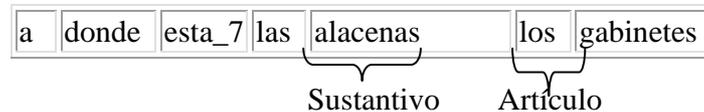


Manual de etiquetación del Nivel POS (Part-of-Speech) del Corpus DIME

El nivel de POS consiste en la identificación y etiquetación de la categoría léxica de cada palabra; la identificación de categorías léxicas puede ser útil para predecir la ocurrencia de una reparación; cuando ocurre una reparación se rompe la estructura normal de la elocución ya que la coherencia gramatical entre las palabras se pierde; por ejemplo, la probabilidad de que un sustantivo sea seguido por un artículo es muy baja, como se ve en el Ejemplo 1. En estos casos, la probabilidad de que haya una reparación es alta; y esta información es útil para predecir la reparación.

Ejemplo 1 (d12 – utt93 POS)



Definición de conjunto de etiquetas.

El conjunto de etiquetas definido es el resultado del análisis de un diálogo del corpus DIME. Como material de apoyo para la definición se emplearon conjuntos de etiquetas propuestos por trabajos anteriores ([1], [2], [3] y [4]). Un conjunto de etiquetas para la identificación de categorías léxicas puede ser tan especializado como sea necesario para resolver una tarea determinada. El resultado de este análisis es muestra en la Tabla 1 que contiene las etiquetas para este corpus.

Etiqueta	Categoría
N	Sustantivo
V	Verbo
VAM	Verbo Auxiliar – Modal
VC	Verbo con Clítico
A	Adjetivo
AD	Adjetivo Demostrativo
TD	Artículo Determinado
TI	Artículo Indefinido
R	Adverbio
RI	Adverbio Interrogativo
RR	Adverbio Relativo
RN	Adverbio de Negación
RA	Adverbio de Afirmación
P	Pronombre
PD	Pronombre Demostrativo

PR	Pronombre Relativo
PI	Pronombre Interrogativo
PC	Pronombre Clítico
S	Preposición
C	Conjunción

Tabla 1: Etiquetas para la identificación de categorías léxicas.

Etiquetación

Para el nivel de etiquetación de POS se utiliza como base el nivel de transcripción de palabras definido en el corpus DIME.; como el resto de las etiquetas, estas etiquetas están alineadas temporalmente a los niveles de etiquetación. A continuación se ejemplifican cada una de las etiquetas utilizadas.

- **Sustantivo**

Es la parte de la oración que expresa la categoría correspondiente a sustancia, es decir, seres o cosas que pueden ser sujeto u objeto de cualquier acción, estado o accidente expresable con un verbo. Morfológicamente, los sustantivos se caracterizan, según las lenguas, por uno o varios de los siguientes rasgos: género y número determinado intrínsecamente y como declinación (en su caso).

Para la etiquetación de un sustantivo en el corpus DIME se ocupa la etiqueta **N** y no se considera ninguna otra propiedad. Ejemplo:

Ejemplo 2 (d12 – utt01 Etiqueta N para sustantivo)

.sil	quieres	que	desplace	o	traiga	algu_7n	objeto	a	la	cocina	.sil
							N			N	

- **Verbo**

Es la parte de la oración que expresa la existencia, acción y estado del sujeto; es el único elemento que funciona como núcleo del grupo predicativo de la oración. El verbo indica procesos, o implica la noción de tiempo, mientras que el sustantivo (elemento no verbal por excelencia) indica objetos y no supone temporalidad. No siendo universales estas caracterizaciones semánticas, se define al verbo como el elemento indispensable para la constitución de un enunciado.

Las propiedades que tiene un verbo son: tipo, modo, tiempo y persona. Para la etiquetación de esta categoría sólo son considerados los siguientes tipos: verbo auxiliar o modal y verbo principal. Los verbos auxiliares o modales, como *querer* y *poder*, son los que se combinan con formas nominales de otros verbos para ampliar el esquema aspectual y temporal de la conjugación.

Los verbos principales se etiquetan con la letra **V**.

Ejemplo 3 (d12 – utt01 Etiqueta V para verbo)

.sil	quieres	que	desplace	o	traiga	algu_7n	objeto	a	la	cocina	.sil
			V		V						

En ocasiones el verbo está unido a un pronombre clítico; cuando se presente en esta forma el verbo la etiqueta que se utiliza es **VC**, ejemplo:

Ejemplo 4 (d01 – utt21 Etiqueta VC para verbo)

podri_7as	juntarlo	un	poco	ma_7s
	VC			

Para los verbos auxiliares – modales la etiqueta a utilizar es **VAM**.

Ejemplo 5 (d12 – utt33 Etiqueta VAM para verbo auxiliar – modal)

.bn	me	puedes	mostrar	el	cata_7logo	de	fregaderos	y	ma_7quinas	.sil
		VAM								

Un verbo auxiliar – modal está parcial o totalmente desprovisto de su significado original (no tiene agente) y se utiliza como un mero morfema por otros verbos [6]. Por ejemplo, el verbo *poder* es un auxiliar si la intención no es preguntar sobre la capacidad del agente de realizar una acción, como se muestra en el Ejemplo 5, y por el contrario, si hay un agente que *puede* el verbo desempeña su función normal; otro ejemplo es el verbo *querer* si no se pregunta por la creencia o el deseo, es decir, no hay un agente que *quiere* se etiqueta como verbo auxiliar

- **Adjetivo**

El adjetivo es una parte de la oración o categoría de rango secundario (adjunto del sustantivo) denotativa de cualidad, grado de comparación y situación; es una de las dos clases de palabras (la otra es el verbo) que modifica al sustantivo. Los rangos morfológicos que presenta el adjetivo son los de género y número.

En el corpus DIME solamente se consideran la distinción entre un adjetivo demostrativo del resto de los adjetivos. Para los adjetivos se utiliza la etiqueta **A**.

Ejemplo 6 (d12 – utt102 Etiqueta A para adjetivo)

.bn	dame	la	blanca	superior	doble	.bn
				A	A	

En este ejemplo la palabra “blanca” es un adjetivo nominalizado, la palabra pierde su función original (adjetival) por una nominal. El artículo que le precede es un morfema del sustantivo que corrobora la afirmación de que es un sustantivo.

En caso de los adjetivos demostrativos se utiliza la etiqueta **AD**.

Ejemplo 7 (d12 – utt11 Etiqueta AD para adjetivo demostrativo)

.sil	quieres	que	mueva	.sil	este	objeto	.sil	hacia	aca_7	.sil
					AD					

- Artículo

Es el elemento gramatical que puede anteponerse o no a las palabras con función sustantiva, con las que concuerda generalmente en género y número, y cuya presencia puede variar la función del sustantivo.

Según el grado de mayor o menor determinación, el artículo, en español puede ser determinado e indeterminado. La gramática moderna niega el carácter de artículo al indeterminado, considerándolo como indefinido. El artículo es un morfema exclusivo del sustantivo; por ello cuando se antepone a una palabra, la sustantiviza: el cantar, la rica, etc.

La etiqueta que corresponde al artículo determinado es **TD**.

Ejemplo 8 (d12 – utt06 Etiqueta TD para artículo determinado)

.sil	hacia	la	derecha	.sil
		TD		

Al artículo indefinido (indeterminado) le corresponde la etiqueta **TI**.

Ejemplo 9 (d12 – utt65 Etiqueta TI para artículo indefinido)

.sil	quieres	un	mueble	como	estos	.sil	aqui_7	.sil
		TI						

Existen casos en los que se debe tener cuidado al etiquetar un artículo indefinido, ya que en ocasiones se presentan ocurrencias en las que se trata, en realidad, de un adjetivo o de un pronombre. Como se muestra en el Ejemplo 10 y Ejemplo 11.

Ejemplo 10 (d12 – utt32 Adjetivo y no artículo indefinido)

.sil	hay	un	cata_7logo	de	alacenas	.sil	uno	de	estantes	.sil	uno	de	estufas	Y	extractores	.sil	...
		A															

Como Adjetivo

Ejemplo 11 (d12 – utt32 Pronombre y no artículo indefinido)

.sil	hay	un	cata_7logo	de	alacenas	.sil	uno	de	estantes	.sil	uno	de	estufas	y	extractores	.sil	...
						P				P							

Como Pronombre

En el Ejemplo 10 se puede ver como la palabra “un” se emplea para hacer notar el número de objetos; en este caso, el número de catálogos de alacenas, mientras que el Ejemplo 11 “uno” se emplea como pronombre de la palabra “cata_7logo”.

- **Adverbio**

Parte invariable de la oración, que sirve para modificar la significación del verbo, del adjetivo, de otro adverbio e incluso puede aplicarse al sustantivo y a toda una frase.

Hay varias clases de adverbios; los que se consideran para la etiquetación son: adverbios pronominales (los que desempeñan la función de pronombres), interrogativos (“cuándo”, “cuánto”), relativos (“donde”, “como”, “cuando”), adverbio de afirmación (“sí”, “también”, “efectivamente”) y adverbios de negación (“no”, “tampoco”).

Para adverbios interrogativos se utiliza la etiqueta **RI**.

Ejemplo 12 (d12 – utt53 Etiqueta RI para adverbio interrogativo)

do_7nde	quieres	que	la	ponga	.sil
RI					

Para adverbios relativos se usa la etiqueta **RR**.

Ejemplo 13 (d13 – utt79 Etiqueta RR para adverbio relativo)

mejor	ponme	la	estufa	.sil	eh	donde	esta_7	la	ventana	.sil
						RR				

Para adverbios de afirmación se emplea la etiqueta **RA**.

Ejemplo 14 (d12 – utt50 Etiqueta RA para adverbio de afirmación)

.bn	s_7	.sil
	RA	

Y para adverbios de negación la etiqueta **RN** es empleada.

Ejemplo 15 (d12 – utt90 Etiqueta RN para adverbio de negación)

.sil	no	.sil
	RN	

Cualquier otro adverbio que no pertenezca a alguna de las clasificaciones mencionadas se etiqueta con la letra **R**, por ejemplo:

Ejemplo 16 (d17 – utt32 Etiqueta R para adverbios en general)

.sil	ahí_7	esta_7	bien	.sil
	R		R	

- **Pronombre**

Es una categoría que se define tradicionalmente como la parte de la oración cuya función es suplir al sustantivo, evitando así y por razones de economía una repetición.

Hay varias clases de pronombres; para la etiquetación del corpus DIME se consideran los pronombres demostrativos, interrogativos, relativos y clíticos; cualquier otro tipo de pronombre se considera como pronombre general.

Para los pronombres en general se utiliza la etiqueta **P**.

Ejemplo 17 (d12 – utt89 Etiqueta P para pronombres en general)

.sil	quieres	alguno	.sil
		P	

Para los pronombres demostrativos se usa la etiqueta **PD**.

Ejemplo 18 (d12 – utt42 Etiqueta PD para pronombre demostrativo)

.sil	e_7ste	junto	a	la	estufa	.sil
	PD					

- **Conjunción**

Es la parte de la oración o clase de palabras cuya función es unir dos frases o dos miembros de una misma frase. La conjunción cumple la función de enlazar miembros gramaticalmente, aunque a veces signifique contrariedad o separación de sentido entre unos y otros.

Para la conjunción se emplea la etiqueta C.

Ejemplo 24 (d12 – utt35 Etiqueta C para conjunciones)

.bn	e_7ste	es	el	cata_7logo	de	fregaderos	.sil	y	ma_7quinas	lavatrastes	.sil
								C			

En ocasiones el pronombre ‘que’ funciona como conjunción y no como pronombre relativo como se muestra en el Ejemplo 25.

Ejemplo 25 (d12 – utt53 Etiqueta C para “que” con función de conjunción)

do_7nde	quieres	que	la	ponga	.sil
		C			

Esto es porque la palabra ‘que’ no hace referencia a alguien, algo o algún lugar y solamente sirve como nexos. En contraste tenemos:

Ejemplo 26 (d12 – utt9 Ejemplo de “que” como pronombre relativo)

.sil	a	la	mitad	del	espacio	que	hay	entre	la	ventana	y	la	pared	.sil
						PR								

aquí, ‘que’ esta haciendo referencia a un algo (el espacio).

- **Silencio y ruido**

En el caso de presencia de silencio o ruido, se mantienen las etiquetas que vienen desde el nivel de palabras, “.sil” y “.bn”, respectivamente.

Ejemplo 27 (d13 – utt11 Etiquetación de un silencio)

po_7nmelo	.sil	aqui_7	.sil	en	esta_7	esquina	.sil
	.sil		.sil				.sil

Ejemplo 28 (d13 – utt77 Etiquetación de un ruido)

.sil	e_7ste	.sil	y	estos	otros	tres	.bn	son	los	modelos	de	extractores	que	tenemos	.sil
							.bn								

Para concluir, en la Figura 1 se muestra la etiquetación de las categorías léxicas de una elocución completa.

Ejemplo 29 (d13 – utt68 Nivel de etiquetación POS)

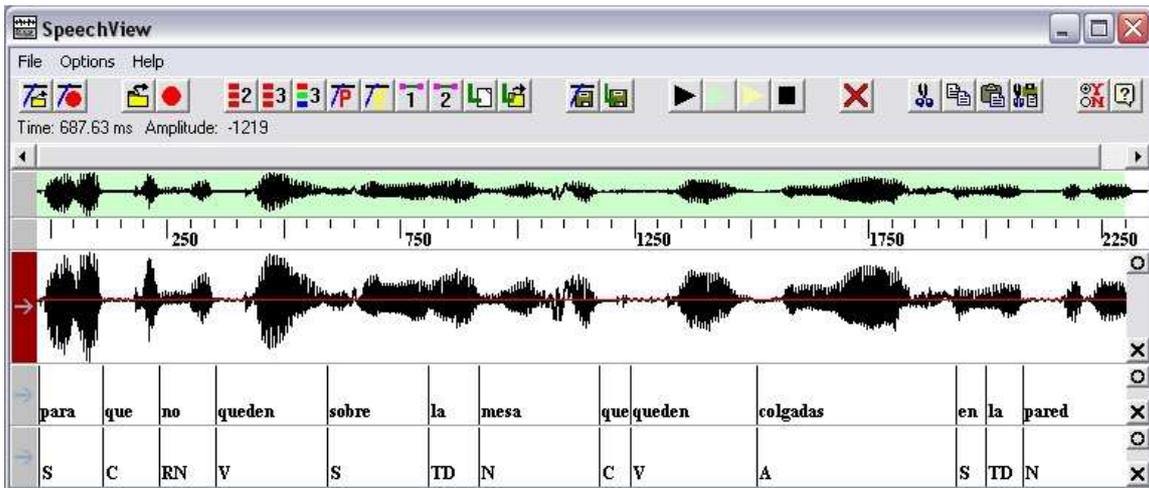


Figura 1: Nivel de POS

En el caso de que alguna palabra funcione como un marcador del discurso no se asigna una etiqueta de categoría léxica, un ejemplo se muestra en la Figura 2.

Ejemplo 30 (d01 – utt01 Nivel de etiquetación POS con marcadores del discurso)

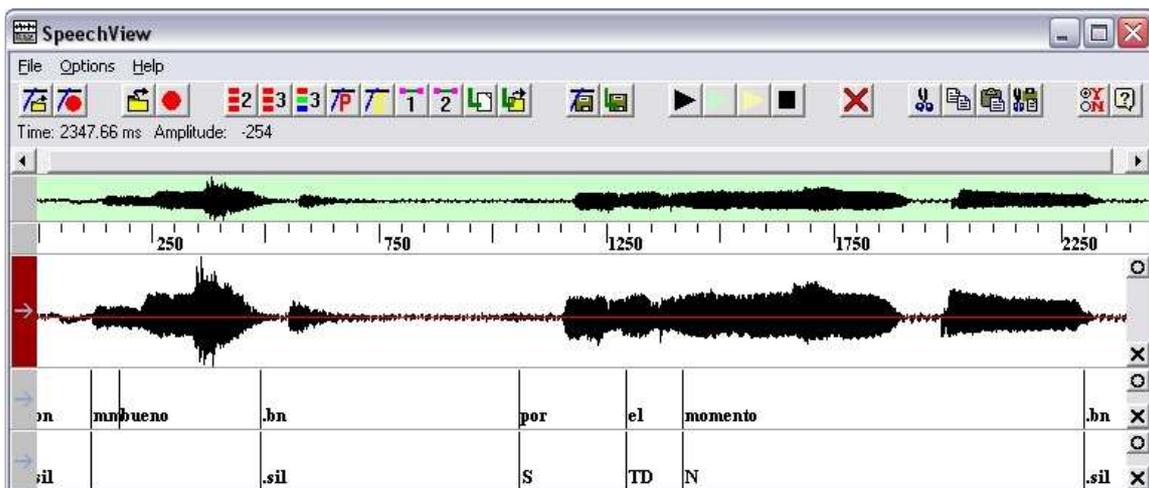


Figura 2: Etiquetación de POS con marcadores del discurso

Sugerencias

Para poder realizar esta etiquetación, una muy buena ayuda es la utilización de un diccionario. Si no cuentan con uno que consideren confiable pueden acceder a la siguiente dirección: <http://www.rae.es>. Se trata de la página de Internet de la Real Academia Española, en la cual se encuentra online el Diccionario de la Real Academia Española.

Referencias

- [1] M. Civit & M. A. Martí. Design Principles for Spanish Treebank. En proceedings of the First Workshop on Treebanks and Linguistics Theories (TLT2002), páginas 61 – 77, septiembre 2002.
- [2] M. Civit. Criterios de Etiquetación y Desambiguación Morfosintáctica de Corpus en Español. Tesis Doctoral. Universidad de Barcelona, 2003.
- [3] P. A. Herman. Speech Repairs, Intonational Boundaries and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog. Tesis Doctoral. Universidad de Rochester, 1997.
- [4] D. Farwell, S Helmreich & M. Casper. SPOST: a Spanish Part-of-Speech Tagger. http://crl.nmsu.edu/Publications/farwell/far_etal95.html
- [5] Enciclopedia Salvat.
- [6] <http://www.amerschmad.org/spanish/gram/conjuga.htm>