

# Léxico y morfología

## Procesamiento del Lenguaje Natural

José María Gómez Hidalgo

<http://www.esp.uem.es/~jmgomez/>

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Índice

- Introducción y definiciones
- Tipos de morfología
- Técnicas de análisis morfológico
- Etiquetado sintáctico estocástico (POS-TAGGING)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Léxico y morfología

## Introducción y definiciones

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Introducción y definiciones

- **Morfología**
  - Se ocupa de la formación de palabras a partir de las unidades más básicas de significado denominadas morfemas
  - Parte de la lingüística que estudia la estructura interna de las palabras, su flexión, derivación y composición
  - En ocasiones a las unidades mínimas de significación se les denomina monemas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Introducción y definiciones

- Los monemas son de dos tipos
  - Lexemas
    - Monemas con significado pleno (representan un concepto o idea)
  - Morfemas
    - No tienen significado pleno, sino un significado gramatical
    - Relacionan a los lexemas o modifican su significación
- Pensamos = pens + amos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Introducción y definiciones

- Léxico o lexicón
  - Vocabulario de una lengua – lista de todos sus elementos léxicos
  - Diccionario típico
    - Las entradas se identifican mediante una forma base o forma canónica
      - Inglés: forma canónica = raíz no flexionada
      - Castellano o francés: los verbos se representan con una forma flexionada (infinitivo) comer
- Informan de pronunciaciones, categorías gramaticales, definiciones, información etimológica o estilística

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Introducción y definiciones

- **Diccionarios electrónicos**
  - Los más elementales cuentan con
    - listas de formas plenas o léxicos desplegados (listas de palabras con todas las formas)
      - walk, walks, walked, walking
    - la información gramatical correspondiente
  - **En lenguas con flexión rica y compleja**
    - El lexicón proporciona una raíz
    - La información gramatical correspondiente
    - El componente morfológico se encarga de generar las posibles formas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Introducción y definiciones

- **Ventajas del análisis morfológico con respecto al uso de léxicos desplegados**
  - En lenguas de flexión rica y en lenguas aglutinantes el uso de léxicos desplegados es inviable
  - Reconocer palabras desconocidas o formas de palabras que no están incluidas en el diccionario
  - A partir de la identificación de flexiones gramaticales pueden inferirse funciones sintácticas
  - Se puede conseguir una descripción del idioma a tratar

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Léxico y morfología

## Tipos de morfología

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Tipos de morfología

- Hay 3 mecanismos para la formación de palabras
  - flexión
  - derivación
  - composición

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Tipos de morfología

- **Morfología flexiva**
- **En las gramáticas tradicionales las variaciones se agrupan en “paradigmas”**
- **Ejemplo – paradigma latino**
  - dominus, dominum, domini, domino, etc.
  - Raíz = domin- se combina con diferentes terminaciones (-us, -um, -i, -o, etc.)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Tipos de morfología

- **Morfología flexiva**
  - Variación en la forma de las palabras según su función
  - No modifica la función sintáctica de la raíz
- **Ejemplos**
  - Nombres en singular y plural (mesa, mesas )
  - Verbos en tiempo presente y pasado (viene, vino)
- **Flexión (o desinencia) – sistema que define las variaciones posibles de la raíz**

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Tipos de morfología

- **Morfología flexiva**
  - Inglés – grado de variación flexiva relativamente pobre
  - Ejemplo
    - La mayor parte de los verbos cuentan únicamente con los morfemas gramaticales -s, -ed, -ing
  - Castellano – grado de variación flexiva mayor

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Tipos de morfología

- **Morfología flexiva**
  - Las lenguas se puede clasificar según el mayor o menor uso de la flexión
    - Lenguas aislantes – Casi sin flexión (chino)
    - Lenguas flexivas – Afijos con significados complejos (castellano)
    - Lenguas aglutinantes – Añaden múltiples sufijos a la raíz (turco, euskera)
  - Lenguas polisintéticas significado gramatical a partir de la flexión (esquimal)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Tipos de morfología

- **Morfología flexiva – ejemplo**
  - Castellano
    - En la punta de la punta de la rama del manzano de la cuesta
  - Euskera
    - Aldapeko sagarraren adarraren puntaren punta
- Si comparamos ambas mediante su traducción euskera-castellano
  - Aldapeko(de la cuesta) sagarraren (del manzano) adarraren (de la rama) puntaren (de la punta) punta (en la punta)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Tipos de morfología

- **Morfología derivativa**
  - Formación de nuevas raíces (flexionables) a partir de otras raíces que suelen pertenecer a categorías gramaticales diferentes
  - Puede provocar un cambio de categoría
    - Nombre nación
      - adjetivo nacional
      - verbo nacionalizar
    - nombre nacionalismo
    - verbo internacionalizar

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Tipos de morfología

- **Morfología de composición**
  - **Combinación de palabras completas para dar origen a nuevas formas**
    - El significado puede deducirse a partir de los significados de las partes: pelirrojo
    - El significado puede variar ligeramente: peliagudo (complicado)
    - El significado puede no estar motivado por el de las partes: boquirrubio (incauto, ingenuo, joven presumido)
  - **Su tratamiento suele ser más complejo**

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Tipos de morfología

- **Morfología de composición**
  - **Puede resultar un problema en algunas lenguas que**
    - basta con escribir dos palabras juntas para formar palabras compuestas
    - no intercalan un carácter (guión) entre ambas (no hay evidencias de composición)
    - se pueden crear nuevas palabras compuestas que no aparecen en el diccionario
  - **Ejemplo: en alemán Lufthansafrachtflüge (vuelos de carga Lufthansa)**

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Léxico y morfología

## Técnicas de análisis morfológico

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Técnicas de análisis morfológico

- **Análisis morfológico**
  - Perseguiamos
    - Capturar las regularidades morfológicas del lenguaje humano
    - Aprovecharlas para reducir el tamaño del léxico
      - Es preferible poder derivar raiz+forma-verbal que listar todas las formas del verbo
      - Solo se listan las formas irregulares

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Técnicas de análisis morfológico

- **Analizador morfológico**
  - Entrada => forma
  - Salida => lema + rasgos morfológicos

Entrada	Salida
cat	cat + N + sg
cats	cat + N + pl
cities	city + N + pl
merging	merge + V + pres_part
caught	(catch + V + past) o (catch + V + past_part)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Técnicas de análisis morfológico

- **Tipos de técnicas**
  - **Técnicas de estados finitos**
    - Autómatas (analizadores de un nivel)
    - Transductores (analizadores de dos o más niveles)
  - **Técnicas basadas en reglas**
    - Equivalentes en expresividad a las anteriores
    - Reglas de reconocimiento y transformacionales
    - Gramáticas regulares, contextuales, basadas en unificación
      - Indicadas por expresividad => simplifican el desarrollo

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Técnicas de análisis morfológico

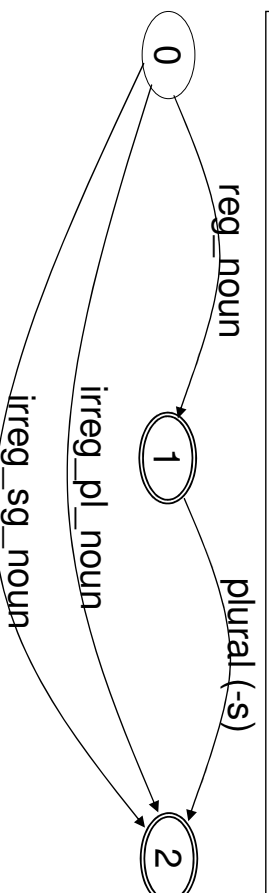
- Técnicas de estados finitos
  - Elementos del analizador
    - Léxico de morfemas
      - raíces + afijos
    - Morfoláctica = qué combinaciones de morfemas son válidas
      - cats = cat + s
      - “el plural del nombre se denota por una s al final”
    - Alteraciones fonológicas = reglas ortográficas = cambios al producirse la combinación
      - city + s = cities

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Técnicas de análisis morfológico

- Estados finitos (léxico, morfoláctica)

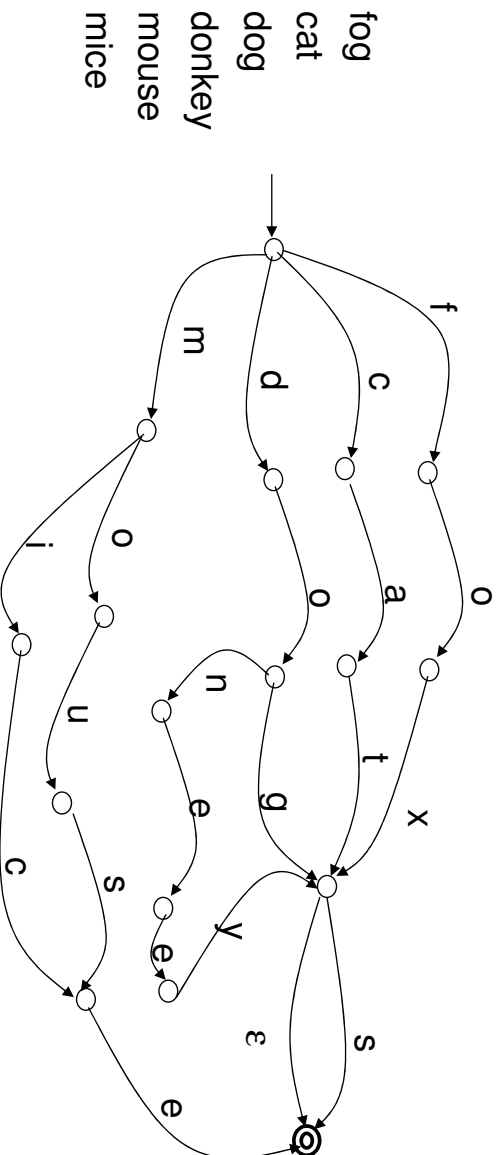
reg_noun	irreg_pl_noun	irreg_sg_noun	plural
fox	sheep	sheep	-s
cat	mice	mouse	
dog			



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Técnicas de análisis morfológico

- Estados finitos (autómata compilado)

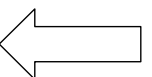


Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Técnicas de análisis morfológico

- Técnicas de estados finitos
  - Morfología de dos niveles

upper level	léxico	cat + N	cat + N + pl
lower level	superficie	cat	cats



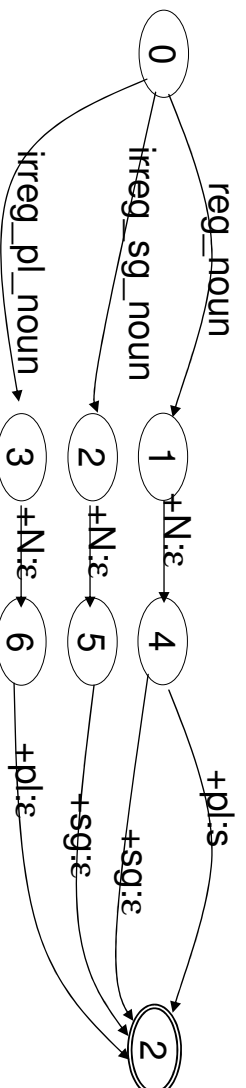
c:c    a:a    t:t    +N:ε    +pl:s

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Técnicas de análisis morfológico

- Técnicas de estados finitos (transductores)

reg_noun	irreg_pl_noun	irreg_sg_noun	plural
fox	sheep	sheep	s
cat	m o:i u:i: ce	mouse	
dog	g o:e o:i: e se	goose	



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Técnicas de análisis morfológico

- Técnicas de estados finitos (transductores)
  - Usos posibles
    - Como reconocedor
      - Recibe dos cadenas de entrada (una léxica y una superficial) y responde cierto o falso según una sea transducción de la otra
    - Como generador
      - Genera pares de cadenas
    - Como traductor
      - Recibe una cadena superficial y genera su transducción léxica

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Técnicas de análisis morfológico

- Técnicas basadas en reglas
  - Equivalentes en expresividad a autómatas
  - Ejemplos

name	description	example
consonant doubling	single letter consonant doubled before -ing/-ed	beg/begging
e deletion	silent e dropped before -ing/-ed	make/making
e insertion	e added after -s,-z,-x,-ch,-sh before -s	watch/watches
y replacement	-y changes to -ie before -s, to i before -ed	try/tries
k insertion	verbs ending with vowel +c add -k	panic/panicked

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

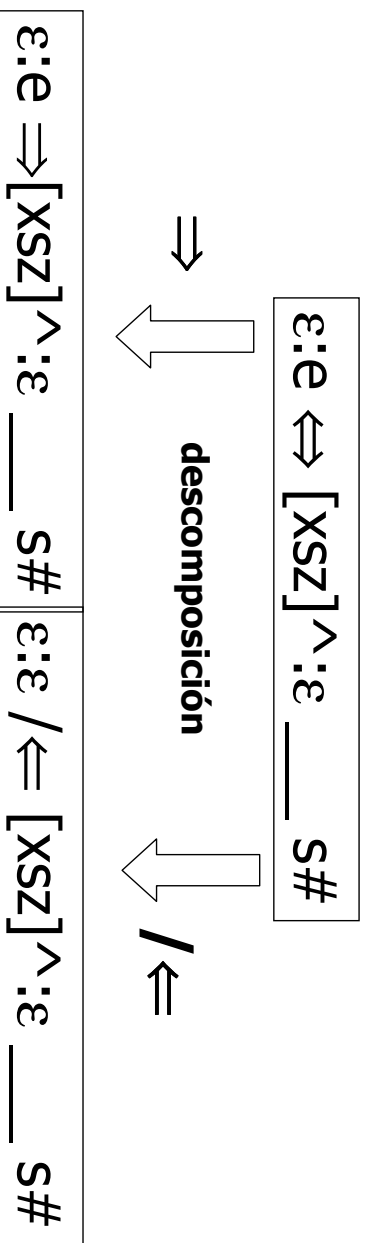
# Técnicas de análisis morfológico

- Técnicas basadas en reglas
  - Estructura de reglas
    - a:b ⇐ Contexto\_izquierdo \_\_\_\_ Contexto\_derecho
      - el item léxico a debe corresponder al item superficial b cuando se encuentra en el contexto
    - a:b ⇒ Contexto\_izquierdo \_\_\_\_ Contexto\_derecho
      - el item léxico a sólo puede corresponder al item superficial b cuando se encuentra en el contexto
    - a:b ⇔ Contexto\_izquierdo \_\_\_\_ Contexto\_derecho
      - el item léxico a debe corresponder al item superficial b cuando se encuentra en el contexto y sólo entonces
    - a:b /⇐ Contexto\_izquierdo \_\_\_\_ Contexto\_derecho
      - el item léxico a no puede corresponder al item superficial b cuando se encuentra en el contexto

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Técnicas de análisis morfológico

- Técnicas basadas en reglas
  - Ejemplo (e-insertion)



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Técnicas de análisis morfológico

- Reglas en gramáticas de unificación
  - Ejemplo = DCGs de Prolog
  - Sistema APRES
- Sistema de recuperación de documentos en el entorno del manual de Unix
  - Los documentos son las páginas del manual, en inglés
- Integra técnicas clásicas de recuperación de información y técnicas avanzadas de PLN
- Parcialmente codificado en Prolog y hace uso de un analizador morfológico

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Técnicas de análisis morfológico

- Reglas en gramáticas de unificación
  - Ejemplo = DCGs de Prolog
  - Sistema APRES
    - Analizador morfológico
      - Categorías flexivas del inglés (nombre, verbo, adjetivo)
      - Deja una palabra en forma canónica (nombre singular, verbo infinitivo, adjetivo en forma base)
      - Se basa en un léxico obtenido de WordNet (que es una base de datos léxica, o diccionario conceptual)
      - Cubre el 100% de las palabras del manual de Unix

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Técnicas de análisis morfológico

- Reglas en gramáticas de unificación
  - Ejemplo = DCGs de Prolog
  - Sistema APRES
    - Analizador morfológico
      - `morf(noun,R,[l]) --> root(R),[s].`
      - `morf(noun,R,[s]) --> root(R),[s,e,s].`
      - `morf(noun,R,[x]) --> root(R),[x,e,s].`
      - `morf(noun,R,[z]) --> root(R),[c,e,s].`

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# Léxico y morfología

## Etiquetado sintáctico estocástico (POS-TAGGING)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Etiquetado sintáctico estocástico POS-TAGGING

- Introducción
- Aplicaciones
- Evaluación
- Taxonomía de métodos
- Modelos de Markov
- Resumen

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Introducción

- Etiquetado sintáctico = *Part-Of-Speech Tagging*, *POS-Tagging*
- Uno de los problemas más populares en PLN
  - Prerrequisito de análisis del LN
    - Primera fase del análisis sintáctico
  - Resurrección del PLN estadístico (90's)
    - Altos índices de efectividad (comparativamente)
      - Los etiquetadores alcanzan efectividades superiores al 95% de acierto

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Introducción

- Definición
  - Selección de la etiqueta sintáctica más probable para una palabra en un contexto
  - O de la secuencia de etiquetas para una secuencia de palabras
  - Las palabras aisladas son ambiguas respecto a su etiqueta sintáctica
    - Etiquetas = Nombre, Verbo, Adjetivo, etc.
    - Problema de desambiguación (sintáctica)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Introducción

- Ejemplo [Rodríguez]

Yo        bajo    con        el hombre    bajo    a  
PP        VM        VM        SP        TD        NC        VM        VM        NC        NC        SP

          VM        VM        AQ        NC        SP

tocar    el        bajo    bajo    la        escalera    .  
VM        VM        TD        VM        VM        TD        NC        NC        NC        FP

          VM        AQ        NC        SP

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Introducción

- Ejemplo [Rodríguez]

Yo        bajo    con        el hombre    bajo    a  
PP        VM        VM        SP        TD        NC        VM        VM        NC        NC        SP

          VM        VM        AQ        NC        SP

tocar    el        bajo    bajo    la        escalera    .  
VM        VM        TD        VM        VM        TD        NC        NC        NC        FP

          VM        AQ        NC        SP

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Introducción

- Las etiquetas sintácticas no son un concepto artificial
  - Dos palabras pertenecen a la misma categoría si y solo si reemplazar una con otra no cambia la “gramaticalidad” de la oración
  - gramaticalidad. 1. f. Ling. Cualidad de una secuencia de palabras o morfemas por la que se ajusta a las reglas de la gramática. (DRAE)  
“The \_\_\_\_\_ is angry.”
  - Nótese similitud con sinonimia

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Introducción

- Sistemas de etiquetas
  - Dependientes del idioma y/o *corpus*
  - Sistemas básicos (e.g. N, V, etc.)
    - Reflejan sólo el rol sintáctico
    - Más sencillos, reducidos, eficientes
  - Sistemas sofisticados (e.g. NN, NC, etc.)
    - Reflejan clasificaciones (nombre común vs. propio, modalidad, temporalidad, etc.)
    - Más complejos, detallados
    - Se pueden colapsar etiquetas (N\* => N)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Introducción

- Penn Treebank (inglés)

1. CC Coord. Conjunction
2. CD Cardinal number
3. DT Determiner
6. IN Prep. / subord. conj
7. JJ Adjective
8. JJR Comp. adjective
9. JJS Superlative adjective
12. NN Noun, sing. or mass
13. NNS Noun, plural
14. NNP Proper noun, sing.
18. PRP Personal pronoun
20. RB Adverb
22. RBS Superlative Adverb
27. VB Verb, base form
28. VBD Verb, past tense
29. VBG Verb, gerund/pres. partic.
30. VBN Verb, past participle
31. VBP Verb, non-3s, present
32. VBZ Verb, 3s, present
33. WDT Wh-determiner

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Introducción

- Estandarización EAGLES

- A nivel europeo, independiente idioma
- Usada CLIC-TALP
- Etiquetas + atributos  
*deportiva\_aq0fs0* y  
*sentimental\_aq0cso*

Atributo	Valor	Código
Categoría	Adjetivo	A
Tipo	Calificativo	Q
	Ordinal	O
Apreciativo	Sí	A
Género	Masculino	M
	Femenino	F
	Común	C
Número	Singular	S
	Plural	P
Participio	Invariable	N
	Sí	P

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Aplicaciones

- [Manning, Márquez]
- Al servicio de otras tareas
- Análisis del LN basado en conocimiento (comprensión), tareas y aplicaciones
  - Análisis sintáctico
    - Eficiencia = reducción del número de análisis (parciales) potenciales
  - Traducción automática
    - dog\_N => perro / dog\_V => perseguir

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Aplicaciones

- **PLN estadístico (clasificación), tareas**
  - Análisis superficial o parcial
    - Detección de sintagmas nominales – *NP bracketing*
    - Agrupamiento sintáctico – *chunking* (detección de grupos sintácticos no anidados)
      - [NP He ] [VP reckons ] [NP the current account deficit ]
      - [VP will narrow ] [PP to ] [NP only # 1.8 billion ] [PP in ]
      - [NP September ] .
  - Realizable *efectivamente* con etq. sintácticas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Aplicaciones

- **PLN estadístico**
  - Clasificación de documentos (recuperación, categorización, etc.)
  - Unidades de indexación (palabra\_etiqueta, sintagmas nominales, etc.)
    - Especialmente en dominios técnicos (medicina, etc.)
  - Respuesta a preguntas – *Question Answering*
    - Los sintagmas nominales son candidatos a respuestas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Aplicaciones

- **PLN estadístico**
  - Extracción de información
    - Texto => Registros estructurados
      - Qué, quién, cómo, etc. en noticias sobre atentados terroristas en Hispano América (MUC)
      - Catálogos comerciales, noticias bursátiles, etc.
    - Cascada de procesadores lingüísticos, que incluyen (de manera crítica) el *etiquetado sintáctico*, análisis superficial, desambiguación del significado, etc.

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Evaluación

- Deben evaluarse múltiples aspectos
  - Eficiencia (el etiquetador es sólo una parte del proceso)
  - Portabilidad (a otros idiomas, dominios)
- Usualmente centrada en la efectividad
  - Métricas
  - Línea base
  - Dificultad del problema
  - Colecciones etiquetadas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Evaluación

- Efectividad – métricas
  - Basadas en número de aciertos
  - Etiquetado completo => exactitud, error
  - Etiquetado incompleto => cobertura, precisión,  $F_1$
  - Conviene
    - Centrarse sólo en palabras ambiguas
    - Desglosar resultados por
      - Categorías sintácticas (N vs. ADV)
      - Tipos de ambigüedad (NV vs. ADV/ADJ)
      - Secuencias ambiguas (e.g. DET N ADJ)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Evaluación

- **Efectividad – línea base**
  - Asignar a cada palabra su etiqueta más frecuente (en el corpus)
    - Exactitud 90%
    - Sustancialmente más eficaz que otras tareas
    - No hay mucho espacio de mejora
    - Pero e.g. 95% en artículos periodísticos (longitud media de oración = 20 palabras) => un error por oración
- **Efectividad – dificultad del problema**
  - + etiquetas => + ambigüedad => + dificultad

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Evaluación

- **Efectividad – colecciones de evaluación**
  - Inglés
    - Brown Corpus (1M, inglés americano, 1979)
    - London-Oslo-Bergen (1M, inglés británico, 1979)
    - Wall Street Journal (300M, inglés americano)
    - British National Corpus (100M, inglés británico)
  - Español
    - LexEsp (5.5M)
    - Real Academia Española (200M)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

- [Márquez]
  1. Lingüísticos (o basados en conocimiento)
  2. Estadísticos o estocásticos (o basados en modelos del lenguaje)
    - Pueden ser considerados subconjunto de los siguientes
  3. Basados en aprendizaje
    - Problemas críticos de 2 y 3
      - Escasez de datos, sucesos no vistos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

## 1. Etiquetadores lingüísticos

- Conocimiento lingüístico de expertos
- Basados en reglas ( $\approx$  1k)
- Construidos manualmente
  - Ejemplos
    - TOSCA, EngCG (inglés)
    - EusCG (euskera)
    - SpaCG (español)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

## 1. Etiquetadores lingüísticos

- Ventajas
  - Riqueza y expresividad de las reglas lingüísticas
  - Excelentes resultados (EngCG > 99% exactitud)
- Desventajas
  - Alto coste de desarrollo (adquisición del conocimiento)
  - No transportables
  - Menos eficientes

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

## 1. Etiquetadores lingüísticos – EngCG

- *Constraint Grammar* = secuencia de sub-gramáticas
- Sub-gramática = serie de restricciones (*constraints*) que establecen condiciones de contexto (@w =0 VFIN (-1 TO))
  - descarta la categoría VFIN si la palabra anterior es “to”
- ENGCG (ENGTWOL)
  - 1100 restricciones
  - 93-97% de las palabras quedan totalmente desambiguadas
  - 99.7% corrección
  - Reglas heurísticas aplicables sobre el residuo
  - 2-3% ambigüedad residual con 99.6% de precisión

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

## 2. Etiquetadores estadísticos

- Modelos del lenguaje y generalizaciones adquiridos automáticamente
  - A partir de un corpus etiquetado manualmente
  - *Data-driven taggers*
- Uso de inferencia estadística
  - Modelos probabilísticos
- Técnicas procedentes del tratamiento del habla
  - Transmisión de señal sobre un canal con ruido

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

## 2. Etiquetadores estadísticos

- Ventajas
  - Marco teórico bien fundamentado
  - Aproximación clara, modelos simples
  - Exactitud aceptable (> 97%)
  - Independencia de la lengua
- Desventajas
  - Dificultades de aprendizaje del modelo
    - Escasez/inexistencia de datos
  - Menor precisión que lingüísticos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

## 2. Etiquetadores estadísticos

- Modelos del lenguaje de tipo n-gramas
- Modelos de Markov
  - Visibles (algoritmo de Viterbi)
  - Modelos ocultos de Markov (*Hidden Markov Models*)
    - Baum-Welch
    - Los más populares (e.g. Xerox tagger, multi-lenguaje)
- Máxima probabilidad (*Maximum Likelihood*)
- Se basan en estimar la probabilidad de una secuencia observada de sucesos

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

## 3. Etiquetadores basados en aprendizaje

- Se contempla el problema como genérico de aprendizaje
  - Atributos (lingüísticos), ejemplares, selección, algoritmos de aprendizaje
- Enfoque más uniforme
- Ejemplos
  - TreeTagger, MX-POST

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

## 3. Etiquetadores basados en aprendizaje

- Atributos (lingüísticos) en contexto
  - Etiquetas anteriores, unidades léxicas anteriores y posteriores, datos de la palabra actual (morfología, mayúsculas, etc.)
  - Tamaño de la ventana ( $\pm 2$ )
- Ejemplares = vectores atributo-valor
- Selección = unidades léxicas más frecuentes
- Algoritmos de aprendizaje = Árboles de decisión, Entropía Máxima, aprendizaje basado en ejemplares, etc.

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

## 3. Etiquetadores basados en aprendizaje -

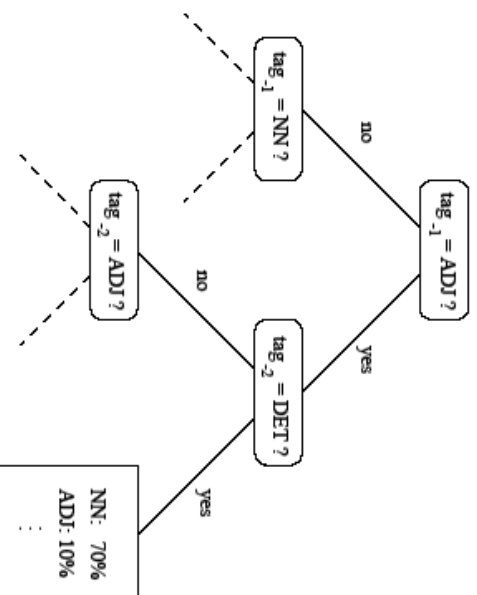
### TreeTagger

- Alemán, inglés, francés, griego, italiano
- Atributos = 2 etiquetas anteriores, sufijos
- Algoritmos de aprendizaje = Árboles de decisión ID3
- Efectividad = 96.32% vs. 96.06% de un etiquetador basado en trigramas

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

# POST - Taxonomía de métodos

## 3. Etiquetadores basados en aprendizaje - TreeTagger



Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- Método de modelado de procesos estadísticos secuenciales
- Desarrollados por Andrei A. Markov (estudiante de Chebyshev), 1913
  - Para modelar secuencias de letras en literatura rusa
- Usados en PLN para
  - Modelos de producción lingüística
  - Etiquetado sintáctico
  - Reconocimiento de habla
  - Múltiples aplicaciones (extracción de información, etc.)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- Usados con secuencias de variables estadísticas *con ciertas propiedades*
  - Sea  $X = (X_1, \dots, X_T)$  variables aleatorias secuenciales, tomando valores en  $S = (s_1, \dots, s_N)$
  - Propiedades de Markov
    - Horizonte limitado
      - $P(X_{t+1}=s_k | X_1, \dots, X_t) = P(X_{t+1}=s_k | X_t)$
    - Invarianza con el tiempo
      - $P(X_{t+1}=s_k | X_t) = P(X_2=s_k | X_1)$
  - $X$  es una cadena de Markov

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- Objetivo = calcular la *secuencia de etiquetas más probable* para una oración dada
- Propiedades de Markov en POST
  - Horizonte limitado
    - $P(t_{i+1}|t_1, i) = P(t_{i+1}|t_i)$
  - Invarianza con el tiempo
    - $P(t_{i+1}|t_i) = P(t_2|t_1)$
  - Simplificaciones que no se cumplen en general

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- Regla de Bayes

$$\begin{aligned} t_{1,n}^{opt} &= \operatorname{argmax}_{t_{1,n}} P(t_{1,n} | w_{1,n}) = \operatorname{argmax}_{t_{1,n}} \frac{P(w_{1,n} | t_{1,n}) P(t_{1,n})}{P(w_{1,n})} \\ &= \operatorname{argmax}_{t_{1,n}} P(w_{1,n} | t_{1,n}) P(t_{1,n}) \end{aligned}$$

- Reducimos probabilidades a parámetros estimables con corpus de entrenamiento

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- Asumiendo
  - Las palabras son independientes entre si
  - Una palabra sólo depende de su etiqueta

$$\begin{aligned} P(w_{1,n} | t_{1,n}) P(t_{1,n}) &= \prod_{i=1}^n P(w_i | t_{1,n}) \prod_{i=2}^n P(t_i | t_{1,i-1}) \\ &= \prod_{i=1}^n P(w_i | t_i) \prod_{i=2}^n P(t_i | t_{i-1}) = \prod_{i=1}^n P(w_i | t_i) P(t_i | t_{i-1}) \end{aligned}$$

- Definiendo  $P(t_1 | t_0) = 1$  por notación

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- En conclusión

$$t_{1,n}^{opt} = \operatorname{argmax}_{t_{1,n}} \prod_{i=1}^n P(w_i | t_i) P(t_i | t_{i-1})$$

- Usamos EMV sobre colección entrenamiento

$$P(w^k | t^j) = \frac{N(w^k, t^j)}{N(t^j)} \quad P(t^m | t^j) = \frac{N(t^m, t^j)}{N(t^j)}$$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

Palabra	AT	BEZ	IN	NN	VB	PER
bear	0	0	0	10	43	0
is	0	10065	0	0	0	0
move	0	0	0	36	133	0
on	0	0	5484	0	0	0
president	0	0	0	382	0	0
progress	0	0	0	108	4	0
the	69016	0	0	0	0	0
.	0	0	0	0	0	48809

Primera etiqueta	Segunda etiqueta					
	AT	BEZ	IN	NN	VB	PER
AT	0	0	0	48636	0	19
BEZ	1973	0	426	187	0	38
IN	43322	0	1325	17314	0	185
NN	1067	3720	42470	11773	614	21392
VB	6072	42	4758	1476	129	1522
PER	8016	75	4656	1329	954	0

P(AT NN BEZ IN AT NN | the bear is on the move) = ?

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- Limitación práctica
  - Número exponencial de secuencias de etiquetas
- Se usa el algoritmo de Viterbi
  - Procede de los modelos *ocultos* de Markov
  - Programación dinámica
  - Cómputo de dos funciones
    - $\delta_i(t)$  = probabilidad de la etiqueta  $t$  en la palabra  $w_i$
    - $\psi_{i+1}(t)$  = etiqueta más probable para  $w_i$  habiendo asignado  $t$  a  $w_i$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- Viterbi
  1. Inicialización:  $\delta_1(PEP) = 1, \delta_1(t) = 0$  si  $t \neq PEP$
  2. Inducción
$$\delta_{i+1}(t^j) = \max_{1 \leq k \leq T} [\delta_{i+1}(t^j) \times P(w_{i+1} | t^j) \times P(t^j | t^k)], 1 \leq j \leq T$$
$$\psi_{i+1}(t^j) = \operatorname{argmax}_{1 \leq k \leq T} [\delta_{i+1}(t^j) \times P(w_{i+1} | t^j) \times P(t^j | t^k)], 1 \leq j \leq T$$
  3. Predicción
$$X_n = \operatorname{argmax}_{1 \leq j \leq T} \delta_n(t^j), \text{ y } X_i = \psi_{i+1}(X_{i+1}) \text{ para } 1 \leq i \leq n$$
$$P(X_1, \dots, X_n) = \max_{1 \leq j \leq T} \delta_{n+1}(t^j)$$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- Tratamiento de *palabras desconocidas*
  - Palabras no presentes en el entrenamiento
  - Frecuentemente marcan la diferencia en efectividad
  - Enfoque simple = etiqueta (abierta) más frecuente
    - Poco efectivo, desaprovecha información lingüística
  - Enfoques más sofisticados usan múltiples fuentes de información
    - Frecuencia, flexión, ortografía (capitalización)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- Ejemplo
  - Frecuencia, flexión, ortografía (capitalización)

$$P(w^k | t^i) = \frac{1}{Z} P(des | t^i) P(may | t^i) P(suf | t^i)$$

des = desconocida, may = mayúsculas, suf = sufijo

- Reducción de errores del 40% al 20%

## POST - Modelos de Markov

- Ejemplo

Atributo	Valor	NNP	NN	NNS	VBG	VBZ
desconocida	sí	0,05	0,02	0,02	0,005	0,005
	no	0,95	0,98	0,98	0,995	0,995
mayúsculas	sí	0,95	0,10	0,10	0,005	0,005
	no	0,05	0,90	0,90	0,995	0,995
sufijo	-s	0,05	0,01	0,98	0,00	0,99
	-ing	0,01	0,01	0,00	1,00	0,00
	-tion	0,05	0,10	0,00	0,00	0,00
	otro	0,89	0,88	0,02	0,00	0,01

$$P(\text{fenestration}^k) = ?, \quad P(\text{fenestates}^k) = ?,$$

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST - Modelos de Markov

- Los modelos anteriores se pueden extender a bigramas y trigramas de palabras
  - Aumentamos efectividad
  - Tagger de Church (1988)
- Modelos ocultos de Markov
  - Razonablemente más sofisticados y muy efectivos
  - Estándar en POS-Tagging

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## POST – Resumen

- Tarea básica para otras tareas de análisis y comprensión
- Problema bien definido, muy popular
- Línea base alta (90%), pero mejorable
- Manifiestamente más fácil que otras tareas
- Métodos lingüísticos, estadísticos y basados en aprendizaje
- Efectividad actual alta (>97%)

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Léxico y morfología

### Bibliografía

## Bibliografía

- Morfología
  - [Martínez] Raquel Martínez. *Niveles de análisis. Análisis morfológico*. Apuntes del curso de doctorado Ingeniería Lingüística aplicada al Procesamiento de Documentos, <http://www.esct.urjc.es/~rmartine/IL.htm>

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Bibliografía

- Técnicas de análisis morfológico
  - [Jurafsky] D. Jurafsky, J. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2000 – Capítulo 3
  - [Rodríguez] Horacio Rodríguez. *Morfología*. Apuntes de PLN, <http://www.lsi.upc.es/~horacio/pln.html>

Procesamiento del Lenguaje Natural – José María Gómez Hidalgo – Universidad Europea de Madrid

## Bibliografía

- POST
  - [Manning] C. Manning, H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999 – Capítulo 10.
  - [Márquez] Lluís Màrquez. *POS Tagging: A Machine Learning Approach based on Decision Trees*. PhD thesis. Dep. LSI. Universitat Politècnica de Catalunya (UPC), 1999.
  - [Rodríguez] Horacio Rodríguez. *Tagging*. Apuntes de PLN, <http://www.lsi.upc.es/~horacio/pln.html>