Análisis Automático de Textos: Reconocimiento de Construcciones dicendi

AUTOMATIC TEXT ANALYSIS: RECOGNITION OF DICENDI CONSTRUCTION

Walter Koza
Grupo INFOSUR-UNR-Becario de CONICET
Rosario, Argentina
Walter_koza@yahoo.com.ar

Abstract

In this article, an automatic-detection-method for the recognition of "dicendi constructions" in natural language texts is presented. The dicendi construction (cdic) is defined as the construction which a speaker uses in his utterance to introduce a speech coming from somebody else´s utterance. The cdic are placed as asides within the clause, or else, as elements preceding or succeeding the canonical order and they are often composed of a dicendi verb, or, there may be occasions in which the verb is elided and the preposition "according to" is used.

For the automatic analysis, the description of the cdic was carried out, which enabled the latter modelling for the implantation in machine. On this occasion, the work was performed with the Xfst (Xerox) finite-state tool, whose tokenizing function enabled the identification of these constructions in the analysed corpus, achieving a 100% precision and a 96% coverage.

Key words: dicendi constructions, Comma, Punctuation, Automatic analysis, Finite states.

Resumen

En este artículo se presenta un método de detección automática para el reconocimiento de 'construcciones dicendi' en textos de lenguaje natural. Se define a la construcción dicendi (cdic) como aquella a la que recurre un emisor para introducir en su enunciado, un discurso proveniente de una enunciación ajena. Las cdic se ubican como incisos dentro de la cláusula o bien como elementos antepuestos o pospuestos al orden canónico y suelen estar conformadas por un verbo dicendi o bien, puede haber ocasiones en que el verbo esté elidido y se indique al autor del discurso citado con la preposición 'según'.

Para el trabajo de análisis automático, se procedió a la descripción de las cdic que permitió la posterior modelización para la implantación en máquina. En esta ocasión, se trabajó con la herramienta de estados finitos Xfst (Xerox), cuya función tokenizadora permitió el balizado de estas construcciones en el corpus analizado, lográndose un 100% de precisión y 96% de cobertura.

Palabras claves: Construcciones Dicendi, Coma, Puntuación, Análisis automático, Estados Finitos.

0. INTRODUCCIÓN

En este artículo se presenta un método para el reconocimiento automático de las **construcciones dicendi** (cdic), mediante la utilización de la herramienta de estados finitos Xfst (Xerox). Se define a la construcción dicendi como aquella a la que recurre un emisor para introducir en su enunciado, un discurso proveniente de una fuente de enunciación ajena. El trabajo se orienta hacia aquellas cdic que están delimitadas (1) o indicadas (2) por comas. [1]

(1) ["Mi esposa"**, dijo Juan,** "ha tenido amoríos con otro hombre.]

(2) ["Mi esposa ha tenido amoríos con otro hombre"**, dijo Juan**.]

En el primero de los casos, la cdic es una construcción incidental, que estaría dentro del grupo de incisos sin antecedentes [2], mientras que en el segundo aparece como un elemento pospuesto y va a estar relacionada con los casos de alteración del orden canónico.

Se intentará mostrar una aplicación de la herramienta informática xfst (Xerox Finite State Tools) para el análisis y balizamiento de las cdic en español. Este software fue desarrollado por *Xerox* y usado por *Xerox Research Centre Europe* (XRCE, Genoble, Francia) y *Palo Alto Research Center* (PARC, California, USA) y otros centros mundiales de investigación lingüística.

El programa permite trabajar los datos lingüísticos en forma independiente de la máquina algorítmica y no requiere que el investigador posea conocimientos de informática.

La aplicación se presenta como una implementación de autómatas de estados finitos, cuyo objetivo es producir análisis morfológico y generación. Xfst trabaja con archivos fuentes en los que se declara la información lingüística se introducen con un editor de textos planos, como se el *notepad* de Windows o el *emac* de Linux.

Entre las herramientas que utiliza este programa se encuentran Tokenizadores de estados finitos que ejecutan la segmentación del texto de acuerdo con la información morfosintáctica almacenada. En este caso, se va a recurrir a ellos para indicar en el corpus la cdic que allí se encuentren.

El artículo se organiza de la siguiente manera: En primer lugar se van a hacer algunas consideraciones teóricas respecto de los incisos y la alteración del orden canónico. En segundo lugar, se hará la distinción pertinente entre el discurso directo y el indirecto. Finalmente, se hará una descripción de aquellas construcciones que permiten introducir un discurso citado, la cual permitirá la posterior modelización para la implantación en máquina con el objetivo de reconocer las construcciones en textos de lenguaje natural, mediante la función tokenizadora de Xfst.

1. ACERCA DEL INCISO Y LA ALTERACIÓN DEL ORDEN CANÓNICO: GENERALIDADES Y ANTECEDENTES

1.1. Sobre el inciso

Como se ha mencionado, el trabajo está focalizado en las construcciones dicendi que se presentan como cláusulas incidentales o elementos pospuestos que remiten a una alteración en el orden canónico.

El inciso ha sido estudiado desde diversos puntos de vista. De acuerdo con lo que plantean Asuaje, Blondet, Mora y Rojas [3], pueden encontrarse análisis sobre él, no solo en la gramática tradicional,

sino también en la sintaxis de la lengua oral, el análisis del discurso, la informática y las tecnologías de síntesis y reconocimiento del habla. Los autores que se han consultado para la elaboración del presente trabajo entablan una estrecha relación entre el inciso y la pausa fónica o el cambio de entonación.

Dentro de la sintaxis en estudios no computacionales, cabe destacar que algunos autores como Alcoba [5] relacionan al inciso con "cualquier expresión que interrumpe la oración". Al respecto, aquí acuerdo con Desinano [6] en que es posible que fuera muy difícil explicar con cierto rigor a qué se refiere el autor cuando dice '*cualquier expresión que interrumpe la oración'*. Pues, tanto en los ejemplos de Desinano, como así también en los del propio Alcoba, los incisos constituyen una parte del *continuum* oracional, el que, más allá de la segmentación, sigue constituyendo una unidad sintáctica reconocible. He aquí uno de ellos:

(3) [Los caballeros, vestidos con sus armas, con el paso firme, se fueron al combate.]

En este caso, los dos incisos correspondientes a 'vestidos con sus armas' y 'con el paso firme' cumplen las funciones de predicativo subjetivo (no obligatorio) y complemento circunstancial de modo. Interrumpen el orden en la medida en que, de acuerdo con el orden regular, tendrían que ir en el predicado, después del verbo. No obstante, estas construcciones no pueden ser consideradas como elementos que interrumpen la oración en la medida en que, precisamente, son parte de ella.

Desde este punto de vista, y de acuerdo con los planteos de Desinano [6]:

> "la *coma* en este caso debe ser considerada como el recurso que encuadra a un inciso, es decir, a un sintagma que se destaca dentro del *continuum*, sintagma de muy variadas estructuras, que cumplen distintas funciones y que puede ser suprimido sin que el sintagma que lo incluye se desorganice como tal".

Ya en el ámbito de la lingüística computacional, Boula de Mareüil y Maillebuau [7] proponen un método de captura automática de los incisos en francés que incluyen verbos dicendi, a la vez que presentan un modelo prosódico específico para la síntesis de la palabra a partir del texto.

A partir de la definición de inciso que propone Grevisse, los autores mencionados distinguen los 'incidentes' de los 'incisos'. Incidente es la subfrase insertada en el interior de otra frase superior pero que no cumple el rol de sujeto o complemento. Por su parte, el inciso es un tipo especial de incidente que indica que se traen las palabras o pensamientos de alguien; se colocan dentro de la cita o después de esta. [7].

Los incisos son capturados automáticamente de acuerdo con criterios léxicos, sintácticos y *puntuacionales*. Para ello, elaboraron un método de detección de expresiones regulares que se apoya en elementos tales como comillas, comas, verbos discursivos, etcétera.

Por otro lado, también presentan un modelo prosódico en el que se reconocen los incisos en la lectura en voz alta de un texto a partir de las variaciones en la entonación y la modulación.

Una particularidad en el trabajo de estos autores consiste en que ellos hacen en su trabajo informático una distinción entre "incisos no final de frase" e "incisos de final de frase". El segundo de ellos correspondería a la expresión con verbo dicendi que se ubica al final de la cláusula:

(4) ["Quiero el divorcio", **dijo Juan**.]

En este aspecto, no acuerdo con el planteo de Boula de Mareüil y Maillebuau en que no parece pertinente hablar de "incisos al final de frase", puesto que, precisamente, la principal razón para que una construcción sea considerada inciso es que esté *insertada* en la cláusula y no al final. A tales efectos, la ubicación de la construcción dicendi ubicada allí, no va a ser considerada incidental, sino como un elemento pospuesto; el verbo principal de la oración que se desplazó a la derecha.

1.2. Sobre el orden canónico

Con orden canónico o no marcado, se remite a la manera "lineal" de exponer las ideas. Cada lengua tiene un orden canónico específico (sujeto-verbo-objeto, objeto-verbo-sujeto, etcétera). En el caso del español, el orden regular consistiría en colocar primero el sujeto con sus adjetivos y complementos, y luego a continuación el verbo y sus complementos: sucesivamente, el complemento directo, el indirecto y los circunstanciales de modo, lugar, tiempo, etcétera, según la voluntad del hablante. El orden marcado, por el contrario, sería cuando alguno de esos elementos no se aparece en su lugar habitual y se antepone o pospone con el objetivo de enfatizarlo. Generalmente, al elemento marcado se lo señala con una coma, de cierre si está al principio de la frase (5), de apertura (6) si está al final de ella. [5]

(5) [**En el parque,** un señor vestido de payaso regala caramelos.]

(6) [Bombardearon un poblado de agricultores**, los soldados del ejército**.]

El trabajo informático se propone detectar, a partir de la puntuación, a aquellas construcciones dicendi que estén delimitadas por comas, en caso de las construcciones incidentales, o indicadas por coma y finalizadas por un punto, en los casos de alteración del orden canónico. Asimismo, también se aplicarán reglas para el reconocimiento de aquellas construcciones iniciadas por la preposición 'según'.

2. ACERCA DE LAS CONSTRUCCIONES DICENDI EN EL DISCURSO DIRECTO E INDIRECTO

2.1. Estructura de la construcción dicendi

Al momento de producir un discurso que incluye más de una fuente de enunciación, el emisor dispone de dos variantes básicas de integración textual, el estilo directo y el indirecto; cada uno con sus propias reglas y condiciones.

En el estilo directo, el hablante reproduce textualmente un mensaje (7), logrando que ambos enunciados mantengan sus marcas de enunciación. En el indirecto, en cambio, el hablante reproduce el mensaje con algunos cambios (8). Tales modificaciones se pueden dar en el nexo, el tiempo verbal, referencias espacio-temporales, etcétera.

(7) [Juan ha dicho: "tengo un problema".]

(8) [Juan ha dicho que él tenía un problema.]

En el discurso directo pueden distinguirse claramente un "discurso citante" que incluye a otro, que se denomina "discurso citado". Los dos mantienen su autonomía y conservan sus propias marcas de

enunciación.

Así por ejemplo, en un caso como el siguiente:

(9) [“Es tarde para arrepentirse”, comentó María con resignación, “el amor se fue muriendo de a poco”.]

El discurso citante estaría dado por el segmento ‘comentó María con resignación’ y el discurso citado referiría a lo que dice María ‘Es tarde para arrepentirse, el amor se fue muriendo de a poco’.

Aquí se va a llamar **construcción dicendi** a aquella parte del discurso citante que cumple función de atribuir la cita textual a quien la dice. Para ello, puede valerse de un sintagma verbal cuyo núcleo sea un verbo dicendi (SVDic) o bien, estos pueden omitirse si se recurre a un sintagma preposicional con ‘según’ (SPsegún) (10).

(10) [Según María, “es tarde para arrepentirse, el amor se fue muriendo de a poco.]

La cita directa puede conectarse de diversas maneras:

- Sintagma Verbal Dicendi (SVDic) + dos puntos + cita textual:

(11) [Dijo María: “Es tarde para arrepentirse, el amor se fue muriendo de a poco.]

- Comienzo de cita textual + coma + SVDic + coma + Fin de cita textual:

(12) [“Es tarde para arrepentirse”, dijo María, “el amor se fue muriendo de a poco”.]

- Cita textual + coma + SVDic:

(13) [“Es tarde para arrepentirse, el amor se fue muriendo de a poco”, dijo María.]

- SPsegún + coma + cita textual. Ver cláusula (10).
- Cita textual + coma + SPsegún:

(14) [“Es tarde para arrepentirse, el amor se fue muriendo de a poco”, según María.]

Asimismo, también puede darse el caso de una combinación entre la preposición según y el verbo dicendi.

(15) [Según dijo María, “es tarde para arrepentirse, el amor se fue muriendo de a poco”.]

Las construcciones con ‘según’ son válidas tanto para el discurso directo como para el indirecto y puede prescindirse de las comillas sin que se altere el significado.

(16) [Según dijo María, es tarde para arrepentirse, el amor se fue muriendo de a poco.]

No obstante, este tipo de construcción no es válido cuando el discurso citado remite al propio autor o al receptor. Por ejemplo, una cláusula como la (17) admitirá la construcción con ‘según’ en (18), pero no en (19).

(17) [“Yo soy un esposo engañado”, dijo Juan.]

(18) [Según /dijo/ Juan, él es un esposo engañado.]

(19) *[Según dijo Juan, “yo soy un esposo engañado”.]

Por otro lado, en algunos casos, puede también recurrirse al adverbio ‘como’.

(20) [Como dijo María, “es tarde para arrepentirse, el amor se fue muriendo de a poco”.]

Al igual que en las construcciones con ‘según’, el uso de comillas es optativo y el adverbio puede aparecer antes, después (21) o en medio (22) del discurso citado.

(21) [“Es tarde para arrepentirse, el amor se fue muriendo de a poco”, como dijo María.]

(22) [“Es tarde para arrepentirse”, como dijo María, “el amor se fue muriendo de a poco”.]

Las construcciones dicendi con ‘como’ se diferencian de las de ‘según’ en que mientras que en este último es posible elidir el verbo dicendi, no ocurre lo mismo con el adverbio.

(23) [Según dijo María, “este es el final”.]

(24) [Como dijo María, “este es el final”.]

(25) [Según María, “este es el final”.]

(26) *[Como María, “este es el final”.]

2.2. Funciones sintácticas

Lo que se puede apreciar con respecto a las construcciones dicendi, es que el verbo (cuando está) es el verbo principal de la cláusula y el discurso citado, el complemento directo. A fin de ilustrar esta

cuestión, se hace pertinente comparar con el discurso indirecto. En él, se suele seleccionar una subordinada completiva con función de complemento directo. [8]

(27) [Juan le contó a Pedro que su esposa le había sido infiel.]

Aquí se tiene un verbo dicendi que selecciona el complemento directo dado por la subordinada 'que su esposa le había sido infiel' (del tipo inanimado) y un complemento indirecto de persona dado por 'Pedro' y reduplicado por el clítico.

En el caso del discurso directo:

(28) ["Mi esposa me ha sido infiel", le contó Juan a Pedro.]

Se observa que la subordinada pasó a ser una oración atributiva, 'Mi esposa me ha sido infiel', que posee la misma función.

3. IMPLANTACIÓN EN MÁQUINA

Se trabajó con un corpus conformado por un grupo de textos periodísticos que sumaban un total de diez mil palabras. Para el reconocimiento automático de las construcciones dicendi, se declaró en el archivo fuente de Xfst los elementos que podían componerlas. Estos eran:

Sintagmas Nominales

Sintagmas nominales comunes

En este caso, puede tratarse de un sintagma nominal núcleo (snn) [9] o no, cuando se extiende más allá de su núcleo. Es decir, se le pueden adicionar al núcleo, diversos modificadores como ser un sintagma adjetivo, un sintagma preposicional o un sintagma adjetivo más un sintagma preposicional. Gráficamente, el SN ofrece las siguientes variables:

- snn (*el libro*)
- snn + sadjn (*el libro nuevo*)
- snn + sp (*el libro de cocina*)
- snn + sadjn + sp (*el libro nuevo de cocina*)

Sintagmas nominales con nombres propios

Los sintagmas nominales con nombres propios estaban conformados de la siguiente manera:

- Un nombre propio (o varios) más un apellido: *José Manuel de la Sota*;
- Un SN más un nombre (o varios) más un apellido: *El abogado Juan Manuel López*;
- Un artículo más un nombre propio: *La corte suprema*.

Preposición 'según'

Se construye un sintagma preposicional cuyo núcleo es ‘según’ adicionándole a la preposición

- Un SN que remite al emisor del discurso citado: *según el médico*
- Un Sintagma Verbal dicendi (SVdic): *según dijo*
- Un SVdic más un SN con función de sujeto: *según dijo el médico*

Sintagma Verbal Dicendi

Para constituir estos sintagmas, en primer lugar se declararon los verbos dicendi y se incluyeron dentro de esta categoría a aquellos verbos que, según el contexto, podían implicar comunicación verbal, por ejemplo, ‘recordar’.

(29) [“No deje de tomar las pastillas”, me recordó el médico, “debe controlar su presión”.]

Posterior a eso, se tuvieron en cuenta la combinación con pronombres clíticos; ‘le dijo’, ‘me habló’, etcétera. Esto significa que el svdic podía estar conformado por:

- Un verbo dicendi: *dijo*
- Un clítico más un verbo dicendi: *me dijo*

A partir de estos elementos, se especificaron reglas para el reconocimiento de las cdic que estuvieran conformadas de la siguiente manera:

- Coma + SVdic + SN + coma: *, dijo Juan,*
- Coma + SVdic + coma: *, dijo,*
- Coma + SVdic + SN + punto: *, dijo Juan.*
- Coma + SVdic + punto: *, dijo.*
- Coma + según + SVdic + SN + coma: *, según dijo Juan,*
- Coma + según + SVdic + coma: *, según dijo,*
- Coma + según + SN + coma: *, según Juan,*
- Coma + según + SVdic + SN + punto: *, según dijo Juan.*
- Coma + según + SVdic + punto: *, según dijo.*
- Coma + según + SN + punto: *, según Juan.*

La notación específica de Xfst es la siguiente:

define cdic [%, " " svdic " " sn %, |

%, " " svdic %, |

%, " " svdic " " sn %. |

%, " " svdic %. |

%, " " seg " " svdic " " sn %, |

%, " " seg " " svdic %, |

%, " " seg " " sn %, |

%, " " seg " " svdic " " sn %. |

%, " " seg " " svdic %. |

%, " " seg " " sn %.];

Con el signo '%', se especifica que el signo siguiente, que en este caso es la coma o el punto, cumple su función original y no es una notación específica del programa. Mediante un blanco entre comillas, se indica el espacio que debe haber entre palabras o signos de puntuación. Con las barras se separan los distintos tipos de cdic que pueden llegar a presentarse.

Una vez definidas las cdic, se propone una regla para el balizamiento:

define FST [cdic] @-> [%< C D I C %> " "] ... [" " %< %/ C D I C %>] || otro1 _ otro2 ;

regex FST;

invert net

save cdic.fst

Esto significa que el programa, si encuentra alguna de las construcciones definidas previamente, debe balizarla. A continuación se presentan ejemplos de las cdic reconocidas por Xfst:

(…) "Las Pascuas las va a pasar acá"**<CDIC> , dijo. </CDIC>**

(…) Siempre ejerció y está ejerciendo el mando en contacto permanente con el Gabinete y tomando decisiones"**<CDIC> , dijo el médico. </CDIC>** Sin embargo, los funcionarios que lo visitaron ayer afirmaron que lo mantienen alejado del teléfono. (…)

(…) Y en los próximos meses**<CDIC> , agregó la especialista, </CDIC>** llegará el varenicline, que ya se comercializa en Estados Unidos, Europa y Brasil. (…)"

(…) "Por eso**<CDIC> , concluyó Schoj, </CDIC>** en términos de salud pública, las medidas más efectivas no son farmacológicas: implementar ambientes laborales libres de humo (…)

4. CONSIDERACIONES FINALES

Se presentó un método de detección automática de las cdic, a partir de la utilización de la herramienta de estados finitos Xfst. Para ello, fue necesaria la descripción lingüística de estas construcciones que permitió la posterior modelización para la implantación en máquina.

Las cdic son elementos que permiten la introducción del discurso directo en un enunciado. Por lo general están conformadas por un verbo dicendi o, en algunos casos, con la preposición 'según'. Pueden ubicarse como incisos dentro de la cláusula o como elementos pospuestos o antepuestos. En esta ocasión se focalizó en aquellas construcciones dicendi en las que estaba implicado el uso de la coma.

A partir de la modelización, se elaboraron reglas para la detección de las cdic en textos de lenguaje natural. Los resultados obtenidos hasta el momento son alentadores, puesto que se logró un 100% de precisión y un 96% de cobertura en el corpus analizado.

El trabajo a futuro se organiza en torno a dos ejes:

- Continuar el análisis de las funciones de la coma, con el objetivo de lograr un método de implantación en máquina;
- Indagar acerca del 'estilo híbrido' y sus posibilidades de detección automática. El estilo híbrido es un modo de traer a colación el discurso ajeno en donde se combinan recursos de cita directa e indirecta.

Referencias

[1] Este trabajo forma parte de una investigación mayor que en estos momentos estoy realizando sobre el análisis de las funciones de la coma en el marco de la elaboración de mi tesis doctoral, bajo la dirección de la Doctora Zulema Solana y financiado por una beca del CONICET.

[2] Koza, W. Análisis automático de textos: Reconocimiento de incisos. Revista Infosur. http://www.infosurrevista.com.ar/biblioteca/INFOSUR-Nro2-2008-Koza.pdf. Agosto, 2008.

[3] Asuaje, R., Blondet, M., Mora, E. et al Codificación Prosódica de la Información Incidental en el Discurso Espontáneo: Un estudio de caso. Rev. Vzlana. de Soc. y Ant. [online]. Vol.15, no.44, p.449-460. 2005

[5] Alcoba, S. "Puntuación y melodía de la frase", en Alcoba (coord.) La expresión oral. Ariel Practicum. Madrid. 2000.

[6] Desinano, N. Puntuación y gramática. En AA.VV. Estudios del lenguaje y enseñanza de la lengua. Ediciones Juglaría. Rosario. 2004.

[7] Boula de Mareüil, P. y Maillebau, E.. 2002, "Traitement des incises en français : capture automatique et modèle prosodique", en XXIVèmes Journées d'Étude sur la Parole, Nancy, 24-27juin.

[8] Delbecque, N. y Lamiroy, B. La subordinación sustantiva: Las subordinadas enunciativas en los complementos verbales. Bosque, I. y Demonte V. (Dirs.) Gramática descriptiva de la lengua española, Tomo III. Espasa Calpe. Madrid. 1999.

[9] Abney, S. Parsing by Chunks. En Berwick et al., Principle-Based Parsing. Dordrecht: Kluwer Academic Publishers. 1991.