Un corpus para el asturiano.
Las tecnologías lingüísticas en la consolidación de las lenguas minorizadas

Roser Saurí Colomer
Computer Science Department
Brandeis University
roser@cs.brandeis.edu

> "*The old quip attributed to Uriel Weinreich, that a language is a dialect with an army and a navy, is being replaced in these progressive days: a language is a dialect with a dictionary, grammar, parser and a multi-million-word corpus of texts-and they'd better all be computer tractable.*"
>
> Nicholas Ostler
> Foundation for Endangered Languages[1]

1 Introducción

En tanto que transmisoras de una visión del mundo y de la historia, el poder de las lenguas como aglutinantes sociales y vehiculadoras de una identidad particular es tan fuerte que las convierte muchas veces en un problema para los gobiernos de imperios o entidades políticas multilingües, los cuales con frecuencia toman medidas para minimizar la diversidad: por ejemplo, relegando la comunidad lingüística y culturalmente discrepante a un rincón ignoto de la realidad que se construye como país; o bien creando una tendencia a la marginación social de los individuos y los pequeños grupos más marcadamente disidentes.

En el primero de los casos, la lengua minorizada ha tenido la posibilidad de sobrevivir hasta nuestros días, apartada a un mundo explícitamente ignorado por la parte dominante y que precisamente por esto ha podido mantener su estructura social. Es por ejemplo el caso de muchas de las lenguas indígenas del continente americano. En el segundo de los casos, irónicamente la lengua pasa de ser un problema para el poder dominante a ser un problema para sus hablantes: la continuada penalización que reciben por usar su lengua, desde sus primeros años de escolarización hasta los ámbitos profesionales y sociales de la vida adulta, hace que ésta deje de ser su medio de comunicación natural y se convierta en una suerte de desgracia de la que avergonzarse, cual una de esas taras familiares que le tocan a uno sin haberlo pedido. A diferencia del

[1] Con motivo de la reseña del *Workshop on Language Resources for European Minority Languages*, que tuvo lugar en la *International Conference on Language Resources and Evaluation,* LREC 1998.

caso anterior, esta situación es generalmente propiciada cuando hay un nivel de contacto mayor entre las dos culturas.

La vitalidad de una lengua se pone pues en peligro cuando se convierte en un obstáculo para sus hablantes. Esta es lamentablemente la situación en la que se encuentran hoy en día muchas de las lenguas que se hablan en el mundo, incluidas aquellas que hasta ahora habían sobrevivido olvidadas en sociedades consideradas arcaicas y sin capacidad de futuro. Esto es debido al contexto globalizador en que nos encontramos, el cual crea un nivel adicional de diglosia donde ya existía uno previo porque, entre muchas otras cosas, impone un proceso de estandarización cultural que pasa también por la creación de una norma de uso lingüístico a favor de las lenguas consideradas de más prestigio; esencialmente, aquellas pocas con capacidad de trascendencia a los sistemas de información y con presencia en el desarrollo tecnológico occidental.

Sin embargo, los mismos instrumentos a través de los cuales se establece este proceso unificador nos abren la posibilidad de garantizar la supervivencia de las lenguas minorizadas. La popularización de las tecnologías de la información y de la comunicación telemática ha dado paso al marco actual de la llamada Sociedad de la Información. Del mismo modo pues, el acceso a los medios de comunicación de masas por un lado, y a las tecnologías para el procesamiento automático del lenguaje humano, por el otro, permitirán en gran medida la normalización social de las lenguas minorizadas, en tanto que las autorizarán como lenguas adecuadas para situaciones comunicativas prestigiadas. El desarrollo de herramientas y recursos para el procesamiento del lenguaje es por tanto un arma para la supervivencia y la relevancia de las comunidades lingüísticas minorizadas.

El presente artículo se centra en la relación potencial de la lengua asturiana con el segundo de estos instrumentos: las tecnologías lingüísticas. El punto de partida que me ha llevado a escribirlo es la convicción de que el asturiano, a diferencia (lamentablemente) de muchas de las lenguas del mundo con un número reducido de hablantes y en situación de subordinación respecto a otra, está aún a tiempo de entrar también en el dominio de las tecnologías lingüísticas. Esto es así debido, por un lado, a su contextualización dentro del marco occidental, lo cual le garantiza el nivel de acceso tecnológico necesario y la disponibilidad mínima de recursos, tanto materiales como humanos; por el otro, a la reciente apertura de un contexto político mínimamente sensible a la normalización de la lengua asturiana, como resultado de las pasadas elecciones del 25 de mayo.

Otras lenguas minorizadas, geográfica y políticamente allegadas al asturiano, han entrado ya en este terreno: el gallego, el euskera y el catalán. Las condiciones de partida de éstas fueron relativamente más propicias que las de la lengua asturiana, debido a su estatus de co-oficialidad dentro del marco político en que se encuentra la mayor parte de su territorio de habla. Sin embargo, la experiencia en todas ellas proporciona un buen punto de referencia para la entrada del asturiano a las tecnologías del lenguaje. Siguiendo estos trabajos previos así como la experiencia adquirida en otras comunidades lingüísticas, vemos que la piedra angular para la entrada de una lengua a las tecnologías del lenguaje es la construcción de un corpus lingüístico. En este artículo planteo pues la posibilidad de un proyecto de corpus para la lengua asturiana.

Naturalmente, la entrada de una lengua en las tecnologías lingüísticas no es la bula papal que le asegura la salvación sin más. El indudable interés que este campo tiene para

las lenguas minorizadas es consecuencia de las posibilidades de desarrollo tanto de investigación como de útiles concretos de base lingüística que permite: desde diccionarios en su tradicional formato de libro para un público general hasta programas de traducción automática. Por esto, la sección 2 introduce al lector en el campo de la Lingüística de Corpus y su objeto de estudio, y analiza las posibilidades que ofrece desde una perspectiva puramente científica. La sección 3 hace un repaso de experiencias previas en proyectos de corpus que pueden servir de referencia: por un lado, las pioneras desarrolladas para el inglés; por el otro, las llevadas a cabo en el ámbito geográfico y político del asturiano. En ambos casos se trata de secciones totalmente prescindibles para los lectores que ya están familiarizados con el campo. Sin embargo he considerado que la vocación divulgadora de mi artículo obligaba a entrar mínimamente en estos aspectos básicos. Finalmente, ahondaré en la relación entre lenguas minoritarias y tecnología lingüística que se ha apuntado más arriba (sección 4.1), y acabaré analizando los elementos que hay que tomar en cuenta para la entrada de la lengua asturiana en el esta área, la cual tiene que pasar necesariamente por la construcción de un corpus lingüístico (sección 4.2).

2 Corpus lingüísticos

2.1 De qué hablamos cuando hablamos de corpus

Un *corpus lingüístico* es una colección de textos orales o escritos de una lengua, que han sido seleccionados a partir de unos criterios lingüísticos explícitos, pasados a soporte electrónico y mínimamente procesados, y que se utilizan como muestra representativa de esta para el abastecimiento de datos destinados a su estudio sistemático.[2] Dentro de varias de las áreas dedicadas al estudio del lenguaje, el uso de corpus como recopilaciones de datos para la tarea científica no es una práctica nueva. Por ejemplo, siguiendo el sentido más allegado a su étimo latino, en el terreno de los estudios literarios se llama corpus a la totalidad de la obra de un determinado autor. Igualmente, se conoce como corpus al conjunto de contextos recolectados por lexicógrafos (muchas veces provinentes de la obra de los clásicos), para la elaboración y ejemplificación de las definiciones de los vocablos de la lengua en cuestión.

Estos usos de corpus difieren sin embargo de lo que actualmente se reconoce como el objeto de desarrollo de la *Lingüística de Corpus*. Un corpus lingüístico se diferencia de un corpus lexicográfico tradicional por el hecho de que este último suele recoger únicamente muestras fragmentadas de la lengua, en algunos casos hasta muy infrecuentes y anecdóticas, mientras que el otro aspira a una mayor *representatividad*. Asimismo, los corpus lingüísticos se diferencian de otros corpus de datos textuales (como los constituidos por la obra de un autor) por su finalidad: los primeros se crean con el propósito específico de servir de base para análisis lingüísticos, y por tanto se diseñan a partir de *criterios lingüísticos explícitos*; mientras que los segundos se elaboran a partir

[2] Esta definición pretende consensuar las distintas caracterizaciones de corpus lingüístico en la bibliografía. Véase por ejemplo Sinclair (1996), Tognini-Bonelli (1996), Torruella y Llisterri (1999) o McEnery y Wilson (2001).

de criterios externos a la lengua de los textos. Finalmente, los corpus lingüísticos se caracterizan por el tratamiento al que es sometida su información con el fin de facilitar la tarea de análisis. Así, independientemente del soporte original de los textos que lo componen –oralidad o escritura— los corpus lingüísticos son *informatizados* y, como se verá en la sección 2.4, en la mayoría de los casos también procesados mínimamente para facilitar el acceso y la recuperación de la información.

El papel que juega la digitalización de la información en los corpus textuales no es menor. Tanto es así que algunos autores se refieren a los corpus lingüísticos actuales como *corpus informatizados*. Esto tiene su justificación en el contexto histórico del área. La utilización de colecciones de datos lingüísticos para el análisis sistemático de una lengua empezó a tener reconocimiento dentro de la tradición de la lingüística de campo a principios del S. XX, así como en la escuela estructuralista posterior. Sin embargo, esta práctica se vio desprestigiada con la llegada en escena de los planteamientos de base mentalista defendidos por Chomsky desde finales de los 50, y que marcan un giro en los estudios lingüísticos contemporáneos. El objetivo de este nuevo paradigma es el estudio de la gramática universal subyacente a la competencia de los hablantes, y dentro de este planteamiento, la actuación se contempla únicamente como el reflejo imperfecto del conocimiento lingüístico de los hablantes. En tanto que reproducen precisamente este nivel de actuación, los conjuntos de datos lingüísticos compilados en la tradición estructuralista y de la lingüística de campo dejaron de interesar.

Pero a pesar de su impopularidad, los estudios de carácter más empírico no se abandonan por completo. No sólo esto, sino que durante las décadas de los 60 y 70 empezaron a incorporar, como elemento indisociable a este tipo de planteamiento, la utilización de recursos tecnológicos. Así, mientras que en los trabajos previos de los estructuralistas y los lingüistas de campo el soporte material de la información era el papel, ahora el medio de almacenaje pasa a ser electrónico y la manipulación de los datos puede beneficiarse de procesos automatizados. Paralelamente, las aproximaciones de base empirista vuelven gradualmente a ganar adeptos, debido a razones tanto de carácter metodológico (por ejemplo, áreas como la adquisición del lenguaje no pueden basarse en intuiciones de lingüista), como de formulación teórica de base sobre la naturaleza del lenguaje natural (una buena introducción a todo este proceso es McEnery y Wilson 2001). De este modo, en los años 80 el campo conocido actualmente como la Lingüística de Corpus hace su eclosión, gracias a la convergencia de esa perspectiva adoptada en algunos ámbitos de la lingüística con las posibilidades de tratamiento de datos que permite el desarrollo tecnológico del momento.

2.2 Seleccionando la información: clases de corpus

En tanto que los corpus lingüísticos tienen como finalidad proporcionar información para el estudio de una o más lenguas, se pretende que constituyan fragmentos representativos de éstas. La representatividad de un corpus es uno de los aspectos que más se ha discutido en la bibliografía del área, alrededor de consideraciones como: qué tamaño debe tener un corpus para que sea realmente representativo; qué clase de textos debe contener; en qué cantidad, etc. Véase a modo de ejemplo Quirk (1992), Biber (1993) o Sinclair (1996).

Durante el auge de la lingüística de corpus, en las décadas de los 80 y 90, la representatividad se planteó esencialmente en términos de *tamaño* (a más información, mayor representación de la lengua), y de *equilibrio* entre los distintos usos de la lengua (variantes dialectales, oralidad versus escritura, diversidad de registros y géneros, etc.). Actualmente, el tamaño sigue siendo valorado como un elemento importante para poder capturar el mayor número de fenómenos posibles de una lengua. Sin embargo, la búsqueda de una representación equilibrada de todas las variedades de uso ha ido quedando relegada únicamente al criterio para la construcción de los llamados corpus generales (aquellos que pretenden reflejar la lengua común en la totalidad de sus variedades y ámbitos). Así, la idea de que las aplicaciones finales determinan diseños diferentes de corpus ha ido cuajando de modo natural, invalidando los planteamientos que imponían una serie de criterios estándares para la obtención de la representatividad.

Veremos las aplicaciones potenciales de un corpus lingüístico en la sección siguiente. Por el momento me limito aquí a introducir las posibles clases de corpus, las cuales se establecen en función de distintos parámetros complementarios entre ellos:[3]

- **Soporte original de los datos:** tenemos *corpus escritos* o *corpus orales*, según si están constituidos por textos escritos u orales.
- **Lenguas que lo constituyen:** se puede diferenciar entre *corpus monolingües* y *corpus multilingües*. Estos últimos se denominan *corpus paralelos* cuando están constituidos por los mismos textos en cada una de las lenguas representadas.
- **Variación lingüística que se pretende representar:** diatópica, diastrática y/o diacrónica. Algunos corpus representan únicamente una de las variantes de la lengua. Al otro extremo estan los *corpus generales*, que pretenden representar a varias de ellas, generalmente delimitados cronológicamente. Por supuesto, en este tipo de corpus es necesario asegurar el equilibrio de representación entre las distintas variantes.
- **Nivel de especialización de los textos:** los *corpus especializados* se concentran únicamente en textos de determinadas áreas de especialidad. Su finalidad suele ser muy específica, generalmente dentro del ámbito de la terminiología y la terminografía.
- **Citerios de selección de los datos:** se distingue entre corpus constituidos a partir de textos enteros y corpus que contienen sólo fragmentos (generalmente de longitud constante), los cuales se conocen como *corpus de muestras*. Actualmente esta distinción está volviéndose obsoleta debido a los problemas que plantean los corpus de muestras para la recuperación del contexto global de uso de las expresiones lingüísticas.
- **Criterios de actualización de los datos:** en algunos casos interesa la renovación periódica de los materiales de un corpus. Esto es especialmente así en corpus que aspiran a representar la lengua actual, en los cuales el fragmento de textos más antiguos es regularmente subsituido por otro de materiales recientes. Este tipo de corpus se conoce como *corpus monitor*.

[3] Mi caracterización es mínima y únicamente destinada a cubrir de modo básico la laguna de que pueda adolecer un lector no familiarizado con el tema. Para una clasificación más exhaustiva y detallada, véase Torruella y Llisterri (1999).

2.3 Utilizando la información: aplicaciones posibles de un corpus

Las utilidades que ofrecen los corpus lingüísticos benefician a distintos campos y disciplinas relacionados con el estudio del lenguaje. Se pueden agrupar en tres grandes niveles:

Desde las ciencias del lenguaje, un corpus sirve de base para estudios de plantemiento teórico. Por ejemplo, como abastecedor de contextos para la validación de modelos de descripción de la lengua, para el análisis de la relación entre dos lenguas (a partir de corpus paralelos), o bien como indicador de la variación y las tendencias de uso, de interés en terrenos como la dialectología y la sociolingüística y que en el caso de las lenguas minorizadas puede derivar en determinadas decisiones de planificación lingüística. Igualmente, un corpus se puede utilizar para el desarrollo de productos de base lingüística. Por ejemplo, para la elaboración o mejora de gramáticas y diccionarios (selección del léxico más frecuente, uso de ejemplos reales, distinción de sentidos en base a los datos y no a la intuición del lexicógrafo, etc.); para la creación de vocabularios especializados; o para la construcción de útiles de apoyo en el aprendizaje de una lengua extranjera.

A un segundo nivel, los corpus se utilizan también para el desarrollo de tecnología lingüística. Por ejemplo, los corpus paralelos se utilizan para entrenar sistemas de traducción automática, los corpus orales se usan en el ámbito de reconocimiento y síntesis de habla, y los corpus textuales, para el entrenamiento de correctores ortográficos y gramaticales.

Finalmente, en el campo del procesamiento del lenguaje natural los corpus están siendo también utilizados a modo retro-alimentativo, para la mejora de las herramientas de procesamiento de corpus (muy particularmente, para el entrenamiento de las de base estadística), tales como analizadores morfológicos y sintácticos, o etiquetadores semánticos. Vemos a continuación cuáles son estas herramientas y para qué sirven.

2.4 Tratando la información: corpus anotados

Los corpus lingüísticos no terminan sin embargo siendo solamente esto: un conjunto de textos caracterizados a partir de determinados criterios, que se almacenan en soporte electrónico, y que son utilizados con una finalidad específica. Precisamente debido a su finalidad como fuente de abasteciemiento de datos, es necesario que la información que contiene sea fácilmente asequible, y para esto la digitalización de textos no es suficiente. Así, en la mayoría de casos los corpus lingüísticos presentan, además de su contenido textual original, un nivel metalingüístico de información que explicita las características gramaticales de los distintos niveles de constituyentes: desde rasgos morfológicos de los elementos léxicos hasta, potencialmente, características de los párrafos o unidades textuales mayores, en el caso de corpus escritos; en los corpus orales, desde información fonética hasta estructura prosódica. Debido a mi experiencia personal dentro de la Lingüística de Corpus, en esta sección y lo que sigue del artículo me concentraré casi exclusivamente en la parte de corpus escritos. El lector sin embargo encontrará suficientes referencias bibliográficas para ampliar sus conocimientos en lo que se refiere a corpus orales. Véase por ejemplo Llisterri (1997, 1999).

2.4.1 Niveles de anotación

Los corpus lingüísticos que incorporan este nivel de metainformación suelen referirse como *corpus etiquetados*, dado que la información se codifica en el texto utilizando códigos específicos o conjuntos de etiquetas (también conocidos como *etiquetarios*). De modo equivalente pueden ser también referidos como *corpus anotados*.[4] Por un lado, el etiquetaje de un corpus facilita la tarea de manejo y recuperación de información. Por ejemplo, para un lexicógrafo o un gramático que quiera analizar los contextos de uso de un verbo determinado, disponer de todas las formas de este verbo identificadas por su relación con un único lema evita tener que buscar los contextos de uso a partir de la enumeración exhaustiva de todas las formas posibles. Por otro lado, el etiquetaje permite una cuantificación de datos más sofisticada que la pura frecuencia de uso de cada forma léxica (por ejemplo, distribución de tiempos verbales, clases de complementos para determinados predicados, tipos de modificación aplicada a una clase de nombres particular, coocurrencias frecuentes de elementos léxicos, variantes dialectales, etc.). De este modo se pueden observar tendencias de comportamineto lingüístico dentro del fragmento de lengua analizado.

El etiquetaje de un corpus suele informar sobre distintos niveles: el *etiquetaje estructural*, por ejemplo, marca la organización estructural del texto: título, subtítulo, capítulo, párrafo, sección, subsección, etc. --hasta frase, generalmente. El *etiquetaje morfológico* se aplica al nivel de los elementos léxicos, indicando su categoría gramatical y, en lenguas como las romances, también la información de flexión verbal y nominal. Generalmente sobre este nivel se aplica el *etiquetaje sintáctico*, que se puede realizar con más o menos profundidad, dependiendo de la aplicación para la que se plantea el corpus. Así, la anotación puede ser superficial, de simples grupos nominales y/o verbales, como más compleja, indicando las funciones sintácticas de los distintos componentes de una frase (sujeto, objeto, modificadores, etc.).[5]

Éstos son a grandes rasgos los niveles de anotación más comunes. Sin embargo, otros niveles de información pueden ser igualmente deseables (además o alternativamente a los que acabo de presentar) en función de la aplicación específica para la que se quiera el corpus. Por ejemplo, se puede introducir también anotación de carácter semántico o pragmático. Pero mientras que en los niveles anteriores se puede trabajar desde un planteamiento bastante neutro respecto a cualquier teoría lingüística, a partir del nivel semántico el etiquetado tiende o bien a basarse en a una visión particular de la semántica o la pragmática, o bien se especializa en un aspecto concreto, como las relaciones

[4] En los últimos años, el uso de corpus no etiquetados se ha convertido también en una práctica relativamente habitual, por la economía de recursos y tiempo que conlleva. Sin embargo, en la mayoría de casos se trata de corpus utilizados para el entrenamiento de herramientas estadísticas de procesamiento del lenguaje (veremos más adelante a qué me refiero), en lugar de corpus pensados para el desarrollo de útiles de base lingüística (diccionarios, gramáticas) o para la descripción de la lengua. El presente trabajo no va a ahondar en esa perspectiva.

[5] Dado que mucha de la bibliografía del campo está en inglés, considero importante introducir aquí la terminología de los distintos niveles de etiquetaje también en esta lengua. A la fase de anotación morfológica se la conoce como *part-of-speech (POS) tagging*. Al nivel de etiquetaje sintáctico, como *parsing*, identificando también el marcaje sintáctico superficial como *shallow parsing* o bien *chunking*.

anafóricas, las unidades de información temporal, los marcadores discursivos, los papeles temáticos de los predicados verbales, etc.[6]

2.4.2 Herramientas de procesamiento de corpus

Todos estos niveles de etiquetaje se realizan a través de varias capas de procesamiento automatizado. Por ejemplo, para la anotación morfológica se utilizan los programas conocidos como *analizadores morfológicos*, que trabajan con la colaboración de un *diccionario* en formato electrónico donde cada elemento léxico va asociado a una o más categorías gramaticales y, en lenguas como las romances, de un *lematizador*, el cual se encarga de descomponer las palabras en su raíz y terminaciones. Para la tarea de etiquetaje sintáctico se usan los llamados *analizadores sintácticos*, que como ya se comentó, pueden trabajar a un nivel más o menos profundo de análisis: desde la simple detección de grupos nominales y verbales, hasta el marcaje de funciones y relaciones de dependencia. Para la anotación del nivel semántico (así como también el pragmático), no hay herramientas específicas que permitan la agrupación bajo un único denominador, como sucede con los niveles previos. Sin embargo merece la pena comentar la existencia de los *etiquetadores semánticos*, sobre los que se ha trabajado bastante últimamente. Se trata de herramientas que asignan etiquetas semánticas a las unidades léxicas a partir de su clasificación en una ontología determinada, lo que naturalmente requiere una tarea previa de desambiguación léxica.

Finalmente, es importante mencionar también las *herramientas para la explotación de corpus*, que son básicas en la investigación lexicográfica y lingüística. Las más destacadas aquí son las que realizan búsquedas de (secuencias de) palabras, presentando los resultados ordenados alfabéticamente y con el contexto oracional, y ofreciendo además la posibilidad de cómputo de frecuencias.

Las herramientas de procesamiento de corpus se dividen en general según el planteamiento de base de que partan: por un lado, están las herramientas de base simbólica; por otro, las de base estadística. Las primeras se desarrollan a partir de conocimiento lingüístico, y están pues en contacto con algunas de las vertientes de la lingüística teórica. Las segundas se fundamentan en modelizaciones estocásticas del fragmento de lengua con que se trabaja. El conocimiento lingüístico utilizado en este segundo caso es mínimo, si no inexistente, y por consiguiente no permite, una vez desarrollada la aplicación en cuestión, hacer una abstracción en términos lingüísticos de los fenómenos tratados. Sin embargo, suele dar resultados más óptimos que el de una herramienta de base simbólica desarrollada durante un período de tiempo equivalente.

Independientemente de la aproximación simbólica o estadística al problema, el campo dedicado al desarrollo de todas estas herramientas para el tratamiento de textos escritos se conoce como *Procesamiento del Lenguaje Natural* (PLN) o, en determinados contextos

---

[6] Ejemplos particulares de corpus en este grupo son: el trabajo presentado en Fligelstone (1992), que se caracteriza por la anotación de relaciones anafóricas; el *RST Corpus* (Carlson *et al*., 2003), en el que los textos se anotan a partir de la teoría de discurso *Rethorical Structure Theory* (Mann y Thompson, 1988); el *TimeBank* (Pustejovsky *et al*. 2003), en el cual sólo se ha anotado las eventividades y las expresiones temporales con la finalidad de servir de base para estudios de razonamiento temporal; o bien también el Corpus de Operaciónes Metalingüísticas Explícitas (Rodríguez, 2003), en el cual se han anotado las operaciones de creación de conocimiento en el marco del discurso científico.

también, *Ingeniería Lingüística*; mientras que en lo que se refiere al procesamiento de textos orales hablamos de *Tratamiento del Habla*. En el artículo presente utilizaré el término genérico de *tecnologías del lenguaje* (o su equivalente, *tecnologías lingüísticas*), para referirme de modo global a esta dos áreas conjuntamente con la Lingüística de Corpus. Tal y como indica el nombre, su denominador común es el uso de tecnología para el tratamiento del lenguaje.

3 Experiencias de referencia

Una vez establecidos los preliminares, esta sección presenta a grandes rasgos las experiencias previas de proyectos de corpus que considero de más interés para lo que aquí nos ocupa: la construcción de un corpus lingüístico para la lengua asturiana.

Las primeras experiencias que abren la nueva era en la Lingüística de Corpus se dan para la lengua inglesa. Actualmente constituyen obras de referencia y es por eso que, a pesar de que están ampliamente documentadas en la bibliografía del área, se hace relevante presentarlas mínimamente (sección 3.1). Su naturaleza de hito es obviamente consecuencia del carácter pionero de la investigación en algunas de las comunidades donde se habla esta lengua. Existen por supuesto corpus en otras lenguas; tanto lenguas cercanas geográficamente, como el francés o el alemán, como más alejadas: desde el japonés, el coreano y el mandarín, hasta las recientemente tan preciadas lenguas habladas en el mundo musulmán, como el árabe y el farsi.[7] Considero sin embargo que queda fuera del alcance de este trabajo hacer un repaso, ni que sea mínimo, de los casos más relevantes para cada lengua.

Por otro lado, revisaré también los proyectos existentes dentro del marco geográfico de la de la Península Ibérica, centrándome exclusivamente en aquellos que se han desarrollado al amparo del contexto político en el que se incluye el asturiano; esto es, el Estado español (sección 3.2).

3.1 Los pioneros

Debido a la misma evolución del campo, los corpus para el inglés pueden clasificarse en dos generaciones. La primera incluye los corpus empezados en las décadas de los 50 y 60, en un momento en que las aproximaciones empíricas al lenguaje dejaron de interesar, particularmente dentro de la lingüística desarrollada en el contexto de habla inglesa. Los corpus de este período inicial se caracterizan por ser de tamaño modesto –comparado con algunos de los corpus construidos más recientemente. Los más destacables son:

- *Survey of English Usage*:

[7] En este sentido, vemos que sigue siendo parte de la comunidad de habla inglesa la que decide promocionar recursos en una u otra lengua. Es indicativo la creciente demanda de lingüistas, traductores e intérpretes en algunas de estas últimas lenguas que se ha generado a partir del 11 de septiembre del 2001. El impulso de la investigación en algunas lenguas se vea propulsado por intereses de la comunidad de la las agencias de espionaje e inteligencia militar, y con objetivos que poco o nada tienen que ver con la investigación puramente científica o el aprecio de valores culturales.

Corpus de inglés británico iniciado en 1955 y realizado a lo largo de aproximadamente 30 años, en el University College London. Contiene un millón de palabras, distribuidas entre textos orales y textos escritos. Interesa remarcar aquí que los materiales de este corpus se utilizaron como base para la constitución de la conocida gramática de referencia del inglés (Quirk *et al.*, 1985).

- *Brown Corpus*, *Lancaster-Oslo/Bergen Corpus* (LOB) y *Kolhapur Corpus*:
  Corpus de textos escritos en las variantes de inglés americano, británico e inglés de la India, respectivamente. El primero se compiló en la década de los 60 y los otros dos en etapas posteriores, como corpus equivalentes al anterior para otras variantes del inglés. Contienen todos un millón de palabras y presentan los textos agrupados en 15 categorías diferentes, procurando así una buena representatividad de los distintos usos de lenguaje escrito. Son pues corpus con vocación de referencia (Francis y Kucera, 1964).

- *London-Lund Corpus*:
  Corpus oral de inglés británico, de 500.000 palabras. Contiene exclusivamente transcripciones de material hablado, originario de dos proyectos diferentes: de la parte oral del *Survey of English Usage* introducido más arriba, y del *Survey of Spoken English*, empezado en la Lund University en 1975 como proyecto hermano del anterior (Svartvik, 1990).

Actualmente la dimensión de los corpus ha variado ostensiblemente debido a los avances tecnológicos. La capacidad de almacenaje es considerablemente mayor de lo era en los inicios de la tecnología digital, y los procesos de tratamiento de datos son mucho más rápidos. A casi 50 años de la creación del primer corpus moderno, los corpus que hay disponibles para el inglés son muchos y de naturaleza variada. De entre todos interesa mencionar los siguientes por su carácter de referencia dentro del área y su trascendencia en el campo de la lingüística aplicada:

- *Bank of English*:
  Iniciado el 1991 por parte de la Birmingham University y COBUILD, una división de la editorial HarperCollins. Su principal objetivo es servir como fuente de datos para estudios lingüísticos y la compilación de diccionarios. Se trata de un corpus monitor; esto es, un corpus al que se le añade material nuevo periódicamente. En su última edición, de 2002, cuenta con 450 millones de palabras (Sinclair, 1987).[8]

- *British National Corpus* (BNC):
  Contiene un total de 100 millones de palabras del inglés moderno, tanto del registro escrito como del oral, el cual hasta incluye conversaciones coloquiales. Fue compilado por parte de un consorcio formado por editores británicos, instituciones académicas como la Oxford University y la Lancaster University. Tanto en este caso

[8] Véase también http://titania.cobuild.collins.co.uk/boe_info.html para una introducción a este recurso.

como en el anterior, su diseño aspira a la mayor representatividad posible (Aston y Burnard, 1998).[9]

3.2 La lingüística de corpus en nuestro ámbito geográfico

El desarrollo de corpus lingüísticos, así como de otras herramientas para el tratamiento del lenguaje natural en el ámbito del Estado español, reproduce la división entre lenguas oficiales o co-oficiales por un lado, y lenguas no reconocidas oficialmente por el otro. En mayor o menor grado, el primer grupo de lenguas (euskera, catalán, español y gallego) dispone en la actualidad de útiles y recursos de procesamiento del lenguaje, grupos de investigación más o menos consolidados dedicados a su desarrollo, y (no por lo último, lo menos importante) la voluntad institucional de fomentar el área. Sin embargo, la Lingüística de Corpus y el Procesamiento del Lenguaje Natural son campos aún vírgenes en el caso de las lenguas sin oficialidad como el asturiano.

El Procesamiento del Lenguaje Natural en nuestro ámbito se inicia a mediados de los 80, catalizado en torno a los proyectos de traducción automática que se llevan a cabo en Barcelona y Madrid, tanto desde empresas privadas como financiados públicamente por la Comisión Europea. El campo pasa entonces por una etapa de euforia paralela a la expectación que existe en el marco europeo más general, pero a principios de los 90 esos proyectos empiezan a abandonarse porque la pobreza de los resultados obtenidos hasta el momento, en comparación con los recursos invertidos, apunta a un fracaso en los planteamientos de base.[10]

Sin embargo, cuando estos proyectos se desmantelan, ya hay la infrastuctura creada, en términos de instituciones y recursos humanos para seguir la investigación en el área, la cual se reorienta ahora hacia la constitución de recursos lingüísticos esenciales (como corpus, diccionarios en formato electrónico y bases de datos léxicos), y hacia la construcción de herramientas básicas para el procesamiento del lenguaje. A esto hay que añadir el hecho de que simultaneamente al desarrollo de proyectos de traducción automática en la década de los 80, ya existe alguna institución trabajando en el área de la Lingüística de Corpus, como es el caso por ejemplo del proyecto del *Diccionari del Català Contemporani* con su *Corpus Textual Informatitzat* (detallado a continuación). Es pues a mediados de los 90 que la Lingüística de Corpus hace su eclosión en nuestro ámbito geográfico, cuajando tanto en el contexto español y en Cataluña, donde se habían empezados las primeras iniciativas en traducción automática, como también en el País Vasco y en Galicia, con el subsiguiente desarrollo de recursos para cada una de las lenguas habladas en estas zonas.

Ofrezco a continuación una muestra representativa, aunque no exhaustiva, de los corpus lingüísticos existentes actualmente para estas cuatro lenguas: euskera, catalán, español y gallego. Aunque ninguna de ellas presenta un panorama semejante al del inglés, tampoco se caracterizan por la situación desoladora en que se encuentran las lenguas sin reconocimiento oficial. Se constata asimismo una cierta disparidad en el nivel de desarrollo del area en cada una de las lenguas. Eso se corresponde en lineas generales con los datos reportados en la parte dedicada al fomento y desarrollo del área dentro del

[9] Véase también: http://www.natcorp.ox.ac.uk.

[10] Véase Abaitua (1999) para una detallada caracterización de este período.

plan nacional de investigación científica del gobierno vasco.[11] En base a un estudio reciente elaborado por parte del Consorcio Europeo EUROMAP, éste detalla que el reparto de esfuerzo de investigación y desarrollo destinada en cada una de las lenguas en cuestión en el marco conjunto del estado es el siguiente: 60 % para el español, 20% para el catalán, 8,5% para el gallego y finalmente 7% para el euskera. Seguiré este orden de lengua de más a menos recursos para la presentación de los proyectos de corpus en cada una de ellas. Sin embargo, me concentraré más en los proyectos desarrollados para las lenguas co-oficiales que en los del español: aunque en menor grado, comparten con el asturiano su situación minoritaria y minorizada, y por consiguiente su experiencia dentro de la Lingüística de Corpus resulta de gran utilidad a la hora de apuntar un proyecto en este terreno, con recursos y apoyo institucional limitado.

3.2.1 Corpus de español

Existen varios corpus para esta lengua. Dos de los más destacados son los corpus de referencia compilados por el Instituto de Lexicografía de la Real Academia Española, uno diacrónico y otro del español actual. Dos referencias que detallan el desarrollo de estos proyectos son Pino *et al*. (1999) y Sánchez-León *et al*. (1999). Ambos corpus incluyen documentos escritos en todas las variantes, tanto peninsulares como extrapeninsulares, y están eminentemente destinados a servir de recurso básico en el trabajo lexicográfico de esta institución:

- *Corpus Diacrónico del Español* (CORDE)
  Corpus que pretende representar la variación del español a lo largo de su historia escrita. Contiene documentos de tres épocas básicas: Edad Media, Siglos de Oro y Época Contemporánea (inicialmente hasta 1975). Actualmente presenta un total de 145 millones de palabras y se puede consultar a través de internet (http://www.rae.es/cordenet.html).

- *Corpus de Referencia del Español Actual* (CREA)
  Corpus monitor, constituido por textos representativos de las distintas variantes del español actual. Abarca un espacio de 25 años, el cual se actualiza periódicamente con materiales más recientes. Esto significa que los documentos más antiguos se traspasan al corpus CORDE. Actualmente contiene 145 millones de palabras.

Existen otros corpus del español desarrollados por parte de editoriales, como por ejemplo:

- *Corpus CUMBRE*
  Desarrollado por la Editorial SGEL para fines lexicográficos. Consta de 20 millones de palabras, provinentes de textos tanto escritos como orales del español peninsular e hispanoamericano. Los textos escritos pertenecen a distintos géneros (literario,

[11] *Plan Nacional de Investigación Científica, Desarrollo e Innovación (2000-2003).* 'Propuesta de acción estratégica en el área sectorial "Sociedad de la Información": Industria de la Lengua e Ingeniería Lingüística'. Eusko Jaurlaritza (Gobierno Vasco). Descargable en: http://www.euskadi.net/euskara/datos/azkeninforme.pdf (última actualización: 26/09/2003).

periodístico, ensayo), y tocan distintas áreas de especialidad (filosofía, historia, ciencia, derecho, economía, etc.). Los textos orales proceden de grabaciones de programas de radio y televisión. Mientras que las muestras orales son de la década de los 90, las escritas alcanzan el período entre los 50 y los 90 (Sánchez *et al*. 1995).

Igualmente, Vox Bibliograf trabaja con su corpus de alrededor de 10 millones de palabras, y la editorial SM con otro de 60.000 palabras.

Además de estos corpus, existen otros de menores dimensiones, ya sean de lengua general pero con fines específicos (los corpus de editoriales que se acaban de mencionar entrarían en esta categoría), o bien de lenguaje restringido, como el *Corpus de vocabulario del niño de 6 a 14 años* de la Universidad de Granada, o bien el *Corpus 92, Lengua escrita por aspirantes a estudios universitarios* de la Universitat Pompeu Fabra.

Finalmente, hay que mencionar también la existencia de corpus estrictamente orales, creados para el desarrollo de aplicaciones en el campo de las Tecnologías del Habla. Como ya anuncié al principio, no me voy a centrar en ellos.

3.2.2 Corpus de catalán

Los corpus de mayor envergadura se datallan a continuación. Información adicional sobre corpus textuales en esta lengua se puede encontrar en Soler i Bou (1998) y, a nivel más general, Llisterri (2000) ofrece una introducción bastante completa sobre los recursos y herramientas de procesamiento del lenguaje existentes actualmente.

- *Corpus Textual Informatitzat de la Llengua Catalana (CTILC)*
  Corpus de textos escritos de registro variado (literario y no literario) desarrollado en el Institut d'Estudis Catalans como componente básico para la creación del *Diccionari del Català Contemporani* (DCC). Se trata por tanto de un corpus de lengua general, que aspira a la representatividad y que ha estado diseñado para servir de base a la investigación lexicográfica. Lo componen alrededor de 4000 textos datados entre 1832 (fecha simbólica del inicio del período conocido como la *Renaixença* de la lengua catalana, con la publicación de la *Oda a la pàtria*, de Bonaventura Carles Aribau) y 1988. En total, recoge 52,3 millones de palabras. El corpus está lematizado y etiquetado morfosintácticamente (Rafel 1992-1993, 1996). Además, se complementa con una base de datos lexicográfica constituida a partir de 13 diccionarios publicados en el mismo período, y ha servido ya para la creación del *Diccionari de Freqüències* (Rafel i Fontanals, 1996-1998).

- *Corpus del Català Contemporani de la Universitat de Barcelona (CUB)*
  Corpus de textos escritos y orales, estructurado en siete subcorpus independientes que se relacionan entre ellos por el objetivo de configurar conjuntamente una caracterización del catalán actual desde su variación diatópica, diastrática y diafásica. Contiene: un subcorpus de variedades geográficas (*Corpus Oral Dialectal*); uno de variedades sociales (*Corpus Oral Social*); y los cinco restantes de variedades funcionales (*Corpus Oral de Conversa Col·loquial*, *Corpus Oral de Registres, Corpus Oral de Publicitat, Corpus d'Informatius Orals, Corpus Escrit de Català Actual*). Los subcorpus escritos están lematizados y anotados morfosintácticamente. Los subcorpus orales están transcritos y, según la naturaleza de los datos, presentan

anotación discursiva, de grupos tonales, o bien lematización y etiquetado morfosintáctico. Si en el caso del CTILC, la utilidad a la que se aspira es la investigación lexicográfica, en este caso se pretende crear una base para estudios específicos sobre la variación del catalán. Buenas introducciones al proyecto son Boix (1996) y Viaplana (2000).

- *Corpus textual plurilingüe especialitzat*
  Corpus elaborado en el Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra (Bach *et al*. 1997). Contiene textos escritos en catalán, español, francés, inglés y alemán, pertenecientes a las siguientes áreas de especialidad: medio ambiente, informática, medicina, derecho y economía. La parte del catalán cuenta con unos 5 millones de palabras, lematizadas y con etiquetado morfosintáctico.

Otros proyectos de corpus del catalán son: el *Corpus Parole*, de 21 millones de palabras desarrollado en el marco de un proyecto europeo, y el *Corpus de Diatopia Diacrònica de la Llengua Catalana*, impulsado desde la Universitat de Barcelona. Así mismo, existen algunos corpus de habla compilados específicamente para el desarrollo de aplicaciones de tecnologías del habla.

3.2.3 Corpus de gallego

Una panorámica general de la situación del gallego la ofrece García-Mateo *et al*. (1998). Aquí interesa destacar los corpus siguientes, todos ellos en estado de desarrollo:

- *Corpus De Referencia Do Galego Actual* (CORGA)
  Este corpus es uno de los proyectos llevados a cabo en el Centro Ramón Piñeiro para a Investigación en Humanidades, una institución creada por parte de la Xunta de Galicia para fomentar la enseñanza, la investigación y el uso de la lengua gallega. El CORGA se plantea como un corpus de referencia de textos escritos y orales, producidos entre 1975 y 2004 (fecha prevista para la finalización del proyecto), y pretende abarcar diferentes registros del gallego actual. Actualmente contiene unos 17,5 millones de palabras aunque se aspira a un total de 25 millones, anotadas con su categoría morfosintáctica. Los detalles del proyecto, así como la última versión de CORGA, emitida en enero del 2003, se puede consultar a través de internet.[12]

- *Corpus do galego moderno* (CGM)
  Corpus compilado en el Instituto da Lingua Galega. Contiene alrededor de 10 millones de palabras, provinentes de textos escritos desde el S.XVII hasta la actualidad. Se pretende incluir en él textos de la literatura oral, historias, refranes, canciones populares, etc. La finalidad principal del corpus es lexicográfica y para el estudio de la lengua gallega en general.

- *Arquivo do Galego Oral*
  Corpus oral, igualmente compilado por el Instituto da Lingua Galega. Se divide en dos subcorpus: el *Arquivo do galego popular* y el *Arquivo do galego culto*. El

[12] http://corpus.cirp.es/corga/

primero contiene más de 1000 horas de grabaciones orales efectuadas entre 1974 y 1998, y abarca hablantes de distintas edades y variants dialectales. El Segundo incluye grabaciones de charlas, conferencias, y mesas redondas sobre temas del ámbito politico y social.

- *Corpus Linguístico da Universidade de Vigo* (CLUVI)
  El CLUVI se desarrolla en el *Seminário de Linguística Informática* de la Universidade de Vigo (Aguirre Moreno *et al*. 2001, 2002, 2003). Se trata de un corpus escrito y oral que recoge textos contemporáneos de distintos registros especializados: jurídico-administrativo, periodístico, informático y literario. Así mismo, los textos pueden ser monolingües en gallego, traducciones gallego-español y traducciones inglés-gallego. La parte de textos de origen escrito se divide en cuatro subcorpus de aproximadamente un millón de palabras cada uno: el corpus paralelo TECTRA, constituido por textos de registro literario inglés-gallego; el corpus paralelo LEGA, de textos jurídico-administrativos gallego-español; el corpus monolingüe XIGA, de textos del ámbito informático escritos en gallego; y finalmente el corpus monolingüe MEGA, del ámbito de los medios de comunicación social. Igualmente, se está trabajando en la contrucción de otra parte adicional constituida por textos paralelos portugués-gallego. Se aspira al nivel de anotación morfosintáctica, así como a la alineación oracional de los subcorpus paralelos. Las aplicaciones que se pretenden construir en base a este corpus son varias: desde la construcción de herramientas básicas como extractores de información léxica, terminológica y fraseológica, hasta aplicaciones más sofisticadas y de largo plazo, como la extracción de información, la sumarización o la recuperación de información *on-line*.

Interesa remarcar también la existencia de otro corpus oral, VOGATEL, desarrollado por parte de la empresa Telefónica I+D en colaboración con la Universidade de Vigo. Su finalidad es estrictamente el entrenamiento de herramientas para el tratamiento del habla.

3.2.4 Corpus de euskera

El euskera tiene dos corpus lingüísticos principales, ambos de textos escritos. Se presentan en Agirre *et al*. (1998), dentro del marco general para el procesamiento del euskera que se ha desarrollado en la Euskal Herriko Unibertsitatea. Existen además otros corpus de carácter menor, confeccionados por parte de grupos de investigación para entrenar las herramientas de base estadística que construyen. Aquí me centraré únicamente en los corpus de referencia, que son:

- *Orotariko Euskal Hiztegia* (OEH, 'Diccionario General Vasco'):
  Corpus textual diacrónico que sirve como base en la creación del diccionario descriptivo de Euskaltzaindia, la academia de la lengua vasca. Contiene un total de 5.800.000 palabras, provinentes de 310 obras escritas en distintas variedades del euskera, desde el S.XVI hasta el momento en que se inicia la estandarización de la lengua, alrededor de 1960.

- *XX. mendeko Euskararen Corpus Estatistikoa* ('Corpus Estadístico del Euskera del S.XX')

Corpus del euskera actual, constituido por muestras de textos procedentes de alrededor de 6000 obras escritas durante el S.XX. Con la intención de recoger la máxima variedad léxica y estructural posible, se estableció en base a una selección aleatoria del inventario de obras escritas en euskera, compilado por UZEI (*Terminologia eta Lexikografiako Zentroa*, Centro Vasco de Terminología y Lexicografía). Estas obras pertenencen a todo tipo de géneros, desde poesía, teatro y literatura infantil, hasta investigación o libros de texto. Además, se procuró representar cada uno de los cuatro períodos relevantes de la historia del euskera del S.XX: 1900-1939 (desde el inicio de siglo hasta las guerras), 1940-1968 (de la posguerra al nacimiento del euskera estándar), 1969-1990 (desde los cambios producidos por el estándar hasta las publicaciones de las normas de Euskaltzaindia), 1991-1999 (posterior a la normativa). El corpus contiene un total de 4,7 millones de palabras, lematizadas con el lema estándar y el de la variante dialectal, y etiquetadas con la categoría gramatical y las distintas acepciones. Se diseñó con la intención de representar las distintas variedades diatópicas y diafásicas del euskera contemporáneo. Está pues pensado como fuente de datos que refleje el uso real de la lengua, evitando una función puramente prescriptiva.[13]

4 Un proyecto de corpus para el asturiano

4.1 Tecnología lingüística y lenguas minoritarias

Un proyecto de corpus lingüístico para el asturiano debe tomar necesariamente en consideración su estatus de lengua minoritaria. Ciertamente, las posibilidades de desarrollo de tecnología lingüística para una lengua de complexión sana son muy superiores a las que tiene una lengua como el asturiano o el euskera. El problema básico de las lenguas identificadas como minoritarias es que el nivel de recursos de que disponen, tanto materiales como humanos, es significativamente menor que el de las lenguas mayoritarias. A este factor hay que añadir al escaso, si no nulo, interés comercial que presentan, lo que supone un coste muy elevado para la investigación y el desarrollo en el ámbito.

El problema no es sin embargo exclusivo de las lenguas conocidas como minoritarias. Hay lenguas mayoritarias y que se encuentran en la misma situación, como por ejemplo el hindi, a pesar de sus 180 millones de hablantes en la India (Somers 2001). Además, la etiqueta de lengua minoritaria presenta cierta ambigüedad. Si por minoritaria entendemos lengua de pocos hablantes, debemos incluir lenguas como el finladés (4,7 millones en Finlandia) o el sueco (7,8 millones en Suecia)[14]. En estos casos, sin embargo, la condición minoritaria no obstaculiza una buena colocación en el ránquing de lenguas para las que existen aplicaciones para el procesamiento del lenguaje: el finlandés se encuentra en la 6ª posición y el sueco en la 7ª, después del inglés, el alemán, el francés, el español y

[13] En la página de UZEI (http://www.uzei.com/default_cas.html) se ofrece más información sobre este proyecto. El corpus es además consultable a través de internet, en la dirección siguiente: http://www.euskaracorpusa.net/XXmendea/Konts_arrunta_fr.html.

[14] Datos procedentes de: http://www.ethnologue.com.

el italiano, según el Natural Language Resource Registry (NLRS).[15] Por otro lado, si como lengua minoritaria entendemos lengua en situación minoritaria dentro del marco político donde se encuentra, el término puede aplicarse también a las lenguas de inmigrantes en relación a su país de adopción, aunque sean de lenguas mayoritarias en su país de origen.

Se hace pues necesario considerar cuáles son los elementos que garantizan o dificultan el acceso de una lengua a las tecnologías del lenguaje. Veremos a continuación como el estatus de cada lengua determina de modo general la relación que ésta mantiene con las tecnologías del lenguaje. Finalmente, identificaremos en la sección siguiente los rasgos que caracterizan la lengua asturiana y aventuraremos una estrategia posible para su entrada en el campo de la tecnología lingüística

Los factores que entran en juego en el acceso de una lengua a las tecnologías del lenguaje son esencialmente las siguientes:

- **Número de hablantes:** A mayor número de hablantes, mayor rendibilidad económica va a suponer el desarrollo de tecnología lingüística y, por consiguiente, mayor interés comercial capaz de mobilizar la inversión en el área por parte empresas privadas. El ejemplo paradigmático de rendibilidad comercial es sin duda el inglés.

- **Apoyo institucional:** En los casos en que no hay una rendibilidad económica visible, el apoyo y promoción de la lengua por parte de la administración son básicos para garantizar el desarrollo de aplicaciones de tecnología lingüística. Naturalmente, el apoyo desde el gobierno facilita también el de otras instituciones, como fundaciones culturales privadas con capacidad de mecenazgo. En esta situación se encuentran lenguas de pocos hablantes, como el finlandés o el holandés.

- **Diglosia:** La entrada de una lengua en el campo tecnológico depende también del grado en que ésta sea empleada en todos los ámbitos y registros de uso, tanto orales como escritos. En una situación de diglosia, una lengua supeditada y relegada estrictamente a la comunicación oral dentro de ámbitos familiares va a tener muy difícil la creación de tecnología lingüística: por un lado porque los recursos existentes como punto de partida (desde diccionarios y gramáticas hasta un cuerpo de textos, escritos o grabados, suficientemente voluminoso) van a ser muy escasos; por el otro, porque el desarrollo de aplicaciones va a pasar prioritariamente por la lengua subordinante. Así, la situación de escasez de tecnología lingüística en que se encuentra el hindi, a pesar de su elevadísimo número de hablantes, es probablemente debida a su supeditación al inglés.

- **Acceso a las tecnologías de la información:** Finalmente, para que una lengua tenga entrada de pleno en el terreno de las tecnologías del lenguaje es necesario que su sociedad tenga también acceso tecnológico de modo más o menos generalizado. Ésto no significa la existencia de ordenadores personales en todos los hogares de habla de la lengua en cuestión (lo que supondría una visión completamente occidente-céntrica

[15] http://registry.dfki.de. Agirre *et al*. (2002) ofrece un análisis de estos datos, orientado a las aplicaciones de PLN independientes de lenguaje.

del problema), sino el uso garantizado de tecnología informática en las instituciones académicas y de investigación, así como en las empresas privadas dedicadas al desarrollo de productos de base lingüística. En la mayoría de casos la imposibilidad de acceso a las tecnologías de la información está condicionada por una seria situación de diglosia. Tal es el caso del quechua, la lengua indígena americana más hablada, con cerca de 10 millones de hablantes.

Cada lengua del mundo se sitúa en un punto de coordenadas determinado en relación a estos cuatro parámetros, lo que conlleva también un nivel determinado de capacidad de desarrollo dentro de las tecnologías lingüísticas. El trabajo en el área de las tecnologías del lenguaje puede clasificarse en función de los tres niveles de desarrollo siguientes, los cuales se corresponden a grandes rasgos con los planteados en varios trabajos que reflexionan sobre las estrategias para el desarrollo de tecnología lingüística para lenguas minoritarias (Sarasola 2000; Agirre *et al*. 2002; Diaz de Ilarraza *et al*. 2003):

- **Fundamentos:**
  Fase de compilación de datos léxicos y corpus textuales –aún sin procesamiento a ningún nivel. Los productos resultantes en este estadio se consideran como el punto de partida necesario para el desarrollo de tecnología lingüística.

- **Herramientas:**
  La bibliografía citada anteriormente considera esta fase como la destinada a la construcción de herramientas para el procesamiento del lenguaje natural: lematizadores, analizadores morfológicos y sintácticos, alineadores de frases para corpus multilingües, ontologías o bases de conocimiento léxico-semántico, etiquetadores semánticos, etc. Se trata de herramientas que tanto van a servir para el tratamiento de la información compilada en la fase previa, como se van a beneficiar de ella (por ejemplo, para el entrenamiento de herramientas de base estocástica).

  Además de la construcción de herramientas, a mi parecer en esta fase hay que añadir también la elaboración de útiles lingüísticos basados en las colecciones de información compiladas en la etapa previa. Me refiero aquí a diccionarios y gramáticas de uso público. Su elaboración en base a corpus textuales permite reflejar más adecuadamente la lengua que se describe con ellos.

- **Aplicaciones:**
  Fase dedicada a la construcción de aplicaciones destinadas a usuarios no especializados. Por ejemplo: correctores gramaticales, sistemas de ayuda a la traducción con cierto nivel de sofisticación, sistemas de extracción o recuperación de información, o sistemas de diálogo. El trabajo en esta etapa se beneficia necesariamente de las herramientas desarrolladas en la fase anterior.

Para las lenguas con un número reducido de hablantes, en clara situación de diglosia y sin soporte institucional, la entrada en la tecnologías de la información va a ser si no imposible, extremadamente costosa. Ésta es por desgracia la situación de la inmensa mayoría de las 6800 lenguas que actualmente aún se hablan en nuestro mundo. En el mejor de los casos, su contacto con la tecnología lingüística va a ser a través de la

lingüística de corpus, si es que algún grupo investigador decide llevar a cabo el proyecto de recogida de materiales léxicos y compilación de corpus (en la mayoría de casos, sin ningún tipo de procesamiento ni anotación) antes de que la lengua en cuestión desaparezca.[16]

En una posición intermedia se encuentran, por un lado, las lenguas pequeñas con respaldo institucional, con un uso generalizado a todos los niveles y una capacidad tecnológica aceptable, como por ejemplo el holandés y el finlandés. Y por el otro, las grandes lenguas con un grado de acceso tecnológico relativamente bajo que ha dificultado hasta hace poco su entrada en el campo de las tecnologías lingüísticas, pero que por razones comerciales o estratégicas están empezando a generar cierto interés. Es el caso, por ejemplo del farsi o el tagalog. En este segundo nivel, el desarrollo de tecnología lingüística supera el nivel primario de compilación de corpus sin etiquetar, para entrar ya en la etapa de desarrollo de herramientas básicas para el procesamiento de corpus, o la generación de útiles de base lingüística (diccionarios y gramáticas) que se han podido beneficiar de corpus textuales básicos y bases de datos léxicos compilados en la etapa previa.

Finalmente, encontramos las lenguas que presentan un gran número de hablantes y que gozan de lo que he calificado como un contexto de fácil acceso tecnológico, como el inglés, el chino o el español –se trata indudablemente de la ínfima minoría. Para estos casos, el nivel de desarrollo de las tecnologías del lenguaje está llegando, si no lo ha hecho ya, a la creación de aplicaciones destinadas a usuarios no expertos como los mencionados en la tercera etapa del desarrollo de la tecnología lingüística.

Naturalmente esta partición en tres grandes bloques de las lenguas del mundo según su grado de participación de las tecnologías del lenguaje es la simplificación de una situación llena de matices. Un ejemplo de esto es el caso de las lenguas minoritarias más cercanas al asturiano y para las que hemos revisado su participación en la Lingüística de Corpus en la sección 3.2: el catalán, el euskera y el gallego. Se trata de tres lenguas con un número reducido de hablantes (aproximadamente 6 millones, 500.000 y 3,5 millones respectivamente de hablantes como primera lengua) y bajo la influencia de cierto nivel de diglosia respecto al español (mayor o menor según el caso), además de complicaciones derivadas de una fuerte variación dialectal en el euskera y, aunque en menor grado, en el caso del gallego. Sin embargo, el desarrollo de tecnología lingüística en estas tres lenguas es remarcable: en los tres casos se ha superado la fase inicial de creación de los fundamentos y se ha entrado ya en la segunda etapa, con la consiguiente generación de productos de base lingüística (como el *Diccionari de freqüències* del catalán, o bien el *XX mendeko Euskararen Corpus Estadistikoa* para el euskera, consultable a través de internet[17]), y la construcción de herramientas de procesamiento del lenguaje, indispensables para la entrada en el nivel de las aplicaciones de usuario no especializado. Además, tanto en el caso del euskera como en el del catalán, se han comercializado ya algunas aplicaciones finales correspondientes al tercer nivel de desarrollo del área (véase Diaz de Ilarraza *et al.* (2003) para el euskera, y Llisterri (2000) para el catalán). No dispongo de información en este sentido en lo que respecta al gallego.

---

[16] A ésto se dedica por ejemplo *SIL International* (entre otros proyectos; véase: http://www.sil.org) o programas de mecenazgo como el *Documentation Programme* del *The Hans Rausing Endangered Languages Project* (http://www.hrelp.org/doc_home.htm).

[17] Para uno y otro trabajo, ver las referencias en las correspondientes secciones 3.2.2 y 3.2.4.

En mi opinión, esta situación responde principalmente a dos factores. Por un lado, el compromiso con la lengua autóctona y la voluntad de trabajar para su normalización desde las esferas académicas y de investigación. Por el otro, la voluntad de las administraciones nacionales catalana, vasca y gallega de fomentar el trabajo en el área, en tanto que se entiende que únicamente así se puede asegurar la competitividad y validez de la lengua en el nuevo marco de la sociedad de la información.[18] A estos dos elementos hay que añadir un tercero: el nivel de desarrollo de las tecnologías lingüísticas de que gozan estas lenguas, a pesar de su condición, es sólo posible para las lenguas minoritarias del mundo occidental, en tanto que tienen un mínimo de estabilidad económica garantizado, acceso a las tecnologías de la información, y están exentas de un contexto de marginación social como es por ejemplo el caso de las lenguas indígenas en todo los países sin excepción del continente americano. El estado de desarrollo de la tecnología lingüística para nuestras tres lenguas no es óptimo si se compara con la situación de lenguas como el inglés o el español, claramente beneficiadas por la inversión desde el sector privado. Sin embargo, es muy superior al estado de la mayoría de lenguas del mundo que se encuentran en una situación comparable –es decir, con un número reducido de hablantes y supeditadas a la lengua mayoritaria y oficial.

En el contexto europeo, la aplicación de las tecnologías del lenguaje en el ámbito de las lenguas minoritarias autóctonas es actualmente un tema de creciente interés. Muestra de ello es: la cada vez más abundante bibliografía sobre el tema, la creación hace pocos años de SALTMIL,[19] un grupo de interés dedicado a ello (Nadeu *et al*. 2001), o bien los seminarios temáticos organizados en el marco de congresos de amplio alcance que se han centrado en esta cuestión.[20] La premisa de partida es la aceptación de la llamada Sociedad de la Información como el nuevo contexto de referencia para el intercambio de conocimiento y el consiguiente avance tecnológico de occidente, así como el papel indispensable de la Ingeniería Lingüística para su progreso. En este marco, la supervivencia de las lenguas minoritarias occidentales pasa necesariamente por el desarrollo de tecnología lingüística (Sarasola 2000, Nadeu *et al*. 2001).

Hemos visto que la complexión vital de cada lengua determina a grandes rasgos su posibilidad de desarrollo en el área de las tecnologías del lenguaje. Sin embargo, hay que considerar también que la entrada de una lengua minorizada en este campo supone, además de la posibilidad de actualización y mejora de sus útiles de descripción lingüística (diccionarios y gramáticas), su entrada en el contexto de la sociedad de la información, lo

---

[18] En este sentido, es relevante la propuesta de acción estratégica en las áreas de Industria de la Lengua e Ingeniería Lingüística por parte del gobierno vasco (véase la nota 11 para la referencia completa), o bien la creación del *Centre de Referència en Enginyeria Lingüística* (CREL) por parte de la Generalitat de Catalunya, y el *Centro Ramón Piñeiro para a Investigación en Humanidades* por parte de la Xunta de Galicia, dos instituciones dedicadas a la investigación y desarrollo de tecnología lingüística para las lenguas autóctonas.

[19] SALTMIL es el *Special Interest Group on Speech and Language Technology for Minority Languages*, creado dentro de la *International Speech Communication Association* (ISCA) (http://isl.ntftex.uni-lj.si/SALTMIL/).

[20] Por ejemplo los *workshops: 'Language Resources for European Minority Languages'* (Granada, LREC 1998), '*Developing language resources for minority languages: re-useability and strategic priorities'* (Atenas, LREC 2000), y '*Natural Language Processing Of Minority Languages And Small Languages'* (Batz-sur-Mer, France, TALN 2003).

que la califica como lengua moderna y capaz de competir a nivel tecnológico, abandonando su imagen arcaica, sin cohesión interna ni capacidad para la comunicación a todos los niveles. En definitiva, contribuye al rompimiento de la situación de diglosia en la que se encuentra, lo que a su vez retroalimenta positivamente la voluntad de apoyo institucional y vitaliza un progresivo interés comercial.[21] El desarrollo de las tecnologías del lenguaje es pues algo que va más allá de un mero ejercicio académico. Es un asunto de interés público y de repercusión social, en el que va de por medio la consolidación y supervivencia de las lenguas minoritarias como lenguas vivas en una cultura global de base tecnológica.

4.2 Punto de partida para el asturiano

La entrada del asturiano en el campo de las tecnologías lingüísticas se convierte por tanto en un proyecto de primera prioridad, a la par con su incorporación en los medios de comunicación de masas. La coyuntura política actual parece abrir una posibilidad al proyecto: por parte de los organismos responsables del área en la Comunidad Europea, existe un reconocimiento de la situación que se plantea actualmente a las lenguas minoritarias autóctonas en este nuevo marco de información; por parte de la administración asturiana, hay un cambio de política respecto a la normalización de la lengua asturiana, traducido recientemente en la garantía de aprobación de un plan de normalización social del asturiano en un futuro próximo, y en la creación de la Oficina de Política Lingüística.[22]

Además, la lengua asturiana goza ya de una mínima base de partida. En primer lugar, dispone de la Academia de la Llingua Asturiana (ALlA), una institución con autoridad lingüística, fundamental para articular proyectos de investigación. En segundo lugar, ha entrado en una fase de normalización gracias a la publicación de la *Gramática de la Llingua Asturiana* (3ª ed., 2001) y del *Diccionario de la Llingua Asturiana* (2000) por parte de esta misma academia. Finalmente, dispone de un decente cuerpo de textos escritos, principalmente en los registros literarios y periodísticos, una parte de los cuales ya ha sido digitalizada a través del *Proyectu Caveda y Nava*[23], así como también de un archivo oral --aunque necesariamente ampliable—compilado en el marco del proyecto *Archivu Oral de la Llingua Asturiana*.[24]

---

[21] En este sentido me parece ilustrativa una anécdota relacionada con la entrada del catalán a la industria televisiva y cinematográfica: Cuando en 1983 se estrenó TV3, el canal de televisión de Cataluña, oír hablar en catalán a *JR*, *Sue Ellen* y sus compañeros de la serie *Dallas* (ya conocidos por el público a través de TVE) era motivo de burla. En el 2002 sin embargo, la indignación popular y su consiguiente boicot comercial, causados por la negativa de la *Warner Bross* a doblar su *Harry Potter* al catalán, consiguió hacer cambiar de opinión a la multinacional estadounidense.

[22] Noticias publicadas en *Asturies.com, Diariu Electrónicu Asturianu* el 8 y el 29 de octubre pasado, respectivamente: http://www.asturies.com/seccion.php?fseccion=llingua#2387 y http://www.asturies.com/seccion.php?fseccion=llingua#2455 (URLs vigentes el 29 de octubre del 2003). No me consta sin embargo una mejora en la actitud del gobierno español respecto al reconocimiento de los derechos del asturiano como lengua perteneciente también al estado.

[23] http://www.cavedaynava.org/ (URL vigente el 29 de octubre del 2003).

[24] http://www.asturies.org/asturianu/archoral/ (URL vigente el 29 de octubre del 2003).

Sin embargo, no hay que ignorar la situación frágil en que actualmente se encuentra la lengua. Su estatus de no oficialidad se traduce a efectos prácticos en una menor capacidad para beneficiarse de apoyo económico. Además, existe la constatación de una pérdida generacional de hablantes, así como una disminución del uso de la lengua durante la última década en áreas y registros hasta ahora exclusivamente relegados a ésta (Llera Ramo, 2003, según la reseña del trabajo hecha en el diario electrónico *Asturies.org*[25]). Estos dos factores suponen una limitación considerable en términos de recursos, y por consiguiente la propuesta para un proyecto que dé entrada al asturiano en el campo de las tecnologías de la información no puede partir de planteamientos maximalistas. Asimismo, hay que atenerse a la experiencia sacada de proyectos con otras lenguas en situación similar al asturiano, como el euskera. En estos trabajos previos se ha visto como no se puede pensar en empezar a desarrollar herramientas ni aplicaciones si no existen los *fundamentos*; es decir, compilaciones léxicas y corpus textuales (Sarasola 2000, entre otros).

Aplicado al caso particular de la lengua asturiana, esto apunta a la creación de un corpus textual --escrito y/o oral-- como prioridad inicial para su entrada en el campo de las tecnologías lingüísticas. Un corpus textual, juntamente con la compilación de datos léxicos (que puede tener el diccionario de la ALlA como punto de partida), garantiza los cimientos para el desarrollo posterior de herramientas y aplicaciones. Así mismo, permite a medio plazo la producción de productos de base lingüística. Por ejemplo, los estudios lingüísticos realizados sobre los datos del corpus pueden derivar en aplicaciones como la mejora de la gramática y el diccionario mencionados anteriormente, la creación de un diccionario descriptivo en base al léxico más frecuente de la lengua, o bien la planificación de determinadas actuaciones sociolingüísticas. Igualmente, es posible pensar en aplicaciones que beneficien el aprendizaje del asturiano: tanto mejorando los materiales existentes a partir de datos procedentes del corpus, como desarrollando herramientas para el aprendizaje asistido por ordenador –por ejemplo, ejercicios que requieran reproducir las categorías morfosintácticas asignadas en determinadas frases del corpus, generar formas flexionadas a partir de un lema y su información asociada, completar perífrasis, frases hechas y colocaciones, etc.

Para ninguna de estas aplicaciones se require un nivel de tecnología excesivamente sofisticado. Un corpus textual lematizado y etiquetado morfosintácticamente, junto con una herramienta de explotación de los datos, es suficiente. Nos situamos sólo en el segundo nivel de desarrollo tecnológico introducido en la sección 4.1, y en cambio las posibilidades de las aplicaciones que se plantean son ya notables.

Adviértase sin embargo que se trata de aplicaciones que presuponen un fragmento particular de lo que es la lengua asturiana: el asturiano actual. La entrada del asturiano en las tecnologías del lenguaje se plantea (además de por el inherente interés a nivel científico) con la intención de permitirle el acceso al nuevo marco de comunicación y legitimarlo como lengua válida en todos los niveles de uso. Teniendo en cuenta su presente situación limitada, un corpus de lengua actual redunda en un beneficio inicial más generalizable a distintas áreas de la lingüística aplicada que no, por ejemplo, un corpus histórico.

---

[25] http://www.asturies.org/seccion.php?fseccion=llingua#2454 (URL vigente el 29 de octubre del 2003).

Asimismo, es especialmente relevante que este primer corpus potencial recoja la máxima variedad dialectal, diastrática y diafásica del asturiano, debido al estadio de asimilación de la norma estándar en el que se encuentra actualmente, a pocos años de la aparición de un diccionario y una gramática editados por la institución con autoridad normativa sobre la lengua. En definitiva, interesa que este primer corpus de la lengua asturiana sea un corpus de referencia de lengua general.

Un corpus para el asturiano tiene que concebirse también para que, a más largo plazo, sirva como base para el desarrollo de aplicaciones tecnológicas más sofisticadas: sistemas de diálogo, sistemas de sumarización, de recuperación o extracción de información, etc. Para ello se requieren dos elementos. Primero, que las herramientas básicas de procesamiento, como lematizadores, analizadores morfosintácticos o sintácticos, se hayan construido de modo que puedan ser fácilmente reutilizables, en caso de ser necesario, por parte de cada una de estas aplicaciones finales (nuevamente Sarasola 2000 y Diaz de Ilarraza *et al*. 2003 defienden este criterio fruto de su experiencia con el euskera). Segundo, que el desarrollo de estas aplicaciones finales parta de un análisis de la realidad del asturiano y la consiguiente adecuación a ella.

Es decir, teniendo en cuenta la condición de lengua minoritaria del asturiano, no es factible plantearse el desarrollo de todas las aplicaciones en las que se trabaja actualmente para lenguas como el inglés. Al contrario, hay que concentrar el esfuerzo en aquellas tareas para las que exista garantía real de uso –y con ella, un mínimo de rentabilidad comercial. Considerando las condiciones de la lengua asturiana, parece claro que, por ejemplo, un sistema de sumarización de documentos va a ser de un interés y nivel de aplicación muy inferior al que puede tener un sistema de traducción automática entre el asturiano y el español. Por este motivo, el proyecto de construcción de un corpus de lengua asturiana debe de considerar, en mi opinión, la posibilidad de que al menos una parte sea bilingüe asturiano-español –preferentemente paralelo.

Finalmente, el proyecto también debe tomar en consideración el estado en que se encuentra el fragmento de lengua que se pretenda representar. Por ejemplo, va a tener que tomar decisiones sobre el tratamiento de la variación dialectal; por ejemplo, la que afecta a la flexión de determinadas categorías, por la repercusión que esto puede tener en el proceso de lematización y asignación de etiquetas morfosintácticas. O bien, va a tener que establecer una política específica para la variación ortográfica, en caso de tratar con textos de una etapa pre-normativa. En este sentido, puede servir de referencia la experiencia de otros proyectos desarrollados para lenguas cercanas al asturiano y que se han presentado en la sección 3.2.

5 Conclusiones

En este artículo se ha planteado la posibilidad de un proyecto de corpus para la lengua asturiana. Con esta finalidad, he presentado brevemente el objeto de estudio de la Lingüística de Corpus y el interés que ofrece dentro de las disciplinas de base lingüística, tanto para la investigación como para el desarrollo de útiles y aplicaciones. Hemos visto que la vocación final de un corpus lingüístico es servir de fuente de datos para el análisis empírico de una lengua y su consiguiente modelización, la cual puede ser utilizada en distintos niveles de desarrollo: para estudios de base teórica, para la elaboración de útiles de base lingüística como diccionarios o gramáticas descriptivas, o para la construcción de

herramientas de base tecnológica, como correctores ortográficos y gramaticales o bien programas de traducción automática.

Un corpus lingüístico es por tanto de indudable interés tanto para la comunidad de profesionales del lenguaje (desde lingüistas teóricos y sociolingüistas, hasta lexicógrafos y traductores) así como, a más largo plazo, para el público general, beneficiario de las aplicaciones resultantes. Toda esta potencialidad hace de los corpus lingüísticos el fundamento necesario para el desarrollo de tecnología lingüística en cualquier lengua.

Por otro lado, se ha analizado la relación entre lenguas minorizadas y tecnologías del lenguaje, y hemos visto como en el marco actual de la sociedad de la información, las tecnologías lingüísticas van a jugar un papel definitivo para lograr una normalización social completa de las lenguas minorizadas. Obviamente, esto es posible en tanto que se satisfagan ciertas condiciones, como un acceso tecnológico mínimamente garantizado o el apoyo institucional a la lengua.

A pesar de la frágil situación en que se encuentra la lengua asturiana, su marco social y político actual parece cumplir mínimamente estos requisitos. Es importante pues aprovechar la coyuntura y asegurar la entrada de la lengua a las tecnologías del lenguaje, la cual debe pasar necesariamente por la construcción de un corpus del asturiano. Asumiendo esto, se han apuntado algunas características que debería cumplir este corpus potencial con la finalidad de garantizar un rendimiento máximo a pesar de la limitación de recursos en que se encuentra la lengua. Específicamente, se ha considerado que un primer corpus para el asturiano debe concebirse como un corpus de lengua general, que dé cuenta de sus variedades diatópicas, diastráticas y diafásicas, y que, a poder ser, incluya también un fragmento de textos paralelos asturiano-español.

6 Bibliografía

ABAITUA, JOSEBA (1999) "Quince años de traducción automática en España". *Perspectives: Studies in Translatology*. 7-2: 221-230. Versió catalana: "Quinze anys de traducció automàtica a l'Estat espanyol". *Digit.HUM*. Universitat Oberta de Catalunya. http://www.uoc.es/humfil/digithum/digithum2/catala/didactica/index.html

AGIRRE E., ALDEZABAL I., ALEGRIA I., ANSA O., ARREGI X., ARRIOLA J., ARTOLA X., DÍAZ DE ILARRAZA A., EZEIZA N., GOJENOLA K., MARITXALAR A., MARITXALAR M., ORONOZ M., SARASOLA K., SOROA A., URIZAR R., URKIA M. (1998) 'A framework for the automatic processing of Basque'. *Workshop on Lexical Resources for Minority Languages*. LREC 1998, Granada.

AGIRRE E., ALDEZABAL I., ALEGRIA I., ARREGI X., ARRIOLA J.M., ARTOLA X., DÍAZ DE ILARRAZA A., EZEIZA N., GOJENOLA K., SARASOLA K., SOROA A. (2002) 'Towards the definition of a basic toolkit for HLT'. *LREC 2002. Third International Conference On Language Resources And Evaluation*. Las Palmas, Islas Canarias.

AGUIRRE MORENO, J.L., N. ANDIÓN, X. GÓMEZ GUINOVART (2001) "Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informática en lengua gallega". *Procesamiento del Lenguaje Natural*, 27: 13-19.

AGUIRRE MORENO, J.L., A. ÁLVAREZ LUGRÍS, X. GÓMEZ GUINOVART (2002) "Etiquetario morfosintáctico del SLI para corpus de lengua gallega: aplicación al corpus paralelo TECTRA". *Procesamiento del Lenguaje Natural*, 28: 23-34.

AGUIRRE MORENO, J.L., A. ÁLVAREZ LUGRÍS, I. BRAGADO TRIGO, L. CASTRO PENA, X. GÓMEZ GUINOVART, A. LÓPEZ LÓPEZ, J. R. PICHEL CAMPOS, E. SACAU FONTENLA, L. SANTOS

SUÁREZ (2003) "Alinhamento e etiquetagem de corpora paralelos no CLUVI (Corpus Linguístico da Universidade de Vigo)". Almeida, J.J. (ed.), *Workshop CP3A 2003, Corpora Paralelos: Aplicações e Algoritmos Associados*. Universidade do Minho, Braga, Portugal.

ASTON, GUY, LOU BURNARD (1998) *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh University Press.

BACH, CARME, ROSER SAURÍ, JORDI VIVALDI, M. TERESA CABRÉ (1997) *El Corpus de l'IULA: descripció*. Sèrie Informes. IULA, Barcelona.

BIBER, D. (1993). 'Representativeness in corpus design'. *Literary and Linguistic Computing*, Vol. 8 (4): 243-257.

BOIX, E. (1996) 'Els materials de llengua oral dels corpus de català contemporani de la UB (CUB)'. LL. PAYRATÓ, E. BOIX, M.R. LLORET, M. LORENTE (Eds.) *Corpus, Corpora. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2)*. Barcelona: Promociones y Publicaciones Universitarias SA. pp. 93-114.

CARLSON, LYNN, DANIEL MARCU, MARY ELLEN OKUROWSKI (2003) 'Building a Discourse-Taggeed Corpus in the Framework of Rethorical Structure Theory'. KUPPEVELT, JAN VAN, RONNIE SMITH (Eds.) *Current Directions in Discourse and Dialogue*. Kluwer (to appear).

DIAZ DE ILARRAZA A., A. GURRUTXAGA, I. HERNAEZ, N. LOPEZ DE GEREÑU, K. SARASOLA (2003) 'HIZKING21: Integrating language engineering resources and tools into systems with linguistic capabilities'. *Workshop on NLP of Minority Languages and Small Languages*. TALN 2003. Batz-sur-Mer, 11-14 juin 2003.

FLIGELSTONE, STEVE (1992) 'Developing a scheme for annotating text to show anaphoric relations'. Gerhard Leitner (Ed.) *New Directions in English Language Corpora: Methodology, Results, Software Developments*. Berlin: Mouton de Gruyter.

FRANCIS, W.N., H. KUCERA (1964) *Manual of Information to Accompany a Standard Sample of Present-day Edited American English, for use with Digital Computers*. Providence RI: Department of Linguistics, Brown University. Edición revisada y ampliada (1979): http://www.hit.uib.no/icame/brown/bcm.html (última actualización: 11/9/1997).

GARCÍA-MATEO, CARMEN, MANUEL GONZÁLEZ-GONZÁLEZ (1998) 'An overview of the existing language resources for Galician'. *Workshop on Language Resources for European Minority Languages*. LREC, Granada, 28-30 May 1998

LLERA RAMO, FRANCISCO JOSÉ (2003) 'Lluces y solombres na evolución sociollingüística del asturianu'. *XXII Xornaes Internacionales d'Estudiu de la llingua Asturiana*. Academia de la Llingua Asturiana, Vigo, 27-29 ochobre 2003.

LLISTERRI, JOAQUIM (1997) 'Transcripción, etiquetado y codificación de corpus orales'. *Seminario de Industrias de la Lengua,* Fundación Duques de Soria, 15 de julio de 1997.

LLISTERRI, JOAQUIM (1999) "Corpus orals per a la fonètica i les tecnologies de la parla". *Actes del I Congrés de Fonètica Experimental*. Tarragona, 22, 23 i 24 de febrer de 1999. Universitat Rovira i Virgili - Universitat de Barcelona: 27-38.

LLISTERRI, JOAQUIM (2000) 'O catalán nas industrias da lingua', *Lingua e cultura catalanas*, Cursos de extensión universitaria, Universidade de Vigo, 5 July 2000.

MCENERY, TONY, ANDREW WILSON (2001) *Corpus Linguistics*. Edinburgh University Press.

MANN, WILLIAM, SANDRA THOMPSON (1988) 'Rethorical Structure Theory. Toward a functional theory of text organization'. *Text*, 8(3): 243-281.

NADEU, CLIMENT, DONNCHA Ó'CRÓINÍN, BOJAN PETEK, KEPA SARASOLA, BRIONY WILLIAMS (2001) ISCA SALTMIL SIG: Speech and Language Technology for Minority Languages. http://gps-tsc.upc.es/veu/research/pubs/download/Nad01b.pdf (28/10/2003)

PINO, M., M. SÁNCHEZ SÁNCHEZ (1999) "El subcorpus del Banco de Datos CREA-CORDE (RAE): procedimientos de transcripción y codificación". *Oralia*, 2: 83-138.

PUSTEJOVKSY, JAMES, PATRICK HANKS, ROSER SAURÍ, ANDREW SEE, ROBERT GAIZAUSKAS, ANDREA SETZER, BETH SUNDHEIM, LISA FERRO (2003) 'The TIMEBANK Corpus'. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, 28-31 March 2003.

QUIRK, R. (1992). "On Corpus Principles and Design". JAN SVARTVIK (ed.) (1990) *The London Corpus of Spoken English: Description and Research*. Lund University Press: 457- 469.

QUIRK, R., S. GREENBAUM, G. LEECH AND J. SVARTVIK (1985) *A comprehensive Grammar of the English Language*. Harlow, Longman.

RAFEL, J. (1992-93) "El 'Diccionari del català contemporani': Treballs realitzats i previsions de futur", *Llengua i Literatura* 5: 733-737.

RAFEL, J. (1996) "El Diccionari del català contemporani i el Corpus textual informatitzat de la llengua catalana". LL. PAYRATÓ, E. BOIX, M.R. LLORET, M. LORENTE (Eds.) *Corpus, Corpora. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2)*. Barcelona: Promociones y Publicaciones Universitarias SA: 71-92.

RAFEL I FONTANALS, J. (Dir.) (1996-1998) *Diccionari de freqüències*. 3 Vols. Barcelona: Institut d'Estudis Catalans. CD-ROM de l'obra completa.

RODRÍGUEZ, CARLOS (2003) "Applying Information Extraction techniques to metalinguistic discourse", *Topics in Computational Linguistics and Intelligent Text Processing*; Lecture Notes in Computer Science. Springer-Verlag. (en prensa)

SÁNCHEZ, A., R. SARMIENTO, P. CANTOS, J. SIMÓN, J. (1995) *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*. Madrid: SGEL.

SÁNCHEZ-LEÓN, F., J. PORTA, J.L. SANCHO, A. NIETO, A. BALLESTER, A. FERNÁNDEZ, J. GÓMEZ, L. GÓMEZ, E. RAIGAL, R. RUIZ (1999). «La anotación de los corpus CREA y CORDE». *Actas del XV Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, SEPLN.

SARASOLA, K. (2000) 'Strategic priorities for the development of language technology in minority languages'. *Workshop on Developing Language Resources for Minority Languages: Re-useability and Strategic Priorities*. LREC 2000. Athens, Greece.

SINCLAIR, J. (ed.) (1987) *Looking Up, An Account of the COBUILD Project*. London: Collins.

SINCLAIR, J. (1996) *Preliminary recommendations on corpus typology*. EAGLES Document TCWG-CTYP/P. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html (21/9/03)

SOLER I BOU, J. (1998) 'Los corpus textuales en lengua catalana'. *Curso de Industrias de la Lengua "Proyectos actuales en procesamiento de lenguaje natural"*, Fundación Duques de Soria, Soria, 13-17 de julio de 1998.

SOMERS, HAROLD (2001) 'Where do we stand?'. Panel Session, *MT Summit VIII*, Santiago de Compostela. 18-22 September 2001.

SVARTKIK, JAN (ed.) (1990) *The London Corpus of Spoken English: Description and Research*. Lund Studies in English 82. Lund University Press.

TOGNINI-BONELLI, E. (1996). *Corpus Theory and Practice*. Birmingham: TWC.

TORRUELLA, JOAN, JOAQUIM LLISTERRI (1999) 'Diseño de corpus textuales y orales'. BLECUA, J.M., G. CLAVERIA, C. SÁNCHEZ, J. TORRUELLA (Eds.) *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona, Universitat Autònoma de Barcelona, Editorial Milenio: 45-77.

VIAPLANA, J. (2000) 'Corpus oral de variació'. *Jornades del Centre de Referència en Enginyeria Lingüística (CREL)*, Institut d'Estudis Catalans, Barcelona, 4 i 5 d'abril de 2000.

ÍNDICE