Sistema para el Etiquetado de Discursos Orales (AVOCES): selección y definición de categorías

Luis Carlos Toro Tamayo
Comisión Nacional de Investigación Científica y Tecnológica (Chile)

Resumen

Tras una caracterización sobre el discurso oral desde la compleja red de elementos y disciplinas que configuran su estudio, se presenta un software para etiquetado y transcripción de los discursos orales en contextos de ciencias humanas y jurídicas (AVOCES). Este sistema se complementa con un algoritmo de extracción de información que permite una búsqueda de términos por contextos discursivos específicos, facilitando los procesos de almacenamiento y recuperación de información para el investigador. El objetivo de la aplicación informática es proveer a los investigadores de un entorno asistido para el análisis del discurso oral desde técnicas automáticas de etiquetado y extracción más fiables, contribuyendo así no solo a la consolidación de una metodología más consistente de los estudios discursivos orales, sino, al mismo tiempo, a una mejor preservación de la tradición oral.

Palabras clave: Análisis del discurso oral. Ingeniería lingüística. Etiquetado textual.

Abstract

After a characterization of oral discourse from the point of view of disciplines that have it as its object of study, a new software program designed to tag and typescript oral discourses in the field of human sciences and law (AVOCES) is introduced. This system is complemented with an information extraction algorithm that allows a search for terms in specific discourse contexts, facilitating the storage and retrieval processes for the researcher. This computer application is aimed at providing researchers with a new tool for oral discourse analysis that features more reliable techniques of labeling and extraction. This software program may make an important contribution to the consolidation of an improved methodology in the field of oral discourse studies and, at the same time, to a better preservation of oral tradition.

Keywords: Oral discourse analysis. Linguistic engineering. Textual labeling.

1. Introducción

El presente trabajo tiene como objetivo presentar las decisiones tomadas en relación con la definición de categorías útiles para el análisis de los discursos orales, específicamente las vinculadas con el etiquetado y la transcripción de los dominios de las ciencias humanísticas y jurídicas. Dichas orientaciones están acompañadas de premisas teóricas y orientadas al desarrollo de un algoritmo de minería textual cuyo objetivo es organizar los procesos de sistematización y almacenamiento de dichos discursos (1).

La minería de texto es considerada como la extensión de la minería de datos y en la literatura se define como el proceso de descubrimiento de patrones interesantes y nuevos conocimientos en una colección de textos; es decir, la minería de texto es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen al relacionar el contenido de varios de ellos. Este proceso consiste de dos etapas principales: una de preprocesamiento y otra de descubrimiento (Tan, 1999). En la primera, los textos se transforman en algún tipo de representación estructurada o semiestructurada que facilita su posterior análisis, mientras que en la segunda se analizan las representaciones intermedias con el objetivo de descubrir en ellas algunos patrones interesantes o nuevos conocimientos. Dependiendo del tipo de métodos usados en la etapa de preprocesamiento variará el tipo de representación del contenido de los textos construidos, y, consiguientemente, este tipo de modelización determinará el tipo de patrones descubiertos.

En definitiva, se trata de un estudio multidisciplinar que intenta, mediante enfoques como el análisis del discurso y la tecnología lingüística, fortalecer las técnicas de sistematización de los discursos orales y su aplicación en campos como la historia, la antropología, la sociología, la lingüística y el derecho, entre otras. Creemos que es menester de los investigadores, y de los centros de investigación, idear métodos alternativos y más fiables de extracción y análisis de la información que nos permitan estudiar los discursos desde técnicas más sistemáticas, y a la vez construir documentos que contribuyan a la preservación de la tradición oral de nuestros antepasados.

2. Principios para la conformación de corpus orales

La configuración de corpus orales no consiste en la simple acumulación de materiales, sino en la reflexión sobre los aspectos inherentes a la representación de los datos. Dicho estudio ha sido definido con el concepto de etiquetado de los discursos orales, e implica un arduo proceso, máxime si tenemos en cuenta lo limitado de los tiempos en la investigación y que, entre transliterar, etiquetar y revisar el material, podemos ocupar diez veces el tiempo de las grabaciones.

Según Amparo Tusón Valls (2003), solo recientemente ha surgido un interés creciente por el estudio de los discursos orales. Dichas reflexiones se han convertido en centro de atención de diversas perspectivas científicas, entre las que destacan la lingüística, la comunicación y las ciencias sociales y humanas. Ya en lo que respecta a la fusión de disciplinas como la lingüística y la ingeniería de sistemas, podemos apreciar una evolución en el terreno de las tecnologías lingüísticas mediante el programa Euromap. Este estudio nos muestra la cantidad de información sobre el estado de investigación y desarrollo de las tecnologías lingüísticas en cada país miembro de la Unión Europea. Otras disciplinas y subdisciplinas, como la antropología, la sociología, la ciencia cognitiva, la filosofía, la sociolingüística, el análisis del discurso, la lingüística textual o la pragmática, comparten el interés por la interacción verbal, en aras de la definición de metodologías de procesamiento del lenguaje natural y la concepción de teorías específicas sobre el tema.

Y es que, por lo general, los registros orales tienen que ser escuchados y analizados desde una perspectiva interdisciplinar, atendiendo a puntos de vista como el control de las estructuras temáticas, la supervisión de la estructura sintáctica, la consideración de la superestructura, el nivel léxico, el control de las formas retóricas, la distribución de los turnos de habla, los roles de los participantes, los juegos de lo implícito y lo explícito en la interacción comunicativa, la ideología, el contexto, los aspectos prosódicos y semióticos, entre otros. Según Régis Debray (2001, p. 15), hablar no consiste solo en trasmitir un mensaje preparado con antelación y mediante una técnica adecuada, sino también en condensar la memoria colectiva de un grupo histórico que perpetúa a través de los años una "personalidad de base", y que forma parte del "momento de un proceso que será más largo y el fragmento de un conjunto más vasto".

Así, para diseñar e implementar el sistema AVOCES, pusimos énfasis en la transcripción ortográfica para pasar luego a un nivel de representación fonética mucho más profundo, tanto en el aspecto segmental como en el suprasegmental. También, planteamos la cuestión del etiquetado de los discursos, que, asociado a otras operaciones como la alineación temporal, permite una utilización posterior del corpus en diversas áreas del conocimiento.

Veamos algunos ejemplos de corpus disponibles en la Red que ilustran la manera en la que otros investigadores procesan la información documental para análisis posteriores. El primero citamos es el de la Real Academia Española, *Corpus de referencia del español actual* (CREA), que recopila, en soporte electrónico, todas formas de uso y manifestación del idioma español, desde sus orígenes hasta nuestros días. Quienes deseen comprender mejor el castellano pueden acceder al enlace web del corpus y realizar la búsqueda de una palabra o expresión, teniendo en cuenta criterios de selección generales como la cronología, el medio, la ubicación geográfica y el tema. Dicha exploración arroja una estadística del número de casos que

se encuentran disponibles y ejemplos que ilustran los usos más comunes de la expresión.

Figuras 1 y 2. Corpus de referencia del español actual (Real Academia Española).

Otro corpus interesante es el *Corpus del español*, creado por el profesor Mark Davies, de la Universidad Brigham Young. Dicho corpus funciona como un motor de búsquedas que se conecta a otros corpus más potentes y posee más de 100 000 000 palabras del castellano, desde épocas remotas hasta nuestros días. También permite efectuar búsquedas que hasta ahora no realizaba ningún otro corpus, como sinónimos, colocaciones, frecuencias y categorías gramaticales, además de sufijos, palabras personalizadas y búsquedas a partir de combinaciones de búsquedas más sencillas.

Figura 3. Corpus del español *(Mark Davies).*

Por su parte, el Centre de Llenguatge i Computació (CLIC) surge por iniciativa de un grupo de investigadores de la Universidad de Barcelona y orienta sus desarrollos en ingeniería lingüística hacia sectores productivos como el de las co-

municaciones, la edición y la enseñanza. Al igual que los dos corpus anteriores, procesa la información digitada y busca en bases de datos las palabras por coincidencia exacta o por su lema. Así mismo, determina las búsquedas por adjetivos, adverbios, determinantes, nombre común, nombre propio, verbo, pronombre, conjunción, interjección, preposición y cifras. Este recurso permite una relación del lenguaje con los sectores productivos y además define nuevos perfiles profesionales en el campo de los servicios, los recursos y la ingeniería lingüística, lo cual constituye una herramienta fundamental para la interacción comunicativa entre la tecnología y la sociedad.

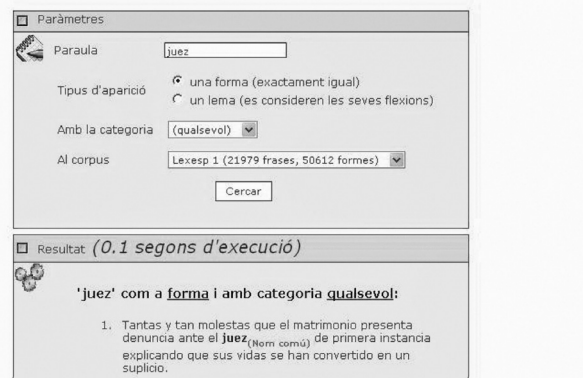


Figura 4. Interfaz del corpus del Centre de Llenguatge i Computació (CLIC).

Tengamos en cuenta que un corpus como el que requiere nuestra investigación está concebido como un conjunto de discursos orales transliterados y etiquetados debidamente por el autor en función de una hipótesis de trabajo predefinida. En este sentido, el etiquetado constituye un enriquecimiento de los discursos mediante información adicional que provee el investigador, con el objetivo de optimizar las búsquedas dado el problema de ruido y silencio de todo sistema de recuperación de información. El establecimiento de etiquetas constituye, pues, un significativo avance en la comprensión de los discursos orales y en la preservación misma de los materiales. No obstante, es necesario definir un sistema de codificación que sirva de marco de referencia normalizador para que otros investigadores comprendan y apliquen correctamente los signos propuestos.

3. Presupuestos y autores de referencia

Tal y como mencionamos anteriormente, los discursos orales tienen tantas particularidades, distintas a las del discurso escrito, que son muchas las ocasiones en

las que el alfabeto no nos alcanza para interpretarlas. Las diferencias se observan no solo en lo referente a la variación fonológica, sino en la segmentación de los signos, las interrupciones, las acotaciones, los silencios, etcétera. Es aquí donde se hace indispensable la definición de categorías de análisis con atributos homogéneos para la gran mayoría de investigadores, de modo que se logre una correcta interpretación de los discursos orales y una conservación fidedigna de las transcripciones y registros discursivos, ya sea en formato impreso o digital.

Esto ha generado una gran polémica entre quienes piensan que la puntuación traiciona al texto oral y que, por consiguiente, no solo afecta su sentido, sino que puede conducir a un análisis equivocado (Blanche-Benveniste, 1998). Por tal motivo, el objetivo de la presente investigación consistió en desarrollar un sistema que permitiera hacer análisis más acertados de discursos orales, a partir de la comparación del original y la interpretación realizada por el investigador. Para ello diseñamos dentro de nuestro sistema un algoritmo especializado en la extracción de los datos.

El propósito del producto de la investigación consistió en integrar en un solo ambiente todas las herramientas necesarias para transcribir discursos orales tales como el video, los signos de transcripción, las voces que en este interactúan y las características propias de un editor de texto, teniendo la posibilidad de generar registros de análisis de la información. Además de poder administrar los proyectos y organizar las entrevistas pertinentes dentro del proceso de los discursos orales.

La metodología empleada en el desarrollo del sistema fue el Custom Development Method (CDM), que consiste en ejecutar una serie de actividades que se especifican como componentes de un entregable. Un entregable, por lo tanto, es un documento donde se consignan los resultados del trabajo correspondiente a una fase del desarrollo de software. Cada entregable deben ser diligenciado por sus responsables inmediatamente después de la culminación de la fase, porque en ese punto se necesita evaluar el desarrollo o la adaptación de software, aplicando ciertas métricas de calidad que pueden variar según la fase que se está llevando a cabo. CDM también trata de impedir la generación de código para la construcción del software hasta que no se tenga una idea muy precisa de lo que se necesita de él. Por lo tanto, todo desarrollo o adaptación de software debe partir de una clara definición de los requerimientos del nuevo sistema informático, tanto de información como de comportamiento.

El desarrollo del software tuvo en cuenta fases como la definición, el análisis y el diseño, en las que se modeló el sistema con la propuesta de Lenguaje de Modelado Unificado (UML). También se utilizaron herramientas de ingeniería del software asistido por computador, conocidas como *herramientas CASE*, entre las cuales se tuvieron en cuenta Visual Paradigm, la cual facilita el diseño gráfico para la implementación y desarrollo del sistema, y Project, que permite controlar, definir y medir las tareas que elabore el sistema.

Además se diseñó la arquitectura del sistema inicial, especificando los recursos de hardware y de software necesarios para el desarrollo. Una vez terminada esta etapa, se procedió al diseño de los módulos y la base de datos de la aplicación utilizando las técnicas de UML y basándose en los requerimientos establecidos por el usuario, de lo cual queda constancia en los diagramas de casos de uso, clases, objetos, estados, actividades, secuencia y colaboración. Posteriormente, y para finalizar, se validó la aplicación evaluando su funcionamiento con un repositorio de los investigadores.

Dentro de las pautas generales que se han de considerar en el tema de la transliteración de las conversaciones y sobre los aspectos fonéticos del discurso conversacional tuvimos en cuenta los estudios llevados a cabo por Sacks, Schegloff y Jefferson (1974), entre otros. Por lo que respecta al análisis del discurso, retomamos los planteamientos de Helena Calsamiglia y Amparo Tusón (2002), Violette Morin (1969, 1974), Teun A. van Dijk (1990, 1991, 1995, 1997, 1998, 2000, 2003), J. Renkema (2000), F. van Eemeren et álii (2000), Anthony Weston (1994), Ruth Wodak (2003) y Norman Fairclough (2003), entre otros. Y todo ello porque para comprender los discursos orales debemos tener en cuenta el conjunto de personas e informaciones que participan en la conversación, en la que aparecen diversos grados de información dependiendo del lugar y las circunstancias sociales, políticas, económicas, etcétera, y hacer una clara distinción entre el texto y su contexto, tal y como lo plantea el método de *análisis de contenido* (Morin, 1969, 1974), y el enfoque de *análisis crítico del discurso* (ACD) (Van Dijk, 1990, 1991, 1995, 1997, 1998, 2000, 2003).

De este modo, análisis del discurso, recuperación de la memoria oral, ideología, cultura y sociedad forman parte de un terreno común de la investigación que merece ser estudiado bajo parámetros sistemáticos por sociohumanistas e ingenieros, a fin de garantizar la validez de las afirmaciones realizadas en las investigaciones y comprobar la legitimidad de las mismas con elementos de juicio, respaldados por la observación empírica.

4. Pasos que debemos cumplir en el etiquetado de discursos orales

Si bien existen diferentes propuestas de categorías para el etiquetado de los corpus orales, entre ellas las que proponen autores como Sacks, Schegloff y Jefferson (1974), Gumperz y Berenz (1990), Du Bois (1991), Payrató (1995), Llisterri (1999), Briz y el Grupo Val.Es.Co (2000), retomamos las orientaciones hechas por Amparo Tusón Valls (2003), quien propuso un método de transcripción con el objetivo de que los materiales orales fuesen objetos manejables, susceptibles de un tratamiento analítico. Según Payrató (1995, p. 45), la transcripción es un procedimiento de adaptación de una producción lingüística oral a una forma gráfica escrita. En este

procedimiento participan tanto la transliteración o transcripción ortográfica como la transcripción fonética, que se lleva a cabo mediante signos diseñados para tal propósito. No obstante, dicha traducción requiere un complejo proceso de reflexión y análisis, en el que se consideran aspectos como el contexto, cuyo significado social depende de la actividad humana, y el entorno comunicativo, entre otros. La competencia comunicativa es, pues, un campo de estudio que ha ido avanzando y que apunta más al desarrollo de los aspectos culturales y situacionales del uso del lenguaje que a las estructuras gramaticales y al vocabulario de los hablantes.

Así pues, entre las categorías de análisis que utilizamos en nuestro Sistema para el Etiquetado de los Discursos Orales tuvimos en cuenta aspectos concernientes a los códigos lingüísticos, paralingüísticos y extralingüísticos (Niño Rojas, 2003, pp. 57-59), justificados en los métodos técnicos de entrevistas que utilizan los investigadores, entre ellos los judiciales. Los códigos lingüísticos son en esencia la lengua o el idioma, constituidos por signos o reglas propios de la gramática de cada habla. Dichos signos son utilizados en el sistema de etiquetado para registrar los elementos verbales del discurso. Es así como, además del discurso emitido por el entrevistado, en el que se registran todas las palabras expresadas por él o los sujetos que intervienen, se tienen en cuenta aspectos como el tema tratado, el lugar donde fue realizada la entrevista, la fecha y la duración, entre otras. En la tabla I se explica cada categoría.

La utilidad de estas categorías radica en la funcionalidad de los descriptores, que ubican al investigador en el contexto de la acción comunicativa y permiten almacenar la información con el rigor y detalle que merece la fuente. La descripción formal y de contenido de los documentos es un proceso importante en las llamadas *ciencias de la documentación*, porque permite administrar mejor la información documental y construir bases de datos electrónicas de fácil reconocimiento. Los analistas del discurso y los académicos que buscamos información en fuentes primarias y secundarias reconocemos el valor de la exactitud en el procesamiento de los datos y sabemos que un documento mal clasificado significa un obstáculo para la investigación. Ahora, si bien los signos lingüísticos operan en niveles como el fónico, el léxico, el semántico y el sintáctico, su objetivo fundamental es evidenciar, o descubrir, los problemas que el lenguaje plantea como medio de relación social. Por su parte, los signos paralingüísticos incluyen recursos prosódicos estrechamente relacionados con los signos lingüísticos pero que apuntan más a la interpretación de aspectos relacionados con la curva melódica de la entonación. La tabla II explica mediante signos cada uno de los códigos que necesitaremos para la transcripción.

Según Pierre Guiraud, citado por Niño Rojas (2003, p. 58), existen tres tipos de códigos del lenguaje que facilitan la interpretación del lenguaje verbal: los relevos, los sustitutos y los auxiliares.

Código o signatura	Referencia dada al documento en el momento de la sistematización según el fondo documental que se le asigne; el código estará conformado por la fecha de realización de la entrevista, la temática, el número consecutivo de la entrevista y el código del investigador
Autor o entrevistador	Autor del documento: director-realizador
Situación comunicativa	Tema: asuntos o materias que representan el contenido del documento (véase Sistema de Clasificación Decimal Dewey, 2000)
	Propósito: objetivo previsto en el momento de entablar una comunicación
Voz principal	Nombre de la persona que interviene en la grabación como protagonista
Género	Masculino o femenino
Edad	≤ 25, 26-55, > 55
Idioma	Lengua materna o dialecto del hablante
Nivel de estudios	Analfabetos, primarios, secundarios, medios, superiores
Nivel sociocultural	Alto, medio, bajo
Otras voces	Nombre(s) de otra(s) persona(s) que interviene(n) en la grabación y cuya voz se escucha en el documento; en esta categoría deben ir definidas las iniciales del nombre o el oficio para caracterizar al informante, todo con el objetivo de abreviar en el momento de la transcripción
Lugar de la grabación	País, departamento o estado, distrito, municipio o cantón, otros
Fecha de la grabación	Datos exactos del día en que fue grabada la entrevista
Duración de la grabación	Tiempo de la entrevista registrado en horas, minutos y segundos; el sistema permitirá al investigador visualizar el tiempo de la grabación mientras transcribe y etiqueta, con el objetivo de hacer un registro fiel de los momentos de la grabación
Soporte de la grabación	Tipo de soporte, tipo de grabación (mono, estéreo, velocidad de reproducción)
Resumen	Dado que el documento sonoro no es directamente accesible sino a través de su reproducción y audición, un resumen que recoja lo esencial del contenido es de gran ayuda para la recuperación y elección del documento buscado

Tabla I. Registro de proyecto. Ficha técnica (Moreiro, 2000, p. 326).

Los códigos de relevos son aquellos que no cambian el significado del lenguaje pero alteran la forma de representarlo. Ejemplos de estos códigos son el sistema braille, el morse y el lenguaje de los sordomudos, entre otros, los cuales permiten recodificar con bastante fidelidad el lenguaje fónico, introduciendo peculiaridades lingüísticas muy significativas propias de la lengua escrita.

Por su parte, los códigos sustitutos del lenguaje son aquellos que no representan el lenguaje, sino que pretenden reemplazarlo. Un claro ejemplo de este tipo de escritos lo encontramos en los ideogramas chinos, donde "los caracteres no representan los signos de la lengua, sino ideas y conceptos" (Niño Rojas, 2003, p. 58). Esto ocurre también con los jeroglíficos, las pinturas y las siglas.

Signos de transcripción	
¿?	Entonación interrogativa
¡!	Entonación exclamativa
<	Tono ascendente
>	Tono descendente
=	Tono sostenido o continuo
...-	Corte abrupto en medio de una palabra
I	Pausa breve
II	Pausa mediana
<...>	Pausa larga, también <pausa> o <09>, indicando el número de segundos que se tarda en hablar
†	Tono agudo
‡	Tono grave
Ac	Ritmo acelerado
Le	Ritmo lento
MAYÚS	Mayor énfasis: la palabra será escrita en mayúsculas, con el objetivo de señalar el tono enfático
Aaa	Alargamiento vocálico
Nnn	Alargamiento consonántico
P	Piano (dicho en voz baja)
Pp	Pianísimo (dicho en voz muy baja)
F	Forte (dicho en voz más alta)
Ff	Fortísimo (dicho en voz muy alta)
Signos relativos a los turnos de palabras	
= =	Al principio de un turno, indica que no ha habido pausa después del turno anterior
=...=	Solapamiento de dos turnos

Tabla II. Signos necesarios para la transcripción (Tusón, 2003, pp. 100-101).

Finalmente tenemos los códigos auxiliares del lenguaje, que acompañan la significación de las palabras según sea la entonación y la expresividad corporal. Este último factor es comúnmente conocido como *código kinésico y proxémico*, e incluye aspectos como los gestos, el movimiento de las manos, la mirada, la utilización del medio ambiente, etcétera. En la tabla III aparecen signos con las características antes descritas, que obviamente deben ser advertidos por el investigador.

A esta última categoría, kinésica y proxémica, la denominamos *código extralingüístico*. Tiene que ver con la capacidad de sentir, el nivel sensorial, y está relacionada estrechamente con el contexto comunicativo y con aspectos como la cultura y la ideología del hablante. Los códigos extralingüísticos nacen de la ex-

[no léxicos]	Fenómenos no léxicos, tanto vocales como no vocales; por ejemplo, [risas], [A mirando a B], [llanto], [duda]...
(ininteligibles)	Palabras ininteligibles o dudosas que deben ser escritas entre paréntesis tal cual suenan

Tabla III. Otros signos que facilitan la interpretación del lenguaje verbal (Tusón, 2003, p. 101).

periencia objetiva, subjetiva y cultural del ser humano, y tiene como propósito significar la relación entre los hombres. Todos estos códigos tienen en común su carácter eminentemente semiótico.

> La semiosis incluye todas las formas de creación de significado —las imágenes, el lenguaje corporal y también el lenguaje—. Podemos entender la vida social como una serie de redes interconectadas de prácticas sociales de diferentes tipos. Todas las prácticas son prácticas de producción que constituyen los escenarios en los que se produce la vida social, ya sea esta económica, política, cultural o de carácter cotidiano. (Fairclough, 2003, pp. 179-180)

5. Software AVOCES

La aplicación permite a los investigadores transcribir discursos orales en un entorno integrado de video y editor de texto, con opciones como análisis de frecuencia de las palabras y búsqueda por contexto. Tal y como se aprecia en la imagen, el sis-

Figura 5. Portada del programa Avoces.

tema contiene dos interfaces: una para el administrador y otra para el investigador. De un lado está el administrador, quien se encarga de registrar y controlar las acciones de los investigadores a la vez que examina el ingreso adecuado de los discursos orales. De otro lado está el investigador, quien posteriormente al registro efectuado por el administrador, debe etiquetar y transcribir las entrevistas, conservando todas las especificaciones dadas por el sistema.

El rigor de la fuente oral exige que no se omita ninguna de estas consideraciones, con el ánimo de lograr una buena preservación de la memoria oral. Existen dos opciones para que los investigadores registren los proyectos: una general, para discursos orales de investigaciones del área sociohumanística, y otra específica del dominio del derecho judicial.

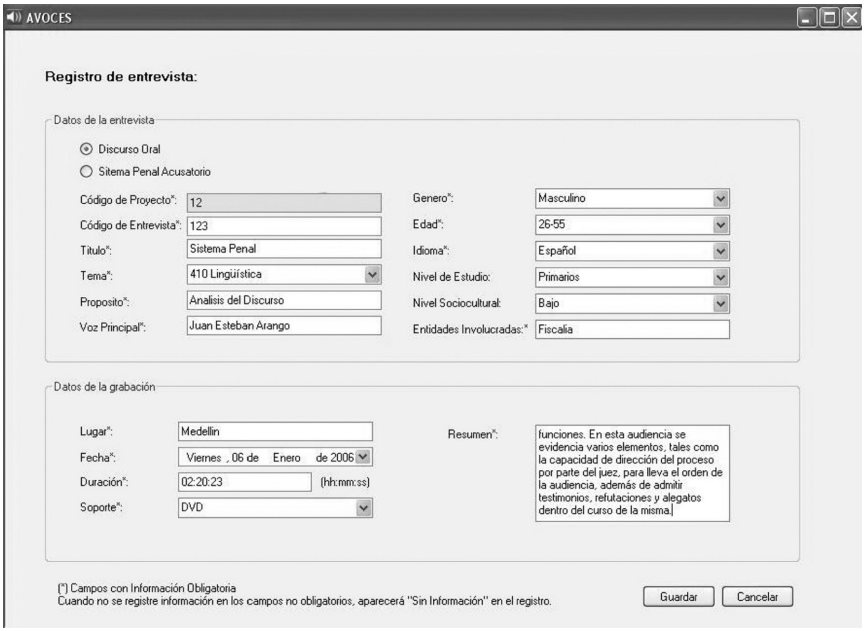


Figura 6. Ventana de registro de la entrevista.

Tal y como ya mencionamos, en caso de que la intervención oral sea una audiencia pública el investigador deberá registrar datos legales en el cuadro que vemos en la figura 7.

Una vez registrados el proyecto y la entrevista, ingresamos en el editor de texto. El editor es quizás la parte más dinámica del sistema, porque incorpora las entre-

Figura 7. Ventana de registro de datos legales de la entrevista.

vistas realizadas en video o audio y la transliteración de las mismas. Además permite al investigador realizar una serie de marcas prosódicas, muy propias del discurso hablado, que servirán posteriormente para el análisis que se quiere realizar.

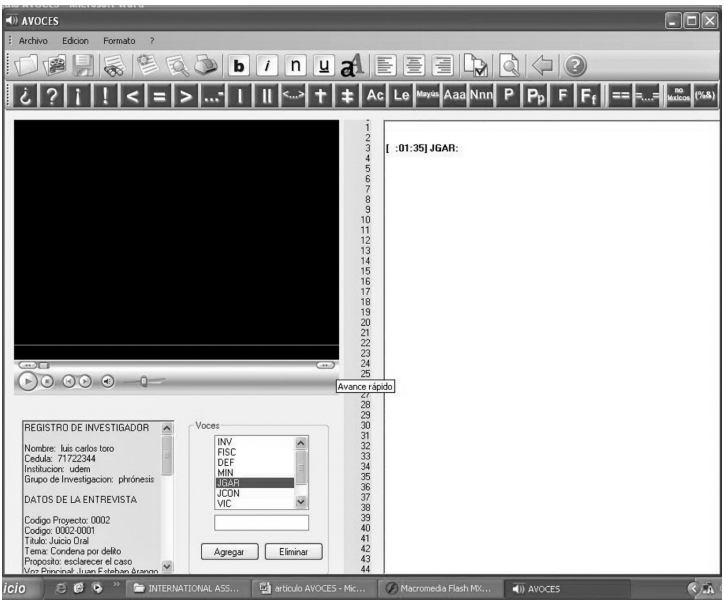
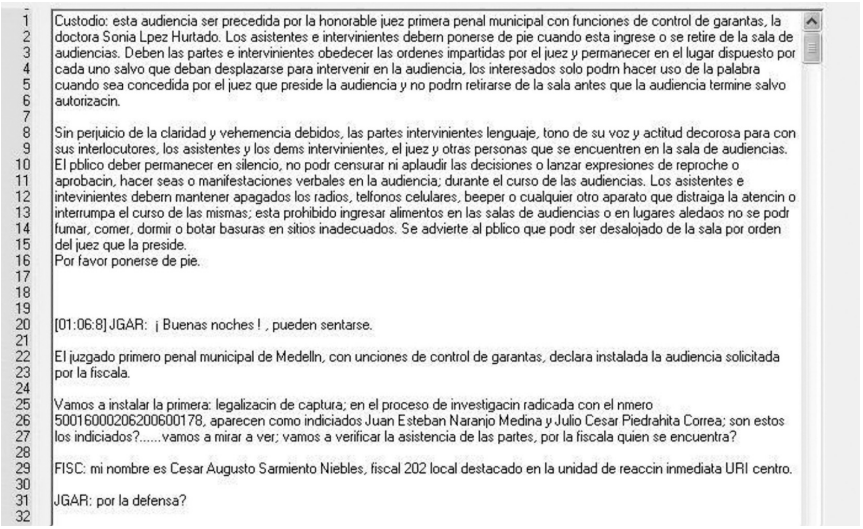


Figura 8. Ventana del editor de texto.

A la hora de transcribir el texto el investigador debe tener en cuenta las etiquetas descritas en el apartado 4, denominadas *signos de transcripción.* Tal y como se aprecia en la figura 9, la información está debidamente cronometrada y numerada, es decir, se identifica el minuto aproximado en el que intervinieron los hablantes y la línea en la que dicha información fue transcrita. Esta sistematización tiene el objeto de facilitar la búsqueda de la información en los casetes y en el texto.


Custodio: esta audiencia ser precedida por la honorable juez primera penal municipal con funciones de control de garantas, la doctora Sonia Lpez Hurtado. Los asistentes e intervinientes debern ponerse de pie cuando esta ingrese o se retire de la sala de audiencias. Deben las partes e intervinientes obedecer las ordenes impartidas por el juez y permanecer en el lugar dispuesto por cada uno salvo que deban desplazarse para intervenir en la audiencia, los interesados solo podrn hacer uso de la palabra cuando sea concedida por el juez que preside la audiencia y no podrn retirarse de la sala antes que la audiencia termine salvo autorizacin.

Sin perjuicio de la claridad y vehemencia debidos, las partes intervinientes lenguaje, tono de su voz y actitud decorosa para con sus interlocutores, los asistentes y los dems intervinientes, el juez y otras personas que se encuentren en la sala de audiencias. El pblico deber permanecer en silencio, no podr censurar ni aplaudir las decisiones o lanzar expresiones de reproche o aprobacin, hacer seas o manifestaciones verbales en la audiencia; durante el curso de las audiencias. Los asistentes e intevinientes debern mantener apagados los radios, telfonos celulares, beeper o cualquier otro aparato que distraiga la atencin o interrumpa el curso de las mismas; esta prohibido ingresar alimentos en las salas de audiencias o en lugares aledaos no se podr fumar, comer, dormir o botar basuras en sitios inadecuados. Se advierte al pblico que podr ser desalojado de la sala por orden del juez que la preside.
Por favor ponerse de pie.

[01:06:8] JGAR: ¡ Buenas noches ! , pueden sentarse.

El juzgado primero penal municipal de Medelln, con unciones de control de garantas, declara instalada la audiencia solicitada por la fiscala.

Vamos a instalar la primera: legalizacin de captura; en el proceso de investigacin radicada con el nmero 500160002062006001 78, aparecen como indiciados Juan Esteban Naranjo Medina y Julio Cesar Piedrahita Correa; son estos los indiciados?.....vamos a mirar a ver; vamos a verificar la asistencia de las partes, por la fiscala quien se encuentra?

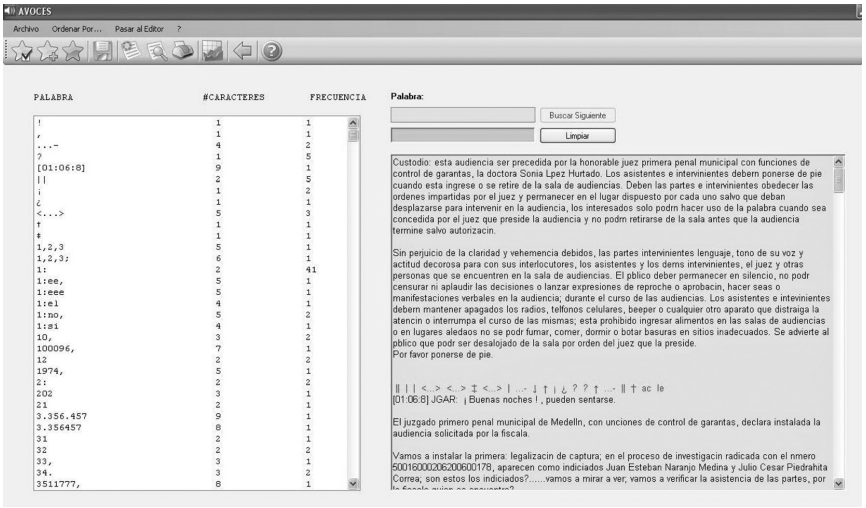
FISC: mi nombre es Cesar Augusto Sarmiento Niebles, fiscal 202 local destacado en la unidad de reaccin inmediata URI centro.

JGAR: por la defensa?

Figura 9. Ejemplo de transcripción de texto.

Los análisis que se derivan de las entrevistas transcritas son el resultado de una lectura cuidadosa, tal y como observamos en la interfaz de las figuras 10 y 11, que permite al investigador ver la ubicación de las palabras utilizadas en el discurso por orden alfabético, por frecuencia y por número de caracteres. Esto a su vez hace posible marcar las palabras de interés y obtener un informe de usos y coincidencias prosódicas del discurso hablado.

Finalmente, un análisis por ubicación permite ver los microcontextos en los que se encuentra el concepto de interés para el investigador. Dicho análisis se hace con el objetivo de buscar coincidencias en un discurso o una serie de discursos orales, lo que hace posible un estudio comparado de fenómenos léxicos y/o fisuras discursivas.



DEF: lo que pasa es que como el se encuentra detenido en las instalaciones del CTI, se coloc esa direccin.

JGAR: bueno, de todas maneras qued en el registro su direccin entonces. Ministerio Publico.

MIN: mi nombre es Cesar Maya Toro, personero delegado en lo penal, cedula de ciudadana 71.583.095.

JGAR: la solicitud de la fiscala es por el delito de hurto calificado y agravado, porte ilegal de arma de fuego, tentativa de homicidio, emmm, como dije, procedo a instalar la primera audiencia, legalizacin de captura; proceda seor fiscal, si es tan amable.

FISC: gracias seor juez... estamos presentes en esta audiencia con el fin de dar control de legalidad a la captura del seor Juan Esteban Naranjo Medina quien se identific a los policas con la cdula de ciudadana. 8.025.854 de Medelln, natural de Medelln, Antioquia, nacido el 8 de mayo de 1974, 21 aos de edad, soltero, hijo de Gilberto y Piedad, residente carrera 65 d 32 almu 34. barrio Beln Ftima, telfono 3511777, es la direccin que l suministra en los datos de registro odontolgico, toda vez que como se explicara mas adelante en acta de derechos del capturado, eeee da el de otra persona..

El da de ayer, 6 de enero, 9:30 de la maana, el seor Jairo Enrique Ftima Gaviria Velsquez, conduca su vehculo mazda matsuri, por el sector de Buenos Aires, sin embargo en su recorrido, llega una camioneta ford explorer, es atravesada a su vehculo, le hurtan el mismo y mas adelante un ciudadano en una motocicleta observa los hechos que se estn aconteciendo, procede a verificar y a seguir a donde llega la ford explorer e inmediatamente se devuelve y da aviso a la polica, sin embargo la fiscala quiere para que su seora tenga mas conocimiento real de los hechos de que acontecieron en el da de ayer a las 9:30 de la maana, se llame a declarar eee, en este estrado se encuentra tanto el seor Jairo Enrique Ftima Gaviria, como el agente de la polica Carlos Mario Parra Pulgarin que es el que realiza la captura, entonces con respeto, solicito se llame a declarar en primera instancia para saber la narracin de los hechos como tal al seor Jaime Enrique Ftima Gaviria, identificado con cedula de ciudadana 3.356.457 de Medelln.

Figuras 10 y 11. Ventana de análisis de frecuencias y detalle de la localización de ocurrencias en un texto.

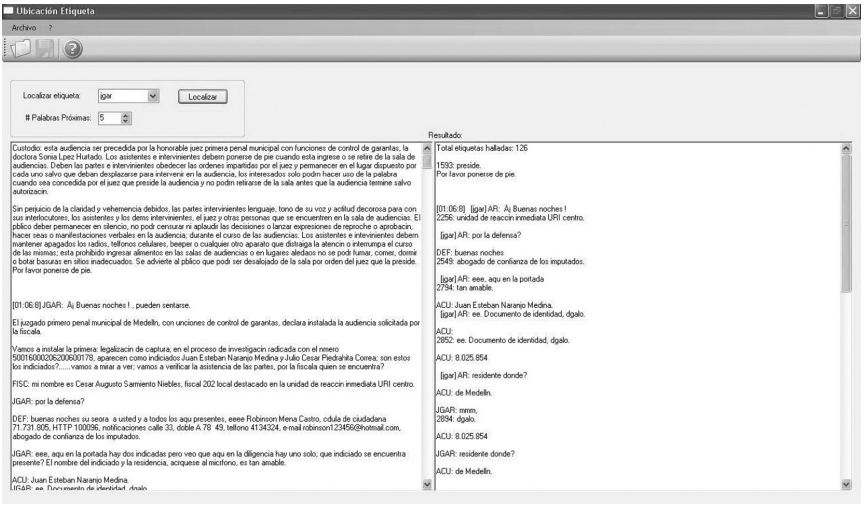


Figura 12. Ventana de análisis de microcontextos.

6. Conclusiones

Si bien todavía es necesario que los discursos orales sean transcritos manualmente, es decir, que pasen por el filtro de la textualidad para un mejor reconocimiento, traducción, resumen y clasificación, debemos reconocer que el etiquetado constituye un avance significativo en el estudio de la interacción comunicativa. Ello amplía nuestra perspectiva sobre el tema tratado, hasta el punto de atrevernos a plantear que el marcado de los discursos es la primera fase del análisis. En este sentido, este sistema permite a los investigadores del área sociohumanística y jurídica precisar aspectos relativos al lenguaje, que sirven como soporte para consolidar un proceso encaminado al análisis del discurso oral en Colombia. Además, dicho avance tecnológico permite conservar todo tipo de manifestaciones orales en soporte magnético o digital, lo que conlleva necesariamente una redefinición de la forma de contar el pasado y procesar la información.

Notas

(1) El proyecto Sistema para el Etiquetado de Discursos Orales surge en 2005 por iniciativa de los grupos de investigación de la Universidad de Medellín Phrónesis, del Departamento de Ciencias Sociales y Humanas, y Arkadius, de la Facultad de Ingeniería. Actualmente forman parte del proyecto el historiador y magíster en Lingüística Luis Carlos Toro Tamayo (investigador principal), las magíster en Ingeniería de Sistemas Diana María Montoya Quintero e Idanis Beatriz Díaz Bolaño (coinvestigadoras) y las estudiantes Catalina Peláez Escobar y Laura María Pareja Georges (auxiliares), de la

Facultad de Derecho y el Programa de Ingeniería de Sistemas respectivamente. La fuente de financiación del proyecto proviene de fondos destinados por la Vicerrectoría de Investigación de la Universidad de Medellín, Colombia.

Referencias

Bazalar Díaz, Jacqueline M. (2006). Características y diferencias entre el código lingüístico oral y el código lingüístico escrito. // El lenguaje. http://www.monografias.com/trabajos32/lenguaje/lenguaje.shtml#codigo (2006-09-04).

Blanche-Benveniste, Claire (1998). Estudios lingüísticos sobre la relación entre oralidad y escritura. Barcelona: Gedisa, 1998.

Briggs, Asa; Burke, Peter (2002). La revolución de la imprenta en su contexto. // De Gutenberg a Internet: una historia social de los medios de comunicación. Madrid: Taurus, 2002.

Briz, Antonio, y Grupo Val.Es.Co (2000). ¿Cómo se comenta un texto coloquial? Barcelona: Ariel ("Ariel Practicum"), 2000.

Booch, G.; Rumbaugh, J.; Jacobson, I. (2002). El Lenguaje Unificado de Modelado (UML). Madrid: Addisson Wesley, 2002.

Civallero, Edgardo (2006). Tradición oral: herramientas y experiencias sobre oralidad, su revitalización, recolección, gestión y difusión. http://www.tradicionoral.blogspot.com/ (2006-08-01).

Calsamiglia, Helena; Tusón, Amparo (2002). Las cosas del decir: manual de análisis del discurso. Barcelona: Ariel, 2002.

Debray, Régis (1997). Transmitir. Buenos Aires: Manatial, 1997.

Debray, Régis (2001). Introducción a la mediología. Barcelona: Paidós, 2001.

Du Bois, John J. (1991). Transcription design principles for spoken discourse research. // Pragmatics. 1 (1991) 71-106.

Dürkeim, Émile (1993). Las reglas del método sociológico. Madrid: Morata, 1993.

Fairclough, Norman (2003). El análisis crítico del discurso como método para la investigación en ciencias sociales. // Wodak, Ruth; Meyer, Michael (comps.). Método de análisis crítico del discurso. Barcelona: Gedisa, 2003.

Febvre, Lucien (1993). Combates por la historia. Barcelona: Planeta Agostini ("Obras Maestras del Pensamiento Contemporáneo", 28).

González, Luis (1972). El arte de la microhistoria. Ponencia presentada al Primer Encuentro de Historiadores de Provincia (San Luis Potosí, 26 de julio de 1972). http://biblioteca.redescolar.ilce.edu.mx/sites/fondo2000/vol1/otra-invitacion/html/1.html (2006-05-16).

Guiraud, Pierre (1971). La semiología. México: Siglo XXI, 1971.

Gumperz, John J.; Berenz, Nadine (1990). Transcribing conversational exchanges. // Edwards Jane A.; Lampert, Martin D. (eds.). Talking Data: Transcription and coding in discourse research. Hillsdale, NJ: Lawrence Erlbaum ("Berkeley Cognitive Science Report", 63), 1990. 91-121.

Habilidades comunicativas del defensor en el juicio oral. Colombia. Unidad de Capacitación, Dirección Nacional de Defensoría Pública. Financiado por el Programa de Fortalecimiento y Acceso a la Justicia.

Halliday, M. A. K. (1989). Spoken and written language. Oxford: Oxford University Press, 1989.

Lyons, J. (1973). Introducción a la lingüística teórica. Barcelona: Teide, 1973, 20.ª ed.

Lyons, J. (1980). Semántica. Barcelona: Teide, 1980.

Llisterri, J. (1999). Transcripción, etiquetado y codificación de corpus orales. // Revista Española de Lingüística Aplicada. Monográfico: Panorama de la Investigación en Lingüística Informática. (1999) 53-82.

Moreiro, José Antonio (2000). Manual de documentación informativa. Madrid: Cátedra, 2000.

Morin, Violette (1969). El análisis de la prensa. París / La Haya, Mouton, 1969.

Morin, Violette (1974). Tratamiento periodístico de la información. Madrid: ATE ("Colección de Libros de Comunicación Social").

Niño Rojas, Víctor Miguel (2003). Competencias en la comunicación. Bogotá: Eco, 2003.

Ong, J. Walter (1994). Oralidad y escritura: tecnologías de la palabra. México: Fondo de Cultura Económica, 1994.

Payrató, Lluís (1995). Transcripción del discurso coloquial. // El español coloquial. Actas del Primer Simposio sobre Análisis del Discurso Oral. Almería: Universidad de Almería, 1995. 45-70.

Renkema, Jan (2000). Introducción a los estudios sobre el discurso. Barcelona: Gedisa ("Cladema Lingüística"), 2000.

Sacks, Harvey; Schegloff, Emmanuel A.; Jefferson, Gail (1974). A simplest systematics for the organization of turn-taking in conversation. // Language. 50 (1974) 731-734.

Searle, John (1980). Actos de habla. Madrid: Cátedra, 1980.

Simone, Raffaele (2001). La tercera fase: formas de saber que estamos perdiendo. Madrid: Taurus ("Pensamiento"), 2001.

Sistema de Clasificación Decimal Dewey (2000). Bogotá: Rojas Eberhard, 2000.

Tan, A. H. (1999). Text mining: the state of the art and the challenges. // PAKDD'99 Workshop on Knowledge Discovery from Advanced Databases. Pekín, 1999. 65-70.

Toro T., Luis Carlos (2006). Análisis argumentativo, retórico y semiótico de discursos publicitarios en medios impresos: el caso de las tecnologías de comunicación masiva en Colombia. Trabajo de grado dirigido por Fabio Mejía Fernández. Medellín: 5.ª Cohorte Maestría en Lingüística, Facultad de Comunicaciones de la Universidad de Antioquia, 2006.

Tusón Valls, Amparo (2003). Análisis de la conversación. Barcelona: Ariel ("Ariel Practicum"), 2003.

Van Dijk, T. A. (1990). La noticia como discurso: comprensión, estructura y producción de la información. Barcelona: Paidós, 1990.

Van Dijk, T. A. (1991). Las estructuras y funciones del discurso. México: Siglo XXI, 1991, 7.ª ed.

Van Dijk, T. A. (1995). Prensa, racismo y poder. México: Universidad Iberoamericana, 1995.

Van Dijk, T. A. (1997). Racismo y análisis crítico de los medios. Barcelona: Paidós, 1997.

Van Dijk, T. A. (1998). Ideología: un enfoque multidisciplinario. Barcelona: Gedisa, 1998.

Van Dijk, T. A. (2000) (comp.). El discurso como interacción social. Barcelona: Gedisa, 2000.

Van Dijk, T. A. (2003). Ideología y discurso: una introducción multidisciplinaria. Barcelona: Ariel, 2003.

Van Eemeren, F. H., et álii (2000). Argumentación. // Van Dijk, T. A. (2000). El discurso como estructura y proceso. Barcelona: Gedisa, 2000.

Weston, Anthony (1994). Las claves de la argumentación. Barcelona: Ariel, 1994.

Wodak, Ruth; Meyer, Michael (2003) (comps.). Método de análisis crítico del discurso. Barcelona: Gedisa, 2003.

Corpus

Real Academia Española: Banco de datos (CORDE) [En línea]. Corpus diacrónico del español. http://www.rae.es (2008-02-02).

Real Academia Española: Banco de datos (CREA). Corpus de referencia del español actual. http://www.rae.es (2008-02-05).

Davies, Mark. Corpus del español. Universidad de Brigham Young. http://www.corpusdelespanol.org/ (2008-02-08).

CLiC. Universidad de Barcelona. Xarxa de Centres de Suport a la Innovació Tecnològica (X-IT). [En línea]. http://clic.fil.ub.es/demo_corpus/busca.php (2006-08-12).

Euromap (2003). Estudio comparativo de la evolución de las tecnologías lingüísticas en Europa. Resumen en español del informe final del proyecto Euromap, de Rose Lockwood y Andrew Joscelyne. Copenhague: Information Society Technologies / Instituto Cervantes, 2003.

Recibido: 2007-04-31. Revisado: 2008-01-24. Aceptado: 2008-05-24
[Retraso debido al gran número de originales en proceso]