Análisis Morfológico

- Introducción
- Morfología
- Análisis Morfológico
- Técnicas de Estados Finitos
- Morfología de dos niveles
- Stemming
- Aprendizaje automático de la morfología

Introducción

- Morfología
  - Flexión
  - Derivación
  - Composición
- Resultado
  - categorización morfosintáctica
    - Ej. categorías Eagles ——— ej. VMIP1S0
    - Ej. Penn Treebank tagset ——— ej. VBD
  - rasgos morfológicos
- Problemas
  - alteraciones fonológicas
  - morfotáctica

Análisis Morfológico

- Problemas
  - sufijos flexivos vs. sufijos derivativos
  - la derivación implica a veces cambio semántico que además no es siempre predecible
    - ej. extensiones de significado
    - reglas léxicas
  - Un sufijo derivativo puede ir seguido de su flexión
    - amar => amante => amantes
  - La flexión no cambia la categoría gramatical, la derivación a veces si
  - La flexión afecta a otras palabras de la oración
    - concordancia

Morfología, Modelos Computacionales

- Funciones
  - Flexión, Derivación, Composición
- Morfotáctica
  - Reglas de formación de palabras
  - Combinaciones posibles entre morfemas
    - Encadenamiento simple
    - modelos complejos raiz/patrón
    - Regularidad y cercanía dependientes de la lengua
- Alteraciones fonológicas (Morfofonología)
  - cambios al unir los morfemas
  - origen: fonología, morfología, ortografía
  - variables en número y complejidad
  - p.ej. armonía vocálica

Morfemas

- 1 morfema:
  - evitar
- 2 morfemas:
  - evitable = evitar + able
- 3 morfemas:
  - inevitable = in + evitar + able
- 4 morfemas:
  - inevitabilidad = in + evitar + able + idad

Morfología Flexiva

- número
  - house houses
  - cheval chevaux
  - casa casas
- tiempo verbal
  - walk walkes walked walking
  - amo amas aman amando
- género
  - niño niña

Morfología Derivativa

- Forma de la derivación
  - sin cambio — barcelonés
  - prefijación — inevitable
  - sufijación — importantísimo
  - infijación
- Origen
  - verbo => adjetivo — tardar => tardío
  - verbo => nombre — sufrir => sufrimiento
  - nombre => nombre — actor => actorazo
  - nombre => adjetivo — atleta => atlético
  - adjetivo => adjetivo — rojo => rojizo
  - adjetivo => adverbio — alegre => alegremente

Morfología Derivativa

- Forma de la derivación
  - sin cambio — barcelonés
  - prefijación — inevitable
  - sufijación — importantísimo
  - infijación
- Origen
  - verbo => adjetivo — tardar => tardío
  - verbo => nombre — sufrir => sufrimiento
  - nombre => nombre — actor => actorazo
  - nombre => adjetivo — atleta => atlético
  - adjetivo => adjetivo — rojo => rojizo
  - adjetivo => adverbio — alegre => alegremente

Análisis Morfológico

- Tipos de analizadores morfológicos
- formarios → Maco+, Freeling; Atserias et al, 1998
  - eficiencia
  - poca variación (ej. inglés)
  - extensibilidad
  - construcción a partir de un generador morfológico
  - Mal para lenguas muy flexivas
  - Mal cuando hay derivación, composición
- técnicas de estados finitos → Roche,Schabes, 1997; Kornai, 1999
  - autómatas
    - analizadores de un nivel → Martí, 1988
  - transductores
    - analizadores de dos o más niveles → Koskenniemi, 1983; Sproat, 1993

Modelos de cómputo

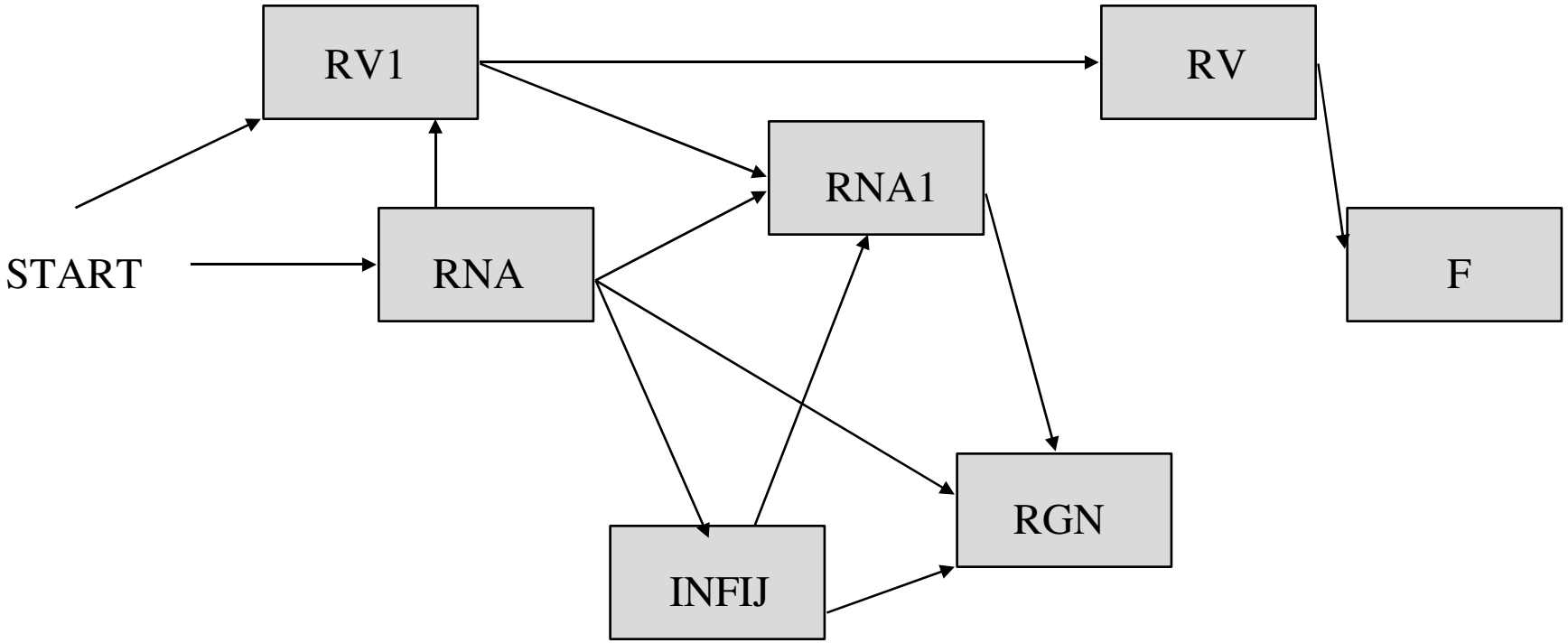
- Mezcla de conocimiento lingüístico y algorítmico
- Sistemas automáticos
- Multilingüismo
- Corpus
- Problema de la eficiencia
- Problema de la sobregeneración

Modelos de cómputo

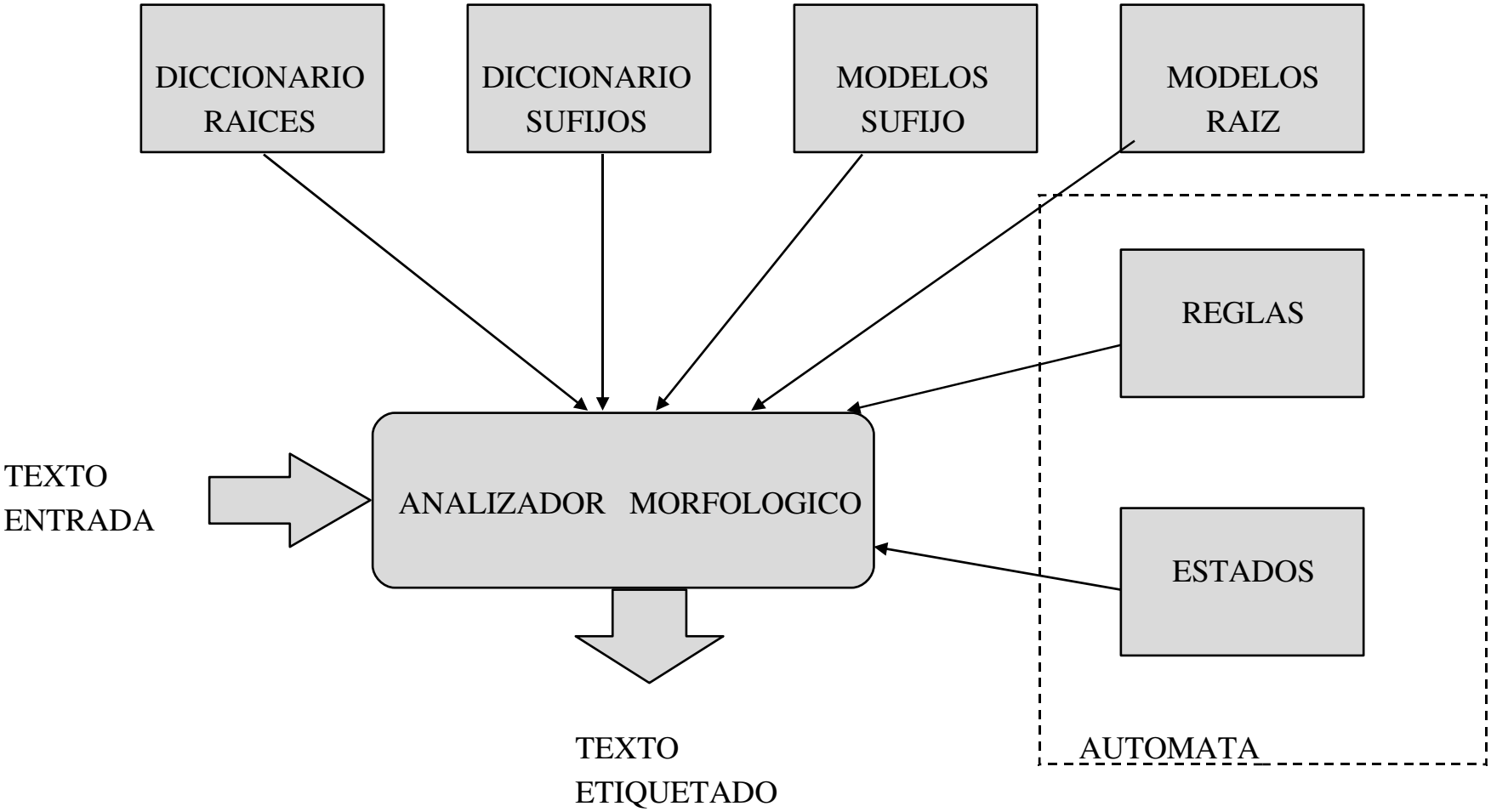
- Criterios de clasificación
  - Poder descriptivo
    - Flexión, Derivación, Composición
    - Análisis y Generación
  - Enfoque
    - Basados en léxico
    - Basados en paradigma (Calder 89)
  - Tratamiento de la morfotáctica
    - Estados Finitos
    - Unificación
  - Tratamiento de la morfofonología
    - Estados Finitos
    - Métodos ad-hoc
  - Elementos del léxico
    - Morfemas
    - Segmentos de palabra

Morfología de un nivel

Sistema AMCAS (Marti 89)



SISTEMA AMCAS (1)



SISTEMA AMCAS (2)

DICCIONARIO DE RAICES (FRAGMENTO)			
RAIZ	MODELO	PROPIEDADES	DIVISIBLE
"d"	D	(("B1" "DORW")("TVM" "VI") ("SEM" "DECIR-1"))	nil
"de"	PREP	()	nil
"del"	PREP	()	nil
"deposit"	AM	(("TGN" "OM") ("BL" "&3"))	nil
"dese"	AM	(("TGN" "OM") ("B1" "OSOJ") ("TVM" "VI"))	nil
"dich"	HECH	(("TGN" "OM") ("CONJ" "3"))	nil
"dich"	DETN	(("DET" "DEM"))	nil
"dich"	PRON	(("PRN" "DEM"))	nil
"diner"	NOM	("B1" "DAF") ("B2" "ALM") ("SEM" "DINERO-1"))	nil
"directori"	NOM	()	nil

SISTEMA AMCAS (3)

DICCIONARIO DE SUFIJOS (FRAGMENTO)			
SUFIJO	MODELO	PROPIEDADES	DIVISIBLE
"a"	AASAM	(("NUM" "SG"))	nil
"a	AASFEM	(("NUM" "SG"))	nil
"a"	GAF	(("GEN" "FEM")("NUM" "SG"))	nil
"a"	GAM	(("NUM" "SG"))	nil
"a"	GBAJ	(("GEN" "FEM")("NUM" "SG"))	nil
"a"	GBAW	(("GEN" "FEM")("NUM" "SG"))	nil
"a"	GN1	(("GEN" "FEM")("NUM" "SG"))	nil
"a"	GOAJ	(("GEN" "FEM")("NUM" "SG"))	nil
"a"	IMP	(("NUM" "SG")("PERS" "2"))	nil
"a"	IPO	(("NUM" "SG")("PERS" "3"))	nil
"a"	SP2	(("NUM" "SG")("PERS" "1/3"))	nil
"aba"	IMA	(("PERS" "1")("NUM" "SG"))	nil
"lo"	PROE	(("ENCL" "LO")("BL" "&1"))	nil
"me"	PROE	(("BL" "&1"))	nil
"&"	GBF&1	(("NUM" "SG"))	nil
"&"	GBM&1	(("NUM" "SG"))	nil

SISTEMA AMCAS (4)

DICCIONARIO DE MODELOS DE RAIZ (FRAGMENTO)	
MODELO	PROPIEDADES
CSS D DETD1	(("CAT" "CONJ") ("TCON" "CSS") ("BL" "SI")) (("CAT" "VERB") ("TV" "D")) (("CAT" "DET") ("PERS" "1") ("TGN" "EAO") ("BL" "SI") ("DET" "DEM"))

SISTEMA AMCAS (5)

DICCIONARIO DE MODELOS DE SUFIJO (FRAGMENTO)	
MODELO	PROPIEDADES
AASAM	(("GEN" "AMBI"))
AASFEM	(("GEN" "FEM") ("CAT" "ADJ"))
GAF	(("CAT" "NOM") ("GEN" "FEM"))
GAM	(("CAT" "NOM") ("GEN" "MASC"))
GBAJ	(("CAT" "ADJ"))
GBAW	(("CAT" "ADJ"))
GN1	()
GOAJ	(("CAT" "ADJ"))
IMP	(("CAT" "VERB") ("TEMP" "PRES") ("PROE" "SI") ("MODO" "IMP"))
IPO	(("CAT" "VERB") ("TEMP" "PRES") ("BL" "SI") ("MODO" "IND"))
PROE	()
SP2	(("CAT" "VERB") ("TEMP" "PRES") ("MODO" "SUBJ"))
&	()

SISTEMA AMCAS (6)

DICCIONARIO DE REGLAS (FRAGMENTO)			
EST_INI	EST_FIN	MODELO	CONDICIONES
RNA1	RGN	OOSMAS	(("TGN" "OAJ"))
RV	F	BL	(("BL" "SI"))
RV	F	&	(("BL" "&1"))
RV	RV	PROE	(("PROE" "SI"))
RV1	INFIJ	CC	(("U" "CCVD"))
RV1	RGN	AASFEM	(("T1" "OAJ"))
RV1	RGN	GAF	(("T1" "AF")("TGN" "AF"))
RV1	RGN	GAM	(("TGN" "AM"))
RV1	RGN	GOAJ	(("B1" "TOJ"))
RV1	RV	IMP	(("TV" "R")("TV" "ACUE) ("TV" "ADC") ("TV" "ADZ")("TV" "D"))
RV1	RV	IPO	(("TV" "PONG")("TV" "R")("TV" "HIZ") ("TV" "PUED"))
RV1	RV	SPB	(("TV" "PONG")("TV" "SEP")("TV" "D"))
START	RV1	D	()

Morfología de dos niveles

- [Koskenniemi 83] definió el modelo computacional de morfología de dos niveles:
  - Es un modelo general aplicable a cualquier lengua.
  - Es válido tanto para el análisis como para la síntesis.
  - Separa claramente el conocimiento lingüístico y el algoritmo.
  - Separa claramente el nivel superficial de la palabra a analizar o generar y el nivel léxico o profundo que es el que se representa en el sistema de diccionario (sistema léxico)
  - Utiliza un sistema de reglas paralelas en lugar de los sistemas de reglas de reescritura.
  - Elementos básicos: reglas y léxico (y el programa!)

Morfología de dos niveles

- Entrada:
  - forma
- Salida
  - lema + rasgos morfológicos

Input	Output
cat	cat + N + sg
cats	cat + N + pl
cities	city + N + pl
merging	merge + V + pres_part
caught	(catch + V + past) or (catch + V + past_part)

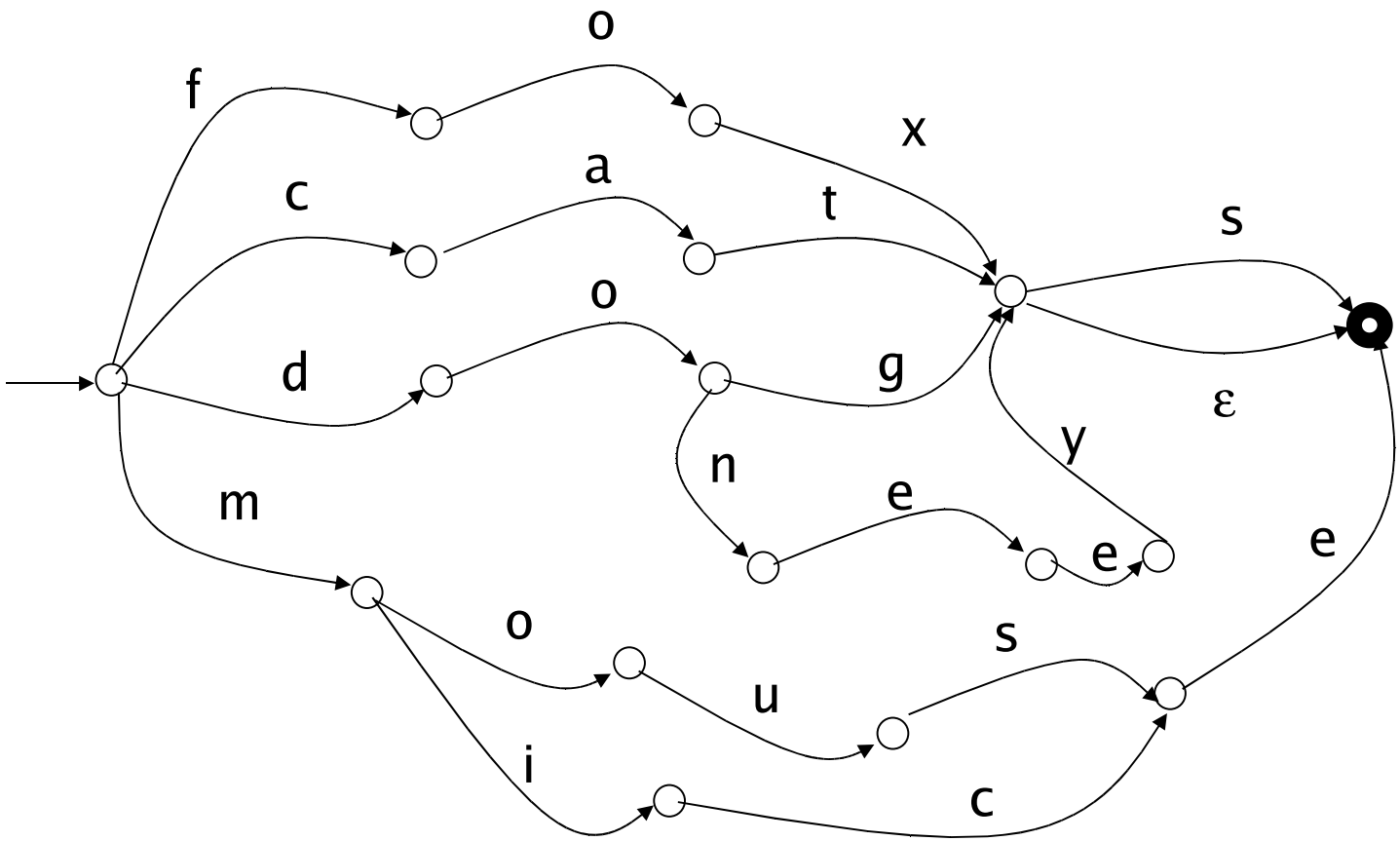
Morfología de dos niveles

Elementos del analizador

- Lexicon de morfemas
  - raiz (stem) + afijos
- Morfotáctica
  - qué combinaciones de morfemas son válidas
    - cats = cat + s
- Alteraciones fonológicas
  - Reglas ortográficas (spelling rules): cambios al producirse la combinación
    - city + s = cities

Morfología de dos niveles

Integración del lexicón y morfotáctica

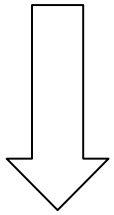


fox
cat
dog
donkey
mouse
mice

Morfología de dos niveles

Integración del lexicón y morfotáctica

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



c:c	a:a	t:t	+N:ε	+pl:s

Morfología de dos niveles

Utilización de un autómata de estados finitos (FST)

- 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

Morfología de dos niveles

Simplificaciones notacionales

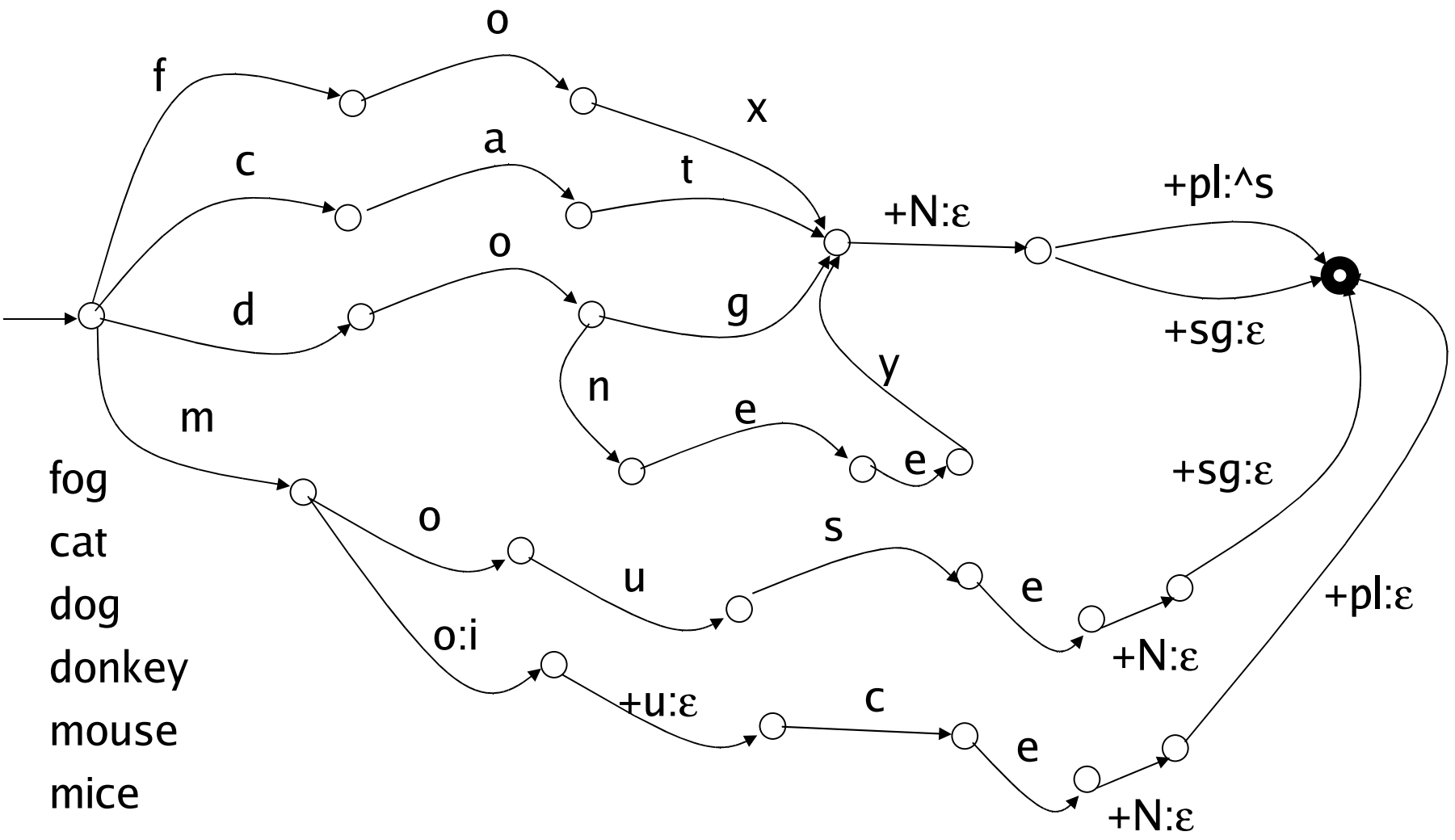
- default pairs
  - a:a
- morpheme separator +
- empty charracter ε or 0
- end of word #
- default correspondence pairs

a	b	c	…	z	‘	+	#
a	b	c	…	z	‘	ε	ε

- feasible pairs
  - default correspondences + explicit correspondences in the rules
- any @

Morfología de dos niveles

Integración del lexicón y morfotáctica (2)



Morfología de dos niveles

El sistema léxico (1)

- Se define un conjunto de morfemas para formar palabras
- Los morfemas se agrupan en subléxicos
- Cada entrada o morfema del léxico se define por:
  - Representación léxica: morfema (raíz)
  - Subléxico: id del subléxico al que pertenece el morfema
  - Clase de continuación: qué subléxicos siguen al morfema
  - Información morfológica: categoria, número, caso, ...
- El orden de los morfemas se controla mediante las clases de continuación
- Se establecen también cuáles son los subléxicos iniciales y finales

Morfología de dos niveles

El sistema léxico (2)

ALTERNATION Begin	RAICES_N RAICES_V
ALTERNATION Raiz_N1	SUFIJOS_N
ALTERNATION Fin	End

- El subléxico inicial es *begin* (cualquiera de los morfemas de RAICES_N RAICES_V) puede comenzar una palabra
- El morfema *coz* definido en el subléxico RAICES_N tiene como clase de continuación Raiz_N1 que representa a los morfemas del subléxico SUFIJOS_N
- +s es un morfema del subléxico SUFIJOS_N y su clase de continuación es Fin (no puede seguirle ningún otro sufijo)
- Por consiguiente, la concatenación de los morfemas coz+s está permitida por el sistema léxico

Morfología de dos niveles

El sistema léxico (3)

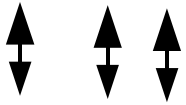
```
; raices_n.lex
\entrada coz
\sublexico RAICES_N
\continuacion Raiz_N1
\atributos
\glosa N(coz)

; sufijos.lex
\entrada +s
\sublexico SUFIJOS_N
\continuacion Fin
\atributos
\glosa +PL
```

Morfología de dos niveles

El sistema de reglas (1)

- maneja dos representaciones: la léxica y la superficial
- Las reglas no ejecutan nada, sólo establecen correspondencias entre los dos niveles.
- Reconocimiento: encontrar una representación léxica válida correspondiente a una forma superficial.
- Generación: parte de la representación léxica conocida y busca representaciones superficiales que se correspondan con ella.
- Reglas: RULE <correspondencia> <operador> <contextos>
- <contextos> = <contexto_izquierdo> _ <contexto_derecho>
- Ejemplo:
  - RULE +:e <=> x:x _ s:s
  - box + s (nivel léxico)

    
  - box e s (nivel superficial)

Morfología de dos niveles

El sistema de reglas (2)

- t:c => _ i "only but not always" (implica)
  - Léxico t corresponde a la superficie c sólo precediendo a i:i, pero no necesariamente siempre en este entorno
- t:c <= _ i "always but not only"
  - Léxico t siempre corresponde a la superficie c precediendo a i:i, pero no necesariamente sólo en este entorno
- t:c <=> _ i "always and only"
  - Léxico t siempre y solamente corresponde a la superficie c precediendo al entorno i:i
- t:c /<= _ i "never"
  - Léxico t nunca corresponde a la superficie c precediendo al entorno i:i

Morfología de dos niveles

El sistema de reglas (3)

- Regla para añadir una “e” epentética
- RULE 0:e <=> C +:0 _ s [+:0|#]
  - C = C:C (consonante a nivel léxico y superficial)
  - + marca de principio de sufijo
  - 0 carácter nulo
  - S = s:s
  - # fin de palabra
  - [+:0|#] opcionalidad
- Indica que cuando a nivel léxico un morfema acaba en consonante y el sihuiente morfema es s, entonces a nivel superficial se inserta una “e” epentética
- Nivel léxico: coz+0s
- Mivel superficial: coc0es

Morfología de dos niveles

El sistema de reglas (4)

- Regla de alteración z:c
- RULE z:c <=> _ +:0 0:e s
  - + marca el principio del sufijo
  - 0 carácter nulo
  - s = s:s
  - No hay contexto izquierdo!
- Nivel léxico: coz+0s
- Nivel superficial: coc0es
- Cuando nos encontramos a nivel léxico un carácter “z” seguido de la correspondencia “+s:es”, entonces a nivel superficial le corresponde una “c”.

Morfología de dos niveles

El sistema de reglas (5)

- Regla de alteración z:c
- RULE z:c <=> _ +:0 0:e s
  - + marca el principio del sufijo
  - 0 carácter nulo
  - s = s:s
- Nivel léxico: coz+0s
- Nivel superficial: coc0es
- Cuando nos encontramos a nivel léxico un carácter “z” seguido de la correspondencia “+s:es”, entonces a nivel superficial le corresponde una “c”.

Morfología de dos niveles

PCKimmo

- International Linguistics Center
- http://www.sil.org/pckimmo
- Conjunto de herramientas para la creación de analizadores morfológicos.
- KGEN
  - creación de autómatas a partir de reglas de alto nivel
- KTEXT
  - Analizador morfológico de texto
- ENGLEX
  - Lexicon con más de 20000 entradas para el análisis del Inglés

Morfología de dos niveles

```
[rigau@adimen PCKimmo]$ pckimmo
PC-KIMMO TWO-LEVEL PROCESSOR
Version 2.1.13 (October 25, 2002), Copyright 2002 SIL
Compiled Mar  5 2007 16:57:25
with PC-PATR functions version 1.3.12 (December 7, 2005)
Type ? for help
PC-KIMMO>load rules castellano.rul
Loading rules from castellano.rul
PC-KIMMO>load lexicon castellano.lex
Loading lexicon from castellano.lex
...
PC-KIMMO>recognize coces
coz+s     [N(coz)+PL]
PC-KIMMO>
```

Morfología de dos niveles

KGEN

- Generador de autómatas a partir de reglas
- Ejemplo: kgen < reglas.txt > castellano.rul
- Regla de alteración z:c
- RULE z:c <=> _ +:0 0:e s

	z	z	+	0	s	@
	c	@	0	e	s	@
1:	2	5	1	1	1	1
2.	0	0	3	0	0	0
3.	0	0	0	4	0	0
4.	0	0	0	0	1	0
5:	2	5	6	1	1	1
6:	2	5	1	7	1	1
7:	2	5	1	1	0	1

Stemming

- El procesamiento morfológico es costoso
- En Recuperación de la Información (IR, del Inglés Informarion Retrieval) puede ser muy interesante normalizar las formas superficiales (lápices => lápiz) para encontrar la raíz o forma canónica.
- En Inglés (al tener una morfología sencilla) puede ser muy útil utilizar un analizador como el de (Porter 1980)
- Se usan en cascada una serie de reglas de reescritura
- ATIONAL -> ATE : relational -> relate
- TIONAL -> TION : conditional -> condition, pero rational -> ration!
- BILITY -> BLE : sensibility -> sensible

Aprendizaje automático de la morfología

- Problema
  - Paradigma raíz + sufijos
  - Obtención de las raíces
  - Clasificación de las raíces en modelos
  - Descubrimiento de patrones o reglas de correspondencia entre pares de palabras
  - Son necesarios grandes volúmenes de texto
- Dos aproximaciones
  - Sin utilizar conocimiento morfológico alguno
    - Goldsmith 2001 (MDL),
    - un buén candidato a raíz lo es de muchas formas!
    - Berent 1999, Snover & Brent 2001, 2002
  - Cuando se dispone de cierto conocimiento morfológico
    - Oliver 2004 (Aplicado al servo croata y ruso)