VARCOM CONCORD – UNA HERRAMIENTA DE ANÁLISIS PARA EL CORPUS MULTIMODAL VARCOM

OLIVER STRUNK
Universidad de Barcelona

RESUMEN. *El corpus multimodal Varcom desarrollado en la Universidad de Barcelona consta de archivos de audio-video y las correspondientes transliteraciones, algunas con marcaje de elementos de entonación. La identificación de elementos gramaticales - como p. ej. los conectores - con herramientas tradicionales exigía una importante dedicación a tareas mecánicas e implicaba potenciales malinterpretaciones. La herramienta Varcom Concord ha sido concebida especialmente para el uso con las convenciones de transcripción adoptadas en Varcom (en su versión catalana) e intenta contribuir a la optimización de los procesos de análisis. La posibilidad de incluir listados de elementos a identificar y desambiguar dentro del corpus permite adaptarla fácilmente a otros corpus de la lengua escrita o corpus de transliteraciones. En el presente artículo planteamos la problemática inicial, describimos la propuesta de procedimiento y detallamos el funcionamiento de la aplicación informática desarrollada para la solución de estos problemas. Finalmente presentamos aspectos de arquitectura abiertos para la adaptación de la herramienta a otros usos de explotación textual.*

ABSTRACT. *Das an der Universität Barcelona entwickelte multimodale Korpus Varcom enthält Audio- und Videodateien und die entsprechende Transliteration, die teilweise zusätzlich mit intonatorischer Auszeichnung versehen wurde. Die Identifizierung bestimmter Worteinheiten, wie z.B. die heterogene Gruppe der Konnektoren, mit traditionellen Mitteln erforderte einen hohen Arbeitsaufwand und war mit möglichen Fehlinterpretationen belastet. Das Werkzeug Varcom Concord wurde spezifisch für die Verwendung mit der katalanischen Version von Varcom entwickelt und optimiert den Auswertungsprozess der Texte. Die Möglichkeit der Einbindung von neuen Listen mit zu identifizierenden Elementen, die innerhalb des Kospurs disambiguiert werden sollen, ermöglicht eine schnelle Anpassung an andere transkribierte Sprachkorpora. In diesem Beitrag wird die ursprüngliche Problematik dargestellt, die vorgeschlagenen Arbeitsprozesse und die entsprechenden Funktionen des Programms, anhand derer die ursprüglichen Probleme gelöst werden. Abschliessend werden jene Aspekte diskutiert, die die Anpassung des Programms an andere Textanalyseverfahren ermöglichen.*

1. INTRODUCCIÓN

El corpus Varcom (Payrató et al. 2003 y Payrató et al. 2005) es un corpus multimodal que incluye dos líneas de creación de corpus principales, una centrada en el catalán (Varcom Catalán), otra en el alemán (Varcom Alemán). Ambas líneas incluyen entrevistas en video semiestructuradas en 3 idiomas: catalán, castellano e inglés en el caso del corpus centrado en el catalán; alemán, catalán y castellano en el caso del corpus centrado en alemán. Las dos líneas son comparables por la estructura de las entrevistas y las tipologías textuales que en ellas se elicitan (Fernández-Villanueva: 2007), pero el almacenamiento de datos, su estructura y por consiguiente las herramientas informáticas utilizadas para la explotación presenta divergencias importantes.

Así, los datos de Varcom Alemán se han transcrito con el editor Exmaralda (Schmidt 2004). Se trata de un editor multimodal que permite el añadido de pistas para un número ilimitado de capas de transcripción. A esta posibilidad técnica, presente también en otras herramientas, como Praat (Boersma, Weenink 2005), se une la posibilidad de sincronización y visualización del vídeo asociado a la pista de audio.

Las dos ventajas diferenciales básicas de Exmaralda son la existencia de herramientas adicionales para la creación de un corpus de archivos y otra de un concordancer integrado (CorpusManager, Zecke) y las numerosas posibilidades de exportación e importación de

archivos provenientes de otros corpus, de modo que se da mayor facilidad para la incorporación de archivos externos al corpus existente.

Las transcripciones de Varcom Catalán en cambio se han elaborado con el editor Elan (Hellwig 2006), que ofrece una mayor facilidad de uso, aunque también una ligera merma de funcionalidad respecto a Exmaralda. Las transcripciones propiamente dichas se han elaborado con Word y están provistas de marcas entonativas.

No obstante las tecnologías utilizadas para el almacenaje, la distribución y la explotación de ambos corpus, las posibilidades técnicas no tardan en agotarse ante las necesidades de análisis de los investigadores que trabajan con estos datos. Éste fue el caso del análisis de conectores llevado a cabo por Cuenca a partir de estudios previos (Cuenca 2001), que, para un análisis de conectores dentro del corpus Varcom Catalán, se encontró con el problema de la identificación de una extensa lista de conectores dentro de un corpus no menos extenso y la desambiguación de esos mismos conectores para evitar que análisis posteriores pudieran basarse en una errónea identidad de conectores y formas homónimas.

La propuesta por parte nuestra fue elaborar una herramienta informática que gestionara los correspondientes subprocesos y asistiera en la toma de decisiones. Para ello la problemática inicial se fragmentó en una serie de subprocesos que abarcaran todo el proceso global:

1. Identificación de conectores en forma de cadenas literales en los tres idiomas del corpus: catalán, castellano e inglés.
2. Identificación de encadenamiento y acumulación de conectores (por posibles interrelaciones semánticas)
3. Extracción de conectores con contexto.
4. Desambiguación y clasificación
5. Eliminación de formas concurrentes homónimas

Estos subprocesos, concebidos ya desde el punto de vista de procesamiento automático, generan un procedimiento que se contempla las siguientes fases:

1. Selección de transcripciones a procesar
2. Selección del grupo de conectores a identificar en el grupo de transcripciones de la fase 1
3. Preprocesamiento de las transcripciones de la fase 1 para la posterior búsqueda de cadenas literales (necesario por la acumulación de signos de transcripción intertextuales)
4. Identificación de las cadenas literales de los diferentes conectores de la fase 2 en las transcripciones de la fase 1
5. Generación del listado de conectores encontrados
6. Desambiguación asistida (añadido de monovalencia-polivalencia, función de concordancia, contexto anterior y posterior)
7. Identificación de acumulación de conectores (marcaje en azul)
8. Exportación de resultados para procesamiento posterior

Este procedimiento ha sido elaborado entre julio y noviembre de 2006 por Oliver Strunk con las ayudas, sugerencias e indicaciones de M. Fernández-Villanueva y M.J. Cuenca. El programa informático correspondiente, Varcom Concord, ha sido realizado con los recursos propios del grupo LADA del Departamento de Filología Inglesa y Alemana de la Universidad de Barcelona. El punto de partida técnico inicial fue que la herramienta informática fuera adaptable a las necesidades de todos los miembros de los proyectos Varcom y Pragmaestil (HUM2005-01936/FILO), el sucesor de Varcom, fácil de ejecutar y ligera de recursos.

Se optó por un entorno de programación para el SO Windows, Microsoft Visual Studio 2005. La integración de este entorno con los datos existentes y la facilidad de su ejecución en

las máquinas de los investigadores fue factor clave para su elección. El programa se ejecuta como aplicación Windows estándar.

Los ocho procedimientos descritos arriba generan casi todos ellos información en pantalla – la única excepción es punto 3, el preprocesamiento de las transcripciones. Por ello optamos aquí por presentar el conjunto de procesos siguiendo la misma estructura del procesamiento informático, de modo que una sola captura de pantalla (véase Ilustración I) nos permitirá describir los procesos de modo más detallado.

La pantalla se divide en tres grandes bloques; a la izquierda aparece en formato de tabla el listado de elementos encontrados (fase 5); a la derecha, el texto original seleccionado en cada momento con su concordancia (fase 4); y en la parte inferior, la lista de archivos que se están procesando en este momento (fase 1). El menú de la parte superior incluye el menú Archivo que permite seleccionar los archivos a procesar (fase 1), el menú Codificación (que permite seleccionar el archivo con el grupo de conectores que se van a aplicar, fase 2), el menú Archivos de Código (visualización y modificación de los archivos de conectores disponibles, fase 2) y un menú Ayuda.

2. FASES

2.1. *Fase 1. Selección de transcripciones*

Tras la selección de uno o varios archivos para procesar, el nombre de los mismos aparece en el listado de la parte inferior. Este listado constituye el corpus momentáneo con el que trabaja el programa. La selección de archivos se realizar con el método estándar del SO.

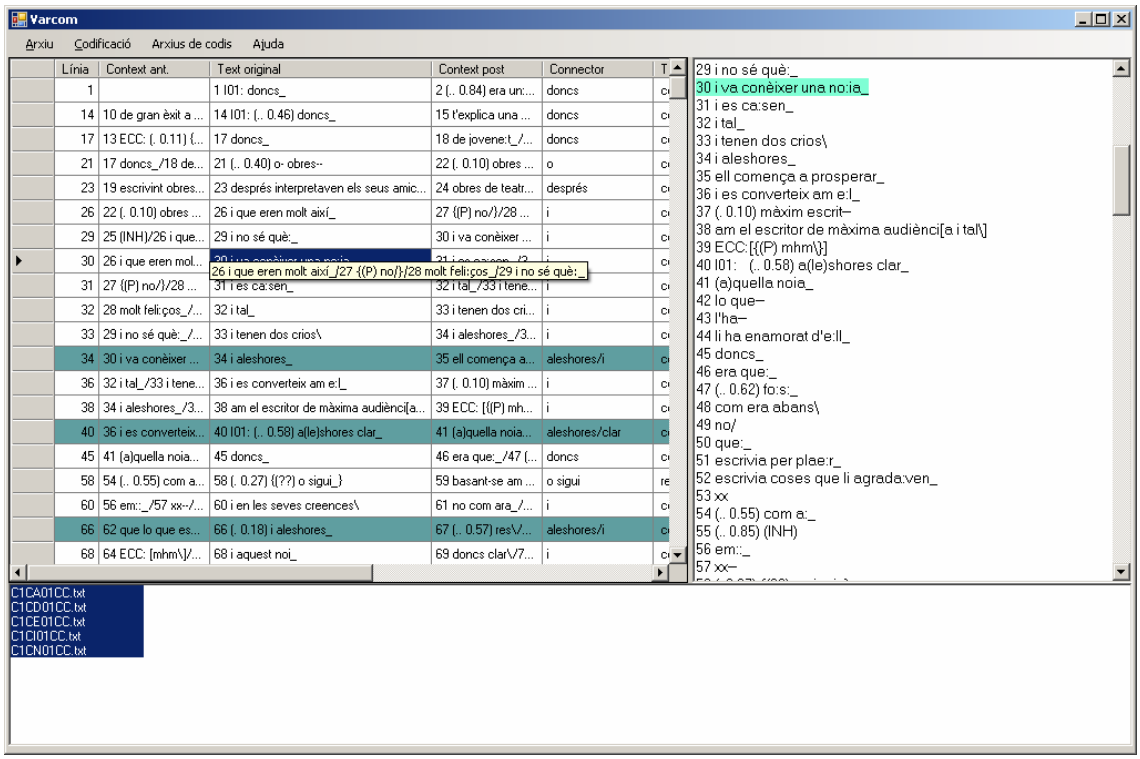

Ilustración I

2.2. *Fase 2. Selección del grupo de conectores*

Los conectores están agrupados como valores CSV en archivos de texto estándar. El archivo incluye una breve cabecera (que, a pesar de su apariencia, no es conforme XML) con la versión del archivo de conectores y la fecha en que fue elaborado. Ello ha de permitir realizar un seguimiento de los cambios introducidos. A cada conector le corresponde una línea y valores codificados de la siguiente manera: cadena literal del conector, tipos a los que puede adscribirse, número de elementos o cadenas separadas por espacios en blanco que contiene, e indicación sobre su valencia: "si" indica que es monovalente, "no" indica que es polivalente. Los conectores se ordenan por número de elementos que los forman y luego alfabéticamente.

```
<header>
<versio>1.0</versió>
<data>02.10.06</data>
</header>

<codis>
i ja està; final; 3; Si;
per dir-li d'alguna manera; reformulació; 3; Si;
a continuació; continua, cp; 2; Si;
a més; distribució, cl; 2; Si;
a part; altres; 2; Si;
a veure; pragmàtic; 2; Si;
encara que; conjunció; 2; Si;
en canvi; parentètic; 2; Si;
en concret; parentètic; 2; Si;
i ja; final; 2; Si;
i tal; continua; 2; Si;
mentres que; conjunció; 2; Si;
molt bé; acord; 2; Si;
```

Ilustración II

Cuando el usuario selecciona un archivo del menú Codificación, los conectores contenidos en este archivo se buscarán dentro de las transcripciones seleccionadas. Este proceso da paso a las fases 3 y 4.

2.3. *Fase 3 y 4. Preprocesamiento de las transcripciones / Identificación de las cadenas literales*

Las transcripciones de Varcom Catalán tienen un formato propio poco compatible con la identificación de cadenas literales. A cada unidad entonativa de un interlocutor le corresponde una línea de texto numerada. Dentro de la transcripción aparecen signos de entonación para marcar pausas y otras características orales. Estos elementos dividen las palabras e imposibilitan una comparación de cadenas directa.

```
10 de gran èxit a xxx--
11 a Estats--
12 Estats Units_
13 ECC:(. 0.11) {(P) mhm\}
14 I01:  (.. 0.46) doncs_
15 t'explica una mica el--
16 que va començar_
17 doncs_
18 de jovene:t_
```

19 escrivint obres de teatre_
20 (. 0.28) m:_
21 (.. 0.40) o- obres--
22 (. 0.10) obres que_
23 després interpretaven els seus amics_
24 obres de teatre\
25 (INH)
26 i que eren molt així_

Ilustración III

Por ello el procesamiento se ha dividido en los siguientes pasos: primero se abre la transcripción. El programa pasa la primera línea a memoria, elimina todos los signos hasta dejar únicamente cadenas de texto (separadas por espacio). En estas cadenas se busca primero el primero de los conectores (véase Ilustración II), luego el segundo, etc.

Si el programa no encuentra ninguno de los conectores, pasa a la línea siguiente. En cambio, si el programa encuentra un conector, lo pasa a una tabla (fase 5; incluye información adicional que referiremos más adelante) y elimina la cadena literal del conector de la línea que tiene en memoria, para evitar repeticiones de elementos: de este modo, se eliminan del texto en memoria primero los conectores de más elementos (razón por la que el archivo de conectores se ordena por número de elementos) y luego los de menos. Luego sigue recorriendo la lista de conectores en busca de más coincidencias. Si encuentra otro conector, lo pasa a la tabla y lo elimina de la memoria. De este modo, "y ya está" se identifica una vez, y cuando el programa busque la conjunción "y", ya no la encontrará en la línea que tiene en memoria y no podrá volver a marcarla.

Cuando ya no encuentra más conectores, el programa salta a la siguiente línea de transcripción y repite el proceso.

2.4. *Fase 5. Generación del listado de conectores encontrados*

Durante las fases anteriores (3 y 4) se han identificado los conectores presentes en las transliteraciones y se han enviado a la tabla que se ve en la Ilustración I. Además de la cadena literal, se envían otros datos a esta tabla: en primer lugar, el número de línea que se había eliminado durante la fase 3 para la identificación de las cadenas, pero también el texto original de la línea en la que se ha encontrado el conector, con todos los signos de entonación que la fase 3 descarta; la forma básica del conector; el potencial tipo al que pertenece y la indicación si ha sido desambiguado o no. Los conectores monovalentes (que sólo pertenecen a un tipo de conector y no tienen homónimos en otras clases de palabra) se marcan como desambiguados, y el investigador puede pasarlos por alto en un posterior proceso de desambiguación; en cambio, los conectores polivalentes se marcan como no desambiguados, y será precisa la intervención humana para determinar si son conectores o no o qué tipo de conector son.

2.5. *Fase 6: Desambiguación asistida*

Una vez terminadas las fases 3, 4 y 5, el investigador dispondrá ya de la información necesaria sobre la monovalencia del conector y las posibles categorías a los que pertenece el conector (estas dos informaciones estaban ya contenidas en el archivo de codificación de los conectores).

Para facilitar la tarea, en esta fase el programa añade el contexto a las unidades entonativas o líneas de las transcripciones. Al ser a veces extremadamente cortas (una o dos palabras) no permiten desambiguar funciones por sí solas. El programa recoge las cinco líneas anteriores y posteriores y las incluye en una celda antes y después de la línea. Cuando

no hay cinco líneas anteriores o posteriores (por ejemplo, porque el conector se encuentra en la segunda línea), sólo recoge como contexto el número de líneas disponible hasta un máximo de 5.

Otra función que se habilita en este momento es la de concordancia: al señalar con el ratón alguna línea de la tabla, el texto (véase a la derecha en la Ilustración I) presenta la línea seleccionada.

Si bien pueden parecer redundantes, las dos funciones descritas tienen usos diferenciados: la primera permite exportar la concordancia a otros formatos, como una hoja de cálculo o una base de datos; la segunda facilita el trabajo en pantalla.

2.6. *Fase 7: Identificación de acumulación de conectores*

Como paso casi final está la identificación de los dobletes o la presencia de dos conectores en una misma unidad entonativa. Al ser un fenómeno frecuente en la lengua hablada requiere de una especial atención para ver si se trata de estrategias de reformulación o de encadenamiento.

Durante este paso, el programa analiza la presencia de dos conectores en una misma línea en base a números de línea repetidos. Si en la tabla aparece dos o más veces el mismo número de línea, asume que se han encontrado dos conectores en una misma línea, y resalta las líneas correspondientes. Esta forma de destacar la presencia de dobletes no queda registrada para la posterior exportación de los datos, puesto que el procesamiento posterior con una base de datos o una hoja de cálculo presenta particularidades en las que no era preciso intervenir.

2.7. *Fase 8: Exportación de resultados para procesamiento posterior*

La exportación de los datos se planteó en un principio como relación dinámica con una base de datos, de modo que los añadidos a la tabla pasaran automáticamente a engrosar la base de datos. No obstante, este procedimiento limitaba la flexibilidad posterior de reestructuración de la base de datos en función de los intereses particulares del investigador. Y dado que los datos de la tabla están en un formato compatible, resultaba más indicado reducir el proceso de exportación a su forma más simple: el contenido de la tabla puede copiarse de forma manual (seleccionando toda la tabla) a una hoja de cálculo o una base de datos, donde luego los datos pueden transformarse libremente.

3. ARQUITECTURA ABIERTA

La aplicación Varcom Concord se ha escrito para la adaptación a un corpus concreto, el Varcom Catalán. Sin embargo, la estructuración interna en módulos (utilizando funciones) encargados de subprocesos concretos permite añadir, quitar o modificar fácilmente algún módulo de la cadena de procesos para adaptar el funcionamiento del programa a otro tipo de corpus.

Este mismo principio se ha aplicado a la estructura de los archivos de cadenas a buscar. La sustitución de las entradas relativas a los conectores por otra tipología libre permitirá buscar en el corpus listados cerrados de cadenas o grupos de cadenas (palabras y grupos de palabras). La detección de formas flexivas puede preverse igualmente en los correspondientes archivos de códigos.

4. DESARROLLO FUTURO

Varcom Concord, en su forma actual, realiza todas las funciones requeridas por los planteamientos de investigación iniciales. Un desarrollo futuro relativo al corpus Varcom será necesario cuando se produzca un cambio en los presupuestos iniciales.
Un desarrollo posterior independiente del corpus Varcom y totalmente abierto a la comunidad de estudios del corpus partiría necesariamente de una delimitación de funciones con respecto a *concordancers* genéricos (WordSmith, Monoconc…). En este sentido, la línea de desarrollo se basaría en la necesidad de detectar dentro de un corpus listados de palabras o grupos de palabras y desambiguar semiautomáticamente sus funciones dentro del texto.

BIBLIOGRAFÍA

Cuenca, M.J. (2001): *Estudi estilístic i contrastiu de l'arquitectura de l'oració. Estil segmentat vs. Estil cohesionat*. Círculo de Lingüística Aplicada a la Comunicación 7, septiembre 2001.

Fernández-Villanueva (2007), *Uses of* also. Catalan Journal of Pragmatics, en prensa

Hellwig, Birgit (2006): *EUDICO Linguistic Annotator (ELAN) version 2.6.*. [Documento de Internet disponible en http://www.mpi.nl/tools/elan.html, recuperado 08.01.2007]

Payrató, L.; Alturo, N.; Juanhuix, M. "Varcom project". BAAL/CUP Seminar on Multimodality and Applied Linguistics, University of Reading, 2003.

Payrató, Lluís; Fito, Jaume; Àlamo, Marina; Juanhuix, Marta (2005): "El projecte VARCOM". VII Jornada sobre la variació lingüística: Línies de recerca en lingüística i comunicació no verbal. Barcelona: Publicacions de la Universitat de Barcelona.

Paul Boersma & David Weenink (2005): *Praat: doing phonetics by computer (Version 4.3.14)* [Programa]. [Documento de Internet disponible en http://www.praat.or, Recuperado 08.01.2007]

Schmidt, Thomas (2004): "EXMARaLDA - ein System zur computergestützten Diskurstranskription." En: Mehler, Alexander / Lobin, Henning (Eds.) (2004): Automatische Textanalyse. Systeme und Methoden zur Annotation und Analyse natürlichsprachlicher Texte. Wiesbaden: Verlag für Sozialwissenschaften. 203-218