000:56)– dirigieron sus esfuerzos hacia otro tipo de gramáticas, como la gramática categorial propuesta por Bar-Hillel en la línea de las ideas que le había inspirado el lógico polaco K. Ajdukiewicz (*cf.* Hutchins 2000b), lo que vino a demostrar la viabilidad de las gramáticas de estructura de frase para describir las lenguas naturales y, en consecuencia, para su aplicación a la LC, al poder dar cuenta de los mismos fenómenos que las gramáticas transformacionales (*cf.* Roeck 1995:154). No obstante, hay que señalar que Chomsky, en coherencia con su propuesta teórica, en ningún momento planteó su teoría con vistas a su implementación informática. De hecho, parece que no tenía interés alguno en toda la línea de trabajos de índole empirista surgidos en torno al ordenador.

Así pues, a mediados de los sesenta empiezan a manifestarse los primeros síntomas de un cambio de tendencia. La desilusión se irá apoderando poco a poco de los investigadores ante la falta de progresos y de resultados, consecuencia en buena medida de: a) la ausencia de teorías lingüísticas en los casos en los que se había optado por un acercamiento estadístico con objetivos prácticos a corto plazo, y b) de la inadecuación de las teorías gramaticales del momento –bien fueran de orientación estructural o generativo-transformacional– en los proyectos

en que sí había un tratamiento lingüístico de los datos, básicamente del nivel sintáctico.

Se toma conciencia, de golpe, de la complejidad del lenguaje y de la dificultad que entraña la descripción de todos y cada uno de sus niveles, lo que siembra un clima de pesimismo y muchos abandonan el campo de la traducción automática: "the more we shall know about linguistic structure, the more complex the description of this structure will be" (Bar-Hillel *apud* Hutchins 2000b:309).

El conocido como *informe ALPAC* ("Automatic Language Processing Advisory Committee") vendrá a ser la "puntilla" a casi dos décadas de trabajo en traducción automática. Ante la falta de logros y dado el alto nivel de inversión que se estaba realizando, el gobierno estadounidense encargó un estudio sobre el estado de esta tecnología. Los resultados, publicados en 1966, van a ser concluyentes: la traducción automática era más lenta, menos exacta y el doble de costosa que la traducción humana, por lo que resultaba infructuoso seguir invirtiendo en ese terreno hasta que no se fijaran metas más modestas y no se dispusiera de una mayor fundamentación teórica en LC[59].

Las conclusiones de este informe fueron decisivas, pues, para el futuro no solo de la traducción automática, sino de la propia LC. Básicamente, hizo explícito lo que para muchos ya era evidente: las expectativas habían desbordado con creces la realidad del momento: los ordenadores con los que se contaba no disponían de la velocidad y potencia de procesamiento ni de la capacidad de almacenamiento suficientes para soportar sistemas mínimamente funcionales; los lenguajes para programar los ordenadores eran muy rudimentarios –código máquina–, resultaban inadecuados para el tratamiento del lenguaje y totalmente incomprensibles para los lingüistas; por otra

[59] *Cf.* HUTCHINS (1996) para un análisis detallado de este informe.

parte, estaba el problema del conocimiento del mundo, para el que la Lingüística no ofrecía ninguna solución. Y es que si queremos simular la conducta lingüística en un ordenador, objetivo último de la LC, es preciso "tomar conciencia tanto de las estructuras propias del lenguaje, como del conocimiento general acerca del universo del discurso**"** (Gómez Guinovart y Palomar 1998:3), lo que implica previamente i) poseer los conocimientos sobre el funcionamiento del lenguaje; ii) disponer de una teoría formal que describa de forma rigurosa y sin ambigüedades las estructuras del mismo, iii) así como contar con mecanismos para representar los conocimientos generales que poseemos las personas y que hacen posible la comunicación. Todo esto será el objetivo prioritario de los subsiguientes trabajos en LC y motivo de que esta dirija su mirada hacia otras disciplinas, como la Inteligencia Artificial, en su búsqueda de fundamentación.

1.5.2. Primera etapa: años cuarenta y cincuenta

Así pues, entre los años cuarenta y los cincuenta se sientan los fundamentos teóricos de la Informática como ciencia de la mano de A. Turing –cuyas ideas tendrán repercusión también en los campos de la Filosofía, de la Inteligencia Artificial o de la Ciencia Cognitiva. Sin embargo, antes de este momento ya se había especulado con la posibilidad de utilizar algún tipo de mecanismo electromecánico para tareas relacionadas con el lenguaje: traducción, consulta de diccionarios, etc. Pero fue la aparición del ordenador digital la que actuó como detonante de toda una serie de líneas de investigación muy próximas, que van a ver en él una nueva herramienta para probar sus teorías sobre el funcionamiento del cerebro y del lenguaje.

Estos primeros investigadores, en su mayoría matemáticos e ingenieros, van a abordar la tarea de usar los ordenadores para tratar el lenguaje dominados por la euforia y el optimismo del momento, desconocedores de los verdaderos problemas que se iban a encontrar, pues los ordenadores hasta entonces únicamente se habían empleado como grandes calculadoras y nunca para tratar las lenguas naturales. Por otra parte, los lenguajes de programación de que disponían tampoco eran especialmente adecuados para trabajar con el lenguaje.

Para Jurafsky y Martin (2000:10-11), dos hechos fundamentales ocurren en esta primera etapa:

1) Se sientan las bases del paradigma estadístico para el tratamiento del lenguaje, que tendrá su mayor auge a partir de los noventa. C. Shannon, matemático e ingeniero estadounidense, propone la conocida como "teoría de la información" (1948), en la que aboga por un modelo estadístico de la comunicación, interesado en la forma de transmitir información de la manera más eficiente posible. Entiende la comunicación como un proceso estocástico en el que la información semántica no desempeña papel alguno: codificación de la información–canal de transmisión-decodificación de la información. Si bien las mayores repercusiones van a tener lugar en la Ingeniería de Telecomunicaciones, la Psicología y la Lingüística también van a hacer uso de este modelo que explica la comunicación en términos de codificación / decodificación.

2) N. Chomsky (1956), adoptando conceptos de la Lógica y las Matemáticas, lleva a cabo los primeros trabajos teóricos sobre la formalización de las lenguas naturales, cuyo resultado es la conocida como "jerarquía de Chomsky", base de la teoría de los lenguajes formales, uno de los pilares de la informática moderna, y que será clave

en el desarrollo de su teoría generativa del lenguaje solo un año más tarde.

Con estas premisas, llegan las primeras aplicaciones de los ordenadores en relación con el lenguaje, marcadas claramente por el contexto bélico de la II Guerra Mundial y, sobre todo, de la guerra fría. En concreto, va a ser la traducción automática la que suscite el interés de los investigadores que, aprovechando las técnicas criptográficas de base estadística empleadas durante la guerra para cifrar y descifrar mensajes en clave, piensan que el proceso de traducción de una lengua a otra es algo parecido.

Aunque la traducción automática fue la principal aplicación, también asistimos a los primeros trabajos en reconocimiento del habla, gracias a los avances en fonética instrumental. Así, los laboratorios BELL, en 1952, probaron un sistema, basado en estadísticas, capaz de reconocer los números del 0 al 9, pronunciados por un locutor, con un porcentaje de éxito del 97-99%, tomando como punto de referencia la frecuencia de los dos primeros formantes vocálicos.

Por último, hay que reseñar que en 1956 surge una nueva disciplina, la Inteligencia Artificial, auspiciada por J. McCarthy, M. Minsky, C. Shannon y N. Rochester, cuyas aportaciones al campo de la LC van a ser considerables. En líneas muy generales, y simplificando en extremo, se puede decir que el surgimiento de la IA fue el producto de una serie de investigaciones que se estaban llevando a cabo de forma paralela en distintos ámbitos del saber y que compartían el interés por la mente humana, ya fuera en la vertiente de proponer leyes que dieran cuenta del pensamiento (lógica), de formular principios matemáticos que recogieran esas leyes o de construir máquinas que simularan la actividad cognitiva.

A. Newell y H. A. Simon (*cf.* McCorduck 1991:56) señalan que en los estudios psicológicos existía un vacío, por lo menos en Estados Unidos, en lo que a la investigación sobre los procesos cognitivos complejos se refería, desde la muerte de W. James (1842-1909), quien había intentado introducir un acercamiento empirista en psicología. Ese acercamiento se llevó a cabo, pero no desde el ámbito de la Psicología, sino desde el de una nueva ingeniería, la Cibernética. N. Wiener vislumbró las analogías que podían establecerse entre los mecanismos electrónicos del momento y los biológicos, partiendo de las ideas y conceptos de la teoría de la información de C. Shannon, quien había aplicado la lógica de Boole a los circuitos eléctricos.

Si la lógica se podía automatizar en unos circuitos eléctricos, ¿por qué no se iba a poder automatizar la inteligencia mediante el uso de circuitos eléctricos y de la lógica? Fueron el neurofisiólogo W. McCulloch y el matemático W. Pitts quienes se encargaron de describir en términos lógico-matemáticos el comportamiento neuronal, al diseñar unas "neuronas electrónicas" en 1943, abriendo de esta forma el camino a la simulación de la cognición humana utilizando sistemas de neuronas artificiales. Sin embargo, con la aparición del ordenador, el planteamiento que se siguió para simular los procesos cognitivos fue otro muy diferente, es decir, no se intentó simular el nivel neurológico, sino el simbólico.

Todas estas líneas de investigación, íntimamente relacionadas entre sí, confluyeron en el Dartmouth College, oficialmente *The Dartmouth Summer Research Project on Artificial Intelligence* (Hanover, New Hampshire), un ciclo de conferencias sobre nuevas perspectivas de investigación en Informática. En el verano de 1956 se reunieron en el campus de la citada universidad una serie de investigadores que,

aunque procedentes de campos distintos[60], coincidían en un mismo interés: los ordenadores y su capacidad para simular la inteligencia humana en alguna de sus diferentes facetas. En total eran diez, entre los que destacamos a: J. McCarthy, principal promotor del encuentro y quien dio nombre a la disciplina, M. Minsky, N. Rochester, C. Shannon, A. Newell y H. A. Simon. Otros participantes fueron: Samuel, quien había diseñado un programa para jugar a las damas capaz de aprender de su propia experiencia; Bernstein, diseñador de un programa para jugar al ajedrez; etc.

La conclusión a la que se llegó fue un tanto decepcionante, y es que se dieron cuenta de que hacer que los ordenadores piensen, diseñar programas de ordenador que se comporten de manera inteligente, era mucho más difícil de lo que esperaban. Requisitos previos para este nuevo acercamiento eran: la colaboración de psicólogos, lingüistas, neurólogos, biólogos, etc. y el diseño de herramientas informáticas adecuadas.

Sin embargo, sus repercusiones se hicieron notar en el ámbito de la Informática, al introducir un cambio significativo en la concepción de estas máquinas. Con anterioridad a su nacimiento, la Informática se regía por el conocido como "régimen Lovelace", que debe su nombre a Ada Lovelace, hija de Lord Byron y una de las primeras programadoras de la historia. De acuerdo con esta concepción, una máquina solo puede hacer lo que se le diga que haga. Sin embargo, desde el surgimiento de la IA, se considera que las máquinas pueden hacer eso, pero también aprender. No existe ninguna forma de predecir cuál será el comportamiento a largo plazo del programa, en el caso de programas muy complejos, ya que su potencia de cálculo excede las posibilidades del cerebro humano.

---

[60] Matemáticas, Física, Psicología, Lógica, Ingeniería electrónica, Cibernética, Economía, etc.

Por lo que respecta al ámbito de la Lingüística, las implicaciones de esta nueva disciplina no fueron menores. Pronto el tratamiento y simulación del lenguaje con medios informáticos se convirtió en objeto de una recién nacida área de la Lingüística Aplicada, la *Lingüística Computacional*, aunque estos primeros intentos por tratar el lenguaje con medios informáticos pronto pusieron de manifiesto que la tarea era mucho más complicada de lo que se pensaba en un principio y que era necesario contar con un modelo teórico sobre el lenguaje. Este se va a buscar en la Lingüística del momento.

1.5.3. Segunda etapa: años sesenta

Precisamente en los años sesenta el lingüista y pensador N. Chomsky va a revolucionar el panorama lingüístico mundial al proponer una teoría del lenguaje que aboga por la formalización, inspirándose en los lenguajes artificiales de la lógica y las matemáticas: la gramática generativa, según la cual con un número finito de reglas es posible generar los infinitos enunciados de una lengua. Los trabajos de N. Chomsky publicados en este período (*Syntactic Structures* en 1957 y *Aspects of a theory of language* en 1965) van a dar un giro a la Lingüística Teórica, pero en LC su verdadera influencia no se dejará sentir hasta los años setenta.

Chomsky, sin embargo, no concibió su teoría generativa pensando en su implementación en ordenadores y, como pronto se probó, tampoco resultaba adecuada, pero la idea de formalizar la gramática resultó atractiva, puesto que la Lingüística, la ciencia del lenguaje, era el punto de referencia teórico que se necesitaba. Además, como señala Kay (2003), la propia Lingüística veía en los ordenadores una forma de introducir consistencia en sus teorías al poder probarlas. De hecho, uno

de los primeros sistemas para el análisis sintáctico lo constituye TDAP ("Transformations and Discourse Analysis Project"), desarrollado por Z. Harris, maestro de Chomsky, en la Universidad de Pensilvania en 1959.

Por otra parte, la importancia concedida al componente sintáctico en las primeras versiones de la gramática generativo-transformacional provocará que en estos años se ponga especial énfasis en el procesamiento del lenguaje basado en la sintaxis y se descuiden los aspectos semánticos.

Para Jurafsky y Martin (2000:11-12), estos trabajos de Chomsky en lingüística y teoría de los lenguajes formales, junto con las aportaciones del recién creado campo de la IA, constituyen el "paradigma simbólico", basado en reglas, de la LC, mientras que en los departamentos de estadística e ingeniería electrónica se va a consolidar otro paradigma bien distinto, el "paradigma estocástico", basado en estadísticas.

En lo que a las aplicaciones se refiere, la traducción automática va a seguir siendo la actividad predominante (*vid.* Hutchins 1986, 2000a, 2000b, 2001; Hutchins y Somers 1995 [1992]):

- Se trabaja sobre todo en los pares ruso-inglés y ruso-francés y, secundariamente, en los pares alemán-inglés, alemán-francés, inglés-francés.
- En general, se desconfía de las teorías lingüísticas de Harris y Chomsky, y se prefieren métodos basados en estadísticas o el método directo.
- Destacan el sistema GAT ("Georgetown Automatic Translation"[61]) –el sistema de IBM para las Fuerzas Aéreas de

[61] *Vid*. HUTCHINS (2005).

> EE.UU. que traducía del ruso al inglés mediante el método directo- y el primer sistema CETA ("Centre d'Études de la Traduction Automatique"[62]), para traducir del ruso al francés.

Sin embargo, los resultados no acababan de ser los esperados y el gobierno estadounidense, debido a las elevadas inversiones que estaba realizando en traducción automática, encargó en 1964 un informe, el famoso informe ALPAC ("Automatic Language Processing Advisory Commitee"), a la Academia Nacional de las Ciencias. Los resultados, publicados en 1966, van a ser concluyentes: dados los conocimientos teóricos del momento y las limitaciones técnicas de los ordenadores y lenguajes de programación, resultaba infructuoso seguir invirtiendo en ese terreno hasta que no se dispusiera de una mayor fundamentación teórica (*vid.* Hutchins 1996).

Como consecuencia, se produjo un drástico descenso de la financiación en traducción automática, área que no volverá a resurgir hasta finales de los setenta, aunque, por otra parte, el informe fue favorable para el campo en general, ya que implicó una revisión crítica del trabajo hecho hasta entonces.

Pero la traducción automática no fue la única aplicación de la LC en esta etapa. El desarrollo de interfaces de cara a obtener una mejor comunicación persona-ordenador en lengua natural ocupó también a algunos de estos primeros investigadores. Las interfaces son programas informáticos que permiten que las personas se puedan comunicar con los ordenadores mediante el uso de una lengua natural, p. ej., español,

---

[62] *Cf.* URL: http://www-clips.imag.fr/geta/historique/. Es interesante destacar que este sistema, partiendo de la imposibilidad de abordar el nivel semántico, se concentró en el tratamiento en profundidad de la sintaxis, prestando especial atención a la formalización, y se erigió en el primero en intentar un acercamiento basado en la interlengua (*vid*. Hutchins 1986:§5).

sin necesidad de acudir a un lenguaje artificial, p. ej., un lenguaje específico para obtener respuestas de una base de datos. No obstante, en los sistemas en que se concretaron las investigaciones de este período, también denominados "sistemas de primera generación", el análisis de las estructuras lingüísticas efectuado por los programas era mínimo. La técnica predominante era el "pattern-matching": identificación de palabras clave en el texto, que estaban asociadas a plantillas que incluían respuestas predeterminadas por parte del sistema. Es decir, ante una pregunta planteada por el interlocutor humano, el programa buscaba palabras clave para proporcionar una respuesta previamente asignada a esa palabra, de tal forma que, aunque en apariencia era capaz de responder y crear la ilusión de que "entendía" lo que se le decía, no llevaba a cabo ningún proceso de análisis del lenguaje.

Entre estas primeras interfaces, únicamente capaces de responder preguntas sobre el dominio que constituía su base de datos[63], destacan:

- BASEBALL, sobre la liga americana de béisbol;
- SAD-SAM, abreviatura de "Sentence Appraiser and Diagrammer-Semantic Analyzing Machine", programa que analizaba frases sobre relaciones de parentesco y las representaba en forma de árbol;
- SYNTHEX, "SYNTHesis of complEX verbal behavior", y PROTOSYNTHEX, un sistema diseñado para proporcionar información a partir de una base de datos textual del inglés constituida por todas las palabras de una enciclopedia;

[63] *Vid.* TENNANT (1981) para una descripción más detallada de estos sistemas.

- DEACON, "Direct English Access and CONtrol", en el dominio de discurso de la localización de unidades militares, comandantes, etc.;
- STUDENT, en el campo del álgebra;
- Etc.

Pero con diferencia es el siguiente sistema el que marcó un punto de inflexión. Aparentemente capaz de sostener una conversación similar a la que podrían mantener un psiquiatra y su paciente, no tiene en cuenta, sin embargo, el significado, sino que se basa en la identificación de palabras clave a las que están asociadas determinadas plantillas con posibles respuestas. Es decir, en realidad no existe un tratamiento del lenguaje, pero es uno de los programas que, mediante la ilusión de inteligencia que genera, más atención ha recibido, entre otras razones porque siempre genera una respuesta. Se trata del programa conocido como DOCTOR o ELIZA, desarrollado por Weizenbaum en 1966[64]. Es la plasmación perfecta del test de Turing. De hecho, como señalan D. Jurafsky y J. H. Martin (2000:8), muchas de las personas que interactuaron con ELIZA pensaron que el programa realmente les comprendía a ellos y sus problemas. Y en una competición que se celebra desde 1991 conocida como *Loebner Prize*, en la que distintos programas intentan pasar el test de Turing, siempre se ha dado el caso de que algunos de los jueces son "engañados" por los programas más toscos, aunque solo sea durante unos minutos. Un ejemplo del diálogo típico que puede sostener ELIZA es el siguiente[65]:

[64] Sitio web para dialogar con Eliza: http://www-ai.ijs.si/eliza/eliza.html
[65] Tomado de JURAFSKY y MARTIN (2000:32).

User$_1$:	*Men are all alike.*
ELIZA$_1$:	*IN WHAT WAY*
User$_2$:	*They're always bugging us about something or other.*
ELIZA$_2$:	*CAN YOU THINK OF A SPECIFIC EXAMPLE*
User$_3$:	*Well, my boyfriend made me come here.*
ELIZA$_3$:	*YOUR BOYFRIEND MADE YOU COME HERE*
User$_4$:	*He says I'm depressed much of the time.*
ELIZA$_4$:	*I AM SORRY TO HEAR YOU ARE DEPRESSED*.

Junto a la traducción automática y las interfaces, la recuperación de información, otra de las principales aplicaciones de la LC, inicia su andadura, aunque fuera del marco de la LC. Y continúan las investigaciones en torno al procesamiento del habla, financiadas principalmente por la agencia estadounidense ARPA, luego llamada DARPA ("Defence Advanced Research Projects Agency").

Asimismo, hay que reseñar que en esta etapa se compila el primer corpus: el *Brown Corpus of American English* (1963-1964), lo que supone el inicio de una tendencia de base empírica para el estudio del lenguaje que hoy en día constituye una disciplina autónoma: la lingüística de corpus.

Por último, los ordenadores también se aplicaron en Lingüística y Filología para tareas relacionadas con análisis cuantitativos y estadísticos: identificar las palabras presentes en un texto, ordenarlas según distintos criterios, calcular su frecuencia de aparición, etc.

En todos los casos, sin embargo, se encontraron dificultades insalvables, por lo que hubo que esperar a la llegada de la siguiente etapa, mejor fundamentada desde el punto de vista teórico y más consciente de la complejidad inherente al lenguaje y a su simulación computacional, algo que en un principio, debido al entusiasmo inicial, se había obviado. Se pensó que bastaban una serie de reglas generales

para dar cuenta del lenguaje. Pero tanto los sistemas de traducción automática como las interfaces pusieron en evidencia la necesidad de un tratamiento semántico, lo que va a dar origen a una línea de investigación que otorgará preeminencia al procesamiento semántico sobre el sintáctico y que, por lo tanto, se va a desviar de los postulados de la Lingüística del momento. Por otra parte, si se querían lograr aplicaciones prácticas, además del conocimiento lingüístico se necesitaba dar cuenta de otro tipo de conocimiento más general, el conocimiento del mundo.

1.5.4. Tercera etapa: años setenta

Durante la década de los setenta, asistimos a una etapa de consolidación en la que se tratarán de paliar las deficiencias observadas en las anteriores. Al optimismo de las fases previas le sucede un período de realismo. Se toma conciencia de la complejidad del lenguaje y las investigaciones se diversifican para intentar cubrir todas sus facetas: sintaxis, semántica, pragmática,… Además, ante la dificultad de tratar el lenguaje en general, los trabajos se restringen a dominios concretos o sublenguajes: estructuras sintácticas y contenidos semánticos empleados en un campo temático muy limitado (rocas lunares, bloques geométricos, etc.). Un único objetivo guiará a los investigadores de esta etapa: demostrar la viabilidad de la simulación computacional del lenguaje.

En primer lugar, hay que destacar los avances tecnológicos que se producen en lo que a la capacidad y potencia de los ordenadores se refiere, así como el desarrollo de nuevos lenguajes de programación más adecuados para el tratamiento del lenguaje: a finales de los cincuenta, J. McCarthy, el padre de la IA, había sentado las bases de

Lisp, un lenguaje de programación de alto nivel basado en la lógica; a principios de los setenta, A. Colmerauer, profesor de Informática de la Universidad del Mediterráneo en Marsella, creará un lenguaje específico para tratar las lenguas naturales, Prolog, que se ha convertido en el lenguaje de programación más extendido en IA.

Desde el punto de vista teórico, a la Lingüística se le sumarán las aportaciones procedentes del nuevo campo de la Inteligencia Artificial, cuya influencia va a ser grande durante la década de los setenta. La gramática generativo-transformacional, pese a cumplir el requisito de formalidad, resultó inadecuada para el tratamiento computacional del lenguaje dada la insuficiente cobertura que prestaba a los fenómenos semánticos, que ya se habían revelado como fundamentales en la etapa previa. Por otra parte, se constató la importancia que desempeñaba el conocimiento del mundo a la hora de interpretar palabras, frases y textos enteros. Precisamente los problemas relacionados con la representación del conocimiento (estructuras de casos, marcos, redes semánticas, primitivos semánticos, etc.) eran el centro de las investigaciones en IA, por lo que aquí se encontró la fundamentación idónea para elaborar formalismos adecuados para representar el significado, pero también para abordar aspectos sintácticos.

Con estas premisas, en lo que a la sintaxis se refiere, tres modelos, alternativos a la gramática generativo-transformacional, centraron los intereses de los investigadores (*cf.* Ramsay 1991:31-33):

a) Adaptaciones de la gramática de casos de Fillmore, un modelo al que la lingüística del momento no prestó demasiada atención, pero que en IA resultaba útil pues establecía una relación entre los papeles semánticos, conceptos de representación del conocimiento y la estructura sintáctica de las oraciones.

b) Modelos que minimizaban el papel de la sintaxis, como los propuestos por la Escuela de Yale; p. ej. la teoría de la *dependencia conceptual* de Schank.

c) Modelos que trataban de ampliar el poder de las gramáticas de estructura de frase con la incorporación de mecanismos procedentes de lenguajes de programación, como las *redes de transición aumentadas* de Woods, formalismo gramatical que incorpora nociones del lenguaje de programación Lisp y que, como dice R. Grishman (1991 [1986]:83), "ha llegado a ser una de las formas más populares de escribir gramáticas de lenguas naturales", lo que desde luego es cierto para esta etapa.

Además de teorías sintácticas, también se desarrollaron programas informáticos para aplicar esas teorías: son los algoritmos de *parsing*, *parsers* o programas para llevar a cabo el análisis sintáctico. Los más importantes son los que siguen una estrategia descendente o *top-down* y los ascendentes o *bottom-up*, cada uno de los cuales presenta ventajas y desventajas (*vid.* Grishman 1991 [1986]:41-47), por lo que generalmente se suelen combinar entre sí y con otros mecanismos, como el *backtracking*, la tabla de subcadenas bien formadas y los *charts*.

Por lo que a la semántica se refiere, hay que partir del hecho de que los conceptos de "significado" manejados en IA y en Lingüística son totalmente diferentes. En IA lo que interesa es que el sistema sea capaz de responder, ya que si genera una respuesta significa que "ha comprendido" la orden que se le ha dado. Ahora bien, el sistema responde de acuerdo con los conocimientos que posee. Estos suelen estar codificados en un lenguaje interno de representación o interlengua, un lenguaje formal exento de las ambigüedades propias

del lenguaje natural. Las propuestas principales en este sentido fueron tres (*cf.* Ramsay 1991:34):

a) Utilizar la lógica como lenguaje de representación (*redes semánticas*, *semántica formal*); las redes se relacionan con los trabajos en psicología de Quillian sobre la organización de los conceptos en la memoria humana (1968) y en semántica generativa.

b) Usar lenguajes de programación (*semántica procedimental*).

c) Servirse de primitivos semánticos (*dependencia conceptual*).

El interés por la semántica lleva a ampliar las miras hacia el terreno del discurso y el diálogo, así como el de la pragmática. Uno de los temas más tratados fue la identificación y asignación de referentes pronominales y otros elementos anafóricos o catafóricos, pero a veces intervienen conocimientos generales o inferenciales a la hora de determinar los referentes, y no información codificada lingüísticamente, por lo que conceptos tales como el de foco, actos de habla, creencias, intenciones, etc. fueron objeto de intentos de formalización.

En cuanto al conocimiento del mundo, sobresalen los trabajos de la Escuela de Yale, formada en torno a la figura de R. Schank, quien aplicó al lenguaje la teoría de los marcos o "frames" que M. Minsky (1975) había desarrollado para la visión artificial. Según M. Meya (1980:155-156), se trata de una "teoría unitaria y global de los procesos cognitivos del ser humano, y por tanto aplicable al lenguaje como vehículo de expresión de estos mecanismos". El marco es una estructura de datos que recoge situaciones estereotipadas, situaciones en las que nuestra experiencia acumulada nos proporciona unas estructuras de

conocimiento en las que encajar o con las que relacionar las nuevas situaciones.

Destacan también los estudios que la citada Escuela de Yale llevó a cabo sobre la comprensión y la organización de la información en la memoria, los primitivos semánticos o conceptos básicos a los que se puede reducir el conocimiento, y la teoría de la dependencia conceptual, lo que plasmaron en su sistema SPINOZA, que combina los primitivos semánticos (once acciones primitivas) con conceptos tomados de la gramática de casos de Fillmore (agente, acción, objeto).

Esta misma Escuela desarrolló la noción de guión o "script" como una manera de codificar conocimientos de tipo general, en este caso, la serie de actos que conforma una situación estereotipada compartida por hablante y oyente, para generar e interpretar historias relacionadas con los mismos. Su modelo no estuvo exento de críticas, por la dificultad de implementarlo en un ordenador y de determinar los elementos primitivos que pudieran dar cuenta de los significados presentes en las lenguas naturales.

En cuanto a las aplicaciones, se trata de prototipos de laboratorio o "toy systems", sin interés práctico. Su única finalidad era demostrar la posibilidad de tratar el lenguaje, aunque ya en este período, como señala M. F. Verdejo (1995:63; Verdejo y Gonzalo 1998:30), "empiezan a surgir los sistemas que después se convertirán en productos comerciales". En su mayoría se trataba de interfaces o sistemas de diálogo hombre-máquina, la aplicación predominante de esta etapa, aunque también asistimos a nuevas aplicaciones, como las ayudas informáticas para el aprendizaje y los sistemas de procesamiento de información textual. Son, básicamente, sistemas basados en la semántica, en la psicolingüística y en las teorías cognitivas. Se atiende, por tanto, a aspectos textuales y extralingüísticos. En este sentido, la

comprensión equivale a "determinar qué función tiene un objeto del mundo real o mundo posible, y qué relación tienen las partes con el todo" (Meya 1980:150). Algunos de esos sistemas son:

- MARGIE, "Meaning Analysis, Response Generation, and Inference in English", sistema que a partir de un enunciado de entrada genera una representación semántica acorde con la teoría de la dependencia conceptual.
- SAM, "Script Applier Mechanism", que, dada una historia, la descompone conceptualmente en primitivos semánticos y después busca guiones relacionados con los conceptos; asimismo puede realizar inferencias, responder a preguntas, elaborar un resumen de la historia o traducirla; etc.
- PAM, "Plan Applier Mechanism", que sigue el mismo funcionamiento que el sistema anterior, pero atendiendo a las metas y planes de los participantes en la comunicación.
- POLITICS, que hace uso de planes y guiones para representar distintas ideologías políticas.
- SOPHIE, "SOPHisticated Instructional Environment", ejemplo de sistema diseñado como ayuda en la enseñanza para los estudiantes de Electrónica.

No obstante, a finales de los setenta y principios de los ochenta, resurgen los sistemas de traducción automática, aunque en este período se desarrollarán principalmente fuera de EE.UU., en Canadá y Europa primero, y luego también en Japón, por lo que se tratarán nuevos pares de lenguas de acuerdo con las necesidades específicas de cada país, y no solo el par ruso-inglés. Además, se experimentará con nuevos acercamientos a la traducción, como el método de la interlengua

inicialmente o el de la transferencia hacia finales de los setenta. Algunos de los principales sistemas, que tendrán su continuidad en la década de los ochenta, son:

- Sistema *Météo*, de la Universidad de Montreal, para la traducción de partes meteorológicos entre inglés y francés.
- *Systran*, programa desarrollado por P. Toma para las fuerzas aéreas de EE.UU. y adoptado por la Comisión Europea para la traducción, también de inglés a francés, aunque con vistas a ampliar las lenguas.
- Proyecto *Eurotra* para la traducción multilingüe entre las diferentes lenguas de la Unión Europea.
- Programa *Ariane*, de la Universidad de Grenoble, basado en la transferencia.
- Programa SUSY, de la Universidad de Saärbrucken, también basado en la transferencia.
- Programa METAL, de la Universidad de Texas, con el mismo enfoque.
- Sistema *Mu*, de la Universidad de Kyoto, sobre los mismos principios, para el par japonés-inglés.

Sin embargo, durante esta tercera etapa dos son los sistemas que sobresalen del resto de aplicaciones por las repercusiones que han tenido y por iniciar lo que Jurafsky y Martin (2000:13) denominan "paradigma de la comprensión del lenguaje": LUNAR y SHRDLU, dos interfaces a bases de datos.

Dos científicos de la NASA, R. Kaplan y W. Woods (1973), desarrollaron el sistema LUNAR o LSNLIS, "Lunar Sciences Natural Language Information System", concebido para que pudiera responder, a partir de la información contenida en su base de datos, a las preguntas que los geólogos le plantearan a la NASA sobre las muestras de rocas recogidas en la Luna por el Apolo. Alcanzó un porcentaje de preguntas respondidas correctamente del 80%, que se elevó hasta un 90% tras algunas mejoras. Utilizaba una gramática de redes de transición aumentadas y una semántica procedimental.

Por su parte, T. Winograd (1972) desarrolló el que quizá sea el sistema más significativo, SHRDLU, basado en este caso en la gramática sistémica de Halliday y en una semántica de tipo procedimental. Simula el brazo de un robot que opera, según las instrucciones que recibe, en un universo de bloques geométricos dispuestos sobre una mesa. Este programa marca la transición hacia los sistemas de segunda generación, al ser el primero en conjugar temas como el análisis sintáctico, el análisis semántico, la identificación de referentes acudiendo al discurso previo, la representación del conocimiento, la solución de problemas, etc., y su respectiva interacción para comprender el lenguaje, eso sí, aplicados a un dominio muy restringido. Demostró que el nivel sintáctico estaba bastante dominado y que era hora de concentrarse en la semántica y en el discurso.

Un ejemplo del diálogo típico que era capaz de llevar a cabo el sistema es este (Winograd 1972:8-9):

1. Pick up a big red block.

OK.

2. Grasp the pyramid.

I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.

3. Find a block which is taller than the one you are holding and put it into the box.

BY "IT" I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

OK.

4. What does the box contain?

THE BLUE PYRAMID AND THE BLUE BLOCK.

5. What is the pyramid supported by?

THE BOX.

…

Por lo demás, el programa de T. Winograd se ajusta perfectamente a los objetivos de la IA:

> *This book describes a computer approach to the understanding of English. There are three kinds of reasons for writing a program such as the one it describes. First, there is the desire to build practical language-understanding systems which can increase the ability of computers to communicate with people. Second, underlying this is an attempt to clarify what language is and how it works in human communication. Finally, an understanding of language is a vital step in discovering the basic principles underlying intelligence* (Winograd 1972:1-2).

Por último, en lo que al tratamiento del habla se refiere, hay que mencionar la importancia que cobra el acercamiento basado en estadísticas aplicado al reconocimiento del habla, debido sobre todo a su empleo en el centro Thomas J. Watson de IBM[66], en la Universidad Carnegie Mellon[67] y en AT & T Labs Research[68].

En definitiva, lo más característico de los sistemas computacionales durante esta etapa era:

- Su inspiración en teorías lingüísticas: gramáticas computacionales basadas en modelos lingüísticos (gramática de casos, gramáticas semánticas, etc.).
- El interés por el tratamiento del nivel semántico.
- El afán por descubrir qué es comprender (orientación psicológica) y cómo hacer explícito el conocimiento general o del mundo.
- La no separación entre conocimiento lingüístico e informático: datos y programas forman un todo inseparable.
- La escasa capacidad para procesar textos incorrectos o incompletos.
- El inglés es la lengua que tratan la mayoría de los sistemas.
- Se refieren a dominios restringidos: se desarrollan ad hoc para ese dominio, por lo que carecen de portabilidad.
- Las aplicaciones son sistemas desarrollados en un laboratorio, sin repercusión en la sociedad.

---

[66] URL: http://www.watson.ibm.com/index.shtml
[67] URL: http://www.speech.cs.cmu.edu/
[68] URL: http://public.research.att.com/index.cfm?portal=1&h=1

- La integración en el marco del PLN y de la Inteligencia Artificial como consecuencia del giro en la base teórica del campo: “From the mid-1960s, most computational natural language work took place under the chapter of NLP [Natural Language Processing], suitably seen as a branch of AI [Artificial Intelligence]” (Roeck 1995:155).

1.5.5. Cuarta etapa: años ochenta

En general, durante los años ochenta, presenciamos un período de crecimiento y consolidación, con énfasis en la investigación básica, pero animado también por el logro de mejores sistemas y la búsqueda de resultados a nivel práctico.

Se vuelve a la lingüística como base teórica, con el desarrollo de nuevos formalismos lógico-gramaticales, más sencillos que la gramática generativo-transformacional, en los que se demuestra un gran interés por el léxico y su recopilación computacional. Por otro lado, cobra auge la lógica como lenguaje de representación del significado y del razonamiento en IA, con el consiguiente desarrollo de lenguajes de programación lógicos o de alto nivel, más adecuados para el tratamiento computacional del lenguaje.

Así es como surge toda una familia de *gramáticas* llamadas *de unificación*, que incorporan la operación de unificación del lenguaje de programación PROLOG. Son teorías lingüísticas directamente aplicables para el tratamiento computacional del lenguaje. La gramática de cláusula definida de Pereira y Warren en 1980 fue la primera de este tipo, a la que le han sucedido en la actualidad diferentes variantes:

gramática de estructura de frase generalizada, gramática léxico-funcional, gramática de unificación funcional, etc.

Por otra parte, las teorías lingüísticas están pensadas específicamente para su implementación informática, debido a su inspiración en lenguajes de programación. Además, también se retoman modelos descartados previamente, como los modelos de estados finitos, sobre todo en el nivel fonológico, morfológico y sintáctico.

La semántica léxica experimenta un gran desarrollo: descripción y organización del léxico y, en consecuencia, se elaboran léxicos computacionales. Este interés por el léxico está impulsado, a su vez, por los nuevos modelos lingüísticos, sobre todo a partir de la teoría de la rección y el ligamiento de Chomsky (1981), y por las necesidades derivadas de la traducción automática, que vuelve a renacer con fuerza, sobre todo gracias al impulso recibido desde la Unión Europea y Japón.

Algunos de los sistemas de traducción automática más representativos del momento son: *Taum-Météo*, que traduce de forma totalmente automática partes meteorológicos entre el inglés y el francés en Canadá, o EUROTRA, el proyecto de traducción multilingüe de la Comunidad Europea. Además, resurge el acercamiento conexionista, con la introducción de estadísticas, dentro de una tendencia general hacia el empirismo: uso de modelos basados en probabilidades como los empleados por IBM para el tratamiento del habla.

La lógica cobra fuerza como lenguaje de programación y de representación del significado: uso de lenguajes de programación declarativos, entroncados con las teorías lingüísticas.

Por otro lado, se intenta igualar la situación de otras lenguas en comparación con el inglés mediante el desarrollo de gramáticas y otros recursos computacionales.

Para K. Sparck Jones (1994:9) este es, además, un período de crecimiento y de consolidación, animado por el logro de mejores sistemas y resultados a nivel práctico. Alcanzado cierto dominio del nivel sintáctico, se tratan otros niveles, como el del discurso o el diálogo, y aparecen más trabajos sobre generación del lenguaje, una faceta hasta entonces apenas abordada.

Quizá este impulso global del campo, tanto a nivel teórico como práctico, es el que conduce a la citada autora, K. Sparck Jones, a observar síntomas de una cierta división entre aquellos que se centraban en cuestiones puramente científicas o de investigación básica y entre los que preferían concentrar sus esfuerzos en las aplicaciones prácticas. Es decir, con la consolidación de la disciplina los intereses de los investigadores se especializan y diversifican, y la parte aplicada va ganando terreno poco a poco. Así, empiezan a llegar al mercado los primeros productos comerciales (interfaces, sistemas de traducción automática y sistemas para el procesamiento textual). En este sentido, asistimos a la aparición de las denominadas *industrias de la lengua*, como consecuencia de la participación de gobiernos e instituciones públicas y privadas.

Como señala M.ª F. Verdejo (1995:65):

> Desde el punto de vista de las aplicaciones, en esta década [la de los 80] se ha iniciado una actividad comercial, han surgido compañías especializadas, grandes corporaciones han creado sus propios productos para sus necesidades específicas y se ha comenzado a hablar de un posible sector de Industrias de la Lengua.

1.5.6. Quinta etapa: años noventa

Por último, a partir de los noventa se observa un cambio de tendencia significativo. El paso en la etapa anterior de la actividad científica a la industrial había puesto de manifiesto la necesidad de una coordinación entre teoría y práctica si se pretendía lograr sistemas capaces de satisfacer las demandas del mercado. Por otra parte, se considera que los avances teóricos y las inversiones realizadas en la década previa no se han concretado en los avances esperados. Para M.ª F. Verdejo y J. Gonzalo (1998:30), las causas son tres:

> i) la ausencia de recursos léxicos generales (se había favorecido el trabajo en dominios específicos);
>
> ii) la nula reutilización de los recursos existentes; y
>
> iii) el elevado coste que supone la elaboración de recursos y aplicaciones a escala real[69].

Para paliar esta situación, en los noventa se tiende, según A. Moreno Sandoval (1998:45), por una parte, a la búsqueda de aplicaciones realistas, de sistemas prácticos (correctores gramaticales, ayudas para la traducción o traducción asistida), y, por otra parte, a la ampliación de la cobertura de los sistemas a cualquier tipo de texto y dominio, en un "claro giro hacia la parte más aplicada y comercial".

Surgen proyectos destinados a cubrir las carencias de recursos básicos: recursos terminológicos multilingües (proyecto *Interval*[70]), corpus y léxicos para catorce lenguas europeas (proyecto *Parole*[71]),

[69] Precisamente en 1992, A. Danzin presenta un informe en la Comisión Europea que aboga por otorgar prioridad a los recursos básicos.

[70] URL: http://www.computing.surrey.ac.uk/ai/new_interval/

[71] URL: http://www.ub.es/gilcub/castellano/proyectos/europeos/parole.html

tratamiento de voz (proyecto *Speechdat*[72]), redes léxico-semánticas para las principales lenguas europeas (proyecto *EuroWordNet*[73]). Y se desarrollan recursos a gran escala (léxicos y gramáticas) estandarizados e independientes de la aplicación: la mayor capacidad de los ordenadores permite procesar grandes cantidades de texto y favorece el empleo de una metodología empírica que utiliza grandes corpus textuales como fuente de datos que, por otra parte, puede procesar con técnicas estadísticas (*vid.* Manning y Schütze 1999). Además, existe una preocupación por la fijación de unas normas de estandarización: TEI ("Text Encoding Initiative"[74]), EAGLES ("European Advisory Group on Language Engineering Standards"[75]), requisito imprescindible para la reutilización de los recursos, otra de las preocupaciones de la etapa.

Otras características de este nuevo giro en la disciplina son: la mejora de los formalismos de la etapa anterior y un especial interés por formalismos adecuados para expresar la información léxica; a su vez, el tratamiento de la sintaxis y la semántica se beneficia del uso de corpus como fuente de datos; el procesamiento textual y la recuperación de información se erigen como áreas destacadas, sobre todo debido al auge de Internet; se introducen técnicas estadísticas, para poder tratar cantidades tan grandes de datos, que se combinan con las basadas en reglas para obtener mejores resultados; se busca ampliar la cobertura de los sistemas para que sean capaces de enfrentarse a cualquier tipo de texto, sin restricciones de dominio; se desarrollan sistemas orales en ámbitos como la traducción y las interfaces, en los que se produce una unión del tratamiento del habla y del texto, así como una preocupación creciente por la evaluación; por último, destaca el papel otorgado a los

[72] URL: http://www.speechdat.org/
[73] URL: http://www.illc.uva.nl/EuroWordNet/
[74] URL: http://www.tei-c.org/P4X/
[75] URL: http://www.ilc.cnr.it/EAGLES/home.html

recursos y a su reutilización, que ha provocado que se empiece a hablar de *ingeniería lingüística*:

> Otra tendencia aún inmadura, pero que constituye una preocupación creciente de empresas, investigadores y fuentes de financiación (Comunidad Europea y Agencias norteamericanas), es la de la reutilización, no solo de los recursos léxicos, sino también del software. Esta tendencia se ha reflejado en un cambio de nombre para la disciplina, que pasa poco a poco a conocerse como «Language Engineering» (Ingeniería Lingüística) (Verdejo y Gonzalo 1998:31).

Algunos ejemplos de los productos actuales del PLN son: *Babel Fish*, traductor automático de la casa *Systran* que opera en el buscador web *Altavista*, que recibe más de un millón de peticiones al día; el proyecto de traducción oral de *British Telecom* en el ámbito de los negocios[76]; *Candide*, sistema de traducción automática de IBM que se apoya en técnicas estadísticas; sistemas de corrección automática de exámenes; asistentes para el aprendizaje de la lecto-escritura[77], etc.

[76] *Vid.* HUTCHINS y SOMERS (1995 [1992]:433 y ss.).
[77] *Vid*. JURAFSKY y MARTIN (2000:9-10), o COLE (1996).

2. ÁREAS DE TRABAJO DE LA LC

2. ÁREAS DE TRABAJO DE LA LC

2.1. Áreas de la LC

Si bien en la práctica los objetivos teóricos y los aplicados mencionados en apartados previos tienden a confluir, es habitual por motivos expositivos dividir el estudio de la LC en dos vertientes, una teórica y otra aplicada, que se plasman en los distintos objetivos de la disciplina.

A lo largo de las siguientes páginas presentaremos una visión general de las principales tareas a las que se debe enfrentar la LC en ambos apartados de su quehacer, en especial en su faceta teórica.

Por lo que respecta al primero de ellos, las bases teóricas de la disciplina se sustentan sobre el hecho de que para elaborar programas informáticos capaces de simular la conducta lingüística es necesario conocer el funcionamiento del lenguaje. Así, igual que en la Lingüística se reconoce la existencia de diferentes áreas en el lenguaje, a la hora de reproducir la conducta lingüística en un ordenador –tanto en el sentido de la comprensión, reconocimiento o análisis como en el de la generación o síntesis– parece fundamental atender a los siguientes niveles básicos o módulos, para poder dar cuenta de toda la complejidad inherente a este objeto de estudio[78]:

1. *Conocimiento fonético y fonológico*. Se ocupa de las realizaciones acústicas, así como de su transcripción. Solo aparece en los sistemas que trabajan a partir del habla (tecnologías del habla).

[78] *Vid.* MORENO SANDOVAL (1998:34-35), MOURE y LLISTERRI (1996:173-174) o LAVID (2005:75).

2. *Conocimiento morfológico*. Trata la formación interna de las palabras. Hay sistemas que lo incluyen con el componente sintáctico o que no lo tratan. Incluirlo o no depende de la aplicación y de la lengua. Transforma la cadena de caracteres de entrada en una secuencia de unidades significativas o unidades léxicas haciendo uso del diccionario y de reglas morfológicas.

3. *Conocimiento sintáctico*. Se encarga de reconocer la estructura de las oraciones. Es un componente básico. Analiza la secuencia de unidades léxicas y produce una representación de su estructura en forma de árbol, red, etc.

4. *Conocimiento semántico*. Asigna significado a las estructuras. Es otro de los componentes básicos. A partir de la estructura generada por el módulo sintáctico genera otra estructura o forma lógica asociada que representa el significado de la oración.

5. *Conocimiento contextual* o *pragmático*. Da cuenta de la información no lingüística que influye en el procesamiento e interpretación. Utiliza la forma lógica o estructura semántica del componente anterior para desarrollar la interpretación final de la oración, en función de las circunstancias del contexto. Se suele subdividir en dos tipos de conocimiento:

> -*Conocimiento del discurso*: información proporcionada por los enunciados emitidos anteriormente, sobre todo de cara a interpretar los pronombres y referencias anafóricas, así como los aspectos temporales.
>
> -*Conocimiento del mundo*: informaciones de tipo general, a menudo implícitas o sobreentendidas.

Los modelos computacionales del lenguaje se basan en la idea de que los procesos complejos se pueden descomponer en *módulos*, relativamente independientes entre sí, que integran procesos más simples[79].

Cada nivel ofrece unas dificultades inherentes a la hora de su tratamiento computacional, motivo por el que se han postulado diferentes acercamientos para intentar dar cuenta de las características propias de las distintas facetas que presenta el lenguaje.

De forma gráfica, los conocimientos implicados en un sistema computacional relacionado con el lenguaje serían:

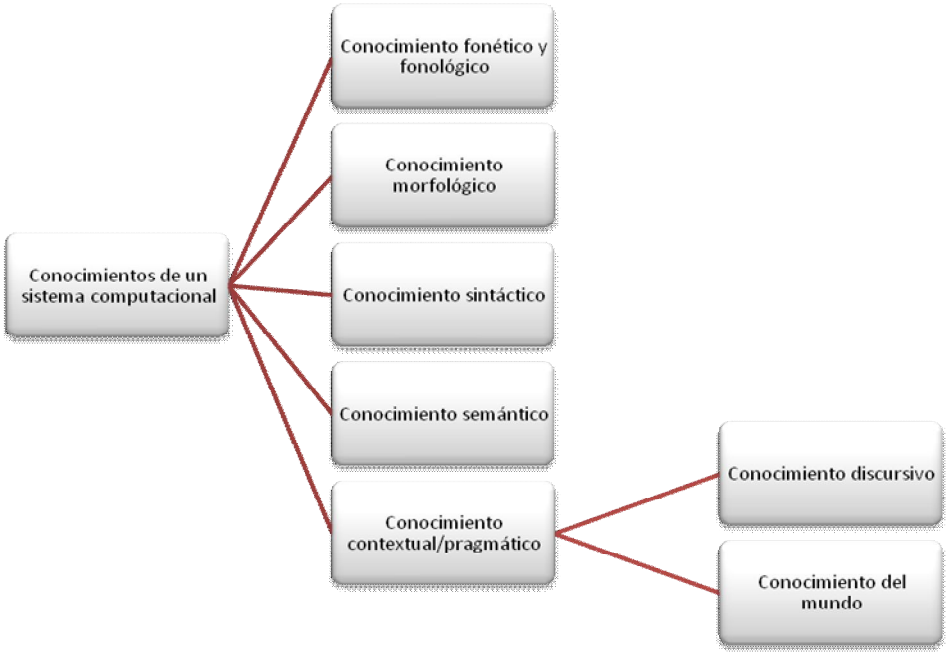


Ilustración 14. Conocimientos de un sistema computacional.

[79] Este planteamiento de partida no es trivial, sino que tiene importantes repercusiones prácticas posteriores. Así, en un sistema concreto para procesar el lenguaje, si este se ha concebido de forma modular, un error, problema, ampliación o modificación de una regla que afecte a un determinado módulo se puede subsanar o llevar a cabo de forma más sencilla, sin afectar al resto del diseño del sistema (*cf.* MORENO SANDOVAL 1998:33). Por lo tanto, la modularidad aporta flexibilidad y facilidades para la ampliación o exportación –a otros ámbitos o aplicaciones diferentes de las finalidades originarias– a los sistemas computacionales para el procesamiento del lenguaje natural.

Por lo tanto, desde la perspectiva teórica, la LC se ocupa del lenguaje en sus diferentes niveles y de cómo tratarlos con medios informáticos. Para ello:

- Elabora teorías que los describen ateniéndose a ciertos requisitos: formalidad, explicitud, no ambigüedad y facilidad de tratamiento computacional.
- Desarrolla programas que los implementan en un ordenador.

Es decir, para reproducir la conducta lingüística, su objetivo fundamental, previamente la LC debe disponer de descripciones adecuadas (teorías, modelos, formalismos) de cada uno de los niveles que conforman el lenguaje. No basta con que estas descripciones sean satisfactorias desde un punto de vista científico[80], sino que además deben ser adecuadas para su implementación en un ordenador, de ahí la necesidad de cumplir una serie de requisitos.

Para llevar a cabo su labor, la LC se beneficia de las aportaciones de la Lingüística Teórica, la Psicolingüística y la Ciencia Cognitiva, aunque también de otras disciplinas como la Lógica, las Matemáticas o la Inteligencia Artificial[81].

Sin embargo, el punto de partida se lo proporciona la ciencia del lenguaje por excelencia, la Lingüística, que divide el conocimiento lingüístico en una serie de niveles que constituyen las principales áreas de investigación en Lingüística y, en consecuencia, en LC.

Cada nivel se caracteriza por una serie de unidades, conceptos y procesos mediante los cuales la Lingüística intenta dar cuenta de las

[80] Lo que sería tarea propia de la Lingüística Teórica.

[81] Recordemos que, desde una perspectiva histórica, la Lingüística no siempre ha sido capaz de proporcionar las descripciones y técnicas de análisis que los trabajos en LC han requerido, motivo por el cual, además de por la naturaleza misma del objeto de estudio, han entrado en juego otros ámbitos del saber.

estructuras y conocimientos que intervienen en el lenguaje: fonemas, morfemas, oraciones, etc.

Partiendo de esta base, los sistemas computacionales que tratan el lenguaje incluyen módulos que abordan cada uno de los mencionados niveles, lo que da lugar a las distintas áreas de estudio dentro de la LC Teórica[82].

- *Fonología computacional*
- *Morfología computacional*
- *Sintaxis computacional*
- *Semántica computacional*
- *Pragmática computacional*

Los posibles módulos de un sistema computacional que quiera dar cuenta de una lengua natural serían:

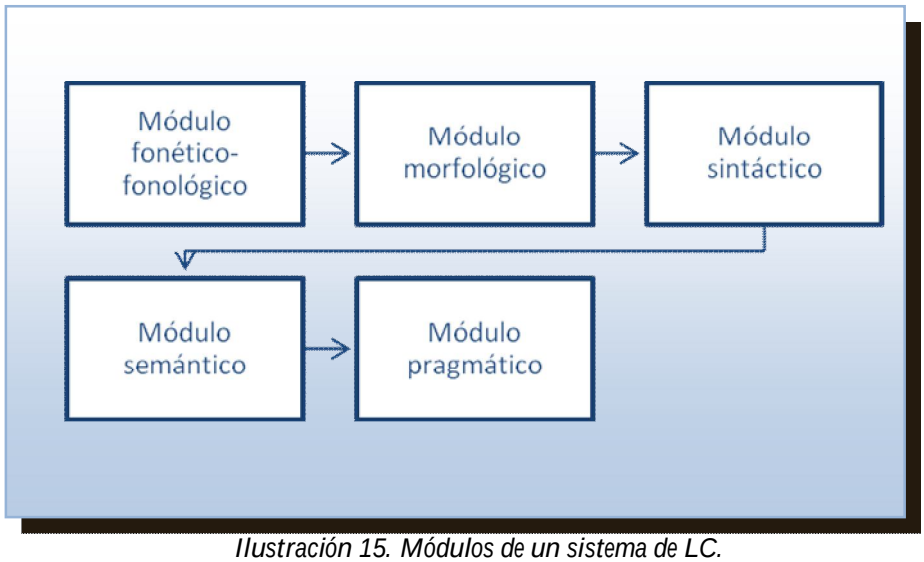


Ilustración 15. Módulos de un sistema de LC.

[82] *Cf.* GRISHMAN (1991 [1986]), KLAVANS (1997), MORENO SANDOVAL (1998), MARTÍ y CASTELLÓN (2000), JURAFSKY y MARTIN (2000) o MITKOV (2003), entre otros.

No obstante, hay que hacer notar que en la mayoría de casos los sistemas de procesamiento del lenguaje natural se centran en la lengua escrita, por lo que no suele ser necesario tratar el nivel fonético-fonológico, módulo que sí deben incorporar los sistemas relacionados con las tecnologías del habla, que son los que abordan de forma específica el tratamiento de la lengua oral.

Además, M.ª A. Martí e I. Castellón (2000:9) consideran que cualquier sistema computacional que trate el lenguaje natural ha de constar de tres componentes básicos:

a) *Textos* en una lengua natural, que son los que queremos procesar o generar.

b) *Datos lingüísticos*, normalmente gramáticas y lexicones, que indican cómo procesar o generar los textos.

   - El diccionario o lexicón es el encargado de recoger las unidades léxicas que pertenecen a la lengua en cuestión junto con la información necesaria para su procesamiento morfológico, sintáctico y semántico. Una mayor complejidad del componente léxico redundará en una simplificación de la gramática.

   - La gramática es la que recoge las reglas necesarias para i) determinar la gramaticalidad de los textos, y ii) efectuar el procesamiento de los mismos, es decir, mostrar su estructura morfológica, sintáctica, semántica, etc.

c) *Programas informáticos* que llevan a cabo el procesamiento o generación de los textos de acuerdo con la información que le proporcionan los datos lingüísticos.

Para cada sistema concreto, en función de su objetivo o de la aplicación que se le quiera dar, se determinan los módulos que precisa, la información que se recoge en cada uno y la forma de actuar de los componentes. Esto es posible gracias a la concepción modular o modularidad, hecho que confiere gran flexibilidad a los sistemas computacionales.

Véase, a modo de ejemplo, la estructura que proponen L. Moreno Boronat y A. Molina Marco (1999:18):

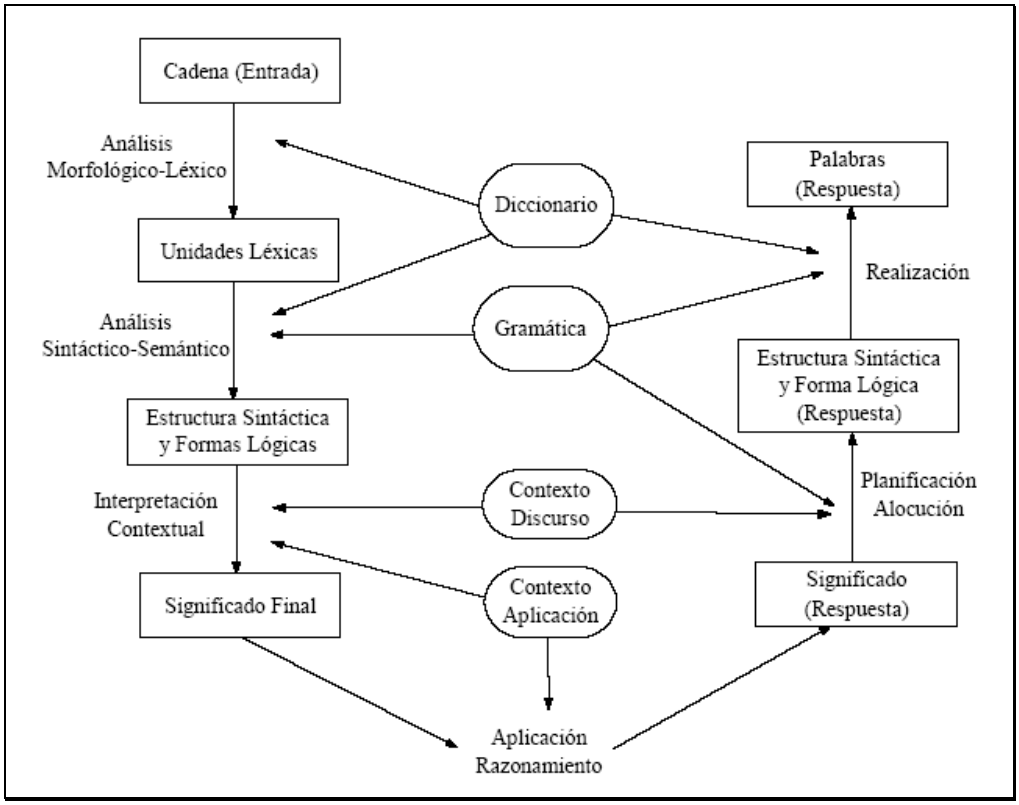


Ilustración 16. Estructura de un sistema de LC.

Para un sistema que trabaja con lengua oral, será preciso, además, uno o varios módulos que den cuenta del conocimiento fonético-fonológico[83].

Véase el esquema que sugiere X. Gómez Guinovart (2000b:85):

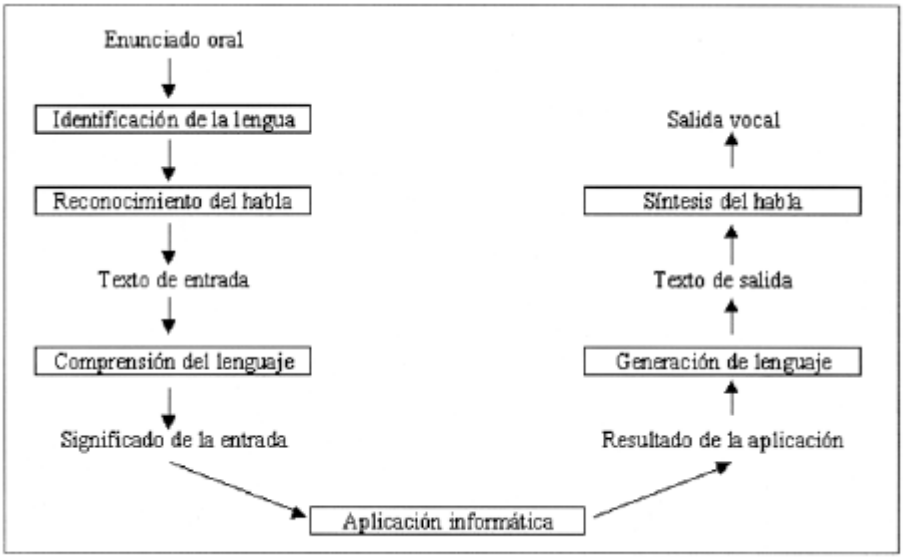


Figura 1: Interacción lingüística oral persona-ordenador

Ilustración 17. Integración del habla en un sistema computacional.

[83] En el ejemplo concreto que se comenta más abajo, se propone un módulo para la identificación automática de la lengua oral, imprescindible en contextos multilingües en los que las comunicaciones pueden realizarse en más de una variedad lingüística (p. ej. en un sistema de consulta telefónica de información en un entorno como la Unión Europea o en un país en el que conviven varias lenguas como España); en otros casos, lo que se precisa es la identificación o verificación de un locutor a partir de una muestra de su voz como paso previo al acceso a determinados servicios (p. ej. de banca telefónica). Identificada la lengua o el locutor –si es preciso–, el siguiente paso es el *reconocimiento del habla*, que suele consistir en la conversión de la onda sonora en su equivalente escrito o simbólico. A partir de esta representación simbólica, si es necesario acceder a los contenidos del mensaje, se aplican las técnicas habituales para tratar el texto escrito. Cuando lo que se quiere es que el sistema genere una respuesta hablada a una demanda del usuario, entonces, además de un módulo de reconocimiento, hay que incorporar uno de generación (*síntesis del habla*), que convierte un mensaje escrito en su equivalente oral, con independencia de que antes haya actuado un módulo de generación del lenguaje (escrito). Para una visión general sobre las tecnologías del habla, *vid.* LLISTERRI (2003, 2004).

De forma esquematizada, lo que se espera de un sistema computacional que trata el lenguaje es que, dado un texto de entrada, el sistema lleve a cabo su procesamiento y produzca un resultado, que variará en función de la aplicación concreta (p. ej. si estamos utilizando un programa de dictado automático, el texto de entrada será nuestra voz y el de salida, el texto que aparece en la pantalla de nuestro ordenador; en el caso del conjugador de un diccionario electrónico, como en el diccionario en línea de la Real Academia, el texto de entrada será el lema y el de salida, la conjugación del verbo que hayamos introducido; etc.). Este planteamiento general se puede trasladar a cada uno de los componentes o módulos, que actúan de la misma manera: texto de entrada → procesamiento computacional → texto de salida.

Por lo que respecta al tratamiento de la lengua escrita, si tomamos como referencia el esquema de L. Moreno Boronat y A. Molina Marco, observamos que se aborda el conocimiento morfológico, sintáctico, semántico y pragmático. A continuación presentaremos los conceptos básicos de algunos de estos niveles lingüísticos y las principales propuestas para su tratamiento computacional[84].

No obstante, hay que señalar que, ante un texto de entrada, en el caso del análisis, por lo general el tratamiento computacional del lenguaje no puede iniciarse de forma inmediata, sino que suele ser necesario pasar por una fase previa (*cf.* Rodríguez Hontoria 2000), que parte de la consideración del texto como una secuencia de caracteres que hay que

[84] H. RODRÍGUEZ HONTORIA (2000) presenta una visión general de las labores implicadas en el tratamiento computacional del lenguaje. En los siguientes manuales y artículos se puede encontrar información más detallada: GRISHMAN (1991 [1986]), JURAFSKY y MARTIN (2000), KLAVANS (1997), LAVID (2005), MARTÍ y CASTELLÓN (2000), MARTÍ (2001 y 2003), MITKOV (2003), MORENO BORONAT *et al.* (1999), MORENO SANDOVAL (1998), VIDAL y BUSQUETS (1996).

segmentar (en párrafos[85], oraciones, intervenciones de locutores), filtrar (desechar información no relevante, como ocurre cuando el texto procede de una página web: enlaces, imágenes, anuncios…), identificar como pertenecientes a una lengua[86] y, por último, localizar las unidades tratables, básicamente, las palabras ortográficas –aquellas delimitadas por espacios en blanco o signos de puntuación. Esta última tarea no está exenta de dificultades, debido a que, además del hecho de que existen lenguas que no se basan en la palabra como unidad ortográfica, en las lenguas que sí lo hacen no siempre existe una correspondencia perfecta entre dichas palabras ortográficas y las unidades gramaticales, como ocurre en el caso de las contracciones (*al, del*) o los átonos pronominales enclíticos (una palabra ortográfica representa dos palabras gramaticales), o con las formas compuestas, perífrasis verbales, conjunciones y locuciones conjuntivas (*no obstante, es decir*), locuciones verbales (*ponerse las pilas*), unidades léxicas multiplabra (dos o más palabras que forman una unidad semántica y/o sintáctica y que suelen coaparecer juntas: *dar los buenos días*). En todos estos casos nos encontramos con secuencias que constan de dos o más palabras ortográficas, pero que desde el punto de vista gramatical y/o semántico forman una unidad. Además, hay que señalar la presencia de nombres propios (*Wall Street*), términos procedentes de otras lenguas diferentes de la lengua objeto de tratamiento ("*A Guardiola le encanta que la* gent *blaugrana sea feliz"*, El País.com, 18/05/2009), siglas (*TVE*, *UE*), abreviaturas (*telecos*), cifras ("*el tesorero del PP recibió de la trama, siempre por orden de Correa, en torno a 650.000 euros"; 19º título de Liga del Barça*, El País.com, 18/05/2009) y fechas (*18/05/2009*), neologismos (*minitrasvase*), símbolos y caracteres especiales (*69%*) o simples errores ortográficos

[85] Más fácil si existen signos de puntación y se distingue entre mayúsculas y minúsculas, lo que no ocurre siempre, por ejemplo, en las transcripciones de la lengua oral.

[86] Especialmente importante en contextos multilingües.

que impiden identificar las unidades como pertenecientes a una lengua por no estar en los diccionarios al uso[87].

Por lo tanto, cuando se inicia el tratamiento del nivel morfológico, previamente ya se han efectuado diversas operaciones que tienen como meta proporcionar los datos de entrada –unidades elementales, por lo general palabras gramaticales– al módulo encargado de la morfología. Así, ante la presencia de p. ej. *del* en un texto, las unidades de entrada en el módulo serán *de* y *el*. De forma gráfica:

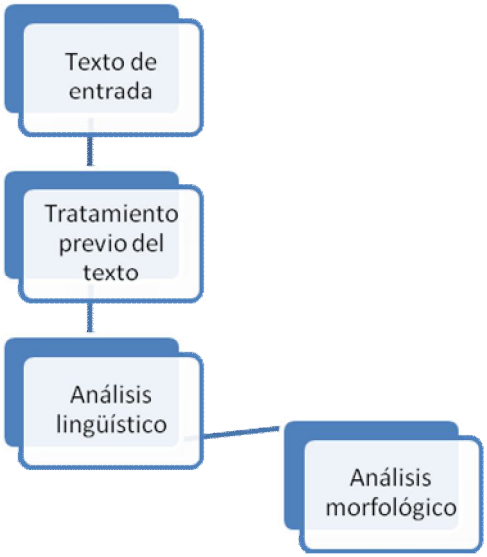


Ilustración 18. Pasos previos al análisis lingüístico.

[87] Existen técnicas para intentar afrontar estas dificultades, como es el uso de lexicones específicos o procesadores para extraer fechas, nombres propios, etc.

2.2. Morfología computacional

Como se puede observar en el gráfico de L. Moreno Boronat y A. Molina Marco (*vid. supra*), ante un texto de entrada (palabra, frase, etc.), el primer nivel lingüístico por el que comienza el tratamiento computacional del lenguaje es el morfológico.

Como es de sobra conocido, la Morfología es la parte de la Lingüística que se ocupa del estudio de la estructura interna de las *palabras*, su unidad superior, mientras que los *morfemas* o elementos constitutivos de las palabras son su unidad básica de trabajo.

Los morfemas son unidades mínimas (no se pueden dividir más sin entrar en otro nivel descriptivo) que poseen significado. Este oscila entre lo léxico (como suele ocurrir con las raíces de las palabras, p. ej. *roj-* en *rojo* es el elemento que aporta el contenido básico de la palabra) y lo gramatical (morfemas flexivos, que producen variaciones en una misma palabra, como se observa en *roj-o*, *roj-a*, *roj-o-s*, *roj-a-s*, y que proporcionan información adicional a la aportada por la raíz, como el género, el número, etc.), con multitud de grados intermedios a medio camino entre lo uno y lo otro (muchos morfemas derivativos, que sirven para formar nuevas palabras, como *en-* y *–ecer* en la palabra *enrojecer*) e, incluso, en ocasiones sin un contenido claro (casos de morfemas a los que resulta difícil asignarles un significado específico). Además, un único morfema puede constituir una palabra (p. ej. *sol*), o ser varios los que la formen (p. ej. *sol-es*); asimismo, hay palabras que son invariables en su forma (p. ej. *hoy*) y otras que la modifican (p. ej. *sal, sales*, *salgamos*); por último, según su estructura, se distingue entre palabras simples (p. ej. *azul*) y palabras complejas (p. ej. *azulgrana*)[88].

[88] *Vid.* PENA (1999) para una visión más detallada de las cuestiones referidas a las unidades de la Morfología.

Ahora bien, la Morfología no solo se ocupa de identificar y aislar morfemas, sino que además da cuenta, mediante reglas, de cómo se combinan entre sí para formar palabras en los procesos denominados:

- *Flexión*
- *Derivación*
- *Composición*

Para poder explicar estos procesos así como el análisis de la estructura interna de las palabras, es necesario tener en cuenta otras unidades y conceptos[89], además de las unidades ya mencionadas –*palabra* y *morfema*, unidad máxima y mínima respectivamente de la Morfología. Así, es normal distinguir dentro de los morfemas entre:

- *Lexema* o *raíz*: es el formante común, sin morfemas flexivos o derivativos, a un conjunto de formas léxicas o palabras. Es el elemento que aporta el significado básico de la palabra y, por lo tanto, la unidad fundamental del léxico de una lengua. P. ej. *am-* es la raíz de *amo, amas, amaba*…; *sol* lo es de *sol*, *soles*, etc. Normalmente es invariable, pero puede presentar modificaciones en función de los morfemas que lo sigan. P. ej. *dormir* presenta tres raíces diferentes: *dorm-*, *duerm-* y *durm-*.

[89] Además de la influencia de la tradición anglosajona en la extensión de determinados términos, las relaciones de la Morfología con otros niveles lingüísticos, como el léxico (cambios en el significado léxico de las palabras como resultado de procesos de derivación o composición; restricciones léxicas que actúan sobre estos procesos), el fonológico (influencia del contexto fonológico en la forma de los morfemas, morfofonología) o el sintáctico (establecimiento de clases de palabras: sustantivo, verbo, etc.; diferencias entre la estructura argumental, tipo de complementos, etc. de palabras base y palabras derivadas a partir de esa base), también juegan un papel importante en la descripción del componente morfológico.

- *Afijo*: es el resto de elementos formativos que acompañan a la raíz y que suelen aportar significados adicionales, léxicos (*afijos derivativos*) o gramaticales (*afijos flexivos*). Los más comunes son *prefijos* (preceden a la raíz; p. ej. *in-* + *feliz*) y *sufijos* (siguen a la raíz; p. ej. *feliz* + *-idad*, *feliz* + *-es*), aunque también se distinguen *circunfijos*, *infijos*, etc. Igual que ocurría con las raíces, su forma se puede ver alterada por el contexto inmediato. La sufijación es el fenómeno más frecuente. Desde la perspectiva computacional con frecuencia solo se consideran prefijos y sufijos (*cf.* Trost 2003:31).
- *Morfo* y *alomorfo*: son términos que aluden a la realización material de un morfema en un contexto dado; cuando existe variación en la forma de los morfemas como consecuencia del contexto, se utiliza el término *alomorfo*. P. ej. el morfema *in-* tiene tres alomorfos: *i-* ante /l/ o /r/: *ilegal*, *irreal*; *im-* ante /p/ o /b/: *imposible*, *impaciente*; *in-* en el resto de contextos. O en el caso del morfema de plural, su expresión formal en español puede ser *–s* (*verde* + *s*), *–es* (*azul* + *es*) o incluso Ø, como en *crisis*.
- *Tema*: es la unidad básica en la descripción de la flexión y la formación de palabras, que resulta de eliminar los morfemas flexivos de una palabra (*cf*. Pena 1999:4308). Lógicamente, puede coincidir con una palabra: *tapa*, *alegre* son temas y palabras; o con una raíz (*mar-* en *mares*, caso en el que coinciden palabra, tema y raíz). A partir de los temas se obtienen las formas flexivas: *tapas*, *alegres*. Asimismo, hay temas que están formados solo por la raíz (*temas simples*, como *mar*) o por raíz y afijos (*temas derivados*: *marin-* en *marino*). Cuando la palabra es el resultado de la combinación de más de un tema (p. ej. *aguamarina*), el *tema* es *compuesto*. Por otra parte, los afijos flexivos se caracterizan por

adjuntarse al tema, mientras que los derivativos forman parte del tema.

Los morfemas no siempre se suman unos a otros, sino que existe también la llamada *morfología no concatenativa* (*cf.* Trost 2003:32-33), como la que se da en las lenguas semíticas, en las que la raíz –en árabe, formada por un número de consonantes que oscila entre dos y cuatro– soporta la carga semántica (p. ej. árabe *ktb* 'escribir'), mientras que las vocales se combinan con las consonantes de la raíz para aportar la información sobre la voz y el aspecto, según un patrón preestablecido: así, en árabe, el patrón CVCVC da lugar a *katab* (voz activa) y *kutib* (voz pasiva). Otros fenómenos de este tipo son el *ablaut* o *alternancia vocálica* –alteraciones en el vocalismo con consecuencias morfológicas, propio de las lenguas indoeuropeas, en las que el cambio de la vocal tónica implica un cambio morfológico (como sucede en el verbo inglés *sing*, *sang*, *sung*, formas del infinitivo, pasado y participio respectivamente); el *umlaut*, *metafonía* o *armonización vocálica*, cambios en el timbre vocálico de la raíz por efectos de vocales cerradas en los sufijos, sufijos que luego pueden desaparecer dejando como única marca la distinción vocálica (que diferencian el inglés *goose* 'oca, ganso' / *geese* 'ocas, gansos' singular y plural respectivamente; el alemán *garten* 'jardín' / *gärten* 'jardines'; el asturiano *pirru* / *perros*), o cambios en los sufijos por efecto del timbre de la vocal tónica (como en turco, lengua en la que el morfema de plural varía, *-lar* o *–ler*, en función del timbre de la vocal tónica: *at* 'caballo' / *atlar* 'caballos'; *yüz* 'cara' / *yüzler* 'caras'); cambios en el tono como marca de morfema o desplazamientos acentuales, como en el inglés *record* (N) *vs. record* (V), /ˈrekɔːd/ /rɪˈkɔːd/ respectivamente; la supleción, o modificación total, frecuente en formas muy usadas (esp. *es, soy, fue*), etc.

Por otra parte, hay que considerar las diferencias entre lenguas, que pueden motivar que lo que en una lengua se expresa morfológicamente en otra se exprese de forma sintáctica, etc. o también de forma distinta (mediante sufijos en un caso, prefijos en otro…). Según la tipología lingüística tradicional, aunque en la realidad los tipos no son tan puros, se distingue entre:

- *Lenguas flexivas*: p. ej. el español, que modifican la raíz mediante la adición de afijos, aunque estos no siempre son fáciles de segmentar (morfemas *portmanteau*), como ocurre en la forma *cantaremos*, en la que *-mos* expresa conjuntamente los morfemas de número y persona, sin posibilidad de separar uno de otro.
- *Lenguas aglutinantes*: p. ej. el húngaro, el turco o el finés, que suelen añadir un gran número de afijos a la raíz; además, cada afijo tiene un significado más o menos claro que se asocia a una forma determinada, por lo que se pueden aislar con relativa facilidad. Es lo que ocurre en el húngaro *a házakban* 'en las casas', donde *a* = 'la', *ház* = 'casa', *ak* = 'plural' y *ban* = 'en' (ablativo).
- *Lenguas polisintéticas*: lenguas sumamente aglutinantes, en las que normalmente los nombres o verbos incorporan sus argumentos en una misma palabra, lo que da lugar a que esta recoja el contenido expresado en otras lenguas por una frase entera, como ocurre en *amanganachquiminchi*, que en algonquino significa 'encina de anchas hojas'.
- *Lenguas aislantes*: p. ej. el chino, en las que los contenidos morfológicos se expresan mediante elementos independientes, palabras nuevas, por lo que no hay afijos propiamente dichos; solo procesos morfológicos de composición.

Teniendo en cuenta estas distinciones, los tres procesos morfológicos mencionados con anterioridad –que pueden a su vez suceder de forma simultánea– se pueden definir como sigue:

- La *flexión* es el procedimiento morfológico que consiste en la modificación de un tema mediante afijos flexivos (sufijos en el caso del español) aunque sin dar lugar a un tema nuevo. Estos afijos únicamente aportan información morfológica, no modifican el significado básico de la palabra ni cambian su categoría morfosintáctica. P. ej. *bocas* < *boca* + *s* (indica que se trata de un plural). Normalmente la flexión permite diferenciar entre clases de palabras invariables en este sentido (partículas: preposiciones, adverbios, conjunciones) y palabras variables (nombres, adjetivos, verbos…).
- La *derivación* es el procedimiento de formación de palabras que altera el tema mediante prefijos o sufijos, o la modificación de la propia base y que da como resultado un tema nuevo. Estos prefijos y sufijos cambian el significado básico de la palabra de una forma que no siempre es predecible y, con frecuencia, dan lugar a cambios en la categoría gramatical. P. ej. *alteración (N)* < *alterar* (V) + *ción*.
- La *composición* es el procedimiento que consiste en la unión de dos temas para producir una palabra nueva. Hay quien considera que se trata de una variante de la derivación. P. ej. *bocacalle* < *boca* + *calle*.
- La *parasíntesis* es el proceso de formación de palabras que aúna la composición y la derivación, como *doceañista* < *doce* + *año* + *-ista* (*cf.* Piera y Varela 1999), o que combina dos o más sufijos, o un sufijo y un prefijo para producir nuevas palabras (*cf.* Gómez Torrego 2007; Pena 1999), como *anochecer* < *a-* + *noche* + *-ecer*, *desmitificar* < *des-* + *mito* + *-ificar*, etc.

Estos procesos permiten una gran productividad en la formación de nuevas palabras: una palabra derivada estará sujeta a las variaciones flexivas de la lengua en que se haya producido y, a su vez, podrá dar lugar a palabras compuestas.

Lo último que hay que destacar es que las contribuciones de la lingüística tradicional al tratamiento de la morfología desde la perspectiva computacional han sido de gran trascendencia, al haberse centrado en ofrecer descripciones sistemáticas de los paradigmas, sobre todo flexivos, de las palabras, pues esta era su unidad de trabajo por excelencia. Esta postura tradicional es la que se suele adoptar en LC al considerar la palabra como una unidad abstracta que se manifiesta de múltiples maneras:

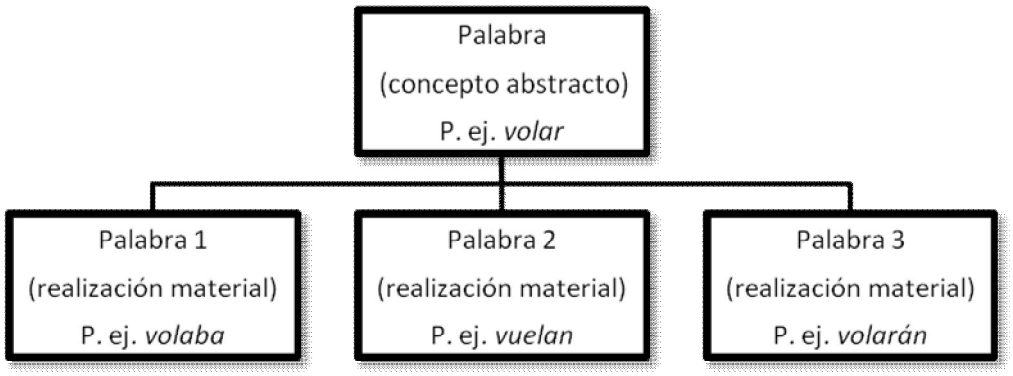


Ilustración 19. La palabra como unidad de la Morfología.

Desde esta perspectiva, los infinitivos para los verbos, el singular para los nombres, la forma masculina singular para los adjetivos, etc. –en el caso del español– son tomados como base para estudiar los procesos morfológicos que afectan a las palabras, entendidas como formas abstractas, generalmente coincidentes con las entradas o lemas de un diccionario.

2.2.1. Las tareas de la morfología computacional

La necesidad de tratar el componente morfológico en LC es en cierto modo reciente, en torno a la década de los ochenta. Hasta ese momento apenas había recibido atención, pues el inglés centraba la mayor parte de los esfuerzos. Dada la relativa simplicidad de su morfología, bastaba con un acercamiento superficial a la misma. Sin embargo, cuando se empieza a trabajar con lenguas mucho más complejas en este sentido, como el finés, a lo que hay que sumar el creciente interés por el desarrollo de sistemas prácticos de procesamiento del lenguaje natural, se estimularán las investigaciones sobre morfología en LC.

Hoy en día se puede decir que es un nivel del lenguaje que está relativamente bien estudiado desde la perspectiva computacional, aunque no por ello exento de problemas. La mayoría de ellos están relacionados con la formación de palabras (reglas que determinan la combinación de raíces y afijos) y las alternancias morfofonológicas.

No obstante, el tratamiento morfológico es decisivo para reducir el volumen del diccionario, componente básico de un sistema computacional que dé cuenta del lenguaje, ya que contiene las palabras que el sistema va a reconocer o generar como válidas en una lengua determinada. Fue precisamente la imposibilidad de listar en el diccionario todas las palabras con sus respectivas variantes la que puso de manifiesto la necesidad de considerar este nivel. De no contemplar la morfología, p. ej. ante cualquier verbo español, deberíamos recoger en el diccionario todas sus formas flexivas (tiempos, modos, personas, números), pues de lo contrario el sistema no las identificaría como palabras de esta lengua: *amo*, *amas*, *amar*... Por otra parte, este tratamiento de las formas de las palabras es mucho más acorde con el

procesamiento cognitivo de las personas: cómo almacenamos en la memoria y recuperamos las palabras.

Así pues, la morfología computacional propone estrategias para no tener que aumentar el diccionario innecesariamente. Además, hay que señalar que el análisis morfológico es un requisito previo al análisis sintáctico, al que aporta informaciones básicas como la clase de palabras, o el género y número para la concordancia.

Teniendo en cuenta estas consideraciones, los objetivos de la morfología computacional son:

- El principal y más general, identificar las palabras de un texto de entrada: es decir, reconocerlas como pertenecientes a una lengua dada o, en el caso de un sistema de generación, producir palabras válidas en la lengua en cuestión. El conocimiento que requiere un sistema computacional es un diccionario, una lista con las palabras que queremos que el sistema identifique o produzca, que en LC se suele denominar *lexicón*[90]. En un caso concreto, como puede ser la palabra *azules* (*input*), el sistema consulta su lexicón. Si esta forma está almacenada en el mismo, la reconocerá y producirá un análisis de la misma con la información que le suministre aquel. Para la generación, partirá de la forma canónica *azul* y le añadirá los afijos correspondientes, en este caso, para la formación de plural (*-es*).
- Como no todas las palabras de una lengua son entradas de un diccionario, la morfología computacional tiene como tarea más específica la de identificar y aislar los diferentes elementos que componen las palabras: morfemas (raíces y afijos). Este proceso se denomina *segmentación*. Las palabras no se suelen presentar como lemas de un diccionario, sino que muestran modificaciones

[90] Influencia de la Psicolingüística y la IA. *Vid.* MORENO ORTIZ (2000).

debidas a la flexión (las motivadas por la derivación y la composición suelen dar lugar a palabras nuevas recogidas como entradas del diccionario). En el caso de *azules*, el sistema debe ser capaz de identificar la raíz *azul-* y el sufijo *-es*. Aquí los conocimientos necesarios se refieren a diccionarios o listas de raíces por una parte y de afijos, por otra.

- Otra tarea propia de la morfología computacional es relacionar los diferentes elementos compositivos de las palabras mediante algún tipo de reglas morfológicas, que establecen cómo se combinan los morfemas entre sí: *concatenación*. La *morfotáctica* es la parte de la morfología que describe estos procesos: en qué orden se combinan los morfemas, qué morfemas pueden combinarse, bajo qué condiciones gramaticales (un afijo solo se concatena con bases de cierta categoría), fonológicas, semánticas (p. ej. un prefijo negativo no se puede aplicar a una base con significado negativo en inglés: *sad* > **unsad*) o léxicas (sufijos que solo actúan sobre bases de origen latino, como en inglés *-ity*) se verifica la combinación, qué modificaciones ortográficas o fonético-fonológicas producen esas combinaciones[91] -tarea esta última de la que se ocupa la *morfofonología*[92]-, etc.

---

[91] Un ejemplo de regla morfotáctica en español relacionada con el orden de los afijos es la que indica que en los adjetivos primero va el morfema de género y luego el de número (*roj-a-s*); la formación del plural con el alomorfo *–es* cuando un nombre acaba en consonante, sería un ejemplo de restricción fonológica (*león* → *leon-es*) que afecta a la morfología; y el cambio de *–z* en *–c-* (*vez* + *-es* → *veces*) reflejaría las consecuencias ortográficas derivadas de un proceso flexivo.

[92] Aunque la morfología computacional normalmente trabaja con textos escritos, las lenguas difieren en el grado en que la escritura se aproxima a la representación fonológica de la lengua en cuestión: pensemos en el español *vs.* inglés. Casos en que la concatenación conlleva un cambio fonético-fonológico son: la asimilación (*en* + *palidecer* → *empalidecer*), la epéntesis (como en los plurales en *es*: *ser* + plural → *ser* + *e* + *es*), la elisión (*hemi* + *esfera* → *hemisferio; inscrito vs. inscripto*) o la armonización vocálica, como ocurre en la terminación del infinitivo en mongol: */-Vx/* (donde V=vocal) se realiza como [-ox], [öx], [-ax] o [ex], según el timbre de la vocal tónica (*cf.* TROST 2003:36-37 y HANNAHS 2001:10054), etc.

Por ejemplo, a partir de la base *azul*, las palabras que se obtienen por derivación (en este caso, mediante la adición de diferentes sufijos: *-ar*, *-ear*, *-ejo*, *-enco*, *-ete*, *-ino*, *-oso*) son las que muestra el siguiente gráfico[93], formas que a su vez se convierten en la base de otros procesos derivativos:

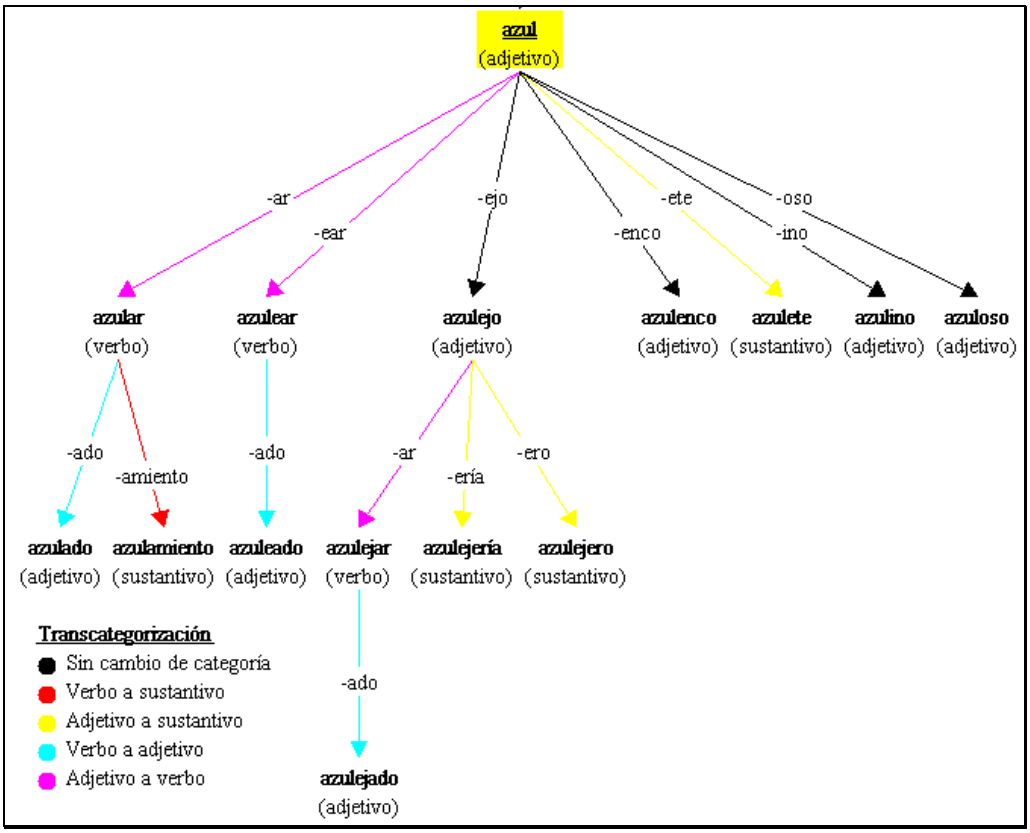


Ilustración 20. Grafo de las relaciones morfoléxicas de la forma "azul".

- Otro proceso del que se ocupa la morfología computacional es, una vez analizada la palabra en sus partes componentes, obtener las formas básicas o palabras tal y como las encontraríamos en un diccionario (*lemas*), es decir, relacionar las palabras concretas con entradas del lexicón o diccionario, que incluye todas las

[93] Obtenido con el generador de relaciones morfoléxicas del Grupo de Estructuras de Datos y Lingüística Computacional de Las Palmas de Gran Canaria. URL: http://www.gedlc.ulpgc.es/investigacion/scogeme02/relmorfo.htm

palabras que el sistema es capaz de reconocer y/o generar, proceso conocido como *lematización*. En el ejemplo de *azules*, el análisis morfológico tendría que llevarnos a las formas canónicas *azul* y *azular*[94]:

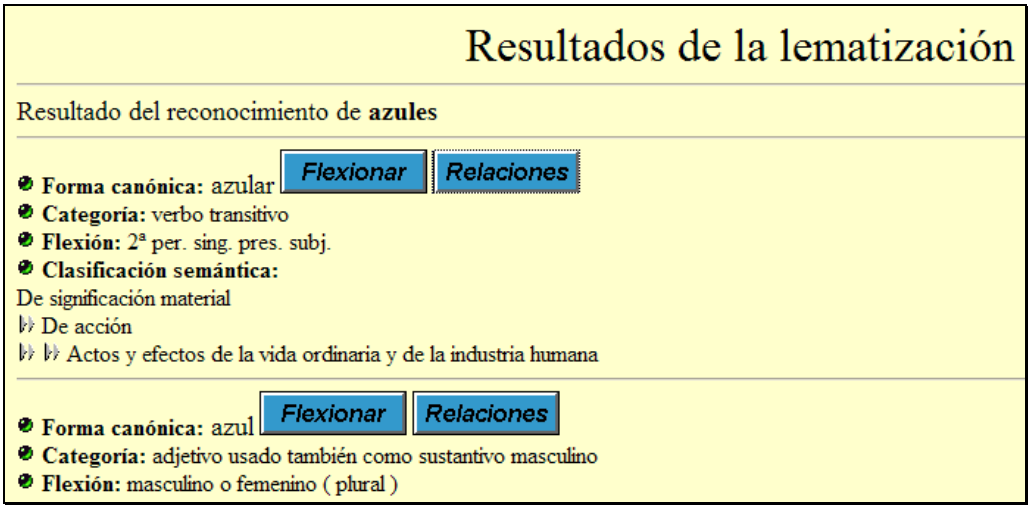


Ilustración 21. Resultado de la lematización de "azules".

- Una de las tareas fundamentales, por ser básica para el posterior análisis sintáctico, es asignar a las palabras (*lemas*) una o varias categorías gramaticales (nombre, verbo, adjetivo, etc.): *categorización* o *POS tagging* (POS = "Part Of Speech"). En el caso de que la palabra pueda pertenecer a más de una categoría, se debe llevar a cabo la *desambiguación* atendiendo a su contexto inmediato. La palabra *azules* podría ser un adjetivo, un nombre o un verbo, según el contexto en que la empleemos. En el caso de

[94] Obtenido con el lematizador del Grupo de Estructuras de Datos y Lingüística Computacional de la Universidad de Las Palmas de Gran Canaria. URL: http://www.gedlc.ulpgc.es/investigacion/scogeme02/lematiza.htm

Las ballenas azules eran abundantes[95], la interpretación válida es su uso como adjetivo:

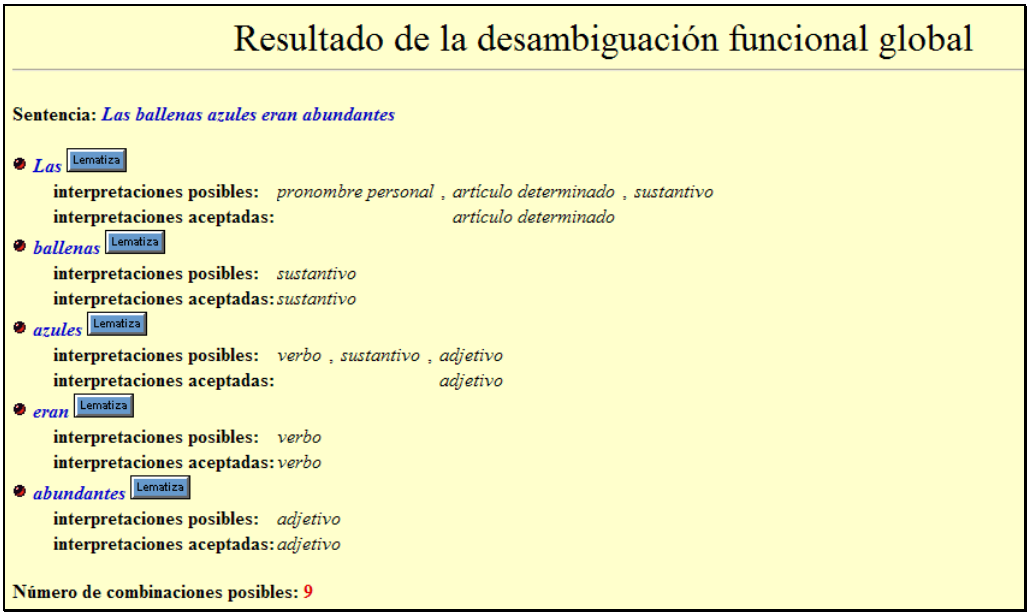


Ilustración 22. Desambiguación de "azules" en "Las ballenas azules eran abundantes".

- Por último, es labor de la morfología computacional explicitar la información morfológica relevante para la aplicación o para el propósito de la investigación (número, género, tiempo verbal, persona, etc.): *etiquetado* ("tagging"). Esta información desempeña un papel fundamental en algunos de los últimos formalismos gramaticales.

---

[95] En este ejemplo, hemos empleado el desambiguador morfosintáctico del Grupo de Estructuras de Datos y Lingüística Computacional de la Universidad de Las Palmas de Gran Canaria. URL: http://gedlc.ulpgc.es/investigacion/desambigua/morfosintactico.htm, que realiza un análisis sintáctico para determinar la categoría de aquellos lemas que pueden pertenecer a más de una. Queremos llamar la atención sobre el número de alternativas (9) o combinaciones posibles para el análisis que genera una frase tan simple y con las que juega el sistema hasta llegar, en este caso, a una única combinación aceptada: (*Las*) artículo determinado (*ballenas*) sustantivo (*azules*) adjetivo (*eran*) verbo (*abundantes*) adjetivo. Este trabajo previo facilita enormemente la tarea de los analizadores sintácticos, aquellos programas que, además de la categoría gramatical, nos proporcionan la estructura sintagmática o agrupaciones de categorías en otros constituyentes de nivel superior.

En resumen, desde el momento en que una palabra se ve afectada por un proceso morfológico de algún tipo, aumentan las dificultades para un sistema computacional: si la palabra en cuestión no está en el lexicón, no será reconocida ni podrá ser generada y, por tanto, será imposible su procesamiento. Por ello, la morfología computacional tiene como cometido general proponer métodos para, dada cualquier palabra (como *cariñito* en la secuencia "buena suerte cariñito" dicha por Carla Bruni a Sarkozy, *Público.es*, 21/05/2009), poder llegar a su forma canónica (*cariño*) e identificar los afijos que la han modificado (sufijo diminutivo *-ito*) o que la pueden modificar, así como las reglas que rigen dicha combinatoria (*cariño* + *-ito* → *cariñito*, con elisión de *–o*) y que, en última instancia, permiten su consulta en el diccionario del sistema, bien sea para el análisis –lo más frecuente- o para la generación. En este cometido, son de gran relevancia las investigaciones sobre morfología en lingüística teórica, ya que proporcionan a la morfología computacional los fundamentos a partir de los cuales describir y formalizar la morfología de las lenguas para su ulterior implementación computacional.

O. Santana *et al.* (1998:13-15) ilustran estos procesos de la siguiente manera[96]: para la segmentación, la palabra de entrada se descompone en posibles pares raíz-terminación (p. ej. *crédulamente* → *crédulament-e* / *crédula-mente* / *crédul-amente*) y los prefijos que pueda haber (*hipersensible* → *hiper* + *sensible*). En los módulos externos se comprueba si la raíz en cuestión admite la terminación, se indica la flexión o derivación que le corresponde, se señala la forma canónica y se proporciona información sobre su categoría gramatical:

[96] *Vid.* gráficos en la página siguiente.

Segmentación:

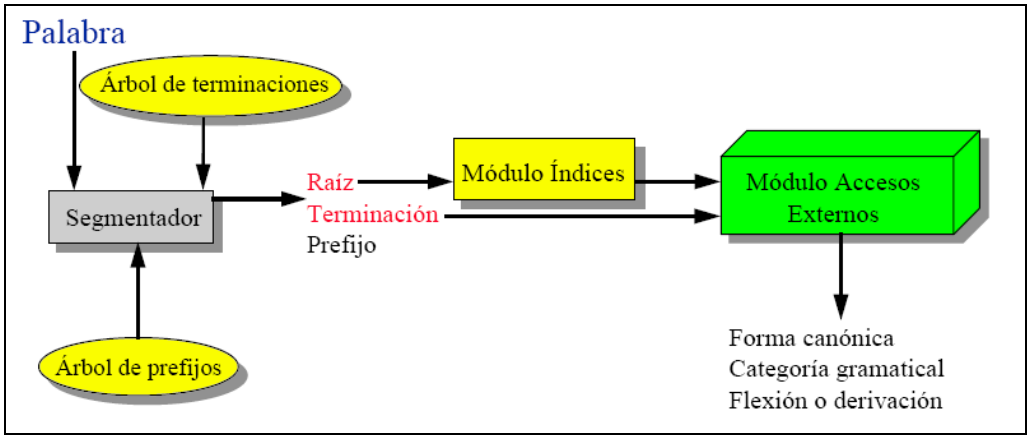


Ilustración 23. Procesos implicados en la segmentación.

Generación:

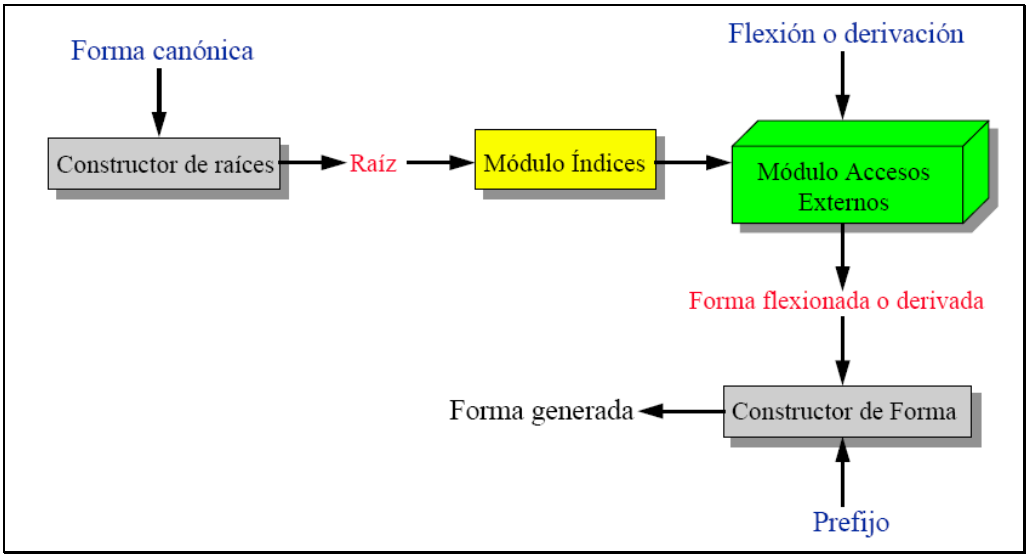


Ilustración 24. Procesos implicados en la generación.

Ante un enunciado como el siguiente: "Estas son las propuestas" (*Público.es*, 11 de mayo de 2009), un programa encargado del análisis morfológico debería ser capaz de proporcionarnos, al menos, la siguiente información:

Texto de entrada	**"Estas son las propuestas"**		
Análisis morfológico	Palabra	Lema	Información morfológica
	estas	este	[determinante demostrativo femenino plural]
	son	ser	[verbo presente de indicativo 3ª persona plural]
	las	el	[determinante artículo femenino plural]
	propuestas	propuesta	[nombre común femenino plural]

Tabla 5. Tareas del análisis morfológico computacional.

De los tres tipos de procesos morfológicos mencionados -flexión, derivación y composición-, la *flexión* es el que ha sido objeto de mayor tratamiento por las propias características que presenta la morfología flexiva: paradigma reducido (número: singular-plural, persona: 1ª, 2ª, 3ª, etc.) y posibilidad de descripción sistemática mediante un conjunto de reglas[97]. Por otra parte, los procesos de flexión no suelen operar un cambio de categoría en la palabra a la que afectan. Es habitual distinguir entre *partículas* (palabras que carecen de flexión, como las preposiciones), *verbos* (conjugación) y *nominales* (declinación), al ser diferentes los procesos flexivos que actúan sobre cada uno de ellos. Y tampoco alteran el significado léxico de las palabras.

La *composición* y la *derivación* no son tan fáciles de describir y sistematizar, además de estar sujetas a una mayor productividad por parte de los hablantes. En la derivación, mediante la adición de afijos a

[97] H. RODRÍGUEZ HONTORIA (2000) comenta que con unos doscientos sufijos flexivos y unas quinientas reglas de combinación es posible describir la morfología flexiva de lenguas como el castellano o el catalán.

una forma base se obtiene una nueva palabra, proceso que suele llevar parejo un cambio de categoría: *azul* (adjetivo, nombre) > *azular* (verbo). Sin embargo, un mismo sufijo no es aplicable a todos los miembros de una categoría, sino que presenta restricciones. P. ej. el sufijo *–ble* actúa sobre verbos para producir adjetivos, pero esta regla no es aplicable a todos los adjetivos[98] (hay adjetivos terminados en *–ble* sin que exista un verbo con el que relacionarlos: *viable*, *inviable* → **viar*). Además, es un fenómeno recursivo, ya que una palabra derivada puede a su vez ser objeto de un proceso ulterior de derivación: *parcial* > *imparcial* > *imparcialidad*. Por otra parte, el significado de la nueva palabra no siempre es claramente la suma de los significados de los morfemas que la forman, a diferencia de la flexión (el morfema de plural siempre aporta el mismo significado). En cuanto a la composición, en este caso, el mecanismo para crear una nueva palabra se sirve de dos formas básicas previamente existentes: *azul + grana > azulgrana*.

Si pensamos en el caso del español, es relativamente fácil establecer reglas para la formación del plural en los nombres o patrones de conjugación verbal. Sin embargo, es más complicado fijar los procesos que se siguen a la hora de crear nuevas palabras. Continuamente los hablantes producen formas nuevas, p. ej. *puenting*, de *puente* y el sufijo inglés *-ing*, adverbios en *-mente* que no estén contemplados en el lexicón, o los neologismos compuestos a partir de *blog*: *bloguero* (del término inglés *blog* y el sufijo español *-ero*), *fotoblog*, *bloguear*, etc.

No obstante, cabe señalar los progresos llevados a cabo en este sentido. En nuestro país, sobresalen los trabajos del Grupo de Estructuras de Datos y Lingüística Computacional de la Universidad de Las Palmas de Gran Canaria, que se han traducido en toda una serie de

[98] O con frecuencia, la presencia de determinada secuencia de caracteres no siempre indica la presencia de un sufijo, como ocurre con *-ble* en *cable*.

aplicaciones en línea para la morfología[99], o las herramientas para la morfología del *Centre de Llenguatge i Computació* (CLIiC)[100] de la Universidad de Barcelona, que han dado lugar a la empresa especializada en el tratamiento computacional de la lengua Thera[101].

Además, hay que destacar que el interés por el tratamiento de la morfología en LC no es meramente teórico: saber cómo funciona este nivel de la lengua para reproducirlo en un ordenador. También la necesidad de desarrollar aplicaciones concretas fomenta su estudio: correctores ortográficos, conjugadores, flexionadores, lematizadores, programas de recuperación de información, programas de etiquetación para corpus, programas de traducción automática, programas relacionados con las tecnologías del habla, programas de ayuda a la redacción de documentos, etc.

Algunos ejemplos de programas de este tipo son:

1. *Lematizadores*: programas que asocian una forma flexionada con su lema o entrada del diccionario, tras la separación de la raíz de los afijos.
2. *Flexionadores*: herramientas que a partir de un lema nos permiten obtener sus formas flexivas. Destacan los *conjugadores*, que generan las formas flexivas de los verbos.
3. *Etiquetadores*: herramientas que asignan de forma automática la categoría gramatical y proporcionan información sobre las características morfológicas.

[99] URL: http://gedlc.ulpgc.es/
[100] URL: http://clic.ub.edu/
[101] URL: http://www.thera-clic.com/site/index-ES.html

En español, algunos ejemplos de lematizadores son el del Grupo de Estructuras de Datos y Lingüística Computacional de la Universidad de Las Palmas de Gran Canaria[102]. Al introducir una forma, como la palabra *notas*, el programa la relaciona con todos los lemas o entradas que tenga en su lexicón según sus características morfológicas: así, propone como lemas *notar*, tras haber analizado la terminación de la forma de entrada como la de la segunda persona del singular del presente del indicativo; el nombre *nota* en su forma de plural; o el femenino plural del adjetivo poco usado *noto*[103] 'público y sabido'. Las capturas de pantalla documentan los resultados:

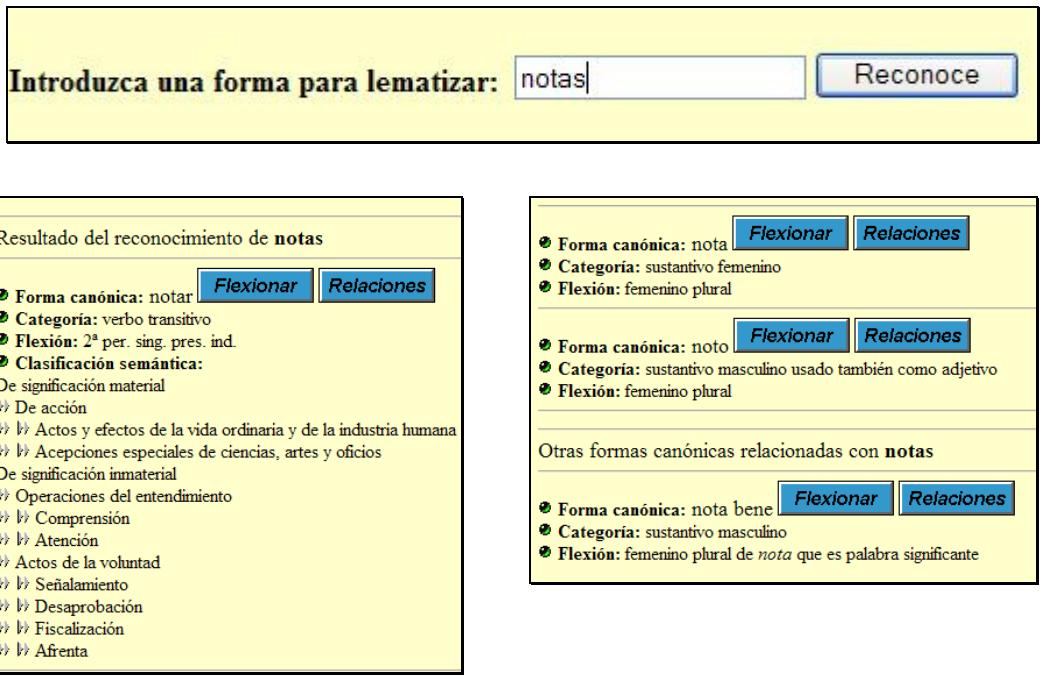


Ilustración 25. Lematización de la forma "notas".

[102] URL: http://gedlc.ulpgc.es/investigacion/scogeme02/lematiza.htm

[103] Como se aprecia, el programa establece conexiones con todas aquellas entradas del diccionario del sistema computacional que estén emparentadas morfológicamente con el texto de entrada, lo que arroja resultados y, por lo tanto, opciones de análisis, que las personas no nos planteamos al tener que procesar la información o que descartamos inmediatamente en función de nuestros conocimientos del mundo, del contexto lingüístico anterior y posterior, etc. En este caso, dada la baja frecuencia de uso del adjetivo, seguramente una persona no habría asociado inmediatamente la forma *notas* con el adjetivo *noto*.

El lematizador de Thera[104] para el catalán y el español (sección Demos, Morfología) ofrece resultados similares:

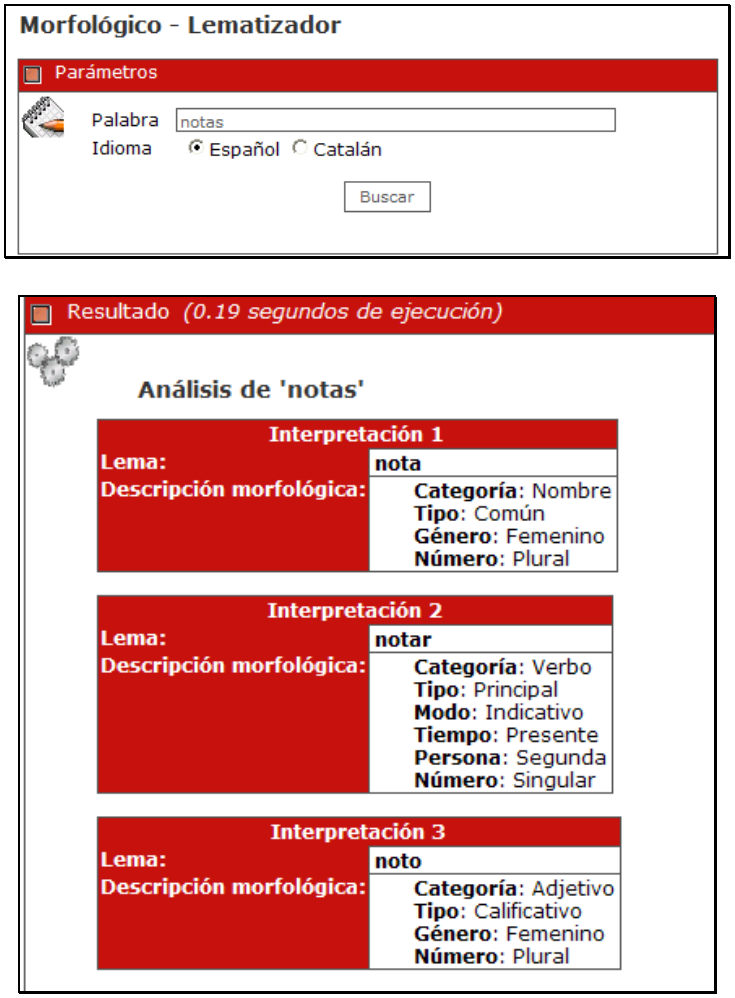


Ilustración 26. Otra lematización de la forma "notas".

Los lematizadores son empleados en la elaboración de diccionarios o para el análisis automático del léxico empleado en un ámbito específico, al facilitar la recuperación de la información cuando se consulta una base de datos documental[105].

---

[104] URL: http://www.thera-clic.com/

[105] Por ejemplo, para la búsqueda automática de conceptos médicos en historias clínicas: *leucocito* y *leucocitarios*. *Vid*. la aplicación *hCod* de Thera para el sector de la salud. URL: http://www.thera-clic.com/site/Productos/hCod-ES.html

Además de los lematizadores, destaca otra herramienta computacional como son los flexionadores. En este caso, ante el texto de entrada, además de obtener la información sobre los lemas, se presentan las formas flexivas de las palabras del *input*. Por ejemplo, al introducir *final* en el flexionador de Thera, nos propone tres lemas (adjetivo, nombre masculino y nombre femenino) y, según las reglas flexivas del español, las formas correspondientes al plural del adjetivo y de los nombres:

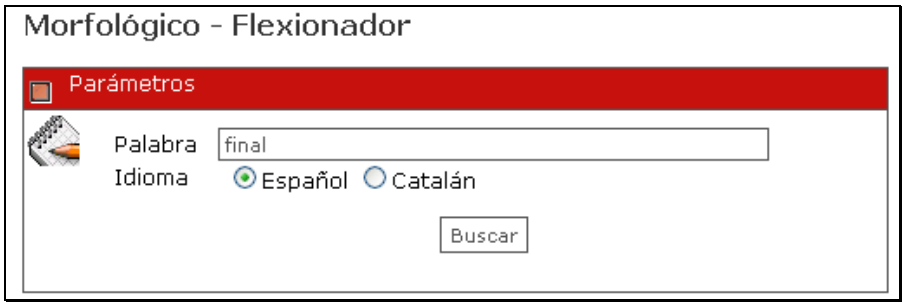


Ilustración 27. Flexión de la forma "final".

La simple incorporación de este tipo de tratamiento lingüístico en buscadores amplía y enriquece la profundidad de los buscadores en páginas web, como ocurre con *W-Lem*[106], de Thera. Obsérvese el funcionamiento:

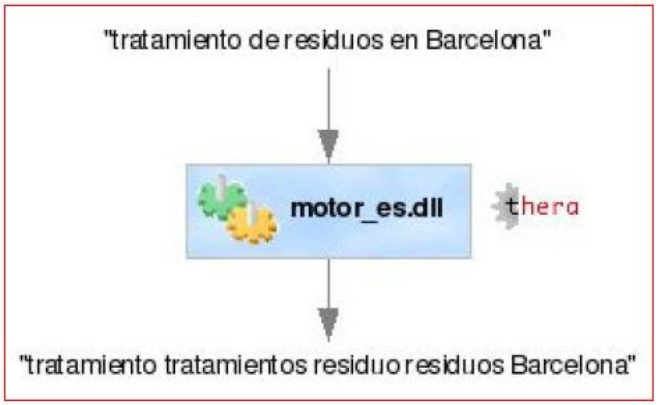


Ilustración 28. Funcionamiento de W-Lem.

El Grupo de Estructuras de Datos y Lingüística Computacional, Universidad de Las Palmas de Gran Canaria, también dispone de un flexionador para sustantivos, adjetivos, verbos y otras formas[107].

Quizá la aplicación más popular de este tipo de programas son los conjugadores, es decir, los flexionadores de verbos, como el que incorpora la versión en línea del *Diccionario de la lengua española* de la Real Academia Española[108], que proporciona en un clic la conjugación de cualquiera de los verbos incluidos en el *DRAE*, con sus correspondientes formas del voseo rioplatense:

---

[106] Para una descripción: http://www.thera-clic.com/pdf/pub_w_lem_cast.pdf

[107] *Vid.* SANTANA *et al.* (1997, 1998, 1999). URL de los flexionadores:

- http://gedlc.ulpgc.es/investigacion/scogeme02/flexsus.htm
- http://gedlc.ulpgc.es/investigacion/scogeme02/flexver.htm
- http://gedlc.ulpgc.es/investigacion/scogeme02/flexadj.htm
- http://gedlc.ulpgc.es/investigacion/scogeme02/flexotra.htm

[108] URL: http://www.rae.es

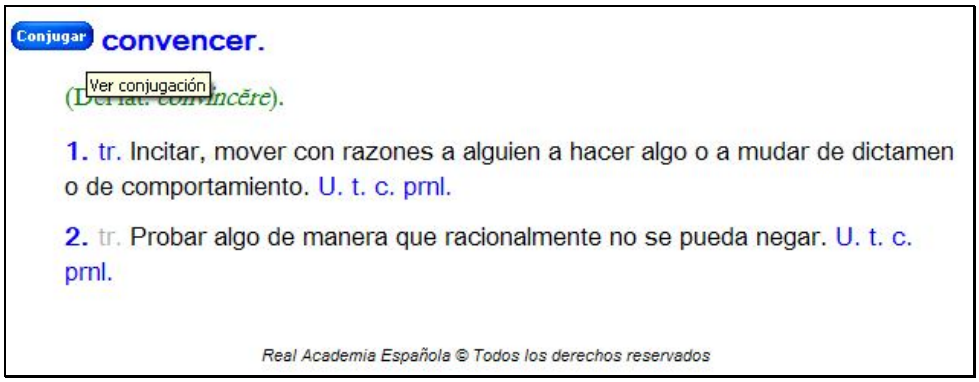


Ilustración 29. Conjugador del DRAE en línea.

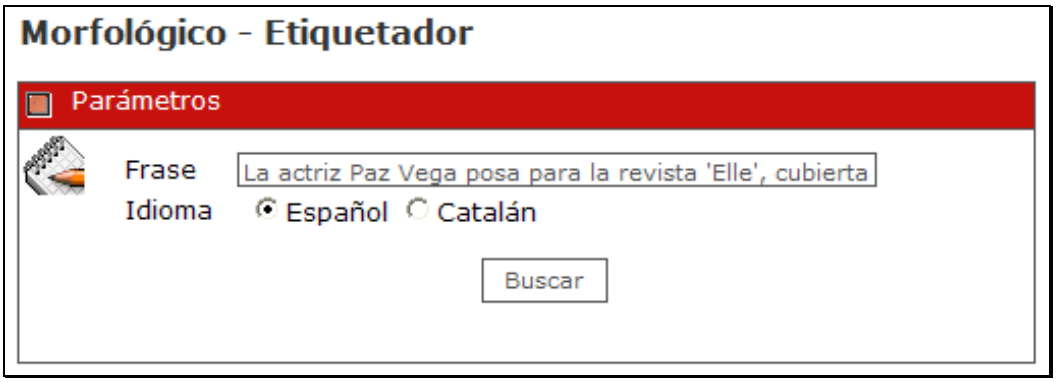
Otros conocidos conjugadores son el *Conjugador de verbos*[109], del Dpto. de Lengua Española de la Universidad de Santiago de Compostela, para su uso en línea; el *Conjugador* de Signum[110] que, además de ser consultable en línea, se puede instalar y utilizar en un ordenador sin necesidad de conexión; o *Verbix*[111], sitio web que nos ofrece acceso a la conjugación de verbos de cientos de lenguas en línea (*WebVerbix*) o sin conexión (*Verbix*), previa descarga del programa en versión para evaluación o de su adquisición, así como otras aplicaciones relacionadas con las lenguas y su morfología verbal.

La principal utilidad de los conjugadores, además de para resolver dudas de los hablantes nativos sobre la conjugación verbal, está en su uso en la enseñanza/aprendizaje de lenguas extranjeras, para hablantes no nativos.

Por último, los *etiquetadores* nos proporcionan información morfológica de las formas de un enunciado, es decir, realizan un análisis morfológico del texto de entrada. Por ejemplo, el etiquetador de Thera nos devuelve el siguiente análisis de la frase "La actriz Paz

[109] URL: http://gramatica.usc.es/conjuga.html
[110] URL: http://www.lenguaje.com/herramientas/conjugador.php
[111] URL: http://www.verbix.com/

Vega posa para la revista 'Elle', cubierta de 6.000 cristales Swarovski" (*El País*, 20/11/2007), hipotetizando en el caso de palabras no incluidas en su lexicón, como ocurre con los nombres propios (Paz Vega, Swarovski, 'Elle'), o de las cifras (6.000), que aparecerán recogidas como palabras, pero no como cifras.



Resultado *(0.32 segundos de ejecución)*

Análisis de la frase 'La actriz Paz Vega posa para la revista 'Elle', cubierta de 6.000 cristales Swarovski'

Palabra	Lema	Descripción morfológica
La	el	**Categoría**: Determinante **Tipo**: Artículo **Género**: Femenino **Número**: Singular
actriz	actriz	**Categoría**: Nombre **Tipo**: Común **Género**: Femenino **Número**: Singular
Paz_Vega	Paz_Vega	*(Hipotetizado)* **Categoría**: Nombre **Tipo**: Propio
posa	posar	**Categoría**: Verbo **Tipo**: Principal **Modo**: Indicativo **Tiempo**: Presente **Persona**: Tercera **Número**: Singular
para	para	**Categoría**: Preposición
la	el	**Categoría**: Determinante **Tipo**: Artículo **Género**: Femenino **Número**: Singular
revista	revista	**Categoría**: Nombre **Tipo**: Común **Género**: Femenino **Número**: Singular
'	'	**Categoría**: Signo de puntuación
Elle	elle	*(Hipotetizado)* **Categoría**: Nombre **Tipo**: Propio
'	'	**Categoría**: Signo de puntuación
,	,	**Categoría**: Signo de puntuación
cubierta	cubierto	**Categoría**: Adjetivo **Tipo**: Calificativo **Género**: Femenino **Número**: Singular **Función**: Participio
de	de	**Categoría**: Preposición
6.000	6000	*(Hipotetizado)* **Categoría**: Cifra
cristales	cristal	**Categoría**: Nombre **Tipo**: Común **Género**: Masculino **Número**: Plural
Swarovski	Swarovski	*(Hipotetizado)* **Categoría**: Nombre **Tipo**: Propio

Ilustración 30. Resultados de la etiquetación morfológica.

Otro ejemplo nos lo proporciona *Spanish Machinese Phrase Tagger*, de la empresa finlandesa Connexor[112], que muestra la importancia que el tratamiento morfológico tiene como paso previo a un análisis sintáctico de un texto:

[112] URL: http://www.connexor.eu/technology/machinese/demo/tagger/

Text	Baseform	Phrase syntax and part-of-speech
La	la	premodifier, determiner
actriz	actriz	nominal head, noun, single-word noun phrase
Paz	Paz	premodifier, proper noun, noun phrase begins
Vega	Vega	nominal head, proper noun, noun phrase continues
posa	poso	postmodifier, adjective, noun phrase continues
para	para	postmodifier, preposition, noun phrase continues
la	la	premodifier, determiner, noun phrase continues
revista	revista	nominal head, noun, noun phrase ends
'Elle'	Elle	nominal head, proper noun, single-word noun phrase
,	,	
cubierta	cubierto	nominal head, adjective
de	de	postmodifier, preposition
6.000	6.000	premodifier, numeral
cristales	cristal	nominal head, plural noun, single-word noun phrase
Swarovski	Swarovski	nominal head, proper noun, single-word noun phrase, sentence boundary

Note: The Connexor Machinese demos are intended for evaluation purposes only.

Ilustración 31. Importancia del etiquetado morfológico para el análisis sintáctico.

2.2.2. Estrategias en morfología computacional

En cuanto a las *estrategias de análisis morfológico* (Martí y Castellón 2000:62 y ss.), o métodos utilizados, varias son las posibilidades, aunque siempre hay que tener en cuenta que las diferentes lenguas del mundo no presentan la misma morfología y, por lo tanto, habrá técnicas más adecuadas para una lengua que para otra.

Por otra parte, los procesos morfológicos tratados también pueden hacer variar la estrategia: no es lo mismo diseñar un programa que dé cuenta de la flexión que uno que lo haga de la derivación o de la composición.

Asimismo, las unidades de trabajo también condicionarán la estrategia que se vaya a adoptar. Recordemos los problemas que

pueden plantear, a la hora de segmentar, números, locuciones, formas verbales compuestas y perifrásticas, etc.

Básicamente, lo que se espera del tratamiento morfológico es que ante una cadena de caracteres o una secuencia de fonemas, el sistema nos proporcione un análisis de la palabra en términos de sus elementos componentes (*cf.* Trost 2003:38).

Para conseguir este objetivo se han propuesto diferentes estrategias:

1) *Diccionario de formas* o *formario*. La técnica o estrategia más sencilla y menos sofisticada es el *pattern-matching*, recurrir a diccionarios en los que se almacenan listas de formas con información morfológica asociada. El algoritmo[113] únicamente tiene que relacionar ('match') la secuencia de caracteres de entrada con un lema del diccionario ('pattern'). De esta forma, es aplicable a todo tipo de palabras. P. ej. (adaptado de Hausser 2001:252):

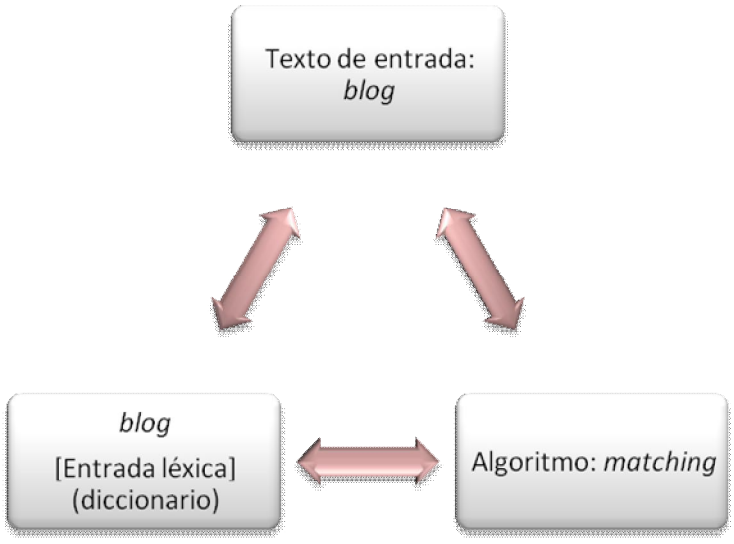


Ilustración 32. "Pattern-matching" de la forma "blog".

[113] *Algoritmo* es un término tomado de las matemáticas que designa un conjunto de instrucciones para solucionar un problema. En este caso, indica al programa informático todos los pasos que ha de seguir para el análisis o la generación de las palabras, es decir, la forma de ir identificando los morfemas hasta llegar al lema o, viceversa, la forma de llegar del lema a la palabra que se quiere generar.

Históricamente, esta estrategia fue la primera en surgir, debido a que la lengua que se tomaba como referencia para el tratamiento computacional del lenguaje era el inglés, lengua con una morfología flexiva "pobre" en comparación con otras lenguas. Como es evidente, el tratamiento morfológico propiamente dicho es prácticamente inexistente: no hay segmentación de la palabra en sus componentes.

Si bien esta técnica es eficiente para lenguas como el inglés, con poca variación morfológica, ya que permite listar en el diccionario las palabras con su flexión (*blog*, *blogs*), para otras lenguas, como el español, que presentan una flexión más compleja, plantea muchos inconvenientes: incluir todas las formas flexivas en el diccionario no sería práctico e impediría utilizar los diccionarios convencionales, que se basan en lemas. Además, esta estrategia introduce redundancia en los sistemas computacionales y resulta ineficaz para tratar formas no incluidas en el lexicón o neologismos. Por otra parte, tampoco da cuenta de la derivación y de la composición, que simplemente trataría como unidades que formarían parte del lexicón, pero sin descomponerlas en sus partes.

Con esta técnica, si la palabra objeto de análisis no está presente en el *formario*, no será reconocida como válida. P. ej. si buscamos la palabra *azules* en un diccionario cualquiera del español no la vamos a encontrar. En el caso de la morfología verbal, el problema aun se complica más (*azular*, *azulo*, *azularé*…). Utilizar este método aumentaría innecesariamente el tamaño del lexicón y haría poco eficiente la implementación computacional. Por lo tanto, en morfología computacional es preciso diseñar estrategias que den cuenta de la relación entre *azul* y *azules* o entre *azular*, *azulo*, *azularé*, sin necesidad de incluir todos los paradigmas flexivos de estas formas en el lexicón. Esto se consigue mediante la formalización de reglas morfotácticas (reglas

que describan la formación de plural en los adjetivos o la flexión verbal, por ejemplo).

Compárese la diferencia entre el diccionario necesario para el inglés y el requerido para el español:

Entrada del diccionario – Inglés	Información morfológica asociada	Entrada del diccionario – Español	Información morfológica asociada
Read	Verbo presente 1ª, 2ª persona singular y plural, 3ª persona plural	*Leo*	Verbo presente indicativo 1ª persona singular
Reads	Verbo presente 3ª persona singular	*Lees*	Verbo presente indicativo 2ª persona singular
		Lee	Verbo presente indicativo 3ª persona singular
		Leemos	Verbo presente indicativo 1ª persona plural
		Leéis	Verbo presente indicativo 2ª persona plural
		Leen	Verbo presente indicativo 3ª persona plural

Tabla 6. Diferencias entre un lexicón para el inglés y otro para el español.

2) *Diccionario de lemas* (“*stemming*” o “*stripping*”) y *de raíces y afijos*. Ante los inconvenientes de la técnica anterior, se puso de manifiesto la necesidad de un verdadero tratamiento morfológico que, mediante reglas, pudiera establecer relaciones entre el lema (o la raíz) y los afijos. Así, en vez de emplear diccionarios con todas las formas, se aprovecha la ventaja que supone contar con diccionarios preexistentes. Puesto que estos toman como principio para recoger palabras los lemas (forma canónica, normalmente la forma base), el algoritmo en este caso

relacionaría el *input* con el lema del diccionario pero, a diferencia del acercamiento ya comentado, tendría que proporcionar una interpretación morfosintáctica, es decir, hay un paso previo (análisis morfológico) entre el *input* y su asociación con un lema, aquel que separa los afijos de la secuencia de caracteres de entrada.

En este acercamiento el algoritmo es sencillo, pero más sofisticado que en el caso anterior. El método de análisis consiste en reducir una palabra a su lema mediante la sucesiva eliminación de afijos. Así, ante el *input*, el algoritmo no busca directamente en el diccionario el término, sino que procede a la segmentación de la palabra en sus componentes, que son los que consulta en el diccionario y, entonces, lleva a cabo la concatenación. La forma de proceder es por eliminación: una vez identificado, por ejemplo, el sufijo flexivo, suprime el afijo y consulta el resultado en el diccionario. Si la entrada está presente, el análisis tiene éxito (*cf.* Martí y Castellón 2000:62), como en el caso de *azules*. Si no, debe proseguir y eliminar ('strip') el siguiente afijo, y así sucesivamente hasta dar con una forma que esté incluida en el diccionario, de la que producirá el análisis morfológico relevante. Pero si el lema no está en el diccionario (*cosmopaletos*), no hay reconocimiento:

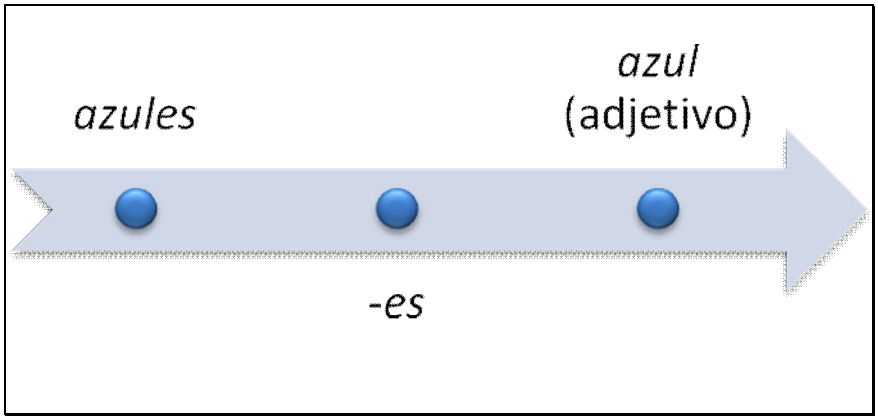


Ilustración 33. Reconocimiento de la forma "azules", tras eliminar el afijo flexivo "–es".

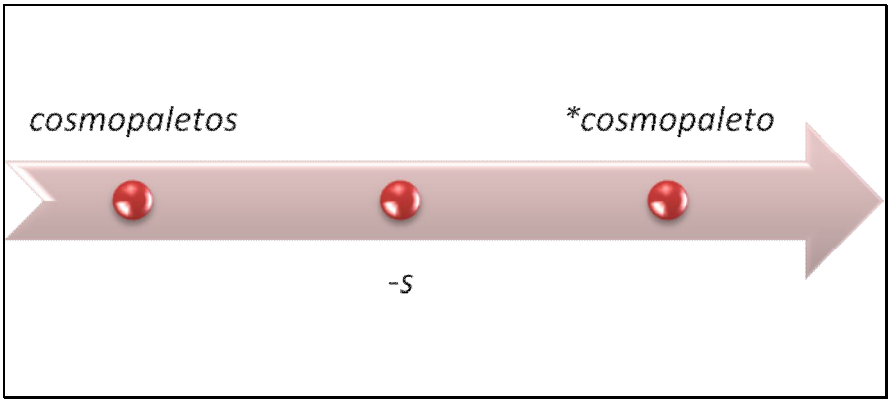


Ilustración 34. No reconocimiento de la forma "cosmopaletos", al eliminar el afijo "-s".

En este caso sí se analiza la estructura interna de las palabras, es decir, hay segmentación, ya que la secuencia de caracteres se descompone en sus partes constitutivas y se explicita la información asociada a las reglas que intervienen en el algoritmo. Asimismo, con este método se elimina la redundancia, ya que no son necesarias varias entradas para palabras del mismo paradigma flexivo. Basta con introducir *azul* en el diccionario: a *azules* se llega mediante la aplicación de las reglas flexivas. Por otra parte, se aprovechan los grandes diccionarios en formato electrónico, cuyas entradas son ya palabras de la lengua, como ocurre en los diccionarios en que se sustentan los programas de corrección ortográfica o los tesauros de los procesadores de texto.

Pese a su mayor utilidad, esta estrategia se enfrenta al problema de la ambigüedad presente en los sufijos. En el ejemplo de *azules*, *-es*, además de un morfema nominal, también es un morfema verbal en español, lo que daría lugar, en un texto dado, a múltiples análisis innecesarios.

Otro problema al que se enfrenta esta técnica son los casos en que la concatenación no es posible (morfología no concatenativa), como cuando ha actuado un proceso de sustitución: *soy-eres-es-somos-sois-son*; *mouse* 'ratón'-*mice* 'ratones'; o la palabra es invariable: *crisis-crisis*. En

estos casos, se hacen necesarios mecanismos ad hoc específicos de la lengua en cuestión para manejar las excepciones y relacionarlas con su lema correspondiente, igual que ocurre con los fenómenos morfofonológicos, que no son contemplados por este modelo.

Otras críticas más generales que se han efectuado a este acercamiento se refieren a que:

- se centra en el análisis, descuidando la generación, para la que se requiere un algoritmo completamente diferente;
- los algoritmos están condicionados por la lengua, al incluir tanto las reglas de concatenación como el tratamiento de las excepciones;
- se desarrolló pensando en un determinado tipo de análisis morfosintáctico, de ahí las dificultades para tratar con los morfos y alomorfos.

Los *diccionarios de raíces* y *afijos*, que se pueden emplear como alternativa a los diccionarios basados en lemas, son más complejos desde el punto de vista de su concepción, ya que no existen recursos previos de este tipo que se puedan aprovechar y hay que construirlos, por tanto, desde cero. En este caso, se clasifican los elementos atendiendo a sus posibilidades combinatorias: cada raíz lleva asociada información sobre su lema y el modelo flexivo que admite (afijos con los que se puede combinar), mientras que los afijos se asocian a su vez con un modelo morfológico y aportan la información gramatical relevante (*cf.* Martí y Castellón 2000:63). Esta es la estructura que encontramos en MACO[114], analizador morfológico de la lengua

[114] Analizador desarrollado por el *Laboratori de Recerca en Lingüística Computacional* de la Universitat de Barcelona y por el grupo de Lenguaje Natural de la Universitat

española estructurado en torno a un diccionario de raíces, un diccionario de sufijos y un conjunto de reglas morfotácticas:

Raíz	Modelo	Lema
Dorm-	M1	Dormir
Duerm-	M2	Dormir
Durm-	M3	Dormir
Salt-	M4	Saltar
Am-	M4	Amar
Zurr-	M4	Zurrar
Ventan-	NF	Ventana

Tabla 7. Diccionario de raíces de MACO.

Sufijo	Modelo	Atributos
-o	1 IP	P=1 T=P N=S
-o	2 MS	G=M N=S
-a	1 IP	P=3 T=P M=I N=S
-a	2 IMP	P=2 M=IMP N=S
-a	3 FA	G=F N=S
-as	1 FA	G=F N=PL

Tabla 8. Diccionario de sufijos de MACO[115].

Ambos diccionarios se relacionan entre sí mediante una serie de *reglas* implementadas en un algoritmo que relaciona las raíces con los afijos y que asigna una categoría gramatical a cada palabra. Estas reglas dan cuenta de la buena formación de las palabras. En el caso de MACO,

Politécnica de Catalunya, válido tanto para el análisis como para la generación. *Vid.* URL: http://gedlc.ulpgc.es/links/Oeilte2.htm

[115] Claves:

-o: 1 IP (indicativo presente) = primera (1) persona (P) tiempo (T) presente (P) número (N) singular (S)
-o: 2 MS (masculino singular) = género (G) masculino (M) número (N) singular (S)
-a: 1 IP (indicativo presente) = tercera (3) persona (P) tiempo (T) presente (P) modo (M) indicativo (I) número (N) singular (S)
-a: 2 IMP (imperativo) = segunda (2) persona (P) modo (M) imperativo (IMP) número (N) singular (S)
-a: 3 FA = género (G) femenino (F) número (N) singular (S)
-as 1 FA = género (G) femenino (F) número (N) plural (PL)

por ejemplo, existe una regla que establece que el modelo de raíz M4 se combina con los modelos de sufijo IP e IMP para formar palabras correctas en español.

Esta estrategia de trabajo en morfología computacional permite reducir considerablemente el tamaño de los diccionarios, al no tener que listar todas las formas posibles de las palabras. Por este motivo, resulta atractiva para lenguas con una flexión rica como el castellano. Además, permite dar cuenta también de la composición y de la derivación, y es capaz de reconocer neologismos, siempre y cuando estos se ajusten a las reglas de formación de palabras previamente especificadas.

3) *Sistema de dos niveles o morfología de dos niveles*. Esta tercera estrategia fue propuesta por K. Koskenniemi en 1983 para el finés. A diferencia de la técnica anterior, que únicamente opera en un nivel, este autor se inspira en la fonología generativa (*cf.* Chomsky y Halle 1968), en la que mediante reglas de reescritura que operan secuencialmente las representaciones fonológicas abstractas (nivel léxico) se convierten en formas superficiales (nivel superficial) pasando por una serie de representaciones o niveles intermedios[116].

La forma general de las reglas es: x → y / z _ w. Es decir, "x" se sustituye por "y" en el contexto "z _ w". Las reglas funcionan para el análisis (en sentido descendente, de la forma abstracta a la forma superficial), pero en la generación (en sentido ascendente, de la forma superficial a la abstracta) pueden producir resultados ambiguos o formas inexistentes en la lengua en cuestión.

[116] Según A. MARTÍ e I. CASTELLÓN (2000:63), este método se ha convertido, en los últimos tiempos, en el modelo estándar para el análisis morfológico, sobre todo cuando se habla de lenguas aglutinantes, aunque no consideran las autoras que pueda aportar demasiado al tratamiento de las lenguas flexivas, como el español o el catalán.

Obsérvese la siguiente ilustración, tomada de L. Karttunen y D. R. Beesley (2005:73)[117].

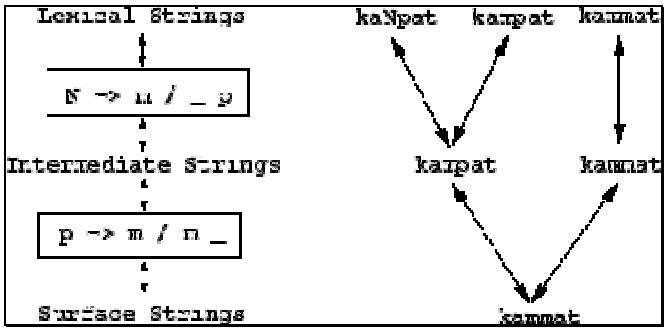


Ilustración 35. Ejemplo de regla de dos niveles.

De modo análogo, Koskenniemi propone distinguir dos niveles[118] en el análisis morfológico:

- Un *nivel léxico*: que da cuenta de la concatenación o combinación válida de morfemas almacenados en el diccionario y también actúa como filtro para relacionar una palabra cualquiera en el nivel superficial con el léxico. Se trata de una representación abstracta, la estructura profunda de la lengua. Cada entrada del diccionario consta de: i) expresión léxica o secuencia de caracteres; ii) clases de continuación o morfemas que pueden concatenarse con la entrada; iii) información morfológica. Por otra parte, el léxico está

[117] La aplicación sucesiva de dos reglas de reescritura: N → m /_ p y p → m / m _ lleva, a partir de la forma léxica *kaNpat* a *kammat* a través de la representación intermedia *kampat*. Pero si se opera en sentido inverso y el análisis parte de la forma superficial *kammat*, las mismas reglas arrojan tres resultados posibles: *kaNpat*, *kampat* y *kammat*. Esto se debe a que, en el primer caso, el orden de aplicación de las reglas es fijo y conocemos la meta a la queremos llegar (determinismo), por lo que una forma léxica solo puede generar una forma superficial; pero cuando se invierte el proceso, una única forma superficial puede haber sido producida de diferentes maneras, ya que no sabemos cuál es la meta (no determinismo) y, por lo tanto, debemos contemplar diferentes posibilidades.

[118] Pues su objetivo fundamental era describir los fenómenos morfofonológicos, aunque la posterior evolución del modelo le ha permitido tratar aspectos puramente morfológicos.

estructurado en subcomponentes, según las propiedades morfotácticas o morfemas con los que se puede combinar cada entrada. Sirva de ejemplo la estructura de la sección del lexicón que incluye los nombres regulares en la formación del plural, los nombres con una forma de plural irregular, los nombres con una forma irregular en el singular y el morfema regular de plural en inglés (*cf.* Jurafsky y Martin 2000:67):

Reg-noun	Irreg-pl-noun	Irreg-sg-noun	Plural
cat	sheep	sheep	-s
dog	mice	mouse	

Tabla 9. Ejemplo de la estructuración del lexicón.

- Un *nivel superficial* (fonológico y ortográfico), que refleja la realización del nivel léxico en forma de palabras concretas. P. ej. *cats*, *dogs*.

Nivel léxico	cat +N +PL
Nivel superficial	cats

Tabla 10. Ejemplo de representación léxica y representación superficial.

Un conjunto de *reglas declarativas* en forma de afirmaciones lógicas se encarga de establecer las correspondencias entre los dos niveles. Las reglas actúan directamente como restricciones, bien sobre el nivel léxico, bien sobre el superficial, o sobre ambos, es decir, establecen condiciones en los morfos en su paso de un nivel de representación a otro. De este modo se describen las clases de alternancias que suceden en las formas superficiales de los morfemas según los diferentes contextos en los que aparecen. Las reglas se componen de los siguientes elementos:

a) *Correspondencia*: consta de un par de caracteres x:y, donde el primero pertenece al nivel léxico (x) y el segundo es su realización en el nivel superficial (y). Estos casos también pueden ser generalizaciones, como, por ejemplo, "C" como símbolo de cualquier carácter consonántico o "V" como una abstracción de cualquier carácter vocálico. Si el contexto no afecta a la correspondencia, la información expresada en el par de caracteres se aplica por defecto. Las cadenas de caracteres en ambos niveles deben ser de la misma longitud, por lo que se postula el carácter cero para dar cuenta de aquellos fenómenos de inserción o de epéntesis. De esta forma se asegura siempre la correspondencia entre los dos niveles.

b) *Contexto*: especifica la situación en la que se verifica la correspondencia mediante expresiones regulares (por ejemplo, los paréntesis indican alternancias; los corchetes, una secuencia, etc.). Se tiene en cuenta el contexto anterior y el posterior. Y se formaliza mediante expresiones regulares.

c) *Operador*: expresa el tipo de relación que existe entre el contexto y el par de caracteres de la correspondencia.

Por ejemplo, la regla 0 : e <=> cons + : s # expresaría que 0 (cero, que se concibe como un símbolo más en este modelo) en la forma léxica se cambia por "e" en la forma superficial (inserción) si y solo si (operador <=>) hay una consonante antes del símbolo "+" y una "s" a continuación que, además, es el final de la palabra (#, contexto). Mediante esta regla sería posible en español dar cuenta de la formación del plural en nombres que terminan en consonante en su forma de singular, del tipo: *bombón* > *bombones*, en los que es preciso insertar una "e" al añadir el morfema de plural: *bombón* + *s* > *bombón* + *e* + *s*, a

diferencia de la formación regular del plural, en la que basta con añadir la "s": *vaso + s > vasos*[119].

#	a	m	o	r	+	s	#
|	|	|	|	|	1	|	|
0	a	m	o	r	e	s	0

Ilustración 36. Funcionamiento de una regla de dos niveles.

En el siguiente ejemplo de L. Karttunen y D. R. Beesley (2005:75), existen dos restricciones o reglas, destacadas con cuadrados: una afecta al nivel léxico y la otra al nivel superficial:

Ilustración 37. Ejemplos de reglas de dos niveles.

- N:m <=> _ p: indica que "N" en el nivel léxico debe sustituirse por "m" en el nivel superficial cuando va seguida, en el nivel léxico, de "p".
- p:m <=> m _ por su parte, establece que "p" en el nivel léxico se sustituye por "m" en el nivel superficial, cuando le precede una "m" en este nivel.

[119] *Vid.* TROST (2003:41-42 y 44-45) para más ejemplos de reglas de dos niveles.

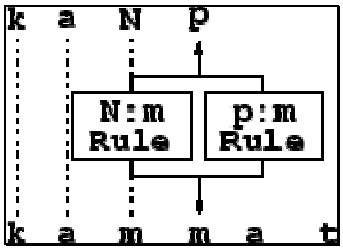


Ilustración 38. Resultado de la aplicación de las dos reglas anteriores.

Este ejemplo nos permite comentar otra de las características más destacadas del modelo de K. Koskenniemi: las reglas que relacionan los símbolos de los dos niveles operan en paralelo, no secuencialmente como ocurre en la fonología generativa. Es decir, no se aplica primero una regla y después otra, sino que todas se aplican de forma simultánea. De esta manera, aunque no se resuelve del todo el problema de la ambigüedad, se facilita su tratamiento: una "m" en el nivel superficial puede ser el resultado del funcionamiento de dos reglas diferentes (N:m y p:m), lo que implicaría postular como posibles formas léxicas "kammat" (en la que las correspondencias se llevarían a cabo sin ningún tipo de restricción), "kampat" (en la que habría operado la restricción p:m) y "kaNpat" (en la que, además de la anterior, habría operado la restricción N:m).

Para evitar cualquier problema relacionado con la "sobregeneración", es decir, resultados no correctos derivados de la aplicación "ciega" de las reglas, la morfología de dos niveles dispone de otra característica interesante: la consulta al diccionario de formas léxicas y el análisis de las formas superficiales se realizan de forma simultánea, de tal manera que el lexicón actúa en realidad como un filtro. Es decir, cada vez que se aplican las reglas se comprueba que la forma léxica resultante es una palabra de la lengua objeto de trabajo, lo que impide que se sigan análisis que lleven a palabras inexistentes. En el ejemplo de antes, para llegar de la forma superficial *kammat* a la

léxica *kaNpat*, el analizador ha tenido que atravesar la cadena del diccionario que contiene la secuencia *kaN*, por lo que la regla N:m bloquearía las posibilidades “kammat” y “kampat”, ya que solo va a contemplar pares de caracteres cuya parte léxica tenga correspondencia con alguno de los arcos que salen del estado en el que se encuentra el procesador en ese momento. Si el análisis no tiene una contrapartida en el lado léxico, se detiene (*cf.* Karttunen y Beesley 2005:76):

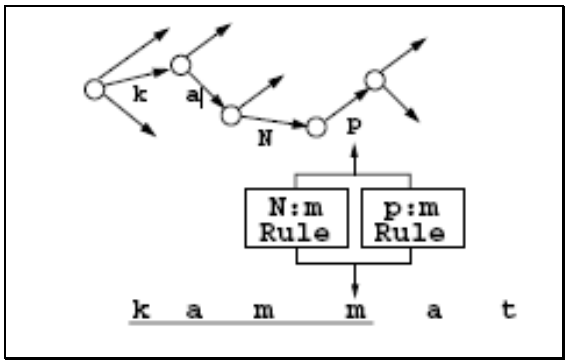


Ilustración 39. Funcionamiento en paralelo de las reglas.

Las reglas de este modelo se implementan computacionalmente como un tipo de autómatas que se encarga de realizar las transformaciones, los llamados *transductores de estados finitos*, que actúan en paralelo. Los transductores de estados finitos se fundamentan en el concepto de *expresión regular*, que es una notación algebraica habitual en informática para caracterizar textos entendidos como secuencias de caracteres alfanuméricos. Fue propuesto por el matemático norteamericano S. C. Kleene en 1956[120].

[120] Este concepto es uno de los pilares en los que se basa el funcionamiento de buscadores web, sistemas de recuperación de información, procesamiento de palabras, cómputo de frecuencias en corpus, etc. (*vid.* JURAFSKY Y MARTIN 2000: cap. 2). Trabaja con una serie de operadores (?, *, ^, …) con los que se establecen los patrones de búsqueda que el sistema es capaz de reconocer y, por tanto, recuperar (**“**pattern-matching**”**). Además de este uso práctico, también se emplea como metalenguaje para describir un tipo de lenguajes formales, las llamadas *lenguas regulares*, es decir,

Las expresiones regulares se implementan en forma de *autómatas de estados finitos*, un mecanismo matemático fundamental en LC, ya que a través de sus diferentes variaciones forma parte de sistemas de reconocimiento y síntesis del habla, de correctores ortográficos o de sistemas de extracción de información. Se suelen representar como grafos que constan de estados o nodos (estado inicial, estados intermedios, estado final) unidos mediante arcos etiquetados que marcan las transiciones de un estado a otro. El reconocimiento o generación de un patrón o secuencia de caracteres se produce cuando es posible recorrer todos los estados, desde el inicial hasta el final:

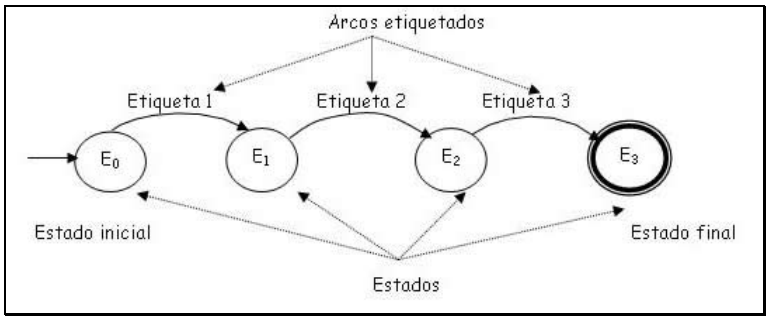


Ilustración 40. Autómata de estados finitos.

Ahora bien, los autómatas de estados finitos son útiles para definir un lenguaje formal mediante un conjunto de caracteres. Pero cuando interesa definir relaciones entre conjuntos de caracteres, entonces se recurre a los *transductores de estados finitos*, un tipo de autómatas capaz de relacionar dos conjuntos de caracteres: leen un conjunto de caracteres como entrada y generan otro diferente como salida. Además, son capaces tanto de reconocer como de generar pares de cadenas de caracteres.

---

aquellas que se pueden caracterizar mediante expresiones regulares. Este tipo de lenguajes se utiliza para modelar ciertos aspectos de las lenguas naturales, como la fonología, la morfología o la sintaxis.

El siguiente ejemplo muestra el transductor de estados finitos que da cuenta de la flexión de número de los nombres en inglés (Jurafsky y Martin 2000:74). Así, p. ej. ante la cadena de caracteres de entrada *cats* (nivel superficial) esperaríamos que el analizador nos devolviera la siguiente cadena: *cat* + N + PL (nivel léxico)[121]:

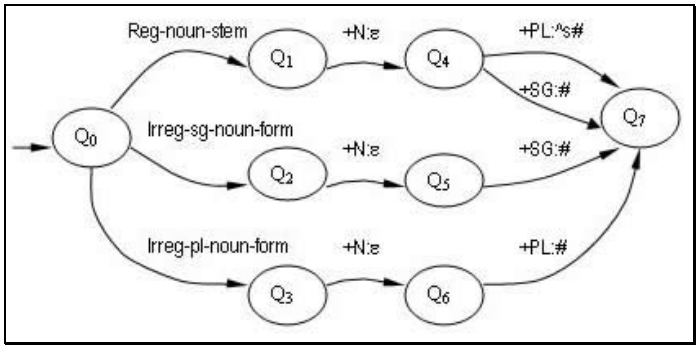


Ilustración 41. Transductor de estados finitos para el inglés.

Un ejemplo de transductor de estados finitos para la morfología del español (*vid*. Hernández, Pérez y Santana 2000b):

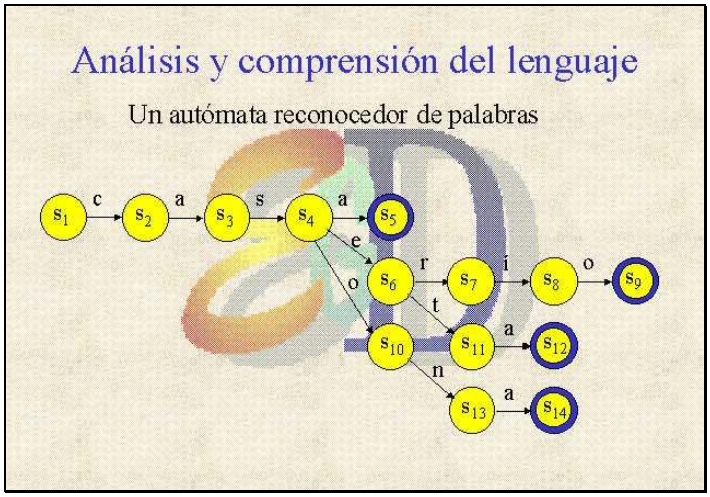


Ilustración 42. Transductor de estados finitos para el español.

[121] Claves:

- Q: estado
- ^: límite de morfema
- ε: cadena vacía
- #: fin de palabra.

Basada en estos principios, la propuesta de la morfología de dos niveles fue pionera para el tratamiento computacional de lenguas con una morfología compleja. Por este motivo, resulta especialmente atractiva para lenguas aglutinantes. Además, aborda aspectos morfofonológicos: alteraciones fonológicas en los morfemas. P. ej. *city* + *s* > *cities*. Asimismo, es una propuesta general que puede ser aplicada a cualquier lengua, ya que combina componentes específicos o dependientes de la lengua (conocimiento lingüístico: reglas y lexicón) y un mecanismo universal aplicable a cualquier lengua (el programa o algoritmo). Otra de sus ventajas es la bidireccionalidad: análisis y generación se llevan a cabo con el mismo procedimiento.

Destaca el programa PC-KIMMO[122] para el tratamiento de la morfología según este acercamiento de dos niveles.

Como resultado del análisis morfológico habremos obtenido unidades léxicas, formas básicas de las palabras tal y como se registran en un diccionario (lemas), así como la información necesaria para que actúen los módulos sintáctico y semántico.

Sin embargo, a veces, el resultado del módulo que da cuenta de la morfología es que a una palabra se le han asignado dos o más análisis morfológicos posibles, pues la morfología se ciñe a la estructura interna de las palabras y, por tanto, no tiene en cuenta el contexto. Pensemos en una palabra como *sobre* en español: ¿qué posibles interpretaciones gramaticales puede recibir? Si la introducimos en el lematizador de Thera obtenemos la siguiente información:

[122] URL: http://www.sil.org/pckimmo/

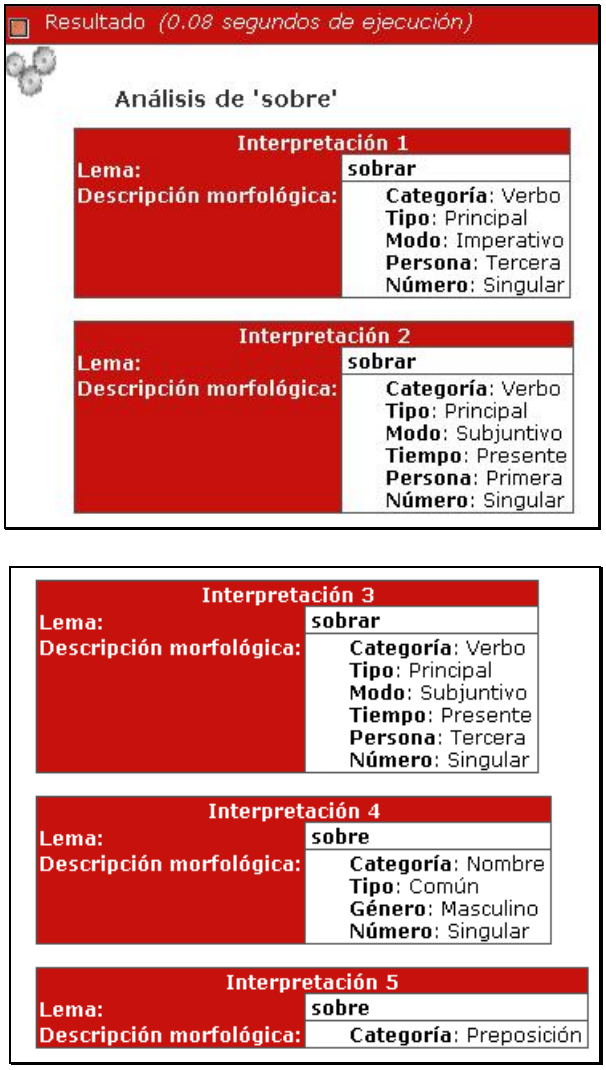


Ilustración 43. Lematización de "sobre".

Por este motivo, antes de proceder al análisis sintáctico y semántico, es necesario llevar a cabo un proceso de *desambiguación morfosintáctica*, del que se ocupan los programas llamados *POS taggers* ("part of speech taggers") o *etiquetadores morfosintácticos*, que evitan que se multiplique el número de estructuras sintáctico-semánticas posibles. Para resolver la ambigüedad, estos programas pueden recurrir a:

- *Conocimiento lingüístico*: mediante reglas que tienen en cuenta el contexto previo y posterior, es decir, la categoría de las palabras vecinas, el *tagger* decide la categoría correcta. P. ej. si a *sobre* le precede un determinante, lo normal es que su categoría sea N. Estos programas arrojan porcentajes de éxito muy elevados: 99,5%, aunque su coste de desarrollo es alto, ya que hay que diseñar las reglas de forma manual. Además, depende de cada lengua y de cada dominio, por lo que es preciso un proceso de adaptación.
- *Conocimiento estadístico*: a partir de la frecuencia de uso en un corpus de muestra, el *tagger* optará por la categoría que sea más frecuente. Los porcentajes de éxito están en torno al 97%, y conllevan un menor coste de desarrollo. Por otra parte, funcionan igual con independencia de la lengua, así que presentan mayor portabilidad.
- *Sistemas mixtos*, que combinan conocimiento lingüístico y estadístico[123].

[123] Para una descripción más detallada del terreno de la morfología computacional, *vid*. KOSKENNIEMI (1983) y SPROAT (1992).

2.3. Sintaxis computacional

Sobre los resultados del módulo morfológico, es decir, sobre las unidades léxicas desambiguadas y la información morfológica a ellas asociada que se ha explicitado mediante un conjunto de etiquetas, actúan los *módulos sintáctico y semántico*[124] (*cf.* Verdejo 1995:59-69):

1) De forma secuencial y separada:

- Primero actúa el módulo sintáctico y, a partir de sus resultados, el semántico.
- Primero actúa el módulo semántico y, después, si se cree necesario, el sintáctico, pues hay teorías que minimizan el papel de la sintaxis.

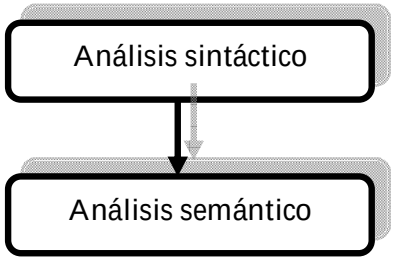


Ilustración 44. Análisis sintáctico y semántico secuenciales.

2) De forma simultánea y conjunta, aunque como módulos independientes. Los módulos sintáctico y semántico actúan en paralelo, dado que a veces es necesario tener en cuenta información semántica para decidir el análisis sintáctico correcto, y viceversa.

[124] El orden y forma de actuación de estos módulos está supeditado a la arquitectura concreta de cada sistema y/o a las necesidades de cada aplicación.

Ilustración 45. Análisis sintáctico y semántico en paralelo.

3) De forma simultánea y conjunta, mediante un único módulo sintáctico en el que las reglas gramaticales llevan asociadas a ellas sus correspondientes reglas de interpretación semántica, de tal forma que el análisis sintáctico está guiado por la información semántica.

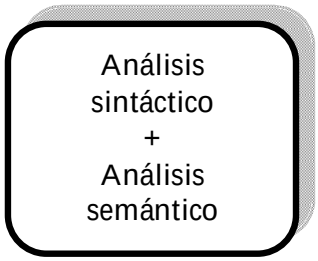


Ilustración 46. Módulo conjunto de análisis sintáctico y semántico.

4) Mediante un módulo único, el semántico, prescindiendo de generar una estructura sintáctica de la secuencia de palabras y, por tanto, eliminando el componente sintáctico, ya sea porque la información semántica se codifica directamente en forma de reglas gramaticales, caso de las gramáticas semánticas, o por realizar un análisis dirigido por la semántica.

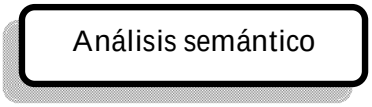


Ilustración 47. Análisis semántico sin análisis sintáctico.

5) Mediante un único módulo, el sintáctico, prescindiendo de generar una estructura semántica y, por tanto, eliminando el componente semántico.

Ilustración 48. Análisis sintáctico sin análisis semántico.

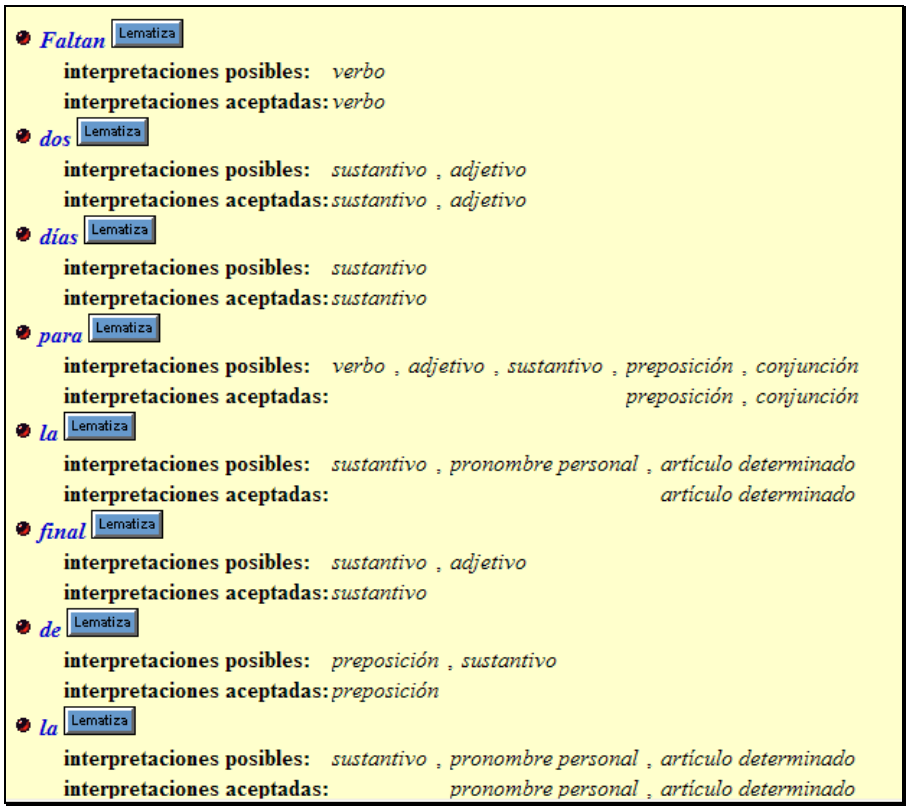
Independientemente de la arquitectura concreta de cada sistema, por lo que se refiere al *módulo sintáctico*, si la morfología se ocupa de la estructura interna de las palabras, la *sintaxis* estudia las relaciones formales de unas palabras con otras dentro del marco de unidades mayores[125]. Se trata de un módulo, igual que el morfológico, también básico en todo sistema computacional; y es el que más atención ha recibido históricamente, de ahí que sea uno de los mejor descritos.

Su punto de partida son las categorías gramaticales o clases de palabras[126]: nombres, verbos, preposiciones… (*cf*. Jurafsky y Martin 2000:288 y ss.), agrupadas a su vez en clases abiertas, clases cerradas, palabras funcionales, etc.

[125] Frases, oraciones o términos equivalentes en función de la teoría lingüística que se adopte como referencia.

[126] Abundan las clasificaciones a propósito del número de clases de palabras, que oscilan entre las ocho que propone Dionisio de Tracia a las ciento cuarenta y seis que manejan algunos etiquetadores para corpus en la actualidad, debido a que esta información es valiosísima para poder determinar la combinatoria de unas palabras con otras: determinante, nombre propio, partícula, gerundio, pronombre interrogativo, coma, punto, etc. Además, también aporta pistas para la pronunciación de las palabras, lo que redunda en, por ejemplo, sistemas de síntesis del habla más naturales; o resulta útil en aplicaciones relacionadas con la recuperación y extracción de información: los buscadores web, de hecho, omiten los determinantes, preposiciones y conjunciones en sus búsquedas, mientras que un programa que haga resúmenes automáticos de documentos se fija en los nombres y verbos para extraer el contenido clave; por último, en el caso de corpus cuyas palabras están etiquetadas gramaticalmente (*vid.* más adelante, el apartado correspondiente), es posible obtener frecuencias de uso de determinados patrones o esquemas sintácticos.

Hay que recordar aquí que el proceso por el que se asignan etiquetas a las palabras de un texto se denomina *tagging* y que, de hecho, la mayoría de las palabras que utilizamos todos los días presentan ambigüedad en las lenguas naturales. Es lo que ocurre en la frase "Faltan dos días para la final de la Liga de Campeones" (*El País*, 25 / 05 / 2009):



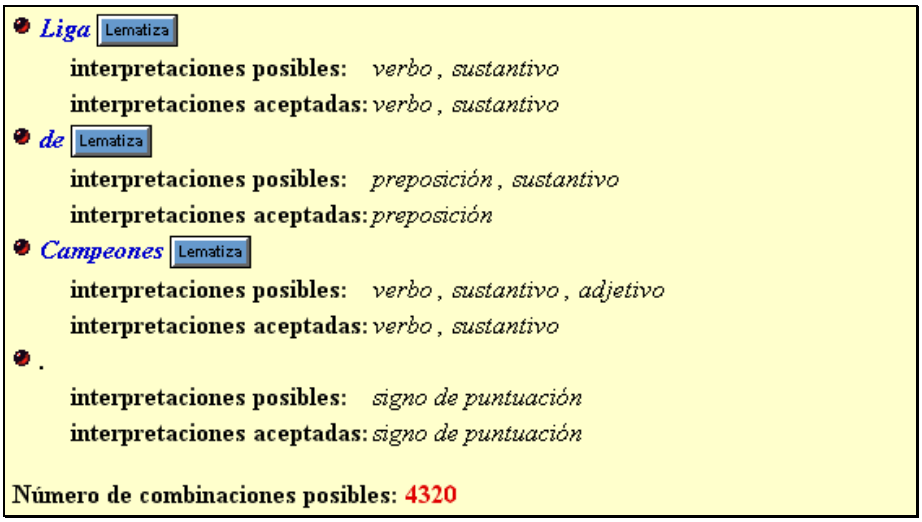


Ilustración 49. Lematización y ambigüedad.

Aunque aparentemente no ofrece grandes dificultades para su análisis, hasta llegar a una combinación de palabras aceptada (en este caso dos), el desambiguador morfosintáctico del Grupo de Estructuras de Datos y Lingüística Computacional de la Universidad de las Palmas de Gran Canaria[127] contempla cuatro mil trescientas veinte posibilidades, debido precisamente a la ambigüedad categorial de las palabras:

Combinaciones aceptadas (2):

1. (*Faltan*) verbo (*dos*) sustantivo (*días*) sustantivo (*para*) preposición (*la*) artículo determinado (*final*) sustantivo (*de*) preposición (*la*) artículo determinado (*Liga*) sustantivo (*de*) preposición (*Campeones*) sustantivo
2. (*Faltan*) verbo (*dos*) adjetivo (*días*) sustantivo (*para*) preposición (*la*) artículo determinado (*final*) sustantivo (*de*) preposición (*la*) artículo determinado (*Liga*) sustantivo (*de*) preposición (*Campeones*) sustantivo

Ilustración 50. Combinaciones aceptadas por el desambiguador[128].

Una vez establecidas las categorías morfosintácticas de las palabras y desambiguadas por alguno de los métodos existentes[129], es posible,

[127] URL: http: / / gedlc.ulpgc.es / investigacion / desambigua / morfosintactico.htm

[128] Observemos que la diferencia entre las dos opciones aceptadas estriba en la categoría propuesta para *dos*, en un caso como sustantivo (opción 1) y en otro como adjetivo (opción 2), a las que se llega tras la desambiguación funcional local y global. En este punto el programa es incapaz de determinar la opción válida entre las dos propuestas. *Vid*. SANTANA *et al.* (2002, 2004, 2005, 2006).

[129] Mediante sistemas basados en reglas confeccionadas de forma manual (por ejemplo, una palabra ambigua será un nombre si está precedida por un determinante), por medio de estadísticas obtenidas de forma automática a partir de un corpus de entrenamiento del que el etiquetador infiere las reglas (asigna a las palabras ambiguas la categoría más frecuente en dicho corpus), o utilizando una combinación de ambas técnicas. Por otra parte, estos etiquetadores también son capaces de adscribir categorías a palabras "desconocidas", es decir, a palabras que no están contenidas en el diccionario que manejan los sistemas, bien por ser muy poco frecuentes, o por ser nombres propios, neologismos, etc. Son tratadas igual que si fueran ambiguas, y se suele utilizar información morfológica (sus terminaciones) u ortográfica (el uso de

finalmente, obtener un análisis sintáctico como el siguiente, en este caso realizado con el analizador sintáctico de Thera:

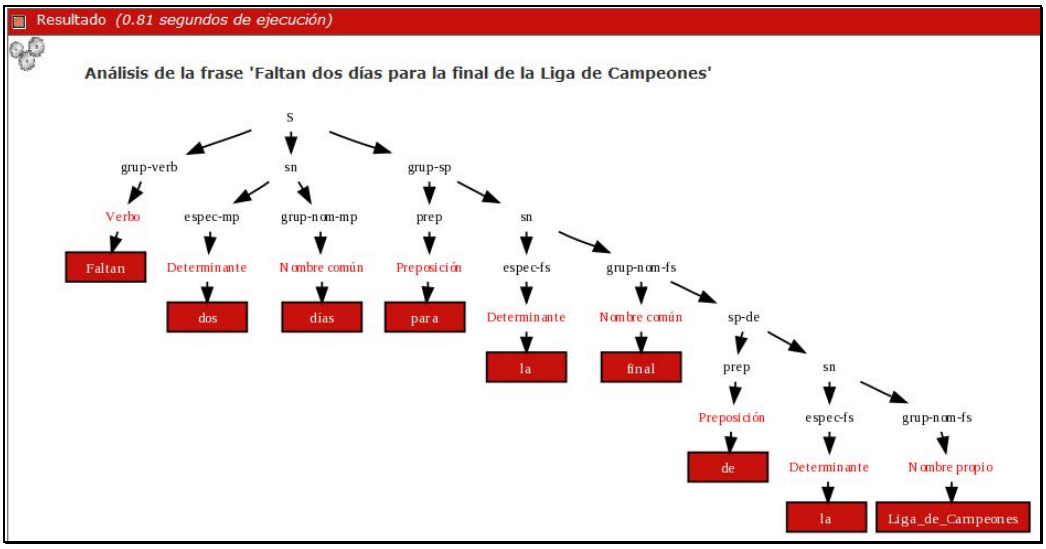


Ilustración 51. Análisis sintáctico con el analizador de Thera.

Sin duda, los trabajos en sintaxis computacional deben mucho a la figura del lingüista N. Chomsky y al papel destacado que este otorgó a la sintaxis en los primeros modelos de la gramática generativa, los cuales influyeron notablemente en la LC por el grado de formalización que introducían, aunque enseguida fueran descartados al no resultar adecuados para el procesamiento computacional de las lenguas naturales que, por otra parte, tampoco formaba ni forma parte del programa de Chomsky. De hecho, uno de los primeros programas de análisis sintáctico (*algoritmo de parsing*) fue el *Transformations and Discourse Analysis Project* (TDAP), desarrollado por Z. Harris entre 1958

mayúsculas iniciales), para deducir su categoría gramatical (*cf.* JURAFSKY Y MARTIN 2000:314).

y 1959 en la Universidad de Pensilvania, que utilizaba catorce reglas compiladas manualmente para llevar a cabo la desambiguación[130].

Por otro lado, hay que mencionar que la necesidad del tratamiento sintáctico en LC surgió por una doble motivación (*cf*. Klavans 1997:676):

1) *Motivación teórica*: se buscaba probar la consistencia de las teorías que proponía la lingüística teórica (sistemas de reglas de aplicación universal), lo que conforma uno de los objetivos teóricos de la disciplina. Desde esta perspectiva, los conceptos clave son: formalización de la teoría –se exige que la teoría sintáctica ofrezca facilidades para ser formalizada, como paso previo a su implementación computacional[131]– y evaluación de los resultados[132].

2) *Motivación práctica*: los sistemas de traducción automática o las interfaces a bases de datos precisaban dar cuenta de este nivel lingüístico para mejorar sus resultados. No eran ya suficientes las traducciones palabra a palabra, centradas casi exclusivamente en cuestiones morfológicas y reglas de reordenamiento local, ni las interfaces basadas en la técnica del *pattern-matching*, que se limitaban a identificar palabras clave en los enunciados de entrada[133] para generar una respuesta. Era preciso dar un paso más en el camino de la emulación computacional del lenguaje.

[130] En las décadas siguientes, el desarrollo de los corpus electrónicos favoreció los acercamientos estadísticos al estudio del lenguaje: para poder manipular cantidades ingentes de datos en un tiempo razonable, la etiquetación manual ya no era una solución válida. Por eso, se desarrollaron etiquetadores y desambiguadores automáticos, capaces de manipular un número mucho mayor de categorías y de reglas con un margen bastante reducido de error.

[131] Requisito que excluye toda teoría lingüística que no esté formulada en términos fácilmente trasladables a un lenguaje informático, de ahí que los caminos de la Lingüística Teórica y de la LC no siempre hayan ido parejos.

[132] El desarrollo de programas informáticos capaces de efectuar análisis sintácticos tiene como finalidad poner a prueba las reglas de la gramática.

[133] *Cf.* el ejemplo del diálogo de Eliza como caso prototípico de este método.

> Desde este acercamiento, los conceptos clave son la eficiencia del sistema –la capacidad para producir resultados de algún tipo– y una cobertura amplia de los fenómenos sintácticos, es decir, una ampliación del universo del discurso en el que se movían para cubrir los casos de enunciados producidos en los más diversos contextos.

En definitiva, los *objetivos* básicos de la sintaxis computacional se pueden resumir en:

- Determinar la *gramaticalidad* o corrección sintáctica de una secuencia de palabras, es decir, si se trata o no de una oración válida de una lengua concreta. En el caso de la generación, producir enunciados que se ajusten a las reglas gramaticales de la lengua en cuestión[134].
- Determinar las *relaciones* (estructura) que se establecen entre las palabras o constituyentes de dicha oración.
- Producir algún tipo de *representación* de la estructura que configuran dichas relaciones. Es decir, generar una interpretación de la oración en términos sintácticos, mostrando la información lingüística pertinente, que puede variar según el tipo de gramática empleado: representación arbórea, estructura de rasgos, formas lógicas, esquemas de actantes, etc.

[134] En ocasiones, se reserva el término de *reconocedores* para los programas que desempeñan esta función, mientras que los que muestran, además, la estructura sintáctica, reciben el nombre de *analizadores* (*cf.* RODRÍGUEZ HONTORIA 2002:92).

Ahora bien, para efectuar el tratamiento sintáctico es preciso contar previamente con un *formalismo* o *lenguaje de representación*, es decir, una información lingüística expresada en términos formales, ya que todo aspecto de una lengua natural que no se pueda formalizar no podrá ser objeto de tratamiento en un sistema computacional. A este propósito es común distinguir entre: conocimiento lingüístico, conocimiento informático y procesamiento lingüístico (*cf*. Badia 2003:196 y ss.).

El *conocimiento lingüístico* lo constituyen la *gramática y* el *léxico*. Por lo general, la gramática consiste en un conjunto de reglas que permite reconocer y generar las secuencias de palabras válidas en una lengua dada, así como describir su estructura sintáctica. Normalmente se define en función de:

- *Vocabulario terminal*: elementos léxicos de la lengua (p. ej. *Sarkozy*, *propone*, *trabajo*, *y*, *autoridad*, *para*, *lograr*, *el*, *cambio*).
- *Vocabulario no terminal*: categorías sintácticas (p. ej. verbo, determinante, nombre común, preposición…).
- *Categoría o símbolo inicial*: categoría superior (p. ej. O = oración; S = *Sentence*).
- *Conjunto de reglas o producciones*: indican las secuencias válidas de palabras. P. ej. la regla SN $\rightarrow$ Det N especifica que un Sintagma Nominal (SN) está formado por un Determinante (Det) y un Nombre (N).

En cuanto al *léxico*, cada vez es más frecuente, en los últimos formalismos gramaticales, que este aporte información fundamental para el análisis sintáctico, como la subcategorización o elementos con los que se puede combinar un elemento del vocabulario. Así, es normal

que un verbo como *amar* se construya con dos argumentos, un SN (sintagma nominal) que actúa como Sujeto y otro SN que lo hace como Objeto Directo. Toda esta información se codifica en el lexicón, donde está disponible para guiar el procesamiento sintáctico cuando se accede al lema correspondiente. Ejemplo de subcategorización para el verbo *amar* (tomado de Moreno Sandoval 1998:81):

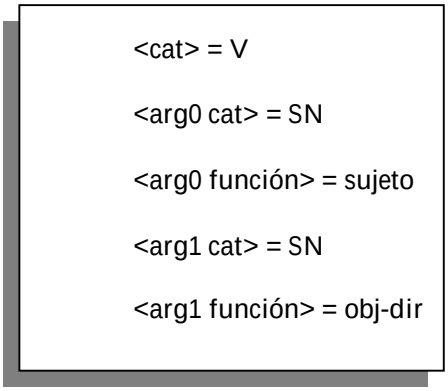


Ilustración 52. Subcategorización del verbo "amar".

El *conocimiento informático* lo conforma el *analizador* o *parser***.** Se trata de un procesador o programa informático que especifica, mediante una serie de instrucciones (*algoritmo*), la estrategia y el orden en que se debe aplicar el conocimiento lingüístico (las reglas de la gramática y la información codificada en el léxico). A veces, como ya se ha apuntado, se distingue entre *reconocedor* y *analizador*, aunque en otras ocasiones el segundo término se emplea indistintamente en ambos casos.

- *Reconocedor*: programa que determina la gramaticalidad de una oración, es decir, si se ajusta a las reglas de la gramática.
- *Analizador*: programa que, además, muestra la estructura sintáctica.

Por último, falta mencionar el *procesamiento lingüístico*: análisis o *parsing*, que se refiere al proceso mediante el cual se lleva a cabo el análisis sintáctico de acuerdo con una gramática y un analizador. Al hacerlo, se genera una serie de estructuras temporales o de trabajo hasta que se llega al análisis definitivo de la oración.

Estas distinciones resultan útiles porque permiten separar el conocimiento lingüístico (gramática y léxico) por una parte y el programa informático (*parser*) por otra. De esta forma, un problema en la gramática o el léxico se puede resolver de forma independiente del *parser*, y viceversa. Los primeros investigadores en LC no trazaron esta división con nitidez. Sin embargo, desde mediados de los ochenta, es habitual hacerlo. Como señala T. Badia (2003:199), este hecho supone una serie de ventajas en términos de eficacia, economía en el número de descripciones necesarias, validez teórica -al permitir aprovechar mejor una teoría lingüística independientemente de la actuación del programa informático- y declaratividad (explicitud del conocimiento).

De manera esquemática, el proceso del análisis sintáctico se puede representar como sigue: dado un texto de entrada, una oración perteneciente a una lengua natural, el procesador o módulo encargado de llevar a cabo el análisis acudirá al conocimiento lingüístico que tenga almacenado (reglas gramaticales, información léxica, etc.) y procederá a analizar el texto (aplicar las reglas de la gramática) de acuerdo con la estrategia especificada en el *parser*. El resultado será una oración analizada sintácticamente, siempre y cuando el texto de entrada se ajuste a las reglas de la gramática de la lengua objeto de análisis.

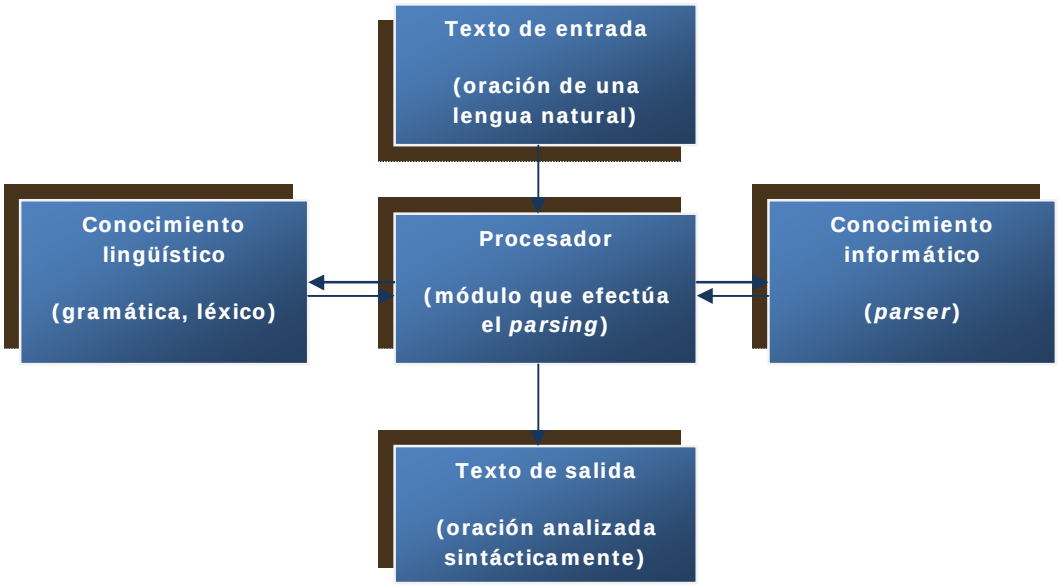


Ilustración 53. Esquema del procesamiento sintáctico de una oración.

Lógicamente, antes habrá actuado el módulo encargado de la morfología y el desambiguador, tal como se aprecia en el siguiente gráfico (tomado del *Centre de Llenguatge i Computació*[135]):

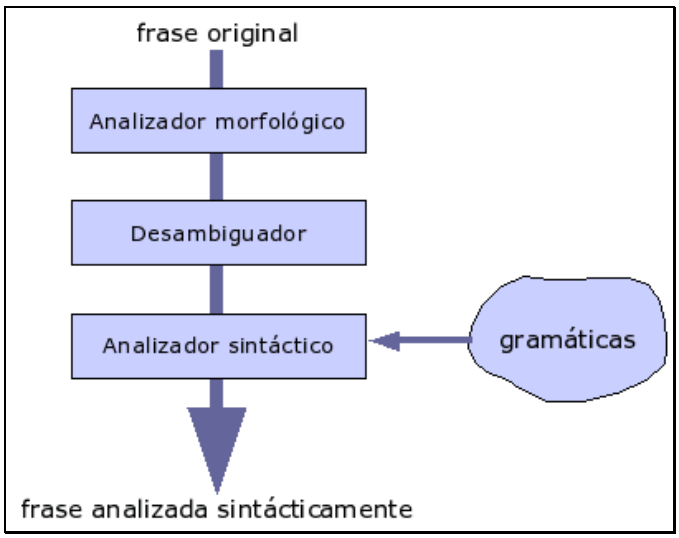


Ilustración 54. Fases previas al análisis sintáctico.

[135] URL: http://clic.ub.edu/

Por lo tanto, ante oraciones como “La crisis monopoliza el debate”, o “La Asamblea francesa ratifica la polémica ley contra la piratería”, un posible análisis sería el siguiente, producido por el analizador del *Centre de Llenguatge i Computació* (Sección Demos > Sintaxis > Analizador):

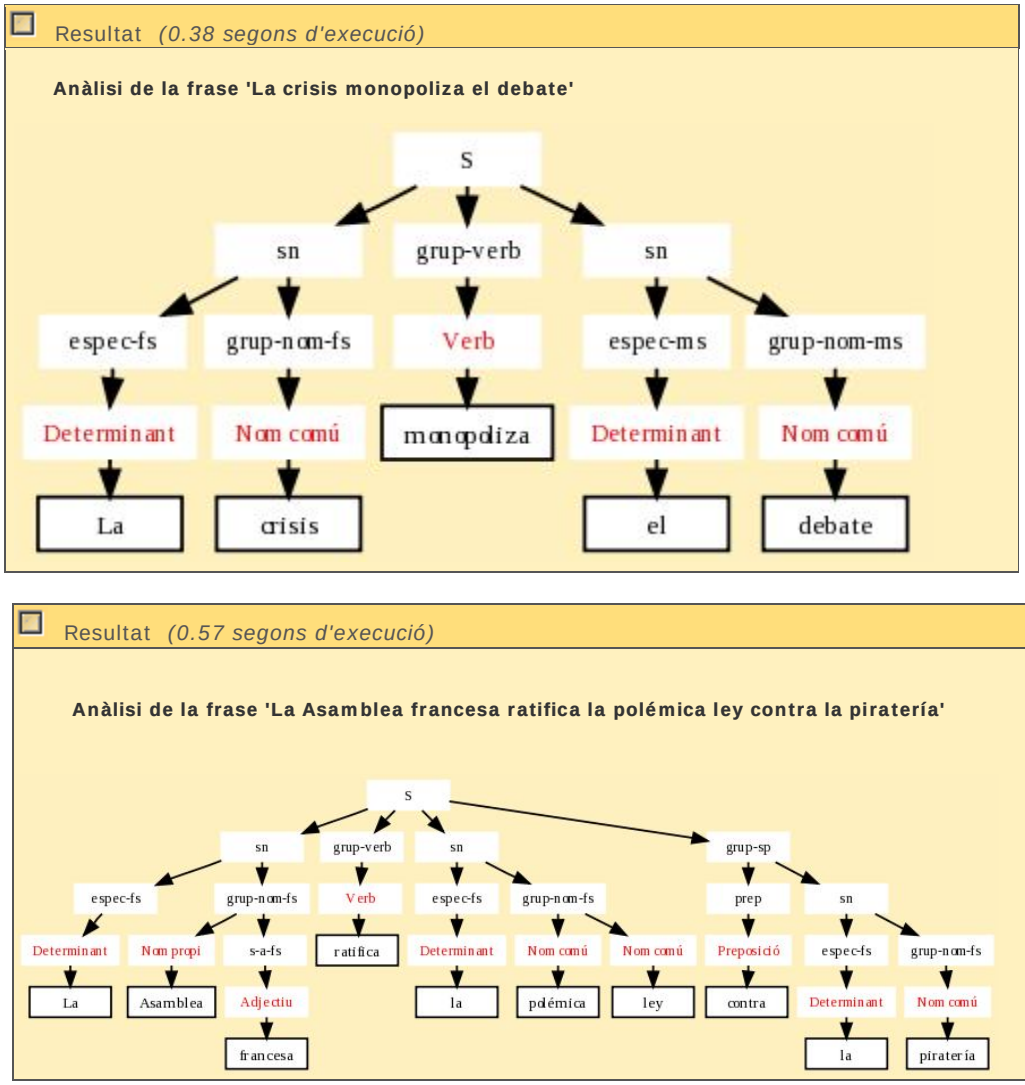


Ilustración 55. Resultados del análisis sintáctico.

Llegados a este punto, cabe plantearse la siguiente pregunta: ¿qué gramática, conocimiento lingüístico, se utiliza como referencia para guiar el análisis sintáctico computacional?

Lo primero que hay que tener en cuenta es que existen diferentes tipos de gramática inspirados en distintas teorías lingüísticas[136]. Es decir, ante una misma secuencia de palabras, en función de la gramática que utilicemos y de la teoría lingüística que sigamos, tendremos una caracterización sintáctica u otra. De ahí la diferenciación que recoge A. Moreno Sandoval (2001:216 y ss.), siguiendo a S. M. Shieber (1988, 1989), entre *teoría*, *formalismo*, *descripción* y *análisis gramatical*:

- *Teoría gramatical*. El conocimiento lingüístico, en concreto la gramática, estará fuertemente condicionado por la teoría lingüística o explicación general sobre el funcionamiento del lenguaje que se asuma. No existe una única teoría que explique dicho funcionamiento, sino múltiples teorías, y cada una de ellas pone el énfasis en determinados aspectos y proyecta su propia visión sobre los fenómenos lingüísticos. Ejemplos de teorías lingüísticas son: la gramática generativa, la gramática funcional, la gramática tradicional, etc. Por lo tanto, el punto de partida es siempre un modelo o teoría lingüística, que establece el marco en el que se inserta el análisis sintáctico.

- *Formalismo gramatical*. Dentro de una teoría gramatical, se denomina *formalismo gramatical* al metalenguaje o notación de que se sirve dicha teoría para expresar la información lingüística y así explicar el funcionamiento del lenguaje: unidades (sintagma, constituyente, cadena…), conceptos (núcleo, dependencia, coordinación…),

[136] Aunque la Lingüística Teórica es la principal fuente de las gramáticas empleadas en LC, no faltan las propuestas de gramáticas dictadas por las aplicaciones o tareas, que se diseñan a medida para un dominio temático muy concreto, o las que se inspiran directamente en los datos, a partir de los cuales, de forma probabilística, se infieren las reglas.

símbolos (O, SN, V, flechas, corchetes, árboles…), reglas (O ⇒ SN SV), etc[137]. En el caso de la LC, los metalenguajes idóneos serán aquellos que se ajusten a los requisitos de formalidad, explicitud, naturalidad, expresividad y facilidad de implementación informática, bien procedan del ámbito de la Lingüística (teorías gramaticales) o de la Informática (lenguajes de programación como Prolog).

- *Descripción gramatical*. Será la especificación formal de las estructuras válidas de una lengua dada de acuerdo con una convención o formalismo, es decir, la gramática particular de una lengua. P. ej. el conjunto de reglas para describir el español. Según A. Moreno Sandoval (1998:143-147), toda gramática se enfrenta a dos problemas generales: la ambigüedad (como la que se deriva de la adjunción de los sintagmas preposicionales, por ejemplo) y la cobertura (la capacidad para producir análisis, que puede verse afectada por la "infrageneración", es decir, que la gramática no dé cuenta de todas las estructuras posibles, o por la "sobregeneración", casos en que las reglas admiten como válidas secuencias de palabras que no lo son).

- *Análisis gramatical*. Es la aplicación de dicha descripción a las oraciones concretas de la lengua para determinar su gramaticalidad y mostrar su estructura, es decir, el análisis sintáctico propiamente dicho.

[137] No obstante, algunos autores identifican *teoría* y *formalismo*. *Vid.* MARTÍ Y CASTELLÓN (2000).

De forma gráfica:

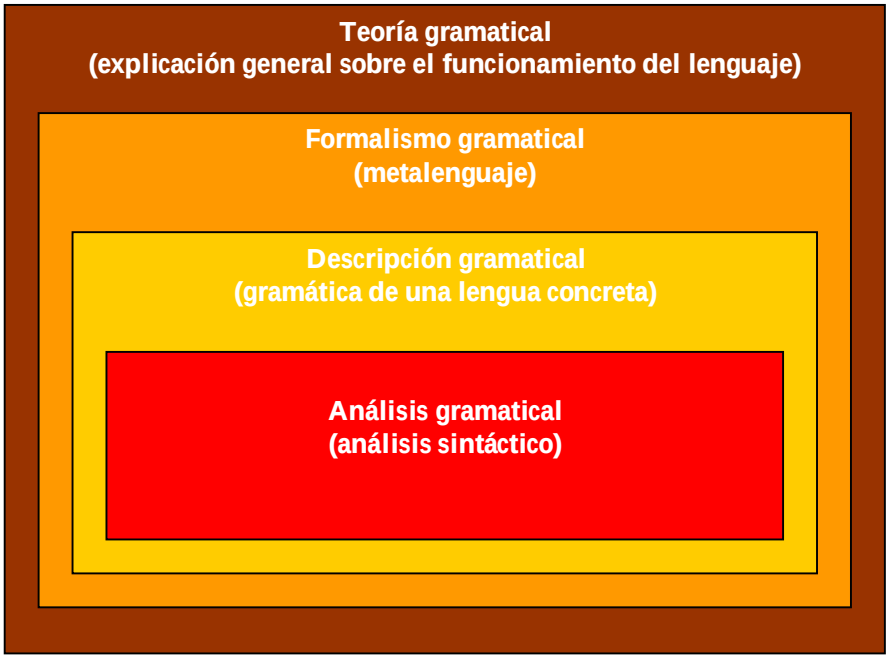


Ilustración 56. Relación entre teoría, formalismo, gramática y análisis sintáctico.

2.3.1. Gramáticas formales y sus tipos

Con independencia de la teoría gramatical que se adopte, el trabajo realizado en sintaxis dentro de la LC no puede entenderse si no se parte de los formalismos gramaticales. Hay que tener presente que desde la perspectiva computacional lo importante es la *adecuación formal* de la gramática[138], uno de los temas centrales en LC Teórica, pues esta cuestión es básica para su *implementación informática*, aunque en la actualidad parece haber cedido terreno a la preocupación por obtener resultados prácticos (*cf.* Moreno Sandoval 1998:59).

[138] En la noción matemática de *gramática formal* subyace parte de la metodología en LC, no solo para describir el conjunto de oraciones posibles de una lengua, sino también las propiedades que presentan en relación con su estructura (sintaxis) y su significado (semántica).

La *teoría de las gramáticas formales* aporta los fundamentos sobre los que se sustenta el diseño de gramáticas para dar cuenta de las lenguas naturales. En Lingüística fue N. Chomsky quien, a mediados de la década de los cincuenta e inspirándose en los lenguajes artificiales de las matemáticas y la lógica[139], se preocupó por formalizar la descripción lingüística. Por una parte, propuso (*cf.* Chomsky 1956) una clasificación de los diferentes tipos de gramáticas formales (la *jerarquía de Chomsky*), base de la *Teoría de los lenguajes formales*, considerada a su vez uno de los pilares de la informática moderna y estrechamente relacionada con la *Teoría de los Autómatas* (*vid.* Fernández y Sáez Vacas 1995).

Por otra parte, Chomsky ha venido desarrollando desde entonces, en modelos sucesivos, la que le pareció más adecuada para tratar las lenguas naturales, la *gramática generativa* -también denominada *gramática de estructura sintagmática* o *de estructura de frase*. Este es el tipo de gramática formal que, sin duda, ha suscitado mayor interés entre los lingüistas computacionales.

---

[139] Hay que señalar el paralelismo entre los trabajos de N. Chomsky en Lingüística y los trabajos previos en Matemáticas, dentro de la corriente "formalista" defendida por D. Hilbert, cuya influencia en la gestación de la gramática generativa no ha sido suficientemente destacada. Las ideas de este matemático pruso-alemán de finales del siglo XIX y principios del XX son precursoras del modelo chomskiano, al proponer un sistema formal que mediante elementos finitos (elementos primitivos) y reglas de inferencia pretendía explicar la gramaticalidad ("buena formación") de las combinaciones de elementos (fórmulas) en axiomas, de acuerdo con unos criterios. Los principios que sustentan este modelo matemático son exclusivamente sintácticos: reglas de inferencia que validan combinaciones de símbolos, sin necesidad de apelar en ningún momento al significado de dichos símbolos. Además, estas reglas introducen la noción de recursividad, que también tendrá un papel relevante en el desarrollo de la gramática generativo-transformacional. La obra del filósofo del Círculo de Viena R. Carnap *The Logical Structure of Language* (1934) recoge las ideas de la corriente formalista al basar su definición de los lenguajes artificiales en ellas. Así, estima que es posible analizar en términos formales la secuencia de palabras supuestamente inglesas "Pirots karulize elatically" como Nombre + Verbo + Adverbio, con total independencia de su significado. En Lingüística, estas ideas formalizadoras están presentes en L. Bloomfield, B. Bloch, Ch. Hockett y, en especial, Z. Harris en lo que a América se refiere, y en L. Hjelmslev en Europa. En definitiva, la teoría de Chomsky se presenta como la culminación de toda una corriente de pensamiento que se había iniciado a finales del siglo XIX (*cf.* TOMALIN 2002).

Una gramática formal como las propuestas por Chomsky, igual que todo *lenguaje artificial*, se caracteriza por las siguientes *propiedades*:

- Explicitud a la hora de especificar las combinaciones de palabras válidas en una lengua dada.
- Expresividad: capacidad para dar cuenta de todas las oraciones posibles de una lengua con un número limitado de reglas y elementos.
- Exhaustividad para describir los fenómenos en detalle.
- Rigor, exactitud en la descripción, de tal modo que la gramática no presente ambigüedad.
- Simplicidad y generalidad en las reglas: elegancia descriptiva.

Siguiendo estos principios, las gramáticas formales se definen mediante un conjunto finito de reglas de reescritura o sintagmáticas que describen las estructuras sintácticas válidas en una lengua. Las reglas son del tipo A → B, donde A representa una categoría sintáctica y B los constituyentes inmediatos de A. P. ej. O → SN SV indica que una Oración está formada por la combinación de un Sintagma Nominal seguido de un Sintagma Verbal. En forma de diagrama arbóreo:

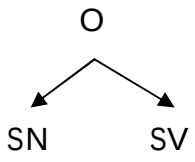


Ilustración 57. Diagrama arbóreo.

Un ejemplo de gramática formal para describir la oración "Ana escucha música" podría ser:

O → SN SV
SN → NP
SN → N
NP → {Ana…}
N → {música…}
SV → V SN
V → V_t
V_t → {escuchar…}

NP = nombre propio

V_t = verbo transitivo

Ilustración 58. Ejemplos de gramática de estructura de frase.

En la citada clasificación, Chomsky estableció cuatro tipos principales de gramáticas formales de acuerdo con una propiedad matemática, el *poder generativo débil* o capacidad de las reglas de la gramática para determinar si una oración es gramatical o agramatical (*vid.* Moreno Sandoval 1998:55 y ss.; 2001: apéndice 2), mientras que el *poder generativo fuerte* hace referencia a la capacidad de la gramática para producir descripciones estructurales.

Poder generativo débil	Gramática	Tipo	Autómata
+ expresivas	▪ Gramáticas enumerables recursivamente o irrestrictas	Tipo 0	Máquinas de Turing
	▪ Gramáticas sensibles al contexto o dependientes del contexto	Tipo 1	Autómatas linealmente finitos
	▪ Gramáticas libres o independientes del contexto	Tipo 2	Autómatas PDS (*Push Down Store*)
- expresivas	▪ Gramáticas regulares	Tipo 3	Autómatas de estados finitos o redes de transición

Tabla 11. Jerarquía de Chomsky.

Para describir las lenguas naturales mediante una gramática formal, esta debe compaginar *expresividad* (construcciones que puede describir) y *no sobregeneración* (restricciones para no admitir como válidas combinaciones agramaticales producidas por la aplicación ciega de las reglas), así como *eficiencia* del *parser*. Es decir, la gramática ideal será aquella que sea lo suficientemente expresiva para dar cuenta de todas las construcciones posibles en una lengua, pero que al mismo tiempo sea suficientemente restrictiva para admitir únicamente aquellas y no otras que sean gramaticales. Además, será fácil de implementar en un programa informático. Por otra parte, también hay que considerar la cobertura o cantidad de texto que puede reconocer.

De los cuatro tipos propuestos por Chomsky, los más empleados en LC son las *gramáticas independientes del contexto* (tipo 2) y las *gramáticas regulares* (tipo 3), es decir, las de menor poder generativo. Y para ambas existen *parsers* eficientes.

Las *gramáticas regulares* se caracterizan por reglas que presentan la siguiente restricción en su forma: solo pueden tener un símbolo no terminal en el lado izquierdo, mientras que en el lado derecho debe haber por lo menos un símbolo terminal. Ejemplos de reglas válidas:

- N → cima
- SN → cima SPrep

Ilustración 59. Ejemplos de reglas de una gramática regular.

Computacionalmente se suelen implementar como *autómatas de estados finitos*[140], un tipo de mecanismo empleado en Informática desde

[140] Es un método procedimental o algorítmico, en el que se explicitan paso a paso las tareas que el programa informático debe realizar para solucionar un problema determinado, en este caso, reconocer una oración. Si las etiquetas de los arcos son

hace tiempo. Y gráficamente se representan mediante unos objetos matemáticos inspirados en la teoría de conjuntos, las *redes de transición*, grafos formados por estados y arcos etiquetados. Un ejemplo de red de transición, para el patrón Art + N + (Adj), es decir, un SN, es el siguiente (Vidal y Busquets 1996:406):

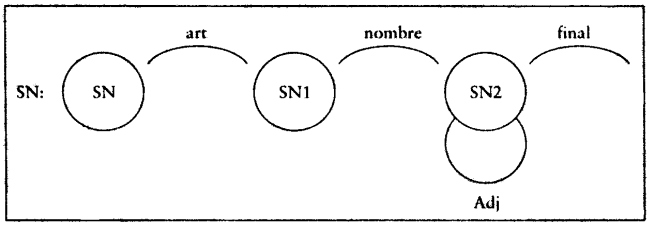


Ilustración 60. Red de transición para reconocer sintagmas nominales.

Dada una secuencia de palabras, esta pertenecerá a la lengua objeto de análisis si es posible trazar un camino desde el estado inicial hasta el final a través de la red. En la red anterior, si introducimos la secuencia "El técnico azulgrana", el autómata primero va a buscar una palabra cuya categoría gramatical sea 'artículo', que es lo que le indica el primer arco; si la encuentra, se produce el reconocimiento y continúa el procesamiento, ahora buscando un nombre, que es la categoría que debe aparecer a continuación, de acuerdo con la gramática; y así sucesivamente, hasta llegar al estado final y dar por concluido el análisis. En la gramática con la que ejemplificamos, como la categoría adjetivo es opcional, existen dos posibles finales: para las secuencias "artículo" + "nombre" el análisis se acabará aquí; si después del nombre aparecen más palabras, el *parser* continuará el análisis buscando un adjetivo, para reconocer secuencias del tipo "artículo" + "nombre" + "adjetivo".

---

símbolos simples, se habla de autómatas; si en estas etiquetas aparecen pares de símbolos, entonces se denominan transductores (*cf.* KARTTUNEN 2003).

El primer círculo representa el estado inicial y corresponderá siempre al primer símbolo terminal de la secuencia de palabras (en el ejemplo, *el*). El resto de círculos representan los diferentes estados de la red (*técnico*). Los arcos que unen los estados están etiquetados con símbolos no terminales (artículo, nombre, adjetivo). Por último, el estado final corresponderá al último elemento de la secuencia de palabras (*azulgrana*).

En el caso de que la secuencia no se ajuste a las reglas de la gramática no se producirá el reconocimiento. Si cambiamos el ejemplo anterior por este otro: "El técnico del Fútbol Club Barcelona", no obtendremos ningún resultado, ya que después del nombre el analizador va a buscar un adjetivo, pero lo que tiene es una preposición.

Como se aprecia en los ejemplos, este tipo de autómata es solo un reconocedor, pero no un verdadero analizador, pues no genera una representación de la estructura sintáctica de la oración, solo nos dice si es válida o no. Por otra parte, presenta problemas para tratar determinados fenómenos, como las construcciones anidadas.

Ante las limitaciones de este tipo de gramática, incapaz de dar cuenta de la regularidad o recursividad presente en las lenguas (p. ej. una oración puede albergar en su interior otra oración, como sucede con las estructuras de relativo), se desarrollaron versiones mejoradas, las *redes de transición recursivas*, en las que las etiquetas de los arcos pueden remitir a otras redes, no solo a una palabra o categoría. Un ejemplo de una red de transición recursiva para reconocer frases del tipo SN + V + SN en castellano sería el siguiente, en el que la etiqueta SN en los arcos llama a la red de transición para reconocer sintagmas nominales (Vidal y Busquets 1996:407):

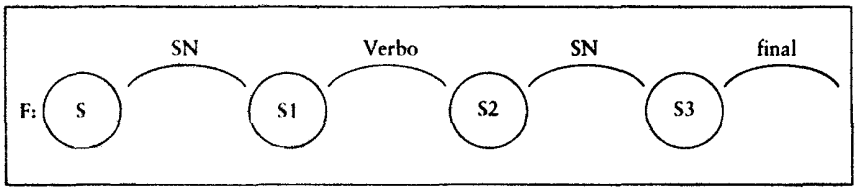


Ilustración 61. Red de transición recursiva.

El problema de este formalismo reside en que puede reconocer como válidas construcciones que no respetan la concordancia ("*Los niños come caramelos") o las especificaciones semánticas del tipo "*Los calcetines cantan habaneras" (Vidal y Busquets 1996:407). Para subsanar este tipo de problemas, W. Woods en 1970 desarrolló las *redes de transición aumentadas*, que son un perfeccionamiento del modelo mediante la introducción de acciones y condiciones en los estados de la red, aumentando así su poder expresivo (*vid.* Grishman 1991 [1986] o Allen 1995 [1987]).

Las redes de transición estuvieron de moda en los años 70 para tratar la sintaxis (*cf.* Grishman 1991 [1986]) y se caracterizan por ser un mecanismo simple pero eficaz desde el punto de vista computacional. Sin embargo, son poco elegantes desde el punto de vista lingüístico, ya que no diferencian elementos procedimentales y declarativos. No obstante, se continúan utilizando en otras facetas del análisis lingüístico, sobre todo en fonología y morfología[141]. Obsérvese el siguiente ejemplo, en el que las redes de transición se combinan con estadísticas para determinar las probabilidades de reconocer o generar oraciones (*cf.* Moreno Sandoval 1998:185-186):

[141] En estas áreas, se combinan con técnicas probabilísticas -como el modelo estadístico de los *n-gramas*, que se basa en la idea de que solo unas pocas unidades anteriores (lo más frecuente, dos o tres, que constituyen el contexto local o *n*) condicionan la probabilidad de aparición de la siguiente- para determinar la categoría gramatical de las palabras en los etiquetadores morfosintácticos o para identificar los fonemas en los sistemas de reconocimiento del habla. Es lo que se conoce como *cadenas de Markov* o *autómatas probabilísticos*.

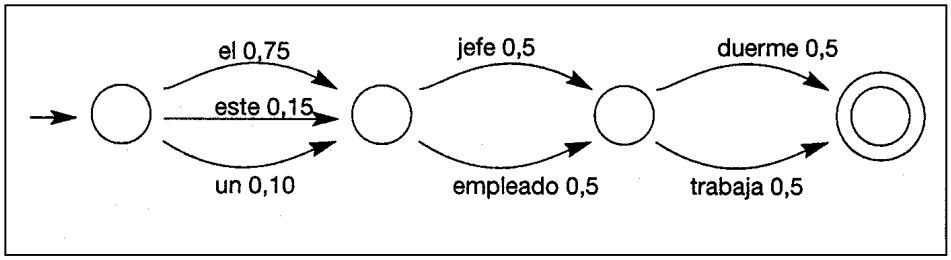


Ilustración 62. Cadena de Markov aplicada al español.

La probabilidad de generar o aceptar una cadena de palabras equivale al producto de las probabilidades de los arcos atravesados (*cf.* Moreno Sandoval 1998:*ibid.*):

Oración	*Probabilidad*
El jefe duerme	0,75 * 0,5 * 0,5 = 0,1875
El jefe trabaja	0,75 * 0,5 * 0,5 = 0,1875
El empleado duerme	0,75 * 0,5 * 0,5 = 0,1875
El empleado trabaja	0,75 * 0,5 * 0,5 = 0,1875
Este jefe duerme	0,15 * 0,5 * 0,5 = 0,0375
Este jefe trabaja	0,15 * 0,5 * 0,5 = 0,0375
Este empleado duerme	0,15 * 0,5 * 0,5 = 0,0375
Este empleado trabaja	0,15 * 0,5 * 0,5 = 0,0375
Un jefe duerme	0,10 * 0,5 * 0,5 = 0,0250
Un jefe trabaja	0,10 * 0,5 * 0,5 = 0,0250
Un empleado duerme	0,10 * 0,5 * 0,5 = 0,0250
Un empleado trabaja	0,10 * 0,5 * 0,5 = 0,0250

Tabla 12. Probabilidades asociadas a oraciones generadas por la cadena de Markov anterior.

Las *gramáticas independientes del contexto*, por su parte, son un tipo de gramática formal que se caracteriza por reglas que presentan menores restricciones: pueden tener cero o más símbolos, terminales o no, en el lado derecho de la regla, aunque en el lado izquierdo solo puede haber un símbolo no terminal, igual que en las anteriores gramáticas. Su nombre se debe a que una regla como N → *equipo* significa que N puede reemplazarse por "equipo" en cualquier lugar en que aparezca, sin importar el contexto, los símbolos que rodean a N.

Una característica destacada es que las reglas independientes del contexto pueden dar cuenta de las estructuras anidadas, es decir, de la recursividad presente en una oración como *El técnico del equipo que ha logrado el triplete es Pep Guardiola* (SN → Det N Orel). También introducen la opcionalidad, mediante el uso de paréntesis como en SN → (Det) N, y la alternancia, mediante el uso de una barra vertical, como en Adj → {*rojo* | *caro*}. Además, generan un diagrama arbóreo que muestra la estructura jerárquica de los constituyentes, como ocurre en FreeLing 2.1[142]:

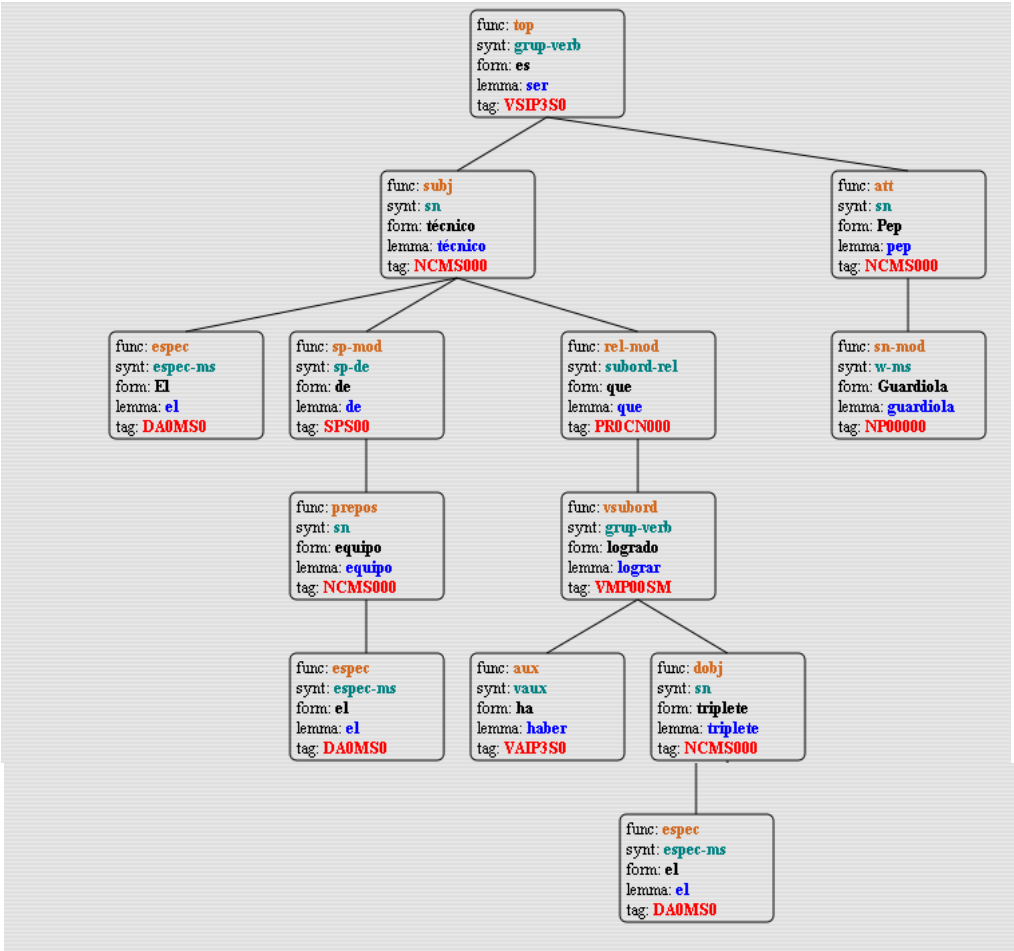


Ilustración 63. Ejemplo de diagrama arbóreo resultado del análisis sintáctico[143].

[142] Analizador desarrollado en el Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla (TALP), Universidad Politécnica de Cataluña (UPC). URL: http://garraf.epsevg.upc.es/freeling/demo.php

[143] Como se observa, el análisis pone de manifiesto uno de los problemas a los que se enfrenta el análisis sintáctico: la ambigüedad en la adjunción de sintagmas

Este tipo de gramáticas son fáciles de manejar, pero no proporcionan soluciones satisfactorias (implican una multiplicación innecesaria de las reglas) para aquellos casos en que es necesario formular restricciones que dependen del contexto, como ocurre con la concordancia, la subcategorización o los constituyentes discontinuos (p. ej. en las estructuras interrogativas, donde se altera el orden respecto a una afirmativa). Por este motivo se han propuesto varios mecanismos para ampliar su poder expresivo, pero sin llegar a la capacidad de las gramáticas de tipo 1, que se considera excesiva para tratar las lenguas naturales.

Las *gramáticas dependientes del contexto o transformacionales* son gramáticas de estructura sintagmática que satisfacen la siguiente restricción: para cada regla de la forma $x \rightarrow y$, la longitud de y (número de símbolos) es mayor o igual a la longitud de x, es decir, en la parte izquierda de la regla puede haber más de un símbolo.

A veces sus reglas tienen una notación diferente: $A \rightarrow y \ / \ x \_ \ z$. Esta notación alternativa hace hincapié en la noción de reescritura de un símbolo dependiendo del contexto: "A" se reescribe como "y" cuando va precedida de "x" y seguida de "z" (contexto).

Por último, las *gramáticas enumerables recursivamente* no presentan restricciones en sus reglas.

---

preposicionales, al hacer depender la oración de relativo (*que ha logrado el triplete*) de *técnico* (el núcleo del sintagma nominal) y no de *equipo*.

De los diferentes tipos de gramáticas formales propuestos por Chomsky, hoy en día existe un acuerdo bastante generalizado de que el poder expresivo de las gramáticas independientes del contexto es suficiente para describir gran parte de las estructuras de las lenguas naturales sin ser necesario tener que recurrir a mecanismos más potentes, como las transformaciones propuestas en los primeros tiempos[144]. Eso sí, es preciso incorporar conceptos como las categorías complejas (con más de un símbolo, p. ej. el género o el número para un nombre), que permitan dar cuenta de fenómenos como la concordancia sin tener que multiplicar las reglas de la gramática. Esto es precisamente lo que hacen las gramáticas de unificación y rasgos (*vid. infra*). Por último, hay que señalar que las gramáticas dependientes del contexto son demasiado expresivas: es decir, contemplan reglas para construir oraciones que no se producen en la realidad.

Por supuesto, existen otros tipos de *gramáticas formales no generativas*, que nos limitamos a enumerar (*vid.* Moreno Sandoval 1998:52 y ss.):

- Gramática de cadenas lingüísticas de Harris, uno de los primeros modelos de gramática implementado en un ordenador, cuyo autor es el maestro de Chomsky.
- Gramática sistémica, desarrollada por el lingüista inglés Halliday y que incorpora aspectos funcionales y pragmáticos.
- Gramáticas de dependencias, que parten de los conceptos de la gramática tradicional.
- Gramáticas categoriales, inspiradas en la Lógica.
- Gramáticas de adjunción de árboles, de orientación claramente computacional, etc.

[144] Pero descartadas desde los años ochenta a favor de otro tipo de mecanismos más elegantes computacionalmente, los rasgos.

En la actualidad, son bastante populares las *gramáticas de unificación y rasgos*, una variante de la gramática generativa que surgió como alternativa al uso de transformaciones y que introduce categorías complejas[145]. Se trata, por lo tanto, de un formalismo muy expresivo: los símbolos pueden ser componentes no atómicos. Por otra parte, su interés radica en que fueron diseñadas específicamente en el marco de la LC para su implementación informática (*cf.* Moreno Sandoval 1998:26-27), en consonancia con los métodos declarativos de los lenguajes de programación lógica, como Prolog, a los que son fáciles de adaptar (*cf.* Vidal y Busquets 1996:409).

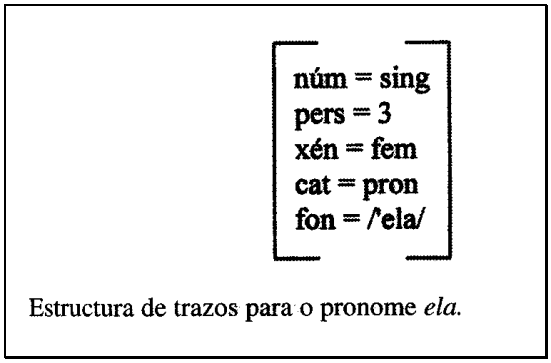
Existen dos conceptos básicos en estas gramáticas:

- *Estructuras de rasgos*: son un sistema para codificar información lingüística que se organiza en pares atributo-valor y que actúan como restricciones. Su punto de partida es la concepción de los objetos lingüísticos como objetos matemáticos que se pueden describir mediante estructuras o conjuntos de rasgos, a modo de ecuaciones. Los rasgos, un concepto conocido en Lingüística, en especial en la tradición generativa[146], "se convierten en los primitivos de la descripción lingüística" (*cf.* Moreno Sandoval 2001) en detrimento de las unidades clásicas de análisis. Ahora, "todas las unidades se interpretan como haces de rasgos, que recogen la información asociada a cada elemento" (*id.*). P. ej. [número = singular] sería una estructura de rasgos, donde *número* sería el atributo y *singular* el valor de ese atributo en un caso

[145] Véase MORENO SANDOVAL (2001) para un monográfico sobre este tipo de gramáticas.

[146] Donde se remontan a Fonología, de la mano de Jakobson, y también destaca su importancia en el Programa Minimalista.

concreto. Los rasgos, además, pueden ser simples o complejos, si el valor de un atributo es a su vez un rasgo. Ejemplos de estructuras de rasgos (una simple y otra compleja), tomados de X. Gómez Guinovart (2000a:227):



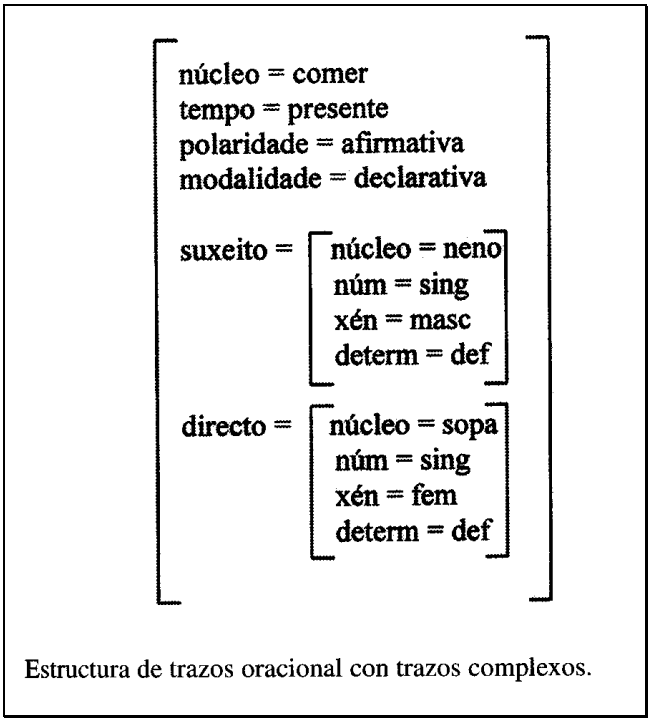


Ilustración 64. Ejemplos de estructuras de rasgos para el gallego.

➲ *Unificación*: es una operación matemática que permite combinar la información codificada en forma de estructuras de rasgos (resolver las ecuaciones), y da como resultado otra estructura de rasgos unificada, siempre y cuando la información de las estructuras parciales sea compatible. Se trata de un mecanismo simple y al mismo tiempo eficiente, tomado de la Informática, en concreto, de la Inteligencia Artificial por M. Kay (1979), quien es responsable de la primera gramática computacional de este tipo[147]: la *gramática de unificación funcional* (FUG, del inglés *Functional Unification Grammar*). Además, cuenta con la ventaja añadida de que se puede implementar en un programa de ordenador con relativa facilidad, lo que "permite la simulación, prueba y evaluación de cualquier propuesta teórica" (*cf.* Moreno Sandoval 2001).

El procedimiento seguido por la unificación se puede esquematizar de la siguiente manera:

- los pares atributo-valor aparecen en la estructura resultante;
- los valores que solo aparecen en una de las estructuras se incluyen en la estructura unificada;
- la unificación no se produce cuando las estructuras son distintas.

[147] La noción de *unificación* rápidamente se extendió del ámbito de la LC a teorías lingüísticas: la *gramática léxico-funcional* (LFG, *Lexical Functional Grammar*), propuesta por Bresnan y Kaplan; la *gramática de estructura sintagmática generalizada* (GPSG, *Generalized Phrase Structure Grammar*), debida a Gazdar *et al.*; o la *gramática sintagmática nuclear* (HPSG, *Head-driven Phrase Structure Grammar*) de Pollard y Sag, que a su vez han servido de inspiracion a gramáticas computacionales, en una relación circular entre la LC, la Lingüística Teórica y de nuevo la LC.

Así, por ejemplo, a partir de las dos estructuras como (*cf.* Vidal y Busquets 1996:412-413):

(1)

{ categoría: SN
concordancia: [persona: plural] }

(2)

{ categoría: SN
concordancia: [persona: tercera] }

si se lleva a cabo la unificación se obtiene:

{ categoría: SN
concordancia: { persona: tercera
número: plural } }

Véase otro ejemplo de unificación, tomado de X. Gómez Guinovart (2000a:228):

ET1 [suxeito = [número = plural]]

ET2 [suxeito = [persoa = 3]
tempo = presente]

ET3 [suxeito = [persoa = 2]
tempo = pasado]

ET4 [suxeito = [persoa = 3
número = plural]
tempo = presente]

Exemplo de unificación de estructuras de trazos.

Ilustración 65. Ejemplo de unificación de estructuras de rasgos para el gallego.

Al codificar la información lingüística[148] en forma de estructuras de rasgos, se descarga el peso del procesamiento del componente de reglas y se traslada al léxico, que es el que incorpora la mayor parte de la información necesaria para el análisis lingüístico. Además, se facilita la integración de la semántica en la sintaxis, de ahí la repercusión que está teniendo este tipo de gramáticas en Lingüística Teórica.

Las gramáticas de unificación y rasgos conforman una amplia familia (*cf.* Gómez Guinovart 2000a) que abarca desde teorías sobre el lenguaje hasta herramientas o formalismos para la descripción lingüística. Las primeras comparten el hecho de partir de la teoría de la rección y el ligamiento de Chomsky, pero la mayoría no contempla dos niveles sino solo uno, el superficial, y rechaza las transformaciones. Destaca el alto grado de formalización que implican y el papel tan importante que se otorga al léxico en el procesamiento lingüístico.

---

[148] Este tipo de formalismo se emplea también en la descripción de otros niveles lingüísticos, como el morfológico. Por ejemplo, el analizador GRAMPAL, aplicado a la morfología del español (*cf.* Moreno Sandoval 1998:103 y ss.), se fundamenta en una gramática con reglas morfológicas (dos para conjugar todas las formas verbales sintéticas y cuatro para la flexión nominal) y un lexicón. Las entradas del lexicón son alomorfos que contienen dos tipos de información codificada mediante rasgos: a) información lingüística gramatical (número, persona, etc.) y léxica (lema y subcategorización), que será utilizada por las reglas sintácticas después; b) información contextual o morfotáctica (tipo de plural, tipo de género, tipo de raíz, tipo de desinencia, conjugación). De acuerdo con este modelo, la forma *pez* contaría con dos entradas (Moreno Sandoval 1998:104-105):

Pez		*pec*	
morfo-cat	= raíz-n	morfo-cat	= raíz-n
sint-cat	= n	sint-cat	= n
lex	= pez	lex	= pez
conc gen	= masc	conc gen	= masc
tipo-plu	= no	tipo-plu	= plu2
tipo-gen	= inherente	tipo-gen	= inherente

Y los morfemas de género y número estarían representados en cinco entradas, también en el lexicón:

Alomorfos de número		*Alomorfos de género*		
s	*es*	*o*	*e*	*a*
morfo-cat = suf-n	morfo-cat = suf-n	morfo-cat = suf-n	morfo-cat = suf-n	morfo-cat = suf-n
conc num = plu	conc num = plu	conc gen = masc	conc gen = masc	conc gen = fem
tipo-plu = plu1	tipo-plu = plu2	conc num = sing	conc num = sing	conc num = sing
		tipo-gen = mas1	tipo-gen = mas2	tipo-gen = fem

Según Shieber (1989 [1986]), quien mejor ha descrito de una manera sistematizada este tipo de formalismos, las ventajas que aportan son tres (Vidal y Busquets 1996:413):

1) proporcionan una herramienta precisa para la descripción de las lenguas naturales (adecuación lingüística);
2) delimitan la clase de las posibles lenguas naturales (expresividad);
3) caracterizan las lenguas naturales de una manera interpretable computacionalmente (efectividad computacional).

Sin embargo, independientemente de la teoría adoptada para dar cuenta de la sintaxis, A. Moreno Sandoval (1998:109 y ss.) considera que existe una serie de obstáculos para la sintaxis computacional:

a) Las *dependencias no acotadas* o *a larga distancia*, como oraciones interrogativas, relativas, etc., uno de los problemas fundamentales de las gramáticas independientes del contexto y que en las gramáticas de unificación se suele tratar con un rasgo llamado *slash*, huella que deja el constituyente que se ha desplazado de su posición normal (una forma de evitar las transformaciones, que dan cuenta del fenómeno pero que no se suelen utilizar en LC).
b) La *coordinación*, que en las gramáticas independientes del contexto supone la duplicación de las reglas, problema que se soluciona en las gramáticas de unificación, donde con una única regla se aborda el fenómeno. Sin embargo, se siguen presentando dificultades: asignación de la estructura interna del elemento coordinado, elipsis de elementos y ambigüedad sintáctica en algunos casos.
c) El *orden de constituyentes*, sobre todo en lenguas de orden más o menos libre. Dado que un tratamiento puramente superficial no funciona, se han efectuado propuestas atractivas teóricamente pero

que chocan con la eficiencia computacional, al tener que construir un segundo nivel de representación sintáctica (estructura profunda). Por lo general, se parte del análisis superficial y de las posibles combinaciones presentes en él y después: i) o bien se asigna función sintáctica a los constituyentes (por ejemplo, se asocian las preposiciones con las funciones o se utilizan los casos en las lenguas que disponen de ellos); ii) o se emplean restricciones semánticas (restricciones seleccionales, *cf.* Chomsky 1965), lo que implica disponer de un modelo semántico para cada verbo que relacione funciones sintácticas y funciones semánticas, algo solo posible en dominios muy restringidos; iii) por último, se puede recurrir a algún tipo de heurística o probabilidades.

d) Los *elementos nulos* o *vacíos*. Tema ya de por sí controvertido en Lingüística Teórica, donde se discuten los argumentos a favor y en contra, y que desde la perspectiva computacional no parece contar con demasiados adeptos debido a que el uso de elementos vacíos redunda en pérdidas de eficiencia.

2.3.2. Analizadores

Por lo que respecta al *parser* o programa informático que determina la forma de aplicar las reglas de la gramática, se han propuesto varias estrategias, con independencia del tipo de gramática que se utilice, aunque los principales analizadores se han diseñado para gramáticas independientes del contexto. Dada una gramática, el *parser*, de acuerdo con la estrategia que siga para encontrar la solución (el análisis correcto), especificará cómo se debe llevar a cabo el análisis sintáctico.

Los *parsers* se empezaron a desarrollar a finales de los sesenta con conocimientos procedentes de la Inteligencia Artificial (p. ej. las redes de transición aumentadas ya comentadas) y de los lenguajes formales, los lenguajes de programación[149].

Las *técnicas de análisis*, también llamadas *estrategias de búsqueda* (*cf.* Vidal y Busquets 1996:413) o *algoritmos de parsing* (*cf.* Verdejo 1995:55), para llevar a cabo el análisis sintáctico se clasifican en función de varios criterios: estrategia, dirección y tratamiento del no-determinismo (*cf.* Verdejo *id.*):

(1) Según la estrategia o sentido en que exploren el espacio de búsqueda, se habla de técnicas de análisis:

- *Descendente* o *top-down parsing* (dirigidas por los objetivos): para producir la representación estructural de una oración, esta estrategia compara las reglas de la gramática con la oración que se le presenta para el análisis empezando por la regla más abstracta, la que contiene el símbolo inicial (gramática), y va descendiendo hasta llegar a los elementos terminales (léxico), es decir, va de la gramática a los datos o, si se prefiere, actúa por objetivos. A partir de la primera palabra de la oración establece una hipótesis sobre una posible estructura (correspondencia entre la parte izquierda de las reglas y las categorías presentes en el *input*) y se mantiene en ella, a menos que se pruebe que es incorrecta. Es la estrategia habitual en las representaciones arbóreas de la gramática generativa.

[149] *Vid.* RODRÍGUEZ HONTORIA (2002) para una descripción detallada de algunos de los principales analizadores.

Por ejemplo[150] (Vidal y Busquets 1996:414):

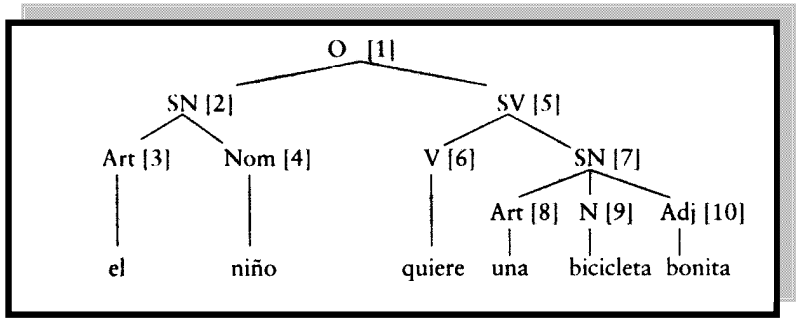


Ilustración 66. Ejemplo del funcionamiento de un "parser" descendente.

- *Ascendente* o *bottom-up parsing* (guiadas por los datos): el proceso es el inverso, se parte de las palabras y se van combinando las reglas hasta alcanzar la unidad de nivel superior, es decir, para producir una representación estructural, esta estrategia compara la oración de entrada con las reglas de la gramática empezando por la última regla, la que contiene los elementos terminales (léxico), y va ascendiendo hasta llegar a la regla más abstracta, por lo tanto va de los datos a la gramática. Para aplicar las reglas busca correspondencias entre las categorías presentes en el *input* y la parte derecha de las reglas. Se trata de una técnica desarrollada en el terreno computacional. Fue, además, el primer tipo de analizador utilizado[151] y, para algunos autores, su forma de actuar reflejaría mejor el procesamiento de las oraciones que efectuamos las personas, aunque esta es una cuestión no exenta de controversias. También se atribuye mayor eficiencia a este tipo de analizadores, ya que al tener conocimientos previos sobre el *input* resuelven el análisis en menos pasos.

[150] Los números entre corchetes indican el orden en que se produce el análisis.
[151] V. Yngve ya sugirió esta estrategia en 1955 (*cf.* JURAFSKY y MARTIN 2000:361).

Por ejemplo (Vidal y Busquets 1996:414):

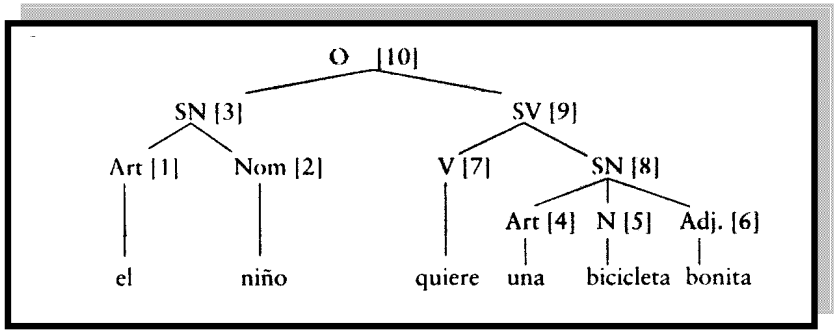


Ilustración 67. Ejemplo del funcionamiento de un "parser" ascendente.

En relación con la estrategia seguida por estos analizadores, J. Vidal y J. Busquets (1996:414) distinguen entre:

- *Parsers deterministas*: el analizador toma un solo camino; es decir, el *parser* actúa condicionado por los datos, que son los que le llevan a proponer una representación estructural y no otras. Es lo que sucede en los analizadores ascendentes, que parten de una secuencia concreta de palabras: están obligados a seguir un único camino, el que le marcan los datos.
- *Parsers no deterministas*: el *parser* no está condicionado por los datos, por lo que contempla varias posibilidades antes de optar por la correcta. Es lo que sucede en los analizadores descendentes, que parten de la gramática: contemplan varios caminos alternativos, bien sea de forma simultánea, en paralelo, o de forma secuencial, primero uno y, si este falla, luego otro. En este último caso, el *parser* tiene que retroceder hasta el último punto analizado con éxito y explorar las otras posibilidades. Este proceso de vuelta atrás se conoce como retrotrazado o *backtracking*.

Las ventajas e inconvenientes de estas técnicas se pueden resumir en la siguiente tabla:

***Parsers* descendentes**	***Parsers* ascendentes**
El análisis descendente es más simple y favorece la eliminación de ambigüedades.	Los constituyentes se construyen de forma definitiva, sin vuelta atrás.
No pierde tiempo analizando estructuras que no conducen al símbolo inicial, ya que este es su punto de partida.	Contempla estructuras que posteriormente tienen que ser descartadas por no corresponder con oraciones de la lengua en cuestión.
Consume recursos analizando reglas que pueden no tener correspondencia en el *input*, ya que este solo se examina al final del proceso.	Parte de los datos.
Presenta el problema de tener que volver atrás constantemente.	Muchas de las agrupaciones son incorrectas.

Tabla 13. "Parsers" descendentes vs. ascendentes.

Debido a los inconvenientes que presenta cada *parser* por separado, se han desarrollado estrategias mixtas, que combinan las ventajas de ambos.

(2) Según la dirección (Verdejo 1995:55) u orden en que se va a analizar la oración, el análisis puede efectuarse de izquierda a derecha, la opción más frecuente, pero también de derecha a izquierda o incluso por expansión de "islas": a partir de determinadas palabras, el análisis se extiende en ambas direcciones (*cf.* Rodríguez Hontoria 2002), o desde el centro, tomando como referencia un elemento nuclear a partir del cual se alterna el análisis hacia la izquierda y hacia la derecha.

(3) Por último, en función del no determinismo (Verdejo 1995:55; Vidal y Busquets 1996:414), es decir, la forma de proceder cuando hay varias alternativas, se distingue entre:

- *Búsqueda secuencial* o *en profundidad (depth-first*): el analizador sigue una alternativa y persiste en esta decisión hasta agotarla, aunque mantiene la pista respecto al grupo de opciones restantes con el fin de poder retroceder (*backtracking*) a fases anteriores si el camino tomado falla y elegir otra opción, es decir, examina de forma sucesiva cada una de las alternativas hasta que encuentra una que tiene éxito.
- *Búsqueda de alternativas en paralelo* (*breadth-first*): cuando hay que tomar una decisión, antes de optar por una alternativa, el analizador explora, de forma simultánea, todas y cada una de las posibilidades, primero las de un nivel, después las del nivel superior, etc. Esta forma de proceder suele consumir bastantes recursos de memoria en los procesadores, dada la ambigüedad presente en las lenguas naturales (en especial, la derivada de la adjunción de sintagmas preposicionales, de la coordinación o del agrupamiento de frases nominales).

El denominado *backtracking*, vuelta atrás o retrotrazado puede llegar a emplear mucho tiempo, al recorrer los mismos caminos una y otra vez. La solución pasa por almacenar las estructuras reconocidas (*charts* o tablas auxiliares que registran las secuencias de palabras ya validadas). Se trata de una estructura geométrica abstracta (red) que consta de unos vértices (estados de análisis) conectados a través de arcos (constituyentes), como en el siguiente ejemplo (Vidal y Busquets 1996:415):

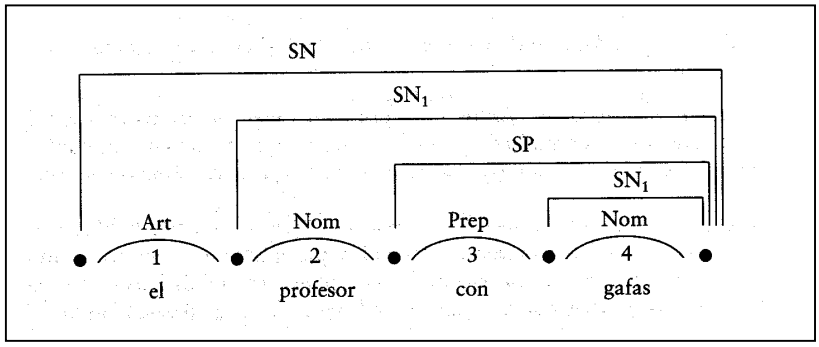


Ilustración 68. Ejemplo de "chart".

Para observar las diferencias derivadas de seguir una estrategia u otra, compárese, a modo de ejemplo, el proceso esquematizado (adaptado de Blackburn y Striegnitz 2002) que implica el análisis de las oraciones del inglés "Mia loved Vincent" y "Vincent shot Marsellus" con un *parser* descendente y con otro ascendente. La gramática de referencia es:

s ---> [np,vp].	s=sentence
np ---> [pn].	np=nominal phrase
vp ---> [iv].	vp=verbal phrase
vp ---> [tv,np].	pn=proper name
pn ---> Vincent | Mia | Marsellus	iv=intransitive verb
iv ---> loved	tv=transitive verb
tv ---> shot	

Tabla 14. Gramática de referencia para los "parsers".

a) *Parser descendente secuencial*. El *parser* parte de la categoría superior, es decir, de la primera regla de la gramática, y busca correspondencias en la parte izquierda de las reglas hasta llegar al vocabulario terminal.

Paso	Categorías buscadas	Cadenas que se corresponden	Comentarios
1.	s	Mia loved Vincent	s --- > np vp
2.	np vp	Mia loved Vincent	np --- > pn
3.	pn vp	Mia loved Vincent	Lex (mia, pn) Correspondencia
4.	vp	Loved Vincent	vp --- > iv Solo se considera de momento esta regla, ya que la búsqueda es secuencial
5.	iv	Loved Vincent	Regla no aplicable. Vuelta al paso 4.
4'.	vp	Loved Vincent	vp --- > tv
5'.	tv np	Loved Vincent	Lex (loved, tv) Correspondencia
6.	np	Vincent	np --- > pn
7.	pn	Vincent	Lex (Vincent, pn) Correspondencia

Tabla 15. Análisis con un "parser" descendente secuencial.

De haber actuado en paralelo, en el paso 4 el *parser* habría analizado simultáneamente todas las reglas para los sintagmas verbales, en vez de primero la regla para los sintagmas verbales formados por verbos intransitivos y, al fallar esta, la de los verbos transitivos; y en el paso 5 habría descartado la regla para los verbos intransitivos y aceptado la de los verbos transitivos, que es la aplicable en este caso concreto.

b) *Parser ascendente*. El *parser* parte de los elementos terminales, del *input*, y trata de llegar a la primera regla de la gramática, por lo que busca correspondencias en la parte derecha de las reglas, ascendiendo de nivel en nivel en la jerarquía, hasta llegar al símbolo inicial.

Paso	*Input*	Regla	Cadena reconocida
1.	Vincent	pn --- > Vincent	pn shot Marsellus
2.	Vincent	np --- > pn	np shot Marsellus
3.	Vincent shot	tv --- > shot	np tv Marsellus
4.	Vincent shot Marsellus	pn --- > Marsellus	np tv pn
5.	Vincent shot Marsellus	np --- > pn	np tv np
6	Vincent shot Marsellus	vp --- > tv np	np vp
7.	Vincent shot Marsellus	s --- > np vp	s

Tabla 16. Análisis con un "parser" ascendente.

De todas formas, lo más habitual es combinar las diferentes estrategias para aprovechar las ventajas que ofrece cada una. P. ej. es habitual utilizar analizadores descendentes en serie con retrotrazado; *parsers* ascendentes en paralelo; *parsers* mixtos ascendentes y descendentes, etc.

Históricamente, los tipos de *parsers* que se han ido utilizando son los siguientes (*cf.* Rodríguez Hontoria 2002:99 y ss.):

- Analizadores basados en la técnica del *pattern matching*, que se limitaban a detectar palabras clave, presentes en la oración de entrada, que estaban asociadas con patrones previamente establecidos, como se puede observar en el sistema Eliza. Por lo tanto, el conocimiento lingüístico implicado era mínimo y en este apartado ni siquiera los hemos comentado.
- Redes de transición. Estos son los primeros analizadores que consideran conocimientos propiamente lingüísticos. Los más representativos son las redes de transición aumentadas, propuestas por Woods a principios de los setenta.
- Analizadores basados en *charts* o tablas auxiliares, como los que aplicó Earley en 1970. Mejoran el tratamiento del retrotrazado en las redes de transición aumentadas al añadir grafos que disponen de memoria temporal donde van almacenando las estructuras ya analizadas –parcial o totalmente– y validadas, por lo que en caso de que el analizador tenga que volver atrás no pierda los análisis válidos, es decir, no tenga que volver a analizarlo todo desde el principio.
- Analizadores basados en la unificación, empleados con éxito desde los ochenta, en relación con el auge de las gramáticas de unificación y que pueden ser implementados directamente en el lenguaje de programación Prolog.

- Analizadores probabilísticos, que asignan probabilidades a las reglas de la gramática (las probabilidades se extraen de pequeños corpus previamente analizados), especialmente adecuados para tratar grandes cantidades de textos sin restricciones. Surgen en los noventa como consecuencia del fracaso de los modelos simbólicos (*vid.* Lavid 2005:121) y en un afán por trasladar los éxitos del paradigma estocástico en el tratamiento del habla al tratamiento del texto.
- Analizadores superficiales, extensiones de etiquetadores morfosintácticos a los que se añaden etiquetas para funciones sintácticas.
- Analizadores parciales, que proporcionan información detallada sobre la sintaxis de un fragmento: un sintagma nominal o preposicional, etc.

El *análisis en profundidad de todo un texto* ("full parsing") proporciona una información detallada de las relaciones entre las palabras, aunque rechaza cualquier oración que no pueda analizarse en su totalidad y, desde el punto de vista computacional, no resulta demasiado robusto ni fiable a la hora de enfrentar el *parser* con textos reales. Por otra parte, como señala H. Rodríguez Hontoria (2000), hay que considerar que no siempre es necesario llevar a cabo un análisis tan completo, tarea que, además, se enfrenta con problemas de diversa índole a la hora de tratar textos no restringidos, de segmentar el texto en unidades de análisis, de elegir entre múltiples análisis posibles o de ampliar la cobertura para manejar oraciones no gramaticales o en las que hay palabras que no forman parte del lexicón.

Un ejemplo de análisis sintáctico en profundidad, en el que los constituyentes se marcan con corchetes etiquetados y la información gramatical se adjunta a las palabras, es el siguiente, tomado del *Lancaster-Leeds treebank*[152]:

[S[Ncs another_DT new_JJ style_NN feature_NN Ncs] [Vzb is_BEZ Vzb] [Ns the_AT1 [NN/JJ& wine-glass_NN [JJ+ or_CC flared_JJ HH+]NN/JJ&] heel_NN ,_, [Fr[Nq which_WDT Nq] [Vzp was_BEDZ shown_VBN Vzp] [Tn[Vn teamed_VBN Vn] [R up_RP R] [P with_INW [NP[JJ/JJ/NN& pointed_JJ ,_, [JJ- squared_JJ JJ-] ,_, [NN+ and_CC chisel_NN NN+]JJ/JJ/NN&] toes_NNS Np]P]Tn]Fr]Ns] ._. S]

Ilustración 69. Texto analizado en profundidad[153].

O estos otros, obtenidos con el uso de *Machinese Syntax*[154] al introducir el enunciado: "Sarkozy takes the crown":

[152] *Treebank* es la denominación que se da a los corpus en los que los textos han sido sometidos a un proceso exhaustivo de análisis sintáctico.

[153] Algunos ejemplos de claves de las etiquetas:

- S = Sentence
- Ncs = noun phrase, count noun singular
- DT = singular determiner
- JJ = adjective phrase
- NN = singular common noun
- Vzb = verb phrase, third person singular *to be*
- BEZ = *is*
- Ns = noun phrase singular
- ATI = article
- & = whole coordination
- CC = co-ordinating conjunction

...

Estas claves están basadas en el sistema de etiquetación diseñado para el corpus *LOB*, el programa *CLAWS*, ya que el *Lancaster-Leeds treebank* es un subconjunto analizado sintácticamente de dicho corpus (*vid.* más adelante, en el apartado de lingüística de corpus).

[154] Desarrollado por la empresa finlandesa Connexor, proporciona un análisis funcional en profundidad con una orientación semántica, aplicable a ámbitos como el

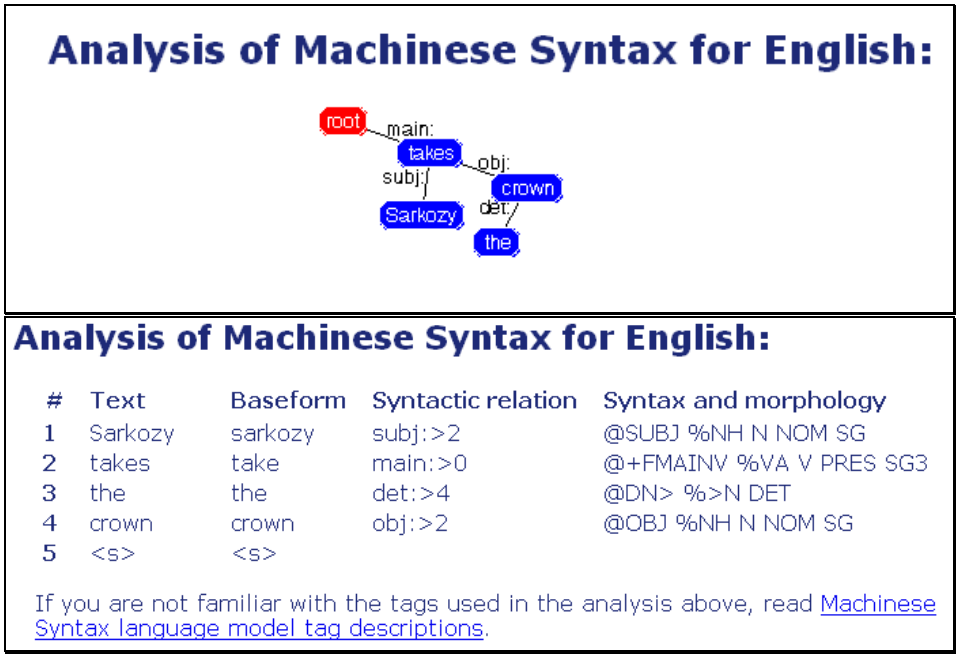


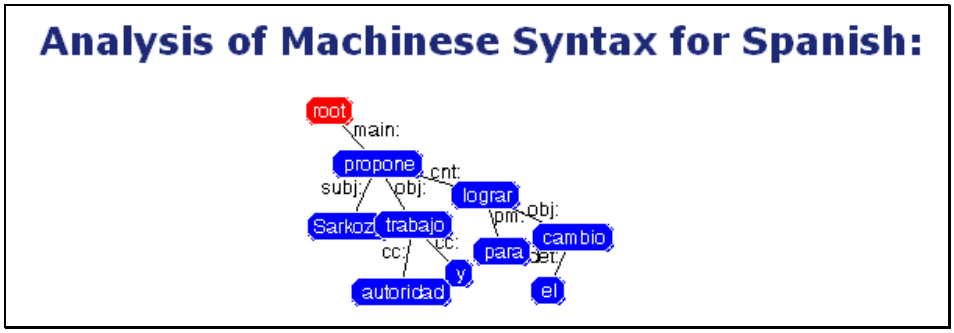
Ilustración 70. Análisis sintáctico en profundidad del inglés.

Como se puede observar, el programa lematiza las palabras (relaciona "takes" con "take"), determina las relaciones sintácticas y la estructura interna de los constituyentes ("Sarkozy" es el sujeto del elemento 2, "takes"; "crown" es el objeto del elemento 2; "the" es el determinante de 4, "crown"), al tiempo que ofrece la información morfosintáctica pertinente: sujeto, núcleo nominal, nombre, singular, etc.

Estos son los resultados obtenidos con el mismo analizador, aplicado al español, del texto "Sarkozy propone trabajo y autoridad para lograr el cambio":

---

análisis de tendencias, los servicios de inteligencia, la traducción o el resumen automático de documentos. Ofrece dos posibilidades: mostrar la estructura en forma de árbol o en forma de texto. URL: http: / / www.connexor.eu / technology / machinese / machinesesyntax /

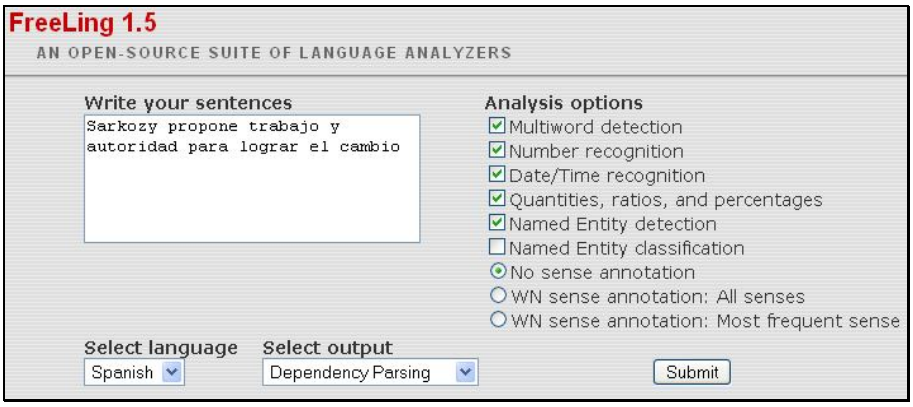


Analysis of Machinese Syntax for Spanish:

#	Text	Baseform	Syntactic relation	Syntax and morphology
1	Sarkozy	sarkozy	subj:>2	&NH <?> <Proper> N SG
2	propone	proponer	main:>0	&+FM V IND PRES SG3
3	trabajo	trabajo	obj:>2	&NH N MSC SG
4	y	y	cc:>3	&CC CC
5	autoridad	autoridad	cc:>3	&NH N FEM SG
6	para	para	pm:>7	&PM> PREP
7	lograr	lograr	cnt:>2	&-FM V INF
8	el	el	det:>9	&DN> DET MSC SG
9	cambio	cambio	obj:>7	&NH N MSC SG
10	<s>	<s>		

Ilustración 71. Análisis sintáctico en profundidad del español.

Otro analizador, ya comentado con anterioridad, es FreeLing (TALP, Universidad Politécnica de Cataluña), desarrollado para varias lenguas (español, catalán, gallego, italiano e inglés), con multitud de características[155] y que, ante el mismo texto, produce el siguiente análisis, basado en una gramática de dependencias:



[155] *Vid.* página web del programa. URL: http://garraf.epsevg.upc.es/freeling/

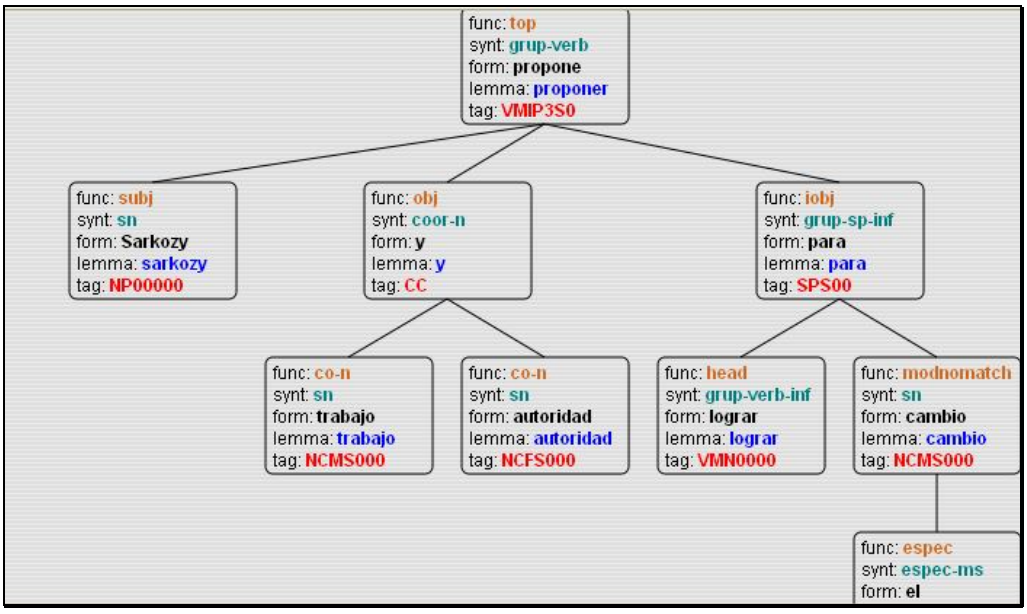


Ilustración 72. Análisis con una gramática de dependencias.

Para evitar las dificultades que pueden surgir a la hora de realizar un análisis en profundidad, en muchas ocasiones los sistemas computacionales se limitan a un *análisis sintáctico parcial* o *fragmental* ("shallow parsing" o "skeleton parsing"), que consiste únicamente en reconocer en el texto de entrada determinadas agrupaciones sintácticas ("chunking") tales como frases nominales, preposicionales o verbales ("chunks"), pero suele ignorar su estructura interna o, como mucho, identifica mediante corchetes los núcleos y sus adyacentes, sin especificar la función sintáctica que desempeñan. Obsérvese el ejemplo siguiente (tomado de Carroll 2003:234):

> *Give me flight times after ten in the morning.*
>
> [$_{V}$ Give ^{H}] [$_{N}$ me ^{H}] [$_{N}$ flight times ^{H}] [$_{P}$ after ^{H}] [$_{N}$ ten ^{H}] [$_{P}$ in ^{H}] [$_{N}$ the morning ^{H}]

Ilustración 73. Análisis sintáctico parcial.

Básicamente el *parser* deduce información sintáctica –agrupaciones (*give*, *me*, *flight times*…)– de los resultados proporcionados por el módulo morfológico y por la desambiguación. A diferencia del tipo de análisis anterior, no establece relaciones sintácticas entre elementos y la información que ofrece es mucho menos completa, pero gana en rapidez, fiabilidad, robustez y reducción del coste computacional. Además, permite analizar grandes cantidades de texto al no tener que efectuar un análisis detallado.

Otro ejemplo de análisis sintáctico superficial es este, de la oración "Sarkozy takes the crown", realizado con "Memory-based shallow parser memo"[156]:

```
[NP Sarkozy/ /NNP NP] [VP takes/VBZ VP] [NP the/DT
crown/NN NP]
```

Ilustración 74. Análisis sintáctico superficial.

En este caso, el *parser* identifica los constituyentes o "chunks" (entre corchetes): frases nominales [NP] y frases verbales [VP], y algunas de sus características principales: nombre propio [NNP], verbo en tercera persona del singular [VBZ], determinante [DT] y nombre singular [NN]. También cuenta con una opción para identificar sujetos y objetos.

Este otro ejemplo es el siguiente, tomado del Servicio de Tecnología Lingüística de la Universidad de Barcelona[157], que muestra el análisis de la oración: "El lenguaje ha sido un instrumento decisivo":

[156] Analizador desarrollado para el inglés. Más información en la URL: http://ilk.uvt.nl/mbsp_demo/mbsp_demo.html

[157] URL: http://www.ub.edu/stel/stel.htm

```
S_[ sn_[
        espec-ms_[ El_da0ms0 ]
        grup-nom-ms_[ lenguaje_ncms000 ] ]
    grup-verb_[ ha_vaip3s0 sido_vsp00sm ]
    sn_[
        espec-ms_[ un_di0ms0 ]
        grup-nom-ms_[ instrumento_ncms000
                      s-a-ms_[ decisivo_aq0ms0 ] ] ]
```

Ilustración 75. Otro ejemplo de análisis sintáctico superficial.

El analizador FreeLing, además de análisis más detallados, como ya hemos comentado, ofrece también la posibilidad de realizar análisis superficiales, como este de la oración "Sarkozy takes the crown", en el que identifica tres bloques o *chunks*: un nombre (*Sarkozy*), un verbo (*takes*) y un sintagma nominal (*the crown*), para los que proporciona la información sobre la forma, el lema y la categoría gramatical.

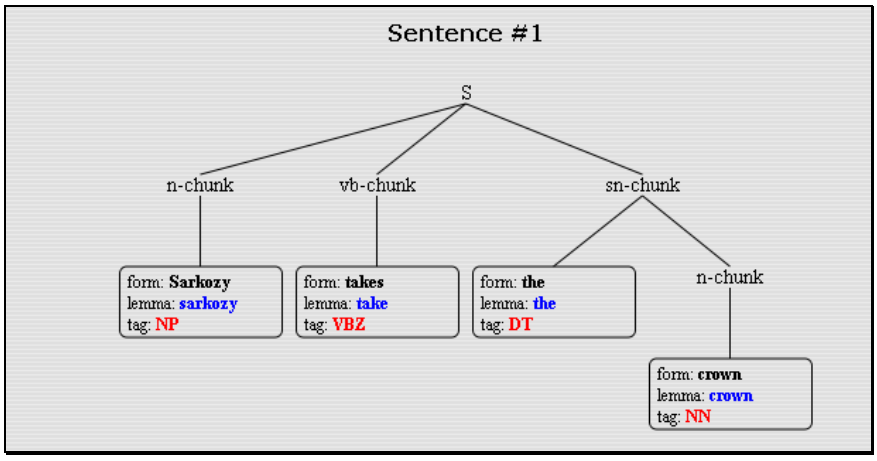


Ilustración 76. Análisis superficial con Freeling.

Contrástese con el resultado del análisis de dependencias, que añade información sobre la función sintáctica de cada bloque:

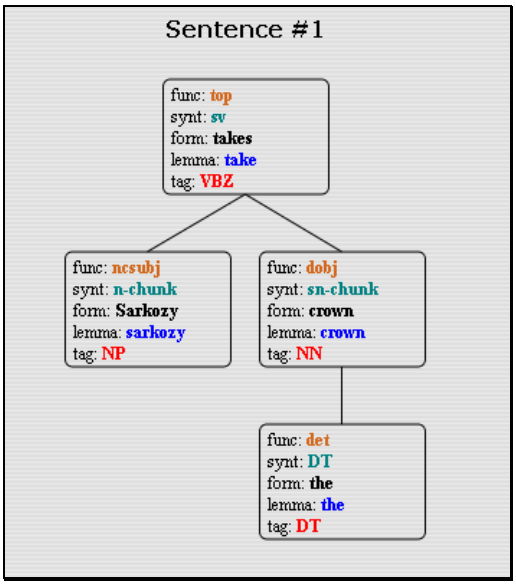


Ilustración 77. Análisis de dependencias con Freeling.

Por último, veamos el análisis superficial que ofrece el analizador del *Centre de LIenguatge i Computació* de la oración *Sarkozy propone trabajo y autoridad para lograr el cambio*:

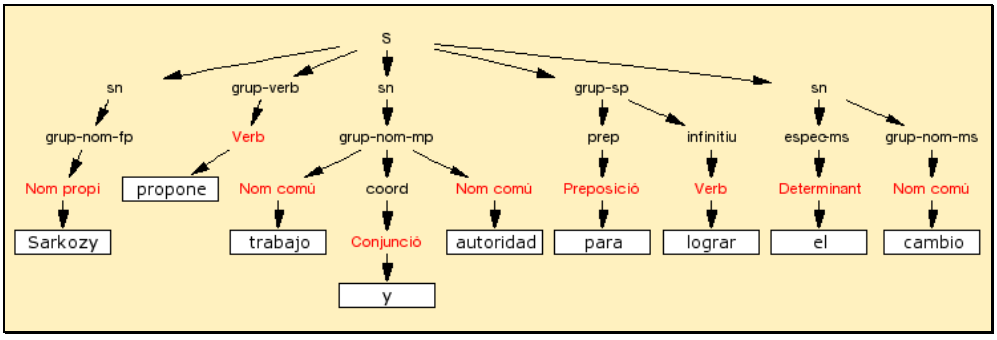


Ilustración 78. Análisis superficial con el analizador de CLiC.

En definitiva, existen diferentes descripciones posibles y analizadores para llevar a cabo el análisis sintáctico de las lenguas naturales. El tipo concreto de análisis, completo o parcial, o la estrategia adoptada, dependerán de la teoría lingüística y del formalismo gramatical en el que se fundamenten, y también de la finalidad de la investigación o del sistema práctico concreto.

2.4. Semántica computacional

Junto a la sintaxis, la Semántica es otro de los módulos centrales del tratamiento computacional del lenguaje. De hecho, según T. Badia (2003:231-232), "es el ideal al cual aspira todo el que se dedica al procesamiento del lenguaje natural", ya que, si queremos una "máquina parlante", el dominio del significado es fundamental para lograr una interpretación del contenido de un texto[158]. Sin embargo, es el área de trabajo de la LC que presenta más dificultades: primero, porque en Lingüística no existe una propuesta general, como ocurría con la sintaxis, que sea apta para su tratamiento computacional; en segundo lugar, porque la suma del significado de las partes no garantiza en ocasiones una interpretación adecuada.

La Semántica estudia el significado sin tener en cuenta el contexto y otros factores que influyen en el sentido último de los enunciados. Por esta razón es preciso considerar también la Pragmática, disciplina que estudia las implicaciones derivadas del uso efectivo del lenguaje en un contexto y en una situación determinados, pues con frecuencia son aspectos de esta índole los que configuran el sentido. Por último, conocimientos no lingüísticos de carácter general también desempeñan un papel relevante en la interpretación final del contenido de una frase o de un texto[159].

---

[158] En aplicaciones como la traducción automática resulta fundamental un buen tratamiento semántico.

[159] Puesto que cuestiones como la identificación de expresiones referenciales y la subsiguiente adscripción de referentes, la recuperación del contexto de elementos elididos o la resolución de la anáfora inciden en la interpretación, los límites entre Semántica y Pragmática en ocasiones se difuminan; otras veces, se conciben como módulos separados, que actúan secuencialmente, primero el semántico y luego el pragmático (*cf.* MARTÍ Y CASTELLÓN 2000:130).

Por otra parte, hay que tener en cuenta que el tratamiento del significado puede abordarse junto con la sintaxis (análisis sintáctico y semántico paralelos), pero también de forma separada, bien como un paso posterior a aquella (análisis semántico guiado por la sintaxis), o directamente sobre los textos (sistemas basados en la semántica), sin requerir un análisis sintáctico previo. En cualquiera de los casos, es el nivel del lenguaje que más esfuerzos está concentrando en la actualidad en LC. Según T. Badia (2003:232-235), las opciones son:

- Semántica en un nivel posterior a la sintaxis.
- Anotaciones semánticas a las representaciones sintácticas.
- Semántica incorporada en la representación sintáctica.
- Anotación semántica (independiente de cualquier teoría sintáctica).

Un primer acercamiento a la semántica postula la necesidad de contar con una representación sintáctica previa como requisito para efectuar la interpretación semántica, es decir, sintaxis y semántica se conciben como dos módulos que actúan de forma independiente y secuencial, primero uno y luego el otro, sin que exista entre ellos relación alguna. Son necesarios, por tanto, dos formalismos diferentes: uno para la sintaxis y otro para la semántica.

Desde otros planteamientos, se parte de un tratamiento de base sintáctico que se enriquece con alguna información semántica mínima. El peso general del procesamiento del lenguaje recae sobre el módulo sintáctico, que interactúa con el semántico para resolver ambigüedades o para ayudar a elegir un análisis sintáctico entre varios posibles. Desde esta postura, primero se efectúa un análisis sintáctico preliminar y, a partir de él, se introducen anotaciones semánticas –sobre todo en los

verbos (argumentos) y los nombres (rasgos semánticos o semas)- que sirven, a su vez, de guía para completar el procesamiento sintáctico. Los problemas surgen a la hora de relacionar ambos niveles.

Un acercamiento próximo es el que, desde los modelos en boga en la actualidad, dispone de un único formalismo lingüístico en el que se integra toda la información necesaria para llevar a cabo las diferentes operaciones implicadas: análisis morfológico, sintáctico y semántico. Es decir, existe una única representación para sintaxis y semántica, en lo que concierne a los aspectos principales del procesamiento lingüístico.

En los sistemas basados en la semántica, a partir del input, se genera una representación semántica y puede que no haya representación sintáctica o, de haberla, se limita a las categorías gramaticales básicas, con frecuencia anotadas con etiquetadores de base estadística. Estos modelos evitan los problemas que pueden plantear una representación sintáctica o la relación entre la sintaxis y la semántica. Estos sistemas suelen presentar un tratamiento semántico ad hoc, limitado al dominio o campo semántico al que se refiere el sistema concreto. Por lo tanto, a diferencia de la sintaxis, no existe un tratamiento semántico general, sino solo parcial.

Con independencia de la postura adoptada, si la Sintaxis tiene como cometido dar cuenta de las relaciones formales entre las palabras, la Semántica, en términos generales, se ocupa de:

- El *significado de las palabras*: los contenidos asociados a cada palabra individual, cómo se obtienen, organizan y codifican para contribuir al proceso de interpretación.

- El *significado de las oraciones*: los contenidos derivados de la combinación de las diferentes palabras en una unidad mayor, en especial la oración; se trata de asignar un significado a las estructuras analizadas, que por lo general se reduce a su contenido proposicional, y a especificar las condiciones bajo las cuales una oración es verdadera, atendiendo únicamente al contexto inmediato.
- El *significado del discurso*: los contenidos que se desprenden de la combinación de unas oraciones con otras en un marco superior, el del discurso; cómo se resuelve la anáfora, la elipsis y la asignación de referentes, según un contexto más amplio; cómo afectan aspectos pragmáticos tales como las intenciones de los participantes en la comunicación.

Según J. Vidal y J. Busquets (1996:416), la semántica computacional consta de tres componentes:

a) Un sistema de representación conceptual, es decir, un lenguaje manipulable por ordenador; de ahí la importancia de desarrollar formalismos adecuados para describir el significado.

b) Un módulo de traducción que relacione las expresiones de una lengua natural con el lenguaje en que esté expresada la representación conceptual.

c) Un sistema que transforme las representaciones conceptuales en entidades y relaciones de un dominio determinado (por ejemplo, que sea capaz de realizar la identificación de referentes).

En cualquier caso, el objetivo final del análisis semántico es la creación de una representación del contenido asociada a los elementos aislados en las fases previas del tratamiento computacional del

lenguaje, en especial los constituyentes en el análisis sintáctico, con el fin de establecer un vínculo con el mundo y nuestro conocimiento acerca de él, es decir, proporcionar una interpretación. Para algunos autores (*vid.* Moreno Sandoval 1998:119), se trata de un componente más universal que el sintáctico y, por tanto, menos dependiente de las lenguas particulares[160].

No obstante, pese a la importancia del tratamiento semántico, hay que señalar que este no alcanza todavía el grado de desarrollo que se ha logrado en sintaxis, quizá por la propia dificultad que entraña la tarea y también debido a que en los primeros modelos generativistas, cuyo peso en LC ya ha sido destacado, el componente semántico se consideraba secundario[161]. Además, como señala T. Badia (2003:232), no existe una teoría general del significado, por lo que solo se han considerado aspectos parciales del mismo.

En las primeras etapas de la LC, la semántica tuvo un papel nulo o mínimo, en unos casos porque se partía de planteamientos demasiado ingenuos sobre la necesidad de incluir conocimientos lingüísticos en general y, en otros, porque el interés suscitado por la sintaxis relegaba a un segundo plano todas las cuestiones relacionadas con el significado. Sin embargo, la aparición en los años ochenta de nuevos formalismos gramaticales –las gramáticas de unificación y rasgos-, en los que las descripciones semánticas de los elementos léxicos juegan un papel central, ha contribuido a una mayor preocupación por el nivel

[160] Sirvan como ejemplo las propuestas de traducción automática basadas en la existencia de una *interlengua*, un lenguaje que representa de la misma manera todas las oraciones que significan "lo mismo", con independencia de la lengua en que ese significado esté expresado. Una convención típica es el uso de papeles temáticos, como *agente* o *evento*, que se asume que son universales (*cf.* JURAFSKY Y MARTIN 2000:812).

[161] Paradójicamente, fue la llamada semántica generativa, una corriente alternativa dentro del propio generativismo, la que llamó la atención sobre la necesidad de dar cuenta de la semántica. De esta importancia se ha hecho eco el propio Chomsky en sus últimos modelos, la teoría de la rección y el ligamiento (1981) y el programa minimalista (1995).

semántico y también ha introducido más realismo en los planteamientos y limitaciones de la tarea.

Las múltiples propuestas de la lingüística teórica se han reducido hasta hace relativamente poco tiempo, en el ámbito computacional, a dar cuenta de dominios restringidos, al empleo de la lógica como convención para expresar el significado y a la representación del conocimiento del mundo dentro del dominio concreto de una aplicación.

Por otra parte, en el tratamiento del significado, hay que resaltar que en general la LC se ha beneficiado de las investigaciones procedentes de la Inteligencia Artificial, ciencia en la que era un tema prioritario cuando la Lingüística estaba ocupada en otras cuestiones, básicamente de índole sintáctica. Y también de la Psicología, preocupada por la organización de los conocimientos en la memoria.

En cuanto a las motivaciones, además de la perspectiva teórica, hay que tener en cuenta que el tratamiento semántico introduce una mejora sustancial en los resultados de los sistemas de traducción automática o de recuperación de información, como los buscadores de información por Internet.

Los principales problemas a los que se enfrenta la Semántica desde la perspectiva computacional son, según Sh. Lappin (2003:92):

- Determinar la naturaleza de la relación entre el significado de una oración y el significado de los constituyentes sintácticos que la conforman, lo que lleva a la cuestión de cómo representar el significado e implica definir la naturaleza de la interfaz entre sintaxis y semántica, si se parte de una postura que concibe el proceso de interpretación vinculado al análisis sintáctico.

- Determinar la naturaleza de la relación entre el significado de una oración y el significado de sus partes, con independencia del nivel sintáctico.
- Determinar el papel de los factores contextuales y discursivos en la interpretación final del significado de una oración.

Otros problemas surgen de las ambigüedades de diverso tipo: homonimia, polisemia, ambigüedad estructural, ambigüedad en la referencia de los pronombres, ambigüedad en el alcance de los cuantificadores; o de la determinación de las unidades de significado, sobre todo en el caso de locuciones y expresiones idiomáticas, formadas por varias palabras ortográficas pero con un significado unitario[162].

A estas dificultades se suma el que hay que partir de un hecho: la respuesta a la pregunta sobre qué es el significado no es única. Incluso dentro de la propia Lingüística no existe un solo concepto de significado. Por ejemplo, la lingüística estructural desde Saussure entiende el significado como una imagen mental[163], que no se relaciona directamente con la realidad a la que se refiere. Pero C. K. Ogden y A. Richards (*apud* Gutiérrez Ordóñez 1981:107) aíslan al menos dieciséis definiciones diferentes de significado. Además, desde la perspectiva de la LC, lo que se entiende por significado no tiene por qué coincidir necesariamente con las propuestas de la Lingüística.

Dada la diversidad de concepciones en torno al significado, no es de extrañar que los métodos para codificar la información semántica en LC

[162] Para más detalles, *vid.* GRISHMAN (1991 [1986]: cap. 3), ALLEN (1995: *part II*), VÁZQUEZ, FERNÁNDEZ y MARTÍ (2002), MORENO ORTIZ (2000) o BLACKBURN y BOS (2001, 2006a y 2006b).

[163] *Vid.* GUTIÉRREZ ORDÓÑEZ (1981), especialmente los capítulos 2 y 3, y GUTIÉRREZ ORDÓÑEZ (1989), capítulos 2, 4 y 6.

procedan de diferentes fuentes (*cf.* Vidal y Busquets 1996:416): de la Lógica, de las Matemáticas, de la Inteligencia Artificial, etc. Lo que no es un inconveniente para que, según Mª. F. Verdejo (1995:49), el tratamiento del significado represente precisamente "una de las contribuciones más importantes del PLN [Procesamiento del Lenguaje Natural] a la Inteligencia Artificial". Esta misma autora (*ibid.*:48) señala que, para obtener el significado de una secuencia de palabras, lo primero es determinar qué se entiende por significado; después, elegir el formalismo adecuado para expresarlo; por último, decidir la forma de llevar a cabo el proceso de interpretación.

2.4.1. Formalismos para la representación del significado

Igual que ocurre con los demás niveles lingüísticos, la representación formal es un requisito indispensable para que el significado pueda ser abordado computacionalmente. Dicha representación debe ajustarse a una serie de parámetros. Jurafsky y Martin (2000:504 y ss.) mencionan:

i) La verificabilidad, es decir, poder determinar la verdad o falsedad de un enunciado. Normalmente se hace comparando la representación del significado con una base de conocimientos sobre el mundo.

ii) La ausencia de ambigüedad, ya que esta puede dar lugar a múltiples interpretaciones, lo que supone un mayor coste de procesamiento computacional.

iii) Cierto grado de vaguedad, útil en ciertos casos. Por ejemplo, puede ser suficiente con especificar "italiano" al lado de un restaurante en un servicio telefónico de consulta de información sobre restaurantes, sin necesidad de detallar el tipo específico de comida que se puede pedir.

iv) El empleo de una *forma canónica*, es decir, asignar la misma representación semántica a oraciones que expresan el mismo contenido proposicional, pese a que difieran en la forma. Así, continuando con el ejemplo anterior, alguien puede formular la pregunta al sistema de información telefónica de diferentes formas: *¿Hay un italiano en la zona X?*, *¿Dónde puedo encontrar un italiano?*, *¿El restaurante Y sirve comida italiana?*, etc.

v) Capacidad para realizar inferencias, es decir, deducir información implícita (generalmente conocimiento del mundo) a partir de la representación del significado. Por ejemplo, a partir del enunciado: *Me encanta la pasta: ¿a qué restaurante puedo ir?*, esperaríamos que un sistema computacional pudiera inferir que debe buscar restaurantes de comida italiana en su base de datos.

vi) Expresividad para que el sistema sea capaz de comprender una amplia gama de asuntos, no solo los referidos a un dominio particular.

Se suele entender la interpretación semántica "como la proyección de las estructuras de la lengua natural en una lengua formal que permita al ordenador hacer un uso directo de la información" (Moreno Sandoval 1998:120). Desde esta perspectiva, el primer paso del proceso de interpretación consiste en traducir el enunciado de una lengua natural a una lengua formal, cuyo resultado se suele denominar *forma lógica*, que luego se completará con la información procedente del nivel del discurso.

El proceso de interpretación se lleva a cabo en dos fases: en la primera, se construye el significado de la oración a partir de los constituyentes, eliminando las ambigüedades; y en la segunda, a partir de la forma lógica no ambigua obtenida, se resuelven las posibles

referencias anafóricas o pronominales, así como aquellos aspectos que dependan del contexto (*vid*. Moreno Boronat *et al.* 1999).

Los formalismos para representar el significado se pueden agrupar en dos grandes familias (*cf.* Verdejo 1995:49):

a) Los basados en la *lógica*, que expresan el significado como una fórmula o un conjunto de fórmulas mediante un lenguaje lógico (la *semántica formal*; *cf.* Vidal y Busquets 1996:421-425).

b) Los basados en *redes semánticas* y en *marcos*, que lo expresan como un conjunto de nodos y arcos o un objeto estructurado respectivamente (las representaciones semánticas basadas en el conocimiento; *cf.* Vidal y Busquets 1996:417-421).

A. Moreno Sandoval (1998:123) también comenta la posibilidad de codificar la información semántica mediante *rasgos*, como los descritos para el tratamiento de la morfología o la sintaxis, que se pueden incorporar en las reglas mediante restricciones de selección. De esta manera, se abortan análisis posibles desde el punto de vista sintáctico, pero no desde el semántico, con lo que se gana en eficiencia y precisión en los sistemas computacionales.

Usar la lógica como lenguaje formal de representación del significado cuenta con una serie de ventajas: ausencia de ambigüedad, uso de reglas simples de interpretación e inferencia o facilidad de traducción de la estructura sintáctica a una forma lógica. Por otra parte, la lógica en sus diferentes variantes (lógica de predicados de primer orden, lógica proposicional, cláusulas de Horn, cálculo lambda, teoría de los modelos de Montague, lógica modal y temporal, lógica de

creencias, etc.) está bien estudiada y ha sido empleada desde hace siglos para el propósito de representar el significado de las oraciones.

La semántica formal se inspira precisamente en la lógica y en las nociones de: i) composicionalidad (establecida por Frege, postula que el significado del todo se deriva del significado de las partes), y ii) significado entendido como condiciones de verdad; es decir, una oración será verdadera o falsa según la correspondencia que se establezca entre los referentes y el mundo real; esta es la respuesta de la lógica a qué entiende por significado. En este caso, el problema del significado se centra en dilucidar qué relación se establece entre las expresiones lingüísticas y la realidad y, sobre todo, en la naturaleza de ese vínculo (*cf.* Vidal y Busquets 1996:421).

Como resulta evidente, la naturaleza de la interfaz sintaxis-semántica juega un papel importante, ya que la representación del significado se construye sobre la representación sintáctica: el significado de una expresión compleja (las condiciones de verdad de una oración) depende del significado de las partes que la componen, lo que implica que las reglas sintácticas y las semánticas están estrechamente relacionadas, es decir, el análisis sintáctico y el semántico deberían ser simultáneos. Es lo que se conoce como *hipótesis regla por regla*: a cada regla sintáctica le corresponde una regla semántica. Así, el proceso de interpretación parte del análisis sintáctico, como se observa en el siguiente ejemplo (Moreno Sandoval 1998:122), hasta llegar a la forma lógica, a partir de la cual se realiza la interpretación semántica:

Lengua natural	*Abelardo ama a Eloísa*
Estructura de constituyentes	[O[SN Abelardo] [SV[V ama] [SP [P a] [SN Eloísa]]]]
Estructura funcional	(PRED ama SUJETO Abelardo OBJ-DIR Eloísa)
Forma lógica	amar (Abelardo, Eloísa).

Tabla 17. De la estructura sintáctica a la forma lógica.

El significado de la oración será verdadero si el significado del SN y del SV lo es.

El uso de la lógica en sus diversas variantes se ha erigido, desde los primeros momentos, en un mecanismo interesante, pese a las limitaciones que presenta, al ser en sí misma un lenguaje de representación del significado de las lenguas naturales. De todas las posibilidades, la lógica de predicados de primer orden es la que más ampliamente se ha empleado: cuenta con una larga tradición, es flexible y se puede trasladar fácilmente a un programa informático, algunos de los principales requisitos exigidos a un lenguaje de representación. Además, impone pocas restricciones en la forma de representar el significado, pues reduce todo a fórmulas lógicas que dan cuenta de los *términos* (u objetos) mediante constantes u objetos específicos (por ejemplo, un restaurante italiano concreto, como puede ser *Piccolo*), funciones o predicados de un único argumento referidos a un objeto (*la dirección del Piccolo*) y variables o generalizaciones sobre objetos que permiten efectuar inferencias y hacer afirmaciones (*Los italianos son caros; Piccolo es un restaurante italiano; El restaurante Piccolo es caro*) (*vid.* Jurafsky y Martin 2000:513 y ss.).

La lógica, sin embargo, no es la única opción que se puede emplear como lenguaje de representación. También se ha hecho amplio uso de las representaciones del significado basadas en el conocimiento, que aprovechan una serie de técnicas tomadas de la Inteligencia Artificial, inspiradas a su vez en teorías de tipo asociativo sobre el almacenamiento de los conceptos en la memoria humana. De esta forma, los significados de las expresiones lingüísticas o conocimiento se representan mediante redes de vínculos que relacionan los conceptos.

Las técnicas para representar el conocimiento pueden ser de dos tipos (*cf.* Vidal y Busquets 1996:417-418):

a) *Métodos declarativos o estáticos*. La mayor parte del conocimiento se representa como una colección de hechos y, además, constan de un pequeño conjunto de procedimientos para poder manipularlos. Sus ventajas son dos: i) cada hecho se almacena una sola vez, independientemente de sus usos; ii) la suma de un nuevo hecho no altera el sistema. Incluyen las redes semánticas, la dependencia conceptual, los marcos y los guiones.

b) *Métodos procedurales o procedimentales*. El conocimiento se representa como un conjunto de instrucciones, como una acción. Es el tipo de semántica de sistemas como el ya comentado SHRDLU. Sus ventajas también son dos: i) la facilidad con que representan el modo en que opera el conocimiento; ii) su capacidad para representar ciertos tipos de conocimiento que no se adaptan a los esquemas declarativos (razonamiento por defecto o el probabilístico, por ejemplo).

Los más empleados son los métodos declarativos. Destacan las redes semánticas, la alternativa más frecuente a las formas lógicas. Se sirven de representaciones gráficas para reflejar las estructuras de significado. Se definen como “nodos conceptuales, en los que cada nodo o célula

simboliza un concepto y cada uno de los conceptos se relaciona con otras palabras mediante un arco. Los arcos están etiquetados según el tipo de relación existente entre cada uno de los conceptos" (Vidal y Busquets 1996:418). Es lo que se denomina *grafos coloreados* en matemáticas y *grafos etiquetados* en LC. De acuerdo con este formalismo, el significado de la oración *Pedro regala un libro a María* tendría la siguiente representación (Vidal y Busquets *id.*):

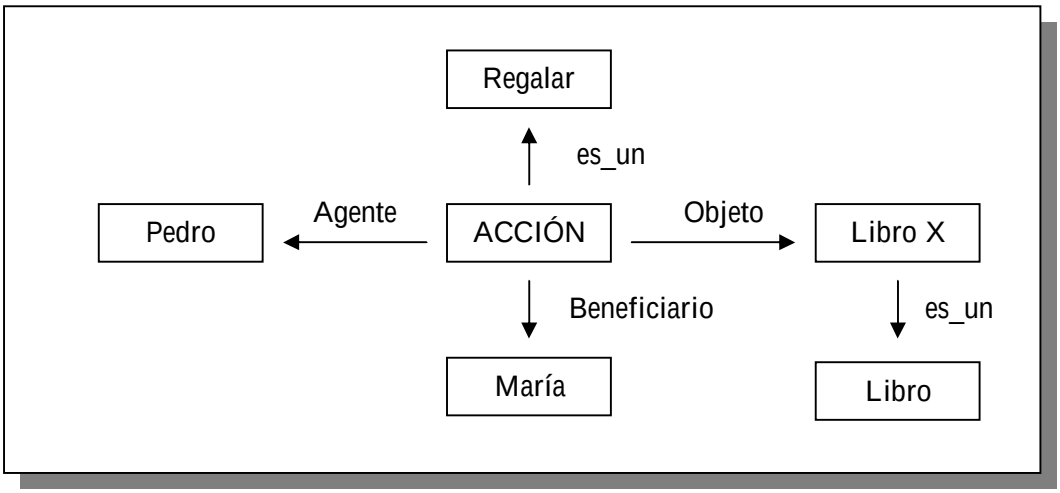


Ilustración 79. Ejemplo de red semántica.

Inspiradas en las teorías sobre la memoria de M. R. Quillian (1968), las redes semánticas constan de dos componentes fundamentales (*cf.* Moreno Ortiz 2000):

a) Estructuras de datos o *nodos* –que representan conceptos– unidos por *arcos* –que representan relaciones entre los conceptos.

b) Procedimientos de inferencia que operan sobre las estructuras de datos.

Existen varios tipos de redes semánticas. Las más populares son las *redes IS-A*, en las que los enlaces entre nodos están etiquetados. Según A.

Moreno Ortiz (*id.*), "una red IS-A es una jerarquía taxonómica cuya espina dorsal está constituida por un sistema de enlaces de herencia entre los objetos o conceptos de representación, conocidos como *nodos*". Su fundamentación hay que buscarla en que el conocimiento se suele estructurar mediante la adscripción de unos elementos a otros más generales, de forma similar a como ocurre en las taxonomías clásicas de las ciencias naturales: "un perro es un cánido", "un cánido es un mamífero", "un mamífero es un animal". De este ejemplo se pueden deducir otras características interesantes de las redes semánticas: la inferencia ("un perro en un animal") y la herencia de propiedades (*perro* heredará todas las características definitorias de las clases superiores en la jerarquía -*cánidos*, *mamíferos* y *animales*-).

La teoría de la Dependencia Conceptual, propuesta por R. Schank a principios de los setenta (1972, 1975), sugiere un nivel de representación semántica independiente de las lenguas particulares, en principio concebido como la interlengua de un sistema de traducción automática llamado MARGIE. Parte de una idea, surgida dentro de la tradición generativa en la línea de Katz y Fodor, según la cual los significados podían ser representados mediante conjuntos estructurados de primitivos semánticos: así, por ejemplo, el significado de *mujer* sería: [+humano +femenino +adulto]. Basándose en estas ideas y en las aportaciones de la gramática de casos de Ch. Fillmore, R. Schank desarrolló la teoría de la Dependencia Conceptual (DC). En dicha teoría sostiene que con tan solo siete primitivos (posteriormente estos llegaron a cuarenta), denominados ACTS, es suficiente para describir la mayoría de eventos, tanto físicos como mentales (*cf.* Schank y Abelson 1977:12-14):

ACCIONES	DEFINICIONES
[MOVE]	Mover cualquier parte del cuerpo de un espacio físico a otro.
[PTRANS]	Cambiar de espacio físico algún objeto de un lugar a otro (puede incluir al agente de la acción).
[MTRANS]	Transferir información de un lugar a otro.
[SPEAK]	Proferir, con algún tipo de entonación, una secuencia bien formada de fonemas de alguna lengua determinada.
[PROPEL]	Aplicar un esfuerzo físico a un objeto con el fin de moverlo.

Tabla 18. Algunas de las acciones primitivas en la teoría de la Dependencia Conceptual.

Además de los ACTS o acciones primitivas, R. Schank completa su teoría con unas escalas de 10 a -10 para indicar cambios en un estado (por ejemplo, en la salud, en el estado mental, en el estado físico, en el grado de consciencia, etc.).

De acuerdo con esto, una oración como *Bill disparó a Bob con una pistola* es representada de la siguiente manera:

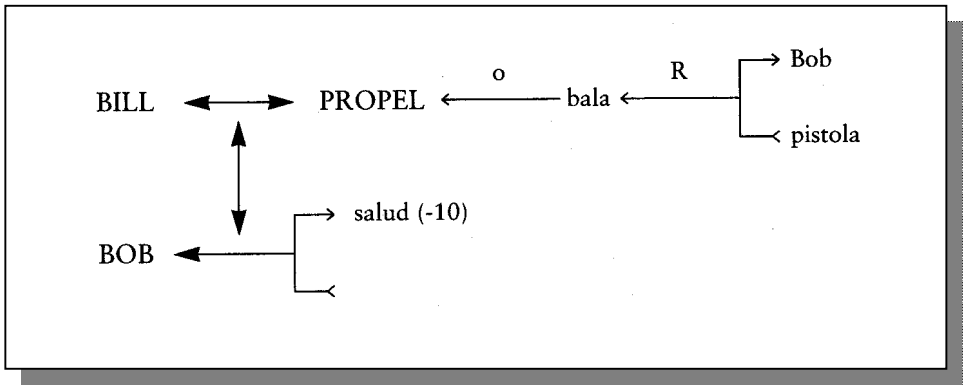


Ilustración 80. Ejemplo de Dependencia Conceptual.

Esta teoría fue criticada por la arbitrariedad de los primitivos escogidos y la falta de sistematicidad en su uso y adopción.

Los marcos o *frames* son una técnica de representación del conocimiento que surgió como consecuencia de que mecanismos como los expuestos anteriormente (Dependencia Conceptual) solo sirven para representar un tipo específico de eventos o experiencias. Mediante los marcos, por lo tanto, se pretende dar cuenta de generalizaciones, así como de la organización y estructura del conocimiento y del razonamiento (*cf.* Vidal y Busquets 1996:419). Tienen su origen en la IA, en la teoría de M. Minsky (1975:211), quien define los marcos de la siguiente manera: "When one encounters a new situation (or makes a substantial change in one's view of a problem), one selects from memory a structure called a *frame*. This is a remembered framework to be adapted to fit reality by changing details as necessary".

Consisten más o menos en redes semánticas complejas, que recuperan estructuras previas almacenadas en nuestra memoria para clasificar o adaptar a ellas situaciones nuevas. Por ejemplo, cuando asistimos a una conferencia, esperamos encontrar una serie de elementos (un orador, una audiencia, etc.), que acomodamos a los estereotipos que ya poseemos sobre este tipo de situación. Por lo tanto, los marcos dan cuenta, de forma jerarquizada, de situaciones estereotipadas mediante unas casillas o *slots* que representan los elementos de que consta la situación. Las casillas se llenan con las instancias concretas en una situación determinada, que también pueden ser valores por defecto.

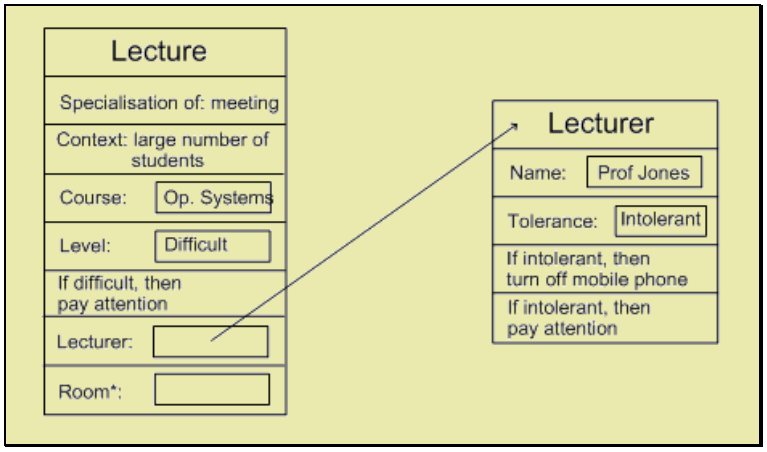


Ilustración 81. Ejemplo de marco[164]*.*

Por último, pero estrechamente relacionados con las redes semánticas y los marcos, están los guiones o *scripts*, que describen "una secuencia de eventos que se desarrollan previsiblemente en un contexto particular". Son el resultado de la propia evolución de la teoría de la Dependencia Conceptual (Schank y Abelson 1977). El ejemplo clásico que proponen sus autores es el del restaurante: entrar, pedir la comida, pagar, etc. serían los esquemas o secuencias de acontecimientos que caracterizan dicha situación. Otro ejemplo es el siguiente, para la compra de un billete de tren de Barcelona a Puigcerdà, que J. Vidal y J. Busquets (1996:429) adaptan de Allen (1995 [1987]):

T1 es un «script» del tren con

[agente: PEDRO1]
[Otro: Otro(T1)]
[Origen: BARCELONA1]
[Destino: PUIGCERDÀ]
[Tren: TR3 desde(TR3, Origen(T1)) y hacia(TR3, PUIGCERDÀ)]
[Billete: Billete(T1) BILLETE_HACIA(Billete(T1), PUIGCERDÀ))
[Taquilla: Taquilla(T1) EN(Otro(T1), Taquilla(T1))]

Precondiciones: TENER(PEDRO1, Precio(TR3))
Borrar lista: TENER(PEDRO1, Precio(TR3))
Añadir lista: EN_TREN(PEDRO1, TR3)

Descomposición:
IR_A(PEDRO1, Taquilla(T1)) →
DAR(PEDRO1, Otro(T1), Precio(Billete(T1))) →
DAR(Otro(T1), PEDRO(T1), Billete(T1)) →
IR_A(PEDRO1) ANDÉN(TR3)) → EN_TREN(PEDRO, TR3)

Ilustración 82. Ejemplo de guion.

[164] Tomado de: http://www.doc.ic.ac.uk/~sgc/teaching/v231/lecture4.html

2.4.2. El tratamiento del léxico

Aunque, en general, a un sistema computacional le interesa la extracción del significado oracional, sin embargo, el tratamiento semántico en sentido amplio gira en torno a las palabras o unidades léxicas, las entradas individuales de los diccionarios, un conocimiento básico en LC, sobre todo si queremos sistemas capaces de trabajar con textos reales sin restricciones.

Es la Semántica Léxica la disciplina lingüística que se ocupa de los contenidos que presentan los lexemas de forma individual.

El léxico de una lengua se concibe como un todo organizado en torno a:

i) la estructura interna de los lexemas y

ii) las relaciones que se establecen entre ellos.

El *lexicón*, entendido como un conjunto finito de lexemas de una lengua, es un componente central en los sistemas computacionales. De hecho, conforma uno de los módulos imprescindibles junto a la gramática, ya que aporta información fundamental para el análisis morfológico y sintáctico, así como para determinar el significado de las oraciones y, por ende, del discurso. Por otra parte, ya se ha señalado el papel destacado que desempeña en los últimos formalismos gramaticales basados en la unificación y los rasgos, así como en la propia Lingüística Teórica.

M.ª A. Martí e I. Castellón (2000:98) consideran que son tres los aspectos centrales en relación con el tratamiento del léxico:

- Determinar el contenido semántico de cada lexema: *conocimiento léxico*.
- Definir un lenguaje de representación para ese contenido: *formalización*.
- Proponer métodos para adquirir el conocimiento léxico de forma automática o semiautomática: *fuentes de conocimiento*.

Por lo que respecta al conocimiento léxico, la información que esperamos encontrar asociada a un lexema es más o menos la que suelen proporcionar los diccionarios tradicionales.

Un diccionario recoge las formas léxicas de las palabras, sobre las que facilita datos como: entrada léxica (con variantes ortográficas si las hay), clase de palabras, estructura sintáctica (uso transitivo/intransitivo, etc.), variantes flexivas o peculiaridades morfológicas, pronunciación, definición (y restricciones para cada posible significado, sinónimos, etc.) y ejemplos (*cf.* Moure y Llisterri 1996:198-199).

A diferencia de los diccionarios convencionales, en un lexicón computacional toda la información tiene que estar explicitada, sistematizada y formalizada. Por este motivo, el traslado de las entradas de los diccionarios en papel o, incluso, de los diccionarios electrónicos, muchas veces meros calcos de los otros en un soporte distinto, es con frecuencia una tarea ardua y salpicada de dificultades, debido a la circularidad que prima en las definiciones. Aun así, los diccionarios son una fuente importante de datos.

La información que un sistema computacional suele demandar del léxico incluye tanto aspectos morfológicos como sintácticos y los específicamente semánticos. En concreto, para M.ª A. Martí e I. Castellón (2000:100-103), debe contemplar:

- El lema transcrito ortográficamente. En caso de tener varias acepciones, cada una se consigna como una entrada diferente del lexicón.
- Transcripción fonética, fundamental para los sistemas de síntesis del habla[165].
- Modelo flexivo.
- Categoría gramatical (p. ej. verbo) y subcategorización (p. ej. verbo transitivo).
- Estructura argumental: patrón sintáctico-semántico que caracteriza los argumentos que acompañan a un verbo, nombre o adjetivo mediante papeles temáticos en consonancia con la propuesta de Ch. Fillmore (1968). P. ej. el verbo *leer* se caracteriza por tener un SN sujeto con el papel temático (o función semántica) de "agente" y un SN objeto directo con el de "tema". Este apartado incluye también las restricciones de selección o exigencias semánticas que deben cumplir los argumentos que acompañan a un determinado lexema. Así, podemos decir que el verbo *comer* normalmente exige un sujeto desempeñado por un SN con el rasgo semántico de [+animado].
- Información semántica propiamente dicha: glosa o definición, categoría semántica, forma lógica, rasgos semánticos…

[165] En el ámbito de las tecnologías del habla, sistemas que transforman un conjunto de caracteres escritos en su equivalente oral (conversión de texto en habla).

- Relaciones léxicas que se establecen entre lexemas, como:
    - Homonimia: lexemas con la misma forma y diferente significado.
    - Polisemia: un lexema con varios significados.
    - Sinonimia: lexemas con significados muy próximos.
    - Hiponimia: relación entre un término específico y uno más general (*uña-dedo*).
    - Hiperonimia: relación entre un término general y uno más específico (*dedo-uña*).
    - Etc.
- Equivalencias en otras lenguas, de especial interés en sistemas multilingües de traducción automática o de recuperación de información.

En los diccionarios computacionales, la cantidad y variedad de información son dos temas centrales. Pero, además, un buen diccionario computacional se debe caracterizar por una clara división de los tipos de información (*cf.* Moreno Sandoval 1998:130). Por otra parte, la dificultad de la tarea de construir un lexicón computacional cambia radicalmente de trabajar en un dominio concreto, en el que el número de elementos léxicos es menor y menos ambiguo, a ampliar las miras hacia un lexicón más general.

La forma de codificar la información puede variar: mediante bases de datos generales de tipo relacional[166], utilizando corpus etiquetados[167] o

[166] Que aprovechan la arquitectura de *software* que les proporcionan productos comerciales ya existentes y que son fáciles de mantener y de interrogar, aunque los datos se encuentran fragmentados y no es posible establecer inferencias o generalizaciones (*cf.* MARTÍ Y CASTELLÓN 2000:103-105).

diccionarios en soporte electrónico, por medio de bases de datos léxicas[168], o mediante estructuras de rasgos[169], que están bastante generalizadas hoy en día para este propósito, como comentaremos a continuación.

Sea cual sea el sistema adoptado para representar la información, este ha de ajustarse a una serie de criterios (*cf.* Martí y Castellón 2000:102-103):

- Expresividad: capacidad de representar todo el conocimiento necesario.
- Idoneidad "representacional": adecuación al tipo de información.
- Idoneidad inferencial: inclusión de mecanismos de inferencia.
- Eficiencia en la forma de acceder al léxico.

Además, hay que tener en cuenta que el lexicón depende habitualmente de la gramática, ya que las entradas léxicas no son otra cosa que los elementos terminales que se insertan en las reglas gramaticales. En el siguiente ejemplo, tomado del lexicón del proyecto PROTEUS[170], se observa cómo se integra información morfosintáctica

---

[167] Conjuntos estructurados de textos, en los que cada palabra porta etiquetas con información de diferente tipo (morfológica, sintáctica o semántica). Así, por ejemplo, se puede obtener una lista con todas las palabras que lleven la etiqueta "verbo". Con programas adecuados (TACT, WordSmith, Word Cruncher…) incluso se puede obtener información cuantitativa sin que el corpus esté etiquetado: frecuencias y contextos de aparición, colocaciones, etc.

[168] Que combinan las ventajas de las bases de datos con fuentes textuales como los diccionarios electrónicos.

[169] A diferencia de otros acercamientos a la representación léxica, los formalismos basados en los rasgos y en la unificación aportan el uso de mecanismos deductivos (como la herencia de propiedades) y de restricciones.

[170] PROTEUS (*PROtotype TExt Understanding System*) es un sistema multilingüe de procesamiento del lenguaje natural, desarrollado desde 1986 en el Dpto. de Informática de la Universidad de Nueva York, bajo la dirección de R. Grishman. La versión española corre a cargo de A. Moreno Sandoval y C. Olmeda. Su objetivo es desarrollar un sistema capaz de analizar textos y extraer información de ellos. Está

(nombre, masculino, singular, objeto directo, ditransitivo…) y semántica (humano) (*cf.* Moreno Sandoval 1998:134):

```
(noun  :masc-sing  ALCALDE    :attributes (HUMAN))
(verb  :baseform   COMPRAR
       :attributes (reflexive)
       :objlist    (DIRECT-OBJ DITRANS DIROBJ-PN (PVAL (PARA)))
```

Ilustración 83. Integración de la información semántica con la morfosintáctica.

En este sentido, se tiende cada vez más hacia un enfoque basado en el conocimiento, acorde con la generalización de los formalismos gramaticales basados en la unificación: parte de la información o conocimiento se representa explícitamente, mediante estructuras jerarquizadas (estructuras de rasgos), y parte se deriva por inferencia, aplicando mecanismos de herencia de propiedades.

Las estructuras de rasgos y la herencia de propiedades son especialmente adecuadas para definir subclases de elementos léxicos, ya que se abrevia considerablemente la codificación de la información para todos los elementos que pertenecen a la misma subclase y se evita la redundancia: basta especificar aquella una vez en plantillas o macros léxicas, que contienen conjuntos de rasgos. Evidentemente, el establecimiento de clases y subclases es de suma importancia, así como la organización de las jerarquías de rasgos.

Un ejemplo de entrada léxica que se ajusta a esta convención es el siguiente (*cf.* Moreno Ortiz 2000), del proyecto Acquilex:

---

basado en la teoría lingüística de Z. Harris y el *Linguistic String Project* de N. Sager (*cf.* MORENO SANDOVAL *et al*. 1993).

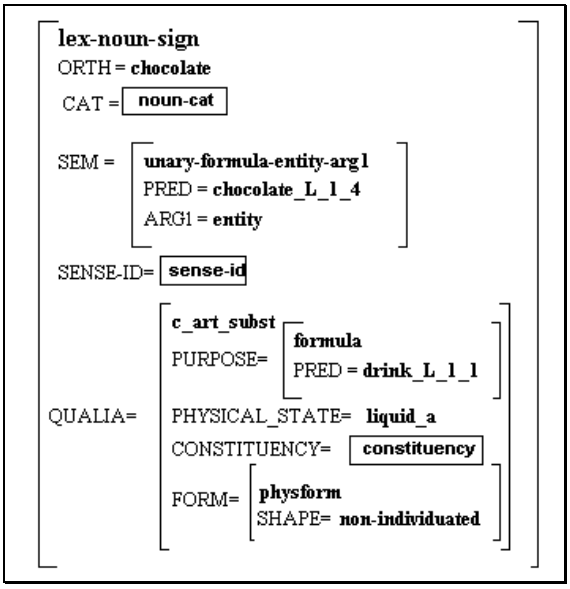


Ilustración 84. Entrada léxica del proyecto Acquilex[171].

En este caso concreto, la entrada léxica (*chocolate*) está representada mediante una estructura de rasgos tipificada[172] que "hereda" la información contenida en el atributo "qualia" (sirve para beber, es un líquido, etc.).

[171] El proyecto Acquilex I y II (*The Acquisition of Lexical Knowledge for Natural Language Systems*) tenía por objetivo la adquisición de información léxica a partir de varios diccionarios electrónicos y corpus de distintas lenguas para integrarla en una base de datos única y multilingüe. Representa un esfuerzo coordinado en los trabajos relacionados con la uniformidad de las entradas y la estructuración de la información, dos temas importantes en lo que a la representación del léxico se refiere (*cf.* MOURE Y LLISTERRI 1996:201). En el proyecto participaron diversas universidades europeas (Ámsterdam, Politécnica de Cataluña, Cambridge, Dublín, Pisa) y editoriales (Cambridge University Press, Biblograf, Garzanti y Van Dale Lexicografie). Se desarrolló entre 1989 y 1995. URL: http://www.cl.cam.ac.uk/research/nl/acquilex/

[172] La negrita de la ilustración representa los tipos (como *lex-noun-sign*, que especifica las propiedades sintácticas y semánticas de los nombres), los atributos van en mayúscula y los recuadros remiten a información externa, almacenada fuera de la entrada. Sigue el modelo de los "qualia", propuesto por J. PUSTEJOVSKY (1991, 1995) en su teoría léxico-semántica del lexicón generativo, que pretendía superar las limitaciones de las restricciones de selección.

Veamos otro ejemplo, tomado de A. Moreno Sandoval (1998:136):

Macro verbos-trans:		*Macro verbos-ditrans:*	
<cat>	= V	Verbos-trans	
<arg0 cat>	= SN	<arg2 cat>	= SP
<arg0 función>	= sujeto	<arg2 valor-p>	= a
<arg1 cat>	= SN	<arg2 función>	= obj-indir
<arg1 función>	= obj-dir		

Tabla 19. Macros para verbos transitivos y ditransitivos.

Aquí se representan dos macros para caracterizar dos tipos de verbos, los transitivos y los ditransitivos. Esta última remite, a su vez, a la macro para los transitivos (*Verbos-trans*), por lo que no es necesario repetir en ella rasgos que ya están especificados en otra macro. Cualquier verbo que se etiquete como transitivo o ditransitivo incluirá, automáticamente en su entrada del lexicón, la información contenida en estas macros.

Aunque las estructuras de rasgos son "monotónicas", es decir, toda la información se conserva, recientemente han sido propuestos acercamientos no "monotónicos" que permiten introducir reglas que admiten excepciones[173]. Es lo que sucede cuando, por ejemplo, un elemento que pertenece a una clase cumple todos los requisitos menos uno; en vez de codificarlo aparte, como una excepción, la propia regla permite dar cuenta de este fenómeno, al incorporar los conceptos de

[173] Aplican redes semánticas, originalmente desarrolladas en el ámbito de la Inteligencia Artificial para la representación del conocimiento general, a la descripción del léxico.

herencia por defecto y "sobrescritura" (cuando un nodo inferior ya tiene asignado un valor para un rasgo concreto, no hereda los rasgos por defecto):

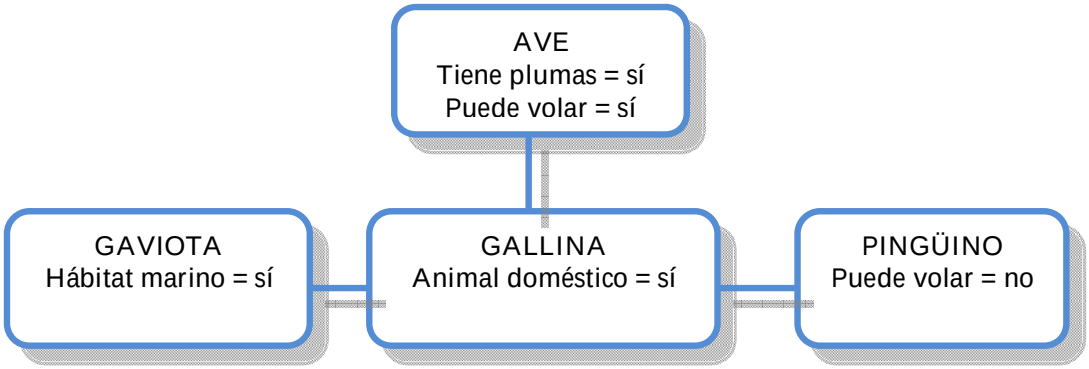


Ilustración 85. Acercamiento no "monotónico" al léxico.

La clase AVE se define por la siguiente estructura de rasgos (herencia por defecto): *tiene plumas = sí* y *puede volar = sí*. Estas características son heredadas por todos los elementos de la clase: *gaviota*, *gallina*, *pingüino*, etc. que, a su vez, pueden poseer especificaciones propias (*hábitat marino = sí*, *animal doméstico = sí*) que sobrescriben las heredadas en caso de conflicto, que es lo que ocurre en el caso de *pingüino* (*puede volar = no*)[174].

R. Evans y G. Gazdar (1996) junto con B. Keller (1995, 1996) han desarrollado un lenguaje de representación léxica, DATR[175], similar a un lenguaje de programación, que es especialmente adecuado para un tratamiento semántico como el que acabamos de describir. Parte de la idea de que las irregularidades en los lexemas se pueden abordar considerando que estos son regulares excepto en una o dos características. Sobre esta asunción proponen un lenguaje de representación no "monotónico", que trata explícitamente la semántica

[174] Un modelo "monotónico" tendría que eliminar este rasgo al entrar en conflicto con otro y, por lo tanto, impedir la unificación.

[175] URL: http://www.informatics.susx.ac.uk/research/groups/nlp/datr/

y la herencia, es fácil de implementar, es eficiente y está en línea con la forma de codificar las entradas léxicas en las gramáticas de unificación.

La no monotonía presenta las siguientes ventajas (*cf.* Moreno Sandoval 1998:139):

1) Permite expresar generalizaciones una sola vez, evitando la redundancia.

2) Utiliza un procedimiento simple y general tanto para las regularidades como para las irregularidades.

3) Las excepciones se marcan como tales, pero incluidas dentro del modelo.

Entre las iniciativas para representar la información léxica, hay que destacar el proyecto *FrameNet*[176], que se está llevando a cabo en el *International Computer Science Institute* de Berkeley (California), bajo la dirección de Ch. Fillmore. Su objetivo es construir una base de datos léxicos para el inglés (10000 unidades), en el que cada verbo (o predicado) incluya en su entrada las posibilidades de combinación que tiene, tanto sintácticas como semánticas (*cf.* Jurafsky y Martin 2000:613). Se inspira en la semántica cognitiva (marcos) y en la metodología de corpus[177].

[176] Inicialmente desarrollado para el inglés, en la actualidad se trabaja en la construcción de bases de datos léxicas equivalentes para otras lenguas, así como en la posibilidad de etiquetar semánticamente textos de forma automática. Más información en la URL: http://framenet.icsi.berkeley.edu/

[177] Para el español, el proyecto se desarrolla en la Universidad Autónoma de Barcelona, dirigido por Carlos Subirats: *Spanish FrameNet*. URL: http://gemini.uab.es:9080/SFNsite

Como ya hemos comentado (*vid. supra*), los marcos son estructuras conceptuales que capturan los elementos y las acciones necesarios para caracterizar una situación. Pueden ser más generales o más específicos; estos últimos heredan las características de los marcos generales (*cf.* Baker, Fillmore y Lowe 1998):

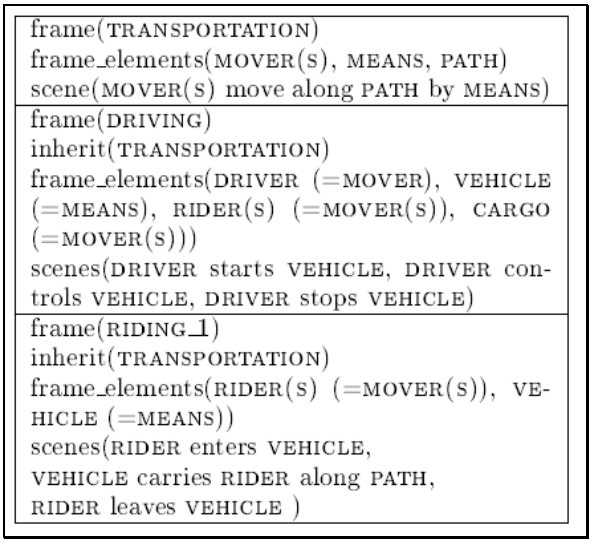


Ilustración 86. Ejemplo de marco conceptual: TRANSPORTATION.

En el ejemplo anterior, existe un marco general, TRANSPORTATION, cuyos elementos son MOVER, MEANS, PATH; y marcos más específicos, como DRIVING o RIDING_1, que heredan los elementos del marco general mediante la instrucción *inherit* (TRANSPORTATION) presente en ambos. En el caso de DRIVING, esta herencia se plasma en DRIVER = MOVER, VEHICLE = MEANS, etc. A cada uno de estos elementos se le puede asignar una función sintáctica (DRIVER = sujeto, VEHICLE = objeto directo…). Así, ante un enunciado, es posible identificar los elementos que participan en el marco asociado a un verbo y guiar de esta manera el análisis sintáctico y, si es preciso, la interpretación de la oración:

Now [D *Van Cheele*] *was driving* [R *his guest*] [P *back to the station*].

Van Cheele	= DRIVER	(Sujeto)
his guest	= RIDER	(Objeto directo)
back to the station	= PATH	(Compl. circunstancial)

Otro ejemplo de marco es el siguiente, tomado de Moreno Ortiz (2000), para dar cuenta de las acepciones del verbo inglés "roar" ('rugir', 'bramar') en el campo léxico del sonido. Fijémonos en las entradas para dos de sus acepciones:

roar 1 ***Frame:*** ROAR
Dimension: *to make a sound like an angry or wild animal*
Parent: *SOUNDS PRODUCED BY ANIMALS (To make a sound like an animal)*
Definition: *to make a very loud noise like a lion*
CP #1 ***SVAM***
S *+Animal Ag BIG_FELINE*
AM +Percep Man PERCEPTUAL_ATTRIBUTE
e.g. *The lion was roaring triumphally*

Ilustración 87. Ejemplos de marco conceptual: ROAR 1.

roar 3 Frame: *LAUGH*
Dimension: *to make a sound expressing happiness*
Parent: *to make a sound indicating an emotion*
Definition: *to laugh loudly and noisily.*
CP #1 SV(AdM)
S *+Hum Ag HUMAN*
AdM *with -Conc Eff POSITIVE_STATE*
e.g. *He threw back his head and roared with laughter.*

Ilustración 88. Ejemplos de marco conceptual: ROAR 3.

La acepción 1 (*to make a very loud noise like a lion*) nos remite al marco ROAR, mientras que la 3 *(to laugh loudly and noisily*) lo hace al marco LAUGH. Este simple hecho nos permite inferir que en el primer caso estamos ante una acepción del verbo relacionada con "sonidos emitidos

por un animal", lo que a su vez implica que el verbo en esta acepción irá acompañado de un sujeto "agente" con el rasgo [+animal] de tipo "grandes felinos" (es decir, un gato nunca podrá ser sujeto de este verbo) y un adverbio de modo con determinados rasgos semánticos. En cambio, la acepción 3 del verbo nos remite a otro marco y, por tanto, a otra información semántica (y sintáctica): en este caso la acepción está relacionada con "emitir un sonido para expresar una emoción", el verbo va acompañado de un sujeto "agente" [+humano] y puede llevar un complemento de modo[178].

Una posible representación en forma de marco de esta segunda acepción sería la que figura debajo, en la que la entrada hereda también las propiedades de los marcos más generales de los que depende (EMIT_SOUND y EMOTIONAL_EVENT) (*cf.* Moreno Ortiz 2000):

---

[178] Como se puede deducir fácilmente de los ejemplos, este tipo de representación de la información léxica de los predicados, aunque vinculada a palabras individuales, sirve de guía para la interpretación semántica de las oraciones y también para su análisis sintáctico. Por eso, en los últimos años, la descripción del componente léxico ha ganado peso en las teorías lingüísticas y, en consecuencia, también en LC, donde ahora mismo ocupa un papel central. El verbo, en este sentido, adquiere una relevancia considerable, como eje en torno al cual se articula toda la estructura sintáctico-semántica de la oración. Además, enfatiza la interrelación entre ambos componentes, el sintáctico y el semántico. En el marco comentado (*vid. supra*), el significado del verbo *roar* condiciona la naturaleza semántica del nombre que va a ocupar la función de sujeto: en un caso, este deberá tener el rasgo [+animal] *–roar 1–*, mientras que en otro, [+humano] *–roar 2*; pero, al mismo tiempo, proporciona indicios sobre la estructura sintáctica que esperamos encontrar: un SN como sujeto y un adverbio de modo como complemento circunstancial.

En palabras de G. VÁZQUEZ, A. FERNÁNDEZ y M. A. MARTÍ (2002:31):

> La semántica del verbo incluye diferentes aspectos, unos relativos a su propio significado (clase semántica y tipo eventivo) y otros al de sus argumentos. Ambos aspectos no son independientes entre sí, sino que se entrelazan, es decir, la presencia de unos implica la de otros. […]
>
> Por otro lado, la eficiencia de los sistemas de PLN depende en parte de la posibilidad de asociar las piezas léxicas que componen una oración con las unidades de significado en que se descompone la interpretación de esta, por lo que es indispensable la inclusión de un sistema adecuado de correspondencias entre ambos niveles.

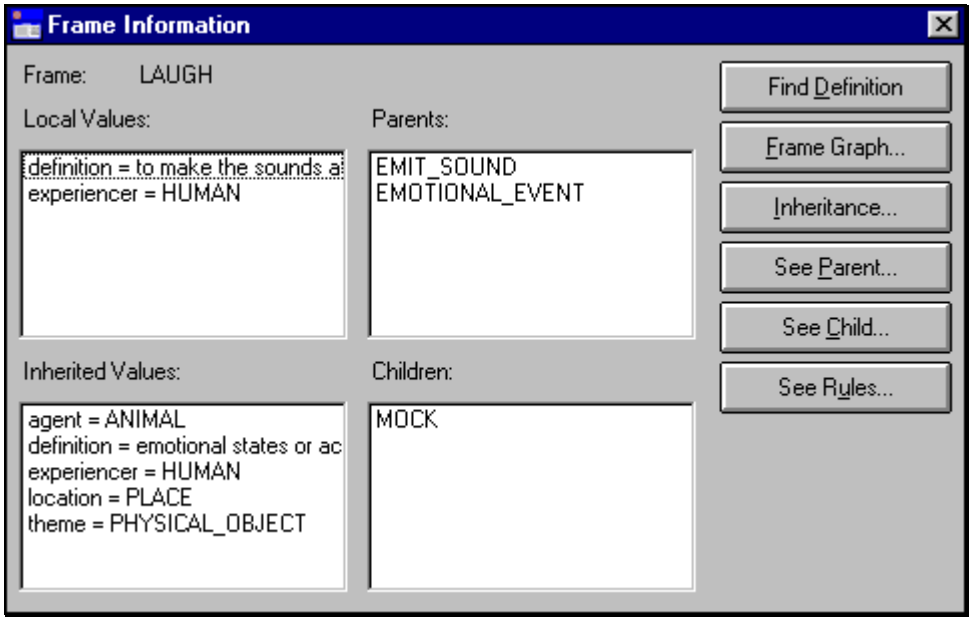


Ilustración 89. Estructura del marco LAUGH.

El marco también recoge de forma gráfica la estructura del léxico, como en este otro ejemplo, en el que elementos concretos como "lion" o "tiger" se adscriben al marco BIG_FELINE, que a su vez depende de otro más general, FELINE y, este, por su parte, de MAMMAL (*cf.* Moreno Ortiz 2000):

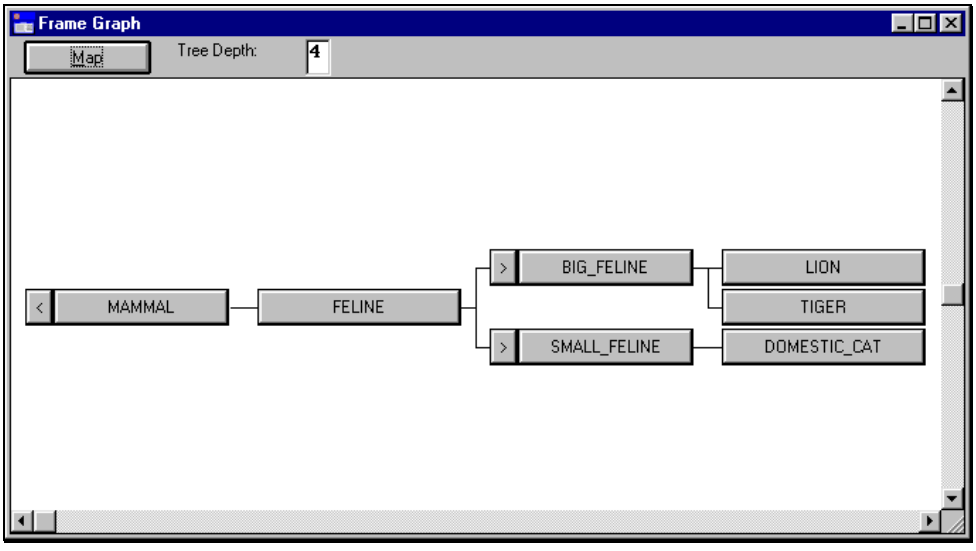


Ilustración 90. Representación gráfica de la estructura del léxico a través de marcos.

Pues bien, *FrameNet* se apoya en estas estructuras de representación del conocimiento (marcos) y en corpus para:

1) Proporcionar definiciones de cada elemento (predicado) de su base léxica (verbo, adjetivo o nombre)[179].

2) Proveer múltiples ejemplos (hasta veinte) de cada unidad léxica para mostrar todas sus posibilidades combinatorias, al tiempo que introduce anotaciones semánticas en los textos del corpus.

3) Trabajar con ejemplos que proceden de grandes muestras textuales de uso de la lengua (corpus), es decir, no son inventados.

4) Organizar el léxico en torno a marcos, no en lemas o palabras.

5) Vincular cada unidad léxica a un marco semántico que la conecta con otras unidades de forma jerárquica, como ocurre en las ontologías o en WordNet (*vid. infra*).

Así, cada unidad léxica consiste en una entrada del diccionario o lexicón que está asociada a un significado, de tal manera que en los casos de polisemia cada sentido o acepción se adscribe a un marco diferente.

Por ejemplo, si buscamos la entrada léxica *leader* en la base de datos, obtenemos información sobre su categoría (nombre), el marco en el que se inscribe (*leadership*) y su definición:

[179] Como se aprecia, además del verbo se consideran otros predicados, como nombres y adjetivos, que se describen en los mismos términos que los verbos: acepción, adscripción a marco, realización gramatical y anotación semántica.

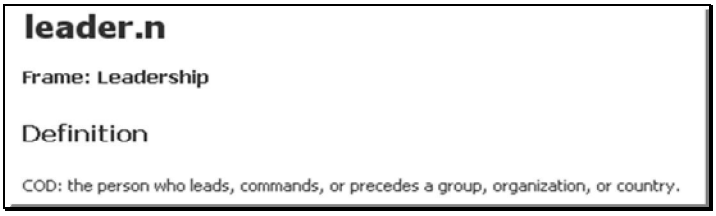


Ilustración 91. Entrada léxica para "leader" en FrameNet.

Aunque para *leader* existe otra entrada, correspondiente a otra acepción y que remite a un marco diferente (*First_rank*):

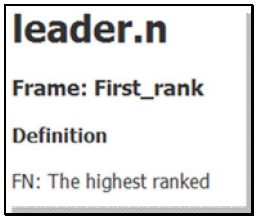


Ilustración 92. Otra entrada léxica para "leader" en FrameNet.

Además, en la descripción de cada entrada léxica se explicitan los elementos que conforman el marco concreto (*jurisdiction*, que marca el ámbito sobre el que alguien ejerce el control; y *leader*, la persona que ejerce el control), así como su realización sintáctica. Véase la muestra para *leader*, dentro del marco *leadership*:

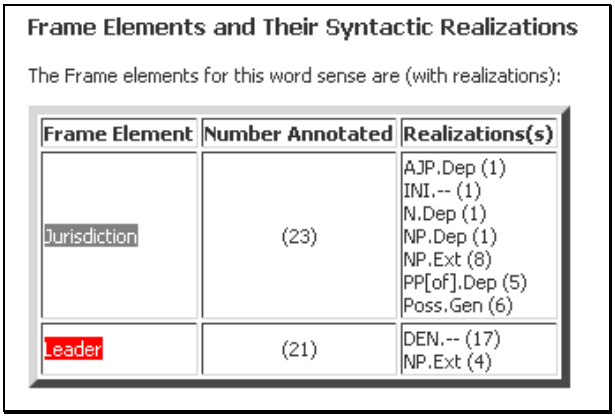


Ilustración 93. Elementos del marco LEADERSHIP y su realización sintáctica para LEADER[180].

[180] Claves: AJP (*Standard Adjective Phrase*), INI (*Indefinite Null Instantiation*), N (*Non-maximal nominal*), NP (*Standard Noun Phrase*), PP (*Prepositional Phrase*), Poss (*Possessive*

Además, FrameNet nos permite ver ejemplos de textos, anotados semánticamente, de cada una de las realizaciones de los elementos léxicos. Así, podemos encontrar casos en que *leader* aparece en estructuras del tipo *leader of* + [*jurisdiction*]:

> *In Italy, the LEADER of the Socialists called this week for the dissolution, in effect, of his own party into a new left-wing alliance with the ex-communists.*

Ilustración 94. Texto anotado semánticamente en FrameNet: "leader of"[181].

O en las que interviene en una estructura posesiva (genitivo sajón, como en *Iran's new leader*, o determinantes posesivos, como en *their leader*):

> *Last month, Iran 's new LEADER, President Hashemi Rafsanjani, offered to work for the release of the American hostages held by Islamic extremists in Lebanon along with 10 other foreigners.*
>
> *Bothwell's men had to carry their LEADER back to Hermitage, where he was greeted by the mortifying news that the prisoners had overpowered their guards and taken charge of the castle.*

Ilustración 95. Texto anotado semánticamente en FrameNet: posesivos.

Aparte de la clave de colores, la información semántica se puede hacer explícita[182]:

Noun Phrase); Ext (*External argument*), Dep (*Dependent*), Gen (*Genitive determiner of noun*), etc.

[181] Cada elemento del marco tiene una clave de color para facilitar su identificación en una frase; y se distinguen elementos nucleares o que tienen una vinculación más directa con el predicado (*core*) de otros que son accesorios o más externos (*peripheral*). En el caso de *leader*, los dos elementos del marco se consideran necesarios.

01. : [Jurisdiction/union] + [Leader/leaders]

1. 409614: On Tuesday [<Jurisdiction>union] [<Leader>*leaders*Tgt] claimed that every French customs post along the Spanish border was closing to all traffic except private cars. Translation

Ilustración 96. Texto anotado semánticamente en FrameNet: información explícita.

Otra opción más que ofrece este proyecto es mostrar gráficamente las relaciones de diverso tipo de un marco concreto con otros marcos[183]:

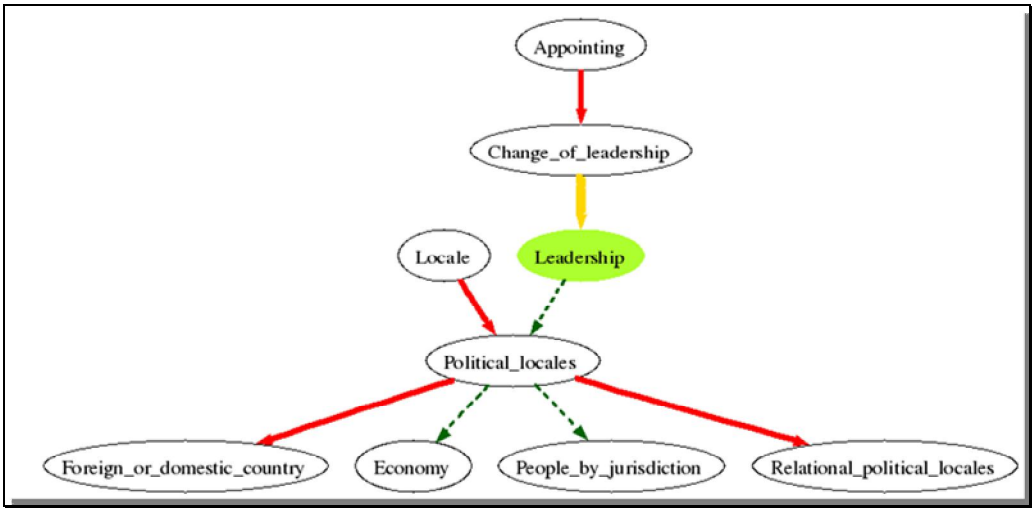


Ilustración 97. Gráfico de las relaciones del marco "Leadership" con otros marcos en FrameNet.

[182] *Union leaders* es la frase nominal en la que se ha materializado el marco *LEADERSHIP* en este texto: *union* se refiere a la jurisdicción o ámbito sobre el que alguien (*leader*) ejerce el liderazgo.

[183] Realizado mediante la aplicación *FrameGrapher*. URL: http://framenet.icsi.berkeley.edu/FrameGrapher/. La flecha roja indica "herencia", la verde que un marco es usado por otro, la amarilla que existe una relación causativa, etc.

Por último, resulta interesante la posibilidad que ofrece de mostrar todos los elementos léxicos adscritos a un tipo semántico determinado. Siguiendo con en el marco *leadership*, que se define como,

Leadership

Definition:

These are words referring to control by a Leader over a particular entity (the Governed) or an Activity. The frame contains both nouns referring to a title or position (e.g. *director, king, president*), and verbs describing the action of leadership (e.g. *rule, reign*). With verbs, it is possible to mention the Role played by the Leader (often a name of a leading position, e.g., *king*)

Sebek em hat was a LEADER of Priests. ca. 1780 BC

In 1789 Fletcher Christian LED the mutiny on HMS Bounty

Louis XIV RULED over his people as king for the longest period of any European monarch.

Ilustración 98. Definición del marco "Leadership" en FrameNet.

las unidades que pueden aparecer en la estructura del marco son nombres (referidos a título o posición: *leader*) y verbos (que describen la acción de liderazgo: *led, ruled*), como:

administer.v, administration.n, authority.n, baron.n, bishop.n, boss.n, caliph.n, captain.n, chairman.n, chairperson.n, charge.n, chief.n, chieftain.n, command.n, command.v, commandant.n, commander.n, congressman.n, crown prince.n, dictator.n, diplomat.n, director general.n, director.n, doyen.n, doyenne.n, drug lord.n, duchess.n, emperor.n, empress.n, executive.n, general.n, govern.v, government.n, governor.n, head.n, head.v, headmaster.n, high_priest.n, imam.n, kaiser.n, khan.n, khedive.n, king.n, lawmaker.n, lead.v, leader.n, leadership.n, legislator.n, legislature.n, maharaja.n, major general.a, mayor.n, minister.n, mogul.n, monarch.n, official.n, overlord.n, pasha.n, power_((govt)).n, power_((rule)).n, premier.n, preside.v, president.n, presidential.a, prime minister.n, prince.n, principal.n, queen.n, rector.n, regime.n, reign.v, representative.n, rule.n, rule.v, ruler.n, run.v, satrap.n, secretary.n, senate.n, shah.n, sheik.n, sovereign.n, spearhead.v, sultan.n, suzerain.n, tsar.n, tsarina.n, vice-captain.n, vice-chairman.n, vice-president.n, vice-principal.n, vice_president.n, viceroy.n, vizier.n

Ilustración 99. Unidades léxicas que se adscriben al marco "Leadership" en FrameNet.

Los principios del proyecto *FrameNet*, originariamente diseñado para el inglés, se han aplicado al alemán, al japonés y más recientemente al español, lenguas todas ellas que disponen en la actualidad de sus propios *FrameNets*[184].

Después de determinar el conocimiento léxico que se va a recoger, así como la forma de representarlo, la siguiente cuestión es cómo lo obtenemos. De hecho, según Mª. F. Verdejo (1995:45), el mayor problema del léxico lo representa su construcción, por lo que las técnicas para elaborar lexicones computacionales son fundamentales. Hoy en día, debido a la gran cantidad de información que necesitan incorporar los sistemas, se prefieren las opciones que realizan este proceso de forma automática, aunque con notables excepciones.

Los lexicones para dominios restringidos no presentan problemas, ya que su tamaño suele ser reducido. Pero cuando se pretende elaborar un diccionario no restringido temáticamente, hay que buscar técnicas y herramientas que faciliten la adquisición de información léxica. Desde los noventa se tiende a partir de recursos ya existentes (*cf.* Moreno Sandoval 1998:130):

-diccionarios en formato electrónico

-corpus textuales

-bases de datos terminológicas

-teorías y descripciones sobre el conocimiento del mundo (tesauros)

[184] Por ejemplo, en el proyecto español (*Spanish FrameNet*), el verbo *dirigir* aparece relacionado con los siguientes nombres, que representan el ámbito sobre el que se puede ejercer la dirección: *película*, *orquesta*, *obra_de_teatro*, *ópera*, *compañía*, *departamento*, *operación*, *proyecto*, *equipo*, *partido*, *grupo*, *trabajo*, *investigación*, *centro*, *organización*, etc.

Una opción es aprovechar la información que contienen diccionarios y tesauros en soporte electrónico utilizando técnicas de extracción de información, lo que ha demostrado ser un acercamiento viable, ya que los diccionarios, a través de sus marcas y estructura interna pueden aportar datos muy valiosos para la elaboración de lexicones computacionales. Destacan los trabajos llevados a cabo sobre diccionarios enfocados a la enseñanza de segundas lenguas, como el *Longman Dictionary of Contemporary English* (LDOCE), el *Oxford Advanced Learners Dictionary* (OALD), el *Webster* y el *Collins-Cobuild Dictionary*, así como los proyectos Acquilex-I y II (*vid.* Martí y Castellón 2000:105 y ss.).

Otra posibilidad es utilizar corpus textuales, es decir, recopilaciones de textos o fragmentos de textos en formato electrónico, sobre todo para suplir información que no suelen contener los diccionarios, como la relativa a patrones sintácticos, restricciones de selección, colocaciones, etc. El caso más conocido es el del diccionario Collins-Cobuild, que dispone de un corpus de 350 millones de palabras para la elaboración de sus diccionarios. Otras opciones pasan por acudir a bases de datos terminológicas multilingües o tesauros.

Por último, es posible construir lexicones computacionales de forma manual, codificando la información que los investigadores crean necesaria. El ejemplo más destacado es *WordNet*[185], una base de datos de relaciones léxicas para el inglés, desarrollada en el Laboratorio de Ciencia Cognitiva de la Universidad de Princeton bajo la dirección de G. A. Miller. Se basa en teorías psicolingüísticas sobre la organización del léxico en la memoria humana.

[185] URL: http://wordnet.princeton.edu/

En su versión 3.0, *WordNet* contiene unas 155000 entradas entre verbos, nombres, adjetivos y adverbios, organizadas en torno al concepto de *synset* (o conjunto de sinónimos) y las relaciones semánticas entre *synsets* (sinonimia, antonimia, hiponimia, meronimia, etc.).

A continuación se muestran los resultados de buscar *arm*[186]: al lado de cada *synset* (S), además de la categoría gramatical (nombre o verbo), se listan los elementos léxicos o sinónimos que se ajustan a la glosa (*branch*, *subdivisión*, *arm*, por ejemplo, se ajustan a la definición "division of some larger or more complex organization").

Noun

S: (n) **arm** (a human limb; technically the part of the superior limb between the shoulder and the elbow but commonly used to refer to the whole superior limb)
S: (n) **arm**, branch, limb (any projection that is thought to resemble a human arm) *"the arm of the record player"; "an arm of the sea"; "a branch of the sewer"*
S: (n) weapon, **arm**, weapon system (any instrument or instrumentality used in fighting or hunting) *"he was licensed to carry a weapon"*
S: (n) **arm** (the part of an armchair or sofa that supports the elbow and forearm of a seated person)
S: (n) branch, subdivision, **arm** (a division of some larger or more complex organization) *"a branch of Congress"; "botany is a branch of biology"; "the Germanic branch of Indo-European languages"*
S: (n) sleeve, **arm** (the part of a garment that is attached at the armhole and that provides a cloth covering for the arm)

Verb

S: (v) **arm**, build up, fortify, gird (prepare oneself for a military confrontation) *"The U.S. is girding for a conflict in the Middle East"; "troops are building up on the Iraqi border"*
S: (v) **arm** (supply with arms) *"The U.S. armed the freedom fighters in Afghanistan"*

Ilustración 100. "Synsets" de "arm" en WordNet.

[186] URL: http://www.cogsci.princeton.edu/cgi-bin/webwno

Al seleccionar un *synset* concreto, se despliegan las opciones de relaciones semánticas: sus hipónimos (*post office*, *executive branch*, *legislative branch*…) o sus hiperónimos directos (*division*), y todas las relaciones heredadas (*administrative unit*, *unit*, *organization*, *social group*…):

- S: (n) branch, subdivision, **arm** (a division of some larger or more complex organization) *"a branch of Congress"; "botany is a branch of biology"; "the Germanic branch of Indo-European languages"*

  - ***direct hyponym*** / *full hyponym*
    - S: (n) post office, local post office ((5))
    - S: (n) executive branch, Executive Office of the President (the branch of the United States government that is responsible for carrying out the laws)
    - S: (n) legislative branch (the branch of the United States government that has the power of legislating)
    - S: (n) judicial branch (the branch of the United States government responsible for the administration of justice)

Ilustración 101. Relaciones semánticas (hipónimos) de uno de los "synsets" de "arm" en WordNet.

o *direct hypernym* / ***inherited hypernym*** / *sister term*

- S: (n) division (an administrative unit in government or business)
  - S: (n) administrative unit, administrative body (a unit with administrative responsibilities)
    - S: (n) unit, social unit (an organization regarded as part of a larger social group) *"the coach said the offensive unit did a good job"; "after the battle the soldier had trouble rejoining his unit"*
      - S: (n) organization, organisation (a group of people who work together)
        - S: (n) social group (people sharing some social relation)
          - S: (n) group, grouping (any number of entities (members) considered as a unit)
            - S: (n) abstraction, abstract entity (a general concept formed by extracting common features from specific examples)
              - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Ilustración 102. Relaciones semánticas (hiperónimos heredados) de uno de los "synsets" de "arm" en WordNet.

Algunos ejemplos del tipo de relaciones semánticas que contempla WordNet son (tomados de Wordnet TreeWalk[187]):

1) Relaciones de *hiperonimia*, es decir, las que se establecen entre un término genérico o hiperónimo y la clase de elementos que designa. P. ej. *árbol* (término genérico) es el hiperónimo de *roble* (término específico), o *caminar* es el hiperónimo de *deambular*.

hypernym			
tree		oak	an oak is a kind of tree
plant		tree	a tree is a kind of plant
dwelling		house	a house is a kind of dwelling
disatisfaction		disappointment	disappointment is a kind of disatisfaction
walk		amble	to amble is to walk in some manner
walk		march	to march is to walk in some manner
say		lisp	to lisp is to say in some manner
hit		slam	to slam is to hit in some manner

Ilustración 103. Ejemplo de relaciones de hiperonimia en WordNet.

2) Relaciones de *hiponimia*, las que se establecen entre un término concreto (hipónimo) y el término genérico que designa la clase a la que pertenece. Y de *troponimia*, entre un verbo y otro verbo que introduce matices en el anterior. P. ej. *roble* (término específico) es hipónimo de *árbol* (término genérico), o *deambular* de *andar*.

		hyponym	
tree		oak	an oak is a kind of tree
plant		tree	a tree is a kind of plant
dwelling		house	a house is a kind of dwelling
disatisfaction		disappointment	disappointment is a kind of disatisfaction

Ilustración 104. Ejemplo de relaciones de hiponimia en WordNet.

[187] URL: http://wntw.sourceforge.net/. Se trata de una interfaz para WordNet desarrollada por Bernard Bou en Francia.

3) Relaciones de *meronimia*, las que se establecen entre un término que designa una parte, una sustancia o un miembro de otro término que designa el todo al que pertenece el primero. P. ej. la relación entre *ala* (parte) y *pájaro* (todo), *oxígeno* (sustancia) y *aire* (todo), *árbol* (miembro) y *bosque* (todo), etc.

meronym			
wing		bird	component-object
mouth		face	component-object
branch		tree	component-object
tree		forest	member-collection
oxygen		air	substance-object
adolescence		growing up	phase-process
paying		shopping	feature activity

Ilustración 105. Ejemplo de relaciones de meronimia en WordNet.

4) Relaciones de *holonimia*, es decir, las que se establecen entre el todo (*pájaro*, *bosque, aire*) y la parte (*ala*), sustancia (*oxígeno*) o miembro (*árbol*):

holonym			
bird		wing	object-component
face		mouth	object-component
tree		branch	object-component

Ilustración 106. Ejemplo de relaciones de holonimia en WordNet.

Otro tipo de relaciones de significado contempladas por WordNet son las de antonimia (*feliz-infeliz*), las de causalidad (*dar-tener*), la de implicación (*roncar-dormir*, *comprar-pagar*), etc.

De toda la gama de relaciones semánticas que comprende WordNet, se han desarrollado implementaciones gráficas y herramientas para su explotación y extensión a otras lenguas[188].

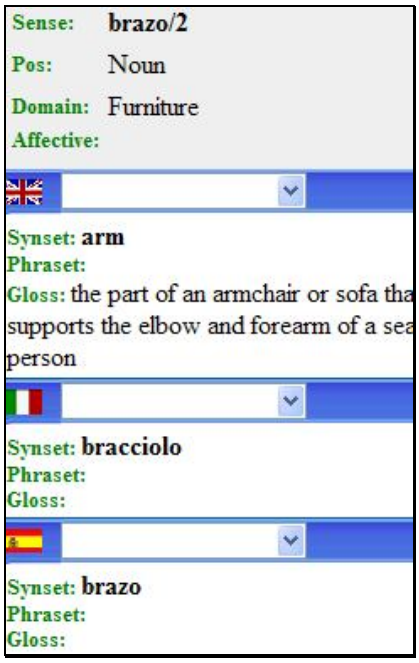


Ilustración 107. Ejemplo de equivalencias léxicas en MultiWordNet.

Destaca especialmente *EuroWordNet*[189], base de datos multilingüe que crea redes conceptuales para diversas lenguas europeas, entre ellas el español y el catalán, proyecto finalizado en 1999. Con posterioridad, el *Centre de Llenguatge i Computació* (CLiC)[190] de la Universidad de Barcelona y el *grupo de Llenguatge Natural UPC-TALP* de la Universidad Politécnica de Cataluña han continuado trabajando en el *WordNet*

[188] *Vid. WordNet 3.0 Vocabulary Helper* (http://poets.notredame.ac.jp/cgi-bin/wn), *The Global WordNet Association* (http://www.globalwordnet.org/), que tiene por objetivo coordinar el desarrollo de *wordnets* para todas las lenguas del mundo; *MultiWordNet* (http://multiwordnet.itc.it/english/home.php), base de datos multilingüe inglés, italiano, español, hebreo, rumano y latín; MEANING, *Multilingual Central Repository* (http://www.lsi.upc.es/~nlp/meaning/meaning.html y también en http://garraf.epsevg.upc.es/cgi-bin/wei4/public/wei.consult.perl).

[189] URL: http://www.illc.uva.nl/EuroWordNet/

[190] URL: http://clic.ub.edu/es/eurowordnet-es

español, y el *Centre de Referència d'Enginyeria Lingüística* (CREL) en el *WordNet* catalán, manteniendo siempre la misma estructura que *WordNet*.

Las entradas (forma ortográfica) muestran los *synsets* asociados con ellas, es decir, el conjunto de sinónimos, definición o glosa, así como ejemplos de uso. En el caso de buscar los sinónimos de *arma* en *EuroWordNet*, nos encontramos con múltiples sentidos, por lo que para determinar las relaciones semánticas, dada la ambigüedad, es preciso elegir uno de ellos:

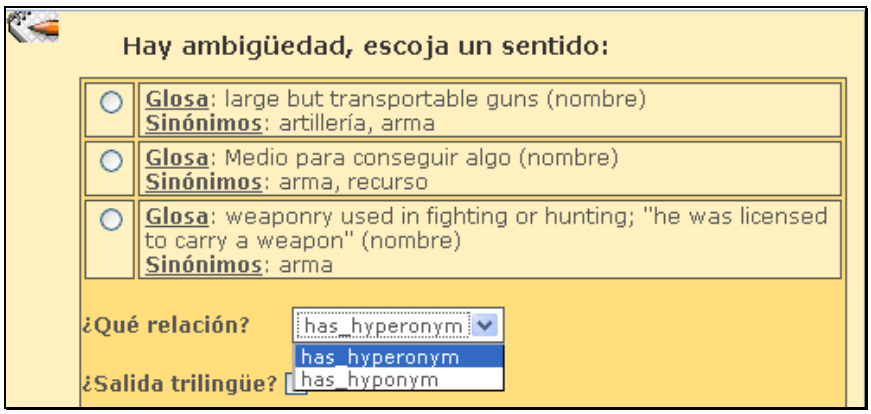


Ilustración 108. Sinónimos de "arma" en EuroWordNet.

Optamos por las relaciones de hiperonimia del segundo sentido, con el siguiente resultado:

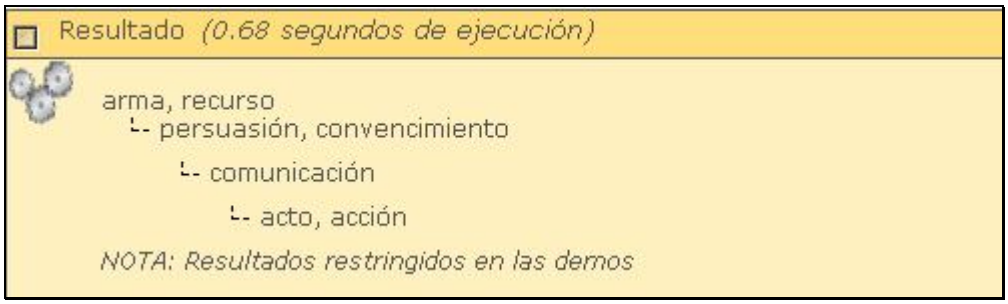


Ilustración 109. Hiperónimos de "arma" en uno de sus "synsets" en EuroWordNet.

2.4.3. El factor discursivo

El proceso de interpretación semántica no se detiene en el nivel léxico ni en el oracional, sino que es preciso tener en cuenta las relaciones que se establecen entre una oración y las que la preceden o siguen, ya que de ellas depende en muchos casos la interpretación final.

Del mismo modo que el nivel supraoracional ha ido cobrando importancia en Lingüística y en Pragmática (Lingüística del texto, Análisis del discurso), también se ha visto reflejado en las preocupaciones de la LC, donde según la aplicación pronto se hizo evidente que para comprender un texto o un diálogo había que atender a algo más que el significado de las palabras aisladas o el significado de las relaciones entre las palabras, es decir, no basta la "competencia lingüística". Es preciso recurrir a otro tipo de factores que intervienen en la determinación del sentido: información sobre la situación o contexto de enunciación, las creencias compartidas por los hablantes, los deseos, las intenciones, las relaciones que se establecen entre las oraciones dentro de un texto. Como dicen T. Moure y J. Llisterri (1996:191):

> La comprensión del texto supera la simple decodificación del contenido literal del mismo para abarcar también la reconstrucción del mensaje en el momento en que fue emitido, las connotaciones que arrastra y lo no-dicho pero implícito en el enunciado. Comprender un texto, al fin, es interpretarlo a la luz de la situación y las condiciones que lo produjeron.

Los sistemas computacionales que pretendan comprender un texto deben enfrentarse a dos problemas (Moure y Llisterri 1996:191):

a) Atender a las conexiones que se establecen entre una oración y las que le preceden o siguen, sobre la base de una gramática del texto.
b) Atender a la identificación de los interlocutores, el momento en que se emite el discurso y las condiciones de la enunciación. Aquí tienen cabida la anáfora y la catáfora. La identificación de antecedentes y consecuentes ha sido uno de los terrenos que más atención han recibido desde la perspectiva computacional, por lo general estableciendo una lista de posibles candidatos a los que se les aplican restricciones de tipo morfológico, semántico o pragmático. A. Moreno Sandoval (1998:124-125) describe cómo partiendo de la forma lógica se va elaborando una lista de *entidades de discurso*, que recoge los posibles candidatos a referentes de las expresiones anafóricas.

Por lo general, los esfuerzos para dar cuenta de la estructura del discurso se han centrado en dos tipos de textos: los narrativos y los dialogados, en los que se ha trabajado sobre los siguientes aspectos (Verdejo 1995:51):

i) Unidades: cuáles son, cómo caracterizarlas, reconocerlas y representarlas.
ii) Estructura del discurso y modelos computacionales para describirla.
iii) Mecanismos para la resolución de referencias.
iv) Papel del conocimiento del mundo.
v) Caracterización del diálogo cooperativo y papel de las intenciones de los interlocutores.

J. Vidal y J. Busquets (1996:427) sintetizan dichos aspectos en tres líneas de investigación, que en su opinión son las que guían el procesamiento computacional del discurso:

1. *Coherencia*, es decir, los mecanismos que dan cuenta de las relaciones *internas* que dan sentido al discurso. Conceptos teóricos como las *implicaturas*, la *relevancia*, las *expectativas* o las *máximas de la conversación* se han aplicado en LC, sobre todo para la generación de oraciones.
2. *Contexto*, es decir, de qué modo el *contexto* afecta al uso y comprensión de los enunciados. Es lo que algunos investigadores pretenden recoger mediante los *planes*.
3. *Estructura del discurso*, es decir, qué unidades básicas lo constituyen y qué relaciones se establecen entre ellas, así como la incidencia que la estructura informativa del discurso (tema, rema, tópicos, focos, etc.) y el contexto tienen en la interpretación semántica de diversos fenómenos lingüísticos, como la referencia de los pronombres, de las descripciones definidas, la elipsis, la cuantificación, etc.

En este ámbito, el trabajo que queda todavía por realizar es mucho, pero, como dice A. Ramsay (2003:132), "nonetheless, computational systems which aim to produce coherent extended discourse, or to take part in extended conversations, will have to pay attention to these issues".

2.5. Las áreas de aplicación de la LC

La vertiente aplicada de la LC tiene como objetivo desarrollar programas y/o sistemas encaminados a dar soluciones a problemas concretos relacionados con las lenguas y la tecnología que se plantean en la sociedad actual. Los conocimientos teóricos obtenidos del tratamiento computacional del lenguaje en sus diferentes niveles sirven de guía a la hora desarrollar aplicaciones concretas que, a su vez, son un estímulo para el desarrollo de nuevas investigaciones.

Las aplicaciones más ambiciosas de la *Lingüística Computacional*, en el sentido de que tratan de reproducir la capacidad humana de procesar el lenguaje, son:

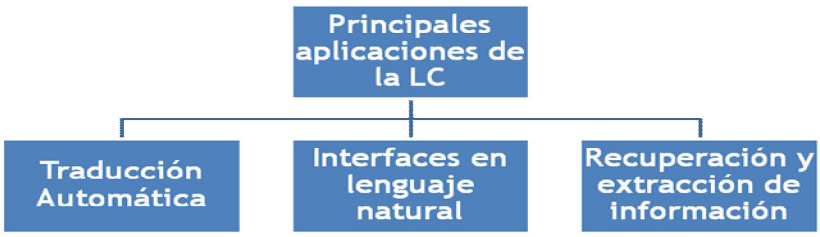


Ilustración 110. Principales aplicaciones de la LC.

- *Traducción automática*: su objetivo es lograr traducir de manera automática un texto, sea oral o escrito, de una lengua fuente a una o varias lenguas meta.
- *Interacción en lenguaje natural (interfaces y sistemas de diálogo)*: su meta es facilitar la comunicación entre personas y ordenadores mediante el uso de una lengua natural (en su modalidad oral o escrita) y no a través de un lenguaje artificial. Normalmente, ante las preguntas de un usuario, el sistema responde en la misma lengua natural en que ha sido formulada la pregunta. La comunicación suele producirse en

el contexto de un dominio restringido –lo más habitual–, aunque también puede darse en uno que no presente restricciones.

- *Recuperación y extracción de información*:
  - *Recuperación*: a partir de la consulta de un usuario a un banco de datos textuales, el sistema se encarga de proporcionar los materiales que se ajustan a los criterios de búsqueda, no solo basándose en la detección de palabras clave sino también llevando a cabo una labor de comprensión lingüística de la consulta. En este ámbito de trabajo, ha sido habitual, hasta época reciente, el predominio de técnicas estadísticas frente a las lingüísticas o simbólicas. Además, también ha pasado de ser una aplicación reducida a determinados terrenos (Derecho, Medicina, Economía) a ampliar su ámbito de trabajo, como consecuencia sobre todo de la generalización del uso de Internet y de la globalización, que han puesto de manifiesto nuevas necesidades en lo que a las demandas de información se refiere.
  - *Extracción*: en este caso también se trata de acceder a información, pero con la finalidad de organizarla de acuerdo con algún tipo de formato preestablecido para su posterior tratamiento o recuperación. Se trata de buscar en un texto cualquiera determinados tipos de contenido. Es habitual distinguir entre la búsqueda de nombres propios y la búsqueda de eventos. P. ej. tareas propias de la extracción de información suelen ser la localización de todos los nombres de empresas que aparezcan en una serie de textos, o todos los asesinatos mencionados para posteriormente clasificar la información según los parámetros previamente fijados (p. ej. nombre de la empresa, nacionalidad, sector, etc.; asesino, víctima, lugar, tiempo, arma, etc.).

2.5.1. Aplicaciones basadas en el tratamiento de información textual

A veces, atendiendo a la modalidad oral o escrita de la lengua, es frecuente distinguir el grupo de aplicaciones que se centra específicamente en el tratamiento de la vertiente escrita. Es la línea de trabajo que se conoce como *Procesamiento del Lenguaje Natural* o *Tecnologías del Texto*.

Además de las aplicaciones anteriores (la traducción automática, la recuperación y extracción de información y la interacción en lenguaje natural), destacan las siguientes herramientas para el tratamiento de la lengua escrita:

- *Herramientas de ayuda a la escritura*, integradas por lo general en los procesadores de texto. Incluyen:
  - *Correctores ortográficos*: programas que revisan la ortografía de un escrito y la comparan con el conocimiento lingüístico previamente almacenado.
  - *Correctores sintácticos y de estilo*: estos programas revisan la sintaxis y el estilo de un escrito, labor para la que se requiere una mayor cantidad de conocimientos lingüísticos, más difíciles de sistematizar, por otra parte, que los relacionados con la ortografía.
- *Creación automática de resúmenes* de uno o varios textos a partir de la información más relevante que contienen estos (títulos, negritas, cursivas, apartados…).
- *Extracción de terminología* de documentos científicos o técnicos. Trabajo fundamental para el análisis, comprensión, generación o traducción de documentos.

- *Indexación automática*: íntimamente relacionada con la tarea anterior, consiste en clasificar un documento dentro de un dominio de acuerdo con las palabras (términos) que aparecen en él y que, además, son útiles para la posterior recuperación de la información.
- *"Data mining" textual* o descubrimiento de datos en textos. Consiste en analizar y descubrir patrones y tendencias en grandes conjuntos de datos textuales, generalmente con el objetivo de tomar decisiones. P. ej. en una investigación sobre la migraña, mediante esta aplicación, se extrajeron y analizaron datos de artículos que permitieron concluir a los investigadores que la migraña aparecía asociada con el estrés y con deficiencias de magnesio, lo que en ese momento era información desconocida.

2.5.2. Tecnologías del habla

Es la línea de trabajo en Lingüística Computacional que se centra específicamente en el tratamiento de la lengua oral. Comprende:

- *Síntesis del habla*: o generación de habla artificial, sobre todo, conversión de textos escritos en su equivalente oral.
- *Reconocimiento del habla*: de forma inversa a la síntesis, transforma un enunciado oral en su contrapartida escrita.
- *Sistemas de diálogo*: combinan las dos tecnologías anteriores para facilitar la interacción oral entre personas y sistemas informáticos.

Entre las aplicaciones de estas tecnologías destacan:

- el dictado automático
- la traducción automática del habla

- la recuperación de información a partir de documentos sonoros
- la identificación y verificación automáticas de la identidad del locutor
- la identificación automática de la lengua en contextos y/o servicios multilingües
- los servicios automáticos de atención telefónica
- los sistemas conversacionales o de diálogo oral entre personas y máquinas
- la atención a personas con discapacidades o con necesidades especiales
- la enseñanza de lenguas asistida por ordenador

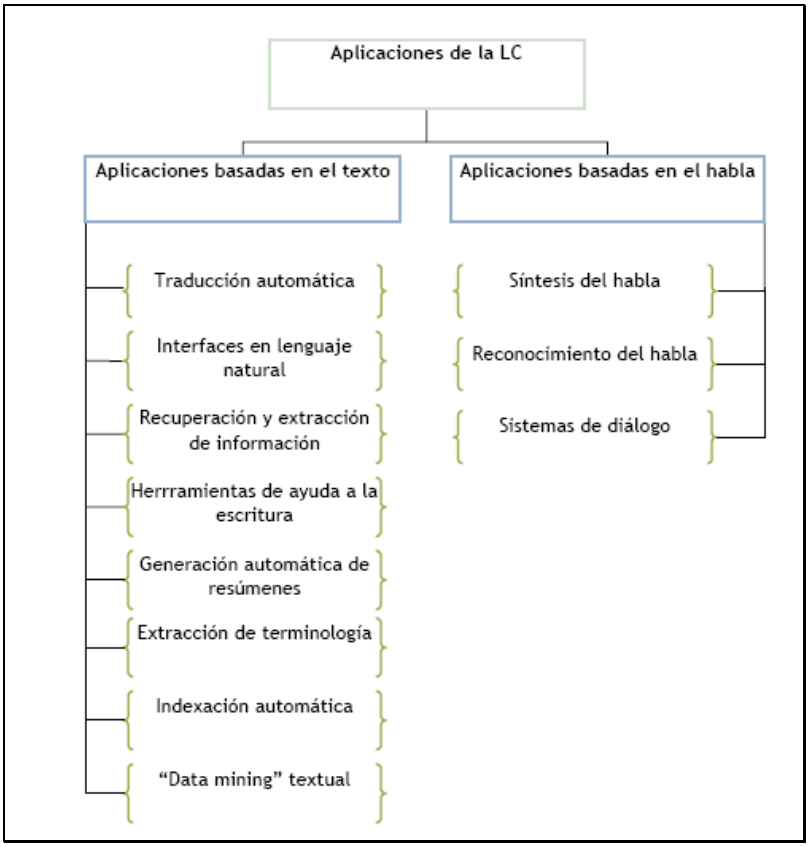


Ilustración 111. Aplicaciones de la LC basadas en el texto y en el habla.

2.5.3. Aplicaciones basadas en el diálogo

Este grupo de aplicaciones reúne aquellas en que existe un intercambio comunicativo entre un usuario y un sistema informático, ya sea de forma oral o escrita. Incluye:

- *Sistemas de acceso a bases de datos* o de pregunta / respuesta en los que se utiliza una lengua natural para interrogar a una base de datos.
- *Sistemas de acceso a otros dominios* (sistemas expertos, sistemas operativos, etc.).
- *Sistemas de diálogo inteligente.*
- *Servicios automáticos a través del teléfono*.
- *Sistemas de enseñanza*, en los que el ordenador interactúa con el estudiante.
- *Control de máquinas* a través de la lengua hablada.
- *Sistemas generales* para la resolución de problemas de forma cooperativa.

Básicamente se trata en todos los casos de formular preguntas o dar instrucciones en lenguaje natural a un sistema que contiene información de algún tipo. El sistema traduce el lenguaje natural a un lenguaje formal y responde al usuario, bien mediante el uso del lenguaje bien realizando la acción que se le pide. Mientras más restringido es el dominio, más limitado es el lenguaje que se puede utilizar y, por lo tanto, menores los problemas lingüísticos. Pero a medida que los sistemas tratan de interactuar con el usuario de forma inteligente, es decir, utilizando las mismas estrategias conversacionales que usamos las personas, las dificultades aumentan al tener que dar cuenta de

aspectos hasta el momento poco formalizados o de difícil formalización, como son todos los relacionados con el ámbito de la pragmática (deseos, creencias, intenciones o conocimiento del mundo en general).

2.5.4. Otras aplicaciones

- *Herramientas informáticas útiles para el lingüista o el filólogo* en diversas tareas relacionadas con el estudio del lenguaje (lingüística de corpus, lingüística estadística, estilometría, lingüística histórica computacional, informática aplicada a la sociolingüística, lexicografía asistida por ordenador, etc…):
  - *Herramientas de análisis textual*: extracción y cómputo de frecuencias de aparición, concordancias, estadísticas. En general, se trata de tareas mecánicas y aburridas que los ordenadores realizan de forma más rápida y precisa que las personas.
  - *Herramientas para el manejo de corpus*: etiquetadores categoriales, desambiguadores, analizadores sintácticos, etc.
  - *Bases de datos lexicográficas y terminológicas*: de gran importancia para la elaboración y gestión de diccionarios.
- *Enseñanza de lenguas asistida por ordenador*. Destacan los programas para la enseñanza de lenguas extranjeras.
- *Aplicaciones multilingües*, sobre todo en conexión con el uso de Internet: identificación de la lengua, alineamiento de recursos terminológicos bilingües y multilingües, recuperación de información en diferentes lenguas a partir de una consulta formulada en una determinada lengua y ayudas para la comprensión.

- *Aplicaciones multimedia y multimodales* para la enseñanza, el entretenimiento, los negocios o el transporte: combinan el lenguaje con otros modos de comunicación (visual, táctil…).

2.5.5. Recursos lingüísticos

Se trata de recursos básicos para el tratamiento computacional de cualquier lengua y, por consiguiente, para el desarrollo de las tecnologías del habla y del texto. Comprenden:

- *Corpus* o conjuntos de muestras textuales que dan cuenta del uso real de una lengua.
- *Bases de datos* léxicos monolingües o multilingües.
- *Redes léxico-semánticas*.
- *Diccionarios en CD-ROM o en línea.*
- *Gramáticas computacionales.*

Ilustración 112. Principales recursos lingüísticos.

3. LOS CORPUS

3. LOS CORPUS

En el apartado primero ya se ha comentado el papel fundamental que se otorga a los corpus en la actualidad como recurso básico para el estudio y descripción de las lenguas. En particular, en el ámbito de la Lingüística Computacional, se han constituido en punto de partida imprescindible para la elaboración de léxicos y gramáticas, y representan una línea de investigación transversal en lo que al tratamiento del lenguaje con medios informáticos se refiere, al ser indispensables para el desarrollo de aplicaciones basadas tanto en el texto como en el habla (*cf.* Moure y Llisterri 1996).

Por otra parte, hay que entender e inscribir el empleo de corpus en Lingüística dentro de una perspectiva metodológica general que adopta el empirismo como forma de concebir el estudio de la lengua. En este sentido, el empleo de datos reales, de muestras de uso lingüístico, resulta el complemento ideal y la referencia ineludible en cualquier investigación que aspire a dar cuenta de algún aspecto relacionado con el lenguaje: los datos son los que apoyan o contradicen una postura teórica, los que permiten inferir reglas y generalizaciones, los que proporcionan informaciones cuantitativas, etc. Y también constituyen el material necesario como punto de partida para el desarrollo de una aplicación práctica.

Es, precisamente, el conjunto de datos –enunciados lingüísticos de algún tipo– lo que se denomina "corpus" en un sentido general del término[191]. Pero ha sido el empleo de ordenadores para reunir,

[191] Observemos la definición que proporciona el *DRAE* (REAL ACADEMIA ESPAÑOLA 2001): "Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación". En el *Panhispánico* (REAL ACADEMIA ESPAÑOLA 2005), se insiste en la misma idea, con el añadido de una nota de uso: "Conjunto de datos o textos de un mismo tipo que sirve de base a una

organizar y procesar esos datos el que ha conferido modernidad a esta tarea, hasta el punto de propiciar el despegue de toda una forma de hacer lingüística, la llamada "lingüística de corpus". Aunque el término en sí mismo es de acuñación relativamente reciente, los trabajos basados en corpus siempre han contado con seguidores. De hecho, esta manera de acercarse al estudio de la lengua está avalada por una amplia tradición de estudios, que desde mediados del siglo XX se han visto potenciados por la posibilidad de emplear ordenadores, lo que ha contribuido al despegue de la lingüística de corpus en un sentido moderno del término "corpus".

El camino que ha tenido que recorrer la lingüística de corpus hasta alcanzar el entusiasmo que despierta hoy en día no ha estado exento de dificultades técnicas y de críticas teóricas que tenían razón de ser cuando fueron formuladas. Sin embargo, el desarrollo espectacular de la informática en un lapso relativamente breve de tiempo y la multitud de estudios publicados que toman como referencia los datos proporcionados por los cada vez más numerosos corpus de todo tipo, avalan la importancia que ha adquirido esta línea de investigación en Lingüística Computacional.

Son T. McEnery y A. Wilson (1996 y 2001), T. McEnery (2003) y T. McEnery, R. Xiao y Y. Tono (2006) los autores que mejor resumen estos primeros pasos de la lingüística de corpus. A grandes rasgos, presentamos algunos de los hitos que han contribuido al proceso de desarrollo y consolidación de la lingüística de corpus[192] tal y como se entiende en la actualidad, con especial atención al concepto de "corpus" manejado en cada caso.

---

investigación'. Es invariable en plural […] No debe usarse, como ocurre por influjo del inglés, el plural latino ⊗*corpora*".

[192] Para más detalles, remitimos a MCENERY (1996:1-19) y MCENERY, XIAO y TONO (2006:1-12).

3.1. Hitos en la lingüística de corpus

3.1.1. Precedentes en el uso de corpus

Hasta el siglo XIX, G. Leech (1991) destaca la existencia de toda una tradición de trabajos lingüísticos basados en corpus, fragmentarios o limitados, para el estudio de lenguas muertas. Así, es posible establecer un primer concepto de corpus, previo al ordenador y caracterizado por:

Corpus-1

- Ser un conjunto de textos escritos (datos).
- Tener como finalidad el estudio y descripción de lenguas muertas (latín, griego…).
- Ser necesario, desde un punto de vista metodológico, para llevar a cabo los estudios lingüísticos: esos datos constituían el único acercamiento posible, pues esas lenguas ya no contaban con hablantes vivos que utilizaran la lengua como vehículo de comunicación en una sociedad.

Con el avance del siglo XIX y hasta mediados del XX (*cf.* McEnery 1996), se siguió empleando esta forma de trabajar, basada en la recopilación de una gran cantidad de datos escritos (corpus) para:

- Dar cuenta del proceso de adquisición del lenguaje infantil a través de la transcripción, en diarios que realizaban los padres, de las interacciones de los niños[193].

[193] Esta forma de trabajar está en la base de los estudios longitudinales que predominan en la actualidad en el campo de la adquisición de la lengua materna: evolución de las muestras de un número reducido de niños (en torno a tres), y

- Establecer convenciones ortográficas[194].
- Obtener listas de vocabulario para la enseñanza de segundas lenguas.
- Realizar estudios comparativos de lenguas.
- Elaborar gramáticas descriptivas[195].

3.1.2. Primera lingüística de corpus

Sin embargo, fueron los trabajos de antropólogos, etnógrafos y, sobre todo, de los lingüistas estructurales norteamericanos[196] –F. Boas, E. Sapir, L. Bloomfield, Ch. Fries…– los que, durante la primera mitad del siglo XX, contribuyeron a sentar las bases de la lingüística de corpus como metodología empírica basada en la observación de datos, aunque el término como tal ("lingüística de corpus") no aparecerá hasta más tarde, a principios de los ochenta.

Algunos de estos investigadores, anteriores a la irrupción de N. Chomsky en el panorama lingüístico, llegaron a considerar el corpus como la "única" herramienta válida para el estudio de las lenguas, ya

también en los estudios basados en una selección amplia de niños, más de moda durante la primera mitad del siglo XX.

194 Destaca, por ejemplo, la labor de Käding, que a finales del XIX recopiló un corpus de once millones de palabras para el alemán.

195 Es el caso de la gramática de Fries para el inglés (*The structure of English: An Introduction to the Construction of Sentences*, 1952), elaborada a partir de un corpus formado por transcripciones de conversaciones telefónicas. El hecho de trabajar con la lengua oral contrasta con lo que ocurre incluso en los corpus actuales, en los que predomina la lengua escrita.

196 El análisis de las lenguas nativas norteamericanas (y sus culturas) no podía basarse en la lengua escrita ni en la metodología de los estudios diacrónicos y comparativos predominante durante el siglo XIX y principios del XX. Se hace preciso un nuevo enfoque sincrónico eminentemente inductivo, que parte de los trabajos de campo para la recogida de muestras orales y su transcripción. Esta inclinación hacia el "habla" posibilita el despegue de los estudios fonético-fonológicos y morfológicos, pero por otra parte deja un poco de lado el recurso al significado y al "mentalismo" en los estudios lingüísticos (*cf.* KOERNER 2002).

que se pensaba que por sí mismo podía proporcionar los datos necesarios para una descripción exhaustiva de las mismas o, por lo menos, de los aspectos que precisaban estos investigadores. De hecho, todo dato que no fuese directamente observable tendía a ser rechazado por los lingüistas "postbloomfieldianos" (B. Bloch, Z. Harris, Ch. Hockett, G. Trager…), aunque no todos ellos defendían esta postura con el mismo radicalismo. Por otra parte, como la fuente más directa de observación era la lengua oral, el análisis de los datos (corpus) será ante todo fonético-fonológico.

Así pues, este nuevo concepto de corpus, el "corpus estructuralista", se caracteriza por:

Corpus-2

- Ser un conjunto de grabaciones y transcripciones en papel (datos), en la mayoría de los casos fichas almacenadas en cajas de zapatos.
- Tener por finalidad el estudio de lenguas vivas, pero no documentadas previamente por escrito (lenguas amerindias).
- Ser necesario, ya que la recogida de datos orales era la única forma de acceder al conocimiento de esas lenguas, nunca antes estudiadas.
- Centrarse en los aspectos fonéticos y (morfo)fonológicos, niveles en los que es posible realizar un inventario de todos los elementos implicados dada la naturaleza finita de sus paradigmas.
- No atender a cuestiones de representatividad: debido a que el análisis de los datos debía efectuarse de forma manual y visual era imposible manejar un número elevado de datos, de ahí una de las principales críticas que recibirá la metodología, por su parcialidad a la hora de describir la realidad.

3.1.3. Críticas a la primera lingüística de corpus

La irrupción de la figura de Chomsky (1957, 1965) en el panorama lingüístico a finales de los cincuenta va a suponer un cambio radical de enfoque en los estudios lingüísticos, ya que con él se impone el racionalismo como filosofía de fondo que aspira a guiar las investigaciones relacionadas con el lenguaje. El resultado será que el trabajo basado en corpus, una metodología ante todo empírica, va a ser objeto, durante los años sesenta y setenta, de duras críticas que provocarán que no haya una continuidad entre los trabajos en lingüística de corpus iniciados por los estructuralistas americanos y la lingüística de corpus actual.

Como consecuencia de estas críticas, se produjo un desprestigio general de la metodología basada en corpus (empirismo) a favor de una nueva ortodoxia en los estudios lingüísticos: un acercamiento basado en las intuiciones del lingüista (racionalismo). El resultado fue que la lingüística de corpus se convirtió en una tendencia marginal durante un par de décadas.

Empirismo	Racionalismo
Actuación	**Competencia**
Corpus	**Intuición**

Tabla 20. Empirismo vs. racionalismo.

No obstante, hay que destacar que en determinados campos seguía siendo imprescindible el uso de muestras reales de la lengua, por lo que durante estas décadas se siguieron elaborando corpus en ámbitos como:

- la fonética: requiere datos, no introspecciones;

- la adquisición de lenguas: los niños no han desarrollado una capacidad metalingüística, por lo que no pueden emitir juicios de valor sobre su lengua;
- la lingüística histórica: en muchos casos, no se podía recurrir a los hablantes nativos.

Las críticas vertidas sobre la lingüística corpus fueron:

-de índole teórica (Chomsky)

-de índole práctica (Abercrombie)

3.1.3.1. Críticas teóricas (Chomsky)

Las críticas de Chomsky a la lingüística de corpus parten de dos postulados fundamentales:

a) La apelación al recurso a la intuición, a la introspección del lingüista, como el único criterio válido para el estudio de la lengua.

b) El papel central otorgado a la sintaxis en las primeras versiones del modelo generativista[197].

Con estas premisas, los corpus no se consideran instrumentos válidos, ya que desde la perspectiva del modelo teórico propugnado por Chomsky:

1) Los corpus dan cuenta de la actuación, que es la evidencia externa de la lengua, sujeta a variaciones o desviaciones de la norma de diverso

[197] En contraposición a los trabajos que venían desarrollando los estructuralistas americanos sobre el plano fonético-fonológico.

tipo debidas a limitaciones de memoria, distracciones, errores, etc. Sin embargo, la labor del lingüista es reflejar la competencia: el conocimiento interiorizado de la lengua que posee un hablante-oyente ideal y que le permite discriminar las secuencias gramaticales de las agramaticales; la competencia es ajena a las circunstancias materiales o de otro tipo que puedan afectar a la comunicación.

2) La parcialidad de los corpus: son incompletos, ya que no contienen todas las oraciones de una lengua, y son sesgados, ya que la inclusión de un elemento lingüístico vendrá determinada por su frecuencia de uso, es decir, los elementos más habituales estarán mejor reflejados en un corpus que aquellos más raros. El corpus es por definición cerrado, finito, tiene unos límites; por lo tanto, no puede dar cuenta de la naturaleza no finita, ilimitada de las lenguas. Estas se caracterizan por su infinita capacidad generativa: con un inventario limitado de elementos, son capaces de "generar" infinitas combinaciones de los mismos (capacidad creativa del lenguaje), especialmente en lo que a la sintaxis se refiere. Lo único finito y, por lo tanto, susceptible de estudio, son las reglas, pero no las combinaciones de elementos obtenidas con esas reglas.

3) Por último, los corpus tampoco son la mejor forma de trabajar, ni siquiera en términos de metodología, ya que el recurso a la competencia, a la intuición del hablante, nos ahorra tiempo frente a la búsqueda en un corpus: no necesitamos contrastar los datos, ya que con los conocimientos de que disponemos aquellos son redundantes e innecesarios. Por otra parte, solo la introspección nos permite determinar la gramaticalidad de un enunciado o resolver ambigüedades. En el corpus están los datos, pero la decisión última sobre su validez corre a cargo del lingüista y de su opinión cualificada, su intuición.

Primera lingüística de corpus	Chomky
– Se centra en la fonética y la fonología.	– Se centra en la sintaxis.
– La lengua se concibe como finita.	– La lengua se concibe como no finita.
– Los corpus como única explicación.	– La intuición como única explicación.
– Los corpus son completos.	– Los corpus son parciales.

Tabla 21. Chomsky vs. la primera lingüística de corpus.

3.1.3.2. Críticas prácticas (Abercrombie)

Aparte de las críticas teóricas de Chomsky, existían problemas prácticos en la primera lingüística de corpus. El procesamiento de datos era lento, propenso al error y caro, al tener que ser realizado por personas. D. Abercrombie (1965) resumió el acercamiento basado en corpus como "pseudo-técnicas".

Estas críticas prácticas eran correctas si tenemos en cuenta que la "primera lingüística" de corpus requería recursos técnicos que no estaban disponibles en aquel entonces. Los corpus debían ser necesariamente de unas dimensiones reducidas, ya que de otra forma era imposible satisfacer las demandas del procesamiento de datos, que tenía que ser realizado por personas, lo que encarecía la tarea y reducía la fiabilidad de los análisis: cada persona podía tener criterios diferentes a la hora de interpretar los datos; por otro lado, hay que tener en cuenta los límites de atención y concentración del ser humano, lo que puede traducirse en cansancio, con los consiguientes errores y merma en la calidad de los resultados.

Por otra parte, el tamaño reducido de los primeros corpus favorecía que fueran sesgados o parciales en cuanto al reflejo de las variables lingüísticas. Además, sin las herramientas técnicas adecuadas, que no existían en el momento, solo con el análisis manual de los datos, tampoco tenía sentido reunir y analizar ingentes cantidades de datos.

Será la llegada del ordenador la que dé un nuevo impulso a la lingüística basada en corpus.

3.1.4. Segunda generación de lingüística de corpus

En este clima poco favorable de opinión de los años sesenta y setenta, se empezó a gestar, aunque al margen de la corriente lingüística dominante, lo que sería la segunda generación de trabajos en lingüística de corpus, marcada ahora por la presencia del ordenador, aunque algunos de los corpus que se recopilan durante este período no fueron diseñados para su informatización.

Fue en Estados Unidos donde se abordó la compilación del primer corpus informatizado sistemáticamente organizado. Desde entonces, los corpus electrónicos han llegado a erigirse en recursos imprescindibles para diversos fines relacionados con la investigación lingüística. Las características más destacadas de los corpus de estas décadas son:

Corpus-3

- La presencia del ordenador: solo en los años sesenta los ordenadores alcanzaron una potencia de procesamiento y una capacidad de almacenamiento suficientes para poder albergar grandes cantidades de texto, aunque en un principio no todos los proyectos para recopilar corpus se concebían pensando en su

informatización. No obstante, el vínculo entre los corpus y los ordenadores ya había sido establecido a finales de los cuarenta por R. Bussa (*cf.* McEnery 2003:452).

- Carácter representativo de los datos: la mayoría de los proyectos de elaboración de corpus pretendía recoger textos escritos que dieran cuenta del estado de la lengua en ese momento. Durante la década de los cincuenta (*cf.* McEnery:*ibid.*), A. Juilland estableció los conceptos de marco de la muestra, representatividad y equilibrio, básicos en el concepto actual de corpus.
- Tendencia a desfavorecer los datos orales por las dificultades técnicas y de transcripción. Predominan los corpus de textos escritos, aunque con notables excepciones.
- Tamaño: un millón de palabras.

Algunos corpus destacados de este período son:

En Inglaterra, R. Quirk (University College, London) sentó en 1959 las bases para la elaboración del *Survey of English Usage Corpus* (SEU)[198], corpus amplio y variado estilísticamente (de un millón de palabras, acorde con los estándares de la época), que empezó a recopilarse en 1961 con la intención de servir como base para una descripción sistemática del inglés británico hablado y escrito –los textos comprenden el período 1955-1985–, proyecto que sería completado por S. Greenbaum. No se diseñó como corpus electrónico, sino que se ciñó a la metodología de trabajo previa a la aparición del ordenador[199],

[198] URL: http://www.ucl.ac.uk/english-usage

[199] Las grabaciones se efectuaron en cintas magnetofónicas, y la transcripción de las mismas se realizó de forma manual, para posteriormente ser mecanografiada y almacenada en fichas de papel, en las que, además, se anotó información gramatical, ya que uno de los principales objetivos del proyecto es dar cuenta de la sintaxis del inglés. Lingüistas de la talla de D. Crystal, S. Greenbaum, G. Leech o J. Svartvik han contribuido al desarrollo de este proyecto a lo largo de los años, uno de cuyos

aunque en la actualidad todo el material está informatizado, en lo que se conoce como *London-Lund Corpus* (LLC). Sin embargo, sobresale por haber marcado las pautas de la futura lingüística de corpus debido a su cuidado diseño, con la selección de quinientas muestras textuales de cinco mil palabras cada una, y por haber trabajado con la lengua oral, en contraposición a la tendencia de la época. Además, hay que señalar que fue el primer proyecto de esta índole en Europa y todavía hoy en día sigue muy vivo, proporcionando de forma constante material nuevo para el estudio de la gramática[200], *software* para extraer la información contenida en los corpus, y también contribuyendo a nuevas recopilaciones, como el *Diachronic Corpus of Present-Day Spoken English*[201] (DCPSE), que ha tomado cuatrocientas mil palabras del LLC con el fin de estudiar la evolución de la lengua a lo largo del tiempo.

Sin embargo, el mérito de ser el primer corpus concebido específicamente para tener un formato electrónico hay que otorgárselo al trabajo que llevaron a cabo N. Francis y H. Kučera[202] en EE. UU. con el nombre de *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*, conocido de forma abreviada como *Brown Corpus*[203], por haberse recopilado en la universidad americana de Brown. Se trata de un corpus de un millón de palabras, diseñado con el objetivo de dar cuenta del inglés americano escrito en prosa. Los textos que lo conforman proceden de quinientas muestras de unas dos mil palabras cada una obtenidas de publicaciones de Estados Unidos de

---

productos destacados es *A Comprehensive Grammar of the English Language* (QUIRK *et al.* 1985), la gramática más importante realizada a partir de un corpus.

[200] Como *The Internet Grammar of English*, una aplicación gratuita en línea para el estudio de la gramática. URL: http://www.ucl.ac.uk/internet-grammar/. Otros aspectos gramaticales que centran las investigaciones amparadas por el SEU son el estudio de la frase nominal y verbal en inglés, o la subordinación.

[201] URL: http://www.ucl.ac.uk/english-usage/projects/dcpse/

[202] Este investigador checo afincado en Estados Unidos también fue pionero en el desarrollo de correctores ortográficos, gramaticales y de estilo.

[203] URL: http://icame.uib.no/brown/bcm.html (manual).

1961. Para decidir las categorías y proporciones temáticas, así como otros criterios de diseño del corpus (tamaño y número de las muestras) se celebró una conferencia en la que se debatieron esas cuestiones. El corpus se codificó en tarjetas perforadas de IBM siguiendo el estándar de procesamiento de datos empleado por la oficina de patentes de EE.UU., que después fueron transferidas a una cinta magnética. Existen varias versiones del corpus, una de ellas etiquetada morfosintácticamente, en la que a las palabras se les añadieron códigos con información sobre su categoría gramatical y otra información morfológica relevante, con vistas al análisis sintáctico automático o semiautomático. El corpus, que representó un hito por aquel entonces, sirvió de base para numerosos estudios y publicaciones. Por ejemplo, en 1967, los análisis a los que sus autores sometieron el corpus se tradujeron en la publicación de *Computational Analysis of Present-Day American English*, seguida en 1982 por *Frequency Analysis of English Usage: Lexicon and Grammar*. Por otra parte, estos trabajos de Kučera sobre la frecuencia léxica despertaron el interés editorial y dieron lugar a uno de los primeros diccionarios construidos a partir de las frecuencias e informaciones extraídas de un corpus que recogía el uso de la lengua, el *American Heritage Dictionary*, publicado en 1969. Obsérvense las diez palabras más frecuentes en el corpus[204]:

[204] URL: http://www.edict.com.hk/lexiconindex/frequencylists/words2000.htm

Words listed by frequency: the first 2000 most frequent words from the Brown Corpus (1,015,945 words)

	Word	Instances	% Frequency
1.	The	69970	6.8872
2.	of	36410	3.5839
3.	and	28854	2.8401
4.	to	26154	2.5744
5.	a	23363	2.2996
6.	in	21345	2.1010
7.	that	10594	1.0428
8.	is	10102	0.9943
9.	was	9815	0.9661
10.	He	9542	0.9392

Ilustración 113. Palabras más frecuentes en el "Brown Corpus".

Por otra parte, el diseño de este corpus sirvió de modelo para otros corpus compilados con posterioridad, como el LOB, de inglés británico, o el Kolhapur, de inglés de la India, que se guían por los mismos parámetros con el fin de comparar variedades de la lengua[205]:

El *Lancaster-Oslo/Bergen Corpus* (LOB)[206] es el resultado de los esfuerzos coordinados de G. Leech (Universidad de Lancaster), S. Johansson (Universidad de Oslo) y el *Norwegian Computing Centre for the Humanities* en Bergen: se trata de un corpus de un millón de palabras que recoge muestras de inglés británico escrito en 1961. Se elaboró ajustándose lo máximo posible a los parámetros de diseño del *Brown Corpus*. Se compiló entre 1970-1978. Existe también versión anotada, en la que a cada palabra se le ha añadido una etiqueta indicativa de su categoría gramatical[207]:

[205] Para observar el impacto del factor tiempo, se han compilado después *The Freiburg - Brown Corpus of American English* (*Frown Corpus*) y *The Freiburg – LOB Corpus of British English* (*FLOB Corpus*) en la Universidad de Friburgo, siguiendo los mismos criterios de diseño pero con textos del año 1991.

[206] URL: http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM (manual).

[207] URL: http://khnt.hit.uib.no/icame/manuals/lobman/INDEX.HTM.

```
A012    ^ *'_*' stop_VB electing_VBG life_NN peers_NNS **'_**' ._.
A013    ^ by_IN Trevor_NP Williams_NP ._.
A014    ^ a_AT move_NN to_TO stop_VB \OMr_NPT Gaitskell_NP from_IN
A014    nominating_VBG any_DTI more_AP labour_NN
A015    life_NN peers_NNS is_BEZ to_TO be_BE made_VBN at_IN a_AT
                meeting_NN
A015    of_IN labour_NN \OMPs_NPTS tomorrow_NR
A016     ^ \OMr_NPT Michael_NP Foot_NP has_HVZ put_VBN down_RP a_AT
A016    resolution_NN on_IN the_ATI subject_NN and_CC
A017    he_PP3A is_BEZ to_TO be_BE backed_VBN by_IN \OMr_NPT Will_NP
A017    Griffiths_NP ,_, \OMP_NPT for_IN Manchester_NP
```

Ilustración 114. Anotación gramatical del "LOB corpus".

También parte del corpus ha sido analizada sintácticamente de forma manual (*Lancaster-Leeds Treebank*) y otra parte de forma automática (*Lancaster Parsed Corpus*).

Con el mismo propósito se desarrolló *The Kolhapur Corpus of Indian English, for Use with Digital Computers*[208], compilado asimismo en la Universidad de Lancaster bajo la dirección de G. Leech. Su finalidad fundamental es la de servir de base para estudios comparativos entre las diferentes variedades del inglés y, al mismo tiempo, para efectuar una descripción del inglés hablado en la India, lo que marca una ligera diferencia con los dos corpus anteriores, así como la fecha de las publicaciones, que en este caso se refieren a 1978.

Otro proyecto destacable que se inicia a finales de este periodo es el de J. Svartvik en la Universidad de Lund, quien en 1975 inició la informatización de la parte correspondiente a textos orales que se habían transcrito en el marco del SEU en el proyecto conocido como *Survey of Spoken English* (SSE), que daría lugar al *London-Lund Corpus of Spoken English* (LLC) [209]. Se trata de un recurso todavía inigualado para el estudio del inglés hablado, ya que se han incluido marcas con información prosódica y paralingüística. Contiene medio millón de

[208] URL: http://khnt.hit.uib.no/icame/manuals/kolhapur/INDEX.HTM.
[209] URL: http://icame.uib.no/london-lund/.

palabras de inglés británico oral procedente de grabaciones realizadas entre 1953 y 1987.

Estos corpus son muy pequeños si los comparamos con los estándares de la actualidad, pero todavía se usan en la investigación debido a la utilidad de una estructura bien planeada y representativa, un rasgo que contrasta fuertemente con algunas de las colecciones a gran escala del momento.

3.1.5. Revisión de las críticas de Chomsky y Abercrombie

Así pues, la metodología de corpus continuó durante los sesenta y los setenta, pero como una metodología minoritaria, pese a la importancia que desde una perspectiva actual tienen algunos de los proyectos emprendidos en la época.

En el resurgir de la lingüística de corpus que ocurrió en la década de los ochenta tuvieron especial importancia diversos autores –entre los que sobresale la figura de G. Leech (1992)–, que rebatieron las críticas teóricas y prácticas que se habían formulado contra la primera lingüística de corpus. Si bien dichas críticas eran parcialmente válidas en su momento, según la opinión de G. Leech, la mayoría de las desventajas de los corpus se habían exagerado o habían dejado de ser ciertas, sobre todo gracias a la evolución de los ordenadores. Algunos de los principales argumentos de G. Leech a favor del uso de los corpus son:

1) El corpus como metodología científica: desde el punto de vista del método científico, el corpus ofrece una serie de ventajas frente a la intuición[210], ya que está sujeto a verificación, lo que descarta el recurso a ejemplos inventados por los lingüistas de forma interesada. Además, en el caso de datos cuantitativos, como la frecuencia, la intuición no es un recurso válido: nuestra percepción de la frecuencia es totalmente subjetiva.

2) La gramaticalidad de los enunciados de un corpus: la mayoría de enunciados de un corpus es gramatical, por lo que los corpus reflejan la competencia[211]. Según Chomsky, los corpus, como muestras de uso de la lengua (actuación), no eran más que un pobre reflejo de la competencia. Sin embargo, los trabajos de W. Labov (1969) mostraron el alto porcentaje de secuencias gramaticales en un corpus.

3) La utilidad de los datos cuantitativos: los corpus son una fuente inigualable para la extracción de este tipo de datos[212]. Si el corpus está bien diseñado, los datos relativos a la frecuencia de uso serán representativos de la lengua en su totalidad.

4) Con el uso del ordenador, el procesamiento de los datos de un corpus no es un conjunto de pseudo-técnicas[213]. Los ordenadores son capaces de procesar gran cantidad de datos a un coste reducido, de forma mucho más rápida que las personas y sin cometer errores.

[210] Rebate la tercera crítica de Chomsky.
[211] Rebate la primera crítica de Chomsky.
[212] Rebate la segunda crítica de Chomsky.
[213] Rebate la crítica de Abercrombie.

De forma resumida, la siguiente tabla refleja las dos posturas enfrentadas:

Lingüística de corpus	Generativismo
Datos	Juicios del hablante
• Externos	• Internos
• Públicos	• Privados
• Observables	• No observables
• Verificables	• No verificables
• Naturales	• Artificiales
• Noción exacta de frecuencia	• Noción vaga de frecuencia

Tabla 22. Datos vs. juicios del hablante.

3.1.6. Renacer actual de la lingüística de corpus

Superadas las críticas -teóricas y prácticas-, y con las nuevas ventajas y posibilidades que ofrecían los ordenadores, los corpus electrónicos se convierten desde la década de los ochenta en un recurso indispensable para el estudio del lenguaje, para probar hipótesis lingüísticas y para construir sistemas prácticos de procesamiento del lenguaje natural. Solo entonces se generaliza el término, especialmente a partir de 1984, año en que J. Aarts y W. Meijs editaron el volumen titulado *Corpus Linguistics I: Recent Developments in the Use of Computer Corpora*, y se empieza a hablar de lingüística de corpus en el sentido actual del término. Algunos hechos que favorecieron este renacer de la lingüística de corpus como metodología de trabajo en Lingüística fueron:

1) El auge de las áreas aplicadas de la Lingüística en general y de la Lingüística Computacional en particular, lo que ha puesto en evidencia la necesidad de contar con datos de uso de la lengua, con datos procedentes de la actuación, tanto de hablantes nativos como no nativos. Por una parte, los corpus reflejan la variedad de la lengua y, por otra, pueden recoger estructuras nuevas o que no se ajustan a las descripciones teóricas –a las que el lingüista no podría haber accedido desde su competencia– y que, sin embargo, requieren una explicación. Además, en el caso de hablantes no nativos, los corpus son una muestra excelente de evidencias de uso de la lengua (por ejemplo, para el análisis de errores o para la elaboración de materiales didácticos).

2) El eclecticismo: el uso de corpus no se concibe como incompatible con el recurso a los juicios del lingüista; por sí solos ni los corpus (postura de los estructuralistas americanos) ni los juicios o intuiciones del hablante-oyente ideal (postura de Chomsky) son suficientes para explicar los fenómenos lingüísticos. En la actualidad, se reconoce que los corpus, al suplir datos textuales de primera mano, no se pueden analizar válidamente sin la intuición y la facultad interpretativa del analista, que usa conocimientos de la lengua (como hablante nativo o no nativo competente) y conocimientos acerca del lenguaje (como lingüista).

3) La mayor disponibilidad de corpus electrónicos, sobre todo gracias a las posibilidades que ofrece Internet para la obtención de textos en dicho formato.

4) El desarrollo de nuevas tecnologías para la introducción de textos de forma más rápida, como el OCR o reconocimiento óptico de caracteres, el dictado automático, etc.

5) La utilidad de los datos cuantitativos en el estudio de determinados aspectos del lenguaje.

6) En LC en particular, el desarrollo de productos comerciales que se empieza a producir en los ochenta pone de manifiesto que los formalismos gramaticales del momento[214], que tan elegantes resultaban desde una perspectiva puramente teórica, no eran útiles para tratar los textos reales producidos por los hablantes, de ahí la necesidad, por una parte, de contar con vocabularios o diccionarios más extensos con el fin de ampliar la cobertura de los sistemas (sistemas capaces de trabajar con cualquier tipo de texto y no solo con sublenguajes o lenguajes limitados a dominios muy restringidos, como el de los partes meteorológicos); y, por otra parte, de manejar frecuencias, estadísticas y cálculos de probabilidades para manipular cantidades cada vez más grandes de texto y para desarrollar sistemas de extracción automática de reglas, así como desambiguadores estocásticos[215].

Como consecuencia de estos hechos, los grandes corpus textuales se erigen como uno de los recursos fundamentales de la ingeniería lingüística o de las tecnologías del lenguaje, áreas o líneas de trabajo en LC en las que son imprescindibles para desarrollar sistemas prácticos; por otro lado, sin ellos tampoco se concibe hoy en día el diseño de gramáticas y de lexicones computacionales. Sobre estas premisas nace la lingüística de corpus tal y como se entiende en la actualidad, definida como "el área de la lingüística especializada en el aprovechamiento de los corpora [*sic*]" (Abaitua 2002:62), "the study of language on the basis

[214] Es entonces cuando surge la familia de gramáticas de unificación.

[215] Estos programas toman decisiones de forma automática en casos de ambigüedad categorial, semántica, sintáctica, etc. La decisión depende del peso estadístico que cada una de las opciones tenga. Esta información en ocasiones se combina con conocimiento lingüístico.

of text corpora" (Aijmer y Altenberg 1991:1) o "the use of large collections of text available in machine-readable form" (Svartvik 1992:7). Estos corpus se van a caracterizar por:

Corpus-4

- Formato electrónico: conjunto de textos informatizados.
- Tamaño en aumento progresivo, hasta superar los cien millones de palabras, aunque existen corpus de tamaños inferiores.
- Carácter abierto: corpus no cerrados, sino en continua actualización (corpus monitor).
- Vertiente comercial: los corpus no se limitan a centros de investigación o instituciones vinculadas a la lengua, sino que muchos proyectos son desarrollados por consorcios comerciales, principalmente editoriales, o empresas de telecomunicaciones en el caso de los corpus orales.
- Se amplía el repertorio de lenguas que disponen de corpus, y también se elaboran corpus multilingües.
- Automatización de las diferentes tareas de procesamiento de los textos de un corpus gracias a los avances técnicos que permiten realizar, de forma automática o semiautomática, la asignación de categoría gramatical a cada una de las palabras del corpus, la desambiguación, la extracción de concordancias o ejemplos en contexto, el alineamiento de las grabaciones de audio con su correspondiente transcripción, etc.
- Estímulo para el desarrollo de nuevos modelos y campos de investigación en Lingüística, así como para la realización de estudios sobre los más variados aspectos lingüísticos, desde los gramaticales hasta los discursivos, pasando por los históricos, los psicolingüísticos o los culturales.

Algunos ejemplos de corpus que responden a estas características son *The British National Corpus*, *The Bank of English* y los bancos de datos de la Real Academia Española, CREA y CORDE, entre otros.

The British National Corpus (BNC)[216] es un corpus de unos cien millones de palabras de inglés británico contemporáneo tanto escrito como hablado[217], a las que se han añadido marcas relacionadas con su categoría gramatical, así como otras informaciones sobre las características estructurales de los textos. El proyecto depende de un consorcio académico-industrial liderado por la editorial Oxford University Press junto con otras editoriales especializadas en diccionarios, la Universidad de Lancaster, la Universidad de Oxford y la British Library. Se inició en 1991 y finalizó en 1994, por lo que es de naturaleza cerrada. Las múltiples posibilidades que ofrece se extienden a cualquier tipo de investigación sobre el lenguaje que se apoye en el uso de medios informáticos. Las aplicaciones principales que se han hecho de él tienen que ver con: la publicación de libros de referencia (gramáticas y diccionarios); la realización de estudios teóricos y aplicados sobre las diferentes áreas de la Lingüística (léxico, sintaxis, morfología, semántica, análisis del discurso, sociolingüística, estilística, etc.); el diseño de sistemas de procesamiento del lenguaje natural (etiquetadores, analizadores, correctores ortográficos…); y la elaboración de materiales didácticos para la enseñanza del inglés.

The Bank of English[218] es un corpus de más de quinientos veinticuatro millones de palabras[219] de inglés moderno tanto escrito como oral de

---

[216] URL: http://www.natcorp.ox.ac.uk/

[217] Está formado por un 90% de textos escritos y un 10% de textos orales, con el objetivo general de representar el inglés británico de finales del siglo XX con el objetivo de desarrollar materiales de referencia y llevar a cabo investigaciones lingüísticas.

[218] URL: http://www.titania.bham.ac.uk/. El corpus comprende textos de diferentes variedades de inglés: británico, americano, canadiense y australiano. Igual que en el BNC, se han introducido marcas para indicar la categoría gramatical; por

diferentes procedencias. Asumido en la actualidad por el Proyecto COBUILD[220], se desarrolla en la Universidad de Birmingham bajo la dirección de J. Sinclair en colaboración con la editorial Collins COBUILD. El corpus se lanzó en 1991, pero COBUILD ya llevaba desde 1980 recopilando textos electrónicos para elaborar sus diccionarios. Está integrado en la *Collins World Web*, una base de datos de más de dos billones y medio de palabras, a la que cada mes se le añaden otros treinta y cinco millones. Se trata del recurso de este tipo más grande existente en el mundo.

En el ámbito del español, son referencia obligada el *Corpus de Referencia del Español Actual* (CREA) y el *Corpus Diacrónico del Español* (CORDE), iniciados a mediados de los noventa en el seno de la Real Academia Española[221], bajo la dirección académica de Guillermo Rojo.

El primero de ellos, el CREA, dispone, según el último anuncio oficial (mayo de 2008), de 160 millones de palabras procedentes de textos escritos y orales, producidos desde 1975 hasta 2004 en los distintos países de habla hispana. Sigue el mismo modelo que el BNC, en el sentido de que el 90% de los textos son escritos y el 10% restante, orales. Considera criterios geográficos (el 50% de los textos proceden de España y el otro 50%, de Hispanoamérica), temáticos (se distinguen diferentes hipercampos, como ciencias y tecnología, artes, salud, ficción,

---

otra parte, se han analizado sintácticamente unos doscientos millones de palabras. Otra característica reseñable es que constantemente se introducen nuevos materiales para mantener el corpus lo más actualizado posible. Además del corpus mismo, el equipo de lexicógrafos y lingüistas con que cuenta el proyecto ha desarrollado herramientas adicionales para analizar el corpus y extraer todo tipo de información: patrones de combinación de palabras o colocaciones, frecuencias de aparición, ejemplos de uso, etc. La meta es examinar esos datos para conseguir diccionarios y materiales de referencia sólidos.

219 Aunque "solo" se pueden consultar 450 en la actualidad.

220 URL: http://www.collins.co.uk/books.aspx?group=140.

221 URL: http://www.rae.es/ (sección "Banco de datos").

noticias, retransmisiones deportivas…), cronológicos y de medio de publicación (prensa, libro, radio, televisión, etc.) a la hora de seleccionar los textos. Como el *Bank of English*, es un corpus que pretende ser representativo del español contemporáneo. Se trata del recurso de este tipo más importante para el español, por lo que, además de servir como fuente de datos reales para las obras académicas[222], sus aplicaciones, tanto para la investigación como para el diseño de productos comerciales, son numerosas.

El segundo, el CORDE, dispone de unos 300 millones de palabras, procedentes de textos escritos desde los orígenes del idioma hasta 1974. La mayoría de formas procede de libros (el 97%), por lo que respecta al medio de publicación; en cuanto al origen geográfico, los textos de España se aproximan al 75%, y el 25% restante provienen de Hispanoamérica; según el género, la mayoría de los textos son de prosa (85%) y el resto de verso; por último, se reparten cronológicamente entre los orígenes y 1491, 1492-1712 y 1713-1974. Se trata de un corpus complementario del CREA. Su principal aplicación está relacionada con su uso para los estudios diacrónicos, en especial para proveer material al relanzado proyecto del *Diccionario histórico* de la RAE.

Aunque todavía en fase de elaboración, no podemos hablar de los bancos de datos académicos sin mencionar el último proyecto[223] que la RAE ha emprendido en este sentido, el *Corpus del Español del Siglo XXI* (CORPES), en colaboración con el resto de Academias de la Lengua, también promovido por Guillermo Rojo, al que ha sucedido en esta

[222] Tanto el CREA como el CORDE han suministrado información para la redacción de las entradas de las últimas obras académicas, desde la vigésima segunda edición del *DRAE* hasta el *Diccionario esencial*, pasando por el *Diccionario del estudiante* y el *Diccionario panhispánico de dudas*. Y continúan proporcionando aportaciones a los nuevos proyectos académicos: la vigésima tercera edición del *DRAE*, la *Nueva gramática de la lengua española*, la *Ortografía* y el *Diccionario de americanismos*.

[223] Propuesto por la RAE en el Congreso de Academias celebrado en Medellín en marzo de 2007.

labor recientemente José Antonio Pascual. Su misión fundamental es proporcionar material suficiente para que la Academia pueda fundamentar sus decisiones sobre datos actualizados del uso de la lengua. En su primera fase, que comprenderá textos del periodo 2000-2011, el corpus constará de 300 millones de formas, el 70% procedente de Hispanoamérica y el 30% restante, de España, por lo que se erigirá en el recurso fundamental para el estudio del español contemporáneo.

Quizá no tan importantes en tamaño, pero sí en cuanto al objetivo que persiguen, existen algunos proyectos relacionados con otras lenguas peninsulares:

El *Corpus de Referencia do Galego Actual* (CORGA)[224], recopilado en el *Centro Ramón Piñeiro para a Investigación en Humanidades* bajo la dirección del académico Guillermo Rojo. Con sus veintitrés millones de formas, pretende ser representativo de una lengua, la gallega, motivo por el cual recoge textos de diferente procedencia temática, geográfica y cronológica, publicados desde 1975 a la actualidad, con el objetivo de proporcionar una base para la descripción de aspectos morfológicos, sintácticos y léxicos del gallego de los últimos tiempos.

El *Corpus Textual Informatitzat de la Llengua Catalana* (CTILC)[225], con cincuenta y dos millones de palabras procedentes de textos escritos en catalán entre 1832 y 1988, se inscribe dentro de las actividades lexicográficas del *Instituto d'Estudis Catalans*. Terminado en 1997, se recopiló inicialmente para elaborar el *Diccionari descriptiu de la llengua catalana*, aunque con el tiempo se han ampliado sus usos y en la actualidad se inscribe en el proyecto del *Diccionari del Català Contemporani* (DCC).

[224] URL: http://corpus.cirp.es/corga/
[225] URL: http://ctilc.iec.cat/

Por último, el *Corpus estadístico del euskera del siglo XX*[226], con más de cuatro millones y medio de palabras procedentes de textos escritos en vasco durante el siglo XX (entre 1900 y 1995), se desarrolló entre 1987 y 1999 en el Centro Vasco de Terminología y Lexicografía (UZEI) con el fin de reflejar el uso de la lengua vasca durante el período mencionado.

[226] URL: http://www.uzei.com/antbuspre.asp?nombre=1901&cod=1901&sesion=1

3.2. Ventajas e inconvenientes del trabajo con corpus

Algunas ventajas que justifican el interés actual por los corpus electrónicos tienen que ver con que:

a) Proporcionan objetividad y ofrecen la posibilidad de verificar las teorías construidas a partir de ellos.

b) Aportan rapidez, precisión y consistencia en el procesamiento de los datos a un bajo coste.

c) Facilitan el acceso y manipulación de los materiales.

d) Permiten el procesamiento automático de textos así como la explicitación de diferentes tipos de información (p. ej. la categoría gramatical de las palabras) que multiplican sus posibilidades de explotación ulterior.

e) El mismo recurso puede tener variados usos y aplicaciones.

f) Responden a la necesidad de contar con grandes cantidades de datos reales fácilmente accesibles como una base más realista para el estudio del lenguaje y también para el desarrollo de sistemas prácticos de procesamiento del lenguaje natural.

g) Han permitido automatizar total o parcialmente muchas tareas –que antes debían efectuarse de forma manual– mediante programas diseñados para extraer información de los corpus. Destacan especialmente los relacionados con el análisis gramatical y sintáctico que son, a su vez, el punto de partida para el desarrollo de modelos probabilísticos sobre el funcionamiento del lenguaje o para probar modelos propuestos por la Lingüística Teórica.

h) Son un recurso muy fructífero para estudios contrastivos y para explorar los aspectos cuantitativos y probabilísticos del lenguaje, de especial utilidad en el ámbito de la lexicografía y de la elaboración de materiales para la enseñanza de la lengua.

i) Son la única vía posible para los estudios diacrónicos, en los que no existe la posibilidad de recurrir a hablantes vivos.

j) Constituyen muestras importantes de uso para los estudios de variación.

En cuanto a las desventajas, se han señalado las siguientes[227]:

1) En determinadas áreas los corpus no son suficientes y es necesario acudir a los análisis manuales, como en el caso de la pragmática.

2) En los corpus de lengua oral, si se trabaja con transcripciones, existe el peligro de alejarse demasiado del texto original.

3) El tamaño no es tan decisivo como la adecuación del corpus a la finalidad para la que ha sido planeado.

No obstante, hoy en día no se puede negar la repercusión que los corpus están teniendo en diferentes áreas relacionadas con el estudio de la lengua y la literatura, en las que cada vez es más frecuente incorporar este tipo de recursos: así ocurre en el área de la enseñanza de lenguas, donde aportan muestras de uso de la lengua para la elaboración de materiales didácticos; en el ámbito de la lexicografía, donde su impacto ha sido decisivo y es inconcebible un diccionario que no se sustente sobre los datos que suministran los corpus (listas de frecuencias, concordancias, colocaciones, ejemplos, información gramatical,

[227] *Vid.* MCENERY, XIAO y TONO (2006:131 y ss.) para las controversias que el uso de corpus ha suscitado entre diversos investigadores.

semántica o sobre el ámbito temático, el lugar o el registro); en el análisis del discurso y análisis crítico del discurso, campos en los que el empleo de corpus permite extraer patrones o tendencias, así como el estudio de la ideología presente en los textos; en la lingüística forense, para determinar casos de plagio o de autoría en contextos legales; en los estudios sobre la variación lingüística o sobre la estilística, etc. (*cf.* Adolphs 2006).

3.3. El concepto de corpus

Como se ha podido apreciar en apartados anteriores, en la actualidad el concepto de corpus ha cambiado mucho con respecto al que manejaban los primeros lingüistas que lo empleaban como recurso para sus investigaciones.

Hoy en día se considera que los corpus deben cumplir los siguientes requisitos (*cf.* McEnery y Wilson 2001:21 y ss.):

1) *Textos en formato electrónico*: un corpus, para ser una herramienta útil al lingüista, debe estar informatizado, es decir, los textos de que consta tienen que estar en formato electrónico (corpus informatizado o automatizado). El hecho de que para los primeros corpus no se pudiera disponer de ordenadores motivó la crítica de las pseudo-técnicas: el procesamiento de los datos debía efectuarse de forma manual, con los errores y problemas que eso ocasionaba. Sin embargo, el empleo del ordenador permite automatizar tareas tales como:

- *Búsqueda de información*: un corpus informatizado permite localizar de forma rápida una palabra, una secuencia de palabras o incluso una categoría gramatical en décimas de segundo.
- *Recuperación de información*: un corpus informatizado permite obtener todos los casos de una palabra, secuencia de palabras, etc. registrados en el corpus, normalmente con su contexto inmediato anterior y posterior (*concordancia*).
- *Cómputo de la frecuencia* de aparición de una palabra, secuencia de palabras, etc.
- *Clasificación de los datos* contenidos en el corpus según diferentes criterios: orden alfabético, frecuencia de aparición, autor, procedencia geográfica, tema, medio de publicación, etc.

2) *Autenticidad de los datos*: los textos recogidos en el corpus deben ser muestras reales de uso de la lengua objeto de estudio. A partir de ellas se construyen (o verifican) de forma empírica las teorías que tratan de explicar el funcionamiento de la lengua o las aplicaciones computacionales.

3) *Criterios de selección*: los textos que forman parte del corpus deben haber sido elegidos de acuerdo con unos determinados criterios –lingüísticos y/o extralingüísticos– para la finalidad concreta que persiga el corpus[228].

4) *Representatividad*: la selección de los textos, además de a unos criterios adecuados, debe responder a parámetros estadísticos que garanticen que los textos "representan" la variedad de lengua objeto de estudio (*muestra representativa*). Esta variedad puede referirse a la obra de un autor determinado, a un período de tiempo, a un género, etc. Cuando lo que nos interesa es la lengua en su conjunto, la opción de reunir en un corpus todas las muestras de esta se hace impracticable, a diferencia, p. ej., de lo que ocurre si queremos recoger todas las obras de Cervantes, que son un universo cerrado. La única solución posible, entonces, es tomar una muestra más pequeña de esa lengua, que refleje, a pequeña escala, el funcionamiento del todo que es la lengua. Como Chomsky criticó con acierto, los corpus corren el riesgo de ser sesgados. Para subsanar este problema se recurre a la selección, según criterios estadísticos, de textos de diversos géneros, tipologías, temas, medios de publicación, etc.

[228] Precisamente el uso de unos criterios previos y explícitos diferencia un corpus de otras recopilaciones de textos tales como archivos, colecciones o bibliotecas electrónicas.

5) *Tamaño*: por lo general, los corpus constan de un tamaño finito, que se suele medir en millones de palabras (o formas) y que se fija antes de empezar la recogida de los textos (p. ej. un millón de palabras); una vez alcanzado ese número, se da por terminada la recopilación del corpus, que no es más que el primer paso de todo el proceso[229]. Sin embargo, también existen corpus abiertos o monitor, como el del proyecto COBUILD dirigido por J. Sinclair en la Universidad de Birmingham, de especial interés para la lexicografía. En el pasado se pensaba que el tamaño era muy importante: mientras mayor fuera el corpus, más posibilidades de reflejar el funcionamiento real de la lengua en todas sus variedades, pero en la actualidad priman los criterios de diseño, es decir, el tamaño solo es importante en la medida en que así lo exija la finalidad del corpus[230].

A continuación se recogen algunas definiciones de corpus que ilustran estas características[231]:

(i) *A* collection of texts *assumed to be* representative *of a given language, dialect, or other subset of a language, to be used* for linguistic analysis (Francis 1982:7 *apud* Francis 1992:17).

(ii) *A* collection of pieces of language *that are* selected *and* ordered *according to* explicit linguistic criteria *in order to be used as* a sample *of the language* (Sinclair 1994:14).

---

[229] Una vez recogidos los textos, estos se codifican, anotan y explotan de múltiples formas, por lo que la recopilación en sí no es más que una primera fase necesaria en todo proyecto de corpus.

[230] Lógicamente, un corpus que pretenda ser representativo de una lengua en toda su variedad (español, inglés, francés…) no podrá conformarse con unos pocos millones de palabras, mientras que un corpus cuyo objetivo sea describir un sublenguaje (jurídico, informático….) puede permitirse un tamaño más reducido. Además, la disponibilidad de los textos es otro factor que puede influir en el tamaño.

[231] Se han destacado tipográficamente las características relevantes.

(iii) A finite-sized *body of* machine-readable texts sampled *in order to be maximally* representative *of the language variety under consideration* (McEnery & Wilson 1996:24).

(iv) *Un corpus és una* mostra *d'una llengua que habitualment s'ha construït a partir d'una* selecció *de textos realitzada segons uns determinats* criteris *i amb un determinat* objectiu (Martí y Castellón 2000:151).

(v) *The term* corpus *should properly only be applied to a* well-organized collection of data*, collected within the boundaries of a* sampling frame *designed to allow* the exploration *of a certain linguistic feature (or set of features) via the collected data* (McEnery 2003:449).

(vi) Un corpus es un *conjunto de textos* de *lenguaje natural e irrestricto*, almacenados en un *formato electrónico homogéneo*, y *seleccionados y ordenados*, de acuerdo con *criterios explícitos*, para ser utilizados como *modelo* de un estado o nivel de lengua determinado, en estudios o aplicaciones relacionados en mayor o menor medida con el *análisis lingüístico* (Santalla 2005:45-46).

(vii) *Rather, the term* corpus *as used in modern linguistics can best be defined as a* collection of sampled texts*, written or spoken,* in machine-readable form *which may be annotated with various forms of linguistic information* (McEnery, Xiao y Tono 2006:4).

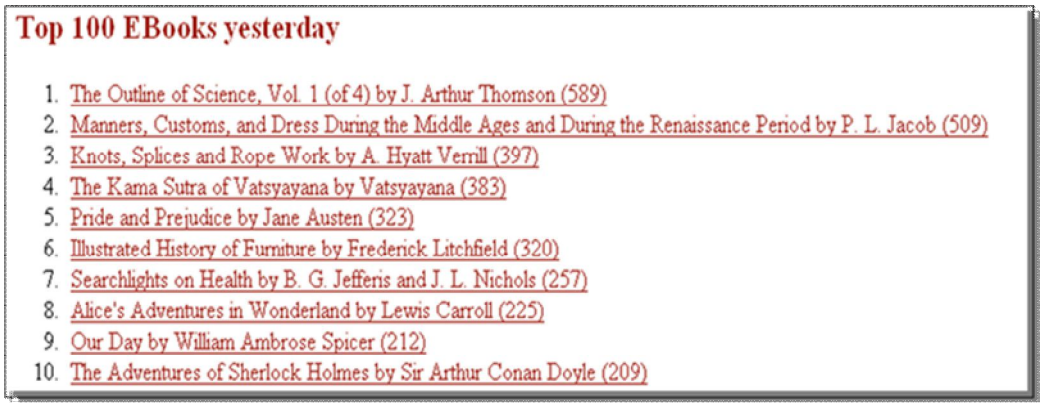
Estos criterios y definiciones permiten discriminar los corpus, en el sentido que se maneja en la lingüística de corpus, de otras colecciones de textos electrónicos[232] (*cf.* Sinclair 1996; Torruella y Llisterri 1999:51-52):

[232] En lingüística de corpus, el término *texto* se refiere tanto a una muestra de lengua escrita como a una de lengua oral (*cf.* SINCLAIR 1996).

- Archivo (o colección) informatizado: se trata de un simple conjunto de textos electrónicos sin estructurar. El único criterio que prevalece a la hora de conformarlo es la disponibilidad de los textos.
- Biblioteca de textos electrónicos: se trata de un conjunto de textos electrónicos recogidos sin seguir criterios lingüísticos, pero guardados en un formato estándar[233].

Algunos ejemplos de archivos y bibliotecas que reúnen textos en soporte electrónico, pero que no pueden calificarse como corpus en sentido estricto, son:

- *Proyecto Gutenberg*[234]: proyecto pionero en lo que se refiere a la recopilación de libros o textos electrónicos. Data de 1971, momento en que fue fundado por Michael Hart y, desde entonces, recoge textos clásicos, textos de literatura ligera y obras de referencia anteriores a 1923. Por ejemplo, estos son los diez libros electrónicos más descargados (04/06/2009), que también pueden leerse *on-line*:



Top 100 EBooks yesterday

1. The Outline of Science, Vol. 1 (of 4) by J. Arthur Thomson (589)
2. Manners, Customs, and Dress During the Middle Ages and During the Renaissance Period by P. L. Jacob (509)
3. Knots, Splices and Rope Work by A. Hyatt Verrill (397)
4. The Kama Sutra of Vatsyayana by Vatsyayana (383)
5. Pride and Prejudice by Jane Austen (323)
6. Illustrated History of Furniture by Frederick Litchfield (320)
7. Searchlights on Health by B. G. Jefferis and J. L. Nichols (257)
8. Alice's Adventures in Wonderland by Lewis Carroll (225)
9. Our Day by William Ambrose Spicer (212)
10. The Adventures of Sherlock Holmes by Sir Arthur Conan Doyle (209)

Ilustración 115. Lista de los diez libros más descargados en el Proyecto Gutenberg.

[233] A diferencia de archivos y bibliotecas, en los corpus informatizados los textos que los componen se han seleccionado según unos criterios lingüísticos y/o extralingüísticos y se han codificado de acuerdo con un estándar, con la finalidad de proporcionar una imagen real de la lengua a partir de la cual se pueden extraer generalizaciones para su estudio.

[234] URL: http://www.gutenberg.org/wiki/Main_Page

- *Búsqueda de libros de Google*[235], antes *Google Print*: proyecto para digitalizar y poner a disposición de los usuarios de Internet los fondos bibliográficos de las universidades de Stanford, Harvard, Princeton, Oxford, Michigan, Complutense de Madrid y de la Biblioteca Pública de Nueva York, entre otros centros que se suman al proyecto cada año.

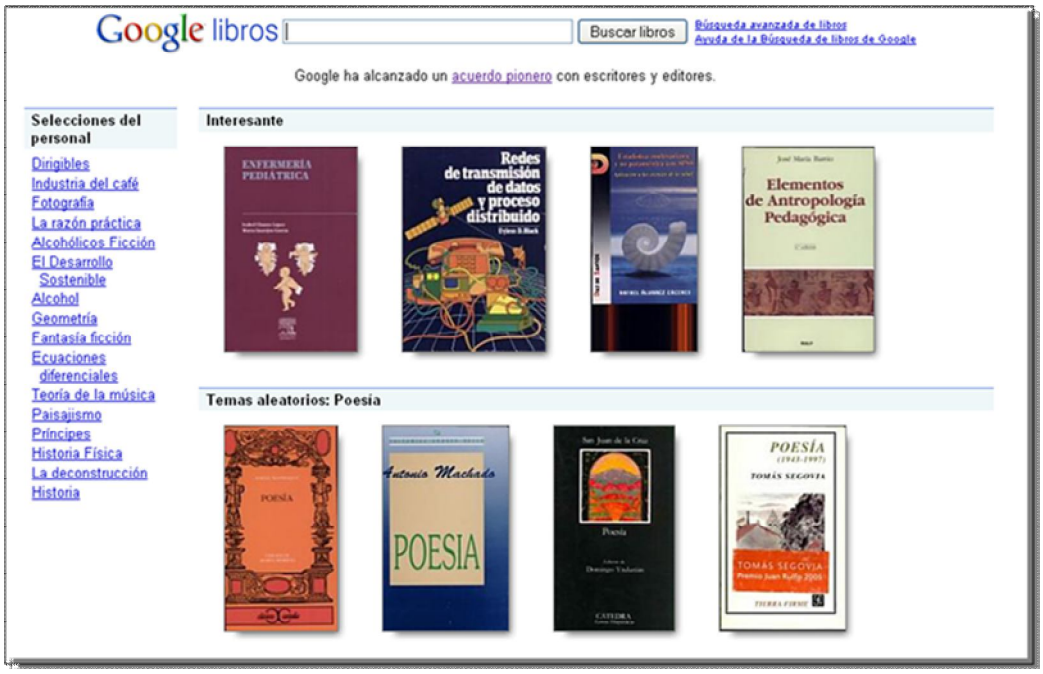


Ilustración 116. Google libros.

- *The Universal Digital Library. Million Book Collection*[236], un proyecto de la Universidad de Carnegie Mellon que nace con el ambicioso objetivo de poner en Internet todos los libros publicados en la historia de la humanidad y que ha empezado por la digitalización de un millón de ellos.

[235] URL: http://books.google.com/
[236] URL: http://www.ulib.org/

- *The Oxford Text Archive*[237]: ubicado en la Universidad de Oxford, desde 1976 recoge textos en veinticinco lenguas de autores individuales, obras de referencia y corpus.
- *Electronic Text Center*[238], Biblioteca de la Universidad de Virginia. Desde 1992 recoge textos en quince lenguas.
- *Electronic Text Collections in Western European Literature*[239]: reúne textos literarios en diferentes lenguas europeas distintas del inglés. Es mantenida por James Campbell, *Western European Studies Section*, *Association of College and Research Libraries*, *American Library Association*.
- *Biblioteca Virtual Miguel de Cervantes*[240]: proyecto que nace en 1999 por iniciativa de la Universidad de Alicante y el Banco Santander con el objetivo de recopilar las principales obras de literatura en español, así como otros recursos relacionados con la historia de España e Hispanoamérica.

Por último, hay que hacer mención de Internet como un corpus (*cf.* Adolphs 2006:33), no en el sentido estricto del término que hemos expuesto (*vid. supra*), porque no sigue unos criterios de diseño y en muchos casos falta información sobre el número y procedencia de los textos. Sin embargo, hay que reconocer su utilidad, aunque sea con los debidos filtros, como fuente de información para los estudios lingüísticos[241].

---

[237] URL: http://ota.ahds.ac.uk/
[238] URL: http://lib.virginia.edu/digital/collections/finding_digital.html
[239] URL: http://www.lib.virginia.edu/wess/etexts.html
[240] URL: http://www.cervantesvirtual.com/
[241] *Vid.* el caso que comenta MORALA (2002) para la palabra *fuereño*.

La web ofrece la posibilidad de acceder a un conjunto de textos, en formato electrónico, que son muestras reales de uso de la lengua de todo tipo y materia y que constituyen un proyecto abierto en cambio continuo, que pueden ser recuperados mediante las diferentes funciones de búsqueda de cualquier navegador.

Sirva de ilustración de este uso de la web como corpus y su contraste con un corpus propiamente dicho, el siguiente caso, que planteamos como parte de una de las prácticas de la asignatura Lingüística Computacional, en 2005, que tenía por objeto el uso de recursos electrónicos para obtener información lingüística. El punto de partida fue la localización de una forma, *bibidí*, en las respuestas a una encuesta del proyecto VARILEX[242]:


B092 [ATHLETIC SHIRT] Camiseta sin mangas.

1)bibidí; 2)Bid; 3)camiseta; 4)camiseta de hombreras; 5)camiseta de resaque; 6)camiseta de sport; 7)camiseta de tirantes; 8)camiseta de tiras; 9)camiseta imperio; 10)camiseta malla; 11)camiseta manga sisa; 12)camiseta sin mangas; 13)camisilla; 14)esqueleto; 15)esqueletos; 16)franela; 17)franelilla; 18)malla; 19)musculosa; 20)playera; 21)camiseta de tirillas; 22)camisola; 23)playera sin mangas.
&)Otros: _____; #)No se me ocurre.
$)Comentario:

Ilustración 117. Localización de la forma "bibidí" en VARILEX.

Sin facilitar esta información a los estudiantes, se les pidió que realizaran una investigación, utilizando todos los recursos electrónicos a su alcance (en especial, diccionarios y corpus), para determinar: el significado de la palabra, su origen, su ámbito geográfico, sus variantes formales, etc.

---

242 URL: http://gamp.c.u-tokyo.ac.jp/~ueda/varilex/cues/cues2000.pdf. Para más información sobre el proyecto VARILEX, *Variación léxica del español en el mundo*, remitimos a la URL: http://lingua.cc.sophia.ac.jp/varilex/index.php

La experiencia resultó muy interesante y, gracias a ella, pudimos comprobar cómo en el *DRAE* esta palabra aparece recogida en su variante *bivirí*:

Ilustración 118. Localización de la forma "bivirí" en el DRAE.

Sin embargo, esta información fue la última a la que llegaron, pues, al buscar *bibidí* inicialmente en el *DRAE*[243], este nos devuelve el mensaje de que la palabra no está en el Diccionario[244]:

Ilustración 119. Resultados de la búsqueda de la forma "bibidí" en el DRAE.

[243] Por defecto, el DRAE selecciona la opción de "Búsqueda por aproximación", es decir, no solo va a intentar recuperar la forma propuesta, sino otras que estén en el diccionario y tengan una ortografía o pronunciación similar, por si hemos cometido una equivocación al teclear el término deseado. No es posible utilizar comodines en el diccionario académico.

[244] Lo mismo ocurre en el resto de diccionarios electrónicos para el español. En ningún caso se ofrecen alternativas válidas (palabras con una forma parecida, bien por su pronunciación o por su grafía), ya que la variación fonético-ortográfica que refleja esta palabra no parece ajustarse a las normas del español, sino a la pronunciación del fonema /d/ como [ɾ] en posición intervocálica en el inglés americano (*cf.* EDWARDS 2003:92), lo que nos da indicios de que su forma de entrar en el español es por vía oral y no escrita.

Continuando las indagaciones, procedemos a realizar la misma consulta en el CREA, con la misma respuesta negativa: "No existen casos para esta consulta". Sin embargo, gracias al empleo de comodines (*?i?idí*) –que los bancos de datos académicos sí permiten–, obtenemos los primeros frutos de nuestra búsqueda: 16 casos en 13 documentos.

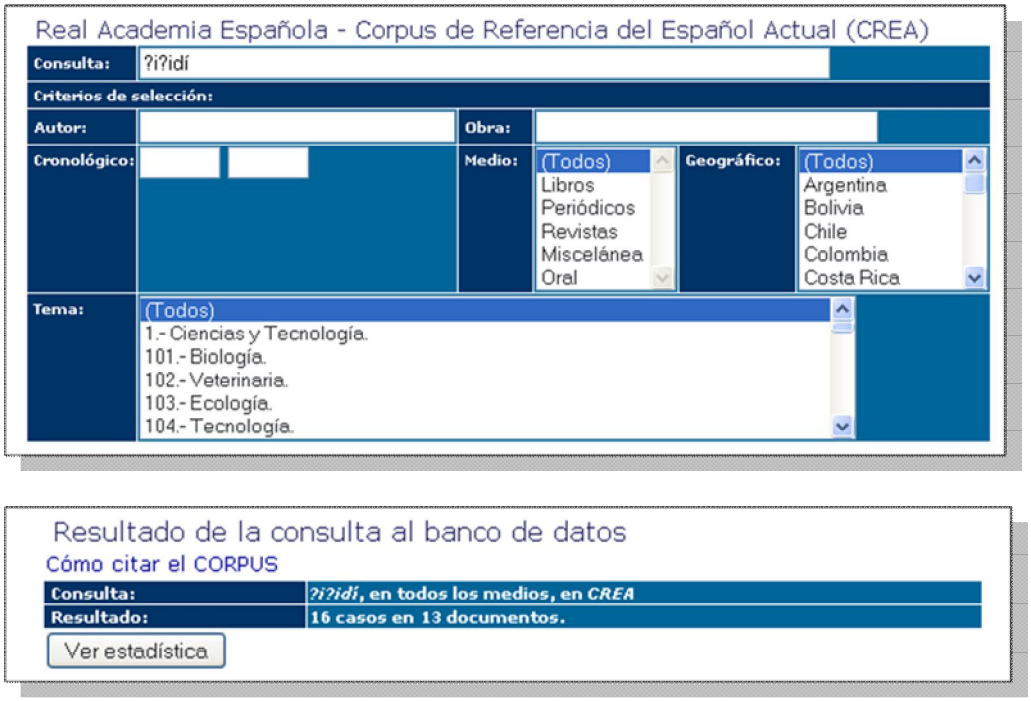


Ilustración 120. Resultados de la búsqueda de la forma "?i?idí" en el CREA.

Al recuperar las concordancias de estos 16 casos, descubrimos que, una vez excluidos los no pertinentes (*dividí* como forma del verbo *dividir*), en el CREA[245] hay cuatro casos de la variante *bividí* (ya que ni *bibidí*, la forma que buscamos, ni *bivirí*, la entrada que después encontramos en el *DRAE*, aparecen documentadas):

[245] Probamos también en el CORDE, pero sin éxito, lo que da cuenta de que la incorporación de esta palabra al caudal léxico del español ha debido de ser reciente.

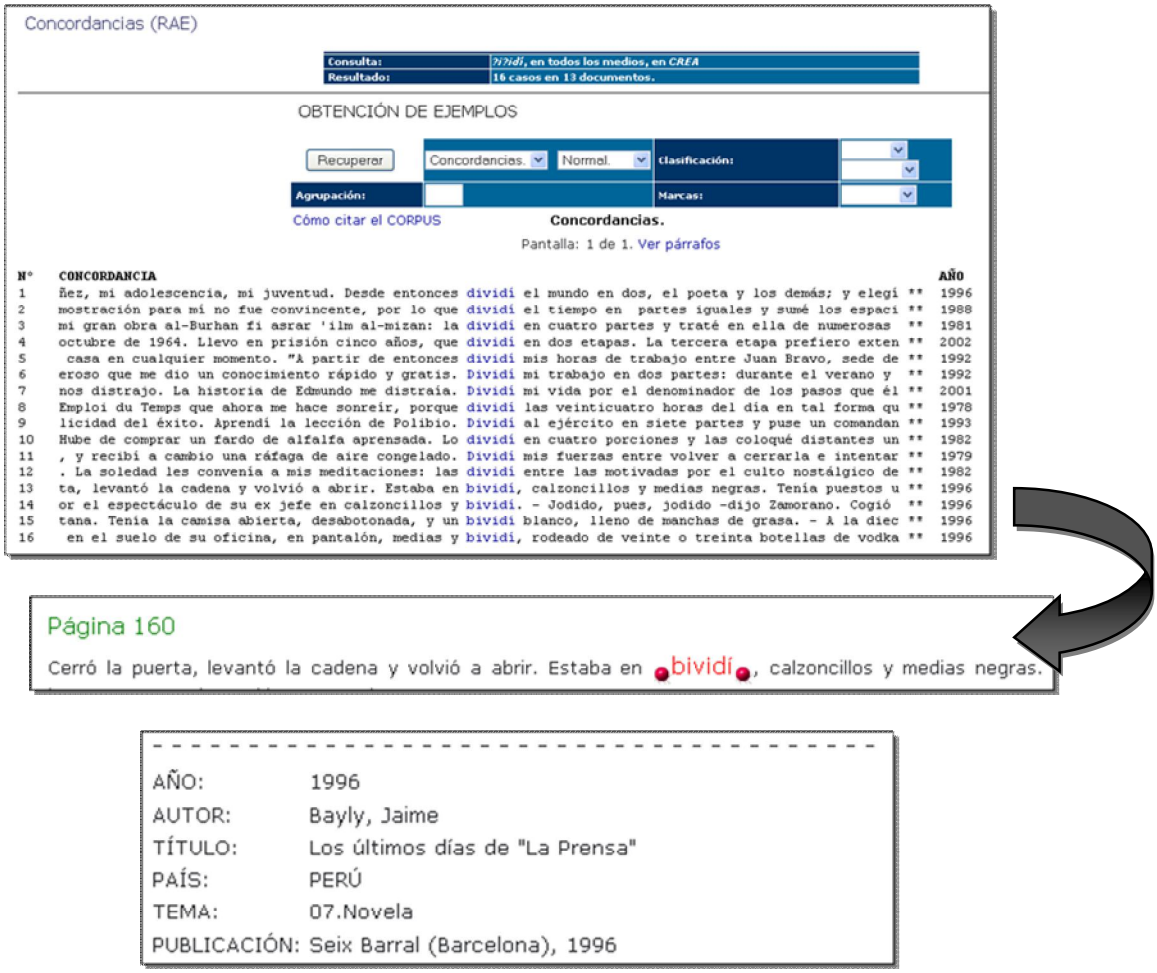


Ilustración 121. Concordancias correspondientes a la búsqueda de "?i?idí" en el CREA.

Los cuatro casos relevantes que mostraron las estadísticas aportan la siguiente información: documentan el uso del término en una época determinada, mediados de los noventa, pues pertenecen todos al año 1996; lo circunscriben a Perú (lo que concuerda con la marca geográfica de la entrada *bivirí* en el *DRAE*); y, por último, se refieren al ámbito de la ficción, pues todos ellos aparecen en la obra del autor peruano Jaime Bayly. Asimismo, nos proporcionan una aproximación a su significado: por el contexto (concordancias), esta forma parece referirse a una prenda de ropa, posiblemente de ropa interior:

- *"En bividí, calzoncillos y medias negras"*
- *"En calzoncillos y bividí"*
- *"Un bividí blanco"*
- *"Pantalón, medias y bividí"*

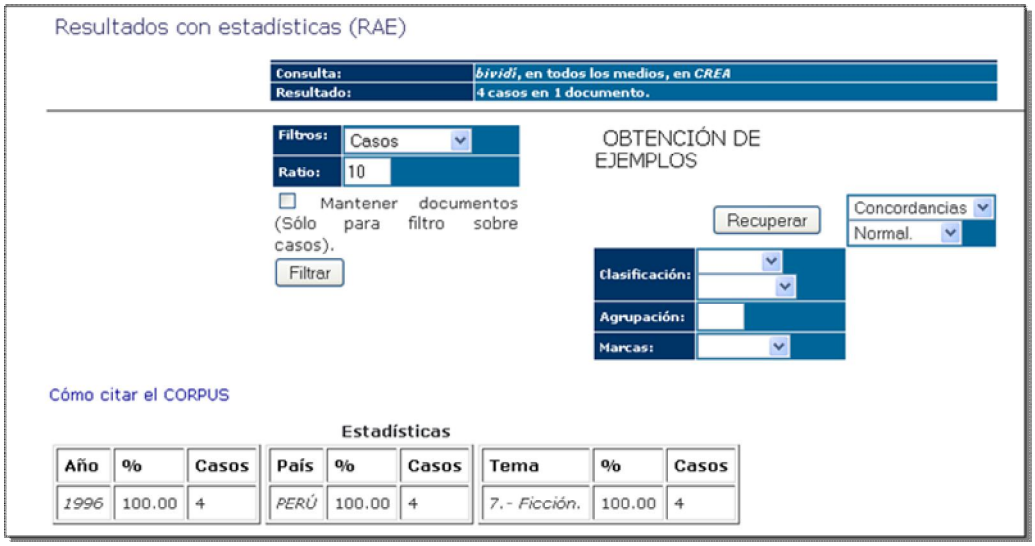


Ilustración 122. Estadísticas de la búsqueda de "bividí" en el CREA.

Con esta información inicial sobre sus variantes formales y su ámbito de uso, extendimos a Google[246] la búsqueda de las dos variantes localizadas, *bividí* y *bibidí*, ya que el significado no se deducía con claridad de los ejemplos del CREA. Hubo que descartar mucho "ruido", en especial con la forma *bibidí* debido a la existencia de un personaje homónimo del *anime* japonés *Dragon Ball Z*. Pero los datos extraídos del buscador nos permitieron dar con el origen de la palabra, obtener una definición clara del término, constatar su empleo en otros países diferentes de Perú y encontrar nuevas variantes formales.

[246] Que nos ofrece, además, la posibilidad de buscar imágenes, lo cual en este caso es particularmente interesante.

Por lo que respecta al origen del término, se trata de una palabra formada a partir de una marca registrada, que pertenece a una firma de ropa interior masculina creada en 1876 en Estados Unidos por Bradley, Voorhees y Day, de cuyas iniciales proviene el nombre (BVD):

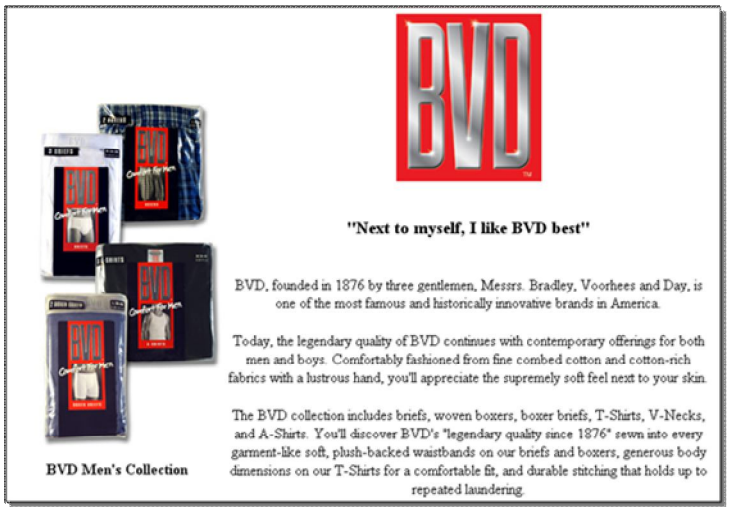


Ilustración 123. Captura de pantalla de la página web de la marca BVD[247].

Este nombre, *BVD*, es propio del inglés americano, de lo que da fe su inclusión en diversos diccionarios de esta lengua. Véanse dos ejemplos bien conocidos:

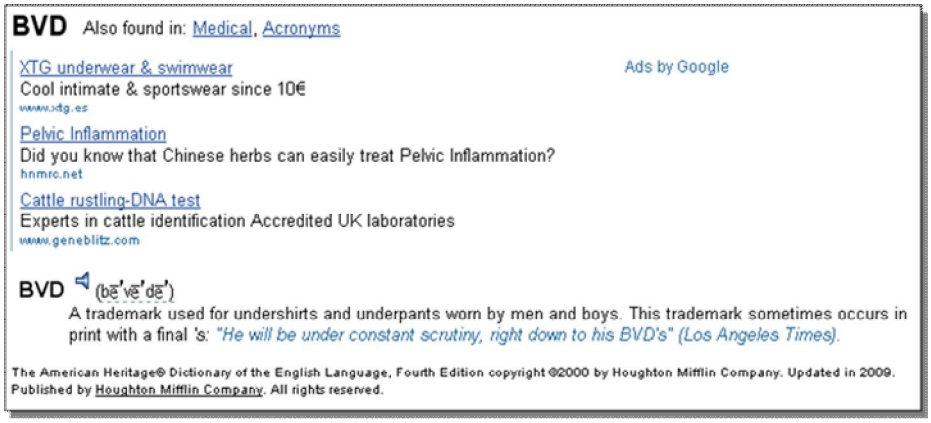


Ilustración 124. "BVD" en el American Heritage Dictionary of the English Language[248].

[247] URL: http://www.bvd.com/.

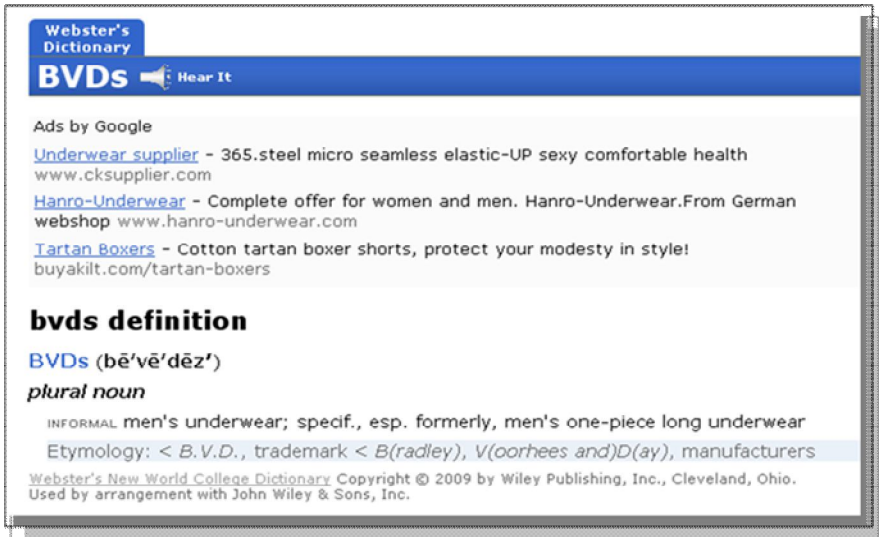


Ilustración 125. "BVD" en el Webster's Dictionary[249].

Como dato curioso, pero al mismo tiempo representativo de su importancia, llama la atención el hecho de que esta palabra está recogida en WordNet, la conocida base de datos léxica del inglés (*vid. supra*):

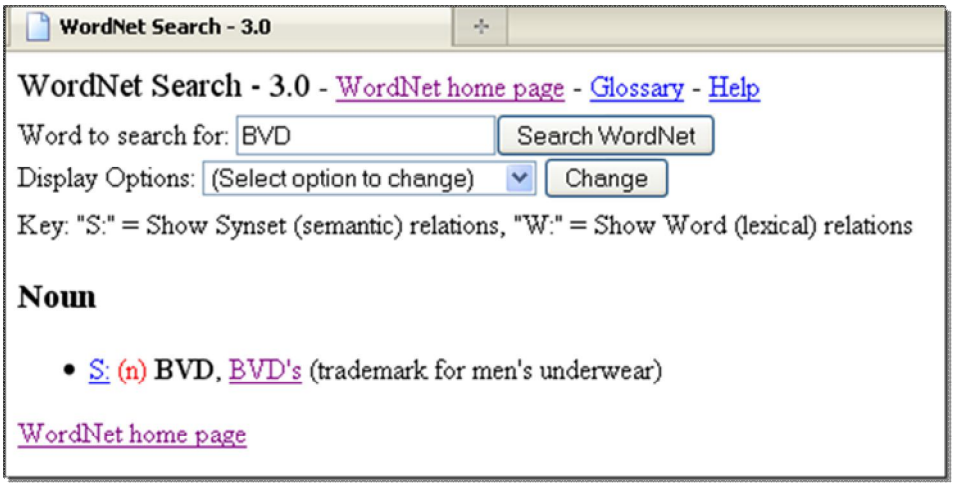


Ilustración 126. "BVD" en WordNet[250].

---

[248] URL: http://www.thefreedictionary.com/BVD
[249] URL: http://www.yourdictionary.com/bvds
[250] URL: http://wordnet.princeton.edu/. Acceso a la consulta en línea a través de la URL: http://wordnetweb.princeton.edu/perl/webwn.

El empleo del término en español documenta este origen, como el siguiente ejemplo, extraído de una página web[251], que recoge *bvd* como una variante de *bibidí*:

> Al Sr. Enrique \"No paro de meter cizana\" Martinez: seguramente Frank deberia encontrar una foto del recuerdo tuya de los anos 1988-1989 para comprobar que tus prendas favoritas eran: **andar en bvd** (o mejor conocidas bajo **bibidi**) y en jeans como los que turqueaba Vanilla Ice.

Cuando buscamos en Google definiciones en español de alguno de estos términos, las dos primeras apariciones son de WordReference, un conocido sitio de referencia para diccionarios de traducción: en concreto, el término *bividí* aparece en el diccionario español-inglés (que incluye la marca de uso Perú antes de proporcionar las traducciones al inglés, ¡de un término inglés!) y en un comentario de su foro, donde un hablante pregunta por el significado de la palabra inglesa *BVD*[252].

Ilustración 127. "BVD" en WordReference[253].

[251] Datos aportados por la estudiante Nuria Martínez. En la actualidad, la mayoría de estos textos no se pueden recuperar, debido a que estaban extraídos de blogs o páginas personales cuyos enlaces han caducado, uno de los inconvenientes que presenta el uso de Internet como banco de datos: la vida efímera de muchas de las páginas, sobre todo las de índole personal, aunque sean las que más proliferan en la actualidad.

[252] Los resultados son los mismos tanto si introducimos en el buscador *bividí* como *bibidí*.

[253] URL: http://www.wordreference.com.

Los datos obtenidos a través de Google también nos ofrecieron muestras de uso del término que sugieren una ampliación de su definición respecto a la que ofrece el *DRAE*, ya que se emplea no solo como prenda interior, y la usan tanto hombres como mujeres. Como botón de muestra, véase el resultado de buscar "bividí" en "Imágenes" de Google. Muchas de las imágenes pertenecen a firmas de venta de ropa que describen las características de sus productos (camisetas interiores, camisetas deportivas, camisetas de vestir…):

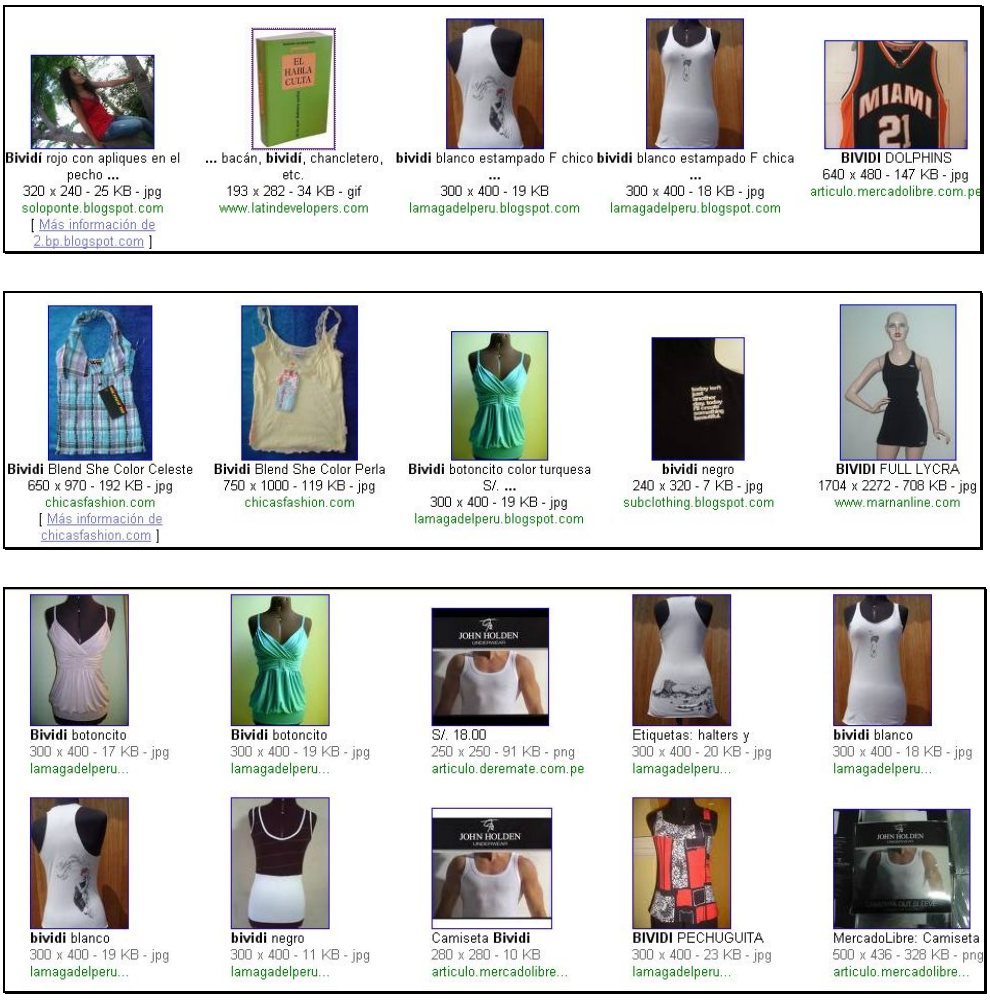


Ilustración 128. Algunos resultados de la búsqueda de "bividí" en Google Imágenes[254].

[254] URL: http://images.google.es.

Por otra parte, ya en su propio empleo en la variedad de inglés americano las siglas se han generalizado como un nombre común[255], empleo que se ha transferido al español, en concreto, al español hablado en Perú:

> En el español del Perú muchas marcas extranjeras ya registradas en el mercado internacional han llegado a hacerse nombre comunes: […] bividí (B.V.D.), cuya variante popular es bivirí atribuida a pronunciación de japoneses, frecuentemente empleados o dueños de lavanderías… (Hildebrandt 1994:155-158, *apud* Universidad de Piura PLANCAD 2001).

J. Calvo Pérez (2007:41) también comenta esta tendencia del español hablado en Perú a la generalización de marcas comerciales, citando la misma fuente: "En el español del Perú, existen diversas marcas ya registradas en el mercado internacional, que se han convertido en nombres comunes. Hildebrandt (1994, s.v. *cuáquer*), anota las siguientes", y recoge una enumeración de diferentes grupos (*id.*): "las perfectamente integradas", "las muy difundidas", "las usuales", "las anticuadas", "las modernas" y "las no totalmente integradas". *Bividí* está integrada en el grupo de las usuales y lleva, además, una nota aclaratoria sobre su inexistencia en España. Más adelante (*ibid.*:42), J. Calvo vuelve a incluir el término objeto de análisis en una lista de entradas "en que la marca se ha tomado como generalizador del

[255] De acuerdo con la respuesta que proporciona el *Daily Mail* (Londres) a una consulta de un lector a propósito del origen de este término que ha oído en algunas películas ambientadas en el Oeste (*QUESTION In some westerns, an old man is seen capering in a suit of long underwear that he calls his 'BVDs' or 'bivvy dees'. Where does this name come from?*), a principios del siglo XX la marca BVD dominaba de tal manera el mercado que el término pasó de designar un producto específico de ropa interior a ser el genérico para cualquier prenda de este tipo. Incluso, ya se había perdido entonces la conciencia de su origen en unas siglas (*cf.* URL: http://www.thefreedictionary.com/BVD, sección *References in periodicals archive*).

producto, perdiendo su especificidad y ganando en universalidad". Se indica adicionalmente que es un peruanismo[256].

A continuación, incluimos un fragmento del testimonio –que ya no se encuentra en la web– de un hablante de Lima en el que compara la palabra estudiada con la variante que se encuentra en otra zona del español (Santa Cruz de la Sierra, Bolivia), con plena conciencia de las diferencias léxicas entre su variedad del español y la que encuentra en este lugar de Bolivia, por otra parte, país limítrofe con Perú:

> Hay otros nombres también de cosas diferentes, de las que me acuerdo como achujcha es caigua, churikis es molleja, no se dice gaseosa sino soda, polera es polo, bibidi es camiseta, casaca es campera. Despues voy a apuntar más nombres.

Resulta curiosa la evolución de la lengua, que utiliza las iniciales de un nombre propio para formar las siglas de una marca comercial que, con el tiempo, se convierten en nombre común, tanto que puede llegar a ser el "apodo-apellido" de alguien y, así, volver a la esfera del nombre propio, como en la historia sobre Pedro Vivirí[257] que encontramos en el diario oficial del departamento de Cajamarca, en el norte de Perú:

> Pedro era un hombre de cierta edad, esa que se vuelve indefinida después que uno pasa los 40. Él vivía en Hualgayoc, un pueblo no muy lejos de aquí, Hualgayoc es un pueblo que amanecía frío y anochecía igual. Sus calles sin luz, apenas eran iluminadas por las luces tenues de

[256] Es interesante la entrada porque recoge todas las variantes que hemos localizado del término: *bividí*, *bibidí*, *bivirí*, *bibirí* y *BVD*.

> BIVIRÍ (bividí) ¶ (m.c.: BVD) [pr.] (camiseta {interior}), *bibirí (bibidí)*.

[257] URL: http://balconinterior.blogspot.com/2009/07/pedro-viviri.html.

lámparas de camiseta, llamadas Petromax, porque esa era su marca registrada y nadie decía una lámpara a kerosene, simplemente se decía Petromax. La marca se había convertido en un nombre, como sucede a veces con algunos apodos.

Pedro tenía una cantina y vendía aguardiente en las mañanas frías, en las tardes heladas y en las noches gélidas, sin embargo él tenía una rara afición, usaba esas (sospecho que solo era una) camisetas llamadas bivirí. Pedro usaba esas camisetas que alguna vez aparecieron en Norteamérica y se registraron bajo la marca BVD (bividi, según su pronunciación), marca registrada, como el frío lo era a Hualgayoc el nombre de ese tipo de camisetas se peruanizó en bivirí.

Camiseta interior masculina (como la denomina la Real Academia de la Lengua Española y le antepone Perú) se convirtió en la prenda favorita de Pedro. Cuando por las noches los parroquianos bebían sendas copas del licor en su cantina, Pedro atendía como si fuese un cantinero del oeste, sin piano, claro; sin chicas con faldas de encajes; sin pistoleros, pero sí con una pléyade de mineros que gastaban su semana y Pedro los atendía en bivirí.

Era una rareza, un hombre atendiendo en ese frío casi infinito llevando puesto una camiseta sin mangas, sin temblar siquiera. Los hualgayoquinos, dados a poner sobrenombres no tardaron en llamarlo "Pedro bivirí" y todos olvidaron su apellido, él mismo empezó a olvidarlo y muchas veces decía "Soy Pedro bivirí" como quien dice soy "Juan casaca" o "Luis Pantalón". Otros olvidaron hasta su nombre y simplemente lo llamaban "El bivirí".

Entonces, el pueblo asumió a bivirí como un nuevo apellido y su esposa era la señora de bivirí y sus hijos eran los bivirís y la calle era del bivirí y el pueblo era el pueblo del bivirí.

Pero la edad no pasa en vano y Pedro empezó a sufrir de achaques primero, ya no podía ponerse su bivirí de antaño y se metía bajo un poncho inmenso que lo abrigaba. La edad no perdona. Nadie sabe por

> qué cierto día empezó a perder la vista y se quedó ciego totalmente, ya no era ni la sombra de aquel cantinero rudo que atendía en la madrugada puesto el bivirí americano.
>
> Las viejas del pueblo entonces empezaron a especular, decían que el frío intenso producía ceguera, otras decían que era el bivirí, otras creían que vender cañazo causaba la atroz ceguera.
>
> Pedro Biviri murió una tarde cuando otra generación había tomado el pueblo, cuando casi nadie ya lo conocía porque él no podía ver a nadie y los demás no lo veían porque no querían ver a un viejo ciego del que nada sabían. Y hoy que es abril y han pasado años de tu muerte me acuerdo de ti, Pedrito, de tu camiseta con marca registrada y sin mangas, de tu última y vacía mirada, literalmente vacía como esta soledad que hoy me ha hecho recordarte.
>
> Tomado de "Hualgayoc, historia y Tragedia de un pueblo minero" (JAP)

Por lo tanto, parece que hay pruebas suficientes de que el término objeto de estudio, procedente del inglés (*BVD*), se introdujo en español por vía oral a través de Perú (peruanismo), mediante un proceso de adaptación de la pronunciación inglesa (*bividí-bibidí*) a la local (bien por "peruanización" de aquella o por deformación popular atribuida a los japoneses, de ahí *bivirí*). En este sentido, es interesante el siguiente comentario de un usuario que mantiene un blog sobre jerga peruana[258], puesto que, además de redundar en el origen de la palabra, registra la variante formal *bivirí*, que, por otra parte, sabemos que es la recogida en el *DRAE-01*:

---

258 URL: http://cultureando.blogspot.com/2007/06/leccion-3-religion.html. El término se encuentra también en otros sitios web que lo mencionan entre los peruanismos y americanismos del *DRAE*.

> bivirí. s. prenda de ropa interior masculina consistente en una camiseta escotada, sin mangas y sin cuello. Esta palabra viene de BVD, siglas de una marca registrada estadounidense de ropa interior para hombres, que se generalizó en el Perú en la década del cincuenta. En inglés estas éstas se leen /bividí/, lo que se ha adaptado como /bibirí/ en la pronunciación local.

Sin entrar a discutir las razones por las que se ha producido esta adaptación fónica, algunos de los textos encontrados dan muestras de una conciencia del hablante sobre el carácter normativo de la forma *bivirí*, pese a no ser la más frecuente en el uso, como comentaremos a continuación:

> Toquemos la cuestion fashion, agg que palabra más rosquetona. Dime que no has sufrid@ como cojud@ para bajar de peso y te entre esa blusita o puedas tú compadre, andar hecho un imbécil con tu bivirí (de BVD, así se escribe. no quiero responder a bestias que me corrigen sin saber) por la calle, ridículo de mierda[259].

> RESPONDERLE A ESTOS TRICICLEROS, ES ENSUCIARSE LA BOCA. FELIZMENTE QUE SON DEMASIADO BESTIAS PARA GANAR, YA LES DIJE QUE LES DEN GRACIAS A SUS AMOS ESPAÑOLES QUE LOS DEJARON BAILANDO YUNSA Y COMIENDO QUESO CON LLAMA, JAJAJAJA, PORQUE DE HABER SIDO ALEMANES O MIS ANTEPASADOS GALESES LA MAYORIA NO EXISTIRIA. PERO LO QUE MAS ME CALIENTA ES QUE SE EXPRESAN COMO SABELOTODOS RIDICULOS HIJOS DE PUTA (“ERUROPA”, “PAGINA WED”, “NGNORANTE”, “ALOGADO”

[259] Tomado del blog personal Memorias del Olvidado, de un peruano. URL: blog: http://mystic-place.blogspot.com/.

"OFENZA"ANTISIPACION", SEGURO QUE TAMBIEN DICEN "BIBIDI" JAJAJAJA TREMENDOS INDIASOS COQUEROS)[260].

Por lo que se puede observar, parece que hay hablantes para los que *bivirí* es la variante escrita que se debe usar, porque esa es la normativa, lo cual no quiere decir que sea la más usada. Además, como se colige del segundo texto, parece ser que para algunos hablantes *bibidí* es usada por personas occidentalizas o cholos y que, por lo tanto, dicho término no debe utilizarse. Sin embargo, los datos de uso de nuevo contradicen esta censura, ya que la forma *bibidí* aparece ampliamente documentada.

Si indagamos un poco más con Google, pronto encontramos muestras procedentes de otras geografías (España, Argentina, Estados Unidos…), con un papel destacado para Ecuador, expansión que parece lógica dada la proximidad de este último país a Perú. De hecho, la entrada de la Wikipedia[261] inglesa para *BVD* alude a su uso en ambos lugares: "In Ecuadorian and Peruvian Spanish, the term *bividi*, pronounced like the English initials, is an eponym for a man's sleeveless underwear T-shirt".

En el momento de realizar la búsqueda[262], se encontraron ejemplos claros en Ecuador y Argentina, y en otros lugares (México, Chile) donde resultaba más dudosa la adscripción geográfica. En cualquier caso, parece claro que el ámbito geográfico en que se emplea el término se ha ampliado. Sin embargo, es cierto que Perú es el país que arroja más casos, lo que parece volver a confirmar este como país de entrada del

[260] Del Centro de Medios independientes Perú. URL: http://peru.indymedia.org/.
[261] URL: http://en.wikipedia.org/wiki/BVD.
[262] El origen se determinó bien a partir de la extensión de las páginas web (.pe), bien a partir de información explícita sobre el país que aparecía en los sitios de Internet, aunque lógicamente no tenemos certeza de la procedencia del autor del texto.

término en español. Además, en Perú conviven *bibidí*, *bivirí, bividí* e, incluso, *bvd*.

En una rápida comparación de los resultados obtenidos por región en Google[263], destaca el empleo del término en Perú, Estados Unidos y España, por orden de casos, lugares donde además se registran las diferentes variantes formales, como se observa en el gráfico que sigue:

[263] Utilizando la "Búsqueda avanzada", opción "Región", que busca solo páginas ubicadas en la región especificada. A continuación resumimos los datos:

	Argentina	**Bolivia**	**Chile**	**Colombia**	**Costa Rica**
bibidí	434 (ruido)	2 (ruido)	268 (ruido)	56	2 (ruido)
bividí	336	1	5	1	0
	Cuba	**Ecuador**	**El Salvador**	**España**	**EE.UU.**
bibidí	0	452	3 (ruido)	1730 (ruido)	20400 (ruido)
bividí	0	148	0	1300	3600
	Honduras	**México**	**Nicaragua**	**Panamá**	**Paraguay**
bibidí	1 (ruido)	208 (ruido)	1 (ruido)	85 (ruido)	6 (ruido)
bividí	0	9	0	4	2
	Perú	**Puerto Rico**	**R. Dominic.**	**Uruguay**	**Venezuela**
bibidí	457	2 (ruido)	5 (ruido)	132 (ruido)	115 (ruido)
bividí	6030	0	0	0	4

Tabla 23. Datos para la búsqueda de "bibidí" y "bividí".

	Argentina	**Bolivia**	**Chile**	**Colombia**	**Costa Rica**
bibirí	80 (ruido)	0	3 (ruido)	4 (ruido)	0
bivirí	151	0	1	0	0
	Cuba	**Ecuador**	**El Salvador**	**España**	**EE.UU.**
bibirí	1 (ruido)	0	0	118 (ruido)	5050 (ruido)
bivirí	1	0	0	118	1860
	Honduras	**México**	**Nicaragua**	**Panamá**	**Paraguay**
bibirí	0	614 (ruido)	0	6 (ruido)	1 (ruido)
bivirí	0	7	0	2 (ruido)	0
	Perú	**Puerto Rico**	**R. Dominic.**	**Uruguay**	**Venezuela**
bibirí	114	1 (ruido)	0	0	1 (ruido)
bivirí	4220	0	0	2	4

Tabla 24. Datos para la búsqueda de "bibirí" y "bivirí".

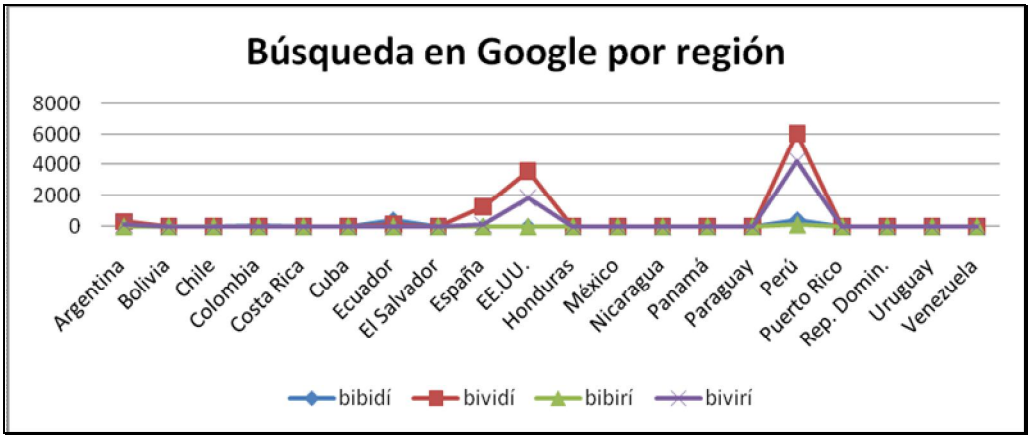


Ilustración 129. Resultados de la búsqueda por región en Google de "bibidí", "bividí", "bibirí" y "bivirí".

En segundo lugar, hay que comentar que, aunque inicialmente las variantes *bibidí* y *bibirí* arrojaban un número más elevado de casos –sobre todo la primera forma– que *bividí y bivirí*, hubo que descartar mucho ruido. En un análisis más profundo de dichos resultados, se detectó que los primeros (los de *bibidí* y *bibirí*) no se correspondían con el empleo del término para designar una camiseta –con las excepciones notables de Perú (457 casos), Ecuador (452 casos) y Colombia (56 casos)–. De estos tres países, en Ecuador (452-148 casos) y Colombia (56-1 casos) la variante *bibidí,* de hecho, supera a *bividí*[264].

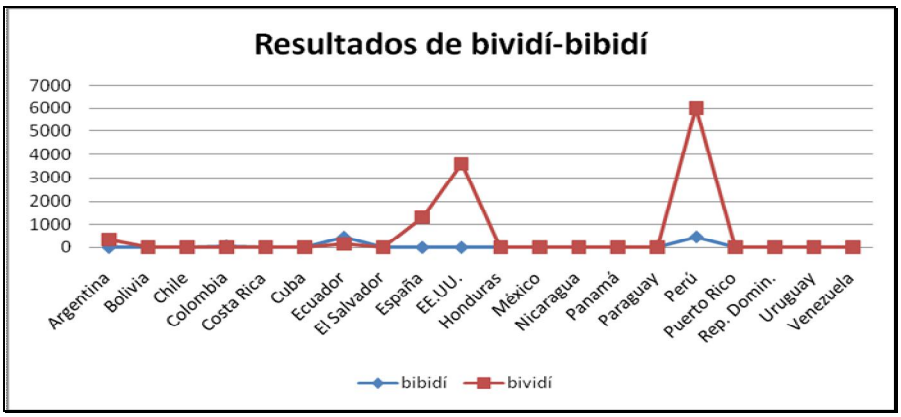


Ilustración 130. Comparación de los resultados de "bibidí" y "bividí".

[264] El tamaño reducido de la escala del gráfico impide apreciar claramente el caso de Colombia.

Por lo tanto, la forma *bividí*, la que más se ajusta a la pronunciación original inglesa (de *BVD*), es la que cuenta con más casos, pese a no ser la forma normativa recogida en el *DRAE*[265]. Perú (6030 casos) encabeza la lista de registros, seguida de Estados Unidos (3600 casos), España (1300 casos), Argentina (336 casos) y Ecuador (148 casos). La propagación del término parece obedecer a una doble tendencia: por una parte, hacia el norte, de la zona andina al Caribe continental (de Perú a Ecuador y Colombia) y, por otra, hacia lugares tan distantes como Estados Unidos, España y Argentina, en los que la proximidad geográfica no puede ser el criterio, sino la emigración. En Ecuador y Colombia con predominio, además, de la variante *bibidí*, frente a la forma más extendida en el resto de lugares, *bividí*.

---

[265] Recordemos que es *bivirí*. Este hecho resulta doblemente llamativo: por un lado, no es la forma más empleada, tal y como se desprende de los datos aportados –y ni siquiera la documentada en los bancos de datos académicos (que es precisamente *bividí*)–; por otro lado, contradice la doctrina académica a propósito de términos que siguen un patrón muy similar, como *DVD* o *CD*. Si bien para estas siglas, a partir de las cuales se han creado también nombres comunes, la forma de entrada en español parece ser la lengua escrita, en ambas se documentan pronunciaciones "a la inglesa" ([dividí], [sidí]) en América, que la RAE, a través del *Diccionario panhispánico de dudas*, desaconseja, igual que las grafías que se corresponden con sus lecturas inglesas (*dividí* y *cidí* respectivamente):

> **DVD**. […] En español debe leerse [deubedé] o [debedé], dependiendo del nombre con que se denomine la letra *v* (→ v, 1); se desaconseja la pronunciación ⊗[dividí], propia del inglés, a pesar de su extensión en algunas zonas de América. […] A partir de la lectura española de la sigla, se han creado los sustantivos *devedé* (pl. *devedés*), en América, y *deuvedé* (pl. *deuvedés*), en España: *«En los devedés, memorables representaciones operísticas»* (Glantz *Rastro* [Méx. 2002]); *«Yo podría haber vivido con Eduardo comprando deuvedés»* (Gopegui *Lado* [Esp. 2004] 133); se desaconseja la forma ⊗*dividí,* por corresponder a la lectura inglesa de la sigla.
>
> **CD**. […] En español debe leerse [sedé, zedé]; se desaconseja la pronunciación ⊗[sidí], propia del inglés, a pesar de su extensión en algunas zonas de América. […] A partir de la lectura española de la sigla se ha creado el sustantivo *cedé* (pl. *cedés*): *«En las tiendas ya se vendían cedés con canciones sobre el tema»* (PzReverte *Reina* [Esp. 2002]). Se desaconseja la forma ⊗*cidí*, por corresponder a la lectura inglesa de la sigla. En cualquier caso, se recomienda usar con preferencia el equivalente español *(disco) compacto.*

La forma normativa *bivirí* sigue el mismo patrón de distribución de los datos que *bividí*, aunque con un recuento menor de muestras en todos los países: en primer lugar, destaca de nuevo Perú (4220 casos), seguido de Estados Unidos (1860 casos), Argentina (151 casos) y España (118 casos). Además, Perú es el único país donde alterna con *bibirí* (114 casos), aunque a una distancia considerable de la otra variable. En el resto de países ha habido que descartar las apariciones de *bibirí* por no ser pertinentes para la búsqueda.

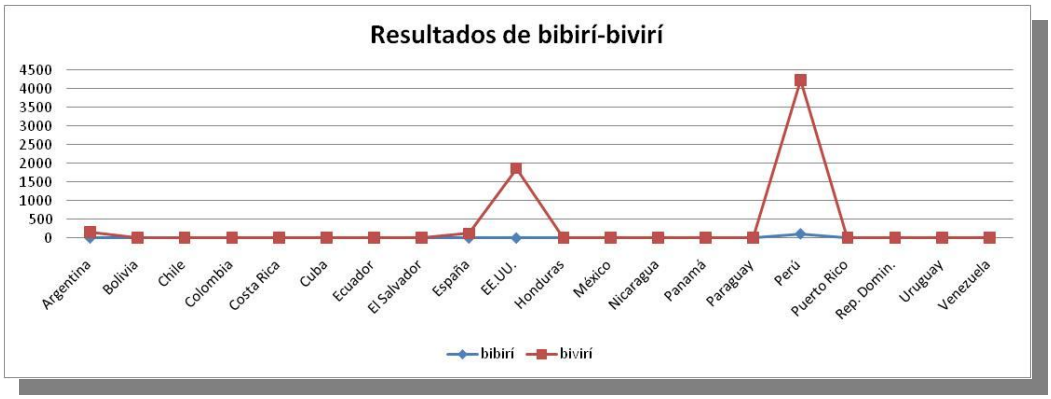


Ilustración 131. Comparación de los resultados de "bibirí" y "bivirí".

Por lo tanto, de forma global, se observa que las dos variantes más frecuentes, *bividí* y *bivirí*, se encuentran concentradas en los mismos lugares –excepto en Ecuador, donde no hay casos de la segunda– con una marcada supremacía de la primera.

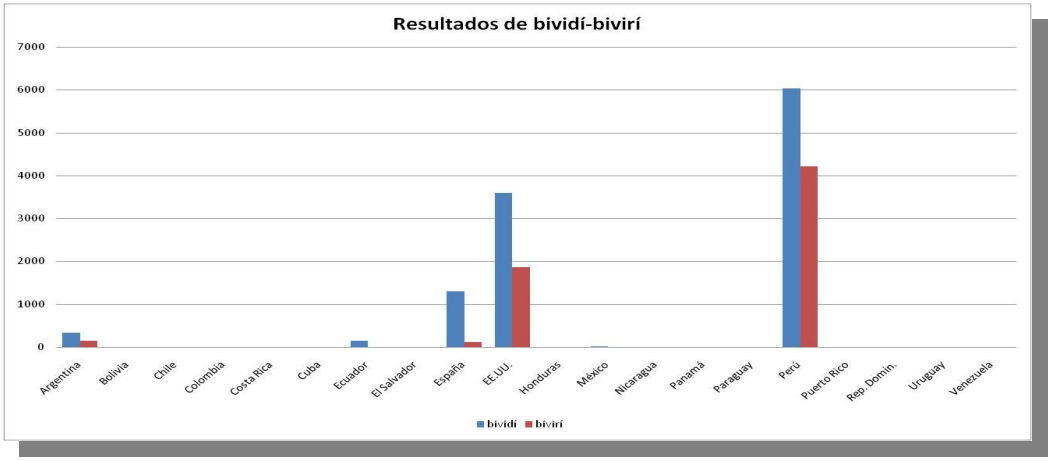


Ilustración 132. Comparación de los resultados de "bividí" y "bivirí".

En resumen, con este pequeño ejemplo, hemos querido demostrar que las posibilidades que ofrecen para la investigación los corpus son apasionantes.

Los inconvenientes de utilizar la web como corpus se pueden resumir en los siguientes puntos:

1) Vida efímera de los textos en la web: muchas páginas desaparecen en un lapso relativamente breve de tiempo, por lo que ni siquiera pueden ser citadas.
2) Caducidad de los enlaces, sobre todo los de las páginas de índole personal (*blogs*), que son las que más proliferación tienen en la actualidad.
3) Falta de información sobre los textos: autores anónimos, ausencia de datación, etc., factores todos ellos que no permiten identificar con certeza la procedencia de las páginas.
4) No aval institucional, lo que puede disminuir la relevancia de los datos obtenidos y arrojar dudas sobre su fiabilidad, al no estar respaldados por ningún tipo de autoridad.
5) Ausencia de planificación y control sobre la selección de los textos (páginas), que no obedecen a unos criterios lingüísticos previamente establecidos, por lo que no existe garantía de que sean representativos de la lengua objeto de estudio.

Sin embargo, no todo son obstáculos; el uso de la web como corpus implica ventajas, en especial en lo que se refiere al número ingente de muestras reales de lengua que proporciona y a las diferentes posibilidades para realizar búsquedas que ofrece: por región, mediante imágenes, por dominio (término site:pe, por ejemplo), en noticias, en vídeos, en libros, en definiciones (define: término), con comodines, etc.

3.4. Clasificación de los corpus

Establecidos la metodología y el concepto de corpus, así como su ejemplificación con un caso concreto, en este apartado pasamos a comentar algunos de los principales tipos de corpus, ya que no todos son iguales. Autores como J. Sinclair (1996) o J. Torruella y J. Llisterri (1999) han propuesto clasificaciones de los distintos tipos de corpus en función de una serie de criterios, aunque en la práctica no siempre está clara ni se hace explícita la tipología de un corpus.

En general, los principales parámetros para clasificar los corpus se centran en:

- La modalidad de la lengua
- El número de lenguas a que pertenecen los textos
- El tamaño o cantidad de textos que conforman el corpus
- Los límites del corpus
- La variedad lingüística o el grado de especialización de los textos
- El período temporal que abarcan los textos
- El tratamiento aplicado al corpus

Con frecuencia, estos criterios vienen determinados por la finalidad u objetivo que se persigue con el corpus: el estudio de la obra de un autor (Cervantes) o de la producción literaria de una época determinada (el Barroco), la descripción de una lengua en general (el español contemporáneo) o de una variedad, sublenguaje o aspecto lingüístico concreto (p. ej. la norma culta en México, el lenguaje de los partes meteorológicos, el léxico jurídico, etc.), la obtención de un

determinado producto comercial (un diccionario, una aplicación telefónica relacionada con las tecnologías del habla, etc.).

Teniendo en cuenta los criterios y consideraciones anteriores, se puede establecer la siguiente tipología:

1) Según la modalidad de la lengua, se distinguen tres tipos de corpus: corpus escritos, corpus orales y corpus mixtos.

- Los *corpus textuales* o *escritos* están conformados exclusivamente por muestras de lengua escrita. Es el caso, por ejemplo del *Corpus Textual Informatitzat de la Llengua Catalana* (CTILC)[266].
- Los *corpus orales*, por su parte, únicamente recogen muestras de lengua hablada, que pueden ser:
    - Transcripciones ortográficas de grabaciones (*corpus de lengua oral*), utilizadas sobre todo en lingüística de corpus para obtener una representación simbólica de una muestra natural de habla. Ocasionalmente se añade información prosódica, pero no se accede a la señal sonora más que para transcribir los textos. El objetivo no es tanto el análisis de las características de tipo fonético, sino contar con una transcripción ortográfica de la lengua hablada. Esta transcripción constituye el punto de partida para el tratamiento posterior del corpus (añadir marcas sobre categorías gramaticales, extraer índices de frecuencia, etc.) y para efectuar diferentes análisis lingüísticos: sociolingüísticos, discursivos, etc. Tratan de reflejar, sobre todo, la variación con textos representativos de los distintos usos de la lengua hablada, por lo que las grabaciones se realizan en entornos

[266] URL: http://ctilc.iec.cat/

naturales y se favorecen las muestras espontáneas, no planificadas, aunque no son las únicas recogidas (diálogos, conversaciones, discursos, grabaciones procedentes de medios de comunicación, etc.). *The Bergen Corpus of London Teenage Language* (COLT)[267] es un corpus de medio millón de palabras conformado por las transcripciones ortográficas de conversaciones espontáneas. Su objetivo fundamental es dar cuenta de una variedad de lengua, la de los adolescentes de Londres y, por tanto, servir como punto de referencia para estudios de índole lingüística (marcadores pragmáticos y discursivos, vocabulario típico, estudios sociolingüísticos, etc.). Para el español, podemos mencionar un proyecto de características similares, el *Corpus Oral de Lenguaje Adolescente* (COLA)[268], el *Corpus de Conversación Coloquial* del Grupo Val.Es.Co[269] o el *Corpus Oral de Referencia de la Lengua Española Contemporánea* (CORLEC)[270]. También hay que destacar el proyecto PRESEA[271] para la creación de un corpus representativo de las variedades geográficas y sociales del español.

[267] El corpus, compilado en 1993 en la Universidad de Bergen, Noruega, pretende dar cuenta de la variedad de lengua oral inglesa empleada por adolescentes (entre 13 y 17 años) de Londres. En la actualidad es un componente del *British National Corpus* (BNC). URL: http://www.hf.uib.no/i/Engelsk/COLT/index.html

[268] URL: http://www.colam.org/om_prosj-espannol.html. Emprendido en la misma Universidad de Bergen, con el objetivo de dar cuenta del habla de los jóvenes (entre 13 y 19 años) de Madrid y de otras capitales latinoamericanas. Su finalización está prevista en julio de 2010. El proyecto está en relación con COLT y con UNO, también llevado a cabo en Bergen y centrado en el lenguaje juvenil en los países nórdicos. URL: http://www.uib.no/uno/unoEng/

[269] URL: http://www.uv.es/~valesco/

[270] URL: http://www.lllf.uam.es/corpus/corpus_lee.html#A. El proyecto, dirigido por F. Marcos Marín en el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid, se realizó entre 1991 y 1992.

[271] URL: http://80.38.130.7/Default.aspx?alias=80.38.130.7/portalpreseea

- Grabaciones (*corpus orales*), empleadas en fonética y tecnologías del habla, que conforman el punto de partida de los análisis extraídos del corpus. Estos corpus suelen realizarse en entornos controlados y estar formados por enunciados o palabras, pero no habla espontánea. La transcripción de los textos es fonética y ortográfica, alineada con la señal sonora.

  Los corpus orales orientados hacia la descripción fonética de las lenguas suelen consistir en inventarios de sistemas fonéticos y fonológicos de las lenguas del mundo a modo de bases de datos de sonidos; o en grabaciones realizadas en condiciones óptimas de segmentos aislados, frases aisladas o textos leídos. En general, se diseñan con mucho cuidado para recoger el fenómeno objeto de estudio y tienen un tamaño reducido, al no utilizar un número elevado de hablantes. También pueden incluir habla espontánea e, incluso, grabaciones de medios de comunicación. En ocasiones, recogen materiales equivalentes en diferentes lenguas para estudios contrastivos y de interferencia fonética en la adquisición de lenguas.

  Los corpus orales que se compilan para el desarrollo de sistemas en el ámbito de las tecnologías del habla (*vid.* Llisterri *et al.* 2005) consisten en inventarios de unidades de síntesis para convertir texto a habla (síntesis del habla): cada grafía se relaciona con una unidad de síntesis y posteriormente se unen para producir la onda sonora; en grabaciones con unidades fonéticas o con elementos específicos, como números de teléfono o de tarjetas de crédito para los sistemas de reconocimiento del habla; en

transcripciones (fonéticas y ortográficas) de grabaciones de lengua oral con información lingüística añadida que se utilizan para elaborar modelos estadísticos del lenguaje; o en grabaciones y transcripciones de diálogos naturales entre personas o entre personas y simulaciones de sistemas informáticos que se emplean para desarrollar servicios automáticos a través del teléfono (venta de entradas, consulta de horarios de transportes públicos, servicios bancarios, etc.).

Ejemplos de corpus orales son *Albayzín* (*vid.* Casacuberta *et al.* 1992), gran base de datos oral desarrollada en España, entre 1992 y 1998, por un consorcio de grupos de investigación en tecnología del habla coordinado por la Universidad Politécnica de Cataluña. Además de los objetivos relacionados directamente con la síntesis y el reconocimiento del habla, también se recopiló con vistas al desarrollo de estudios fonéticos sobre la variabilidad inter- e intra- locutor, la variabilidad contextual y la variabilidad condicionada por las condiciones ambientales. También destaca el proyecto *EUROM* (*vid*. Chan *et al.* 1995), base de datos oral multilingüe, en la que las grabaciones se llevaron a cabo bajo las mismas condiciones, con el mismo número de sujetos y un corpus equivalente para once lenguas de nuestro entorno. En el caso del proyecto *SpeechDat*[272], el objetivo era el desarrollo de "teleservicios" (servicios de información, de transacciones, correo hablado, centralitas…), sistemas de ayuda a la conducción mediante el habla, recursos para el entrenamiento de sistemas de reconocimiento del habla, etc. Un último ejemplo es *Gaudí* (*cf.* Battaner *et al.* 2005), corpus

[272] URL: http://www.speechdat.org/

para la identificación y verificación de hablantes, desarrollado entre la Escuela Universitaria de Ingenieros de Telecomunicaciones de la Universidad Politécnica de Madrid y el Servicio de Policía Judicial de la Dirección General de la Guardia Civil.

- Los *corpus mixtos* combinan ambas modalidades de lengua, aunque siempre favoreciendo la lengua escrita, ya que su obtención es menos costosa que la de la lengua oral que, además, requiere un proceso posterior de transcripción de las grabaciones. El *Corpus de Referencia del Español Actual* (CREA)[273] o el *British National Corpus* (BNC)[274] pertenecen a este tipo de corpus: el 90% de sus textos son escritos y el 10% restante, orales.

2) Según el número de lenguas, los corpus se clasifican fundamentalmente en monolingües y bilingües o multilingües.

- Los *corpus monolingües* están compuestos por textos en una sola lengua. Se recopilan con el objetivo de dar cuenta de dicha lengua o variedad lingüística (o de un subconjunto de la misma). Es el caso del CREA (para el español), del CORGA[275] (para el gallego), etc.
- Los *corpus bilingües* o *multilingües* están formados por textos de dos (bilingües) o más lenguas (multilingües) sin que, en principio, sean traducciones unos de otros y sin compartir criterios de selección. No obstante, este tipo de corpus son raros; son más habituales los corpus de dos o más lenguas que

[273] URL: http://www.rae.es/
[274] URL: http://www.natcorp.ox.ac.uk/
[275] URL: http://corpus.cirp.es/corga/

contienen textos elegidos según unos mismos criterios o que son traducciones mutuas:

- *Corpus comparables* ("*paired texts*"): consisten en un conjunto de textos en más de una lengua o variedad lingüística, parecidos en cuanto a sus características y que comparten criterios de selección. Se utilizan sobre todo para comparar variedades de la lengua en estudios contrastivos. El ejemplo más destacado es el *International Corpus of English* (ICE)[276], un corpus en el que desde 1990 se están recopilando materiales escritos y orales posteriores a 1989 pertenecientes a diferentes variedades del inglés a lo largo del mundo. En la actualidad están en marcha veinte proyectos en otros tantos países, desde Australia hasta Estados Unidos, pasando por Jamaica, Nueva Zelanda o Pakistán: ICE-GB (inglés de Gran Bretaña), ICE-NZ (inglés de Nueva Zelanda), ICE-IRE (inglés de Irlanda), ICE-PHI (inglés de Filipinas), etc. Cada corpus, de un millón de palabras, consta de quinientas muestras (orales y escritas) de dos mil palabras cada una. Todos siguen el mismo esquema de diseño y de anotación. Otro ejemplo de corpus de este tipo es *C-Oral-Rom*[277], un corpus multilingüe de habla espontánea de cuatro lenguas romances (italiano, francés, portugués y español).

- *Corpus paralelos* ("*bi-texts*"): recogen textos en más de una lengua (bilingües o multilingües) pero, a diferencia de los anteriores, se trata del mismo texto y sus traducciones o equivalentes en una o más lenguas. El más sencillo consta del original y su traducción. Son especialmente útiles para

[276] URL: http://ice-corpora.net/ice/
[277] URL: http://lablita.dit.unifi.it/coralrom/

los estudios de traducción, para el desarrollo de sistemas de traducción automática y en entornos bilingües o multilingües, como la ONU, la OTAN, la UE o el parlamento de Canadá, en los que los documentos deben publicarse obligatoriamente en todas las lenguas oficiales. Desde un punto de vista metodológico, son discutidos por algunos autores, ya que se pueden producir interferencias entre las lenguas objeto de traducción. Se remontan a la Edad Media, cuando se hacían "biblias políglotas", que contenían textos uno al lado de otro en hebreo, latín y griego, y a veces también versiones vernáculas. Un ejemplo muy conocido es el *Hansard Corpus*[278], con textos en inglés y en francés (en su variedad canadiense) procedentes de las actas de las sesiones del parlamento canadiense. Otro ejemplo de corpus paralelo es el *Corpus Lingüístico da Universidade de Vigo* (CLUVI)[279], de unos veintitrés millones de palabras, elaborado en el Seminario de Lingüística Informática de la Universidad de Vigo bajo la dirección de Xavier Gómez Guinovart.

- *Corpus alineados*: son corpus paralelos en los que, para facilitar su explotación, los textos están dispuestos unos al lado de otros por párrafos o frases, de tal forma que sea más fácil extraer las

[278] URL: <www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20>

[279] URL: http://sli.uvigo.es/CLUVI/. Compuesto a su vez por diferentes subcorpus:

- *Corpus literario TECTRA inglés-galego* (1.476.020 palabras)
- *Corpus literario FEGA francés-galego* (1.648.272 palabras)
- *Corpus xurídico LEGA galego-español* (6.582.415 palabras)
- *Corpus UNESCO inglés-galego-francés-español de divulgación científica* (3.724.620 palabras)
- *Corpus LOGALIZA de localización de software inglés-galego* (3.526.850 palabras)
- *Corpus CONSUMER español-galego-catalán-euskara* de información sobre consumo (5.586.431 palabras)

> equivalencias de traducción: aquellos elementos que son traducciones mutuas. Aunque no siempre es un proceso simple, el alineamiento de oraciones y palabras se puede conseguir automática o semiautomáticamente con un alto grado de exactitud. Se utilizan, sobre todo, como entrenamiento para sistemas de traducción automática basados en estadísticas o en la docencia sobre traducción. El CLUVI también ilustra perfectamente este tipo de corpus.

3) Según la cantidad, la proporción y la distribución de los tipos de textos, se habla de:

- *Corpus grandes*: no tienen un límite de palabras o este es muy elevado en comparación con otros tipos de corpus; no suelen atender a cuestiones de equilibrio o de representatividad. Cada vez es mayor la tendencia al aumento de volumen gracias a los medios y facilidades técnicas disponibles; no obstante, en la actualidad existen corpus de gran tamaño diseñados con criterios que garantizan la representatividad de los datos.
- *Corpus equilibrados*: recogen la misma proporción de diferentes tipos de textos.
- *Corpus piramidales*: contienen textos distribuidos en estratos o niveles, de tal forma que un nivel consta de pocas variedades temáticas pero con muchos textos para cada una; un segundo nivel, de textos más variados temáticamente, pero con menos cantidad de cada uno; etc.

- *Corpus léxicos* ("*sample corpus*"): recogen fragmentos de textos muy pequeños y de longitud constante en cada documento. Era lo habitual en los primeros corpus, debido a las limitaciones de tamaño que los medios técnicos de la época imponían. Hoy en día han vuelto a cobrar importancia debido a lo cuidado de su diseño[280].

4) Según los límites establecidos, los corpus se clasifican en corpus cerrados y corpus abiertos o monitor.

- Los *corpus cerrados* constan de un número finito de palabras, que se establece de forma previa a la recopilación del corpus. Una vez alcanzado ese número, el corpus se da por finalizado, sin añadir más material posteriormente. Es lo que ocurrió, por ejemplo, con el corpus *Brown*[281]. Este tipo de corpus son útiles cuando interesa estudiar fenómenos estáticos o estados de lengua.
- Los *corpus abiertos* o *corpus monitor*, por el contrario, son corpus dinámicos, que se mantienen en constante crecimiento, normalmente mediante la introducción periódica de nuevas cantidades de textos según unas proporciones previamente definidas. Cuando la capacidad de almacenamiento no lo permitía, se iban retirando los textos más antiguos a medida que se introducían los nuevos. Son un material excelente para los estudios diacrónicos, para observar tendencias de uso, cambios de significado, frecuencias de distribución, etc. No obstante, no están exentos de críticas frente al modelo predominante de corpus, basado en una concepción estática (tamaño finito) y más

[280] Se oponen a corpus formados por textos enteros.
[281] URL: http://icame.uib.no/brown/bcm.html

preocupado por ser equilibrado en cuanto a sus muestras. En cambio, el modelo del corpus monitor suele centrarse en alcanzar un tamaño considerable y prefiere incluir textos enteros en vez de simples muestras. Es el caso del *Bank of English*[282].

5) Según la especificidad de los textos, los corpus pueden ser generales o especializados; también genéricos y canónicos.

- Los *corpus generales* o *de referencia* pretenden reflejar la lengua o variedad lingüística de la forma más equilibrada posible; cuantos más tipos de textos, modalidades (textos orales, textos escritos), géneros y materias, mejor. Por este motivo también tienen que ser lo suficientemente amplios como para reflejar todas las variedades relevantes de una lengua y su vocabulario, de forma que se puedan tomar como base para la elaboración de gramáticas, diccionarios, tesauros, etc. El CREA sería un ejemplo de corpus de este tipo.
- Los *corpus especializados* recogen textos que puedan aportar datos para la descripción de un tipo particular de lengua ("sublenguaje"). P. ej. el *Corpus Técnico do Galego* (CTG)[283] del Seminario de Lingüística Informática de la Universidad de Vigo, que contiene textos jurídico-administrativos, de informática y telecomunicaciones, de ecología y ciencias ambientales, de economía, de sociología y de medicina. O el *Corpus textual especializado plurilingüe*[284], proyecto desarrollado por el Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra, que consta de textos en catalán, castellano, inglés, francés y alemán sobre economía, derecho, medio ambiente, medicina e

[282] URL: http://www.titania.bham.ac.uk/docs/about.htm
[283] URL: http://sli.uvigo.es/CTG/
[284] URL: http://www.iula.upf.edu/corpus/corpuses.htm

informática, con la meta de estudiar cómo funciona la lengua en cada una de esas áreas y extraer información útil para detectar neologismos, elaborar diccionarios y tesauros, estudiar la variación lingüística, etc.

- Los *corpus genéricos* recogen textos pertenecientes a un único género, ya que el objetivo es caracterizar ese género frente a otros. Por ejemplo, el *York-Helsinki Parsed Corpus of Old English Poetry*[285], que contiene solo poesía.
- *Corpus canónicos*: están formados por todos los textos que configuran la obra completa de un autor.

6) Según el periodo temporal que abarcan los textos, las principales tipologías de corpus que encontramos son:

- Los *corpus periódicos* o *cronológicos* recogen textos de unos años determinados o de unas épocas concretas con el objeto de estudiar la lengua producida durante ese período, como en los casos de los corpus *Brown* o LOB, que recogen textos publicados exclusivamente en 1961 en Estados Unidos y el Reino Unido respectivamente.
- Los *corpus diacrónicos* o *históricos* incluyen textos de diferentes etapas temporales sucesivas con el fin de poder observar evoluciones de la lengua en un período largo, lo que los diferencia de los corpus monitor, que no abarcan períodos temporales tan amplios. Para el español, por ejemplo, además del CORDE, destaca el *Corpus del español*[286], un corpus de cien

[285] URL: http://www-users.york.ac.uk/~lang18/pcorpus.html
[286] URL: http://www.corpusdelespanol.org/

millones de palabras recopilado por Mark Davis en la Universidad de *Brigham Young*, y que contiene textos en español desde el siglo XIII hasta el XX.

- *Corpus sincrónicos*: su finalidad es permitir el estudio de una o más variedades lingüísticas en el momento presente, sin prestar atención a su evolución excepto en lo que se refiere a los cambios rápidos que ocurren en la actualidad. Es el caso del *Corpus of Contemporary American English*, de más de trescientos ochenta y cinco millones de palabras procedentes de textos de diferentes fuentes de los años 1990 a 2008[287].

7) Según el proceso al que se someta el corpus, se distingue entre:

- *Corpus simples*, en bruto, no anotados o no codificados: consisten en textos guardados sin formato alguno y sin añadir ningún tipo de información adicional, como pueden ser códigos o anotaciones. Un corpus así ofrece unas posibilidades muy limitadas para los estudios lingüísticos.
- *Corpus verticales*: son el resultado de disponer en forma de columna las palabras de un texto ordenadas según criterios alfabéticos o de frecuencia. Las palabras se consideran aisladamente, sin contexto.

[287] Compilado por Mark Davies en la *Brigham Young University*. URL: http://www.americancorpus.org/

Orden	Frec.absoluta	Frec.normalizada
1. de	9,999,518	65545.55
2. la	6,277,560	41148.59
3. que	4,681,839	30688.85
4. el	4,569,652	29953.48
5. en	4,234,281	27755.16
6. y	4,180,279	27401.19
7. a	3,260,939	21375.03
8. los	2,618,657	17164.95
9. se	2,022,514	13257.31
10. del	1,857,225	12173.87

Ilustración 133. Lista de frecuencias. 10 formas más frecuentes en el CREA.

- *Corpus codificados o anotados*: están formados por textos a los que se les han añadido, de forma manual o automática, determinadas informaciones. Estas pueden referirse a datos bibliográficos o a la estructura de los textos: etiquetas especiales para indicar el autor, el título, los capítulos, los párrafos, etc. (*codificación*); o, lo que es más interesante, a aspectos puramente lingüísticos, como la categoría gramatical, la estructura sintáctica, etc. (*anotación*). La explicitación de estos datos enriquece los corpus y aumenta considerablemente las posibilidades de explotación que ofrecen.
    - *Corpus analizados morfológicamente* ("tagged"): los textos del corpus han sido anotados con información morfológica. Cada palabra del corpus tiene asociada una lista de sus posibles categorías morfosintácticas. Es posible incluir más o menos detalles en este apartado (nombre, verbo; nombre común, nombre propio, verbo principal, verbo auxiliar, etc.). La mayoría de corpus hoy en día cuenta con este tipo de anotación, que se inserta mediante un sistema de códigos al lado de cada palabra.

<s>Before<w CS> we<w PPIS2> begin<w VV0> this<w DD1> morning<w NNT1> formally<w RR> with<w IW> our<w APPGE> agenda<w NN1> ,<w ,> I<w PPIS1> 'd<w VM> like<w VVI> to<w TO> take<w VVI> just<w RR> one<w MC1> minute<w NNT1> to<w TO> welcome<w VVI> you<w PPY> all<w DB> and<w CC> say<w VVI> that<w CST> this<w DD1> is<w VBZ> wonderful<w JJ> that<w CST> you<w PPY> 're<w VBR> all<w DB> here<w RL> .<w .> </s> <s>I<w PPIS1> think<w VV0>

Ilustración 134. Muestra de texto etiquetado del "Corpus of Spoken, Professional American-English"[288].

- *Corpus "parentizados"*: son aquellos que se han sometido a un proceso de análisis sintáctico superficial, marcado entre paréntesis o corchetes. Normalmente se identifican los constituyentes principales: por ejemplo SN (sintagma nominal), SV (sintagma verbal), etc. Un ejemplo es el *Lancaster Parsed Corpus* (LPC)[289], que representa un subconjunto del LOB de unas ciento cuarenta mil palabras que han sido analizadas sintácticamente.
- *Corpus analizados* ("*treebanks*"): los textos que los conforman están procesados sintácticamente de manera completa. Cada oración del corpus ha sido analizada de forma exhaustiva: p. ej. SN sujeto animado. Cada vez son más habituales este tipo de corpus. Destaca la *Base de Datos Sintácticos del Español Actual* (BDS)[290] o, más recientemente, los corpus CESS-ECE[291] para el español, el catalán y el euskera, y *AnCora*[292], para el español y el catalán.

[288] URL: http://www.athel.com/cpsa.html. El texto anotado se corresponde con el siguiente texto sin anotar:

> *Before we begin this morning formally with our agenda, I'd like to take just one minute to welcome you all and say that this is wonderful that you're all here.*

[289] URL: http://khnt.hit.uib.no/icame/manuals/LPC/LPC.PDF
[290] URL: http://www.bds.usc.es/
[291] URL: http://clic.ub.edu/cessece/index.php
[292] URL: http://clic.ub.edu/ancora/index.php

3.5. El desarrollo de un corpus (I): diseño y constitución

La constitución del corpus se refiere al proceso mediante el cual se fijan los criterios que han de guiar el diseño del corpus y, de acuerdo con ellos, se recopilan los textos.

La selección de los textos que formarán parte de un corpus (*cf.* Sinclair 1996) se puede efectuar según criterios internos o criterios externos.

3.5.1. Criterios internos o lingüísticos

Están basados en la distribución de palabras o de características gramaticales. Se centran en la aparición de patrones o elementos diferenciadores de la variedad lingüística de un texto, como p. ej. la longitud de las oraciones. Dan lugar a los llamados *tipos de texto*. A menudo se han propuesto como la mejor solución para elaborar corpus representativos: mientras más palabras o rasgos lingüísticos incluya el corpus, más representativo será de la lengua o variedad objeto de estudio. Sin embargo, también han recibido críticas relativas a su cientificidad, por no presentar independencia respecto de los objetivos de la investigación: los textos se seleccionan de forma intuitiva, buscando la presencia de determinados rasgos. Destacan entre estos criterios:

- *Tema*: dominio o ámbito al que pertenece el texto. Es un parámetro que clasifica un texto a partir de su contenido. Corresponde prototípicamente a los criterios internos. Se trata de un tipo de clasificación que ha sido muy criticada, al estar basada en descripciones y clasificaciones de las áreas de conocimiento

que han de reducir forzosamente una realidad compleja a una estructura jerárquica monodimensional. Es decir, el problema es el subjetivismo que representa una división arbitraria del saber humano. Sin embargo, en la práctica es un tipo de criterio muy utilizado, especialmente en los corpus más extensos. Por ejemplo, es uno de los criterios que se manejan en CREA para elegir textos. Así, se clasifican estos en seis hipercampos (que agrupan, a su vez, un gran número de áreas temáticas), en el caso de los textos informativos, y un séptimo hipercampo para los textos de ficción:

- *Hipercampo 1. Ciencias y tecnología.*
- *Hipercampo 2. Ciencias sociales, creencias y pensamiento.*
- *Hipercampo 3. Política, economía, comercio y finanzas.*
- *Hipercampo 4. Artes.*
- *Hipercampo 5. Ocio y vida cotidiana.*
- *Hipercampo 6. Salud.*
- *Hipercampo 7. Ficción: Novela, relatos y teatro.*

Véase también la lista de temas que utiliza el BNC para seleccionar sus textos:

Table 3. Written Domain

	texts	w-units	%	s-units	%
Imaginative	476	16496420	18.75	1352150	27.10
Informative: natural & pure science	146	3821902	4.34	183384	3.67
Informative: applied science	370	7174152	8.15	356662	7.15
Informative: social science	526	14025537	15.94	698218	13.99
Informative: world affairs	483	17244534	19.60	798503	16.00
Informative: commerce & finance	295	7341163	8.34	382374	7.66
Informative: arts	261	6574857	7.47	321140	6.43
Informative: belief & thought	146	3037533	3.45	151283	3.03
Informative: leisure	438	12237834	13.91	744490	14.92

Ilustración 135. Campos temáticos del British National Corpus para textos escritos.

- *Estilo*: tipo de modelo de lengua que sigue. Es lo que ocurre, por ejemplo en el corpus C-ORAL-ROM, que subdivide los registros de habla espontánea teniendo en cuenta el estilo dialógico (*vid.* Campillos, Gozalo y Moreno 2007):

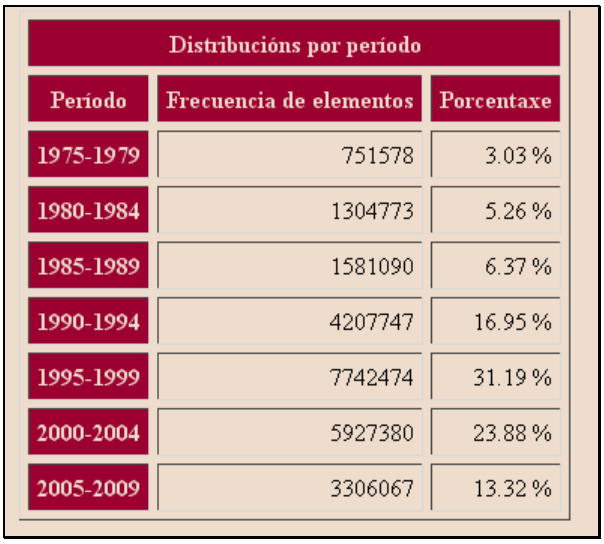
Tipo	*Subtipo*	*Sub-subtipo*
Informal	Privado	Diálogo y conversación
Informal	Privado	Monólogo
Informal	Público	Diálogo y conversación
Informal	Público	Monólogo
Formal	Contexto Natural	Discurso político, debate político, sermones, discurso de enseñanza, charla profesional, conferencias, discurso de negocios y discurso jurídico
Formal	Medios de comunicación	Programas de entrevistas, prensa científica, reportajes, entrevistas, deportes, noticias, pronósticos del tiempo
	Teléfono[1]	Conversación humana, interacción humano-máquina

Ilustración 136. Estilos en el corpus C-ORAL-ROM.

3.5.2. Criterios externos o situacionales

Estos criterios no tienen en cuenta características internas presentes en los textos, sino cuestiones relacionadas con el entorno de los mismos. En la actualidad se suelen preferir a los criterios internos, ya que aportan independencia: los textos se eligen de forma objetiva, sin hacer predicciones sobre los aspectos lingüísticos que van a presentar. Además, son el tipo de criterios que se utilizaron en los primeros corpus y que más difusión han tenido. Destacan:

- *Cronología*: uno de los datos más objetivos que se puede dar sobre un texto es su fecha de elaboración. Para la gran mayoría de los textos se puede determinar fácilmente con mucha precisión cuál es el año o período cronológico que corresponde a su datación. Esto hace que la fecha sea una de las informaciones primarias con que se caracterizan los textos que forman parte de un corpus.



Distribucións por período		
Período	Frecuencia de elementos	Porcentaxe
1975-1979	751578	3.03 %
1980-1984	1304773	5.26 %
1985-1989	1581090	6.37 %
1990-1994	4207747	16.95 %
1995-1999	7742474	31.19 %
2000-2004	5927380	23.88 %
2005-2009	3306067	13.32 %

Ilustración 137. Periodos en que se estructuran los textos de CORGA.

- *Origen*: considera aspectos de la producción del texto que puedan afectar a la estructura o al contenido; datos diversos sobre el autor o los autores, el editor, el lugar de publicación, etc. Por ejemplo, uno de los criterios para elegir textos escritos que sigue el BNC es el nombre del autor, su edad, sexo, etc. Se trata de otro de los criterios más empleados en los corpus.
- *Estado*: cuestiones relativas al aspecto físico del texto y a su soporte en el momento en que se selecciona para el corpus; modo de transmisión (oral, escrito o electrónico), etc.

- *Objetivo*: tiene en cuenta la motivación del texto y las finalidades que persigue; el tipo de audiencia o de público al que se dirige (niños, adolescentes, adultos, todo tipo de audiencia), resultados que se espera obtener de su difusión, etc.
- *Género literario*: este parámetro se ha aplicado sobre todo a los corpus que contienen un número importante de textos literarios. Esta clasificación suele seguir la división tradicional en cuatro géneros básicos: ensayo, narrativa, poesía y teatro, como ocurre en la parte literaria del *Corpus Textual Informatitzat de la Llengua Catalana*. En la actualidad, no es uno de los parámetros más importantes para el diseño de corpus, ya que precisamente las obras literarias o de ficción tienden a tener un menor peso frente al material de tipo informativo, que se considera que representa mejor el uso mayoritario de la lengua.
- *Medio de publicación*: los parámetros centrados en el medio de publicación del texto son uno de los criterios externos más utilizados hoy en día. Este hecho obedece a la tendencia que manifiesta preferencia por los criterios externos, ya que no presentan el subjetivismo inherente a los criterios internos. El carácter limitado (pese a las posibilidades de variedad actuales) del número de medios de difusión textual y el que se trate de características no inherentes al texto, facilitan un alto grado de homogeneidad. P. ej. en el caso de un corpus de textos escritos, se distinguirían con facilidad los siguientes grupos: libros, publicaciones periódicas, miscelánea (folletos publicitarios o de otro tipo, catálogos…), material escrito no publicado (correspondencia, diarios…), textos electrónicos, material escrito para ser leído (discursos políticos, guiones para programas de radio o televisión, etc.).

Las últimas tendencias parecen inclinarse por compaginar criterios externos e internos:

Criterios externos	Criterios internos
a) Cronología b) Origen c) Estado d) Objetivo e) Género literario f) Medio de publicación	a) Tema b) Estilo

Tabla 25. Criterios para la selección de textos.

Por ejemplo, el BNC (*British National Corpus*) se guio por el tema (criterio interno), el medio de publicación y la fecha (criterios externos) para elegir los textos escritos que iban a conformar el corpus; el CREA utiliza criterios muy similares: tema como criterio interno y medio de publicación, fecha y, además, procedencia geográfica como criterios externos.

3.5.3. Otras cuestiones de diseño

Además de los criterios que se van a seguir a la hora de recopilar el corpus, la primera fase en el desarrollo de un corpus también implica decisiones relativas a:

- Finalidad
- Límites temporales, geográficos y lingüísticos
- Tamaño y tipo de textos
- Proporciones temáticas

En cuanto a la finalidad del corpus, es el objetivo o motivo por el que se lleva a cabo la recopilación. Puede ser una finalidad muy concreta (un corpus para un servicio automático de telefonía) o más general (un corpus para estudiar el funcionamiento de la lengua española). Mientras más amplio sea el objetivo o finalidad, más complicado es el diseño. También se consideran en este momento inicial cuestiones sobre la posible reutilización del corpus (lexicografía, procesamiento del lenguaje natural, fonética, etc.). Asimismo, hay que destacar que la finalidad del corpus condiciona muchas de las decisiones posteriores: no es lo mismo un corpus que va a servir como base para la elaboración de materiales de referencia (gramáticas, diccionarios) que otro cuya finalidad principal es el diseño de programas de procesamiento del lenguaje natural (correctores ortográficos, etiquetadores…).

Decidida la finalidad, el siguiente paso es fijar los límites temporales, geográficos y lingüísticos de los textos que contendrá el corpus, así como la proporción de los mismos para cada período, zona o variedad, lo que lógicamente vendrá dictado por la finalidad.

Los límites temporales se refieren al período que han de comprender los textos. Algunos ejemplos son:

Corpus of Contemporary Spanish	español posterior a 1990
Corpus Textual Vox-Biblograf	español posterior a 1950
Corpus Textual Informatitzat de la Llengua Catalana	catalán posterior a 1833
Longman/Lancaster English Language Corpus	inglés posterior a 1899

Tabla 26. Ejemplos de distribuciones temporales.

Los límites geográficos comprenden las zonas y porcentajes de cada una que deben integrar el corpus. Por ejemplo:

Corpus de Referencia del Español Actual	50% de español de España 50% de español de América
Corpus del Español del Siglo XXI	30% de español de España 70% de español de América
Corpus of Contemporary Spanish	25% de español peninsular 25% de español de Argentina 50% de español de otras zonas de América del Sur
Bank of English (proyecto COBUILD)	70% de inglés británico 20% de inglés americano 10% de inglés de otras zonas
Longman/Lancaster English Language Corpus	50% de inglés británico 40% de inglés americano 10% de inglés de otras zonas

Tabla 27. Ejemplos de distribuciones geográficas.

Los límites lingüísticos hacen referencia a las lenguas o variedades lingüísticas a las que deben pertenecer los textos: español, inglés, gallego, catalán, etc. No es lo mismo un corpus monolingüe que uno bilingüe o multilingüe. Así, CRATER (*Corpus Resources and Terminology Extraction Project*) contiene textos en inglés, francés y español; y el *Corpus textual especializado plurilingüe* del Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra de Barcelona, está conformado por textos en cinco lenguas diferentes (catalán, castellano, inglés, francés y alemán).

En cuanto al tamaño y tipo de textos del corpus, aunque los corpus de gran volumen están de moda, hay que recordar que el tamaño es una cuestión relativa, condicionada por la finalidad del corpus. Además del tamaño del corpus en general, también se debe decidir en esta fase sobre la tipología textual de la que se nutrirá el corpus (tipos de textos y su distribución, así como el tamaño de las muestras):

- Tipo de textos: generales o específicos.
- Textos concretos que integrarán el corpus: una lista que dé cuenta de los documentos que conformarán el corpus (nómina).
- Cantidad de texto que se tomará de cada documento para las muestras: textos íntegros o fragmentos; si son fragmentos de textos, la parte inicial, la central o la final de un texto, etc.

El incluir textos enteros:

(i) Facilita el estudio de un amplio abanico de aspectos, en especial en los estudios de lingüística textual.

(ii) Permite recortar los textos posteriormente.

(iii) Se enfrenta con problemas de derechos de autor, sobre todo en el caso de textos recientes.

(iv) Conlleva la presencia de textos de tamaño desigual.

(v) Tiene el peligro de que se tomen como usos generales peculiaridades estilísticas o temáticas de un autor o ámbito.

Los fragmentos, en cambio, parecen ser preferibles pues:

(i) Favorecen el trabajo con textos de un tamaño constante, lo que dota de mayor equilibrio al corpus.

(ii) Además, los rasgos lingüísticos más frecuentes son constantes en su distribución, por lo que una muestra de dos mil palabras es suficiente, según han demostrado diversos experimentos; solo en el caso de buscar rasgos muy raros es necesario acudir a muestras más amplias.

(iii) Se recomienda que haya equilibrio entre los fragmentos tomados del inicio, de la mitad y del final de los textos.

Algunos ejemplos de la política de los corpus en cuanto a este aspecto son:

Corpus of Contemporary Spanish	Muestras textuales de 70.000 palabras
Longman/Lancaster English Language Corpus	Muestras textuales de 40.000 palabras
International Corpus of English (ICE)	Muestras textuales de 2.000 palabras
Brown Corpus of American English	Muestras textuales de 2.000 palabras
Lancaster-Oslo/Bergen Corpus (LOB)	Muestras textuales de 2.000 palabras
Bank of English (proyecto COBUILD)	Muestras textuales de textos enteros

Tabla 28. Ejemplos de tamaños de muestras.

El siguiente paso aborda cuestiones relativas a la proporción y el número de muestras para cada categoría textual. Es especialmente importante cuando se aplica una técnica de muestreo estratificado (*vid. infra*) para seleccionar los textos. No obstante, entraña cierta dificultad, pues no existe unanimidad sobre los grupos temáticos (arte, deporte, etc.) que deben establecerse o sobre el porcentaje adecuado de cada

uno. Además, los temas se expresan de diferente forma (género): prosa, debate, etc., por lo que también deben tomarse decisiones sobre los géneros que abarcará el corpus. P. ej.

- *Brown Corpus* y *LOB Corpus*: muestras de 15 géneros elegidas al azar.
- *Longman Lancaster English Language Corpus*: muestras elegidas siguiendo el criterio de los libros más leídos en las bibliotecas, pertenecientes a 10 grupos temáticos:

1. Ciencias puras y naturales	6%
2. Ciencias aplicadas	4,3%
3. Ciencias sociales	14,1%
4. Cuestiones mundiales	10,4%
5. Comercio y finanzas	4,4%
6. Artes	7,9%
7. Creencias y pensamientos	4,7%
8. Pasatiempos	5,7%
9. Ficción	40%
10. Poesía, teatro y humor	2,3%

Tabla 29. Ejemplo de proporciones temáticas.

- *CORGA* contiene la siguiente proporción de temas:

Distribucións por área temática principal		
Área temática	Frecuencia de elementos	Porcentaxe
Economía e política	5693850	22.94 %
Cultura e artes	1836747	7.40 %
Ciencias sociais	3284502	13.23 %
Ciencia e tecnoloxía	1622688	6.54 %
Ficción	10048912	40.49 %
Outros	2334410	9.40 %

Ilustración 138. Proporciones temáticas en CORGA.

3.5.4. Representatividad del corpus y muestreo

Decididas las cuestiones anteriores, se pasa a obtener los textos que van a integrar el corpus. Para efectuar la selección de los mismos, es necesario aplicar una serie de principios estadísticos que garanticen que las muestras, a partir de las cuales se va a efectuar una generalización sobre la lengua, son *representativas* de la *población* (en este caso, la población es la lengua o variedad lingüística en cuestión). De hecho, hoy en día, la representatividad es la principal cuestión a la hora de compilar un corpus, característica que, además, lo diferencia de archivos o simples colecciones de textos.

La representatividad se refiere a que, puesto que, salvo en casos particulares –obras de un autor determinado, lenguas muertas, etc.–, es imposible recoger en un corpus todas las muestras de una lengua, se hace preciso elegir un subconjunto de ellas que presente las características de la lengua como un todo.

Un corpus trata, por definición, de mostrar a pequeña escala cómo funciona el todo que es una lengua natural; pero para ello es necesario que esté diseñado sobre unas bases estadísticas apropiadas que aseguren que el resultado sea efectivamente un modelo de la realidad.

A menudo, en Lingüística no estamos interesados en un texto o autor individual, sino en toda la variedad de la lengua. Entonces tenemos dos opciones para reunir datos (*cf.* McEnery, Xiao y Tono 2006:13 y ss.):

a) Analizar cada enunciado de esa variedad. Opción impracticable salvo las excepciones ya señaladas.

b) Seleccionar una pequeña parte de esa variedad, tomar una muestra. Opción más realista.

Una de las críticas de Chomsky al acercamiento al estudio del lenguaje mediante corpus era que la lengua es infinita y, por lo tanto, cualquier corpus será sesgado: algunos enunciados no aparecerían en el corpus por ser raros; otros más comunes podrían quedar excluidos por casualidad; o también se podría dar el caso de que enunciados raros aparecieran en el corpus dando una falsa imagen de la realidad. Por lo tanto, en el diseño de corpus se deben establecer medidas para garantizar que el corpus no ofrece una visión parcial de la realidad que pretende reflejar.

Sin embargo, la representatividad no es una cuestión sencilla y es objeto de debates dentro de la lingüística de corpus: ¿hasta qué punto un corpus es representativo de la realidad lingüística que quiere describir? ¿En qué medida se pueden generalizar los descubrimientos realizados a partir de la muestra? ¿Hay diseños de corpus más representativos que otros? ¿Cuáles son los criterios que determinan esta representatividad?

La respuesta es que la representatividad es una noción relativa**.** En términos generales, se puede decir que el factor que determina la representatividad es la relación que existe entre el diseño de un corpus (criterios para seleccionar los textos) y las finalidades que se han previsto como objetivo fundamental de su explotación. De esta forma, el hecho de que haya corpus más o menos extensos en cuanto al número de palabras, más o menos restringidos en el parámetro cronológico, de unas u otras proporciones de tipologías textuales, obedece a las diferencias sobre el tipo de resultados que se espera obtener.

En concreto, en términos prácticos, la representatividad de un corpus respecto a la lengua o variedad que tiene como referente está en función del equilibrio entre las diferentes categorías o tipologías textuales

(textos escritos, orales, generales, específicos…) que, tomadas juntas, puedan considerarse el "promedio" y proporcionar una imagen razonablemente exacta de toda la población en que se esté interesado. Evidentemente, tanto representatividad como equilibrio son cuestiones relativas, pues ambas dependen de la finalidad del corpus. Por este motivo, a menudo se siguen las pautas marcadas por corpus ya existentes, aceptados como modelos de representatividad por la comunidad investigadora. Es lo que ocurre, p. ej. con el BNC, cuyo diseño está fundamentado sobre los siguientes parámetros:

- 90% de textos escritos, elegidos según los criterios de
    - dominio o campo
    - tiempo o período
    - medio de publicación
- 10% de textos orales transcritos, elegidos según los criterios de
    - demografía (sexo, edad, clase social), para los textos informales
    - contexto, para los textos formales

Por otra parte, para asegurar la representatividad de un corpus también hay que tener en cuenta los cambios debidos al tiempo, de ahí la necesidad de efectuar actualizaciones. Se distingue en este sentido entre corpus estáticos y corpus dinámicos:

- Los llamados en inglés "sample corpus" (corpus léxicos) emplean el mismo diseño para estudiar la misma variedad en diferentes etapas de la lengua. Es lo que sucede con el LOB para el inglés británico de los años 60 y el Freiburg-LOB para el de los

90. En cada caso el corpus presenta una imagen estática del estado de la lengua en un período determinado.

- Los corpus monitor son modelos dinámicos, pensados sobre todo para detectar cambios rápidos en la lengua, cambios que suceden en un período corto de tiempo, como p. ej. el estudio de los neologismos.
- Los corpus diacrónicos, aunque modelos estáticos en su concepción (cada subcomponente representa un período amplio), se emplean, en cambio, para estudiar períodos más amplios, para observar la evolución de la lengua a lo largo del tiempo.

Por otra parte, atendiendo a las categorías textuales, no es lo mismo un corpus general que uno especializado. El primero, al tener como objetivo proporcionar una imagen lo más completa de la lengua, suele contener una amplia gama de géneros, mientras que el segundo se limita a un dominio (p. ej. el Derecho) o a un género (p. ej. el periodístico). Sin embargo, ambos tipos de corpus deberán mantener cierta proporcionalidad entre los tipos textuales que pretenden representar (p. ej. un corpus especializado en Derecho deberá contener en la debida proporción textos legales, resoluciones judiciales, etc.).

No obstante, hay que señalar que el equilibrio entre los diferentes tipos de textos es más importante para los corpus estáticos que para los monitor, más interesados por lograr un tamaño considerable.

En cualquier caso, es aconsejable que se documenten de forma explícita los criterios que se han seguido a la hora de diseñar el corpus[293].

Una vez consideradas y decididas las cuestiones anteriores, se está en disposición de empezar a compilar el corpus. En este momento entran en juego las técnicas de muestreo, técnicas de base estadística cuya finalidad es asegurar que los textos que van a integrar el corpus contendrán las características objeto de estudio. Para ello, es necesario:

1º) Definir la unidad de muestreo (p. ej. periódicos, libros, conversaciones…), la población o conjunto de unidades (p. ej. periódicos publicados en España) y el marco de muestra o lista de unidades (p. ej. *El País*, *ABC, Público, El Mundo*…).

2º) Aplicar una técnica de muestreo:

- Muestreo aleatorio simple, en el que las unidades se numeran y se eligen mediante una tabla de números aleatorios. El inconveniente de esta técnica es que el investigador carece de control, por lo que elementos raros, aquellos que tienen poca frecuencia, pueden quedar fuera de la muestra, o estas pueden no recoger determinados aspectos que pueden ser decisivos, como el sexo, la edad, la procedencia geográfica, etc.
- Muestreo aleatorio estratificado, que divide la población en grupos homogéneos, según las pautas marcadas por el investigador, y luego toma muestras aleatorias de cada grupo establecido.

[293] MCENERY, XIAO y TONO (2006:13 y ss.) ejemplifican con detalle estas cuestiones y comentan los debates suscitados por el tema de la representatividad.

3.6. El desarrollo de un corpus (II): codificación y anotación

Un corpus tal y como se ha descrito en los apartados anteriores es un corpus "crudo" ("raw corpus" en inglés) o corpus formado únicamente por las muestras textuales seleccionadas. Este tipo de corpus, aunque útil para analizar determinados aspectos, es limitado en cuanto a las posibilidades que ofrece al estudioso de la lengua. Por ello, son preferibles los corpus codificados y anotados, es decir, corpus en los que se ha explicitado información lingüística (y no lingüística) presente ya antes –de forma implícita– en el corpus, que se ha añadido a los textos. P. ej., en una oración como "El lujo es la cultura" se podría explicitar la información relativa a la categoría y características gramaticales de los elementos que la constituyen:

El_DMS lujo_NMS es_VIP3S la_DFS cultura_NFS

Etiqueta	Información
DMS	Determinante Masculino Singular
NMS	Nombre Masculino Singular
VIP3S	Verbo Indicativo Presente 3ª Persona Singular
DFS	Determinante Femenino Singular
NFS	Nombre Femenino Singular

Tabla 30. Clave de las etiquetas.

Este proceso de enriquecimiento de un corpus, que en muchos casos se puede llevar a cabo de forma totalmente automática con un porcentaje muy elevado de éxito, aumenta el potencial de los corpus

para efectuar investigaciones sobre diversos aspectos del lenguaje. Además, permite contextualizar los textos y agruparlos por variables (género, autor,…), hecho fundamental para la posterior explotación del corpus: toda la información explícita se puede recuperar.

En concreto, el procedimiento que se sigue para anotar un corpus consiste en introducir una serie de códigos o etiquetas que pueden referirse tanto a aspectos propiamente lingüísticos (anotación) como a otros aspectos (codificación o *mark-up*).

3.6.1. Codificación: estándares

La *codificación* trata de documentar tanto el corpus en general como los textos concretos que lo forman. Refleja:

- *Aspectos extratextuales*. Se codifica información externa al texto, referida al contexto en que se efectuó su producción. Así, es normal que se especifiquen datos de carácter bibliográfico, temática o género del texto, autor de la recopilación, lugar en que se tuvo lugar la producción del texto, fecha de finalización o última actualización, etc. Ofrece grandes ventajas para la recuperación de información, al permitir obtener únicamente los textos que responden a un determinado criterio o serie de criterios (textos de un autor, de una fecha, de un origen geográfico, de un tema, etc.). Todo este tipo de informaciones se suelen recoger en una parte del documento denominada "cabecera.
- *Aspectos textuales*. En este caso se codifica información básica sobre la estructura interna del texto (títulos, capítulos, párrafos, oraciones, citas textuales, tablas, fotos, notas a pie de página,

intervenciones de un hablante, etc.). Igual que en el caso anterior, esta información se debe diferenciar claramente de lo que es el texto propiamente dicho.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE CORPES SYSTEM
"file:/C:/Documents%20and%20Settings/a/Escritorio/CORPES_XXI/CORPESXXIEXT.dtd">
<CORPES id="">
   <cabecera fecha_electrónica="">
      <título_principal autor_título_principal=""></título_principal>
      <título_secundario autor_título_secundario=""></título_secundario>
      <edición lugar_de_publicación="" editorial="" fecha_de_publicación=""/>
      <numpal n=""/>
      <criterio_clasificación_CORPES criterio="" año=""/>
      <clasificación_textual bloque="" tema="" medio="" país="" zona="" origen=""/>
      <codificación equipo_codificación="" persona_codificación="" fecha_codificación=""/>
      <validación valor_validación="" persona_validación="" fecha_validación=""/>
      <revisión_RAE valor_revisión_RAE="" persona_revisión_RAE="" fecha_revisión_RAE=""/>
      <notas></notas>
   </cabecera>
   <texto>
      <p></p>
   </texto>
</CORPES>
```

Ilustración 139. Ejemplo de cabecera para un corpus: CORPES XXI[294].

```
PA2009_0016_001.xml
1  <?xml version="1.0" encoding="UTF-8"?>
2  <!DOCTYPE CORPES SYSTEM "file:/C:/RAE/CORPESXXIEXT.dtd">
3  <CORPES id="PA2009_0016_001">
4      <cabecera fecha_electrónica="2009-05-07">
5          <título_principal autor_título_principal="">Voanoticias.com</título_principal>
6          <título_secundario autor_título_secundario="">Reserva Federal analiza medidas económicas</título_secundario>
7          <edición lugar_de_publicación="Washington D. C." editorial="voanews.com/spanish" fecha_de_publicación="2009-03-18"/>
8          <numpal n="168"/>
9          <criterio_clasificación_CORPES criterio="Primera_edición" año="2009"/>
10         <clasificación_textual bloque="No_ficción" tema="Política_economía_justicia" medio="Prensa" país="Estados_Unidos" zo
11         <codificación equipo_codificación="ULE" persona_codificación="Llanos Casado, Laura" fecha_codificación="2009-05-07"/>
12         <validación valor_validación="1" persona_validación="Villayandre Llamazares, Milka" fecha_validación="2009-05-09"/>
13     </cabecera>
14     <texto>
15         <p>La Reserva Federal estadounidense tiene la esperanza de que las medidas que serán analizadas en su reunión de este
16         <p>Se tiene previsto que la Reserva mantenga una tasa de interés clave en su históricamente bajo nivel de cero a 0,25
17         <p>Según economistas, funcionarios del Banco Central están considerando nuevos esfuerzos para ayudar a los mercados c
18         <p>El mercado crediticio es fundamental para el crecimiento económico pero se estancó cuando colapsó el mercado inmob
19         <p>La Fed tiene nueva información por considerar. Un informe divulgado el miércoles muestra que los precios al consum
20     </texto>
21 </CORPES>
```

Ilustración 140. Ejemplo de texto de CORPES XXI codificado según el esquema anterior[295].

[294] En el modelo de cabecera propuesto para CORPES XXI, se observa perfectamente el tipo de información que se recoge en esta parte del corpus: fecha en que se lleva a cabo la codificación, título y autor del texto, lugar y fecha de publicación, editorial, número de palabras del texto, tema, medio de publicación, país, zona, equipo y persona que lleva a cabo la codificación, persona que la valida, etc.

[295] Aquí se muestra el resultado de completar las etiquetas anteriores para un texto concreto, que aparece recogido después de la cabecera, separado en párrafos.

Para propiciar el intercambio de información y que cada grupo de investigación no emplee un sistema propio, existen hoy en día recomendaciones sobre el uso de unos esquemas de etiquetación concretos para garantizar la uniformidad en el sistema de codificación y el intercambio de documentos. Todos ellos comparten muchas características, aunque ninguno se ha erigido en estándar absoluto. Algunos de los más conocidos son (*cf.* McEnery, Xiao y Tono 2006:23 y ss.):

1) *Referencias COCOA* ("word Count and COncordance on Atlas"): las primeras históricamente. Se trata de un programa informático antiguo desarrollado por Donald B. Russell en el Atlas Computer Laboratory de Chilton, England. Fue usado originariamente para extraer índices de palabras de textos electrónicos en un formato que más tarde fue adoptado por conocidos programas de concordancias[296], como OCP (*Oxford Concordance Program*). Corpus que se han codificado siguiendo este sistema son el Longman-Lancaster Corpus y el Helsinki Corpus.

Una referencia COCOA consta de una estructura muy sencilla: <nombre etiqueta valor>. Un ángulo marca la apertura de la etiqueta y otro marca el cierre. En su interior se recogen dos entidades:

- Una etiqueta o código: nombre de una variable o atributo (p. ej. A representa la variable "autor").
- Un conjunto de caracteres que representan los valores del atributo, ejemplos de esa variable (p. ej. Charles Dickens).

[296] Los programas de concordancias son una de las herramientas más utilizadas para extraer información de los corpus: palabras en su contexto inmediato anterior y posterior. Ya se han mostrado algunos ejemplos en el apartado anterior, a propósito de la palabra *bividí*.

Así, la referencia COCOA para indicar que el autor de un texto es Charles Dickens sería: <A CHARLES DICKENS>[297].

El siguiente ejemplo es la cabecera de un documento del Helsinki Corpus. En la parte izquierda está codificada la cabecera en formato COCOA, mientras que en la derecha está la glosa del código (una "X" indica que la información relacionada con ese código es irrelevante o no está disponible):

<B CMPOLYCH>
<Q ME3 NN HIST TREVISA>
<N POLYCHRONICON>
<A TREVISA JOHN>
<C ME3>
<O 1350-1420>
<M 1350-1420>
<K CONTEMP>
<D SL>
<V PROSE>
<T HISTORY>
<G TRANSL>
<F LATIN>
<W WRITTEN>
<X MALE>
<Y 40-60>
<H PROF>
<U X>
<E X>
<J X>
<I X>
<Z NARR NON-IMAG>
<S SAMPLE X>

(1) <B = `name of text file'
(2) <Q = `text identifier'
(3) <N = `name of text'
(4) <A = `author'
(5) <C = `part of corpus'
(6) <O = `date of original'
(7) <M = `date of manuscript'
(8) <K = `contemporaneity'
(9) <D = `dialect'
(10) <V = `verse' or `prose'
(11) <T = `text type'
(12) <G = `relationship to foreign original'
(13) <F = `foreign original'
(14) <W = `relationship to spoken language'
(15) <X = `sex of author'
(16) <Y = `age of author'
(17) <H = `social rank of author'
(18) <U = `audience description'
(19) <E = `participant relationship'
(20) <J = `interaction'
(21) <I = `setting'
(22) <Z = `prototypical text category'
(23) <S = `sample'
(24) <P = `page'
(25) <R = `record'

Ilustración 141. Ejemplo de texto codificado mediante referencias COCOA.

[297] Ejemplo tomado de MCENERY y WILSON (1996:27).

Las referencias COCOA únicamente permiten codificar tipos limitados de información (autores, fechas, títulos). Por eso se han propuesto esquemas más ambiciosos para codificar textos, entre los que destacan:

2) *TEI* (*Text Encoding Initiative*)[298]: iniciativa que nace en 1987 promovida por la ACL (*Association for Computational Linguistics*), la ALLC (*Association for Literary and Linguistic Computing*) y la ACH (*Association for Computers and the Humanities*) con el propósito de constituirse en un estándar no solo para la codificación de corpus, sino de textos electrónicos en general. Se puede expresar en diferentes lenguajes formales, como SGML (*Standard Generalized Markup Language*)[299] y actualmente XML (*Extensible Markup Language*)[300]. A diferencia de las referencias COCOA, la TEI permite codificar más elementos, que se marcan con una etiqueta de inicio <> y otra de cierre </>.

Los textos se conciben como documentos que se estructuran en dos partes:

- *cabecera* (*head*), donde se incorpora de forma obligatoria la información bibliográfica completa del archivo electrónico y, opcionalmente, otro tipo de informaciones (como fecha de los cambios, lenguas del texto, etc.);
- y *cuerpo* (*body*) o texto propiamente dicho, en el que se codifican básicamente elementos estructurales (capítulos, párrafos, oraciones, palabras).

[298] URL: http://www.tei-c.org/index.xml.
[299] URL: http://xml.coverpages.org//sgml.html.
[300] URL: http://www.w3.org/XML/.

```
<obra> A guerra das linguas e as políticas lingüísticas </obra> <autor>
Louis-Jean Calvet </autor> <tradutor> Xoán Manuel Garrido Vilariño
</tradutor> <capítulo> <tit_cap> Nas orixes do conflicto </tit_cap>
<sección> <tit_sec> A cuestión das orixes </tit_sec> <parágrafo> <oración>
Os animais non falan porque non teñen nada que dicir... </oración> ...
</parágrafo> ... </sección> ... </capítulo> </obra>
```

Ilustración 142. Ejemplo de texto codificado mediante el sistema etiquetas SGML[301]*.*

```
<anthology>
  <poem><title>The SICK ROSE</title>
     <stanza>
          <line>O Rose thou art sick.</line>
          <line>The invisible worm,</line>
          <line>That flies in the night</line>
          <line>In the howling storm:</line>
     </stanza>
     <stanza>
          <line>Has found out thy bed</line>
          <line>Of crimson joy:</line>
          <line>And his dark secret love</line>
          <line>Does thy life destroy.</line>
     </stanza>
  </poem>

          <!-- more poems go here    -->

</anthology>
```

Ilustración 143. Otro ejemplo de texto codificado en SGML[302]*.*

Todos los textos que siguen el esquema de codificación propuesto por la TEI deben ajustarse a una estructura particular. Así, la cabecera –enmarcada entre la etiqueta de apertura <teiHeader> y la de cierre </teiHeader>– está conformada por una serie de elementos[303], unos obligatorios y otros opcionales:

- <fileDesc> </fileDesc>
- <encodignDesc> </encodignDesc>
- <profileDesc> </profileDesc>
- <revisionDesc> </revisionDesc>

[301] Tomado de: http://sli.uvigo.es/dout_cluvi/contidos/tema1.html#1.3.4.
[302] Tomado de la URL: http://www.isgmlug.org/sgmlhelp/g-sg13.htm.
[303] Tomado de: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

```
<teiHeader>
 <fileDesc>
<!-- ... -->
 </fileDesc>
 <encodingDesc>
<!-- ... -->
 </encodingDesc>
 <profileDesc>
<!-- ... -->
 </profileDesc>
 <revisionDesc>
<!-- ... -->
 </revisionDesc>
</teiHeader>
```

Ilustración 144. Estructura de la cabecera de un documento según la TEI.

De todos ellos, solo el primero, <fileDesc> </fileDesc> es obligatorio; los otros son opcionales. Contiene la descripción bibliográfica completa del archivo electrónico, registrada mediante otra serie de subelementos:

```
<teiHeader>
 <fileDesc>
  <titleStmt>
<!-- ... -->
  </titleStmt>
  <editionStmt>
<!-- ... -->
  </editionStmt>
  <extent>
<!-- ... -->
  </extent>
  <publicationStmt>
<!-- ... -->
  </publicationStmt>
  <seriesStmt>
<!-- ... -->
  </seriesStmt>
  <notesStmt>
<!-- ... -->
  </notesStmt>
  <sourceDesc>
<!-- ... -->
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

Ilustración 145. Estructura de la etiqueta <fileDesc> </fileDesc>.

Tres son obligatorios:

- <titleStmt>: título de la obra, con subelementos como el título, el autor o el compilador.

```
<titleStmt>
 <title>Two stories by Edgar Allen Poe: electronic version</title>
 <author>Poe, Edgar Allen (1809-1849)</author>
 <respStmt>
  <resp>compiled by</resp>
  <name>James D. Benson</name>
 </respStmt>
</titleStmt>
```

Ilustración 146. Estructura de la etiqueta <titleStmt> </titleStmt>.

- <publicationStmt>: donde se codifica información sobre la publicación o distribución de la obra, con subelementos como la editorial, el lugar de la edición, la fecha, el ISBN o los derechos de autor.

```
<publicationStmt>
 <publisher>Oxford University Press</publisher>
 <pubPlace>Oxford</pubPlace>
 <date>1989</date>
 <idno type="ISBN">0-19-254705-4</idno>
 <availability>
  <p>Copyright 1989, Oxford University Press</p>
 </availability>
</publicationStmt>
```

Ilustración 147. Estructura de la etiqueta <publicationStmt> </publicationStmt>.

- <sourceDesc>: que describe la fuente de la que se ha tomado el texto electrónico. Suele consistir en una breve frase descriptiva entre dos etiquetas elemento de naturaleza digital, elemento sin fuente… .

```
<teiHeader>
 <fileDesc>
  <titleStmt>
   <title>Thomas Paine: Common sense, a
       machine-readable transcript</title>
   <respStmt>
    <resp>compiled by</resp>
    <name>Jon K Adams</name>
   </respStmt>
  </titleStmt>
  <publicationStmt>
   <distributor>Oxford Text Archive</distributor>
  </publicationStmt>
  <sourceDesc>
   <bibl>The complete writings of Thomas Paine, collected and edited
       by Phillip S. Foner (New York, Citadel Press, 1945)</bibl>
  </sourceDesc>
 </fileDesc>
</teiHeader>
```

Ilustración 148. Cabecera mínima de un texto según la TEI.

3) *CES* (*Corpus Encoding Standard*)[304]: diseñado específicamente para la lingüística de corpus y no para todos los documentos electrónicos en general, esta nueva propuesta de estándar supone, por una parte, una simplificación de las recomendaciones de la TEI, al limitarse a aquellas que son de interés para los corpus, y, por otra parte, una ampliación de dicha iniciativa, ya que propone etiquetas específicas para tratar elementos lingüísticos. También se expresa tanto en SGML como en XML (XCES).

3.6.2. Anotación: tipos

El proceso conocido como *anotación* da cuenta, mediante etiquetas, de aspectos exclusivamente lingüísticos. En este sentido, las informaciones que se hacen explícitas pueden referirse a cualquiera de los distintos niveles (y aspectos) que se distinguen en el tratamiento del lenguaje.

[304] URL: http://www.cs.vassar.edu/CES/

Así, se habla de: anotación ortográfica, fonética, prosódica, morfológica, sintáctica, semántica, discursiva, etc.

Por otra parte, el proceso de anotación puede ser (*cf.* McEnery 2003:455-456):

- *Totalmente automático*: un programa informático anota el corpus según las reglas y algoritmos programados o aprendidos. El desarrollo de este tipo de programas es costoso, pero una vez logrado, introduce rapidez y consistencia en la anotación. P. ej. el programa no olvidará poner la etiqueta N a ningún nombre, mientras que a una persona se le puede olvidar. Estos programas alcanzan porcentajes de error inferiores al 3% en la lematización (*vid. infra*) y la asignación de categorías morfológicas en lenguas como el inglés, el francés y el español.
- *Semiautomático*: en ocasiones, los resultados que ofrecen los programas informáticos no son fiables o exactos, por lo que exigen la intervención de analistas humanos expertos (*post-edición*). Es decir, primero actúa el programa y, después, las personas revisan los resultados o interactúan con él cuando les solicita ayuda. Lógicamente, el proceso es más lento y costoso que si se efectuara totalmente de forma automática, pero aporta mayor fiabilidad en los resultados y es más rápido que la anotación solo manual.
- *Totalmente manual*: en algunos casos, no es posible efectuar la anotación de forma automática, bien sea porque no existen las herramientas informáticas para ello o bien porque los porcentajes de error que arrojan son muy altos, por lo que son personas las que llevan a cabo todo el proceso. Dado su coste y lentitud, se suele limitar a corpus de tamaño reducido.

La anotación de los textos de un corpus es un proceso de enriquecimiento del mismo, que no se realiza mediante la introducción de nueva información, sino haciendo explícito lo que antes estaba implícito, para favorecer la posterior explotación del corpus. Esto implica, en algunos casos, optar por una interpretación de los datos: dado un texto, existen multitud de análisis posibles que surgen del grado de variación que producen, por una parte, las ambigüedades del lenguaje natural y, por otra, las fronteras borrosas entre las categorías. Al final, la anotación de un corpus no es más que uno de esos posibles análisis que se "impone" sobre el corpus. Normalmente este análisis está codificado en relación con una teoría lingüística específica. Los rasgos o características se representan mediante elementos mnemotécnicos que se introducen en el corpus. Esta información adicional derivada o extraída del corpus es introducida por el investigador en función de sus objetivos y, lo que es más importante, de su interpretación lingüística de los materiales recogidos. Como ya se ha señalado, la anotación añade valor a un corpus, ya que lo habilita para responder a más preguntas sobre el funcionamiento del lenguaje.

Algunas de las ventajas que se han señalado para los corpus anotados son (*cf.* McEnery 2003:454-455):

a) *Facilidad de explotación*: con un corpus anotado, las posibilidades y velocidad de explotación aumentan considerablemente. Es decir, el corpus proporciona mayor facilidad para extraer información y aporta mayor fiabilidad a los resultados.

b) *Reutilización*: dado lo costoso del proceso, una vez realizado, permite efectuar análisis una y otra vez.

c) *Multifuncionalidad*: el análisis o explotación que se ha hecho con un propósito específico en mente, cuando el corpus se reutiliza puede

ser totalmente distinto, pero los datos de partida serán el mismo corpus anotado para un fin determinado.

d) *Análisis explícitos*: la anotación es un excelente medio de dejar constancia de una visión sobre el lenguaje, lo que es un claro beneficio frente a la subjetividad.

Por supuesto, no faltan las críticas (*cf.* McEnery 2003:456-457), como que: i) la anotación puede dificultar el acceso al texto, al mezclarse con él; ii) que impone un análisis lingüístico a los usuarios del corpus; iii) que puede restringir el acceso, la actualización y la expansión del corpus; iv) que puede hacer peligrar la exactitud o precisión de los análisis, etc. No obstante, todas estas críticas han sido rebatidas.

Ahora bien, ¿qué tipos de información lingüística se pueden etiquetar en un corpus? Los principales tipos de anotación se refieren a:

1) *Anotación categorial* o *gramatical* (*POS tagging*). Se trata del tipo más básico y más común hoy en día de anotación lingüística de un corpus. Consiste en asignar a cada unidad léxica del texto un código que indica su categoría o parte de la oración ("Part Of Speech" > POS), por lo que también es conocida como "anotación gramatical" o "anotación morfosintáctica". Es útil en sí misma para distinguir entre homógrafos (*coma*, *prueba*, *vino*) o como base fundamental para otros tipos de análisis (sintáctico, semántico). Los programas que efectúan este proceso de forma automática se denominan *taggers* o *etiquetadores*, como CLAWS[305] para el inglés. Las etiquetas se adjuntan a las palabras de

[305] URL: http://ucrel.lancs.ac.uk/claws/. Este programa, desarrollado a principios de los ochenta en el *University Centre for Computer Corpus Research on Language* de la Universidad de Lancaster, alcanza porcentajes de éxito del 96-97%. CLAWS ("Constituent Likelihood Automatic Word-tagging System") fue utilizado, por ejemplo, para etiquetar el BNC.

diferentes formas: usando referencias de entidades de la TEI (&), con un guion bajo, con códigos SGML o XML, etc.

A continuación se muestran algunos ejemplos de textos con este tipo de anotación:

A move to stop Mr Gaitskell from nominating any more labour life peers is to be made at a meeting of labour MPs tomorrow.

```
A014    ^ a_AT move_NN to_TO stop_VB \OMr_NPT Gaitskell_NP from_IN
A014    nominating_VBG any_DTI more_AP labour_NN
A015    life_NN peers_NNS is_BEZ to_TO be_BE made_VBN at_IN a_AT
                meeting_NN
A015    of_IN labour_NN \OMPs_NPTS tomorrow_NR
```

Ilustración 149. Fragmento etiquetado del LOB corpus[306].

[306] Tomado de http://khnt.hit.uib.no/icame/manuals/lobman/LOB2.HTM#29. En este corpus, las etiquetas gramaticales se asignaron a continuación de las palabras mediante un guion bajo. Existen dos versiones anotadas del corpus: una como la mostrada en el ejemplo y otra en la que las palabras del corpus están dispuestas de forma vertical, cada una en una línea, precedidas de un código numérico y seguidas de la información gramatical. Se emplearon etiquetadores que realizaron la tarea de forma automática combinados con una labor manual previa y posterior de edición. Las claves a las que se corresponden las etiquetas son (*vid.* para más detalles el manual: http://khnt.hit.uib.no/icame/manuals/lobman/LOBAPP4.HTM):

AP: post-determiner/pronoun
AT: article
BE: *be*
BEZ: *is*, *'s*
DTI: singular or plural determiner
IN: preposition
NN: singular common noun
NNS: plural common noun
NP: singular proper noun
NPT(S): plural titular noun with word-initial capital
NR: singular adverbial noun
TO: infinitival *to*
VB: base form of verb
VBN: past participle
VBG: present participle, gerund

CLAWS ofrece tres tipos posibles de etiquetación: horizontal, vertical y XML. Así, al introducir la oración "Poll shows Conservative party still short of clear majority" en la versión de prueba en línea[307], obtenemos los siguientes resultados:

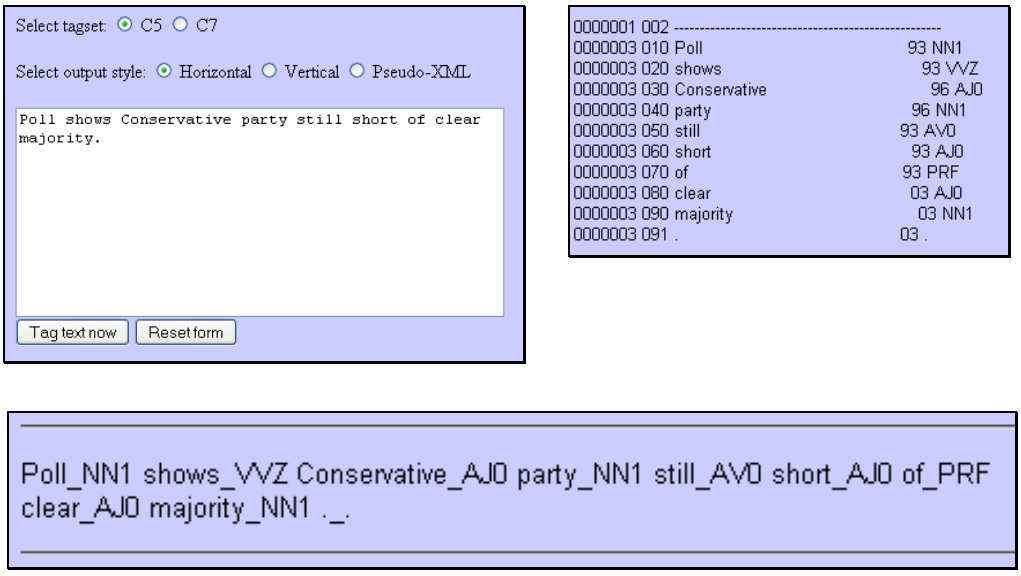


Ilustración 150. Anotación gramatical con CLAWS: vertical, horizontal y pseudo-XML[308].

[307] URL: http://ucrel.lancs.ac.uk/cgi-bin/claws7.pl

[308] El etiquetario utilizado es muy similar al visto para el *LOB corpus*. Las claves de las etiquetas se corresponden a (*vid.* http://ucrel.lancs.ac.uk/claws5tags.html):

AJ0: adjective (unmarked)
AV0: adverb (unmarked)
NN1: singular noun
PRF: the preposition *OF*
VVZ: *-s* form of lexical verb
PRF: the preposition *OF*

Veamos a continuación un par de casos para el español: el *parser* para el proyecto *Spanish FrameNet*[309] incorpora un etiquetador que muestra la información gramatical relevante de cada una de las unidades léxicas de un enunciado. En el ejemplo se muestran los resultados al introducir la oración "Mi novio me vendió por un Mercedes":

Texto a analizar

Mi novio me vendió por un Mercedes.

Texto analizado

[o mi.APOS:m:f:s novio.N:m:s me.CLI:1s vender.VPRED:IPIND:3s por.PREP un.DET:m:s Mercedes.NPROP_N_PILA .PUNTO]

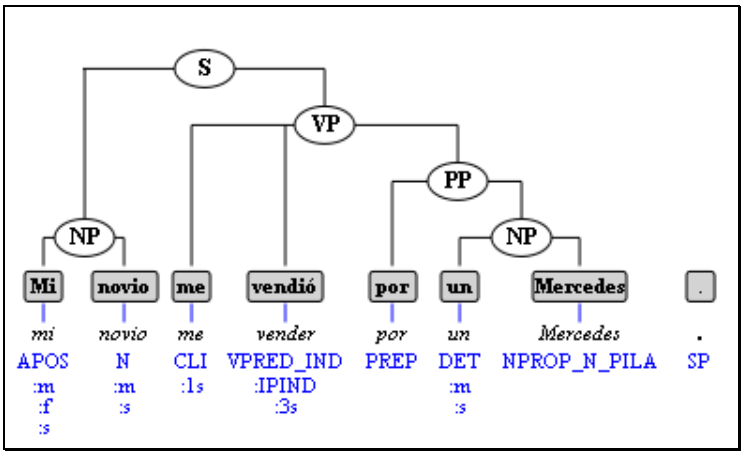


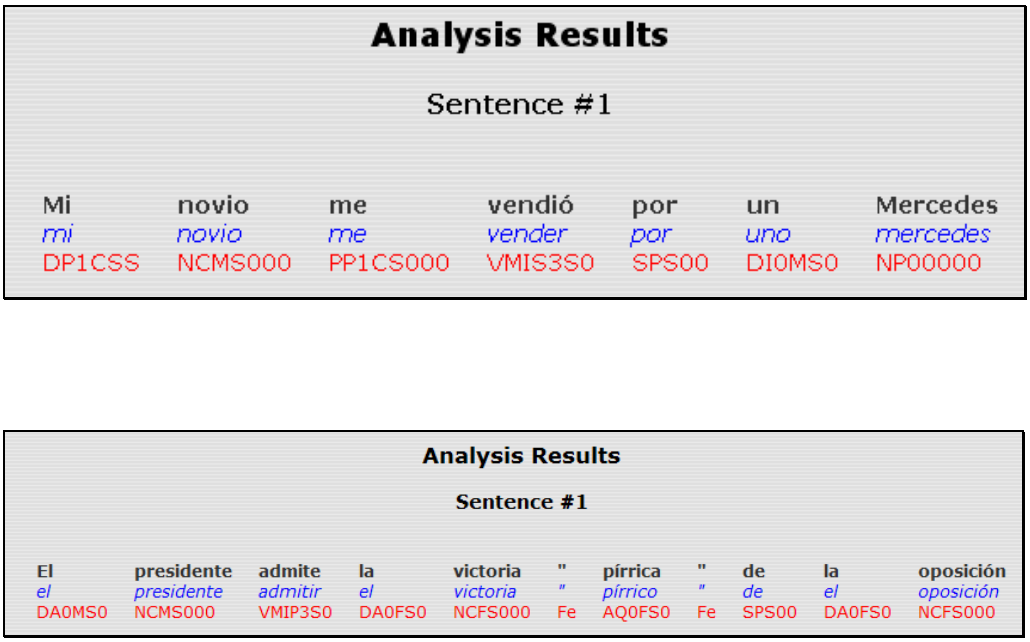
Ilustración 151. Etiquetador del proyecto Spanish FrameNet[310].

[309] URL: http://gemini.uab.es:9080/SFNsite/sfn-tools/sfn-parser

[310] La clave de las etiquetas empleadas es la siguiente (*vid.* para más detalles URL: http://gemini.uab.es:9080/SFNsite/taggers-chunkers):

APOS: adjetivo posesivo
N: nombre
CLI: pronombre clítico
DET: determinante
NPROP_N_PILA: nombre propio nombre de pila
PREP: preposición
VPRED_IND: verbo predicadivo
m: masculino
f: femenino
s: singular
1s: primera persona del singular
3s: tercera persona del singular
IPCIND: indicativo pretérito indefinido

Por último, veamos el funcionamiento de *FreeLing*[311] a propósito de las oraciones "Mi novio me vendió por un Mercedes" y "El presidente admite la victoria 'pírrica' de la oposición":



Analysis Results

Sentence #1

Mi	novio	me	vendió	por	un	Mercedes
mi	*novio*	*me*	*vender*	*por*	*uno*	*mercedes*
DP1CSS	NCMS000	PP1CS000	VMIS3S0	SPS00	DI0MS0	NP00000

Analysis Results

Sentence #1

El	presidente	admite	la	victoria	"	pírrica	"	de	la	oposición
el	*presidente*	*admitir*	*el*	*victoria*	*"*	*pírrico*	*"*	*de*	*el*	*oposición*
DA0MS0	NCMS000	VMIP3S0	DA0FS0	NCFS000	Fe	AQ0FS0	Fe	SPS00	DA0FS0	NCFS000

Ilustración 152. Etiquetador gramatical de FreeLing[312].

[311] URL: http://garraf.epsevg.upc.es/freeling/demo.php.

[312] *FreeLing 2.1*, TALP Research Center, Universitat Politècnica de Catalunya. El etiquetario usado por este anotador se basa en las propuestas del grupo EAGLES para la anotación morfosintáctica de lexicones y corpus para todas las lenguas europeas. Hay atributos que no se especifican, bien por no existir en español o por no ser relevantes. Estos llevan la marca 0. Clave de las etiquetas (*vid.* http://garraf.epsevg.upc.es/freeling/doc/userman/parole-es.html):

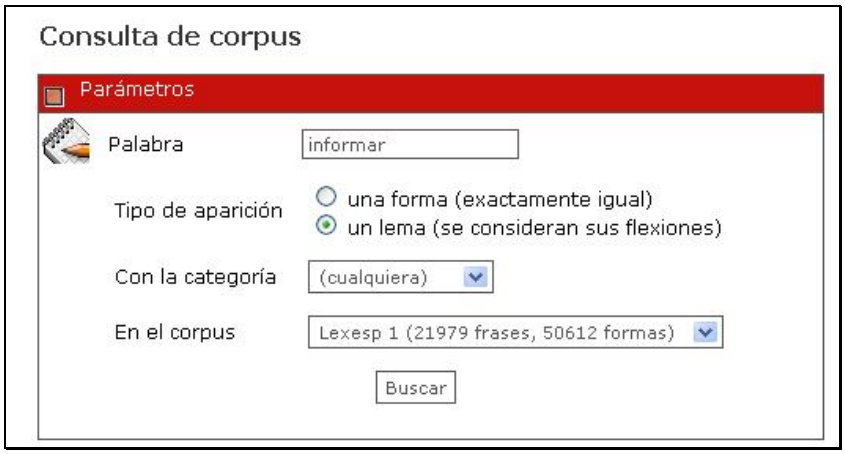
AQ0FS0: adjetivo calificativo femenino singular
DA0FS0: determinante artículo femenino singular
DA0MS0: determinante artículo masculino singular
DI0MS0: determinante indefinido masculino singular
DP1CSS: determinante posesivo primera persona género común singular un poseedor
NCFS000: nombre común femenino singular
NCMS000: nombre común masculino singular
NP00000: nombre propio
PP1CS000:
SPS00:
VMIP3S0: verbo principal indicativo presente tercera persona singular
VMIS3S0: verbo principal indicativo pasado tercera persona singular

Lógicamente, el proceso de anotación gramatical no está exento de problemas, como la segmentación del texto en palabras gramaticales, las expresiones idiomáticas, las contracciones, etc.

En general, los programas de etiquetado categorial buscan la facilidad de uso: un conjunto de etiquetas mnemotécnicas, fáciles de identificar sin tener que consultar el significado de la etiqueta en una tabla. El etiquetario o conjunto de etiquetas que los programas manejan varía en función del grado de detalle que se quiera dar al análisis.

2) *Lematización*. Este tipo de anotación está íntimamente ligada a la identificación de partes de la oración, es decir, al proceso anterior. Consiste en la reducción de las palabras de un corpus a sus respectivas formas básicas (entradas o lemas que aparecen normalmente en un diccionario). El *lema* está formado por las variantes flexivas de esa palabra. P. ej. *amar* sería el lema que incluiría las formas flexivas *amo*, *amas*, *amó*... Se trata de un tipo de anotación especialmente importante para la lexicografía y los estudios de vocabulario. Si el corpus está lematizado, el usuario puede examinar todas las variantes de una palabra de una vez sin necesidad de buscar explícitamente una por una, así como extraer toda la información sobre frecuencia y distribución. Igual que ocurría en el caso anterior, existen programas que realizan el proceso de forma automática con bastante precisión para lenguas como el inglés, el francés o el español. No obstante, para lenguas con poca flexión, como el chino o el inglés, la lematización no parece ser muy útil, de ahí que p. ej. para el inglés haya pocos corpus lematizados.

A continuación se muestra cómo funciona la lematización en el corpus Lexesp[313]: al introducir el término sobre el que queremos hacer la consulta (*informar* en esta búsqueda), seleccionamos la opción "un lema", que nos devolverá todos los casos en que aparezca el verbo en cualquiera de sus formas flexionadas (*informarse*, *informa*, *informado*, etc.) y no solo la forma del infinitivo que hemos introducido:



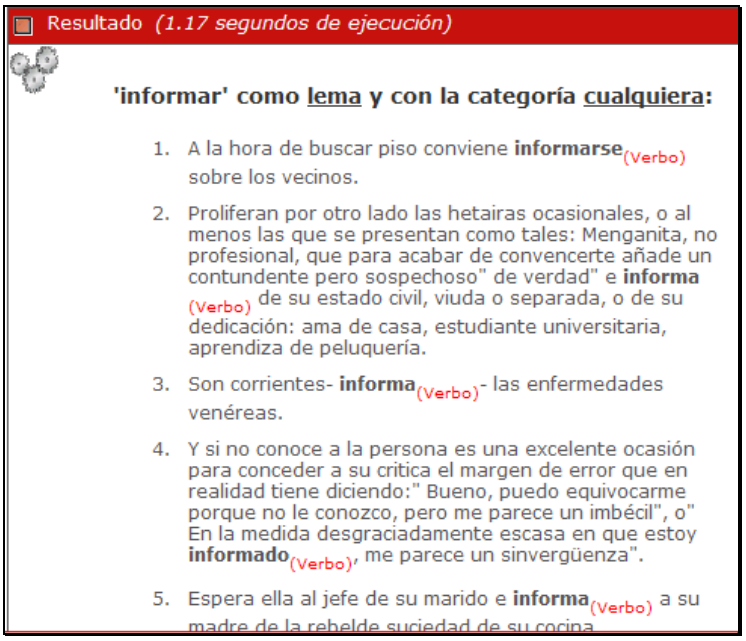


Ilustración 153. Búsqueda en el corpus Lexesp de Thera.

[313] URL: http://www.thera-clic.com/ (sección "Demos" → "Corpus").

3) *Anotación sintáctica* o "*parsing*". El etiquetado gramatical y la lematización son solo una parte de una empresa mayor, el *análisis sintáctico* o *parsing* de un corpus. De un corpus analizado sintácticamente es posible recuperar información más abstracta que no puede especificarse en términos de palabras o clases léxicas: se trata de la referida a la estructura interna de las oraciones, por ejemplo, frases, cláusulas, constituyentes (frases nominales, adjetivas, adverbiales, preposicionales). El etiquetado sintáctico es la forma más común de anotación después de la categorial y es especialmente útil en la mayoría de aplicaciones del Procesamiento del Lenguaje Natural.

Los corpus analizados sintácticamente se conocen a veces como "treebanks", término que alude a los diagramas arbóreos o marcadores de frase utilizados en el análisis, como los siguientes, resultado de someter al analizador sintáctico de Thera[314] la frase: "La eurozona da 80.000 millones a Grecia":

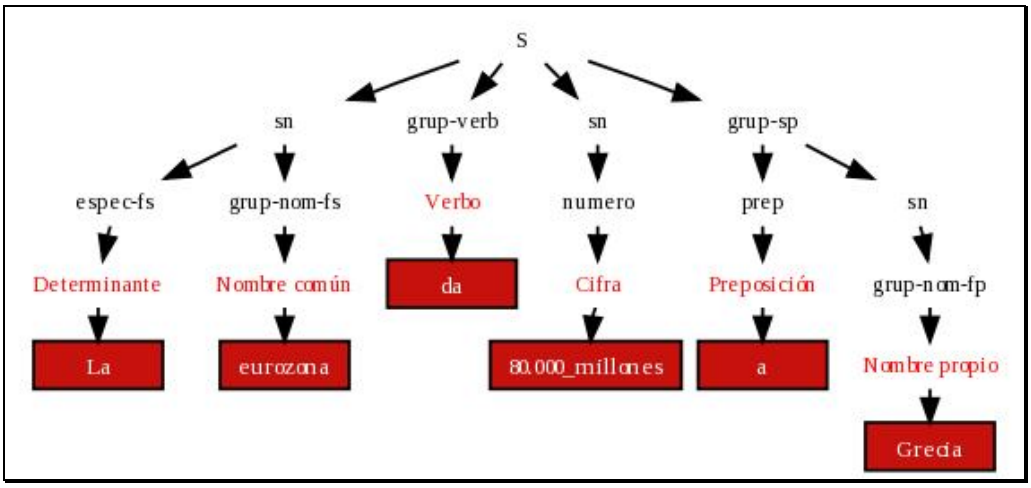


Ilustración 154. Diagramas arbóreos del analizador de Thera.

[314] URL: http://www.thera-clic.com/site/Demos/Sintactico-ES.html

Sin embargo, estos diagramas visuales rara vez se encuentran en la anotación de un corpus. Con más frecuencia, la misma información se representa usando corchetes o paréntesis, de ahí la denominación de *parentizado* ("*bracketing"*). La información morfosintáctica se adhiere a las palabras por medio de guiones bajos en forma de etiquetas categoriales, mientras que los constituyentes se indican abriendo y cerrando corchetes anotados al principio y al final de cada frase. Obsérvese el siguiente ejemplo tomado del *Lancaster Parsed Corpus*, donde los corchetes van marcando las agrupaciones sintácticas[315] del enunciado "It would become easy to be cynical and to despair":

```
B06 666
[S[N it_PP3 N][V would_MD become_VB V][J easy_JJ J][Ti&[Vi
to_TO be_BE Vi][J cynical_JJ J][Ti+ and_CC [Vi to_TO
despair_VB Vi]T+]Ti&] ._. S]
```

Ilustración 155. Parentizado en el "Lancaster parsed corpus".

El *parsing* puede llevarse a cabo de forma automática, pero a causa de los bajos porcentajes de éxito, a menudo requiere la corrección por parte de analistas humanos o, incluso, la anotación totalmente manual.

Por otra parte, el análisis puede ser más o menos detallado:

[315] "S" es la etiqueta empleada para marcar el inicio y el final del enunciado, constituido por una frase nominal –N (*it*)–, una frase verbal –V (*would become*)–, una frase adjetiva –J (*easy*)– y una frase de infinitivo compuesta, es decir, formada por la coordinación de dos infinitivos –Ti& (*to be cynical and to despair*)–, cuya composición interna es una frase verbal de infinitivo y un adjetivo –Vi (*to be*) y J (*cynical*)– y otra frase verbal de infinitivo –Vi (*to despair*)–. Más detalles en el *Manual of Information for the Lancaster Parsed Corpus* (Garside, Leech y Váradi). URL: http://khnt.hit.uib.no/icame/manuals/LPC/LPC.PDF

(a) *Full parsing*, o análisis detallado, como en este ejemplo tomado del *Lancaster-Leeds treebank* (*cf.* McEnery y Wilson 1996:45), en el que, además de marcar los constituyentes, se añade información gramatical a cada una de las palabras[316]:

```
[S[Ncs another_DT new_JJ style_NN feature_NN Ncs] [Vzb is_BEZ Vzb]
[Ns the_AT1 [NN/JJ& wine-glass_NN [JJ+ or_CC flared_JJ HH+]NN/JJ&]
heel_NN ,_, [Fr[Nq which_WDT Nq] [Vzp was_BEDZ shown_VBN Vzp]
[Tn[Vn teamed_VBN Vn] [R up_RP R] [P with_INW [NP[JJ/JJ/NN&
pointed_JJ ,_, [JJ- squared_JJ JJ-] ,_, [NN+ and_CC chisel_NN
NN+]JJ/JJ/NN&] toes_NNS Np]P]Tn]Fr]Ns] ._. S]
```

Ilustración 156. Ejemplo de análisis sintáctico en profundidad en el "Lancaster-Leeds treebank".

Dada la dificultad que implica este tipo de anotación, con frecuencia se efectúa a mano. Es lo que sucede, por ejemplo, en el corpus AnCora[317] (corpus de catalán y español), en el que el análisis sintáctico en profundidad (constituyentes y funciones) se hizo de forma totalmente manual. Observemos su funcionamiento: como resultado de buscar en el corpus todos los enunciados en los que aparezca el lema "informar", obtenemos 172 casos, de los que podemos ver el árbol sintáctico completo en forma de gráfico o de texto jerarquizado.

---

316 Así, junto con etiquetas como las correspondientes a enunciado (S), frases de relativo (Fr), frases adjetivas (JJ), frases preposicionales (P) o frases adverbiales (R), se incorporan etiquetas mucho más específicas, como las referidas al tipo de frase nominal (con un nombre contable singular como núcleo –Ncs–, con un nombre singular –Ns–, con un nombre plural –Np–, con un pronombre relativo del tipo *wh-* –Nq–) o verbal (con un verbo que es un participio pasado –Vn–, o la tercera persona del verbo *to be* –Vzb–, o la tercera persona singular de la pasiva –Vzp–), entre otras.

317 URL: http://clic.ub.edu/ancora/

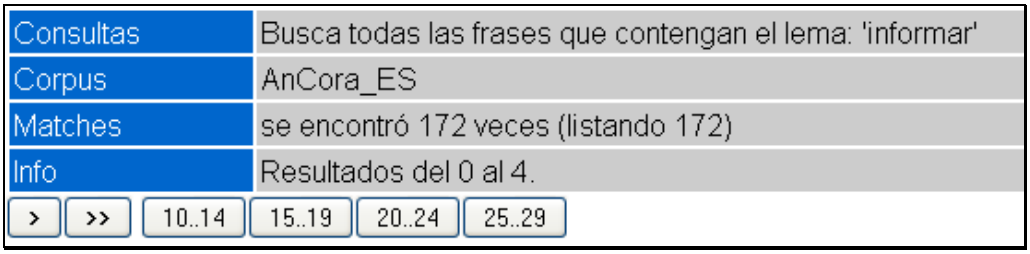


Ilustración 157. Búsqueda de "informar" en el corpus AnCora (subcorpus español).

A continuación, seleccionamos uno de los ejemplos: "La federación también *ha informado* al jugador y al Espanyol":

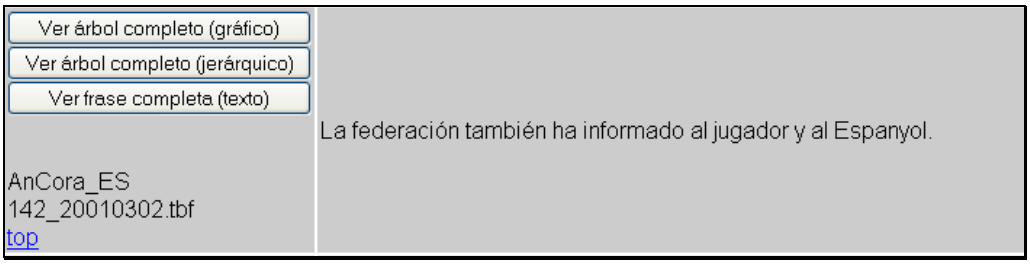


Ilustración 158. Uno de los 172 casos en que aparece el lema "informar" en el corpus AnCora.

Como resultado del análisis manual que se ha efectuado de este enunciado, obtenemos el diagrama arbóreo que muestra la agrupación de los elementos en constituyentes (sintagma nominal *–la federación–*, sintagma adverbial *–también–*, grupo verbal *–ha informado–*, sintagmas preposicionales *–al jugador*, *al Espanyol–*, etc.), así como la categoría y otra información gramatical detallada de cada unidad:

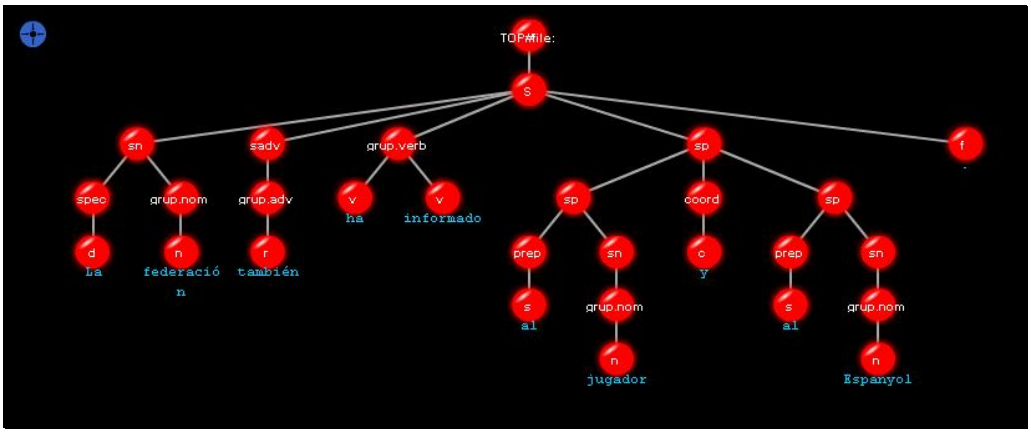


Ilustración 159. Representación arbórea del enunciado seleccionado en el corpus AnCora.

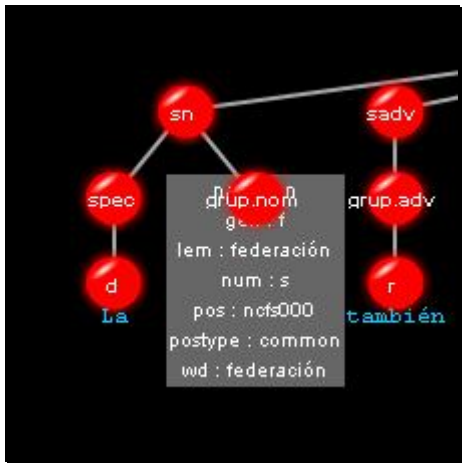


Ilustración 160. Información gramatical del lema "federación" en la oración seleccionada.

Al utilizar la opción de visualización mediante texto jerarquizado, además de los constituyentes obtenemos información sobre las funciones sintácticas de cada argumento verbal[318]: sujeto (*la federación*), modal (*también*) y complemento directo (*al jugador y al Espanyol*).

[318] Y también información sobre las funciones semánticas o papeles temáticos: agente (sujeto), paciente (complemento directo); así como sobre la clase semántica del verbo.

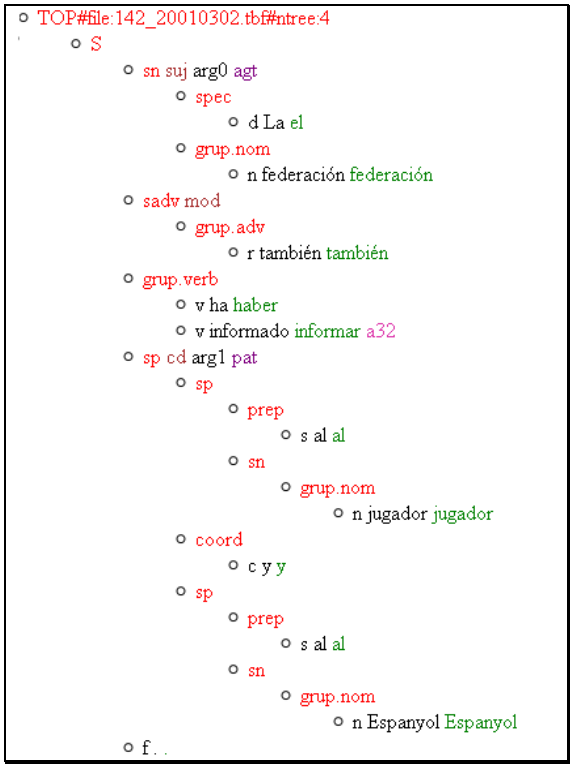


Ilustración 161. Representación en forma de texto jerarquizado del enunciado seleccionado.

(b) *Skeleton parsing, shallow parsing* o análisis menos detallado, que es aquel que utiliza menor número de categorías gramaticales o que se limita a identificar solo algunos constituyentes o *chunks* (sintagmas nominales, preposicionales, etc.). Véase el siguiente ejemplo, realizado con Freeling, a propósito del enunciado "Mi novio me vendió por un Mercedes":

Ilustración 162. Análisis superficial con FreeLing.

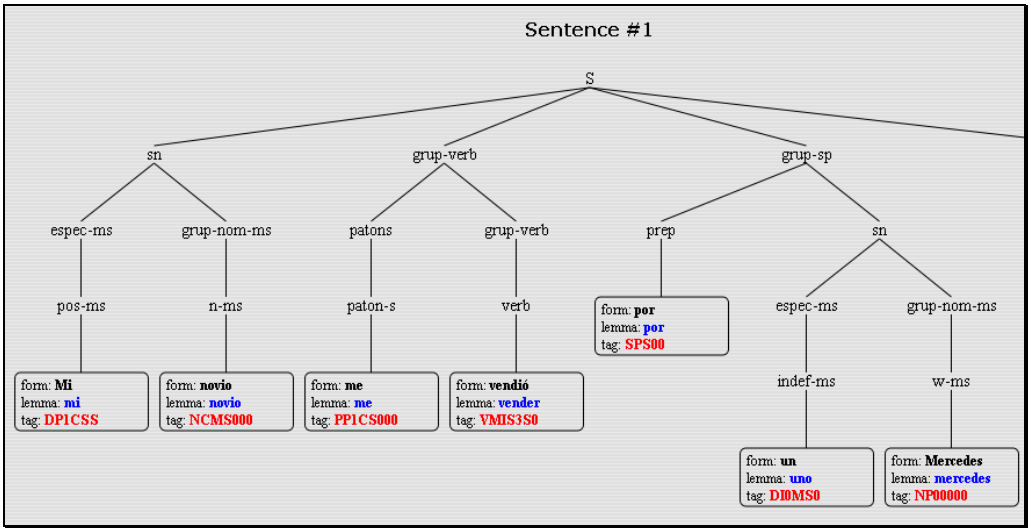


Ilustración 163. Resultados del análisis superficial con FreeLing.

4) *Anotación semántica*. Consiste en asignar etiquetas que indican información sobre el significado de una palabra. Se distinguen dos tipos de anotación semántica:

- La que tiene que ver con las *relaciones semánticas* entre los elementos de un enunciado (agente, paciente). Comienza a utilizarse ahora, aunque algunos formalismos de *parsing* la incluyen, por lo que en realidad está más relacionada con el nivel sintáctico. P. ej. en el corpus AnCora encontramos este tipo de anotación:

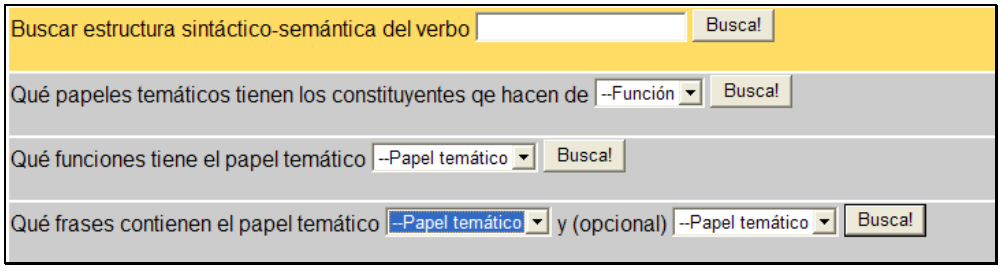


Ilustración 164. Información sobre papeles temáticos en el corpus AnCora.

Esto nos permite buscar las estructuras sintáctico-semánticas en que se utiliza un verbo. P. ej., continuando con el verbo *informar*, nos muestra la estructura sintáctica de cada uno de los casos del corpus así como los papeles temáticos de cada:

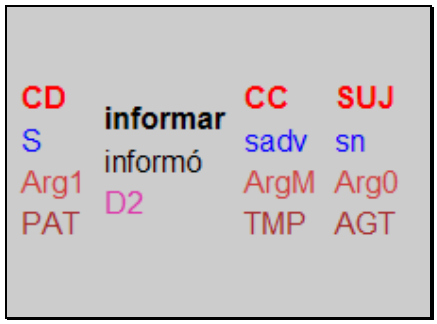


Ilustración 165. Información semántica en el corpus AnCora.

- Otro tipo de anotación semántica es la relacionada con los diferentes *sentidos de una palabra* o con su adscripción a un *campo semántico*, con vistas en ambos casos a la desambiguación, más común en la actualidad que la anterior. P. ej. en la Universidad de Lancaster se revisa el texto manualmente para determinar los campos semánticos que predominan, lo que implica que se da prioridad a la acepción relacionada con ese campo a la hora de efectuar la desambiguación. Existen programas que alcanzan un porcentaje de éxito del 92% en la realización automática de esta tarea.

El siguiente ejemplo del *LOB corpus* muestra la anotación semántica del enunciado "I like a particular shade of lipstick"[319]:

[319] Tomado de la URL: http://ucrel.lancs.ac.uk/annotation.html#acamrit

Etiqueta gramatical	*Texto*	*Etiqueta semántica*
PPIS1	*I*	Z8
VV0	*like*	E2+
AT1	*a*	Z5
JJ	*particular*	A4.2+
NN1	*shade*	O4.3
IO	*of*	Z5
NN1	*lipstick*	B4

Tabla 31. Anotación semántica en la Universidad de Lancaster.

En este fragmento, el texto se lee en sentido descendente, con las etiquetas gramaticales a la izquierda y las semánticas a la derecha. Las etiquetas semánticas constan de:

- una letra mayúscula que indica el dominio general de discurso (Z, E, A, O, B);
- un dígito que indica una primera subdivisión del campo (Z8, E2, Z5, A4, O4, B4);
- (opcionalmente) un punto seguido de un dígito que indica una subdivisión más fina (A4.2, O4.3);
- (opcionalmente) uno o más signos "+" o "–" que indican una posición positiva o negativa en una escala semántica (E2+, A4.2+).

Así, por ejemplo, la etiqueta E2+ codifica la siguiente información semántica sobre la palabra *like*:

- E: categoría de "estados, acciones, eventos y procesos emocionales".
- E2: subcategoría "gustos y aversiones".
- E2+: "gustos" / "aversiones".

Otro ejemplo de anotación semántica es el siguiente, tomado de McEnery y Wilson (1996:51). En este caso, las categorías semánticas se representan por series numéricas de ocho dígitos[320]:

And	00000000
the	00000000
soldiers	23241000
platted	21072000
a	00000000
crown	21110400
of	00000000
thorns	13010000
and	00000000
put	21072000
it	00000000
on	00000000
his	00000000
head	21030000
and	00000000
they	00000000
put	21072000
on	00000000
him	00000000
a	00000000
purple	31241100
robe	21110321

Tabla 32. Anotación semántica en la Universidad de Lancaster.

La anotación semántica también puede referirse a las categorías o rasgos semánticos del tipo "animado", "humano", etc., de especial importancia para guiar el análisis sintáctico.

[320] El código 00000000 indica que se trata de una palabra de escaso contenido léxico (*and*, *the*, *a*, *of*, *on*, *his*, *they*, etc.); 13010000, pertenencia al mundo vegetal en general; 21030000 tiene que ver con el cuerpo y sus partes; 21072000, con la actividad física orientada a objetos; 21110321, con ropa exterior masculina; 21110400 se refiere a un sombrero; 23231000, a la guerra y conflictos en general, etc.

Así, por ejemplo, en la *Base de Datos Sintácticos del español actual* (BDS)[321], además de los esquemas sintácticos en que aparece un verbo (por ejemplo, *informar*), tenemos acceso a información sobre características semánticas, como la naturaleza animada o no del sujeto (*San* o *Snan*), del complemento directo (*Dan*), del suplemento (*SPan*); así como la frecuencia de uso en cada uno de los esquemas y subesquemas sintáctico-semánticos:

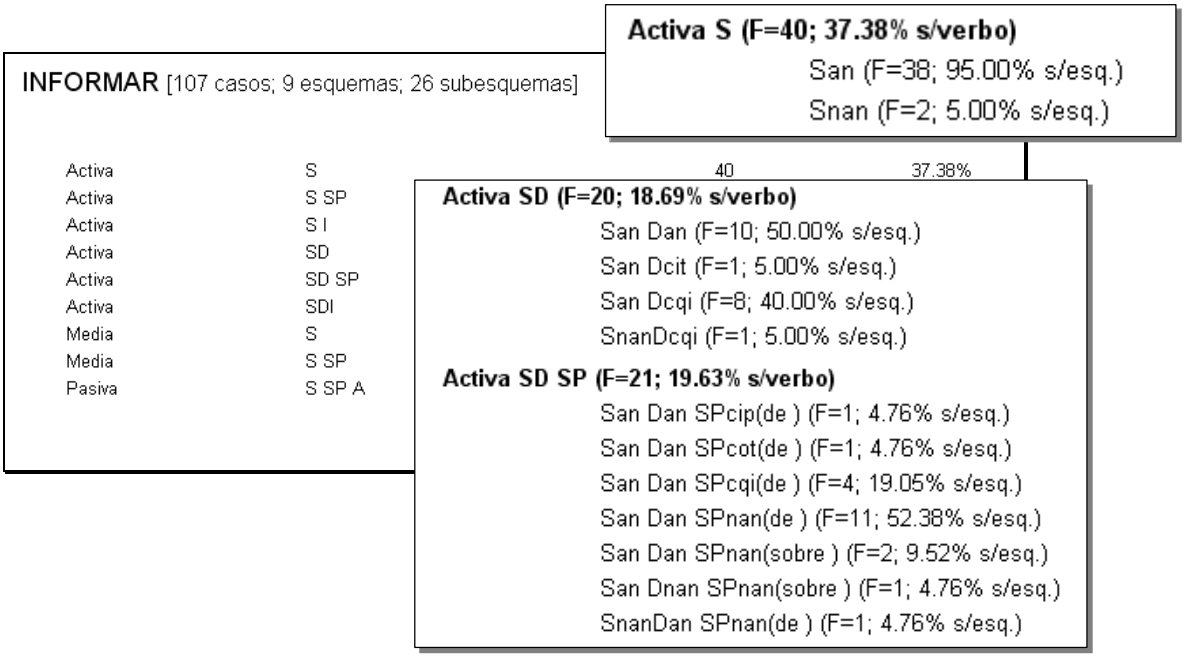


Ilustración 166. Esquemas y subesquemas en que aparece el verbo "informar" en la BDS.

5) *Anotación anafórica*. Este es un tipo de anotación relacionada con el nivel del discurso y que se puede encontrar en varios corpus. Consiste en la identificación de los referentes de las expresiones anafóricas, tales

[321] URL: http://www.bds.usc.es/. ADESSE (Base de datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español) viene a completar la BDS, haciendo precisamente hincapié en la semántica verbal: acepción de los verbos, clase semántica, roles semánticos de los argumentos, etc. Por ejemplo, *informar* es encuadrado dentro de la clase semántica de los verbos de "Comunicación", aparece frecuentemente con un argumento explícito que representa el "emisor", con los rasgos semánticos de "animado" y "concreto" y desempeña la función de sujeto (*vid*. URL: http://adesse.uvigo.es/).

como pronombres, frases nominales, etc. Es decir, en la determinación de qué elementos de un texto aluden al mismo referente (correferencia), hecho de suma importancia para la cohesión de los textos. La anotación anafórica se suele llevar a cabo asignando un mismo índice al pronombre y a la frase nominal que son correferenciales, aunque existen diversos esquemas de anotación. Hasta hace poco no existían programas que permitieran efectuar esta tarea de forma automática o semiautomática. Destaca el programa XANADU, desarrollado por Garside en la Universidad de Lancaster. En el siguiente ejemplo (*vid.* McEnery, Xiao y Tono 2006:38-39), tanto *they* como *their* aluden a una frase nominal mencionada anteriormente (*the married couple*):

(6 the married couple 6) said that <REF=6 **they** were happy with <REF=6 **their** lot

Ilustración 167. Ejemplo de anotación anafórica.

Otro ejemplo de corpus con anotación anafórica[322]:

S.1 (0) The state Supreme Court has refused to release {1 [2 Rahway State Prison 2] inmate 1}} (1 James Scott 1) on bail .
S.2 (1 The fighter 1) is serving 30-40 years for a 1975 armed robbery conviction .
S.3 (1 Scott 1) had asked for freedom while <1 he waits for an appeal decision .
S.4 Meanwhile , [3 <1 his promoter 3] , {{3 Murad Muhammed 3} , said Wednesday <3 he netted only $15,250 for (4 [1 Scott 1] 's nationally televised light heavyweight fight against {5 ranking contender 5}} (5 Yaqui Lopez 5) last Saturday 4) .
S.5 (4 The fight , in which [1 Scott 1] won a unanimous decision over (5 Lopez 5) 4) , grossed $135,000 for [6 [3 Muhammed 3] 's firm 6], {{6 Triangle Productions of Newark 6} , <3 he said .

Ilustración 168. Anotación anafórica en The Associated Press Treebank (AP).

[322] Tomado de la URL: http://ucrel.lancs.ac.uk/annotation.html

Aquí, también se recurre al uso de índices (1, 2, *n…*) para relacionar distintos constituyentes sintácticos correferenciales o equivalentes desde el punto de vista semántico:

(i i)	índice
[i…]	constituyente (normalmente una frase nominal) que alberga alguna equivalencia semántica
<i	indica un pronombre que tiene un antecedente
>i	indica un pronombre cuyo referente aparece después
{{i i}	frase nominal que mantiene algún tipo de relación con una frase nominal precedente
{i i}}	frase nominal relacionada con otra frase nominal posterior
(0)	barrera anafórica, por lo general, el inicio de un nuevo texto.

6) *Otros tipos de anotación***.** Por último, es posible anotar otro tipo de informaciones lingüísticas en un corpus, como p. ej. los actos de habla que aparecen en diálogos (anotación pragmática); los rasgos de estilo de textos literarios (anotación estilística); los errores presentes en corpus formados por textos de aprendices de una lengua extranjera (anotación de errores); anotación fonética y prosódica (solo en corpus orales); y, finalmente, anotaciones diseñadas específicamente para marcar un fenómeno muy concreto, que depende de la finalidad de la investigación.

El fin último de la lingüística de corpus es poder extraer información sobre el funcionamiento real del lenguaje a partir de muestras de su uso por parte de los hablantes de una lengua determinada. Hoy en día, con un corpus anotado, disponemos de una herramienta fundamental para este fin. En general, para recuperar la información presente en un corpus, se utilizan diferentes herramientas informáticas que proporcionan datos estadísticos como la frecuencia de uso; generan listas de palabras o grupos de palabras; crean concordancias, ejemplos en los que la palabra o secuencia buscada aparece destacada, rodeada de su contexto anterior y posterior; buscan palabras clave por su frecuencia de aparición; o sugieren posibles "colocaciones", combinaciones de palabras que frecuentemente aparecen juntas sin llegar a constituir una unidad de significado. J. Lavid (2005:324-345) o J. Llisterri[323] proporcionan una lista detallada de programas para el tratamiento de corpus, tales como Wordsmith, Concordance, MonoConc Pro, Collocate, ParaConc, Concorder Pro, TextWorld.com, Word List Maker, UltraFind, etc. Con ellos, con nuestros conocimientos sobre el lenguaje y con nuestra capacidad de interpretación estamos en disposición de dar respuesta a interrogantes como qué palabras debe incorporar un diccionario, con qué variante, en qué estructura se emplean, cuántas acepciones tienen, cómo se usan, qué errores cometen con más frecuencia los aprendices de una lengua, sobre qué aspectos hay que incidir en la enseñanza de idiomas, qué caracteriza el estilo de un autor o de una variedad lingüística, qué carga ideológica existe detrás de una palabra…

[323] *Cf.* http://liceu.uab.es/~joaquim/language_resources/lang_res/Herram_TecnTex.html

4. CONCLUSIONES

4. CONCLUSIONES

A lo largo de las páginas precedentes hemos intentado esbozar una modesta aproximación al campo de la Lingüística Computacional. En primer lugar, intentando delimitar el objeto, finalidad, líneas de investigación e historia de la disciplina. En segundo lugar, deteniéndonos en las áreas que configuran el ámbito de trabajo de la LC. Por último, centrándonos en uno de los aspectos que más interés han suscitado, el de la lingüística de corpus. Todos estos puntos se han articulado en sendos materiales en línea (*vid.* anexos para algunas muestras) que sirven desde hace algunos años como apoyo de la docencia de las asignaturas correspondientes de la licenciatura en Lingüística y que son de consulta libre para todos aquellos que quieran utilizarlos en su docencia: Universidad de Santiago, Universidad de Vigo, Universidad de La Coruña, etc.

La aparición del ordenador digital ha marcado un cambio en la forma de contemplar el objeto de estudio de la Lingüística y, con ello, el surgir de una nueva disciplina, de una nueva ciencia (*cf.* nota 5), que ya cuenta con medio siglo de andadura, más todo el saber acumulado antes por el hombre sobre el lenguaje; que propone reglas y formalismos para describirlo de forma exhaustiva; que parte de datos reales, verdaderos; que, mediante su análisis y descripción, busca mejorar los conocimientos de que disponemos sobre él, para lo que elabora teorías nuevas, adapta otras ya existentes, verifica su funcionamiento, descarta las que no sirven –plantea hipótesis que contrasta continuamente con los datos–, y así sucesivamente hasta encontrar aquella teoría o modelo general sobre el lenguaje más satisfactorio, bien sea desde una perspectiva meramente científica o con

vistas al desarrollo de una aplicación determinada, pues también contribuye a la obtención de productos, como fase final de todo el proceso.

El interés por emular el lenguaje humano no es nuevo, sino que sus orígenes hay que buscarlos ya en la Antigüedad. Ese empeño humano por entenderlo y reproducirlo en algún tipo de ingenio artificial ha estado presente en toda nuestra historia. Por eso, el ordenador solo ha sido el catalizador de todos esos intentos que han contemplado los siglos precedentes. Después de la II Guerra Mundial, se aplican las técnicas ideadas para descifrar códigos al lenguaje, pensando que la tarea de decodificarlo iba a ser sencilla. Sin embargo, pronto surgieron las primeras dificultades pero, al mismo tiempo, estímulos para desarrollar nuevas vertientes de trabajo: no solo los aspectos más formales del lenguaje (sintaxis) debían recibir tratamiento, también otros en principio menos sistematizados (semántica, pragmática), pero fundamentales para la comunicación humana: sin atender al significado, sin considerar el conocimiento del mundo, sin atender a las intenciones de los hablantes, a la forma de urdir el discurso no se puede entender el lenguaje.

Estos fracasos fueron necesarios, en cierta medida, para que los investigadores se dieran cuenta de que tenían que ajustar los objetivos iniciales a otros más modestos o más realistas. Al fin y al cabo, la finalidad última de la LC es entender mejor el lenguaje y desarrollar herramientas útiles para su tratamiento, así como aplicaciones en las que el lenguaje es una parte fundamental. Un requisito previo resulta fundamental en esta tarea: contar con formalismos de representación de la información lingüística, aplicados a los diferentes niveles en que se divide el estudio del lenguaje: desde el fonético-fonológico hasta el

pragmático, pasando por el morfológico, el sintáctico y el semántico. Esta concepción modular configura las distintas áreas de interés de la LC, de las que hemos intentado presentar una visión panorámica, deteniéndonos más en algunos aspectos en detrimento de otros. Y es que a medida que progresábamos en las implicaciones del tratamiento computacional del lenguaje, nos íbamos dando cuenta de la complejidad inherente al estudio del mismo, más aún cuando se quiere que un programa informático sea capaz de capturar las sutilezas del mismo. Somos conscientes de que apenas hemos mencionado las tecnologías del habla o las líneas de trabajo más aplicadas (traducción automática, interfaces en lenguaje natural, etc.). Sin embargo, como una aproximación que esta tesis es o pretende ser al campo de la LC, confiamos en poder seguir ahondando en la apasionante intersección de la Lingüística y la Informática.

Aunque la perspectiva adoptada en este acercamiento es la de la Lingüística, nos ha parecido especialmente interesante descubrir las aportaciones de otras disciplinas, como la Informática, la Inteligencia Artificial, la Lógica, las Matemáticas, la Psicología, etc., en un marco de trabajo más general e interdisciplinar, el de la ciencia cognitiva. Y también, dentro de la propia Lingüística, llama la atención la riqueza de matices que se aprecia y la integración de enfoques: desde la gramática tradicional hasta la cognitiva, pasando por la estructural o la generativa, todas las corrientes han contribuido al desarrollo de la disciplina. El resultado son las diferentes líneas de investigación que conforman el campo: cada una pone el énfasis en aspectos de índole diversa: unas más en el lado cognitivo, otras en el psicológico, en el humano, en la naturaleza lógica, en los aspectos matemáticos, estadísticos, tecnológicos, etc. Así, se distinguen trabajos desde una

vertiente más teórica (Lingüística Computacional propiamente dicha), más aplicada (Procesamiento del Lenguaje Natural), más cognitiva (Inteligencia Artificial), más auxiliar (Lingüística Informática), más centrada en la lengua hablada (Tecnologías del habla), más comercial (Industrias de la lengua), más preocupada por el desarrollo de recursos básicos (Ingeniería lingüística). Precisamente, el diseño y recopilación de corpus lingüísticos es un apartado que nos ha llamado especialmente la atención, en tanto que hoy en día constituyen una fuente de información privilegiada para los estudios lingüísticos. Por este motivo, nos hemos detenido en particular en presentar una visión más exhaustiva del mismo: desde los primeros corpus que mostraban cierta modestia en unos planteamientos, por otra parte, muy bien fundamentados, hasta los recursos más recientes, en los que los nuevos medios (Internet principalmente) permiten compilar cantidades ingentes de datos; pasando por los problemas y ventajas que supone trabajar con corpus, la necesidad de definir claramente los criterios de selección de los textos o las posibilidades de explotación que implica disponer de corpus anotados con distintos tipos de información.

Con la Lingüística como referente en este acercamiento, hay que reconocer que las teorías surgidas en su seno en ocasiones han resultado insuficientes para los propósitos de la LC, que ha tenido que desarrollar sus propias propuestas o acudir a los conocimientos y técnicas que, con otros objetivos, se habían empleado en otros campos del saber. Especialmente notable ha sido la incidencia de la Inteligencia Artificial, que en su afán por simular las capacidades cognitivas humanas, ha otorgado un papel destacado al lenguaje, conducta inteligente por excelencia. Sus principales contribuciones se han dejado

sentir sobre todo en los aspectos relacionados con la representación del conocimiento.

Las dificultades del tratamiento computacional del lenguaje son muchas (ambigüedades de todo tipo, elipsis, información que depende del contexto, conocimientos generales, etc.), pero basta echar la vista atrás para ver el recorrido que se ha realizado en unas pocas décadas. Los resultados están muchas veces presentes en nuestras vidas diarias: un ascensor que nos indica mediante una voz sintetizada el piso en el que estamos, un teléfono móvil que nos permite la marcación por voz, un programa que lee en voz alta los contenidos de un documento a una persona con problemas de visión, buscadores de información a través de la web que integran tratamiento del lenguaje para mejorar los resultados de las búsquedas, diccionarios electrónicos, programas de ayuda a la traducción, servicios automáticos de información operados mediante la voz, etc.

En el ámbito específico de la Lingüística, los beneficios que se derivan del empleo de ordenadores son numerosos: bases de datos para la elaboración de diccionarios, conjugadores, lematizadores, analizadores de diverso tipo, programas de concordancias, programas para la síntesis y el reconocimiento del habla, corpus electrónicos, etc. Sin duda, el abanico de posibilidades es tan amplio como abierto es el lenguaje. Como dice José Saramago, “la importancia que puede tener usar una palabra en vez de otra, aquí, más allá, un verbo más certero, un adjetivo menos visible, parece nada y finalmente lo es todo”.

Si leemos el diccionario de nuestros doctos académicos, nos dice: "ENSEÑAR. Instruir, doctrinar, amaestrar con reglas o preceptos", definición esta que no cambió desde Autoridades. Y: "ESTUDIANTE. Persona que cursa estudios en un establecimiento de enseñanza". Estas definiciones, claras, sí, pero ¡oh cuánto desprovistas de humanidad!, no me gustan y no son las perlas de las que hablaba hace un momento.

Solo nos hablan de reglas y preceptos y, si bien la Academia no se equivoca en su etimología que propone al derivar "enseñar" del latín "insignāre", esto es, "señalar", se equivoca del todo en lo que en la Facultad de León y en el Departamento de Lengua entendemos por enseñanza. Nosotros preferimos equivocarnos con Covarrubias en la etimología. ¿Qué dice el buen canónigo?

> ENSEÑAR. Dotrinar, cuasi ensenar, vel insinuare; porque el que enseña mete en el seno (conviene a saber en el corazón) la dotrina, y el que la oye la guarda allí y en su memoria.

> ESTUDIANTE. El que estudia. Algunas veces se toma por el que es oyente, y otras por el muy docto, que aunque lo sea, siempre estudia y nunca le parece que ha llegado a saber lo que basta, descubriendo cada día cosas nuevas.

[…] nos unimos a Covarrubias en el sentido en que hemos intentado, cada uno desde nuestro saber y campo, enseñaros las frías reglas de la Gramática desde el corazón, para que os queden en la memoria.

(Janick Le Men Loyer)

5. ANEXOS

5. ANEXOS

A continuación se recogen algunos de los materiales utilizados en la docencia de las asignaturas Lingüística Computacional I y Lingüística Computacional II[1]:

-Anexo I. Contiene un esquema general de la organización de la docencia de las asignaturas de Lingüística Computacional:

-Breve descriptor de las dos asignaturas.

-Estructura del curso (contenidos teóricos, actividades o prácticas presenciales, actividades o ejercicios de evaluación para el alumno, referencias bibliográficas, enlaces de interés y pantalla de avisos relacionados con la secuenciación del curso e informaciones de interés), con algunos ejemplos de las diferentes partes.

-Esquema con cierto detalle de los diferentes temas teóricos que componen el curso.

-Ejemplos de los diferentes tipos de actividades que se proponen (de respuesta breve, de verdadero o falso, de opción múltiple, cuestiones para desarrollar, etc.), bien para que el alumno las lleve a cabo en clase bajo la guía del profesor (prácticas presenciales), bien para que las efectúe de forma autónoma para su posterior evaluación.

-Modelo de cuestiones planteadas a los estudiantes para evaluar la asignatura.

[1] Todo el material está disponible en Internet, en la web del Dpto. de Filología Hispánica y Clásica de la Universidad de León. URL: http://www3.unileon.es/dp/dfh/Milka/Milka.htm

-Anexo II: contiene con más detalle los aspectos referidos específicamente a la docencia de lingüística de corpus:

-Presentación: objetivos, lecturas recomendadas, materiales complementarios.

-Esquema del tema: propuesta de organización de contenidos teóricos.

-Prácticas y actividades: actividades presenciales y de evaluación propuestas en relación con los contenidos de la materia.

5.1. Anexo I: Lingüística Computacional

5.1.1. Asignaturas

A. Lingüística Computacional I (Plan 2001)

- Carácter: Troncal
- Curso: 4º de Lingüística (1er curso de Lingüística)
- Créditos: 6 (3 teóricos y 3 prácticos)
- Cuatrimestre: 2º

Contenidos:

Introducción a los conceptos, teorías, métodos y herramientas básicos relativos a la aplicación de los ordenadores al estudio del lenguaje.

Objetivos:

- Presentar una visión general del campo de la Lingüística Computacional (LC).
- Introducir las bases teóricas de la disciplina.
- Describir algunas de las aplicaciones menores de la LC, como los correctores ortográficos, los diccionarios electrónicos o los programas informáticos para la enseñanza-aprendizaje de lenguas asistido por ordenador.

B. Lingüística Computacional II (Plan 2001)

- Carácter: Optativa
- Curso: 5 º de Lingüística (2º curso de Lingüística)
- Créditos: 6 (3 teóricos y 3 prácticos)
- Cuatrimestre: 1º

Contenidos:

Principales aplicaciones y recursos relacionados con el uso de ordenadores para el estudio de la lengua.

Objetivos:

- Principales aplicaciones de la Lingüística Computacional, tales como la traducción automática y las interfaces en lenguaje natural.
- Algunos recursos lingüísticos imprescindibles en cualquier trabajo computacional, en especial los corpus electrónicos.
- Otros recursos relacionados con el tratamiento computacional del lenguaje: programas de traducción automática, programas de concordancias, sistemas de dictado automático, etc.
- Tratamiento computacional de la lengua hablada.
- Análisis y detección de los problemas que presenta el lenguaje natural para su tratamiento computacional en sus diferentes niveles.

5.1.2. Estructura del curso

1. CONTENIDOS TEÓRICOS (*vid. infra* ejemplo 1)
    - Conceptos clave
    - Tablas resumen
    - Sugerencias bibliográficas
2. ACTIVIDADES PRÁCTICAS (*vid. infra* ejemplo 2)
3. EJERCICIOS DE EVALUACIÓN (*vid. infra* ejemplo 3)
4. BIBLIOGRAFÍA
5. ENLACES DE INTERÉS (*vid. infra* ejemplo 4)
6. AVISOS (*vid. infra* ejemplo 5)

Ejemplo 1: contenidos teóricos

1.2. OBJETIVOS DE LA LC: LC TEÓRICA Y LC APLICADA

En este apartado nos centraremos en:

- Distinguir los objetivos de la LC: teóricos y aplicados

- Relacionarlos con las dos orientaciones principales de la disciplina: LC Teórica y LC Aplicada

---

La doble vertiente, lingüística e informática, que hemos comentado en el apartado anterior, se concreta en los dos objetivos o motivaciones con los que se puede abordar el trabajo en LC, objetivos teóricos y objetivos aplicados, que han dado lugar a que la LC se divida en dos ramas:

- LC Teórica, más vinculada a la Lingüística
- LC Aplicada, más relacionada con la Informática y la Inteligencia Artificial

Lingüística Computacional (LC)	
LC Teórica	LC Aplicada
Objetivos teóricos	Objetivos aplicados
Perspectiva de la Lingüística	Perspectiva de la Informática

OBJETIVOS TEÓRICOS

- También llamados “científicos”.
- Son generales, independientes de cualquier aplicación.
- Constituyen el ámbito de trabajo de la LC Teórica.
- Según R. Grishman (1991:16-17), se concretan en:
  - Probar las gramáticas que propone la Lingüística Teórica.
  - Investigar los procesos psicológicos que intervienen en la producción y comprensión del lenguaje dentro del marco general de la Ciencia Cognitiva.
  - Estudiar la forma de representar el conocimiento general o del mundo.

OBJETIVOS APLICADOS

- También llamados “tecnológicos” o “aplicaciones orientadas a la ingeniería”.
- Se trata de sistemas prácticos o programas informáticos específicos.
- Constituyen el ámbito de trabajo de la LC Aplicada.
- Según R. Grishman (1991:15-16), las tres aplicaciones principales de la LC son:
  - la traducción automática
  - la recuperación de información
  - las interfaces hombre-máquina.

Completa este punto con la siguiente lectura:

R. Grishman (1991 [1986]): “Los objetivos de la lingüística computacional”, *Introducción a la lingüística computacional*, Madrid: Visor, págs. 15-17.

LC TEÓRICA

- Es lo que se entiende por LC en sentido estricto o LC por antonomasia.
- Toma sus temas de trabajo de la Lingüística Teórica y de la Ciencia Cognitiva.
- Las aportaciones de la Psicología Cognitiva, en especial de la Psicolingüística, también son de especial relevancia, lo que se ha traducido en el surgimiento de una nueva ciencia, la Psicolingüística Computacional.
- Su objetivo es proporcionar una explicación del funcionamiento del lenguaje en alguno de sus niveles: fonético y fonológico, morfológico, sintáctico, semántico, léxico, pragmático, etc. (*vid. supra* la definición 11 de la ACL y la 18 de Wilson y Keil).
- Este objetivo general, según X. Gómez Guinovart (2000a:223), se concreta en:
  - La elaboración de teorías o modelos lingüísticos que cumplan dos requisitos
    - ser formales
    - ser adecuados para su implementación informática
  - La descripción de fenómenos lingüísticos concretos en el marco de las teorías o modelos anteriores.
  - La comprobación automatizada de la consistencia de una teoría lingüística.

LC APLICADA

- Se trata de una vertiente de la LC que posee una clara orientación tecnológica, lo que ha motivado que hoy en día con frecuencia se aluda a ella con nombres como *ingeniería lingüística* o *tecnología lingüística o del lenguaje (humano)*.
- Se centra en los aspectos prácticos que se puedan derivar de la simulación de la conducta lingüística con medios informáticos.
- Su objetivo es crear productos informáticos que incorporen algún componente en el que intervenga el lenguaje, oral o escrito.
- Uno de sus principales retos es mejorar la comunicación entre personas y ordenadores mediante el uso del lenguaje natural.
- Consiste en métodos, técnicas, herramientas y aplicaciones.
- Según X. Gómez Guinovart (2000a:223-224), las principales aplicaciones (*vid.* también la definición 11 de la ACL y la 18 de Wilson y Keil), que este autor agrupa en 4 categorías, son:
    - Programas para la comprensión y generación de enunciados: consulta a bases de datos, sistemas de diálogo, etc.
    - Programas relacionados con las tecnologías del habla: dictado automático, conversión de texto en voz, etc.

- Herramientas para el procesamiento documental: correctores ortográficos y estilísticos, programas para la generación automática de resúmenes, sistemas de extracción y recuperación de información textual, etc.
- Herramientas para el procesamiento plurilingüe: programas para la enseñanza de lenguas asistida por ordenador o para la creación de ejercicios, programas de ayuda a la traducción, etc.

Completa este punto con las siguientes lecturas:

	F. Uszkoreit (1996, 2000): "What is Computational Linguistics?".
	X. Gómez Guinovart (2000a): "Lingüística Computacional", en F. Ramallo, G. Rei Doval e X. P. Rodríguez Yáñez (coords.), *Manual de Ciencias da Linguaxe*, Vigo: Edicións Xerais de Galicia, págs. 223-224.

El siguiente cuadro, elaborado tomando como referencia las opiniones de R. Grishman (1991 [1986]:15-17), F. Uszkoreit (1996, 2000) y M. A. Martí e I. Castellón (2000:3-4), sintetiza las principales características de la LC Teórica y la LC Aplicada:

LC TEÓRICA	LC APLICADA
• Es lo que se entiende por LC propiamente dicha	• También se denomina ingeniería lingüística o tecnología del lenguaje
• Trata de objetivos científicos generales	• Trata de objetivos tecnológicos concretos
• Toma sus temas de la Lingüística Teórica, de la Psicolingüística y de la Psicología Cognitiva	• Las disciplinas en las que se apoya son la Informática y la Inteligencia Artificial
• Elabora modelos computacionales que intentan simular el lenguaje	• Desarrolla sistemas informáticos que incorporan módulos que tratan el lenguaje
• Su meta es explicar los procesos que intervienen en la comprensión y generación del lenguaje	• Busca obtener productos útiles, prácticos

Algunas líneas de investigación son:

- Elaborar teorías y formalismos que den cuenta del lenguaje en toda su complejidad y al mismo tiempo sean susceptibles de recibir un tratamiento informático
- Simular aspectos concretos del lenguaje
- Evaluar sus teorías o las propuestas por la Lingüística Teórica
- Investigar los procesos cognitivos que intervienen en el procesamiento del lenguaje
- Estudiar los problemas de representación del conocimiento del mundo

- Algunas aplicaciones son:
- Traducción automática
- Recuperación y extracción de información
- Interfaces hombre-máquina
- Enseñanza de lenguas asistida por ordenador
- ...

Ejemplo 2: actividades prácticas

Práctica 4. Diccionarios en línea: el caso del español

Diccionarios electrónicos en línea

- *Diccionario de la Lengua Española*, Real Academia Española, Madrid: Espasa-Calpe, 2001, 22ª ed. URL: http://www.rae.es
- *Diccionario panhispánico de dudas*, Real Academia Española, Madrid: Santillana, 2005. URL: http://www.rae.es
- *Nuevo Tesoro Lexicográfico de la Lengua Española*, Real Academia Española. URL: http://buscon.rae.es/ntlle/SrvltGUILoginNtlle
- *Clave, Diccionario de uso del español actual*, Madrid: SM, 1997. URL: http://clave.librosvivos.net
- *Diccionario Salamanca de la Lengua Española*, Salamanca: Santillana-Universidad de Salamanca, 1996. Se puede consultar en el sitio web "Centro Nacional del Información y Comunicación Educativa" (Ministerio de Educación, Cultura y Deporte), sección "Recursos para el aula: Diccionarios". URL: http://fenix.cnice.mec.es/diccionario/
- *Diccionario General de la Lengua Española*, Barcelona: Vox-Biblograf, 1997. También ofrece otros diccionarios: sinónimos y antónimos, ideológico, español-inglés, español-francés, español-alemán, español-italiano, castellano-catalán y castellano-portugués. URL: http://www.diccionarios.com/
- Diccionarios del periódico *El Mundo* (de la editorial Espasa Calpe, 2001). URL: http://www.elmundo.es/diccionarios
- Diccionario de sinónimos y antónimos en línea, Signun Cía Ltda. URL: http://www.lenguaje.com/herramientas/tesauro.php
- Diccionarios en línea, Universidad de Oviedo. URL: http://www.etsimo.uniovi.es/dic/
- Diccionarios de variantes del español, página con enlaces de José Ramón Morala (Universidad de León). URL: http://www3.unileon.es/dp/dfh/jmr/dicci/0000.htm
- Varilex, Diccionario de español de geosinónimos, Signum Cía. Ltda. URL: http://www.lenguaje.com/glosario/glosario.php. Desarrollado a partir del Proyecto Varilex (Variación léxica del español en el mundo), coordinado por Hiroto Ueda (Universidad de Tokio).
- Diccionario Argentino-Español para españoles, Alberto J. Miyara. URL: http://www.elcastellano.org/miyara/
- Diccionario del Español usual en México, Luis Fernando Lara. URL: http://mezcal.colmex.mx/Scripts/Dem/principal.htm. También en: http://www.cervantesvirtual.com/FichaObra.html?Ref=3161
- DiCE, Diccionario de Colocaciones del Español, desarrollado en la Universidad de La Coruña bajo la dirección de Margarita Alonso Ramos. URL: http://www.dicesp.com/
- Vademécum de Español Urgente, Agencia EFE, Fundéu. URL: http://www.fundeu.es/esurgente/lenguaes/
- Cambridge Dictionaries Online, Cambridge University Press. URL: http://dictionary.cambridge.org/
- WordReference Dictionaries: http://www.wordreference.com/
- OneLook. Dictionaries. URL: http://www.onelook.com
- Logos, Multilingual E-Translation Portal. Diccionario, diccionario para niños, glosarios. URL: http://www.logos.it/lang/transl_en.html
- ForeignWord.com, El Portal del Idioma. Diccionarios, glosarios, enlaces a diccionarios. URL: http://www.foreignword.com/es/
- The Alternative Dictionaries, Diccionario multilingüe de jergas y palabrotas. URL: http://www.alternative-dictionaries.net/

Trabaja con los diccionarios en línea: (*práctica presencial*)

a) Consulta algunos de los diccionarios electrónicos de la página anterior y observa su funcionamiento.

b) ¿Qué diccionario o diccionarios recomendarías para…?

-Conocer los adjetivos que más frecuentemente se utilizan con *odio*.

DICCIONARIO:

-Saber cómo hay que escribir: *medio ambiente* o *medioambiente*

DICCIONARIO:

-Saber qué palabras tienen un significado similar a *fulgor*

DICCIONARIO:

-Buscar una imagen de una avispa

DICCIONARIO:

-Encontrar el significado y la traducción de la expresión *vivir del cuento*

DICCIONARIO:

-Conocer los tacos del español

DICCIONARIO:

-Saber cómo escribir en español el nombre de la capital de Islandia en un medio de comunicación

DICCIONARIO:

c) Busca, compara y analiza:

Busca "líbido" en el *Diccionario* de la RAE, en el de *El Mundo*, en el *Cla* y en el *Diccionario Salamanca*. Comenta los resultados.

Analiza las posibilidades de búsqueda que ofrece el *Vademécum*.

Busca palabras terminadas en *-itis* utilizando el asterisco (*itis) en *Diccionario* de la RAE y en el *Clave*. Comenta los resultados.

Compara las posibilidades de búsqueda del *Diccionario* de la RAE del *Diccionario del español usual en México*.

Compara los resultados de buscar "saco" en el *Diccionario de sinónimos y antónimos en línea* de Signun y en el *Diccionario de sinónimos y antónimos* de Vox.

Busca "alfajor" en el *Diccionario argentino-español para españoles*, en el *Diccionario* de la RAE y en el *Clave*. Analiza la presencia de hipervínculos en la definición, la información que proporcionan, la presencia o ausencia de ejemplos de uso y de etimología.

Analiza las posibilidades de búsqueda de *Onelook*.

Busca "camión" en *Varilex*. Analiza los resultados.

d) Ahora presta especial atención a:

- Clave
- Salamanca
- El Mundo
- WordReference

Analiza los siguientes aspectos de cada uno de ellos:

- Forma de acceso a la información: alfabética, por lema, otras opciones (palabras que empiezan por, terminan en…).
- Entrada:
  - Etimología
  - Información sobre pronunciación
  - Información gramatical: clase de palabra, género…
  - Información sintáctica: estructuras
  - Frases y locuciones
  - Ejemplos
  - Marcas de uso: geografía, registro…
- Hipertextualidad: enlaces a otras partes del diccionario, enlaces a otros sitios de Internet.
- Multimedia: imágenes, sonido, posibilidad de modificar el tamaño, color, etc. de las fuentes, opciones de presentación.
- Interacción: posibilidad de contribución del usuario

e) Observa las siguientes frases.

1) Una casa ecológica que lo **flipas**.
   http://cvc.cervantes.es/aula/palabra_por_palabra/pie_pagina.asp?pxp=9
2) Los vecinos del centro de A Coruña están hartos del **botellón** que cada fin de semana organizan cientos de jóvenes bajo sus ventanas.
3) Según la Fundación del Español Urgente, Fundéu, la pronunciación correcta de 'Guei' es '**Gay**'.
4) El beso tradicional es el simple **ósculo** que se realiza sobre los labios de otra persona.
5) Con Z de Zapatero: "**ProsperidaZ**, **competitividaZ**, empleo de **calidaZ**..." y un largo etcétera de logros del Gobierno son así citados en un vídeo en el que el presidente aparece con chaqueta, camisa blanca… **[Palabras terminadas en -dad]**
6) Ayer, los populares no podían disimular el «**subidón**» que les ha proporcionado la Conferencia Política del fin de semana.
7) **Aeromoza** de Singapur Airlines muestra Airbus A380 con 12 compartimientos lujosos.
8) Londres **hace de menos** al Barça.
9) Aguirre: "En el PP me consideran **lideresa** nacional"

Busca los términos destacados en negrita en los diccionarios de d). En WordReference, además de la definición en español, busca el equivalente en otra lengua.

Ejemplo 3: ejercicios de evaluación

Actividad 6. Hitos en la historia de la LC

Versión de texto

Lee la información del apartado 1.4., Evolución de la LC, y, después, completa los siguientes enunciados.

1. J. _____, uno de los fundadores de la IA, desarrolló el lenguaje de programación _____.
2. El corpus _____ de inglés americano representa el inicio de la disciplina que más tarde se conocerá como Lingüística de corpus.
3. En la Universidad de _____, IBM llevó a cabo la primera demostración pública de un sistema de traducción automática.
4. Las teorías del lingüista N. _____ y de su maestro Z. _____ han ejercido una gran influencia en el campo de la LC.
5. Las primeras investigaciones en traducción automática aplicaron técnicas _____.
6. W. _____ elaboró en _____ un famoso _____, "Translation", que se considera el punto de inicio de los trabajos en LC.
7. Los primeros investigadores en el campo de la LC eran muy _____ sobre los resultados que esperaban obtener.
8. Los laboratorios _____ fueron pioneros en el reconocimiento del habla.
9. Los _____ son el léxico y las estructuras gramaticales propias de un ámbito determinado, p. ej., de los partes meteorológicos.
10. El método _____ traduce de una lengua a otra palabra a palabra.
11. El paradigma _____ es aquel que se basa en el empleo de estadísticas.
12. C. _____ desarrolló la _____ de la _____, con la que sentó las bases del acercamiento _____ al tratamiento computacional del lenguaje.
13. La conocida como Escuela de _____, liderada por R. _____, destaca por los estudios sobre representación del conocimiento del mundo.
14. Durante la segunda etapa de la LC, el nivel lingüístico que recibió más atención fue el _____.
15. El paradigma denominado _____ se caracteriza por el uso de reglas.
16. _____ es quizá el lenguaje de programación más extendido en IA.

Ejemplo 4: enlaces de interés

Asociaciones

- ACL, The Association for Computational Linguistics
- EACL, European Chapter of the ACL
- NAACL, North American Chapter of the ACL
- SEPLN, Sociedad Española para el Procesamiento del Lenguaje Natural
- AAAI, American Association for Artificial Intelligence
- AEPIA, Asociación Española de Inteligencia Artificial
- ACH, Association for Computers and the Humanities
- OESI, Oficina de Español en la Sociedad de la Información

Centros de investigación

- CLiC, Centre de Llenguatge i Computació, Universitat de Barcelona
- TALP, Centre de Tecnologies i Aplicacions del Llenguatge i la Parla, Universitat Politècnica de Catalunya
- SLI, Seminario de Lingüística Informática, Universidad de Vigo
- LLI, Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid
- Lab.Lingua, Laboratorio de Lingüística Informática, Universidad de Alicante

Grupos de investigación

- Grupo de Estructuras de Datos y Lingüística Computacional, Dpto. de Informática y Sistemas, Universidad de Las Palmas de Gran Canaria
- GilcUB, Grup d'Investigació en Lingüística Computacional, Universitat de Barcelona
- Natural Language Research Group, Departament de Llenguatges i Sistemes Informatics, Universitat Politècnica de Catalunya
- Natural Language Processing Group, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia y Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante
- Julietta Research Group in Natural Language Processing, Dpto. de Filología Inglesa, Universidad de Sevilla
- GPLSI, Grupo de Investigación en Procesamiento del Lenguaje y Sistemas de Información, Universidad de Alicante
- Grupo de PLN de la UNED

Publicaciones

- *Computational Linguistics*, revista oficial de la ACL
- *Revista de Procesamiento del Lenguaje Natural*, SEPLN
- *ACL Anthology*, A Digital Archive of Research Papers in Computational Linguistics

Ejemplo 5: avisos

5.1.3. Contenidos teóricos

A. Temas de Lingüística Computacional I

i. INTRODUCCIÓN A LA LINGÜÍSTICA COMPUTACIONAL

- Definición y características básicas.
- Objetivos. LC Teórica y LC Aplicada.
- Principales líneas de investigación.
- Evolución histórica.

ii. LINGÜÍSTICA COMPUTACIONAL TEÓRICA

- Las áreas de la Lingüística Computacional.
- Morfología computacional.
- Sintaxis computacional.
- Semántica computacional.
- Pragmática computacional.

iii. DICCIONARIOS ELECTRÓNICOS

- El concepto de diccionario electrónico: características básicas.
- Ventajas e inconvenientes: los diccionarios electrónicos vs. los diccionarios en papel.
- Los diccionarios en línea: el caso del español.
- Los diccionarios en CD-ROM: algunos ejemplos para el español.
- Análisis de algunos diccionarios.

iv. CORRECTORES ORTOGRÁFICOS

- La corrección ortográfica.
- La corrección gramatical.
- La corrección de estilo.
- Su incorporación a los procesadores de texto.
- Correctores ortográficos en línea.

v. NUEVAS TECNOLOGÍAS Y ENSEÑANZA DE LENGUAS

- La enseñanza de lenguas asistida por ordenador (ELAO).
- Internet como recurso didáctico: usos didácticos de la red para la elaboración de actividades.
- Los programas de autor para la creación de actividades.
- Integración de texto, imágenes, sonido, vídeo e hipervínculos.
- La evaluación: criterios para evaluar recursos existentes.

B. Temas de Lingüística Computacional II

i. INTRODUCCIÓN A LA LINGÜÍSTICA COMPUTACIONAL APLICADA

ii. LA LINGÜÍSTICA DE CORPUS

iii. LAS TECNOLOGÍAS DEL HABLA

iv. LA TRADUCCIÓN AUTOMÁTICA

v. LAS INTERFACES EN LENGUAJE NATURAL

5.1.4. Actividades

A. Prácticas (ejemplos)

A.1. Práctica 2

A.2. Práctica 6

B. Ejercicios de evaluación (ejemplos)

B.1. Respuestas breves

B.2. Verdadero / Falso

B.3. Opción múltiple

B.4. Otros

A.1. Práctica 2

Lingüística computacional. Nuevas tecnologías y enseñanza de lenguas 1

Los sistemas de autor (II). Hot Potatoes

Hot Potatoes, Versión 6, Half-Baked Software, University of Victoria, Humanities Computing and Media Centre, Canadá.

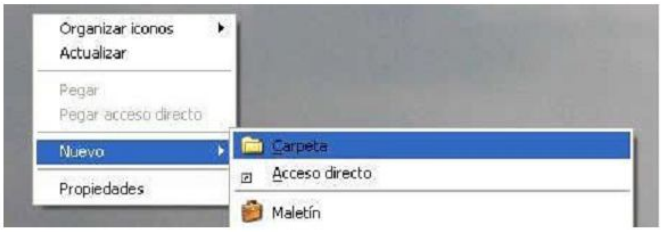
- Es uno de los sistemas de autor más empleados para el diseño de actividades interactivas en el ámbito de la enseñanza de idiomas.
- Se trata de un paquete informático que incluye 6 aplicaciones para crear ejercicios interactivos.
- La última versión del programa consta de seis aplicaciones para crear diferentes tipos de ejercicios:
  - **JMatch**: permite crear actividades de relacionar.
  - **JCloze**: permite crear actividades de rellenar huecos.
  - **JQuiz**: permite crear actividades basadas en preguntas de 4 tipos diferentes (respuesta breve, elección múltiple…).
  - **JCross**: permite crear crucigramas.
  - **JMix**: permite crear actividades basadas en la reconstrucción de frases o textos previamente desordenados.
  - **The Masher**: permite integrar diferentes actividades en una misma unidad.

- Programa: http://hotpot.uvic.ca/
- Descarga: http://hotpot.uvic.ca/index.php#downloads
- Creación de actividades para el autoaprendizaje, Noelia González Verdejo, Universidad de León: http://www3.unileon.es/dp/dfh/noelia/met/trab2/trabajo2.htm
- Tutorial en HTML para leer en línea: http://platea.pntic.mec.es/~iali/CN/HotPot60/tutorial.htm
- Aula de letras, José M.ª González-Sema Sánchez: http://www.auladeletras.net/potatoes.html
- Materiales interactivos para atención a la diversidad en la ESO. Incluyen tutorial y ejemplos de actividades: http://www.educarm.es/materiales_diversidad/start.htm
- Actividades Hot Potatoes: http://www.educa.madrid.org/portal/web/educamadrid/hotpotatoes
- Rincón Hot Potatoes Aragón: http://www.catedu.es/gestor_recursos/public/hotpotatoes/principal.php
- Hot Potatoes, Aula 21: http://www.aula21.net/segunda/hotpotatoes.htm
- Tutorials and other resources on Hot Potatoes: http://hotpot.uvic.ca/tutorials6.php
- Hot Potatoes and TextToys Exercises, Cyberteacher's Website: http://www.cyberteacher.it/esercizi_eng.htm
- Technically interesting Hot Potatoes pages, Glenys Hanson, Centre de linguistique appliquée, Université de Franche-Comté: http://eolf.univ-fcomte.fr/

Sigue las instrucciones para diseñar ejercicios con Hot Potatoes.

1. Para empezar, crea una **carpeta** en el escritorio del ordenador para guardar tus ejercicios y el material relacionado con ellos: imágenes, sonidos, etc.

Botón derecho del ratón sobre el escritorio > Nuevo > Carpeta. Ponle un nombre identificativo.



2. Si no está instalado en tu ordenador el **programa**, vete al sitio web de Hot Potatoes, descarga la última versión (sección "Downloads" > Hot Potatoes 6.2 installer) e instálala: http://hotpot.uvic.ca/index.php#downloads

"abrir"/"ejecutar", "setup", "spanish".

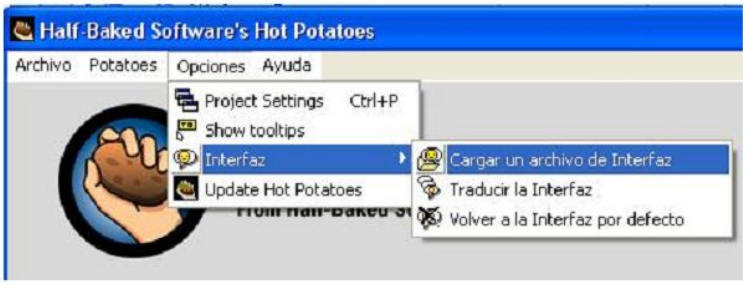
3. Abre el programa pulsando sobre el **icono** de Hot Potatoes que aparecerá en el escritorio de tu ordenador: una mano con una patata. Una vez instalado el programa, para mantenerlo actualizado cada vez que esté disponible una nueva versión, es suficiente con seguir las instrucciones del gráfico que tienes a continuación: **"opciones" > "update Hot Potatoes"**.

4. Antes era necesario registrarse para a través de la opción **"Register"** de la página web del programa para poder utilizar todas las posibilidades que ofrecen los programas y no sólo un número limitado de las mismas. Ahora el nombre de usuario se descarga automáticamente con el programa. Desde el menú de **Ayuda** se accede al registro.

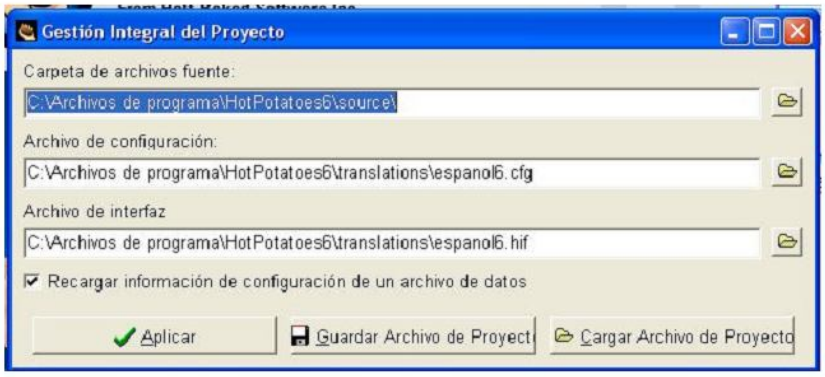
5. Por defecto, las instrucciones y comentarios para el alumno aparecen en inglés. Para visualizarlos en español, abre Hot Potatoes, selecciona la pestaña **Opciones** > **"Project settings"**.

6. En la ventana que aparece, en **"Archivo de configuración"** está seleccionado el archivo en inglés. Pulsa sobre la carpeta y vete a **"translations"**. Selecciona el archivo en **español** y pulsa sobre **"Abrir"** y después sobre **"Aplicar"**.

7. En el mismo menú de **Opciones**, selecciona después la opción **"Interfaz"** > **"Cargar un archivo de Interfaz"** > **"espanol6.hif"**.



El resultado debería ser:



3.1. Creación de un ejercicio con JMatch

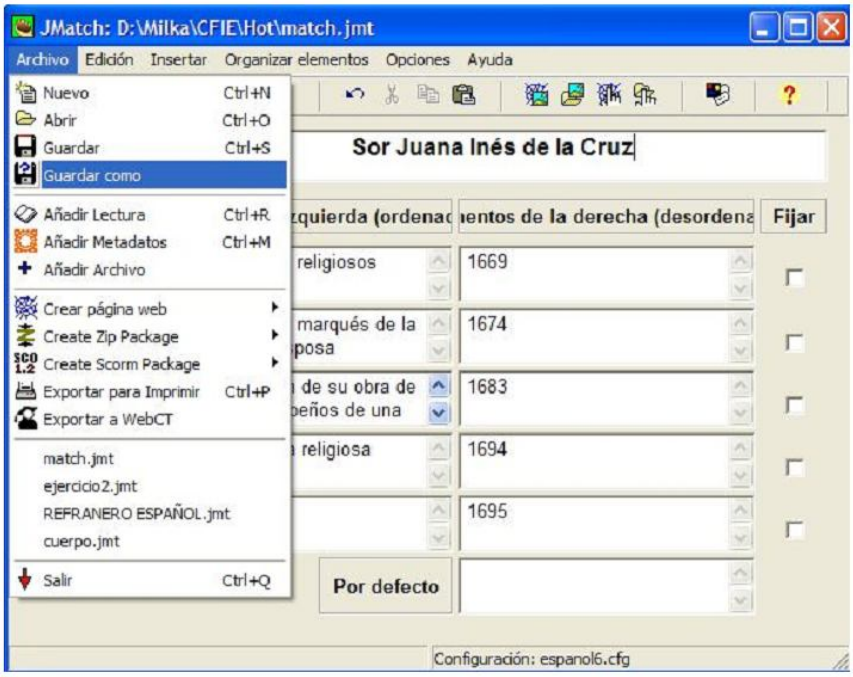
Este programa nos permite crear actividades para relacionar información: texto, imágenes, sonidos, vídeos, etc.

Vamos a crear un ejercicio en el que nuestros alumnos, tras leer la biografía de Sor Juana, relacionen acontecimientos de su vida con las fechas en que estos ocurrieron.

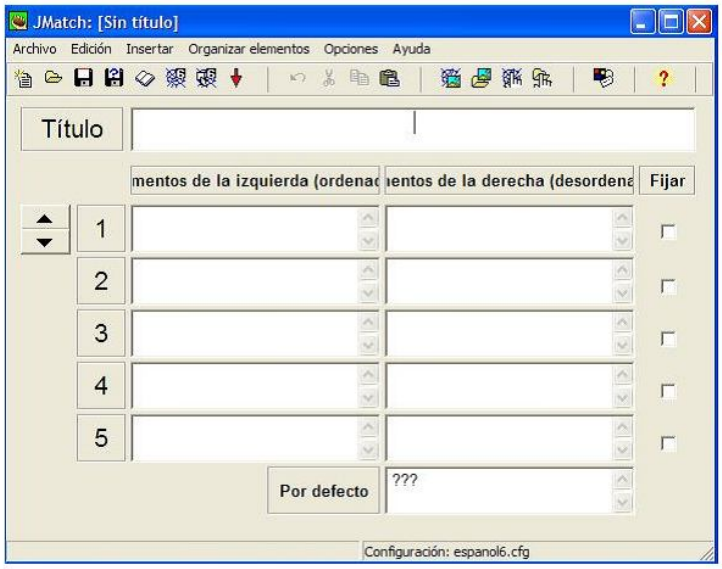
Antes de empezar a crear el ejercicio necesitamos recopilar el **material** necesario. En este caso:

- Página web. Localiza en Internet una página con la biografía de la autora.
- Sonetos de amor y discreción, Biblioteca Virtual Miguel de Cervantes: http://www.cervantesvirtual.com/multimedia/Sor_Juana_Enrique/UNAM02.asx
- Imagen de Sor Juana. Utiliza la opción de búsqueda de imágenes de Google o de otro buscador para localizar una. Guárdala en la carpeta de trabajo que has creado antes.

Abre la patata correspondiente a **JMatch** y da un **nombre** al archivo en el que vas a trabajar: **Archivo > Guardar como**. No utilices tildes ni otros símbolos diacríticos; tampoco dejes espacios en blanco en el nombre. Elige la carpeta que has creado antes para guardar el archivo.



La pantalla de trabajo:



Título: en el espacio en blanco que aparece al lado, escribe un título. P. ej. *Sor Juana.*

Ejercicio: en la columna de la izquierda vamos a escribir los acontecimientos de su vida; y en la columna de la derecha, las fechas en que ocurrieron. Por ejemplo:

Columna de la izquierda:	*Columna de la derecha:*
1. *Nacimiento*	*1648*
2. *Llegada a la ciudad de México*	*1660*
3. *Toma los votos religiosos*	*1669*
4. *Amistad con el marqués de la Laguna y su esposa*	*1674*
5. *Representación de su obra de teatro "Los empeños de una casa"*	*1683*
6. *25 años de vida religiosa*	*1694*
7. *Muerte*	*1695*

Resultado: básicamente, el ejercicio ya está creado. Si quieres ver el resultado, pulsa sobre el menú **Archivo > Crear página web > Página web para navegadores V6.** Dale un nombre y guárdala. Igual que antes, evita los diacríticos y los espacios en blanco. Pulsa sobre la opción **"Ver el ejercicio en mi navegador.**

- ✓ Esta operación puedes efectuarla tantas veces como quieras. Mantén siempre el mismo nombre para la página web y limítate a reescribirlo para ver el aspecto que va teniendo.
- ✓ Ten en cuenta que una cosa es el archivo en el que trabajamos y diseñamos el ejercicio y otra la página web donde se visualizan los resultados, que es la que se cuelga en Internet y donde posteriormente nuestros alumnos realizan las actividades.

Instrucciones y formato. Para personalizar el ejercicio, pulsa sobre el menú **Opciones > Configurar el formato del archivo originado.**

Imágenes. Tanto en el título como en las columnas podemos insertar imágenes. Vamos a insertar la imagen que tenemos de Sor Juana justo debajo del título. Para ello, haz clic con el ratón en el recuadro en blanco que aparece bajo "Subtítulo del ejercicio". Después, pulsa sobre **Insertar > Imagen > Imagen desde un archivo local.** Selecciona tu carpeta de trabajo y la imagen de Sor Juana que guardaste antes en ella. Alinea la imagen en el centro.

Hipervínculos. Igual que con las imágenes, podemos insertar enlaces a otras páginas o documentos en cualquier parte del ejercicio. Vamos a insertar dos hipervínculos en las instrucciones.

Abre tu navegador y vete a la página que contiene la biografía que has seleccionado para realizar el ejercicio. Copia la dirección.

Escribe el siguiente texto en el recuadro en blanco que aparece bajo "Instrucciones":

> *Lee la biografía de Sor Juana en el siguiente enlace. Mientras trabajas, escucha sus poemas: pulsa aquí.*

Ahora, selecciona "siguiente enlace" y pulsa sobre **Insertar** > **Enlace** > **Vínculo a una dirección de Internet**. Pega la dirección que has copiado donde dice **"Ruta URL"**.

Haz lo mismo con el texto "pulsa aquí". En este caso, inserta un enlace a: http://www.cervantesvirtual.com/multimedia/Sor_Juana_Enrique/UNAM02.asx

Otros aspectos. Puedes elegir el tipo de letra, el color de las diferentes partes de la página web, insertar una imagen de fondo, escribir diferentes mensajes para que aparezcan a lo largo de la realización de los ejercicios, etc. Todo ello se define en el menú **Opciones > Configurar el formato del archivo originado** y **Opciones > Fuentes.**

- Elige las que más te gusten.
- Pon un **contador** de 15 minutos para realizar el ejercicio.

ENTREGA DE LA PRÁCTICA

Cuando hayas terminado, guarda los resultados y crea la página web definitiva. Cierra todas las aplicaciones, comprime la carpeta de trabajo y envíamela, junto con tu nombre, a la dirección de correo: **milka.villayandre@unileon.es**

3.2. Creación de un ejercicio con JCloze

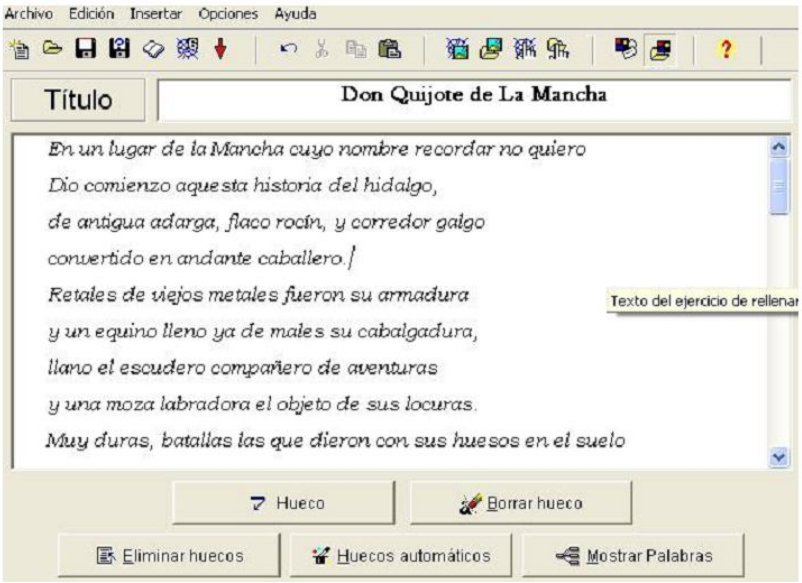
Este programa nos permite crear actividades basadas en rellenar huecos.

Vamos a crear un ejercicio en el que nuestros alumnos van a ver y escuchar un vídeo de un grupo de rap interpretando un tema inspirado en El Quijote. Después, tendrán que completar los huecos en el texto de la canción.

Antes de empezar a crear el ejercicio necesitamos recopilar el **material** que vamos a utilizar. En este caso:

- Vídeo de You Tube: "Quijote hip hop". Vete a la siguiente dirección y añádela a tus favoritos: http://www.youtube.com/watch?v=jPRa_JyEe7M&mode=related&search=]
- Texto de la canción. Lo encontrarás en: http://www3.unileon.es/dp/dfh/Milka/LC/quijote_rap.doc
- Busca un par de imágenes relacionadas con don Quijote para decorar la actividad y guárdalas en tu carpeta de trabajo.

Abre la patata correspondiente a **JCloze** y da un **nombre** al archivo en el que vas a trabajar: **Archivo > Guardar como**. Crea una carpeta con tu nombre_JCloze y guarda el archivo. P. ej. *Milka_JCloze.*



La pantalla de trabajo

Título: en el espacio en blanco que aparece al lado, escribe un título. P. ej. "*Don Quijote de La Mancha*".

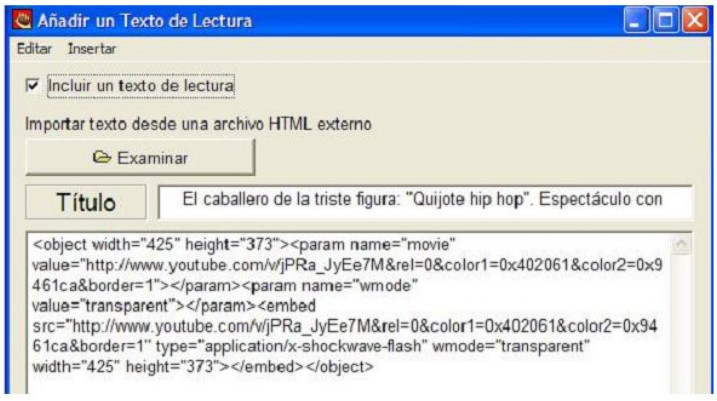
Ejercicio: en el espacio en blanco que aparece debajo del título es donde escribiremos el **texto**, frases, etc. en las que queremos crear huecos. En este caso, necesitamos el texto de la canción. Vete a la dirección indicada más abajo, selecciona el primer texto, cópialo y pégalo: **http://www3.unileon.es/dp/dfh/Milka/LC/quijote_rap.doc**

Huecos. Para crear huecos en el texto, únicamente tenemos que seleccionar la palabra que queremos eliminar y pulsar sobre el botón "**Hueco**". Si nos equivocamos, pulsamos sobre "**Borrar hueco**" (elimina el hueco seleccionado) o "**Eliminar huecos**" (elimina todos los huecos creados). Alternativamente, el programa puede generar huecos de forma aleatoria ("**Huecos automáticos**"). Con la opción "**Mostrar palabras**" podemos editar una palabra para la que hayamos creado un hueco (p. ej. para introducir una pista, una respuesta alternativa, etc.).

Crea huecos para las palabras que aparecen subrayadas y en negrita en la segunda versión del texto que aparece en el documento con el texto de la canción: *historia, caballero, armadura...*

Crea la página web para ver los resultados. Recuerda: **Archivo > Crear página web / tecla F6 / icono**

Añadir lectura. En todos los ejercicios con Hot Potatoes es posible insertar un texto o similar, que aparecerá en la parte izquierda de la pantalla. En este caso, vamos a ocupar el espacio reservado para el texto con un **vídeo.** Pulsamos sobre el menú **Archivo > Añadir lectura.** Aparecerá la siguiente pantalla:



Marcamos la casilla: **Incluir un texto de lectura.**

Ponemos un **título** al texto:

> *"El caballero de la triste figura: Quijote hip hop".*

Seleccionamos el **texto**: el texto puede ser un documento que nosotros previamente hayamos redactado o buscado y guardado en nuestro ordenador (Opción **Examinar**) o un fragmento que hemos copiado y queremos pegar en el espacio en blanco. En este caso, utilizaremos esta segunda opción: vamos a la página previamente guardada en **Favoritos**; seleccionamos y copiamos el código que aparece bajo el rótulo **"Embed"** (podemos personalizar el color de los controles del vídeo); por último, pegamos este código en el espacio en blanco destinado para el texto de la lectura.

> http://www.youtube.com/watch?v=jPRa_JyEe7M&mode=related&search=]

Crea otra vez la página para ver el resultado: **Archivo > Crear página web / tecla F6 / icono**

Opciones > Configurar el formato del archivo originado.

Ya solo te queda personalizar el formato: avisos e indicaciones; botones; aspecto; contador (pon un **contador de 15 minutos**); colores e imagen de fondo; otros aspectos; y tipo de letra (Opciones > Fuentes).

En el espacio reservado para el **"Subtítulo del ejercicio"** vamos a escribir las instrucciones:

> *Escucha la canción y completa, después, los espacios en blanco con las palabras que aparecen más abajo. Pulsa "Play" para empezar.*

En el espacio reservado para las **instrucciones**, escribe:

> *Las siguientes palabras aparecen varias veces: caballero (3), derrotas (2), don (2), Quijote (2)*

Inserta alguna **imagen** de las que has buscado en un lugar apropiado: en el título, en las pistas, en las instrucciones, en los botones, etc.

ENTREGA DE LA PRÁCTICA

Cuando hayas terminado, guarda los resultados y crea la página web definitiva. Cierra todas las aplicaciones, comprime la carpeta de trabajo y envíamela, junto con tu nombre, a la dirección de correo: milka.villayandre@unileon.es

3.3. Creación de un ejercicio con JCross

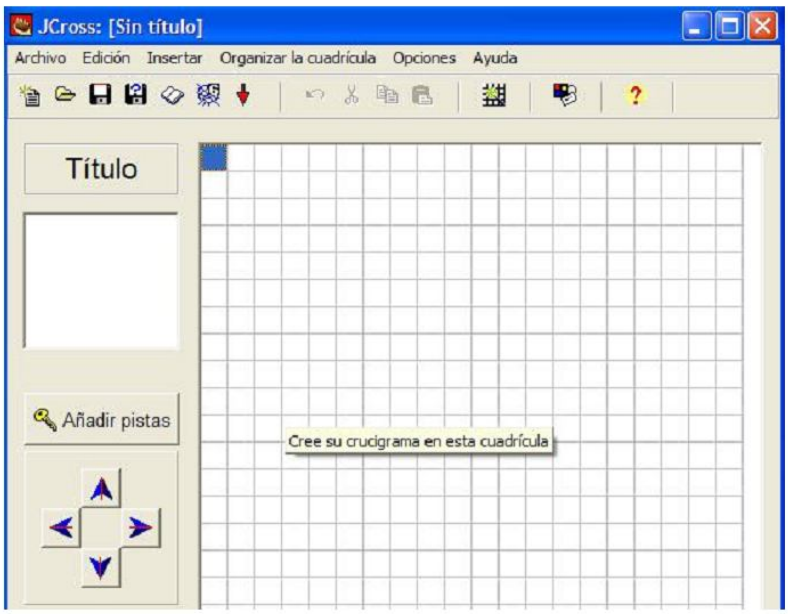
Este programa nos permite crear crucigramas.

Vamos a crear un ejercicio en el que nuestros alumnos van a leer y escuchar un poema de Lorca. Después, tendrán que realizar un crucigrama.

Antes de empezar a crear el ejercicio necesitamos recopilar el **material** que vamos a emplear. En este caso:

- Texto del poema. Vete a la siguiente dirección y añádela a tus favoritos: http://antologiapoeticamultimedia.blogspot.com/2006/08/romance-de-la-luna-luna.html
- Vídeo de You Tube con una interpretación de Camarón. Vete a la siguiente dirección y añádela a tus favoritos: http://www.truveo.com/romance-de-la-luna-camaron/id/3261816094
- Diccionario de la lengua española, Real Academia Española. Añade a tus favoritos esta dirección: http://www.rae.es
- Imágenes: busca un par de imágenes ilustrativas para decorar la actividad y guárdalas en tu carpeta de trabajo. También puedes necesitar imágenes para las pistas (*ver más adelante*)

Abre la patata correspondiente a **JCross** y da un **nombre** al archivo en el que vas a trabajar: **Archivo > Guardar como**. Crea una carpeta con tu nombre_JCross y guarda el archivo. P. ej. *Milka_JCross.*



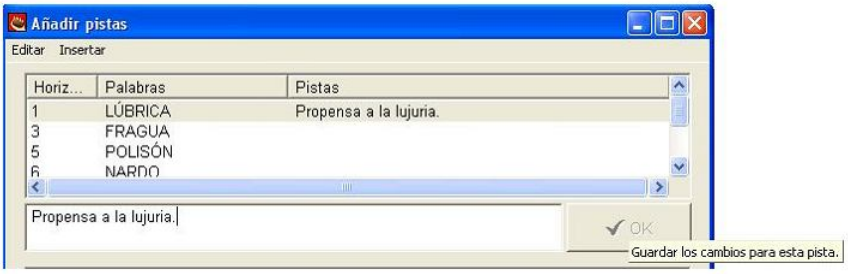
La pantalla de trabajo

Título: en el espacio en blanco que aparece debajo, escribe un título. P. ej. *Romance de la luna, luna*.

Ejercicio: coloca en la cuadrícula las palabras que los alumnos van a tener que localizar después (*ver más abajo*). Pulsa con el ratón en la cuadrícula y escribe una letra por casilla. No necesitas disponer las palabras tal y como van a quedar. Basta con que las escribas una en cada fila, dejando una fila en blanco entre ellas. Después, pulsa sobre **"Organizar la cuadrícula" > Crear automáticamente.**

fragua, polisón, nardo, lúbrica, estaño, yunque, blancor, olivar, zumaya, velar

Añadir pistas. Pulsa sobre este botón para introducir las definiciones de las palabras.



Selecciona la palabra en cuestión y, en el espacio en blanco que aparece debajo, escribe una definición (o copia y pega una desde el diccionario de la Real Academia u otro diccionario de tu elección). P. ej. seleccionamos "*lúbrica*". Debajo escribimos "*Propensa a la lujuria*". Por último, pulsa sobre "OK" para guardar la pista.

También puedes ilustrar las definiciones o las pistas con imágenes o con vínculos. Recuerda: **Insertar > Imagen > Desde un archivo local / Insertar > Enlace.**

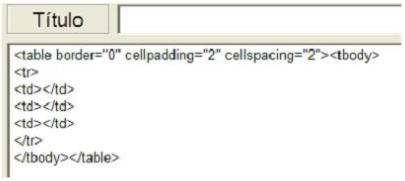
Inserta definiciones o pistas para el resto de palabras.

Crea la página web para ver los resultados hasta el momento. Recuerda: **Archivo > Crear página web / tecla F6 / icono**

Añadir lectura. Pulsamos sobre el menú **Archivo > Añadir lectura.**

- Marcamos la casilla: **Incluir un texto de lectura.**
- Vamos a la página que figura debajo, que previamente hemos guardado en nuestros Favoritos, seleccionamos el texto del poema y copiamos: **http://antologiapoeticamultimedia.blogspot.com/2006/08/romance-de-la-luna-luna.html**
- **NOTA.** Como el programa no nos ofrece la posibilidad de dar formato al texto, este va a aparecer a la izquierda de la pantalla y quedará mucho espacio vacío. Podemos **insertar una tabla** (pincha con el ratón en el espacio en blanco para el texto), de una fila y tres columnas, sin bordes (borde=0), para organizar mejor la información: **menú Insertar > Tabla HTML**

- Al insertar la tabla, aparece el siguiente **código**:



<table…> indica el inicio de una tabla
<tr> indica el principio de una fila
</tr> indica el final de una fila
<td> indica el principio de una celda
</td> indica el final de una celda
</table> indica el final de una tabla

- Colocamos el ratón entre la segunda etiqueta <td> y </td> y pegamos el texto del poema.
- Tras la primera etiqueta <td> insertamos una de las imágenes elegidas para ilustrar la actividad. Recuerda: **Insertar > Imagen > Desde archivo local.**
- Tras la tercera etiqueta <td> insertamos otra imagen.
- Por último, damos a **OK.**

Guarda los cambios y crea otra vez la página para ver el resultado: **Archivo > Crear página web / tecla F6 / icono**

Opciones > Configurar el formato del archivo originado.

Ya solo nos falta personalizar el formato: avisos e indicaciones; botones, aspecto; contador; colores e imagen de fondo; otros aspectos; y tipo de letra (Opciones > Fuentes). Configúralo a tu gusto.

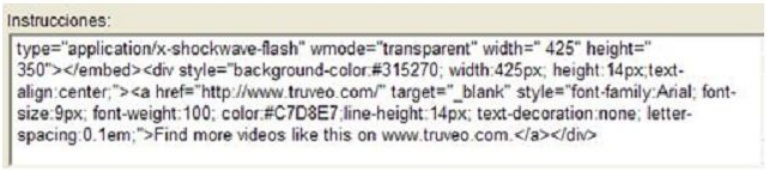
En el espacio reservado para el "**Subtítulo**" escribe:

Lee y escucha el poema de Federico García Lorca. Después, completa el crucigrama.

En el espacio reservado para las "**Instrucciones**" vamos a insertar el vídeo. Vete a la página de tus favoritos: **http://www.truveo.com/romance-de-la-luna-camaron/id/3261816094**

Justo debajo del vídeo verás lo siguiente: selecciona y copia lo que aparece a continuación de "**Embed**". Pégalo en el espacio reservado para las instrucciones.

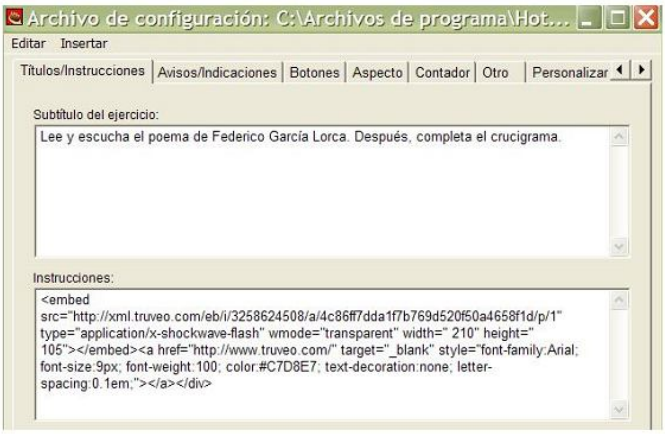
Observa el código que aparece:



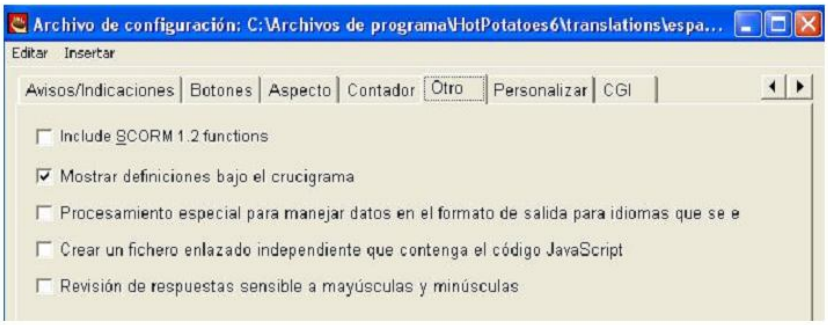
Vamos a modificar algunos parámetros para que el reproductor de vídeo sea más pequeño y no aparezca la publicidad:

- Cambia los valores del ancho y del alto del reproductor:
  - *width: 425 > 210*
  - *height: 350 > 105*
- Elimina el código desde <*div style…* hasta *align center;*">
- Elimina el código *line-height 14px*
- Elimina "*Find more videos like this on www.truveo.com*"

El resultado final:



En el submenú **Otro**, activa la opción **"Mostrar definiciones bajo el crucigrama"**.



Una vez que hayas terminado, crea de nuevo la página para ver cómo ha quedado.

ENTREGA DE LA PRÁCTICA

Cuando hayas terminado, guarda los resultados y crea la página web definitiva. Cierra todas las aplicaciones, comprime la carpeta de trabajo y envíamela, junto con tu nombre, a la dirección de correo: **milka.villayandre@unileon.es**

3.4. Creación de un ejercicio con JQuiz

Este programa nos permite crear actividades basadas en preguntas y respuestas.

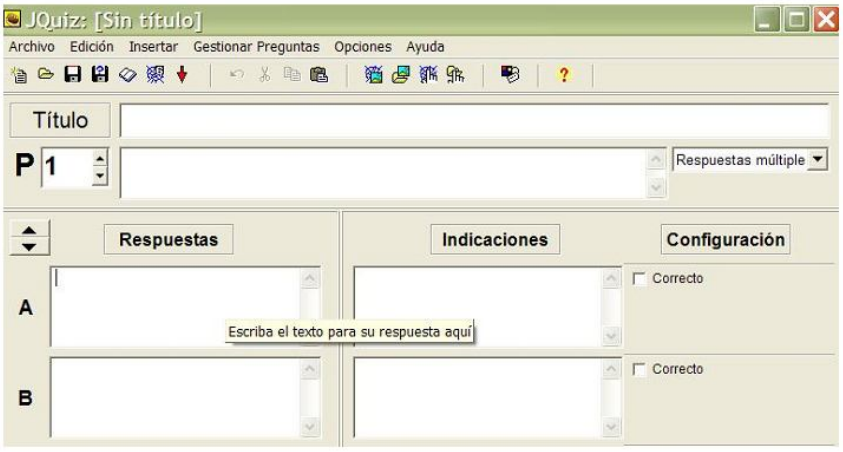
Vamos a crear un ejercicio en el que nuestros alumnos van a leer un texto sobre el español en América. Después, tendrán que responder a una serie de preguntas sobre el mismo.

Antes de empezar a crear el ejercicio necesitamos recopilar el **material** que vamos a utilizar. En este caso:

- Texto: "Historia del español en América". Vete a la siguiente dirección y añádela a tus favoritos: http://szamora.freeservers.com/america.htm
- Enlace: "Idioma español" (Wikipedia). Vete a la siguiente dirección y añádela a tus favoritos: http://es.wikipedia.org/wiki/Castellano
- Enlace: "Familia algonquina". Vete a la siguiente dirección y añádela a tus favoritos: http://www.proel.org/index.php?pagina=mundo/amerindia/algonquin
- Enlace: "Aprende náhuatl". Vete a la siguiente dirección y añádela a tus favoritos: http://mexica.ohui.net/glosarios/2/
- Sonidos: "Banco de imágenes y sonidos" (MEC). Vete a la siguiente dirección y añádela a tus favoritos: http://bancoimagenes.isftic.mepsyd.es/

Abre la patata correspondiente a **JQuiz** y da un **nombre** al archivo en el que vas a trabajar: **Archivo > Guardar como**. Crea una carpeta con tu nombre_JQuiz y guarda el archivo. P. ej. *Milka_JQuiz.*

La pantalla de trabajo



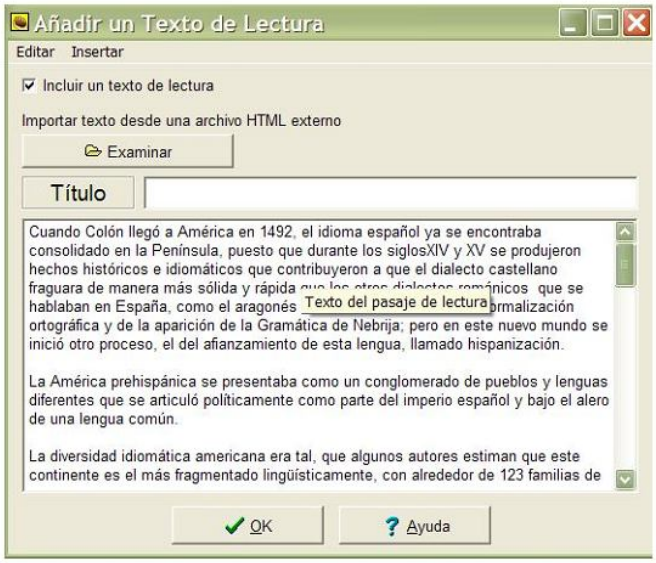
Título: en el espacio en blanco que aparece al lado, escribe un título. P. ej. "*El español en América*".

Ejercicio:

Texto. En una actividad de este tipo lo primero que necesitamos es el **texto** (aunque, como ya sabes, también podemos trabajar con un documento de sonido o de vídeo, que podemos insertar donde más nos convenga).

Vete a la página web indicada en primer lugar y copia el texto que allí aparece: **http://szamora.freeservers.com/america.htm**

Añadir lectura. En todos los ejercicios con Hot Potatoes es posible insertar un texto o similar, que aparecerá en la parte izquierda de la pantalla. Pulsa sobre el menú **Archivo > Añadir lectura.** Aparecerá la siguiente pantalla:



- Marca la casilla: **Incluir un texto de lectura.**
- Si lo deseas, pon un **título** al texto: "Historia del español en América".
- Pincha con el ratón en el espacio en blanco debajo del título y pega el texto previamente seleccionado: **Editar > Pegar.**
- Al final del texto pon la referencia: "*Texto tomado de*: **http://szamora.freeservers.com/america.htm**"
- Por último, añade el siguiente texto debajo de la referencia: "*Para saber más sobre el idioma español, pulsa aquí*".
- Selecciona "pulsa aquí" e inserta un enlace a: **http://es.wikipedia.org/wiki/Castellano**. **Insertar > Vínculo > Vínculo a una dirección de Internet.** Al lado de **Ruta URL** introduce la dirección.
- Pulsa **OK** para salir de la pantalla **Añadir lectura.**

Preguntas.

- Las **preguntas** se insertan en el espacio en blanco al lado de P (1, 2, 3, etc. indican el número de la pregunta).
- Debajo de cada pregunta escribimos las posibilidades de **respuestas** (A, B, C…).
- Por último, en el espacio de **indicaciones** podemos escribir un comentario para el alumno, insertar imágenes, sonidos, enlaces, etc.

Tipo de preguntas. Opciones: (se pueden combinar en un mismo ejercicio)

1) Preguntas de respuestas múltiples. Formulamos una pregunta y damos varias opciones de respuesta. El alumno debe seleccionar la correcta.
2) Preguntas de respuestas cortas. Formulamos una pregunta y el alumno debe responder de forma breve. No hay opciones.
3) Preguntas de respuestas híbridas. Primero se formulan como una pregunta de respuesta corta. Tras varios intentos fallidos, se transforman en una pregunta de respuestas múltiples.
4) Preguntas de respuestas multiselección. En este caso, ante una pregunta, se le ofrecen al alumno varias opciones. Debe marcar todas las correctas y no marcar las incorrectas.

Preguntas para el texto: sigue estas instrucciones para hacer preguntas sobre el texto que hemos insertado.

P1 (Respuestas múltiples). Selecciona "Respuestas múltiples" en el menú desplegable de los tipos de respuestas.

En el espacio en blanco al lado de P1 escribe la **pregunta**:

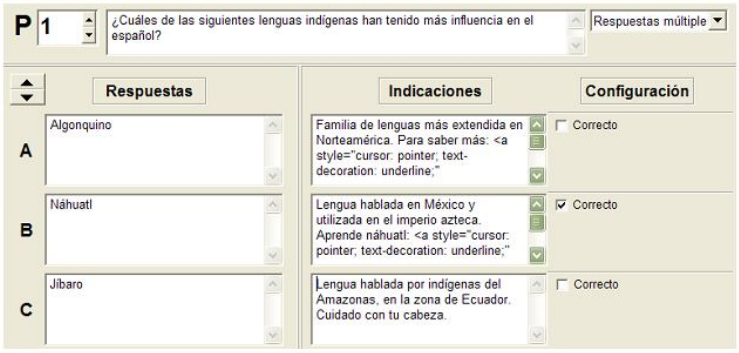
"¿Cuál de las siguientes lenguas indígenas ha tenido más influencia en el español?"

En los espacios en blanco debajo reservados para las respuestas (A, B, C…) escribe, en cada hueco, las siguientes **respuestas**:

- *Algonquino*
- *Náhuatl (marca como respuesta correcta: "**Correcto**")*
- *Jíbaro*

Indicaciones: en el espacio en blanco reservado para las indicaciones al lado de cada una de las anteriores respuestas, escribe:

1. *Familia de lenguas más extendida en Norteamérica. Para saber más: familia algonquina. Selecciona "familia algonquina" e* ***inserta un enlace a:*** **http://www.proel.org/index.php?pagina=mundo/amerindia/algonquin (Insertar > Vínculo)**
2. *Lengua hablada en México y utilizada en el imperio azteca. Aprende náhuatl: aquí. Selecciona "aquí" e inserta el siguiente* ***enlace:*** **http://mexica.ohui.net/glosarios/2/ (Insertar > Vínculo)**
3. *Lengua hablada por indígenas del Amazonas, en la zona de Ecuador. Cuidado con tu cabeza…*

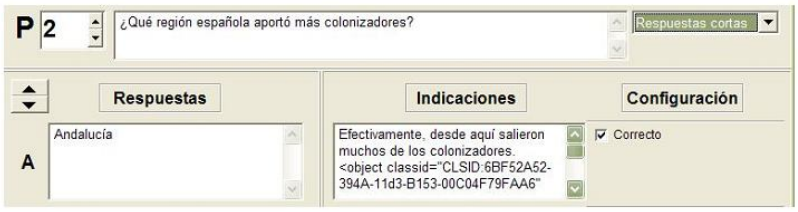


➡ **P2** (Respuestas cortas). Selecciona "Respuestas cortas" en el menú desplegable de los tipos de respuestas.

Pregunta: *¿Qué región española aportó más colonizadores?*

Respuesta: *Andalucía (marca* ***Correcto****)*

Indicaciones: *Efectivamente, desde aquí salieron muchos de los colonizadores.*

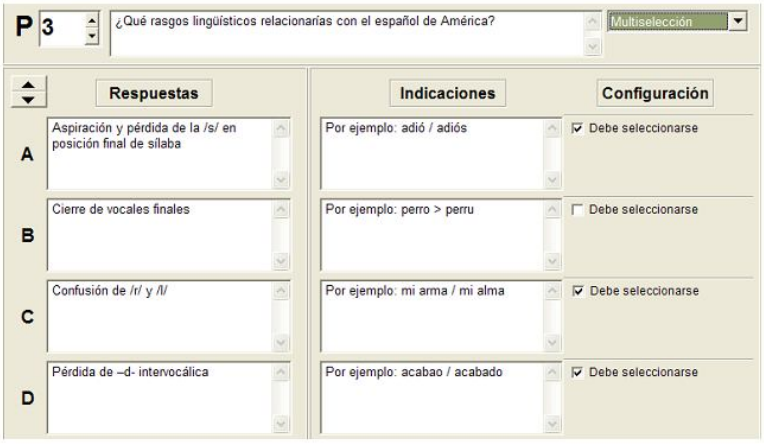


➡ **P3** (Multiselección). Selecciona "Multiselección" en el menú desplegable de los tipos de respuestas.

Pregunta: *¿Qué rasgos lingüísticos relacionarías con el español de América?*

Respuestas:

- *Aspiración y pérdida de la /s/ en posición final de sílaba (marca Debe seleccionarse)*
- *Cierre de vocales finales*
- *Confusión de /r/ y /l/ (marca Debe seleccionarse)*
- *Pérdida de -d- intervocálica (marca Debe seleccionarse)*



P4 (Híbrida). Selecciona "Híbrida" en el menú desplegable de los tipos de respuestas.

Pregunta: *¿Qué ciudad fue el punto de partida de los viajes a América?*

Respuestas:

- *Madrid (marca **Incluir en opciones M/C**)*
- *Sevilla (marca **Correcto** e **Incluir en opciones M/C**)*
- *Sta. Cruz de Tenerife (marca **Incluir en opciones M/C**)*

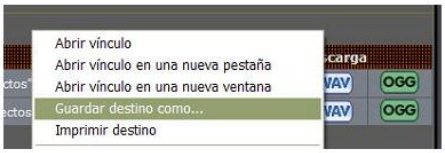
Básicamente ya tenemos el ejercicio. Crea la página web para ver el resultado hasta ahora. Recuerda: **Archivo > Crear página web.**

Multimedia. Ahora vamos a insertar unos sonidos para que cada vez que el alumno seleccione o escriba una respuesta correcta se reproduzca un sonido; y cuando seleccione o escriba una respuesta incorrecta, también se reproduzca un sonido diferente.

Los sonidos para los mensajes los vamos a obtener del **Banco de sonidos del MEC**, cuya dirección has almacenado antes en tus favoritos: **http://bancoimagenes.isftic.mepsyd.es/**

Vete a la dirección indicada y busca un sonido adecuado para las respuestas correctas y otro para las incorrectas. Puedes buscarlos escribiendo en la casilla en blanco al lado de "Buscar" un término. P. ej.: "*error*", "*bien*"...

Para escuchar los sonidos, pulsa sobre el enlace. Cuando encuentres uno que te guste, pulsa sobre el icono de "Descargar archivo" adecuado (formato .wav, .mp3 u .ogg), según el tipo de reproductor que tenga tu ordenador. Para los sonidos con formato **.wav** suele servir el reproductor de sonidos que incorpora Windows. Da un nombre al archivo y guárdalo en tu carpeta de trabajo.



Ya tenemos los sonidos. Ahora vamos a insertarlos en el lugar adecuado. P. ej. vete a la pregunta 1. En las indicaciones, sitúa el ratón después del código que indica el final del enlace: y pulsa sobre **Insertar > Objeto multimedia.**

Aparecerá el siguiente menú:

- **Archivo multimedia:** pulsa sobre buscar y selecciona el archivo de sonido que has guardado para las respuestas incorrectas. Pulsa sobre **Abrir**. El nombre del archivo aparecerá en pantalla.
- **Anchura y altura:** se refiere al tamaño del reproductor. No queremos que se vea, así que escribimos **"0"** en ambas casillas.
- **Reproductores:** selecciona **"Añadir Windows Media Player"**, por ejemplo.
- **Incluir simplemente un enlace:** desactiva esta opción.
- **OK:** pulsa "OK" cuando hayas terminado.

El resultado que aparecerá en el ejercicio será un código:

- *<object classid="CLSID:6BF52A52-394A-11d3-B153-00C04F79FAA6" width="100" height="30"> <param name="url" value="" /> <param name="**autostart**" **value="false**" /> <param name="showcontrols" value="true" /></object>*
- En el parámetro **"autostart"** cambia **"value = false"** por **"value = true"**.

Realiza esta misma operación con la otra respuesta incorrecta de esta pregunta (o, simplemente copia y pega el código). Después, inserta el sonido para la respuesta correcta.

Por último, solo te falta configurar algunas opciones en los menús **"Gestionar preguntas"**, **"Opciones > Configurar el formato del archivo originado"** y **Modalidad** (en la modalidad **Avanzada** puedes ponderar el peso otorgado a cada pregunta en la nota final).

ENTREGA DE LA PRÁCTICA

Cuando hayas terminado, guarda los resultados y crea la página web definitiva. Cierra todas las aplicaciones, comprime la carpeta de trabajo y envíamela, junto con tu nombre, a la dirección de correo: **milka.villayandre@unileon.es**

3.5. Creación de un ejercicio con JMix

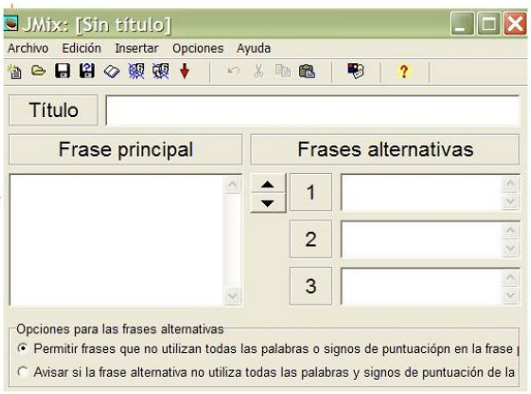
Este programa nos permite crear actividades basadas en ordenar información previamente desordenada: puede ser una frase, un pequeño texto, etc.

Vamos a crear un ejercicio en el que nuestros alumnos van a escuchar un texto y luego tendrán que ordenarlo.

Antes de empezar a crear el ejercicio necesitamos recopilar el **material** que vamos a utilizar. En este caso:

Texto: rima VII de Bécquer
Del salón en el ángulo oscuro,
de su dueña tal vez olvidada,
silenciosa y cubierta de polvo,
veíase el arpa.
¡Cuánta nota dormía en sus cuerdas,
como el pájaro duerme en las ramas,
esperando la mano de nieve
que sabe arrancarlas!
¡Ay!, pensé; ¡cuántas veces el genio
así duerme en el fondo del alma,
y una voz como Lázaro espera
que le diga «Levántate y anda»!

Audio:
http://www.cervantesvirtual.com/multimedia/archivo/GustavoAdolfoBecquer/Rimas_007.asx
Imagen: http://es.wikipedia.org/wiki/Becquer



Abre la patata correspondiente a **JMix** y da un **nombre** al archivo en el que vas a trabajar: **Archivo > Guardar como**. Crea una carpeta con tu nombre_JMix y guarda el archivo. P. ej. *Milka_JMix.*

La pantalla de trabajo

Título: inserta la imagen de Bécquer. Vete a la página web de Wikipedia sobre Bécquer, selecciona y guarda la imagen que allí aparece: **http://es.wikipedia.org/wiki/Becquer**

Después, pincha con el ratón en el espacio en blanco reservado para el título y vete al menú **Insertar > Imagen**.

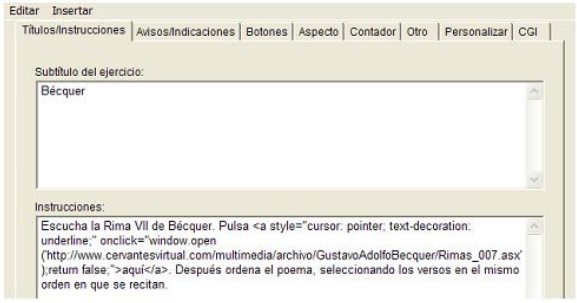
Frase principal. Aquí vamos a escribir el poema de Bécquer de forma ordenada: una frase por línea, tal y como aparece más arriba. Si fuera una frase, escribiríamos una palabra por línea, etc.

Vete al menú **Opciones > Configurar el formato del archivo originado.** En la pestaña **"Títulos/Instrucciones"** escribe:

Subtítulo del ejercicio: *"Bécquer".*

Instrucciones: *"Escucha la Rima VII de Bécquer. Pulsa aquí. Después ordena el poema, seleccionando los versos en el mismo orden en que se recitan".*

Selecciona **"aquí"** e inserta un enlace **(Insertar > Vínculo)** a: **http://www.cervantesvirtual.com/multimedia/archivo/GustavoAdolfoBecquer/Rimas_007.asx**



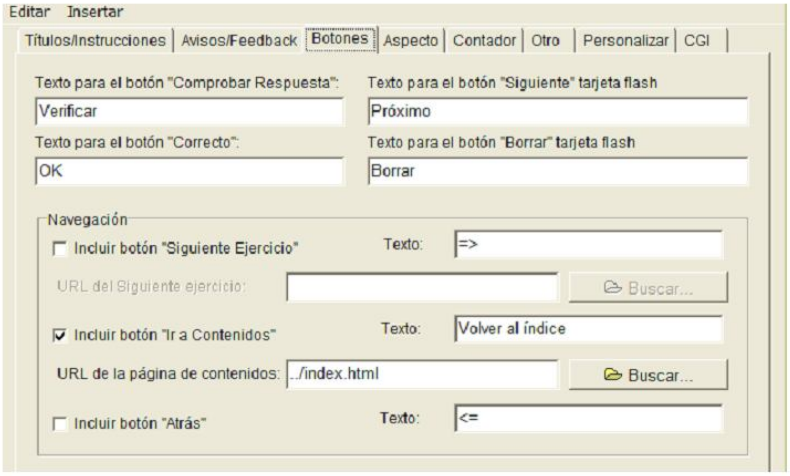
ENTREGA DE LA PRÁCTICA

Cuando hayas terminado, guarda los resultados y crea la página web definitiva. Cierra todas las aplicaciones, comprime la carpeta de trabajo y envíamela, junto con tu nombre, a la dirección de correo: **milka.villayandre@unileon.es**

3.6. Enlazar actividades en una misma unidad:

Es posible encadenar actividades:

1) De forma manual: desde el menú **Opciones** > **Configurar el formato del archivo originado** de cualquiera de los programas de Hot Potatoes. Pestaña: **"Botones"**.



Activaríamos la casilla **"Incluir botón 'siguiente ejercicio'"** y en el espacio en blanco al lado de **URL del Siguiente ejercicio**, buscaríamos la página web correspondiente al ejercicio que queramos que los alumnos realicen a continuación del actual.

También podemos incluir el **botón "Atrás"** marcando esta casilla: nos permite ir al último ejercicio que hayamos realizado antes de aquel en el que nos encontremos.

2) De forma automática: con el programa **"The Masher"**. La versión sin licencia del programa únicamente nos permite encadenar 3 ejercicios. Si queremos encadenar más, necesitamos adquirir una licencia.

Pulsa a **Añadir archivos** para ir seleccionando los ejercicios que quieres enlazar. Puedes seleccionarlos ya por orden u ordenarlos después con la flecha que aparece a la derecha.

Una vez que tengas los tres ejercicios, pulsa sobre **Construir unidad**. El programa genera automáticamente un **índice** con las actividades seleccionadas y permite realizar los tres ejercicios de forma consecutiva con los botones de navegación que genera en la parte superior de las actividades.

También puedes utilizar esta aplicación para generar un **índice** de una serie de actividades, aunque no las enlaces.

O para subir directamente la unidad a **Hotpotatoes.net**.

Cuando estés conforme con el resultado, guarda el proyecto: **Archivo > Guardar como.**

ENTREGA DE LA PRÁCTICA

Cuando hayas terminado, guarda los resultados y crea la página web definitiva. Cierra todas las aplicaciones, comprime la carpeta de trabajo y envíamela, junto con tu nombre, a la dirección de correo: **milka.villayandre@unileon.es**

A.2. Práctica 6

Práctica 6. Problemas para el tratamiento del texto *(práctica presencial)*.

Lee el siguiente artículo: señala y comenta los casos que plantearían problemas para el tratamiento computacional de este texto. Presta especial atención a:

- Contracciones, átonos pronominales enclíticos, formas compuestas y perífrasis verbales
- Conjunciones y locuciones conjuntivas
- Nombres propios y términos extranjeros
- Cifras, abreviaturas y siglas
- Locuciones preposicionales, adverbiales y verbales
- Expresiones multipalabra

Calorías a la vista contra la obesidad

Los restaurantes de comida rápida de Nueva York, obligados a etiquetar sus productos

AINHOA [illegible] - Madrid - 03/05/2006 22:12

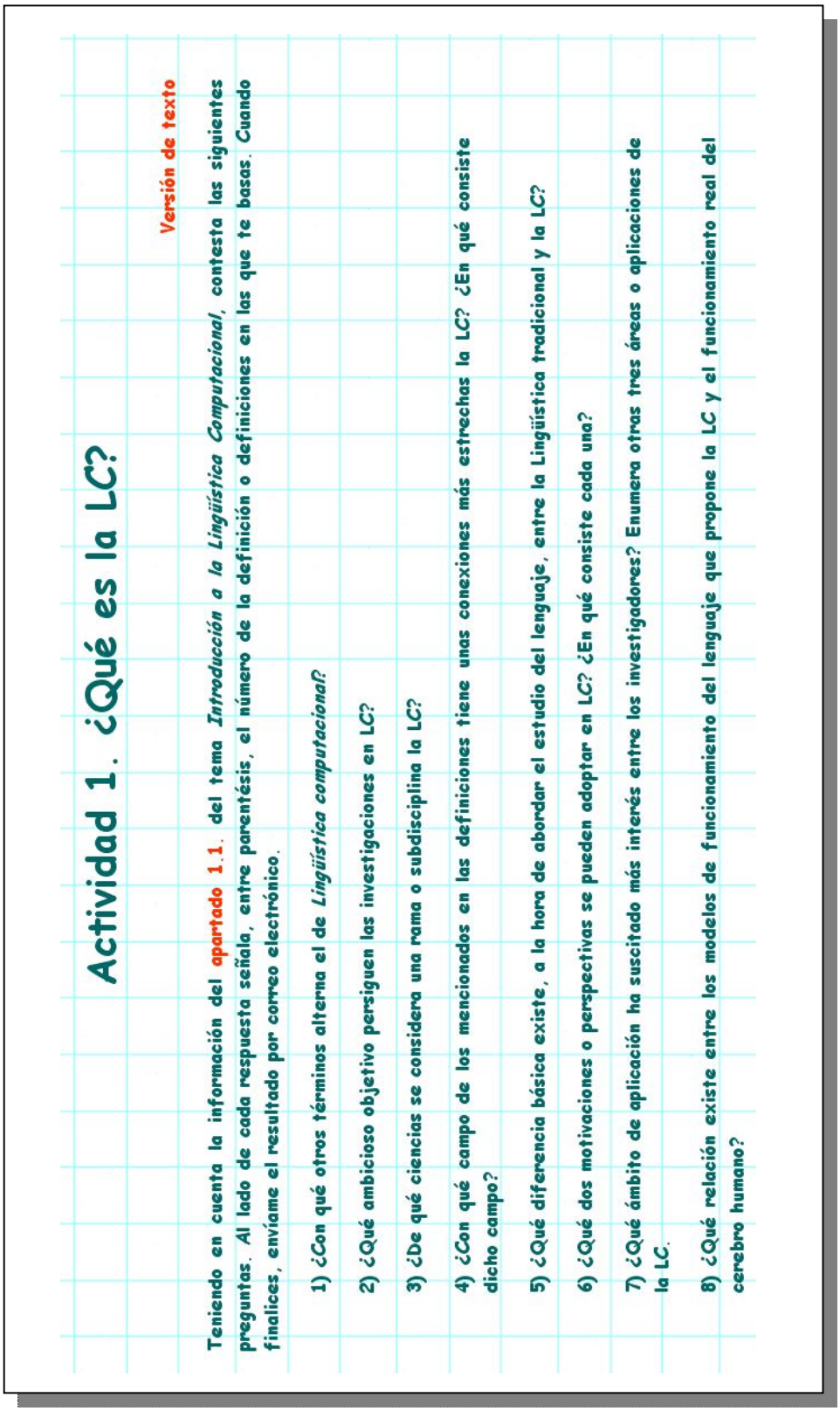
El combate de las autoridades sanitarias contra la obesidad no ha hecho más que empezar y EEUU, en ésta como en otras muchas guerras, quiere llevar la voz cantante. La última iniciativa se ha puesto en marcha en Nueva York esta semana y obliga a los restaurantes con más de 15 establecimientos a lo largo del país (la mayoría de cadenas de la denominada comida rápida o *fast food*) a especificar las calorías de cada uno de los platos presentes en su carta.

La medida no ha hecho ninguna gracia a los restauradores que, hasta última hora, intentaron paralizarla judicialmente al considerar que violaba la primera enmienda de la Constitución, que protege la libertad de expresión frente a la intromisión del Gobierno. Sin embargo, la ley ha salido adelante y, desde ahora, al ojear el menú en restaurantes tan populares como Starbucks Coffe o T.G. Friday's los clientes verán –al mismo tamaño y con similar tipografía– las calorías que se esconden tras sus platos.

Producto de Starbucks Coffe con las calorias indicadas - AP

Eso sí, la Asociación de Restaurantes de Nueva York ha conseguido que se establezca una moratoria y que hasta julio no se pueda poner multas. Además, tienen la posibilidad de volver a recurrir la norma en unos meses.

B.1. Ejemplo de actividad de respuestas breves



Actividad 1. ¿Qué es la LC?

Versión de texto

Teniendo en cuenta la información del apartado 1.1. del tema *Introducción a la Lingüística Computacional*, contesta las siguientes preguntas. Al lado de cada respuesta señala, entre paréntesis, el número de la definición o definiciones en las que te basas. Cuando finalices, envíame el resultado por correo electrónico.

1) ¿Con qué otros términos alterna el de *Lingüística computacional*?

2) ¿Qué ambicioso objetivo persiguen las investigaciones en LC?

3) ¿De qué ciencias se considera una rama o subdisciplina la LC?

4) ¿Con qué campo de los mencionados en las definiciones tiene unas conexiones más estrechas la LC? ¿En qué consiste dicho campo?

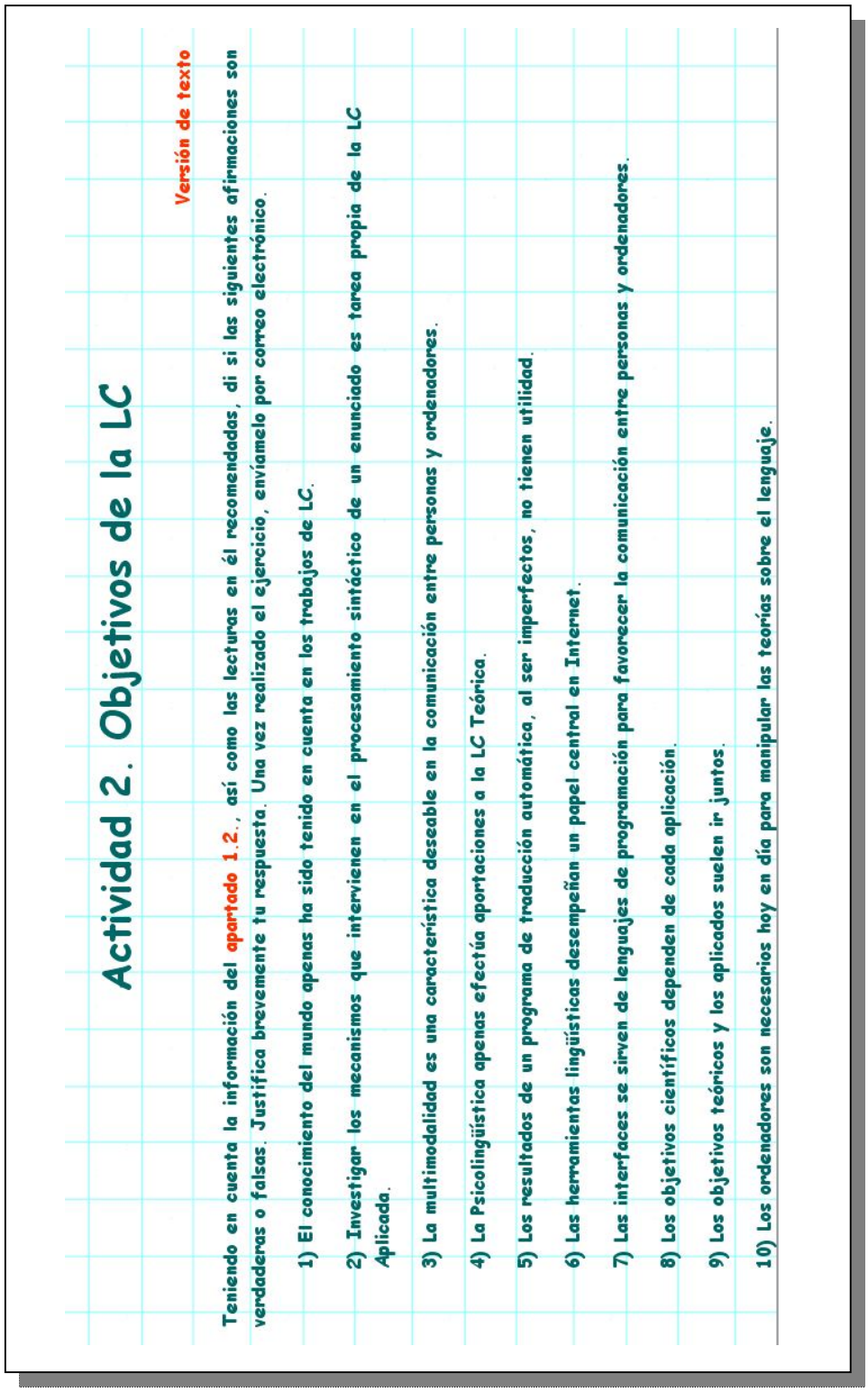
5) ¿Qué diferencia básica existe, a la hora de abordar el estudio del lenguaje, entre la Lingüística tradicional y la LC?

6) ¿Qué dos motivaciones o perspectivas se pueden adoptar en LC? ¿En qué consiste cada una?

7) ¿Qué ámbito de aplicación ha suscitado más interés entre los investigadores? Enumera otras tres áreas o aplicaciones de la LC.

8) ¿Qué relación existe entre los modelos de funcionamiento del lenguaje que propone la LC y el funcionamiento real del cerebro humano?

B.2. Ejemplo de actividad de tipo verdadero / falso



Actividad 2. Objetivos de la LC

Versión de texto

Teniendo en cuenta la información del apartado 1.2., así como las lecturas en él recomendadas, di si las siguientes afirmaciones son verdaderas o falsas. Justifica brevemente tu respuesta. Una vez realizado el ejercicio, envíamelo por correo electrónico.

1) El conocimiento del mundo apenas ha sido tenido en cuenta en los trabajos de LC.

2) Investigar los mecanismos que intervienen en el procesamiento sintáctico de un enunciado es tarea propia de la LC Aplicada.

3) La multimodalidad es una característica deseable en la comunicación entre personas y ordenadores.

4) La Psicolingüística apenas efectúa aportaciones a la LC Teórica.

5) Los resultados de un programa de traducción automática, al ser imperfectos, no tienen utilidad.

6) Las herramientas lingüísticas desempeñan un papel central en Internet.

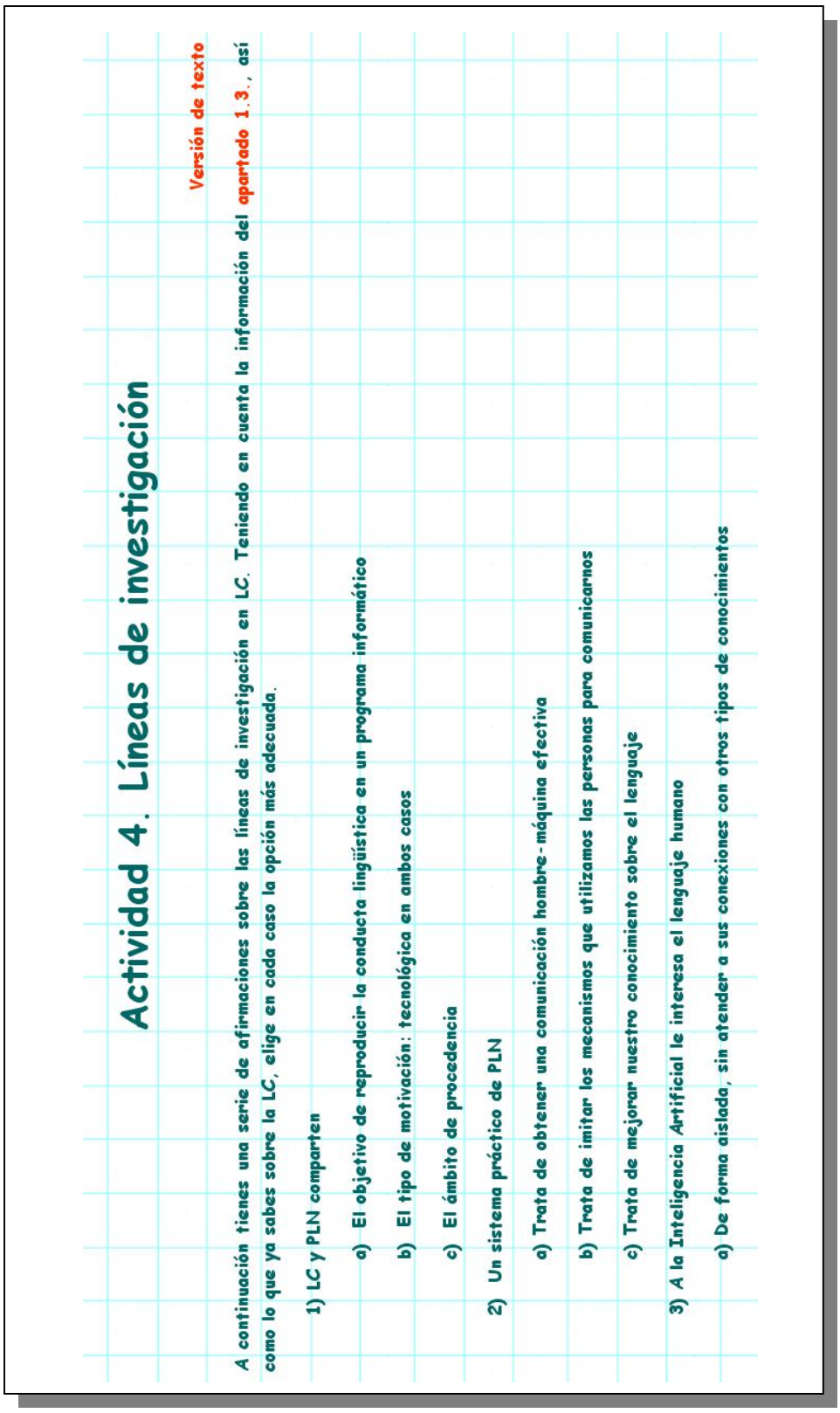
7) Las interfaces se sirven de lenguajes de programación para favorecer la comunicación entre personas y ordenadores.

8) Los objetivos científicos dependen de cada aplicación.

9) Los objetivos teóricos y los aplicados suelen ir juntos.

10) Los ordenadores son necesarios hoy en día para manipular las teorías sobre el lenguaje.

B.3. Ejemplo de actividad de opción múltiple



Actividad 4. Líneas de investigación

Versión de texto

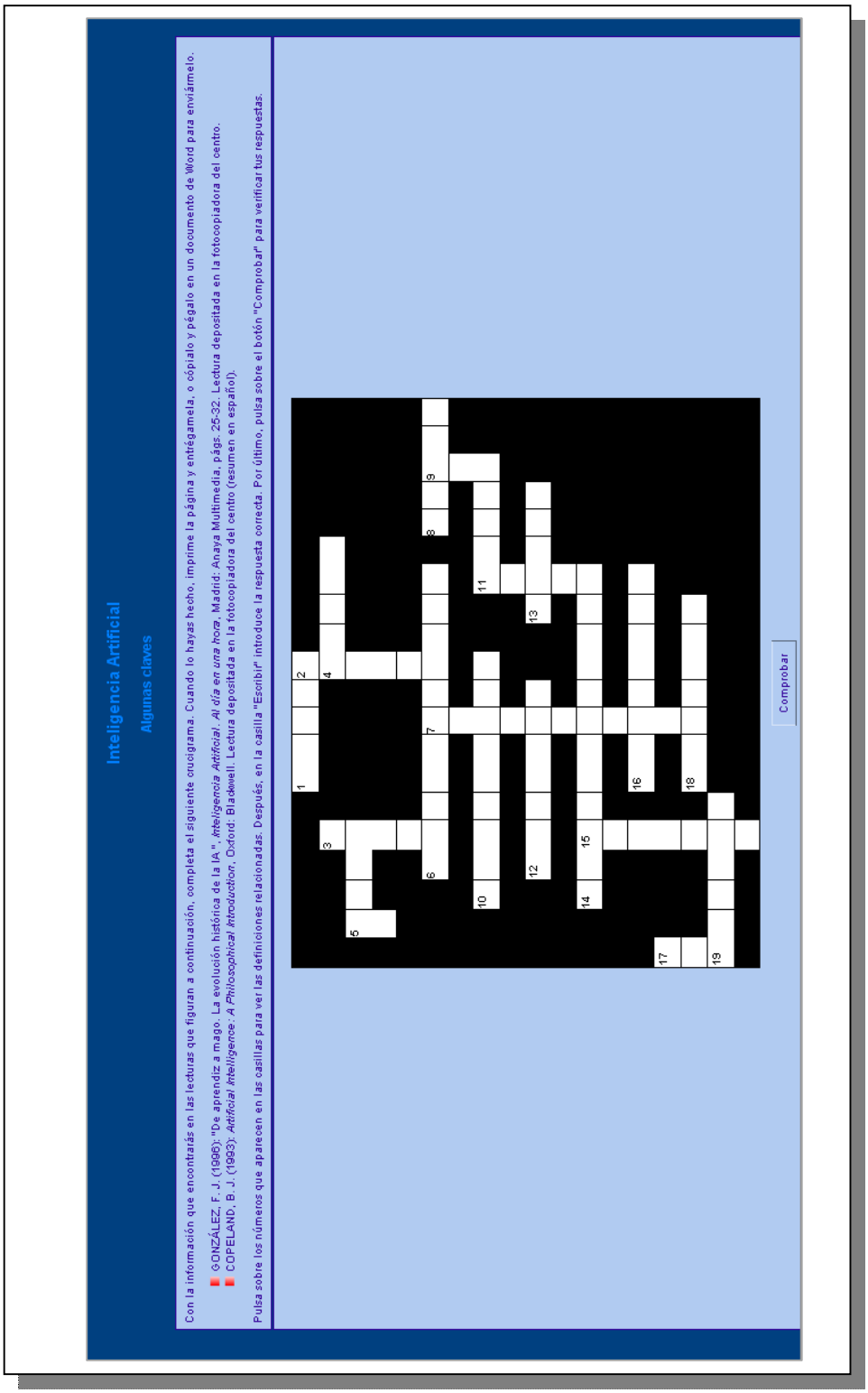
A continuación tienes una serie de afirmaciones sobre las líneas de investigación en LC. Teniendo en cuenta la información del apartado 1.3., así como lo que ya sabes sobre la LC, elige en cada caso la opción más adecuada.

1) LC y PLN comparten
   a) El objetivo de reproducir la conducta lingüística en un programa informático
   b) El tipo de motivación: tecnológica en ambos casos
   c) El ámbito de procedencia

2) Un sistema práctico de PLN
   a) Trata de obtener una comunicación hombre-máquina efectiva
   b) Trata de imitar los mecanismos que utilizamos las personas para comunicarnos
   c) Trata de mejorar nuestro conocimiento sobre el lenguaje

3) A la Inteligencia Artificial le interesa el lenguaje humano
   a) De forma aislada, sin atender a sus conexiones con otros tipos de conocimientos

B.4. Ejemplo de actividad de otro tipo



Horizontal

1. Samuel diseñó un programa que le venció a este juego.
4. Torres de _ _ _ _ _. Popular juego que ejemplifica una forma que tenemos las personas de resolver problemas y que fue imitada por el General Problem Solver. Si quieres intentar encontrar la solución al problema, pulsa aquí.
5. Autómata móvil diseñado por Vaucanson en el s. XVIII.
6. Programa de ajedrez de IBM que inauguró la actual serie Deep Blue, capaz de derrotar al campeón de ajedrez Gary Kasparov.
8. Uno de los primeros sistemas expertos.
10. Nombre del centro universitario donde nació oficialmente la Inteligencia Artificial.
11. Nombre con el que se bautizó al primer ordenador en EE.UU. Visita su museo.
12. País de procedencia de Von Neuman.
13. Junto a McCulloch, representante de la corriente que investigaba el funcionamiento lógico del cerebro.
14. Tipo de lógica empleada en el proyecto emprendido en 1984 por Microelectronics and Computer Technology Corporation.
16. Nombre del investigador que promovió la reunión en la que se acuñó el término "Inteligencia Artificial".
18. El programa Logic Theorist fue capaz de demostrar algunos de los teoremas de este lógico.
19. Matemático inglés que sentó las bases teóricas de la informática.

Vertical

2. Programa desarrollado por T. Winograd que demostró la capacidad de los ordenadores para entender enunciados en un dominio concreto: un mundo de bloques.
3. Autor del álgebra en la que se basaron los trabajos de Turing.
5. Nombre clave del proyecto para desarrollar el primer ordenador.
7. Instrucciones que le indican a un programa los pasos que debe seguir para resolver un problema.
9. Proyecto que pretende recoger en una base de datos los conocimientos referidos a la cultura occidental.
11. Popular programa interactivo diseñado por J. Weizenbaum que simula a una psicoterapeuta. ¿Quieres hablar con el programa? La conversación tendrá que ser en inglés.
15. "¿Puede pensar una _ _ _ _ _ _ _?" es el título del trabajo en que se publicó la definición de inteligencia para un programa informático. ¿Quieres saber más sobre la prueba conocida como "test de Turing"? Pulsa aquí.
17. Nombre actual del centro de investigación donde se ideó el primer ordenador.

5.1.5. Evaluación del curso

- **Cuestionario**
    - Estructura de la asignatura
    - Claridad de los contenidos
    - Interés de las lecturas
    - Utilidad de las actividades
    - Dificultad de
        - contenidos
        - lecturas
        - actividades
        - ejercicios de evaluación
        - utilizar otros idiomas: inglés, catalán, gallego...
    - Secuenciación de contenidos teóricos y prácticos
    - Tiempo para realizar las actividades de evaluación
    - Interés general por la materia
    - Forma de impartir la asignatura
    - Medios para realizar el curso
        - en la Facultad
        - en casa
        - en otros lugares

5.2. Anexo II: Lingüística de corpus

5.2.1. Presentación

Los corpus constituyen uno de los principales recursos lingüísticos de hoy en día: además de servir como fuente inigualable para diversas aplicaciones por la cantidad de datos que aportan sobre el uso de la lengua, también son una forma empírica de acercarse al estudio de la misma y de extraer conclusiones sobre su funcionamiento.

Los objetivos del tema son:

i) presentar la metodología de trabajo basada en corpus;

ii) enumerar los requisitos que debe cumplir un corpus para ser una muestra representativa de una lengua;

iii) comentar los diferentes tipos de corpus posibles y su clasificación en función de varios criterios;

iv) describir las fases necesarias para el desarrollo de un corpus;

v) por último, conocer algunos de los principales corpus que existen, tanto para el español como para otras lenguas.

Lecturas recomendadas

- McEnery, T. y Wilson, A. (1997 [1996]): *Corpus Linguistics*, Edinburgh: Edinburgh University Press. Suplemento web: http://www.lancs.ac.uk/fss/courses/ling/corpus/

Materiales complementarios

- *Informe sobre recursos lingüísticos para el español (II): Corpus escritos y orales disponibles y en desarrollo en España*, Alcalá de Henares: Observatorio Español de Industrias de la lengua, Instituto Cervantes.

- Lavid, J. (2005): "Los ordenadores y la investigación lingüística en la era de la información", en *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*, Madrid: Cátedra, págs. 281-362.

- Martí Antonín, M. A. y Castellón Masalles, I. (2000): "La lingüística de corpus", en *Lingüística Computacional*, Barcelona: Universitat de Barcelona, págs. 151-160.

- McEnery, T. (2003): "Corpus Linguistics", en R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, págs. 448-463.

- Rafel i Fontanals, J. y Soler i Bou, J. (2003): "El procesamiento de corpus", en M. A. Martí Antonín (coord.), *Tecnologías del lenguaje*, Barcelona: UOC, págs.41-73.

- Torruella, J. y Llisterri, J. (1999): "Diseño de corpus textuales y orales", en J. M. Blecua, G. Clavería, C. Sánchez y J. Torruella (eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos,* Barcelona: Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona - Editorial Milenio, págs. 45-77. Disponible también en .pdf: http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf

- Enlaces de interés (Seminario de Lingüística Informática, Universidad de Vigo): http://webs.uvigo.es/sli/paxinas/enlaces.html#corpus

- Gateway to Corpus Linguistics on the Internet (Yvonne Breyer, Macquarie University, Sydney, Australia): http://www.corpus-linguistics.de/

5.2.2. Esquema del tema

1. La lingüística de corpus como metodología lingüística

2. Concepto de corpus

3. Tipos de corpus

4. El desarrollo de un corpus (I): diseño y constitución

5. El desarrollo de un corpus (II): etiquetación. Normas y estándares

6. El desarrollo de un corpus (III): explotación. Programas de concordancias

    - Lectura de J. Lavid (2005): *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*, Madrid, Cátedra, págs. 324-345.

    - J. Llisterri: Herramientas de análisis textual. URL: http://liceu.uab.es/~joaquim/language_resources/lang_res/Herram_TecnTex.html

7. Algunos casos concretos de corpus

5.2.3. Prácticas y actividades

Actividad 1:

Actividad 1. Los trabajos con corpus.

De acuerdo con la información de los apartados 1 y 2, di si las siguientes afirmaciones son verdaderas o falsas. Justifica brevemente tus respuestas.

1) En la actualidad, no es suficiente disponer de un conjunto de datos para poder hablar de corpus.

2) En los estudios históricos, los corpus siempre han sido una herramienta imprescindible.

3) La metodología basada en corpus es una forma de trabajar exclusiva de los siglos XX y XXI.

4) Los lingüistas estructurales creían que los corpus podían recoger todos los datos que precisaban para llevar a cabo sus descripciones.

5) Chomsky rechaza los corpus porque contienen errores.

6) Los corpus no permiten determinar si una estructura es correcta o no en una lengua dada.

7) Pese a las críticas de Chomsky y Abercrombie, los corpus se siguieron empleando en sintaxis.

8) La primera generación de lingüística de corpus se caracteriza por la recopilación de corpus de lengua oral.

9) La validez científica de la lingüística de corpus como método se sustenta sobre todo en la importancia de los datos cuantitativos.

10) Uno de los requisitos más importantes que deben cumplir los corpus en la actualidad es tener un tamaño considerable, de cientos de millones de palabras.

11) La segunda generación de lingüística de corpus se preocupó especialmente por la recopilación de corpus representativos.

12) El SEU fue el primer corpus electrónico de la historia.

13) El resurgir de la lingüística de corpus en los años 80 estuvo motivado por los avances de la Lingüística Aplicada.

14) La informatización de los corpus anula la crítica de Abercrombie.

15) Si reúnes varios textos procedentes de Internet ya tienes un corpus.

Actividad 2:

Actividad 2. Tipos de corpus.

I. De acuerdo con los contenidos del apartado 3 del tema, explica por qué son falsas las siguientes afirmaciones sobre los corpus mencionados. Necesitarás visitar el sitio web de cada uno.

1) British National Corpus (BNC). URL: http://www.natcorp.ox.ac.uk/

1. Es un corpus oral.
2. Es un corpus monitor.

2) CLUVI. URL: http://sli.uvigo.es/CLUVI/

3. Es un corpus general.
4. Es un corpus bilingüe o multilingüe.

3) C-Oral-Rom. URL: http://lablita.dit.unifi.it/coralrom/ y http://www.lllf.uam.es/c-oral-rom/index.html

5. Es un corpus monolingüe.
6. No es un corpus comparable.

4) Hansard Corpus. URL: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20

7. Es un corpus general.
8. Es un corpus comparable.

5) The International Corpus of English. URL: http://www.ucl.ac.uk/english-usage/projects/ice.htm

9. Es un corpus histórico.
10. No es un corpus analizado.

6) Corpus del español. URL: http://www.corpusdelespanol.org/

11. Es un corpus periódico.
12. Es un corpus analizado.

7) COLT, The Bergen Corpus of London Teenage Language. URL: http://www.hf.uib.no/i/Engelsk/COLT/index.html

13. Es un corpus oral orientado a las tecnologías del habla.
14. Es un corpus “parentizado”.

Práctica 1. Análisis de corpus (I): tipos

I1. A continuación tienes una serie de corpus. Marca a qué tipo pertenecen, de acuerdo con los criterios del apartado 3. Ten en cuenta que un mismo corpus puede adscribirse a más de un tipo.

Corpus	Tipo de corpus													
	Escrito	Oral	Monolingüe	Analizado	Anotado	Paralelo	Alineado	Cerrado	Abierto	General	Especializado	Diacrónico	Periódico	Sincrónico
1. CORPES	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
2. CORDE	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
4. ILSP	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
5. CTILC	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
6. Val.Es.Co.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
7. YCOE	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

1. CORPES XXI. URL: http://www.rae.es/ (sección “Banco de datos”)
2. CORDE, Real Academia Española. URL: http://www.rae.es/ (sección “Banco de datos”)
3. ILSP Greek Corpus, Institute of Language and Speech Processing. URL: http://hnc.ilsp.gr/en/
4. CTILC, Corpus textual informatitzat de la llengua catalana, Institut d’Estudis Catalans. URL: http://www.iec.cat/gc/ViewPage.action?siteNodeId=690&languageId=1&contentId=3284
5. Corpus Val.Es.Co. URL: http://www.uv.es/corpusvalesco/
6. YCOE, York-Toronto-Helsinki Parsed Corpus of Old English Prose. URL: http://www-users.york.ac.uk/~lang22/YcoeHome1.htm

Actividad 3:

Actividad 3. Criterios de selección de textos y más cuestiones de diseño.

Después de leer el apartado del tema correspondiente al diseño y constitución de un corpus, di por qué son falsas las siguientes afirmaciones:

1. El diseño de un corpus para estudiar el español desde sus orígenes hasta la actualidad es sencillo.

2. Los corpus deben tener un tamaño considerable, de cientos de millones de palabras.

3. Los corpus formados por textos enteros son más equilibrados.

4. Las muestras de los textos deben ser superiores a las 2000 palabras para reflejar las características lingüísticas que se desea analizar.

5. Es fácil decidir los grupos temáticos que ha de contener un corpus.

6. Es posible recoger en un corpus todas las muestras de una lengua.

7. La representatividad es una noción absoluta: un corpus es representativo o no lo es.

8. La finalidad es irrelevante a la hora de plantear el diseño de un corpus.

9. Todos los corpus son representativos y se pueden tomar como modelo a la hora de elaborar un nuevo corpus.

10. Un corpus monitor debe prestar más atención a la proporción de textos que incluye que un corpus léxico.

11. En un corpus formado por textos producidos por estudiantes de E/LE de distintos niveles, la técnica más adecuada para recoger los textos es el muestreo aleatorio simple.

Práctica 2:

Práctica 2. Análisis de corpus (II): diseño.

Analiza los corpus que figuran más abajo prestando atención a las siguientes cuestiones:

1) Tamaño del corpus (en millones de palabras).
2) Variedad o variedades lingüísticas de los textos que conforman el corpus.
3) Finalidad o finalidades del corpus: objetivo para el que se ha diseñado.
4) Límites temporales que comprenden los textos.
5) Límites geográficos o zonas dialectales a las que pertenecen los textos.
6) Tipo de muestras: textos orales y/o escritos, textos enteros o fragmentos, etc.
7) Proporciones temáticas: tipos de texto de cada tema.

Después, comenta brevemente qué tipo de criterios de diseño predominan: los internos, los externos o una combinación de ambos.

Corpus:

Corpus 1. CREA: http://www.rae.es/ (sección Banco de datos)
Corpus 2. BNC: http://www.natcorp.ox.ac.uk/
Corpus 3. CORGA: http://nunes.cirp.es/corga/
Corpus 4. CTILC: http://ctilc.iec.cat/

Práctica 3:

Práctica 3. Los estándares de etiquetación.

a) Codifica los textos 1 y 2 mediante etiquetas de tipo COCOA. Da cuenta del título del texto, del autor, del poema y de los versos (texto 1), de los párrafos y líneas (texto 2).

b) Codifica la cabecera de los textos 1 y 2 siguiendo las convenciones de la TEI al respecto.

Texto 1

NUEVO CANAL INTEROCEÁNICO

Te propongo construir
un nuevo canal
sin esclusas
ni excusas
que comunique por fin
tu mirada
atlántica
con mi natural pacífico.

(Mario Benedetti, *Antología poética*, Madrid, Alianza Editorial, 2001).

Texto 2

«Yo vengo de una familia en la que cada miembro dañaba de algún modo a los demás. Luego, arrepentidos, cada uno se dañaba a sí mismo.»

Un ranchero quiere que sus cuatro hijos sean sacerdotes; ellos piensan distinto. Un hombre es humillado por su patrón; su hijo quisiera humillarlo más. Una madre renuncia a su carrera de cantante y se pregunta si valió la pena; su hija renuncia al mundo y vive a través de los reality shows. El hijo del presidente se rebela contra su padre, pero depende de su protección. Una mujer sufre el sadismo de su marido por amor. Una madre dolorosa explica la vida de su hija al hombre que la asesinó. Una pareja sesentona se reencuentra y se pregunta si de veras fueron jóvenes amantes. Un comandante debe escoger quién morirá de sus dos hijos. La vieja madre de un joven mariachi lo rescata. Una fiel pareja gay enfrenta la tentación. Una prima fea hace peligrar un matrimonio. Un cura esconde a su hija en una aldea. Un mujeriego se niega a casarse con su amante por temor a matar el placer. Un actor es obligado a enfrentar la realidad por su hijo minusválido. Un Don Juan juega con dos mujeres que le dan su merecido. Tres hijas se reúnen en torno al féretro de su padre por última vez en diez años.

Historias puntuadas por «coros», algunos humorísticos, la mayoría trágicos, que dan voz a los sin-voz: niños mendicantes, hijas violadas, huérfanos, parientes rivales, traficantes, pandillas asesinas que descienden de las calles de Los Ángeles o ascienden de las selvas de Centroamérica.

(Carlos Fuentes, *Todas la familias felices*, España, Alfaguara. **Fecha de publicación:** 06/9/2006. ISBN: 842047083X

Práctica 5:

Práctica 5. Concordancias en algunos corpus.

El fin último de la lingüística de corpus es poder extraer información sobre el funcionamiento real del lenguaje a partir de muestras de su uso por parte de los hablantes de una lengua determinada. En general, se utilizan diferentes herramientas informáticas que proporcionan datos estadísticos o de frecuencia de uso, así como concordancias, ejemplos en los que la palabra o secuencia buscada aparece destacada, rodeada de su contexto anterior y posterior.

Antes de hacer la práctica, realiza las siguientes lecturas relacionadas con el apartado 6 del tema:

- J. Lavid (2005): *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*, Madrid: Cátedra, págs. 324-345. [En fotocopiadora]
- J. Llisterri, Herramientas de análisis textual. URL: http://liceu.uab.es/~joaquim/language_resources/lang_res/Herram_TecnTex.html

En esta actividad te proponemos que emplees los bancos de datos de la Real Academia Española (http://www.rae.es) –en la sección Real Academia Española, Banco de datos– para dar respuesta a una serie de cuestiones sobre el uso del español. Consulta tanto el CREA como el CORDE, según consideres necesario.

1) Aguirre: "En el PP me consideran *lideresa* nacional" (*El País*, 7/11/2007). ¿Está recogida en el *Diccionario* académico (*DRAE*) la palabra *lideresa/s*? ¿Qué información puedes deducir sobre su uso consultando los bancos de datos de la RAE?

2) "El fútbol es el mejor reality que hay" (*Publico*, 22/11/2009). ¿Desde cuándo se recoge la palabra *reality* en los bancos de datos académicos? ¿Dónde predomina su uso? ¿Con qué otras palabras o expresiones se suele usar? ¿Aparece entrecomillado?

3) *Me importa un…* / *me importa una…* + NOMBRE indica que algo nos interesa poco o nada. ¿Qué nombres suelen aparecer en esta estructura? ¿Qué expresión tiene más vitalidad en el español actual? Compara los resultados en el CREA y en el CORDE.

4) *Blog*, *weblog*, *cuaderno de bitácora*. ¿Qué término se utiliza más en el sentido de 'sitio electrónico personal, actualizado con mucha frecuencia, donde alguien escribe a modo de diario o sobre temas que despiertan su interés, y donde quedan recopilados asimismo los comentarios que esos textos suscitan en sus lectores'? ¿A partir de *blog* se documentan palabras derivadas en español? ¿Cuáles?

5) *Palabras claves* / *palabras clave*. Ambas posibilidades son correctas, pero ¿cuál es más habitual? Compara los resultados del CREA y los del CORDE.

6) *Sobredosis de*. Las colocaciones son secuencias de palabras que, sin formar una unidad, suelen aparecer juntas frecuentemente. Según el *DRAE*, una *sobredosis* es una "dosis excesiva de un medicamento o droga". ¿Se usa solo con medicamentos y drogas? Consulta el CREA.

7) ¿De qué suele ser una *tanda*? Anota las posibilidades que se te ocurran. Según el *DRAE*, se trata de un "número indeterminado de ciertas cosas de un mismo género". Consulta ahora los corpus académicos. ¿Podemos concretar la naturaleza de esas "cosas"?

8) Busca los términos *sismo* y *seísmo* en el DRAE y en el DPD (*Diccionario Panhispánico de Dudas*). Apunta su definición. Después consulta los bancos de datos académicos. ¿Qué información sobre su uso puedes deducir?

9) Consulta la definición de *testar* en el DRAE y en el DPD. Después, busca en los bancos de datos académicos. ¿Qué acepción predomina de este término? ¿Influye el factor geográfico?

10) *Dosier*, *dossier*, *dosieres*, *dossieres*, *dosiers*, *dossiers*. Consulta los diccionarios y los bancos de datos académicos. ¿Qué informaciones puedes aportar sobre el uso de esta palabra y de su forma de plural?

6. BIBLIOGRAFÍA

6. BIBLIOGRAFÍA

6.1. Referencias bibliográficas

AARTS, J. y MEIJS, W. (eds.) (1984): *Corpus Linguistics*, Amsterdam: Rodopi.

ABAITUA, J. (2002): "Tratamiento de corpora bilingües", en M. A. MARTÍ y J. LLISTERRI (eds.), *Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita*, Soria: Fundación Duques de Soria-Barcelona: Edicions de la Universitat de Barcelona, 61-90.

ABERCROMBIE, D. (1965): *Studies in Phonetics and Linguistics*, London: Oxford University Press.

ADOLPHS, S. (2006): *Introducing Electronic Text Analysis. A practical guide for language and literary studies*, London / New York: Routledge.

AIJMER, K. y ALTENBERG, B. (eds.) (1991): *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, Londres: Longman.

ALLEN, J. (1995 [1987]): *Natural Language Understanding*, Redwood City, Ca.: Benjamin / Cummings, 2ª ed.

ARMS, W. Y. (2000 [1999]): *Digital Libraries*, Cambridge, Mass.: The MIT Press. Disponible electrónicamente en: http: / / www.cs.cornell.edu / wya / DigLib / MS1999 / glossary.html

BADIA CARDÚS, T. (2003): "Técnicas de procesamiento del lenguaje", en M. A. Martín Antonín (coord.), *Las tecnologías del lenguaje*, Barcelona: UOC, 193-248.

BAKER, C. F.; FILLMORE, Ch. J. y LOWE, J. B. (1998): "The Berkeley FrameNet project", en *Proceedings of the COLING-ACL*, Montreal,

Canadá. Disponible en formato electrónico: http://framenet.icsi.berkeley.edu/papers/acl98.pdf

BATES, M. (1994): "Models of Natural Language Understanding", en D. B. ROE y J. G. WILPON (eds.), *Voice Communication between Humans and Machines*, Washington: National Academy of Sciences, 238-253.

BATTANER, E. *et al.* (2005): "VILE: Estudio acústico de la variación inter e intra locutor en español", *Segundo Coloquio de Lingüística Computacional de la UNAM*, Facultad de Filosofía y Letras, Universidad Nacional Autónoma de México, México, 11 de febrero de 2005. Documento electrónico disponible en: http://liceu.uab.es/~joaquim/phonetics/VILE/VILE_CoLiCo2.pdf.

BENNETT, W. S. (1995): "Machine Translation in North America", en E. F. K. KOERNER y R. E. ASHER (eds.), *Concise History of the Language Sciences: from the Sumerians to the Cognitivists*, Oxford: Elsevier Science, Pergamon, 445-451.

BLACKBURN, P. y BOS, J. (2001): "Inference and Computational Semantics", *Computing Meaning*, Vol. 2, 27-45.

BLACKBURN, P. y BOS, J. (2006a): *Representation and Inference for Natural Language. A First Course in Computational Semantics*, Stanford, CA.: CSLI. Disponible versión electrónica en: http://homepages.inf.ed.ac.uk/jbos/comsem/

BLACKBURN, P. y BOS, J. (2006b): *Working with Discourse Representation Theory: An Advanced Course in Computational Semantics*. Publicación electrónica en: http://homepages.inf.ed.ac.uk/jbos/comsem/.

BLACKBURN, P. y STRIEGNITZ, K. (2002): *Natural Language Processing Techniques in Prolog*, curso disponible electrónicamente en http://cs.union.edu/~striegnk/courses/nlp-with-prolog/html/index.html

BOGURAEV, B.; GARIGLIANO, R. y TAIT, J. (1995): "Editorial", *Natural Language Engineering*, 1 (1), 1-7.

BORRAJO, D. *et al.* (1997 [1993]): *Inteligencia Artificial: Métodos y Técnicas*, Madrid: Centro de Estudios Ramón Areces.

CALVO PÉREZ, J. (2007): "Marcas comerciales y proyección lexicográfica en el español del Perú", *Boletín de la Academia Peruana de la Lengua*, 43 (43), 25-49. Disponible versión electrónica en: http://academiaperuanadelalengua.org/academia/boletin/43/calvo/marcas-comerciales

CAMPILLOS LLANOS, L.; GOZALO GÓMEZ, P. y MORENO SANDOVAL, A. (2007): "El corpus C-ORAL-ROM en la enseñanza de ELE", en E. BALMASEDA MAESTU (coord.), *Las destrezas orales en la enseñanza del español L2-LE: XVII Congreso Internacional de la Asociación del Español como lengua extranjera (ASELE): Logroño 27-30 de septiembre de 2006*, vol. 2, 1115-1128.

CARROLL, J. (2003): "Parsing", en R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 233-248.

CASACUBERTA, F. *et al.* (1992): "Desarrollo de corpus para investigación en tecnologías del habla (Albayzín)", *Procesamiento del Lenguaje Natural*, 12, 35-42. Disponible en formato electrónico en: http://liceu.uab.es/~joaquim/publicacions/Casacuberta_et_al_92_Corpus_Albayzin.pdf

ČERNÝ, J. (2000 [1996]): *Historia de la Lingüística*, Cáceres: Universidad de Extremadura (Versión española traducida por el autor, 1ª ed. en 1998).

CHAN, D. *et al.* (1995): "EUROM - A Spoken Language Resource for the EU", *Eurospeech'95, Proceedings of the 4th European Conference on*

Speech Communication and Speech Technology (Madrid, 18-21 September, 1995), vol. 1, 867-870.

CHOMSKY, N. (1956): "Three models for the description of language", *IRI Transactions on Information Theory*, 2 (3), 113-124.

CHOMSKY, N. (1957): *Syntactic Structures*, The Hague: Mouton.

CHOMSKY, N. (1965): *Aspects of the Theory of Syntax*, Cambridge, MA.: The MIT Press.

CHOMSKY, N. (1981): *Lectures in Government and Binding. (Studies in generative grammar 9)*, Dordrecht: Foris.

CHOMSKY, N. (1995): *The Minimalist program*, Cambridge, MA: The MIT Press.

CHOMSKY, N. y HALLE, M. (1968): *The sound pattern of English,* New York: Harper and Row.

COLE, R. A. *et al.* (eds.) (1996): *Survey of the State of the Art in Human Language Technology*, Cambridge: Cambridge University Press. Publicación electrónica en: http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

COPELAND, B. J. (2000): "The Modern History of Computing", en E. N. ZALTA (ed.), *Stanford Encyclopedia of Philosophy*. URL: http://plato.stanford.edu/archives/spr2001/entries/computing-history/

CRYSTAL, D. (2000 [1980]): *Diccionario de lingüística y fonética*, Barcelona: Octaedro. Traducción y adaptación de X. Villalba. [Original: *A Dictionary of Linguistics and Phonetics*, Oxford: Basil Blackwell]

CUNNINGHAM, H. (1999): "A definition and short history of Language Engineering", *Journal of Natural Language Engineering*, vol. 5, 1-16.

EDWARDS, H. T. (2003): *Applied Phonetics: The sounds of American English*, NT: Delmar Thomson Learning, 3ª ed.

EDWARDS, J. A. y KINGSCOTT, A. G. (eds.) (1997): *Language Industries Atlas*, Amsterdam: IOS Press, 2ª ed.

EVANS, R. y GAZDAR, G. (1996): "DATR: A language for lexical knowledge representation", *Computational Linguistics*, 22.2, 167-216.

FERNÁNDEZ PÉREZ, M. (1986): "Las disciplinas lingüísticas", *Verba*, 13, 15-73.

FERNÁNDEZ PÉREZ, M. (1996): "El campo de la lingüística aplicada. Introducción", en M. FERNÁNDEZ PÉREZ (coord.), *Avances en Lingüística aplicada*, Universidade de Santiago de Compostela: Servicio de Publicacións e Intercambio Científico, 11-45.

FERNÁNDEZ PÉREZ, M. (1999): *Introducción a la Lingüística: Dimensiones del lenguaje y vías de estudio*, Barcelona: Ariel.

FERNÁNDEZ, G. y SÁEZ VACAS, F. (1995): *Fundamentos de Informática. Lógica, Autómatas, Algoritmos y Lenguajes*, Madrid: Anaya Multimedia.

FILLMORE, CH. (1968): "The case for the case", en E. W. BACH y R. T. HARMS (eds.), *Universals in Linguistic Theory*, New York: Holt, Rinehart & Winston, 1-90.

FRANCIS, W. N. (1992): "Language Corpora B.C.", en J. SVARTVIK (ed.), *Directions in Linguistics: Proceedings of Nobel Symposium 82 (Stockholm, 4-8 August 1991)*, Berlin / New York: Mouton de Gruyter, 17-32.

GARRIDO MORAGA, A. M. (1984): "La lingüística y los ordenadores. Consideraciones sobre lingüística mecanizada", *Analecta Malacitana*, Universidad de Málaga, vol. VII, 2, 213-232.

GÓMEZ GUINOVART, J. (1998): "Fundamentos de Lingüística Computacional: bases teóricas, líneas de investigación y

aplicaciones", en J. BARÓ I QUERALT y P. CID LEAL (eds.), *Anuario SOCADI de Documentación e Información*, Barcelona: Societat Catalana de Documentació i Informació, 135-146. URL: http://www.raco.cat/index.php/Bibliodoc/article/viewFile/56629/66051

GÓMEZ GUINOVART, J. (1999): "Introducción", en J. GÓMEZ GUINOVART *ET AL.* (eds.), *Panorama de la investigación en lingüística informática*, *Monografía de Revista Española de Lingüística Aplicada*, Logroño, 7-9.

GÓMEZ GUINOVART, X. (2000a): "Lingüística computacional", en F. RAMALLO, G. REI-DOVAL y X. P. RODRÍGUEZ YÁÑEZ (eds.), *Manual de Ciencias da Linguaxe*, Vigo: Xerais, cap. 6, 221-268. URL: http://webs.uvigo.es/sli/arquivos/xerais.pdf

GÓMEZ GUINOVART, X. (2000b): "Perspectivas de la lingüística computacional", *Novática: Revista de la Asociación de Técnicos de Informática*, Número especial del 25 aniversario (*Horizonte 2025*), 145, 85-87. Publicación electrónica en: http://www.ati.es/novatica/2000/145/javgom-145.pdf

GÓMEZ GUINOVART, J. y PALOMAR, M. (coords.) (1998): *Lengua y Tecnologías de la Información*, número monográfico de *Novática*, 133.

GÓMEZ TORREGO, L. (2007): *Gramática didáctica del español*, Madrid: Ediciones SM.

GREGORY, R. L. (ed.) (1995 [1987]): *Diccionario Oxford de la mente*, Madrid: Alianza Diccionarios.

GRISHMAN, R. (1991 [1986]): *Introducción a la lingüística computacional*, Madrid: Visor. Traducción de A. Moreno Sandoval.

GUTIÉRREZ ORDÓÑEZ, S. (1981): *Lingüística y Semántica. Aproximación funcional*, Oviedo: Universidad de Oviedo.

GUTIÉRREZ ORDÓÑEZ, S. (1989): *Introducción a la Semántica Funcional*, Madrid: Síntesis.

HALVORSEN, P.-K. (1991 [1988]): "Las aplicaciones informáticas de la teoría lingüística", en F. J. NEWMEYER (comp.), *Panorama de la Lingüística Moderna de la Universidad de Cambridge, vol. II: Teoría lingüística: Extensiones e Implicaciones*, Madrid: Visor. Traducción de J. Gómez Guinovart y A. Tusón Valls. Edición supervisada por L. Eguren, 247-271. [Original: *Linguistics: The Cambridge Survey II: Linguistic Theory: Extensions and Implications*, Cambridge: Cambridge University Press, 198-219].

HANNAHS, S. J. (2001): "Morphophonology", en P. B. BALTES y N. J. SMELSER (eds.), *International Encyclopedia of the Social & Behavioral Sciences*, Amsterdam: Elsevier Science, 10053-10058.

HAUSSER, R. (2001 [1999]): *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*, 2ª ed., revisada y ampliada, Berlin: Springer-Verlag.

HERNÁNDEZ FIGUEROA, Z.; PÉREZ AGUIAR, J. R. y SANTANA SUÁREZ, O. (2000a): "Análisis y comprensión del lenguaje", seminario disponible electrónicamente en: http://protos.dis.ulpgc.es/docencia/seminarios/pln/Analisis_y_comprension/index.htm

HERNÁNDEZ FIGUEROA, Z.; PÉREZ AGUIAR, J. R. y SANTANA SUÁREZ, O. (2000b): "Procesamiento del Lenguaje Natural", seminario disponible electrónicamente en: http://protos.dis.ulpgc.es/docencia/seminarios/pln/Analisis_y_comprension/sld019.htm

HUTCHINS, W. J. (1986): *Machine translation: past, present, future*, Chichester, Ellis Horwood: Ellis Horwood Series in Computers and

their Applications. Disponible en: http://www.hutchinsweb.me.uk/PPF-TOC.htm

HUTCHINS, W. J. (1995): "Machine Translation: A Brief History", en E. F. K. KOERNER y R. E. ASHER (eds.), *Concise History of the Language Sciences: from the Sumerians to the Cognitivists*, Oxford: Elsevier Science, Pergamon, 431-445.

HUTCHINS, W. J. (1996): "ALPAC: the (in)famous report", *MT News International*, 14, June, 9-12. Publicación electrónica en: http://ourworld.compuserve.com/homepages/WJHutchins/Alpac.htm

HUTCHINS, W. J. (1999): "Warren Weaver memorandum: 50th anniversary of machine translation", *MT News International*, 22, vol. 8.1, 5-6. Publicación electrónica en: http://ourworld.compuserve.com/homepages/WJHutchins/Weaver49.htm

HUTCHINS, W. J. (2000a): "The first decades of machine translation: overview, chronology, sources", en W. J. HUTCHINS (ed.), *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*, Amsterdam/Philadelphia: John Benjamins, 1-16.

HUTCHINS, W. J. (2000b): "Yehoshua Bar-Hillel: a philosopher's contribution to machine", en W. J. HUTCHINS (ed.), *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*, Amsterdam/Philadelphia: John Benjamins, 299-312.

HUTCHINS, W. J. (2001): "Machine translation over fifty years", *Histoire Épistémologie Langage*, XXIII, 1, 7-31. URL: http://www.hutchinsweb.me.uk/HEL-2001.pdf

HUTCHINS, W. J. (2005): "The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954".

Publicación electrónica en: http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf

HUTCHINS, W. J. y SOMERS, H. L. (1995 [1992]): *Introducción a la traducción automática*, Madrid: Visor. Traducción dirigida por J. K. Abaitua. [Original: *An Introduction to Machine Translation*, London: Academic Press]

Ingeniería lingüística. Cómo aprovechar la fuerza del lenguaje. Luxemburgo: Anite Systems. Versión española a cargo del Observatorio Español de Industrias de la Lengua, Instituto Cervantes.

JOHNSON, K. y JOHNSON, H. (eds.) (1998): *Encyclopedic Dictionary of Applied Linguistics: A Handbook for Language Teaching*, Oxford: Blackwell.

JOSHI, A. K. (2002 [1999]): "Lingüística computacional", en R. A. WILSON y F. C. KEIL (eds.), *Enciclopedia MIT de ciencias cognitivas*, Madrid: Síntesis, vol. I, 745-748.

JURAFSKY, D. y MARTIN, J. H. (2000): *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Upper Saddle River, New Jersey: Prentice Hall. Publicación electrónica (cap. 1): http://www.cs.colorado.edu/~martin/SLP/slp-ch1.pdf

KARTTUNEN, L. (2003): "Finite-state technology", en R. MITKOV (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 339-357.

KARTTUNEN, L. y BEESLEY, D. R. (2005): "Twenty-five years of finite-state morphology", en A. ARPPE *ET AL.* (eds.), *Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday*, Stanford, California: CSLI Publications, 71-83. Disponible electrónicamente en:

http://cslipublications.stanford.edu/koskenniemi-festschrift/8-karttunen-beesley.pdf

KAY, M. (1979): "Functional grammar", *Proceedings of the 5th Annual Meeting of the Berkeley Linguistic Society*, Berkeley, California, 142-158.

KAY, M. (2003): "Introduction", en R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, xvii-xx.

KELLER, B. (1995): "DATR theories and DATR models", *33rd Annual Meeting of the Association for Computational Linguistics*, 55-62.

KELLER, B. (1996): "An evaluation semantics for DATR theories", *COLING-96*, 646-651.

KLAVANS, J. (1997): "Computational linguistics", en W. O'GRADY, M. DOBROVOLSKY y F. KATAMBA (eds.), *Contemporary Linguistics. An Introduction*, London/New York: Longman, cap. 17, 664-702. [Adaptación de W. O'Grady y M. Dobrovolsky (eds.) (1987), *Contemporary Linguistics Analysis: An Introduction*, Toronto: Copp Clark Pitman]

KOERNER, E. F. K. (2002): *Toward a History of American Linguistics*, London/New York: Routledge.

KOSKENNIEMI, K. (1983): *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*, Helsinki: Universidad de Helsinki.

LABOV, W. (1969): "The logic of non-standard English", *Georgetown Monographs on Language and Linguistics*, 22.

LAPPIN, SH. (2003): "Semantics", en R. MITKOV (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 91-111.

LAVID, J. (2005): *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*, Madrid: Cátedra.

LAWLER, J. y ARISTAR DRY, H. (eds.) (1998): *Using Computers in Linguistics. A Practical Guide*, London and New York: Routledge.

LEECH, G. (1991): "Corpora", en K. MALMKJAER (ed.), *The Linguistics Encyclopedia*, London / New York: Routledge, 73-80.

LEECH, G. (1992): "Corpora and theories of linguistic performance", en J. SVARTVIK (ed.), *Directions in Linguistics: Proceedings of Nobel Symposium 82 (Stockholm, 4-8 de agosto 1991)*, Berlin: Mouton de Gruyter, 105-122.

Lenguaje y tecnología. De la torre de Babel a la aldea global, Luxemburgo: Oficina de Publicaciones Oficiales de las Comunidades Europeas, 1997.

LLISTERRI, J. (1999): "Tecnologías lingüísticas y sociedad de la información", *Economía Industrial (La sociedad de la información en España I)*, 325, 37-56. Publicación electrónica en: http://liceu.uab.cat/~joaquim/publicacions/Llisterri_99_TecnolLing_SocInfo.pdf

LLISTERRI, J. (2003): "Las tecnologías del habla: Entre la ingeniería y la lingüística", *Actas del Congreso Internacional "La Ciencia ante el Público. Cultura humanística y desarrollo científico y tecnológico" (Universidad de Salamanca, 28-31 de octubre 2002)*, Salamanca: Instituto Universitario de Estudios de la Ciencia y la Tecnología, edición en CD-ROM, 44-67. Publicación electrónica en: http://liceu.uab.es/~joaquim/publicacions/TecnolHab_Salamanca_02.pdf

LLISTERRI, J. (2004): "Las tecnologías del habla para el español", en R. SEQUERA (ed.), *Ciencia, tecnología y lengua española: la terminología*

científica en español, Madrid: Fundación Española para la Ciencia y la Tecnología, 123-141. Publicación electrónica en: http://liceu.uab.es/~joaquim/publicacions/TecnolHablaEsp_FECyT03.pdf

LLISTERRI, J. *et al.* (2005): "Corpus orales para el desarrollo de las tecnologías del habla en español", *Oralia. Análisis del discurso oral*, 8, 289-325. Disponible en formato electrónico en: http://liceu.uab.es/~joaquim/publicacions/Llisterri_Machuca_Mota_Riera_Rios_05_Corpus_Orales_Tecnologias_Habla_Espanol.pdf

LLISTERRI, J. y GARRIDO ALMIÑANA, J. M. (1998): "La ingeniería lingüística en España", *El español en el mundo*, Madrid: Arco/Libros y Alcalá de Henares: Centro Virtual Cervantes, Instituto Cervantes. Publicación electrónica en: http://cvc.cervantes.es/obref/anuario/anuario_98/llisterri/

MANNING, D. D. y SCHÜTZE, H. (1999): *Foundations of Statistical Natural Language Processing*, Cambridge, Mass./London, England: The MIT Press. Capítulo 1 disponible en formato electrónico: http://www.csd.uwo.ca/courses/CS442b/Books/StatNatLangProc/chap1.pdf y otros materiales disponibles de forma electrónica en: http://nlp.stanford.edu/fsnlp/.

MARTÍ ANTONÍN, M.ª A. (coord.) (2001): *Les tecnologies del llenguatge*, Barcelona: Universitat Oberta de Catalunya.

MARTÍ ANTONÍN, M.ª A. (coord.) (2003): *Tecnologías del lenguaje*, Barcelona: Editorial UOC.

MARTÍ ANTONÍN, Mª. A. y CASTELLÓN MASALLES, I. (2000): *Lingüística computacional*, Barcelona: Universitat de Barcelona.

MARTÍ, Mª. A. y LLISTERRI, J. (2001): "La ingeniería lingüística en la sociedad de la información", *Digit-HVM*, *Revista Digital*

d'Humanitats, 3. Publicación electrónica en: http://www.uoc.es/humfil/articles/esp/llisterri-marti/llisterri-marti.html

MCCORDUCK, P. (1991 [1979]): *Máquinas que piensan. Una incursión personal en la historia y las perspectivas de la inteligencia artificial*, Madrid: Tecnos (Traducción de D. Cañamero).

MCENERY, T. (2003): "Corpus Linguistics", en R. MITKOV (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 448-463.

MCENERY, T. y WILSON, A. (1996): *Corpus Linguistics*, Edinburgh: Edinburgh University Press. Suplemento web: http://www.lancs.ac.uk/fss/courses/ling/corpus/

MCENERY, T. y WILSON, A. (2001): *Corpus Linguistics*, Edinburgh: Edinburgh University Press, 2ª ed.

MCENERY, T.; XIAO, R. y TONO, Y. (2006): *Corpus-Based Language Studies: an advanced resource book,* London/New York: Routledge.

MEYA, M. (1980): "La inteligencia artificial", *Revista Española de Lingüística*, 10/1, págs. 135-159.

MINSKY, M. (1975): "A Framework for Representing Knowledge", en P. WINSTON (ed.), *The Psychology of Computer Vision*, New York: McGraw-Hill, 211-277.

MITKOV, R. (ed.) (2003): *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press.

MORALA, J. R. (2002): "Nuevas tecnologías y recursos lexicográficos: fuereño", en G. CLAVERÍA (coord.), *Filología en Internet*, Bellaterra, Barcelona: Servei de Publicacions, Universitat Autònoma de Barcelona, 45-53.

MORENO BORONAT, L. *et al.* (1999): *Introducción al procesamiento del Lenguaje Natural*, Alicante: Universidad de Alicante.

MORENO BORONAT, L. y MOLINA MARCO, A. (1999): "Preliminares y tendencias en el Procesamiento del Lenguaje Natural", *Inteligencia Artificial*, 7, Primavera, 65-76. Disponible en PostScript (se puede abrir con Adobe Acrobat). Publicación electrónica en: http://ltcs.uned.es:8080/aepia/Uploads/7/174.gz

MORENO ORTIZ, A. (2000): "Diseño e implementación de un lexicón computacional para lexicografía y traducción automática", *Estudios de Lingüística Española*, 9. Disponible en formato electrónico: http://elies.rediris.es/elies9/index.htm

MORENO SANDOVAL, A. (1998): *Lingüística computacional. Introducción a los modelos simbólicos, estadísticos y biológicos*, Madrid: Síntesis.

MORENO SANDOVAL, A. (2001): *Gramáticas de unificación y rasgos*, Madrid: Antonio Machado.

MORENO SANDOVAL, A. *et al.* (1993): "*PROTEUS*: un sistema multilingüe de extracción de información", *Procesamiento del Lenguaje Natural*, 13, 47-56.

MOURE, T. (2002): *La lingüística en el conjunto del conocimiento: Una mirada crítica*, Lugo: Tris Tram.

MOURE, T. y LLISTERRI, J. (1996): "Lenguaje y nuevas tecnologías: el campo de la lingüística computacional", en M. FERNÁNDEZ PÉREZ (coord.), *Avances en Lingüística aplicada*, Universidade de Santiago de Compostela: Servicio de Publicacións e Intercambio Científico, 147-227. Publicación electrónica: http://liceu.uab.es/~joaquim/publicacions/llisterri_moure_96.html

NEWMEYER, F. J. (1986): *Linguistic Theory in America*, San Diego, Ca.: Academic Press, 2ª ed.

Oficina de Español en la Sociedad de la Información (OESI): http://oesi.cervantes.es//

PAYRATÓ, L. (1998): *De profesión, lingüista. Panorama de la lingüística aplicada*, Barcelona: Ariel.

PENA, J. (1999): "Partes de la morfología. Las unidades del análisis morfológico", en I. BOSQUE y V. DEMONTE (dir.), *Gramática Descriptiva de la Lengua Española (vol. 3. Entre la oración y el discurso. Morfología)*, Madrid: Espasa, 4305-4366.

PIERA, C. y VARELA, S. (1999): "Relaciones entre morfología y sintaxis", en I. BOSQUE y V. DEMONTE (dir.), *Gramática Descriptiva de la Lengua Española (vol. 3. Entre la oración y el discurso. Morfología)*, Madrid: Espasa, 4367-4422.

PRUÑONOSA TOMÁS, M. (1996): "La palabra", en C. MARTÍN VIDE (ed.), *Elementos de Lingüística*, Barcelona: Octaedro, 171-200.

PUSTEJOVSKY, J. (1991): "The Generative Lexicon", *Computational Linguistics*, vol. 17, 4., 409-441.

PUSTEJOVSKY, J. (1995): *The Generative Lexicon*, Cambridge, MA.: The MIT Press.

QUILLIAN, M. R. (1968): "Semantic Memory", en M. MINSKY (ed.), *Semantic Information Processing,* Cambridge, Mass: The MIT Press, 27-70.

RAMSAY, A. M. (1991): "Artificial Intelligence", en K. MALMKJAER (ed.), *The Linguistics Encyclopedia*, London and New York: Routledge, 28-38.

RAMSAY, A. M. (2003): "Discourse", en R. MITKOV (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 112-135.

REAL ACADEMIA ESPAÑOLA (2001): *Diccionario de la lengua española*, Madrid: Espasa, 22ª edición. URL: http://www.rae.es

REAL ACADEMIA ESPAÑOLA (2005): *Diccionario panhispánico de dudas*, Madrid: Santillana. URL: http://www.rae.es

RODRÍGUEZ HONTORIA, H. (2000): "Técnicas básicas en el tratamiento informático de la lengua", *Quark*, nº 19, julio-diciembre, *Monográfico sobre Tecnologías de la lengua*. Publicación electrónica: http://www.prbb.org/quark/19/019026.htm

RODRÍGUEZ HONTORIA, H. (2002): "Técnicas de análisis sintáctico", en M. A. MARTÍ y J. LLISTERRI (eds.), *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita*, Soria: Fundación Duques de Soria y Barcelona: Edicions de la Universitat de Barcelona, 91-132.

ROECK, A. de (1995) : "Computational linguistics", en J. VERSCHUEREN, J.-O. ÖSTMAN y J. BLOMMAERT (eds.), *Handbook of Pragmatics, Manual*, Amsterdam/Philadelphia: John Benjamins, 154-164.

ROJO, G. (1986): *El lenguaje, las lenguas y la lingüística*, Universidad de Santiago de Compostela (Lalia: Serie Lingüística, Departamento de Filología Española, Teoría de la Literatura y Lingüística General, 1).

SANTALLA DEL RÍO, M.ª P. (2005): "La elaboración de corpus lingüísticos", en M. CAL, P. NÚÑEZ e I. M. PALACIOS (eds.), *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas*, Universidade de Santiago de Compostela: Servizo de Publicacións e Intercambio Científico, 45-63.

SANTANA, O. *et al.* (1997): "FLAVER: Flexionador y lematizador automático de formas verbales", *Lingüística Española Actual*, XIX, 2, Madrid: Arco/Libros, 229-282. Disponible en formato electrónico en: http://www.gedlc.ulpgc.es/art_ps/art28.pdf

SANTANA, O. *et al.* (1998): "Reconocedor y generador automático de formas nominales", *Diccionarios e informática*, Jaén: Publicaciones de la Universidad de Jaén, 57-74. Disponible en formato electrónico en: http://protos.dis.ulpgc.es/art_ps/art28b.pdf

SANTANA, O. *et al.* (1999): "FLANOM: Flexionador y lematizador automático de formas nominales", *Lingüística Española Actual*, XXI, 2, Madrid: Arco/Libros, 253-297. Disponible en formato electrónico en: http://www.gedlc.ulpgc.es/art_ps/art29.pdf

SANTANA, O. *et al.* (2002): "Hacia la desambiguación funcional automática en español", *Procesamiento de Lenguaje Natural*, Revista Nº 28, SEPLN, 1-22.

SANTANA, O. *et al.* (2004): "Bases para la desambiguación estructural de árboles de representación sintáctica", *Procesamiento de Lenguaje Natural*, Revista Nº 32, SEPLN, 43-65.

SANTANA, O. *et al.* (2005): "Spanish Morphosyntactic Disambiguator", *The 17th Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (ACH/ALLC 2005), Conference Abstracts*, 207-209.

SANTANA, O. *et al.* (2006): "Functional Disambiguation Based on Syntactic Structures", *Literary and Linguistic Computing*, Vol. 21, No. 2, 187-197.

SCHANK, R. (1972): "Conceptual dependency: A theory of natural language understanding", *Cognitive Psychology*, 3(4), 552-631.

SCHANK, R. (1975): *Conceptual Information Processing*, Amsterdam: North Holland.

SCHANK, R. y ABELSON, R. (1977): *Scripts, Plans, Goals and Understanding. An Inquiry into Human Knowledge Structures*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

SHIEBER, S. M. (1988): "Separating linguistic analyses from linguistic theories", en U. REYLE y C. ROHRER (eds.), *Natural Language Parsing and Linguistic Theories,* Dordrecht: Reidel, 33-68.

SHIEBER, S. M. (1989 [1986]): *Introducción a los formalismos gramaticales de unificación*, Barcelona: Teide. [Original: *An Introduction to Unification-Based Approaches to Grammar*, Chicago: University of Chicago Press]

SIMON, H. A. (1995): "Artificial intelligence: an empirical science", *Artificial Intelligence*, 77, 95-127.

SINCLAIR, J. (1994): *EAGLES*, Document EAG-CWG-IR-2.

SINCLAIR, J. (1996): *EAGLES Preliminary recommendations on Corpus Typology*. Documento electrónico: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

SLAMA-CAZACU, T. (1981): "Sur l'objet de la linguistique appliquée", *Revue Roumaine de linguistique*, XXVI, 5-21.

SLAMA-CAZACU, T. (1984): *Linguistique Appliquée: Une introduction*, Brescia: La Scuola.

SPARCK JONES, K. (1994): "Natural Language Processing: A Historical Review", en A. ZAMPOLLI, N. CALZOLARI y M. PALMER (eds.), *Current Issues in Computational Linguistics: In Honour of Don Walker*, Linguistica Computazionale, vol. IX-X, Pisa: Giardini y Norwell (USA): Kluwer, 3-16.

SPROAT, R. (1992): *Morphology and computation*, Cambridge, Mass.: The MIT Press.

SVARTVIK, J. (1992): "Corpus linguistics comes of age", en J. SVARTVIK (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 (Stockholm, 4-8 August 1991)*, Berlin/New York: Mouton de Gruyter, 7-13.

TENNANT, H. (1981): *Natural Language Processing. An introduction to an Emerging Technology*, New York: Petrocelli Books.

TOMALIN, M. (2002): "The formal origins of syntactic theory", *Lingua*, 112, 827-848.

TORRUELLA, J. y LLISTERRI, J. (1999): "Diseño de corpus textuales y orales", en J. M. Blecua *et al.* (eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos*, Barcelona: Milenio y Universidad Autónoma de Barcelona, Dpto. de Filología Española, 45-77. Disponible electrónicamente en .pdf: http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf

TRASK, R. L. (1993): *A Dictionary of Grammatical Terms in Linguistics*, London-New York: Routledge.

TROST, H. (2003): "Morphology", en R. MITKOV (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 25-47.

Universidad de Piura PLANCAD, Facultad de Ciencias de la Educación (2001): *Fascículo autoinstructivo. Abordar la realidad lingüística*, Lima: Ministerio de Educación/DINFOCAD/UCAD/PLANCAD. Disponible en versión electrónica en: http://ciberdocencia.gob.pe/archivos/fasciculo_Comunicacion_abordar_la_realidad.doc

USZKOREIT, H. (1996, 2000): *What is Computational Linguistics?*, Department of Computational Linguistics and Phonetics, Saarland University, Saarbrücken. Documento electrónico disponible en: http://www.coli.uni-saarland.de/~hansu/what_is_cl.html

VÁZQUEZ, G.; FERNÁNDEZ MONTRAVETA, A. y MARTÍ, M. A. (2002): "Léxicos verbales computacionales", en M. A. MARTÍ y J. LLISTERRI

(eds.), *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita*, Barcelona: Edicions de la Universitat de Barcelona / Soria: Fundación Duques de Soria, 29-60.

VERDEJO, Mª. F. (1995): "Comprensión del lenguaje natural: avances, aplicaciones y tendencias", *Arbor*, *CLI*, 595 (Julio), 39-83.

VERDEJO, M.ª F. y GONZALO, J. (1998): "Del procesamiento del lenguaje natural a la Ingeniería Lingüística: ¿dónde nos encontramos?", *Novática (Monografía: Inteligencia Artificial)*, enero / febrero, 131, 29-36.

VIDAL VILLALBA, J. y BUSQUETS RIGAT, J. (1996): "Lingüística computacional", en C. MARTÍN VIDE (ed.), *Elementos de lingüística*, Barcelona: Octaedro Universidad, 393-446.

WIKIPEDIA, *La Enciclopedia Libre*. URL: http://es.wikipedia.org/wiki/Portada

WIKIPEDIA, *The Free Encyclopedia*. URL: http://en.wikipedia.org/wiki/Main_Page

WINOGRAD, T. (1972): *Understanding Natural Language*, San Diego, CA.: Academic Press.

WINOGRAD, T. (1983): *Language as a Cognitive Process, Volume I: Syntax*, Reading, Mass.: Addison-Wesley.

YNGVE, V. H. (2000): "Early research at M.I.T.: in search of adequate theory", en W. J. HUTCHINS (ed.), *Early Years in Machine Translation: Memoirs and Biographies of Pioneers*, Amsterdam / Philadelphia: John Benjamins, 39-72.

6.2. Otra bibliografía consultada

AI Topics, American Association for Artificial Intelligence. URL: http://www.aaai.org/aitopics/pmwiki/pmwiki.php/AITopics/HomePage

Alan Turing.net, *The Turing Archive for the History of Computing*. URL: http://www.alanturing.net/turing_archive/index.html

ALLEN, J. (1998): "AI Growing Up. The Changes and Opportunities", *AI Magazine*, 19 (4), 13-23.

ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL): *What is Computational Linguistics?*. URL: http://www.aclweb.org/archive/what.html

AZORÍN FERNÁNDEZ, D. (2003): "La lexicografía como disciplina lingüística", en A. M. MEDINA GUERRA (coord.), *Lexicografía española*, Barcelona: Ariel, 31- 52.

BAJO MOLINA, M. T. y CAÑAS DELGADO, J. J. (1991): *Ciencia cognitiva*, Madrid: Debate.

BARNBROOK, G. (1996): *Language and Computers: A Practical Introduction to the Computer Analysis of Language*, Edinburgh: Edinburgh University Press.

BATES, M. y WEISCHEDEL, R. M. (eds.) (1993): *Challenges in Natural Language Processing*, Cambridge, Mass.: Cambridge University Press.

BIBER, D. (2006): *University Language: A corpus-based study of spoken and written registers*, Amsterdam/Philadelphia: John Benjamins.

BIBER, D.; CONRAD, S. y REPPEN, R. (1998): *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge: Cambridge University Press.

BOTT, M. F. (1975 [1970]): "Lingüística computacional", en J. LYONS, *Nuevos horizontes de la lingüística*, Madrid: Alianza Editorial, 227-240.

BRACE, C. y JOSCELYNE, A. (1991): "Henry Kucera", *Language Industry Monitor*, enero/febrero. Publicación electrónica en: http://www.lim.nl/monitor/kucera.html

CAL, M.; NÚÑEZ, P.; PALACIOS, I. M. (eds.) (2005): *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas*, Universidade de Santiago de Compostela: Servizo de Publicacións e Intercambio Científico.

CARAVEDO, R. (ed.) (1999): *Lingüística del Corpus. Cuestiones teórico-metodológicas aplicadas al español*, Salamanca: Ediciones Universidad de Salamanca.

CHOMSKY, N. (1959): "On certain formal properties of grammars", *Information and Control*, 2, 137-167.

CLAVERIA, G. (coord.) (2002): *Filología en Internet*, Universitat Autònoma de Barcelona: Servei de Publicacions.

COLEMAN, J. (2005): *Introducing Speech and Language Processing*, Cambridge: Cambridge University Press.

CORI, M. y MARANDIN, J. M. (2001): "La linguistique au contact de l'informatique: de la construction des grammaires aux grammaires de construction", *HEL*, 23/I, 49-79.

CRUSE, D. A. (1986): *Lexical Semantics*, Cambridge: Cambridge Universtiy Press.

DE KOCK, J. (ed.) (2001): *Lingüística con corpus. Catorce aplicaciones sobre el español*, Salamanca: Ediciones Universidad de Salamanca.

DÍEZ ORZAS, P. L. (1999): *La relación de meronimia en los sustantivos del léxico español: contribución a la semántica computacional*, *Estudios de Lingüística Española*, vol. 2. Publicación electrónica en: http://elies.rediris.es/elies2/index.htm.

ELUERD, R. (2000): *La lexicologie*, Paris: Presses Universitaires de France.

FEIGENBAUM, E. A. y FELDMAN, J. (eds.) (1963): *Computers and Thought*, New York: McGraw-Hill.

FERNÁNDEZ PÉREZ, M. (2005): "Aplicaciones de la Lingüística y nuevas tecnologías. De hecho, pareja", en M. CAL, P. NÚÑEZ e I. M. PALACIOS (eds.), *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas*, Santiago de Compostela: Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico, 29-44.

FORCADA, V. M.; GIL DE CARRASCO, A. y SAGER, J. C. (comp.) (1996): *Estudios computacionales del español y el inglés*, Madrid: Instituto Cervantes.

GLADKIJ, A. V. y MEL'ČUK, Í. A. (1972): *Introducción a la lingüística matemática*, Barcelona: Planeta.

GONZÁLEZ, F. J. (1996): *Inteligencia Artificial*, Madrid: Anaya Multimedia.

GRANGER, S.; HUNG, J. y PETCH-TYSON, S. (eds.) (2002): *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam/Philadelphia: John Benjamins.

HARRIS, Z. S. (1962): *String Analysis of Sentence Structure*, The Hague: Mouton and Co.

Informe sobre recursos lingüísticos para el español (II): Corpus escritos y orales disponibles y en desarrollo en España, Alcalá de Henares: Observatorio Español de Industrias de la lengua, Instituto Cervantes.

KARTTUNEN, L. y BEESLEY, D. R. (2001): "A short history of two-level morphology". URL: http://www.ling.helsinki.fi/~koskenni/esslli-2001-karttunen/

LIBERMAN, M. Y. (1990): "The ACL Data Collection Initiative", *Information Technology*, *'Next Decade in Information Technology', Proceedings of the 5th Jerusalem Conference on (Cat. No.90TH0326-9)*, 781-786.

LLISTERRI, J. (ed.) (1998): "Lingüística teórica, lingüística aplicada i aplicacions de la lingüística", *Límits, Revista d'Assaig i d'Informació sobre les Ciències del llenguatge*, Barcelona, 5, 33-60.

LLISTERRI, J. (1997): "Etiquetado, transcripción y codificación de corpus orales", *Seminario de Industrias de la Lengua*, Curso "Etiquetación y extracción de información de grandes corpus textuales", Fundación Duques de Soria, Soria, 15 de julio de 1997. Disponible electrónicamente en: http://liceu.uab.es/~joaquim/publicacions/FDS97.html

LLISTERRI, J. (2007): "El español y las nuevas tecnologías", en M. LACORTE (coord.), *Lingüística aplicada del español*, Madrid: Arco/Libros, 483-520.

LLISTERRI, J. y MARTÍ, M. A. (2002): "Las tecnologías lingüísticas en la Sociedad de la Información", en M. A. MARTÍ y J. LLISTERRI (eds.), *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita*, Soria: Fundación Duques de Soria-Barcelona: Edicions de la Universitat de Barcelona, 13-28.

LYONS, J. (1977): *Semantics*, 2 vols., Cambridge: Cambridge University Press

MARCOS MARÍN, F. A. (1994): *Informática y humanidades*, Madrid: Gredos.

MARTÍN DE SANTA OLALLA SÁNCHEZ, A. (1999): *Una propuesta de codificación morfosintáctica para corpus de referencia en lengua española*, *Estudios de Lingüística Española*, vol. 3. Disponible en formato electrónico: http://elies.rediris.es/elies3/index.htm

MCCARTHY, J. (2002): "What is Artificial Intelligence?". Notas introductorias al campo de la IA. Publicación electrónica en: http://www-formal.stanford.edu/jmc/whatisai/whatisai.html

MCCORDUCK, P. (1993): "Inteligencia artificial: un aperçu", en S. R. GRAUBARD (comp.), *El nuevo debate sobre la inteligencia artificial. Sistemas simbólicos y redes neuronales*, Barcelona: Gedisa, 81-101.

MEYA LLOPART, M. y HUBER, W. (1986): *Lingüística computacional*, Barcelona: Teide.

OAKES, M. P. (1998): *Statistics for Corpus Linguistics*, Edinburgh: Edinburgh University Press.

OBLER, L. K. y GJERLOW, K. (1999): *El lenguaje y el cerebro*, Cambridge. CUP.

OLOHAN, M. (2004): *Introducing Corpora in Translation Studies*, London / New York: Routledge.

PAYRATÓ, L. *et al.* (eds.) (1996): *Corpus, corpora*, Barcelona: PPU.

PÉREZ GUERRA, J. (1998): *Análisis computarizado de textos. Una introducción a TACT*, Universidade de Vigo: Servicio de Publicacións.

PYLYSHYN, Z. W. (1988): *Computación y conocimiento. Hacia una fundamentación de la ciencia cognitiva*, Madrid: Debate.

QUIRK, R. *et al.* (1985): *A Comprehensive Grammar of the English Language*, London: Longman.

RAFEL I FONTANALS, J. y SOLER I BOU, J. (2003): "El procesamiento de corpus", en M. A. MARTÍ ANTONÍN (coord.), *Tecnologías del lenguaje*, Barcelona: UOC, págs. 41-73.

ROJO, G. (1992): "El futuro *Diccionario de construcciones verbales del español actual*" en C. MARTÍN VIDE (ed.), *Actas del VIII Congreso de Lenguajes naturales y lenguajes formales*, Barcelona: Univ. de Barcelona, 41-50.

ROJO, G. (1993): "La base de datos sintácticos del español actual", *Español actual: Revista de español vivo*, 59, 15-20.

ROJO, G. (1994): "Problemas lingüísticos e informáticos en los diccionarios de construcción y régimen", en *Actas del Congreso de la Lengua Española (Sevilla, 7 al 10 de octubre de 1992)*, Madrid: Instituto Cervantes, 1994, 307-315.

ROJO, G. (2001): "La explotación de la Base de datos sintácticos del español actual (BDS)", en J. DE KOCK (ed.), *Lingüística con corpus. Catorce aplicaciones sobre el español*, Salamanca: Ediciones Universidad de Salamanca, I, 7, 255-286.

ROJO, G. (2003): "La frecuencia de los esquemas sintácticos clausales en español", en VV.AA. (ed.), *Lengua, variación y contexto: estudios dedicados a Humberto López Morales*, Arco/Libros, 413-424.

ROJO, G. y SÁNCHEZ, M. (2010): *El español en la red*, Madrid: Fundación Telefónica/Barcelona: Ariel.

RUIZ ANTÓN, J. C. (2005): "Lenguaje e informática / Lenguaje y ordenadores", en A. LÓPEZ y B. GALLARDO (eds.), *Conocimiento y lenguaje*, València: Universitat de València, 401-436.

RUPPENHOFER, J. *et al.* (2006): *FrameNet II: Extended Theory and Practice*. Publicación electrónica disponible en http://framenet.icsi.berkeley.edu/book/book.html

SAGER, J. C. (1993): *Language Engineering and Translation. Consequences of automation*, Amsterdam/Philadelphia: John Benjamins.

SÁNCHEZ, A. (1995): *Cumbre. Corpus Lingüístico del Español Contemporáneo: fundamentos, metodología y aplicaciones*, Madrid: SGEL:

SANTANA SUÁREZ, O. y PÉREZ AGUIAR, J. R. (2000a): "Morfología computacional en español". Seminario disponible en: http://protos.dis.ulpgc.es/docencia/seminarios/mc/Morfologia/index.htm

SANTANA SUÁREZ, O. y PÉREZ AGUIAR, J. R. (2000b): "El procesador morfológico en Internet", presentación disponible en: http://protos.dis.ulpgc.es/docencia/seminarios/mc/Procesador_en_Internet/index.htm

SAVITCH, W. J. *et al.* (1987): *The Formal Complexity of Natural Language*, Dordrecht, Holland: Reidel.

SERRANO, S. (1975): *Elementos de lingüística matemática*, Barcelona: Anagrama.

SIL: "Computational Morphology and Phonology". Publicación electrónica en: http://www.sil.org/computing/comp-morph-phon.html

SINCLAIR, J. (2005): "Corpus and Text - Basic Principles", en M. WYNNE (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford: Oxbow Books.

STEVENSON, M. y WILKS, Y. (2003): "Word-sense disambiguation", en R. MITKOV (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford: Oxford University Press, 249-265.

TROST, H.: "Computational Morphology". Publicación electrónica en: http://www.ai.univie.ac.at/~harald/handbook.html

USZKOREIT, H.: *Language Technology. A First Overview*. Publicación electrónica en: http://www.dfki.de/~hansu/LT.pdf

VV.AA. (1990): *La Lingüística Aplicada: Noves perspectives - Noves professions - Noves orientacions*, Barcelona: Publicacions Universitat de Barcelona / Fundació Caixa de Pensions.

WERNER, R. (1982): "Léxico y teoría general del lenguaje", en G. HAENSCH *et al.*, *La lexicografía. De la lingüística teórica a la lexicografía práctica*, Madrid, Gredos, 21-94.

WINSTON, P. H. (1993): *Artificial Intelligence*, Reading, Mass.: Addison-Wesley, 3ª ed.

ZAMPOLLI, A.; CALZOLARI, N., PALMER, M. (eds.) (1994): *Current issues in Computational Linguistics: in honour of Don Walker*, Pisa: Giardini / Dordrecht, The Netherlands: Kluwer Academic Publishers.