Nociones básicas de morfología

Fernando Carranza
fernandocarranza86@gmail.com

Clase 6
Sábado 26/04/2025

- Unidad 1: Introducción
- Unidad 2: Los algoritmos supervisados
- Unidad 3: Anotación morfológica y de clase de palabra
  - a. Análisis morfológico
    - **i) nociones básicas de morfología: raíz, lema, tema y forma de palabra, morfología flexiva y derivativa;**
    - ii) stemmatizer y lemmatizer de NLTK;
    - iii) lemmatizer de Spacy basado en reglas y basado en aprendizaje automático;
    - **iv) etiquetamiento en CONLL-U: columnas de ID, de palabra y de lema.**
  - b. Clases de palabras
    - i) etiquetas de BNC y de Universal. Práctica de anotación.
    - ii) etiquetamiento en CONLL-U de clase de palabra con EAGLES y de rasgos morfológicos.
    - iii) Postag NLTK con tagsets BNC y Universal y Postag de Spacy y Stanza.
- Unidad 4: Anotación sintáctica
- Unidad 5: Anotación para propósitos específicos

Presentación

Estructura y temas de la clase de hoy:

1. Introducción
2. 3.a.i) nociones básicas de morfología: raíz, lema, tema y forma de palabra, morfología flexiva y derivativa
   - Morfología flexiva
   - Morfología derivativa
3. 3.a.iv) etiquetamiento en CONLL-U: columnas de ID, de palabra y de lema
4. Recapitulación
5. Bibliografía

Gramática

- En sentido estricto
  - Morfología: estructura de las palabras
  - Sintaxis: manera en que se combinan y se disponen linealmente las palabras, así como de los grupos que conforman.
- En sentido amplio
- Fonética
- Fonología

Semántica y pragmática no son parte de la gramática para Española y de Academias de la Lengua Española (2009).

- Cada parte de la gramática consta de unidades sustantivas y relaciones.
- Las unidades de la morfología son el morfema (unidad menor) y la palabra (unidad mayor).
- El morfo o exponente es la forma concreta en que se realiza el morfema, que es una unidad abstracta.

(1) a. librería $\rightarrow$ libr-er-ía

b. maldad $\rightarrow$ mal-dad

Existen distintos tipos de lengua en relación con el modo en que se componen morfológicamente las palabras.

- **Lenguas aislantes**: como el chino mandarín. Prácticamente no tiene morfemas ligados y cada palabra es invariable.

(2)

Ta	*chi*	*le*	*yi*	*ge*	*yóutiáo*
él	comer	PERFECTIVO	uno	CLASIFICADOR	buñuelo

Él comió un buñuelo.

- **Lenguas aglutinantes**: como el turco, el finlandés, el húngaro, el quichua santiagueño. Los morfemas se pueden delimitar claramente, no presentan morfoamalgamas.

(3) *suya-a-chka-nki*
esperar-1PER.SG.OBJ-PROGRESIVO-PAS-2PERS.SG

- **Lenguas flexivas**: como el español. Los afijos presentan un alto grado de fusión, aparecen morfoamalgamas.
- **Lenguas polisintéticas**: Permite la incorporación y síntesis, de modo tal que una palabra puede incluir morfemas con distintas funciones sintácticas.

(4) a. *Nina-ca-qua*

yo carne como

Existe evidencia de que el tipo morfológico de una lengua afecta el rendimiento de los modelos de lenguaje que se construyen con métodos estadísticos (n-gramas, modelos con redes neuronales) (ver Park *et al.* 2021, Gerz *et al.* 2018).

Los modelos de lenguaje tienen que tokenizar los textos antes de procesarlos. Existen distintas aproximaciones para hacer esto:

- **Byte-Pair Encoding (BPE)**: divide a través de un algoritmo no supervisado el texto en chunks de caracteres en función de su frecuencia de aparición.
- **Morfessor**: divide a través de un algoritmo no supervisado las palabras en morfemas en función de las recurrencias de sus chunks en otras palabras.
- **Finite State Transducer (FST)**: es un método que segmenta las cadenas en función de una serie de reglas dadas de antemano (ejemplo: `https://github.com/wjonasreger/morphological_finite_state_transducer`.

Park *et al.* 2021 estudian el efecto de usar uno u otro de estos tokenizadores para modelizar lenguas tipológicamente variadas en lo que respecta al modo en que se componen morfológicamente las palabras y observan que los FST funcionan mejor que los otros métodos y que algunos de los tipos de codificación morfológica afectan también el rendimiento.

Tipos de morfología

Algunos autores argumentan que el morfema como tal no existe porque existen varios fenómenos en las lenguas del mundo en los que no puede identificarse un morfema en concreto, sino más bien un proceso morfológico (ver Bonet 2016). Son los casos de morfología no concatenativa: alternancias vocálicas o consonánticas, reduplicación, morfología sustractiva.

Alternancias vocálicas

sufijo	open	opened
sufijo + cambio vocálico	sleep	slept
cambio vocálico	give	gave

Cuadro: Pasado en inglés. Adaptado de Bonet (2016)

Alternancias vocálicas

e $\rightarrow$ **ie**	referir	refiero	refirió
o $\rightarrow$ **ue**	morir	muero	murió
e $\rightarrow$ **i**	reír	río	rió
incremento velar	salir	salgo	salió

Reduplicación

takki ‘pierna’	taktakki ‘piernas’
ulu ‘cabeza’	ululu ‘cabezas’

Cuadro: Plural mediante reduplicación en agta. Tomado de Bonet (2016)

Morfología sustractiva

(5) a. ataká:-li-n ‘colgar algo (sg.)’

b. aták-li-n ‘colgar algo (pl.)’

Plural mediante morfología sustractiva en Koasati.
Tomado de Bonet (2016)

Tipos de morfemas

Los morfemas se pueden clasificar en distintos tipos:

- Por su grado de libertad: libres y ligados.
- Por su posición con respecto a la base: prefijos, sufijos, interfijos, infijos, circunfijos.

Tipos de morfemas

- **Raíz**: morfema que constituye el inicio de una construcción morfológica
- **Afijos**: morfemas ligados que se combinan con una base.
- **tema o raíz grecolatina**: son morfemas que requieren combinarse con otro tema, con un afijo o con una palabra. Se caracterizan por proceder de palabras griegas y latinas.

Unidades superiores al morfema

Otras unidades morfológicas superiores al morfema e inferiores a la palabra incluyen las siguientes:

- **tema**: resultado de eliminar a una palabra todos los afijos flexivos
- **tema verbal**: raíz verbal más vocal temática
- **base**: elemento al que se le aplica un proceso morfológico.

- **Lexema**: forma abstracta que se usa para denominar a una palabra sin hacer alusión a ninguna forma de palabra específica.
- **Lema**: forma convencional que hace uno de alguna forma de palabra en particular para referir al lexema.

Mientras sintamos vergüenza habrá esperanza para todos

palabra	lema	tema	raíz
Mientras			
sintamos			
vergüenza			
habrá			
esperanza			
para			
todos			

Gran parte de la ciudad se encuentra bajo agua. Todas las actividades fueron suspendidas. El Municipio no descarta que haya otras víctimas fatales

(https://www.infobae.com/sociedad/2025/03/07/tragico-temporal-en-bahia-blanca-en-vivo-las-ultimas-noticias-minuto-a-minuto/)

palabra	lema	tema	raíz
Gran			
parte			
de			
la			
ciudad			
se			
encuentra			
bajo			
agua			

Todas las actividades fueron suspendidas.

palabra	lema	tema	raíz
Todas			
las			
actividades			
fueron			
suspendidas			

El Municipio no descarta que haya otras víctimas fatales

palabra	lema	tema	raíz
El			
municipio			
no			
descarta			
que			
haya			
otras			
víctimas			
fatales			

- **Alomorfía**: Fenómeno mediante el cual un morfema se realiza por más de un morfo, cuya naturaleza no se puede explicar por meras reglas fonológicas.
- **Supleción**: Fenómeno mediante el cual un morfema se expresa a través de más de una raíz. verbo *ir* con raíces del latín *uadere* e *ire* (Maiden (2013).

Existen dos grandes tipos de ramas para la morfología:

- Según el tipo de morfología
  - Flexiva
  - Derivativa
- Según la perspectiva
  - Sincrónica
  - Diacrónica

Morfología flexiva

Estudia la morfología flexiva las variaciones de las palabras que implican cambios de contenido de naturaleza gramatical con consecuencias en las relaciones sintácticas, como en la concordancia (Ellos trabajan), o en la rección (Para ti). El conjunto de estas variantes constituye la flexión de la palabra o su paradigma flexivo.

En español se reconocen las siguientes categorías flexivas:

- Número
- Género
- Caso
- Persona
- Tiempo
- Aspecto
- Modo

Valor

- Saussure (1916) estableció los lineamientos generales para una teoría gramatical. Entre ellos, definió los signos como unidades que se oponen en términos de valor. Sin embargo, Saussure no llegó a implementar sus ideas en una gramática concreta. Esa obra fue recogida por el estructuralismo.
- La tradición estructuralista posterior codificó matemáticamente la noción de valor mediante el uso de rasgos.

Valor

Adaptando la sistematización de (Barthes, 1993, p. 66), las oposiciones en los valores de estos rasgos con respecto a la relación de los términos de la oposición pueden darse de tres formas:

- Oposiciones privativas
- Oposiciones binarias +/-
- Oposiciones equipolentes

Oposición privativa

Una oposición privativa es una en la que el rasgo está marcado para ciertos elementos, pero no para otros.

- Supongamos que tenemos el rasgo A con el valor [a].
- El valor [a] aparece en las formas marcadas.
- Las formas no marcadas carecen del valor [a]. Sencillamente no hay ninguna especificación para ese rasgo.

Oposición binaria

Una oposición binaria +/- es similar a una privativa, solo que en lugar de ausencia del rasgo en el elemento no marcado, lo que hay es un valor negativo, mientras que la marcación se indica con un rasgo positivo.

- Supongamos que tenemos el rasgo A con los valores [+a], [-a]
- El valor [+a] aparece en algunas de las formas.
- En otras de las formas aparece el valor [-a]

Oposición equipolente

Una oposición equipolente es una en la que los valores de la oposición son en alguna medida cualitativamente diferentes:

- Supongamos que tenemos el rasgo A con los valores [a] y [b].
- El valor [a] aparece en algunas de las formas.
- En otras de las formas aparece el valor [b].

Es sumamente relevante, a la hora de formalizar las categorías morfológicas, decidir qué rasgos contienen y a qué clase de oposición obedecen.

Número I

El número en español es una categoría que cuenta con dos valores, singular y plural. En los nombres, solo el plural recibe una marcación concreta. Esta marcación se realiza mediante los siguientes alomorfos:

1. /s/
   - terminados en vocal átona o -á, -é, ó: *casas*, *calles*, *yanquis*, *libros*, *tribus*, *sofás*, *cafés*, *platós*.
   - terminados en diptongo: *bonsáis*
   - terminados en consonantes distintas de -l, -n, -r, -d, -z, -j: *mamuts*, *cenits*, *tictacs*
2. /es/
   - terminados en -l, -n, -r, -d, -z, -j: *cónsules*, *mieles*, *leones*, *caracteres*, *tutores*, *paredes*, *peces*, *relojes*.
   - terminados en -s y -x agudos o monosílabos: *autobuses*, *compases*, *toses*
   - terminados en -y, salvo extranjerismos crudos: *bueyes*, *leyes*, pero *jerseis*.

Número II

3. / /
   - nombres acabados en -s y -x no agudos ni monosílabos: *tesis*, *lunes*, *torax*
4. casos especiales
   - terminadas en -í, -ú: *jabalíes/jabalís*, *tabúes/tabús*
   - algunas formas populares: *mamases*, *papases*, *cafeses*, *manises*, *cacahueses*, *pieses*.

- *Singularia tantum*: formas que solo aparecen en singular: canícula, caos, cariz, cenit, grima, oeste, salud, sed, tez, tino, zodíaco
- *Pluralia tantum*: adentros, albricias, anales, andas, andadas, andurriales, arras, comestibles, entendederas, exequias, expensas, facciones, fauces, gárgaras, maitines, ojeras, nupcias, preliminares, viatuallas, víveres, etc.

China supervisa el suministro de víveres a buque de Filipinas varado en disputado arrecife

(https://www.infobae.com/america/agencias/2024/11/15/china-supervisa-el-suministro-de-viveres-a-buque-de-filipinas-varado-en-disputado-arrecife/)

palabra	lema	tema	raíz
China			
supervisa			
el			
suministro			
de			
víveres			

Dado que el número es una categoría que marca solo uno de sus valores, el rasgo privativo parece ser la forma más adecuada de tratarlo, si bien la opción más popular es considerar que se trata de una oposición equipolente: singular vs. plural.
En otras lenguas aparecen otros números, como el dual o el paucal.

El género es una categoría gramatical que aparece en muchas lenguas y que puede aparecer en pronombres, deteminantes, nombres, adjetivos y verbos, entre otras categorías. En español aparece en pronombres, nombres, determinantes, cuantificadores y adjetivos.

Los nombres se pueden clasificar según su comportamiento en relación al género:

- No animados
  - Nombres de género arbitrario: *la mesa*, *la mano*, *el dedo*, *el tema*
  - Nombres de género ambiguo: *el/la maratón*, *el/la mar*.
- Animados
  - Nombres que flexionan en género: *león/leona*, *camarero/camarera*.
  - Nombres de género común: *dentista*, *artista*
  - Nombres epicenos: *jirafa*, *víbora*, *culebra*, *tiburón*

- En la mayoría de los nombres del español, el género es enteramente arbitrario y no recibe una interpretación semántica.
- Muchos nombres han sufrido cambios de género en la diacronía, sin que eso conlleve un cambio de significado, como ocurrió con sustantivos *dolor*, *calor*, *color*, etc.
- Algunos nombres de género ambiguo (e.g., *el/la color*, *el/la mar*) están asociados a valoraciones sociolingüísticas o de registro.
- Algunos nombres que refieren a objetos presentan pares masculino y femenino que parecen obedecer a una especie de derivación: *manzana/manzano*, *naranja/naranjo*, *palta/palto*, *cereza/cerezo*. No obstante, estos pares conforman una clase no productiva.

Harris (1991) hace una descripción bastante exhaustiva de cómo se manifiesta el género en español y propone lo que se conoce como la *regla del clonaje humano*:

> ***Regla del clonaje humano***: *Los nombres que refieren a humanos tienen una forma masculina y otra femenina.*

Nombres que marcan -o para el masculino y -a para el femenino:

secretario	secretaria
campesino	campesina
cocinero	cocinera
criado	criada
alumno	alumna
amigo	amiga

Nombres de género común sin ninguna marca particular de género:

estudiante
intérprete
cómplice
esquimal
caníbal
conyuge
mártir
joven

Nombres sin terminación específica para el masculino y -a para el femenino:

(presid)ente	(presid)enta
(sirvi)ente	(sirvi)enta
(profes)or	(profes)ora
colegial	colegiala
doncel	doncella
monje	monja
nene	nena
jefe	jefa

El masculino y el femenino están vinculados “derivativamente”:

duque	duquesa
poeta	poetisa
actor	actriz

Nombres marcados con -a de género común:

(aristó)crata
(art)ista
(mon)arca
camarada
acróbata
patriota
suicida
policía
colega
pirata
guía

Nombres terminados con -o de género común:

contralto
soprano
testigo
modelo

Nombres con pares supletivos o heterónimos:

macho	hembra
hombre	mujer
padre	madre
yerno	nuera

Las excepciones a la regla del clonaje humano que reconoce Harris (1991) son muy pocas:

(6) persona, criatura, víctima

Solo unos pocos animales cumplen asimismo con la regla del clonaje humano.
Algunos mediante pares flexivos:

gato	gata
perro	perra
león	leona

Otros mediante pares supletivos:

toro	vaca
caballo	yegua
carnero	oveja
gallo	gallina

La inmensa mayoría de los nombres que refieren a animales poseen solo género gramatical y no aportan ningún tipo de información respecto del sexo del animal:

Masculino	femenino
camello	*camella
reno	*rena
erizo	*eriza
dinosaurio	*dinosauria
*foco	foca
*cebro	cebra
*ardillo	ardilla
*jirafo	jirafa

En ciertos contextos pragmáticos se puede forzar una interpretación de género:

(7) a. La tortuga y el caracol / salen de paseo con su casa al sol / la tortuga es una señora/ y el caracol es un señor. (Canción infantil anónima)

b. Pero las que estaban hermosísimas eran las víboras. Todas, sin excepción,estaban vestidas con traje de bailarina, del mismo color de cada víbora... Un flamenco dijo entonces: - Yo sé lo que vamos hacer. Vamos a ponernos medias blancas, coloradas y negras y las víboras de coral se van a enamorar de nosotros. (H. Quiroga, Las medias de los flamencos)

Estomba (2016) denomina a estos casos *pares conceptuales con oposición de género*.

Existen también contextos que fuerzan pares genéricos inesperados para lograr un efecto retórico. Es lo que Estomba (2016) denomina pares lúdicos:

(8) a. El cielo canta a la ciela (Huidobro, Poema de VII cantos, V)

b. he aquí que caliente, oyente, tierro, sol y luno,/incógnito atravieso el cementerio. (C. Vallejo, Quedéme a calentar la tinta)

- La visión más extendida y conocida sobre el género en español es que sus valores se vinculan mediante una relación equipolente que opone los valores femenino y masculino.
- Desde un punto de vista teórico, esta noción tiene problemas para dar cuenta del llamado masculino genérico, así como de cómo obtienen masculino las completivas (*el que vaya me preocupa*) o las cláusulas de infinitivo (*el salir el sol es todo un espectáculo*).
- Tampoco puede dar cuenta de por qué el masculino es la forma no marcada y el femenino la marcada.

- En oposición a eso, de acuerdo con Harris (1985), entre otros, el género en español es un rasgo binario [+/-femenino].
- Esto da mejor cuenta del masculino genérico y en principio podría dar cuenta también de la marcación.
- Sin embargo, sigue sin dar cuenta de cómo obtienen masculino las completivas o las cláusulas de infinitivo.

- En Harris (1991) se presenta una alternativa según la cual el género en realidad es un rasgo privativo solo presente en los sustantivos femeninos.
- Esta visión parece ser la más adecuada para dar cuenta de todos los fenómenos vinculados al género observados (el hecho de que el masculino sea la forma no marcada y el femenino la marcada, el masculino genérico, el masculino que aparece en completivas o en cláusulas de infinitivo).

En otras lenguas, existen otros géneros, como el género neutro y otros.

palabra	lema	tema	raíz
No			
la			
encuentra			
ella			

Persona

- **1era persona**: El que habla
- **2nda persona**: El interlocutor
- **3era persona**: Alguien externo al acto de comunicación.

Halle (1997), basándose en ideas de Benveniste, propone formalizar la persona a partir de los siguientes rasgos:

- [± PAAH] (participante del acto de habla)
- [± AAH] (autor del acto de habla)

De este modo, el sistema de personas quedaría formalizado de la siguiente forma:

- 1era: [+ PAAH] [+ AAH]
- 2nda: [+ PAAH] [– AAH]
- 3era: [– PAAH] [– AAH]

Caso

Función	Caso
sujeto gramatical	nominativo
objeto directo	acusativo
objeto indirecto	dativo
término de preposición	terminal u oblicuo
modificador nominal	genitivo

Cuadro: Correspondencias entre casos y funciones gramaticales. Adaptado de Muñoz Pérez y Trombetta (2018).

	Nominativo	**Acusativo**	**Dativo**	**Terminal**
1sg	yo	me	me	mí
2sg	vos	te	te	vos
3sg	él/ella	lo/la	le	él/ella
1pl	nosotros	nos	nos	nosotros
2pl	ustedes	los/las	les	ustedes
3pl	ellos/ellas	los/las	les	ellos/ellas

Tiempo, aspecto y modo

En español, el tiempo, aspecto y modo se expresan de un modo que dificulta su delimitación. En la flexión verbal, estas categorías se expresan en cierta medida separadamente de las categorías de persona y número:

(9) a. Raíz + VT = Tema

b. Tema + TAM + PN

La morfología derivativa, a diferencia de la flexiva, no produce formas de palabra, sino que crea palabras nuevas.

La derivación en español se puede dar por diferentes vías:

- Derivación por prefijación:
- Derivación por sufijación:
- Derivación por infijación:
- Derivación por parasíntesis:
- Derivación por composición:
- Derivación por composición culta:

No es tan frecuente hacer anotación de esquemas derivativos para entrenar un algoritmo no supervisado, como sí lo es la anotación de flexión.

El español cuenta con el corpus AnCora:
`https://github.com/UniversalDependencies/UDSpanish-AnCora`

Este corpus está estructurado siguiendo los lineamientos de CONLL-U, por las Conferences of Natural Language Learning. La U es por Universal Dependencies.

Revisar el link que aparece en el Campus.

En esta clase introdujimos los siguientes temas:

- qué es la morfología
- cuáles son sus unidades básicas (tema, raíz, morfema, afijo)
- qué tipos de morfología existen y sus posibles formalizaciones
- Qué es CONLL-U.

Bibliografía I

Barthes, R. (1993). Elementos de Semiología. En *La aventura semiológica*, pp. 17–83. Paidós, Barcelona. Publicación original: 1985.

Bonet, E. (2016). Morfemas y alomorfos. En Gutiérrez-Rexach, J., editor, *Enciclopedia de Lingüística Hispánica*, pp. 700–709. Routledge, Londres.

Española, R. A. y de Academias de la Lengua Española, A. (2009). *Nueva gramática de la lengua española*. Espasa Calpe, Madrid.

Estomba, D. A. (2016). El género sintáctico y la proyección funcional del nombre. Tesis de Maestría. Universidad Nacional del Comahue.

Gerz, D., Vulić, I., Ponti, E. M., Reichart, R., y Korhonen, A. (2018). On the relation between linguistic typology and (limitations of) multilingual language modeling. En *2018 Conference on Empirical Methods in Natural Language Processing*, pp. 316–327. Association for Computational Linguistics.

Bibliografía II

Halle, M. (1997). Distributed morphology: Impoverishment and fission. En Bruening, B., Kang, Y., y McGinnis, M., editores, *Papers at the Interface*, volumen 30, pp. 425–449. MIT Working Papers in Linguistics, Cambridge, Massachusetts.

Harris, J. (1985). Spanish word markers. En Nuessel, F. H., editor, *Current Issues in Spanish Phonology and Morphology*. Indiana University Linguistics Club, Bloomington.

Harris, J. (1991). The exponence of gender in spanish. *Linguistic Inquiry*, 22(1):27–62.

Maiden, M. (2013). Morphophonological innovation. En Maiden, M., Smith, J. C., y Ledgeway, A., editores, *The Cambridge History of the Romance Languages*, pp. 216–267. Cambridge University Press.

Bibliografía III

Muñoz Pérez, C. y Trombetta, A. (2018). Caso y diátesis. En Giammatteo, M., editor, *Categorías lingüísticas*, pp. 221–250. Waldhuter, Buenos Aires.

Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., y Schwartz, L. (2021). Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Saussure, F. d. (1916). *Curso de lingüística general*. Losada, Buenos Aires. Traducido por Amado Alonso. Año de la edición: 2012.