Proyecto SOTA.: Sistema de Organización de Texto Abierto

Investigador Principal: O. Santana. Colaboradores: Z. Hernández, G. Rodríguez, J. C. Rodríguez, J. D. Gonzalez. Grupo de Investigación de Estructuras de Datos. Departamento de Informática y Sistemas. Universidad de Las Palmas de Gran Canaria.

Objetivos

Desarrollo de un sistema para la indización de documentos textuales débilmente estructurados o sin estructura definida. Además de un alto grado de *flexibilidad* tanto en lo que respecta a los formatos permitidos como a las modalidades de interrogaciones posibles; incluye las cualidades de ser *adaptable* a una amplia gama de configuraciones de recursos informáticos y *transportable* a los entornos operativos más populares con un mínimo esfuerzo de programación. La figura 1 muestra la descomposición modular del proyecto.

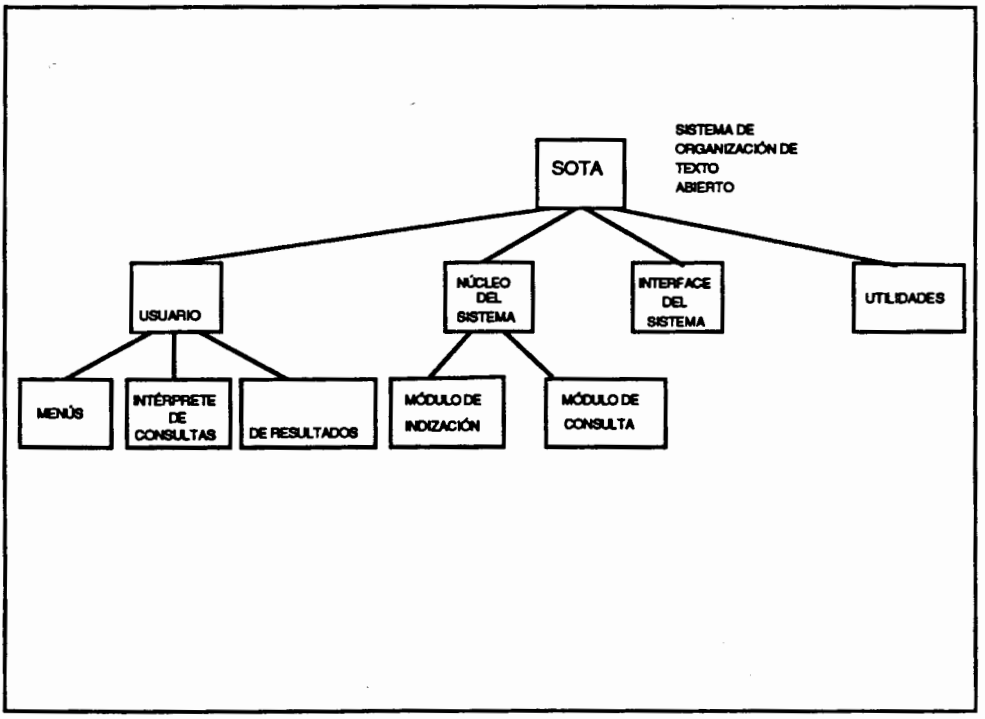

Figura 1

Características

Tratamiento de la información en texto abierto sin restricciones (en cuanto al número de documentos, bibliotecas, palabras, apariciones de los campos, etc.), limitado únicamente por las peculiaridades físicas de la máquina.

- Documentos sin estructura.
- Documentos débilmente estructurados a nivel de campos; puede contener distinto tipo de información como por ejemplo: textual, fecha, numérica, etc.

Sistema multibiblioteca. El volumen documental se gestiona en base a bibliotecas. Cada biblioteca la compone un conjunto de documentos relacionados.

Sistema multiusuario. El sistema permite trabajar simultáneamente a varios usuarios:

- Las bibliotecas son independientes entre sí, y por tanto, no existen restricciones en cuanto a llevar a cabo operaciones con varias al mismo tiempo.
- Dentro de una misma biblioteca distintos usuarios pueden realizar consultas simultáneas, pero estas no pueden coincidir con los procesos de indización.

Formatos de documentos flexible.

- En principio SOTA está pensado para aceptar documentos en formato ASCII y WordPerfect.
- Documentos en otros formatos no presentan mayor inconveniente porque todos ellos permiten conversiones a los permitidos.
- No obstante, se piensa incorporar de manera directa los formatos de documentos más importantes.

Seguridad.. Debido a la heterogeneidad de ambientes de implantación práctica y a que en muchos casos la confidencialidad constituye un aspecto crítico de la organización, la seguridad conforma uno de los objetivos de mayor importancia de SOTA. Se aplica a tres niveles distintos:

1) A nivel de entrada, mediante claves de acceso (contraseñas) de los usuarios de SOTA.
2) Distintos niveles de acceso a cada biblioteca según los permisos que tenga el usuario en cuestión. Atendiendo a las capacidades que tienen, se distinguen diferentes tipos de usuarios.
3) Protección contra robo de la biblioteca. Encriptación de documentos mediante contraseñas de seguridad.

Consultas. Todos los documentos que satisfacen la búsqueda constituyen La respuesta.

- Tipos de consultas: exacta, más similares, truncamientos, máscaras, rango, cercanía.
- Lenguaje de consulta. En la sintaxis intervienen operandos y operadores (lógicos y relacionales).

Entorno de ejecución.

- Windows debido a la extensión de su implantación en el mercado que facilita el manejo de SOTA.
- Posteriormente se extenderá a otros entornos (Unix, Mac, ...).

CORPUS DEL NÁHUATL EN TRANSCRIPCIÓN FONÉTICA

Andrés González Ruiz
Ángel Yanguas Álvarez de Toledo

UNIVERSIDAD DE SEVILLA

Conscientes de la relevancia lingüística y social de las lenguas amerindias, y aprovechando las oportunidades ofrecidas por el Plan Nacional I+D (que incluía, entre sus líneas prioritarios, una específica sobre América Latina), el Área de Lingüística General de la Universidad de Sevilla ha iniciado un programa de investigación en el área lingüística meso-americana. Uno de los primeros resultados tangibles de esta iniciativa es un Corpus del náhuatl hablado, con un tamaño próximo a las 200.000 palabras, compilado *sobre el terreno* por el primero de los autores de este informe [A.G.R.].

Por su naturaleza y por su contenido, el Corpus -uno de los pocos existentes sobre *vernáculos* amerindios- puede ser de gran utilidad en varios campos: estudios sobre el náhuatl y el español mejicano, lingüística general (investigaciones formales, tipológicas, discursivas, sociolingüísticas, ...) y otras disciplinas sociales (etnología, folklore, narratología oral, ...). También puede ser de interés como término de comparación para otros *corpora* y en el desarrollo de herramientas informáticas.

La base de este Corpus oral son 35 horas grabadas *in situ* de habla coloquial y espontánea, fundamentalmente monólogos narrativos y conversaciones sostenidas con A.G.R.. El trabajo de campo se realizó en una zona rural situada al norte del Estado de Guerrero (unos 150 kms. al sur de la Ciudad de México), en el curso alto del Río Balsas. La situación lingüística en la región es característicamente diglósica, con el español como código 'de prestigio' y el mexicano como código 'sincrético'. Las variedades lingüísticas locales se inscriben en los denominados dialectos centrales del náhuatl.

El Corpus consta de 60 textos coherentes y unitarios, en ficheros independientes con encabezamientos tipo TEI que proporcionan información sobre el hablante, fecha y lugar de registro, entorno y situación, términos temáticos y, en su caso, características especiales de la transcripción. Para la mayoría de los textos existe traducción literal o glosa, aunque ésta no se incluye en el Corpus.

La clase y el nivel de codificación, así como las convenciones de transcripción, pueden verse en la muestra ilustrativa que reproducimos de la versión WordPerfect (hay una versión 'gemela' en ASCII).

En la transcripción se sigue la notación AFI, con algunas excepciones habituales y algunas limitaciones impuestas por la disponibilidad de caracteres. Por ejemplo, en la versión ASCII se emplean los dígrafos [c% s%] en vez de [_ ö] (caracteres WP 1.99 y 1.177). La cantidad vocálica y la labialización (fono-lógicamente relevantes en náhuatl) requieren caracteres adicionales, *v.g.*, [i:] y [kw] ([k^] en ASCII; *nota*: el superíndice [w] es necesario por la existencia del aproximante [w]). También tienen representación digráfica [ts] y [tl], correspondientes a los segmentos africados coronales característicos del náhuatl, sibilante y lateralizado respectivamente.

El náhuatl es una lengua de acento fijo (en penúltima sílaba) salvo en algunos procesos morfológicos; en los elementos léxicos importados se tiende a mantener el patrón acentual original. En el corpus sólo se registra el acento en palabras polisilábicas no-paroxítonas (ejemplos en la muestra: 'i:xoleh y 'ko:lerah).

Para preservar la coherencia e integridad de los textos, las intervenciones del entrevistador [A.G.R.] se incluyen, lo mismo que las emisiones erróneas o incompletas de los informantes, pero entre marcas que

posibilitan su exclusión en el procesamiento: corchetes en el primer caso, paréntesis en el segundo (ej.: (o:ti_)). Adicionalmente, se marcan en su lugar las llamadas ([*]) a notas editoriales relevantes para la interpretación de la transcripción. Los párrafos se marcan y numeran (al estilo COCOA) para facilitar la ulterior localización. El formato WordPerfect permite prescindir de la unidad 'línea'.

Entre la información supra-segmental, discursiva y estilística, se anotan las pausas, las citas de estilo directo, las preguntas, las exclamaciones y los incisos, como puede apreciarse en la muestra adjunta. En el corpus no hay 'cristales' (en el sentido TEI): todo lo enunciado por los hablantes nativos se transcribe, incluso los elementos fáticos.

Para más información sobre el Corpus y sobre otros materiales del náhuatl en MRF, contactar por correo electrónico con uilga@cica.es.

1.- Muestra del Corpus transcrito

<p 1> ah / pues un (o:_) o:pe:h un # ke:tlah (pri_) a:man tlaki:ska:tlah / o:tipiökakeh # diah / o:tipiökakeh / o:titlankeh / titlasakahka[*] tosen # diah / (o:ti_) o:tiwa:hlahkeh # diah / a:man in note:lpo:tsin / na: nemi / o:nikti:tlan i:wa:n i:na:n / nikihlia # <«> öwia / un / öpehpenati un sentli wa:n o:noka:h <»> / <«> teh / ma:ski <»> # o:nikti:tlan o:jah # ke a:man ka: kwalka:n / a las sinko o:ki:skeh / ahsikeh pe: komo a las siete # diah / de umpa / note:lpo:tsin / o:pe:h kahsik toto:nki / fuerteh # diah / niman kwa:hwi:kak nokuña:doh # las dose na: nemi / nehkok / na:n ninemi / na: nikihlia # <«> <¿> tli:no:n tikpia <?> <»> # wa:hjetih pan nokaba:joh <«> ah <{> kihta <}> teh / nitoto:nia a:man <»> #

[¿Cuándo fue eso?, ¿en diciembre?]

<p 2> jeh / diah / kineh / nikilwia # <«> kwa:ko:n ötolo pasti:jah <»> # je (nik_) nikmak pasti:jah / kineh / de para un toto:nki / wa:hnote:kak / notla:tlapa_oh ka kobi:xah / pan se: o:rah o:ka:h / wa:hnokwite:h / o:jah ne: paka (o:_) o:noöi:öato # kwa:h nikitak wa:h_o:katih / kihta # <«> jo:pe:h niki:ötia pu:roh jestli <»> # <«> 'i:xoleh <{> nikihtoh <}> kwa:ko:n tlakah /

in kahsik un / 'ko:lerah / jo:kahsik / ah teh <»> # segi:doh jaw / segi:doh jaw / niman i:na:n ke:n a:man / nin öehko #

2.- Glosa de los extractos transcritos

Ä Ah, esto empezó como a pri... como ahorita, en la seca. Cosechamos. Luego que cosechamos, terminamos, acarreamos nuestro maicito. De ahí, nos vinimos. Ahorita a este mi hijito que anda ahí, lo mandé con su mamá. Le dije: 'vete a juntar el maicito que se quedó'. 'Bueno, pues'. Lo mandé y se fue. Como ahorita temprano, a las cinco, salieron. Llegaron allí como a las siete. De ahí, mi hijito empezó a agarrar calentura fuerte. Enseguida lo trajo mi cuñado. A las doce aquí está. Llegué, aquí estoy, le digo: '¿qué tienes?'. Viene montado en mi caballo. 'Ah -dice- ahorita tengo calentura, pues'.
Ä ¿Cuándo fue eso?, ¿en diciembre?
Ä Sí, de ahí, este, le digo: 'entonces tómate una pastilla'. Le dí una pastilla, pues, para la calentura. Se acostó, se tapó con la cobija. En una hora lo dejó. Se levantó. Fue ahí a cagar. Cuando lo fui a ver, viene llorando, dice: 'ya empecé a sacar pura sangre'. 'Híjole -dije- entonces éste ya agarró el cólera ése, ya lo agarró, pues'. Seguido va, seguido va, y luego su mamá, como ahorita, ni llega.