Análisis Léxico de Unidades Léxicas Compuestas

MARC ORTEGA GIL. *Universidad Autónoma de Barcelona*
marc.ortega@uab.es

RESUMEN

En este artículo se quiere mostrar cómo se realiza el análisis de unidades léxicas compuestas como las locuciones, los tiempos verbales compuestos y las locuciones verbales en español, en el marco del sistema de análisis léxico del proyecto FrameNet Español[1] *basado en un diccionario electrónico formado por 634.500 formas, simples y compuestas, y un conjunto de gramáticas y herramientas construidas tomando las máquinas de estado finito como modelo matemático. El análisis de estos elementos se realiza sobre un corpus de oraciones anotadas léxicamente, de modo que cada unidad léxica (palabra) se anota con su correspondiente categoría léxica y sus características morfológicas, como en el caso de los verbos, nombres y adjetivos.*

ABSTRACT

This article aims to show how to perform lexical analysis of multiword lexical units like compound tenses and verbal phrases in Spanish, in the context of the lexical analysis system developed in the Spanish FrameNet project, based on an electronic dictionary consisting of 634,500 forms, simple and compound, and a set of grammars and tools built to take the finite state machine as mathematical model. The analysis of these elements is performed on a corpus of lexically annotated sentences, so that each lexical unit (word) is annotated with its corresponding lexical category and morphological characteristics, as in the case of verbs, nouns and adjectives.

PALABRAS CLAVE: lingüística de corpus, análisis léxico, unidades léxicas compuestas, máquinas de estados.
KEY WORDS: Corpus linguistic, lexical analysis, multiword lexical units, finite state machines.

1. INTRODUCCIÓN

El sistema de análisis en el que se enmarca este trabajo se realiza sobre un corpus de oraciones anotadas léxicamente, de modo que cada unidad léxica (palabra) se anota con su correspondiente categoría léxica y sus características morfológicas, como es el caso de los verbos, nombres y adjetivos, y permite reconocer tanto formas simples como formas locutivas. Dentro de estas últimas se analizan tanto las que se pueden reconocer a partir de un diccionario, como p. ej. *en todo momento*, como las que requieren un análisis sintáctico posterior al análisis léxico inicial para poder ser reconocidas. Este es el caso de locuciones verbales como *dar por sentado*, que puede aparecer como *da [siempre muchas cosas] por sentado*, o de los tiempos verbales compuestos en español. En estos casos el reconocimiento de la unidad léxica no puede llevarse a cabo únicamente a partir de un diccionario o de procedimientos estadísticos (Collins, 1999) y se requiere un análisis sintáctico más profundo que permita identificar como una unidad las formas que constituyen la unidad léxica locutiva

[1] **FrameNet Español**: un recurso léxico para el procesamiento semántico automático del español (FFI2008-0875), es un proyecto de investigación de semántica léxica y lingüística de corpus.

y anotarla con su correspondiente categoría léxica y sus características morfológicas, a la vez que los elementos que no pertenecen a la parte conexa de la locución, como *siempre muchas cosas* del ejemplo anterior, se sitúan en el contexto derecho, o izquierdo si fuera necesario, de la unidad locutiva, p. ej. *[dar/por/sentado] siempre muchas cosas* (Subirats y Ortega 2000).

El análisis de estas unidades locutivas se realiza en el marco de un sistema de análisis textual basado en técnicas de estado finito (*finite state methods*) en el que el análisis de las locuciones y los tiempos verbales compuestos se realiza a partir de un conjunto de gramáticas locales representadas como transductores subsecuenciales (Mohri, 1997) que se aplican, mediante un proceso de transducción, sobre autómatas finitos deterministas que representan oraciones anotadas léxicamente a partir de un diccionario electrónico.

El análisis de estas unidades léxicas locutivas permite, a la vez que son reconocidas y etiquetadas, desambiguar de forma eficiente los casos de ambigüedad (Laporte, 2001) como el que aparece con la forma *sentado* del ejemplo anterior, que se asocia a dos categorías distintas: (1) como forma del participio del verbo *sentar* y (2) como adjetivo. El análisis léxico basado en técnica de estado finito, en el que las oraciones se representan como autómatas finitos, permite representar y manipular de forma eficiente los casos de unidades léxicas ambiguas, es decir, aquellas unidades léxicas que como el caso de la forma *sentado*, están asociadas a dos o más clases de palabra o propiedades morfológicas. El análisis a partir de gramáticas locales, representadas como transductores subsecuenciales, permite eliminar gran parte de estas ambigüedades, con un margen de error prácticamente inexistente, de durante el análisis de las formas locutivas.

2. ANÁLISIS LÉXICO

El análisis de formas locutivas que se presenta se enmarca dentro del sistema de análisis léxico desarrollado en el proyecto FrameNet Español (FNE). Este sistema de análisis está formado por:

1. Un sistema de diccionarios electrónicos que contiene 634.500 formas, simples y compuestas,
2. un proceso de análisis léxico que permite, a partir del diccionario, analizar un texto y anotar con información léxica las formas simples y las formas locutivas que pueden ser reconocidas a partir del diccionario, como por ejemplo *bomba atómica*,

3. un proceso de transducción l que permite reconocer aquellas formas locutivas que por su naturaleza no pueden reconocerse a partir del diccionario ya que contiene construcciones sintácticas entre los elementos que forman la parte conexa de la locución.

2.1. *El diccionario electrónico*

El diccionario electrónico se genera a partir de un conjunto de diccionarios de lemas en los que cada uno de ellos está asociado a una categoría léxica con sus especificaciones flexivas, en los casos en los que estas puedan tener flexión morfológica. El lema es la forma que utilizan los diccionarios como modelo para definir sus entradas, es decir, el verbo en infinitivo y los nombres y los adjetivos, en singular, y en masculino cuando tienen flexión de género. Cuando un lema está formado por una única secuencia de caracteres, se denomina forma simple. Si un lema está integrada dos o más secuencias de caracteres, separadas por espacios en blanco, se denomina forma compuesta o locución. Las cadenas de caracteres que integran las formas compuestas o locuciones forman parte de la lista de formas simples que se integran en el diccionario, de modo que las cadenas que integran las locuciones están construidas sobre las entradas simples del diccionario. El conjunto de diccionarios de lemas que recoge 86.104 formas simples y 25.721 formas compuestas.

A partir de los diccionarios de lemas se utiliza un conjunto de reglas de flexión que formalizan la flexión morfológica del español. En el sistema que se presenta esta flexión se divide en dos grandes clases:

1. La flexión verbal,
2. la flexión que se aplica a nombres, pronombres y adjetivos.

El resultado de esta flexión es un conjunto de entradas en las que cada unidad léxica, simple o compuesta, se asocia a su correspondiente lema e información léxica (Cf. Fig. 1). En el caso de las unidades léxicas ambigua cada entrada lleva asociada un secuencia de dos o más lemas y su información léxica.

Este conjunto de entradas que forman el diccionario electrónico se representa en forma de transductor subsecuencial (Cf. Figs. 2 y 3) en el cual cada camino de estados, desde el estado inicial hasta un estado final, se corresponde con una única entrada del diccionario. Este formato permite representar de forma eficiente el diccionario a la vez que acelera de forma muy significativa el proceso de análisis.

```
ama,ama.N:m:f:s,amar.VPRED:IPRES:3s:IIMPE:2s,amo.N:f:s
amaba,amar.VPRED:IPIMP:1s:3s
amabais,amar.VPRED:IPIMP:2p
amábamos,amar.VPRED:IPIMP:1p
amaban,amar.VPRED:IPIMP:3p
amabas,amar.VPRED:IPIMP:2s
amad,amar.VPRED:IIMPE:2p
amada,amado.APRED:f:s,amado.N:f:s,amar.VPRED:PP:f:s
amadas,amado.APRED:f:p,amado.N:f:p,amar.VPRED:PP:f:p
amado,amado.APRED:m:s,amado.N:m:s,amar.VPRED:PP:m:s
amados,amado.APRED:m:p,amado.N:m:p,amar.VPRED:PP:m:p
amáis,amar.VPRED:IPRES:2p

bomba/atómica,bomba/atómica.N:f:s
bombas/atómicas,bomba/atómica.N:f:p
```

Figura 1. Ejemplos de entradas del diccionario electrónico.

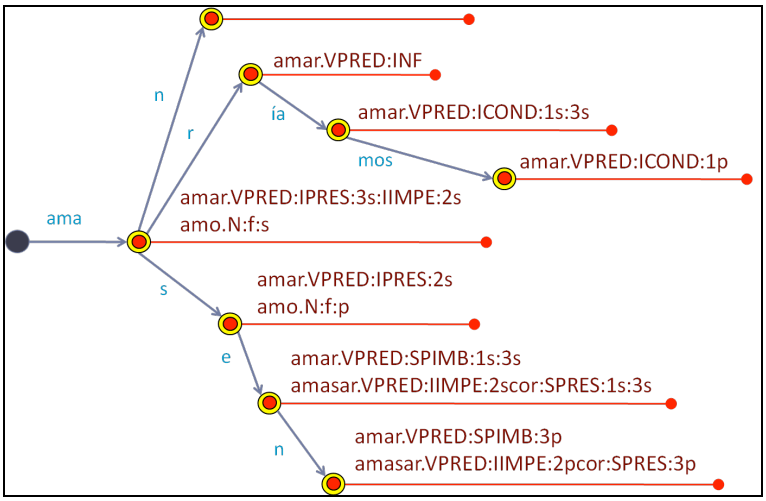

Figura 2. Representación del diccionario electrónico en forma de transductor subsecuencial en la que aparecen parte de las formas del verbo *amar*.

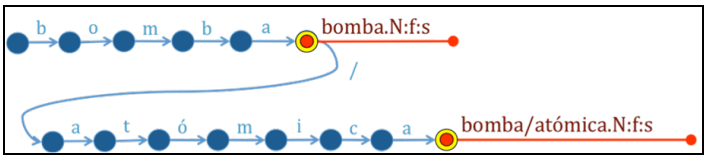

Figura 3. Representación del diccionario electrónico en forma de transductor subsecuencial en la que aparecen las formas de la locución *bomba atómica*.

2.2. *Análisis léxico*

El transductor subsecuencial que representa el diccionario electrónico se utiliza para transducir texto plano, sin formato ni ningún tipo de anotación. En este proceso el texto a analizar se utiliza como entrada del algoritmo de transducción de modo que cada secuencia de caracteres entre espacios en blanco es tratada como la entrada del transductor del

diccionario. El resultado de la transducción de cada una de estas secuencias es una nueva secuencia formada por el lema y la información léxica correspondiente a la unidad léxica presente en el diccionario. Estas salidas no se representan como un texto plano sino que se utilizan como elementos del alfabeto de entrada del autómata finito que finalmente representará el texto analizado.

Como se puede apreciar en la figura 4 el autómata finito que representa el resultado del análisis léxico del texto *al habérselo propuesto a tiempo* contiene la anotación de los elementos simples y locutivos que se pueden analizar a partir del diccionario y permite representar y manipular posteriormente de forma eficiente los casos ambiguos como las formas del verbo *haber*.

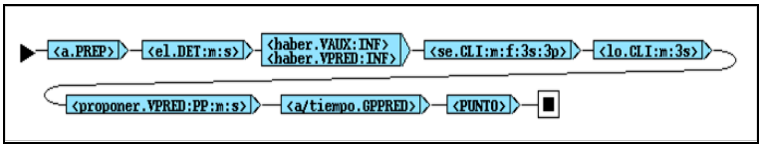

Figura 4. Resultado del análisis de *al habérselo propuesto a tiempo*.

3. TRANSDUCCIÓN LÉXICA

En el análisis de *al habérselo propuesto a tiempo* representado en la figura 4 aparece la forma verbal compuesta del verbo *proponer*, *haber propuesto*, representada como una secuencia cuatro de unidades léxicas en la que entre las formas *haber* y *proponer* que forman el tiempo verbal aparecen dos pronombres clíticos. El hecho de que este tipo de locuciones, a diferencia de lo que ocurre con la forma compuesta *a tiempo*, puedan contener entre las unidades que las forman elementos que requieren un análisis sintáctico (secuencias de pronombres clíticos, grupos nominales, etc.) provoca que no puedan detectarse a partir del diccionario y que por tanto se requiera un proceso de análisis posterior que se realiza sobre el autómata finito que representa el texto.

Dicho proceso se denomina *transducción léxica* y en él se aplican sobre el autómata textual un conjunto de gramáticas léxicas en forma de transductores subsecuenciales (Karttunen 1994) que permiten reconocer los elementos que forman parte de la unidad léxica locutiva, agruparlos en una única entrada dentro del autómata y separarlos de aquellos elementos que no pertenecen a la forma locutiva.

La transducción léxica permite detectar y analizar:

1. tiempos verbales compuestos, como es el caso de “habérselo propuesto”, y

2. locuciones verbales que no pueden analizarse en el paso previo debido a que entre sus elementos pueden aparecer segmentos oracionales que no pertenecen a la locución y que requieren un análisis sintáctico. Este es caso de la locución verbal "dar por sentado". En ella podemos encontrar elementos no locutivos, entre sus elementos conexos (Bobes 2000), como por ejemplo en "dar siempre muchas cosas por sentado" , donde encontramos la secuencia "siempre muchas cosas" entre las partes conexas "dar" y "por sentado".

El resultado de esta transducción permite eliminar la ambigüedad de las formas transducidas, como es el caso de las formas del verbo "haber" en el ejemplo de la figura 5. Otra de las ventajas de este proceso es que se reduce la complejidad del procesamiento del autómata textual en los procesos posteriores agrupando en una única unidad léxica todos los elementos que forman parte del predicado locutivo y situando los elementos externos en su contexto derecho (Cf. figura 5). De este modo durante un análisis sintáctico posterior la forma verbal compuesta del ejemplo anterior se detecta como un único elemento.

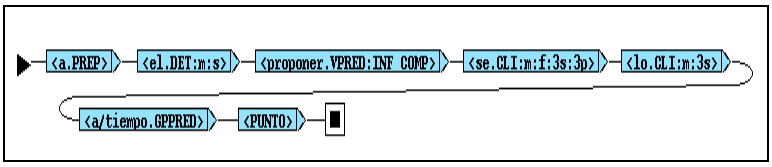

Figura 5. Resultado de la transducción del autómata de la figura 4 en la que se ha reconocido el tiempo verbal compuesto *haber propuesto*.

3.1. *Definición de las gramáticas de transducción*

Una gramática de transducción se define como una expresión regular o como un conjunto de expresiones regulares que especifican los elementos que forman la parte conexa de la unidad locutiva y los elementos no pertenecen a ella, y cómo se agrupa la parte conexa creando un nuevo lema y asociando a este la información léxica que se deriva del análisis de la unidad léxica.

En el ejemplo siguiente puede verse cómo se define la expresión regular que permite detectar y procesar el tiempo verbal compuesto haber + participio:

(1) (<haber.VAUX:INF\1> + <haber.VAUX:GER\1>) (<E> + <CLI\2> (<E> + <CLI\3>)) <VAR-1.VPRED:PP\4> [VAR-1,4-2,1-3&_COMP |2|3]

En este ejemplo se observa cómo se utilizan variables, \1 y VAR-1, por ejemplo, que permiten ampliar la definición del transductor y recolectar la información de las unidades que

forman parte del texto transducido para, posteriormente, construir la salida correctamente si se detecta correctamente la construcción locutiva.

Las variables de tipo VAR-x se denominan **variables posicionales** y se cargan con el valor que se encuentre en esa posición dentro de secuencia de caracteres que define la unidad léxica procesada. De este modo en la salida ([VAR-1,4-2,1-3&_COMP |2|3]) producida por el transductor durante la detección de la *secuencia haber se lo propuesto* la variable VAR-1 se carga con el lema del participio del verbo proponer, VAR-1=proponer

Las variables definidas como \1 se denominan **variables ligada a la transición** y son variables multicampo que se dividen en tantos campos como campos contiene la información de la unidad léxica que aparece en la posición transición definida por el valor de la variable.

Unidad léxica	<haber.VAUX:INF>			<se.CLI:m:f:3s:3p>					
Variable	\1			\2					
Elementos de la información léxica	haber	.VAUX	:INF	se	.CLI	:m	:f	:3s	:3p
Campos de la variable	1-1	1-2	1-3	2-1	2-2	2-3	2-4	2-5	2-6

Tabla 1. Ejemplo de carga de variables ligadas a la transición para la transducción de *haber se lo propuesto*.

Unidad léxica	<lo.CLI:m:3s>				<proponer.VPRED:PP:m:s>				
Variable	\3				\4				
Elementos de la información léxica	lo	.CLI	:m	:3s	proponer	.VPRED	:PP	:m	:s
Campos de la variable	3-1	3-2	3-3	3-4	4-1	4-2	4-3	4-4	4-5

Tabla 2. Ejemplo de carga de variables ligadas a la transición para la transducción de *haber se lo propuesto*.

En las tablas 1 y 2 puede verse como se definen estas variables y qué valores se cargan durante la transducción de *haber se lo propuesto.* La variable 4-2 se carga con la clase de palabra de la cuarta unidad de la secuencia (*4-1=VPRED* y la variable 1-3 con la información morfológica de la primera unidad, que en esta construcción puede ser *INF* (infinitivo) o *GER* (gerundio), de modo que en el ejemplo *1-3=INF*. Por último las variables 2 y 3 se cargan con el valor de los elementos opcionales que se pueden encontrar en las posiciones 2 y 3 de la secuencia.

El resultado producido por la salida del transductor a partir de los valores cargado en las variables es el que puede observarse en la figura 5. El transductor de la gramática del tiempo verbal haber + participio que se construye a partir de (1) es el que puede observarse en la figura 6.

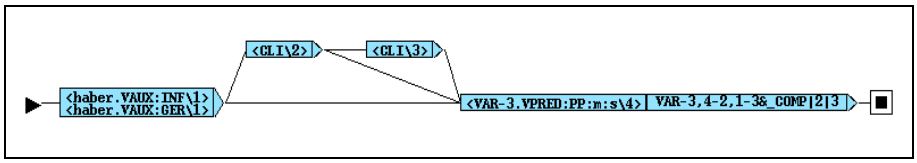

Figura 6. Transductor subsecuencial de la construcción *haber + participio*.

3.2. *Transducción de locuciones verbales*

Del mismo modo que en el apartado anterior se ha mostrado el proceso de detección y transducción de un tiempo verbal compuesto se realiza la transducción de locuciones verbales como *correr un riesgo*. De este modo para el texto:

(2) corrió en todo momento un enorme riesgo

su análisis léxico genera el autómata de la figura 7 en el que puede observarse como la locución en todo momento sí se detecta a partir del diccionario, pero en cambio no ocurre lo mismo con la locución correr un riesgo. Obsérvese como esta última contiene entre sus partes conexas elementos que no forman parte de la locución: *... en todo momento [un] enorme ...*

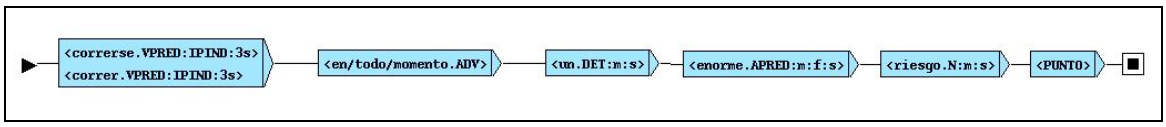

Figura 7. Autómata resultante de la anotación léxica de *corrió en todo momento un enorme riesgo*.

La anotación de esta locución solo es posible mediante la transducción del autómata textual con el transductor de la figura 8 que formaliza la gramática de detección de la unidad locutiva. Esta gramática se define de forma similar a la gramática del tiempo verbal compuesto vista anteriormente.

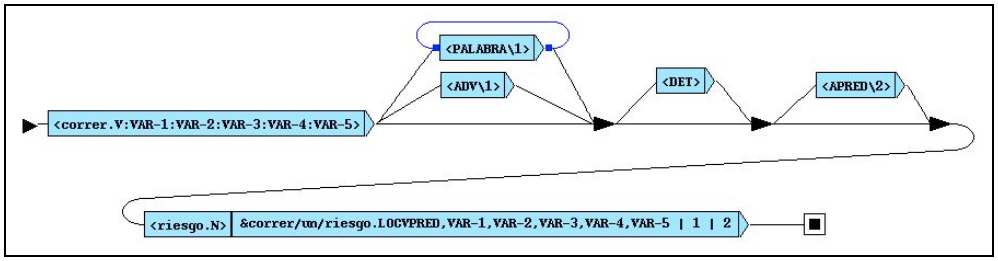

Figura 8. Transductor subsecuencial de la locución verbal *correr un riesgo*.

El proceso de transducción léxica del autómata de la figura 7 produce como resultado un nuevo autómata textual en el que la unidad locutiva correr un riesgo se anota como una única unidad y los elementos que no forman parte de ella se sitúan en su contexto derecho (Cf. Fig. 9).

Figura 9. Resultado de la transducción del autómata de la figura 7.

4. REFERENCIAS

Bobes, E. (2000). *Gramática electrónica de las locuciones verbales*. Laboratorio de Lingüística Informática, Universidad Autónoma de Barcelona.

Collins, M. (1999). *Head-driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Laporte, E. (2001). Reduction of lexical ambiguity. *Lingvisticae Investigationes XXIV:1*, Amsterdam-Philadelphie : Benjamins, pp. 67-103.

Karttunen, L. 1994. Constructing Lexical Transducers, en *Proceedings of the Fifteenth International Conference on Computational Linguistics. Coling 94,* vol. I, pág. 406-411, Kyoto, Japan.

Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics*, *vol. 23(2)*, 269-311.

Subirats, C. y Ortega, M. (2000). Tratamiento automático de la información textual en español mediante bases de información lingüística y transductores. *Estudios de Lingüística Española 10*. Disponible en http://elies.rediris.es/elies10/