Autómatas 2

Escuela Nacional de Antropología e Historia (ENAH)
Agosto – diciembre de 2015

Análisis (Parsing)

- Obtener una estructura lingüística que describa una entrada.
- going → VERB-go + GERUND-ing
- Análisis morfológico

Análisis (Parsing)

- En lugar de analizar, ¿por qué no mejor almacenar todas las formas posibles de todas las palabras de una lengua?

Lexicón (Lexicon)

- Lista de palabras.
- Lista de bases y afijos.

Autómata morfológico

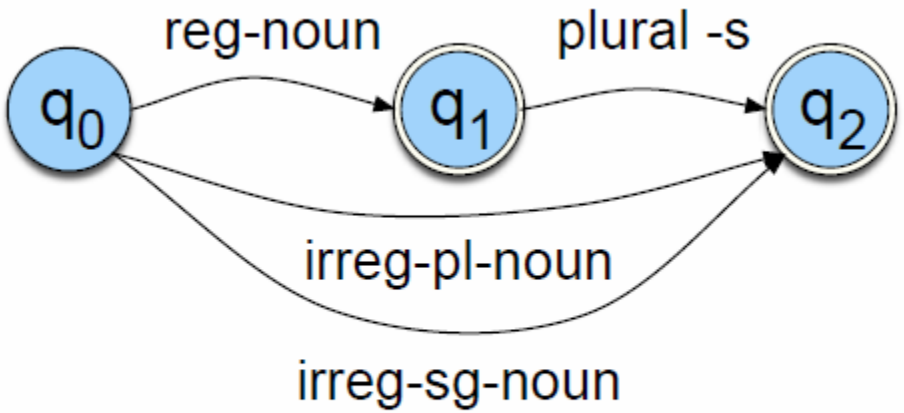
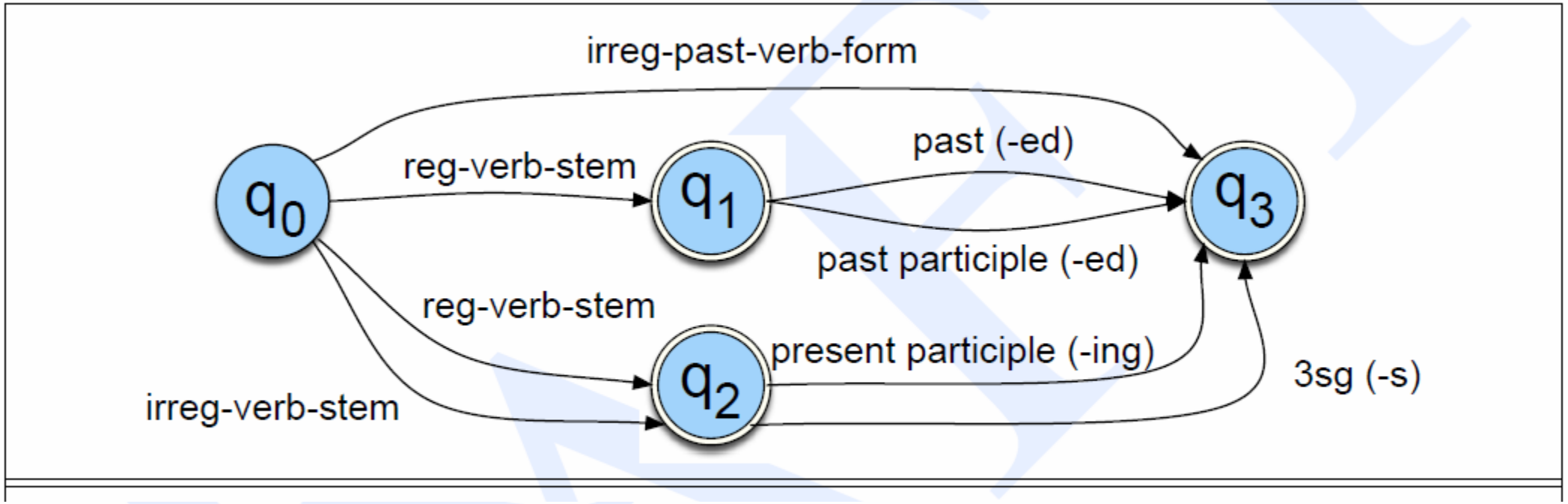


Figure 3.3 A finite-state automaton for English nominal inflection.

reg-noun	irreg-pl-noun	irreg-sg-noun	plural
fox	geese	goose	-s
cat	sheep	sheep	
aardvark	mice	mouse	

Autómata morfológico



reg-verb-stem	irreg-verb-stem	irreg-past-verb	past	past-part	pres-part	3sg
walk	cut	caught	-ed	-ed	-ing	-s
fry	speak	ate				
talk	sing	eaten				
impeach		sang				

Autómata morfológico

big, bigger, biggest,	cool, cooler, coolest, coolly
happy, happier, happiest, happily	red, redder, reddest
unhappy, unhappier, unhappiest, unhappily	real, unreal, really
clear, clearer, clearest, clearly, unclear, unclearly	

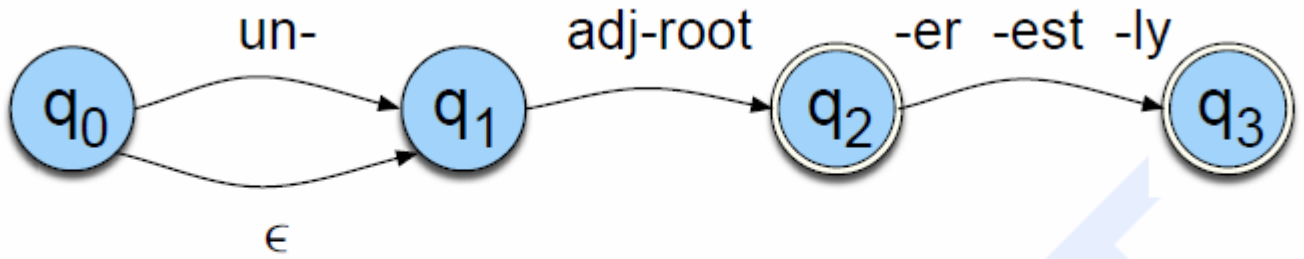


Figure 3.5 An FSA for a fragment of English adjective morphology: Antworth's Proposal #1.

Autómata morfológico

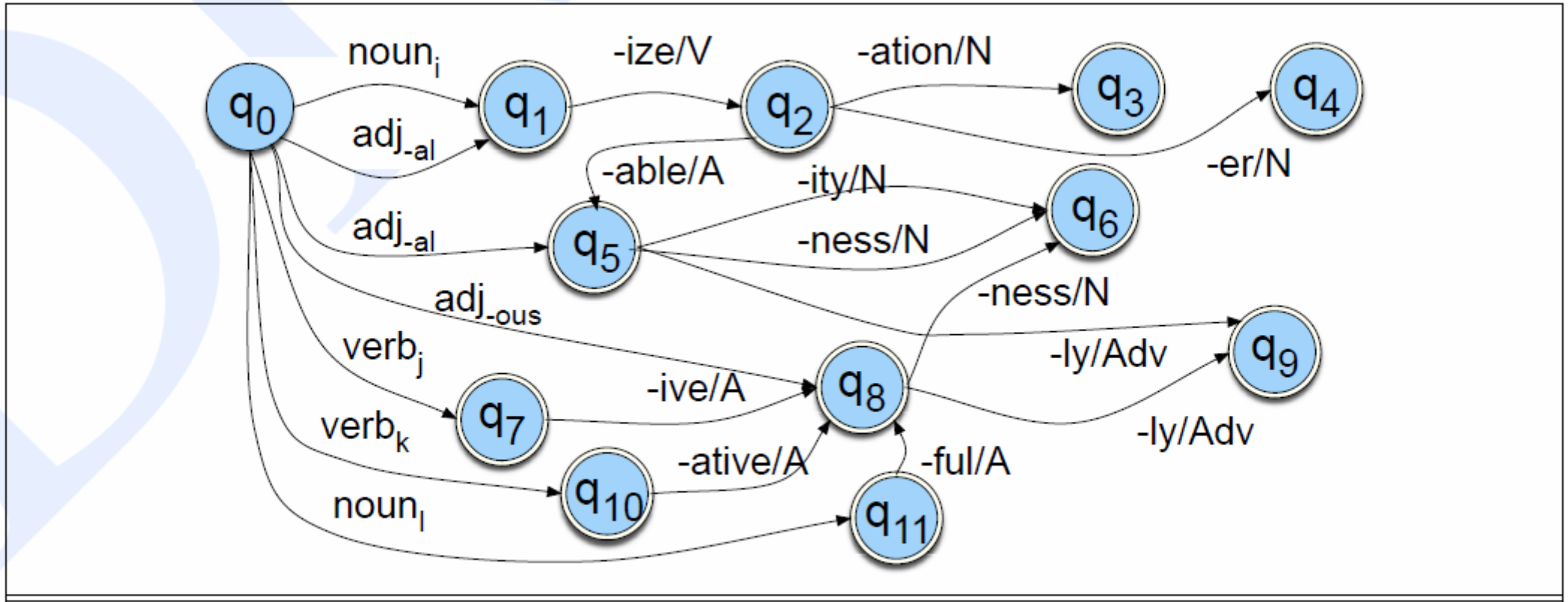


Figure 3.6 An FSA for another fragment of English derivational morphology.

Transductores de estados finitos (FST)

- Tipo de autómata de estados finitos que hace correspondencia entre dos conjuntos de símbolos.
- Aceptador: recibe pares de cadenas y los acepta.
- Generador: genera pares de cadenas.
- Traductor: recibe una cadena y genera otra de salida.

Transductores de estados finitos (FST)

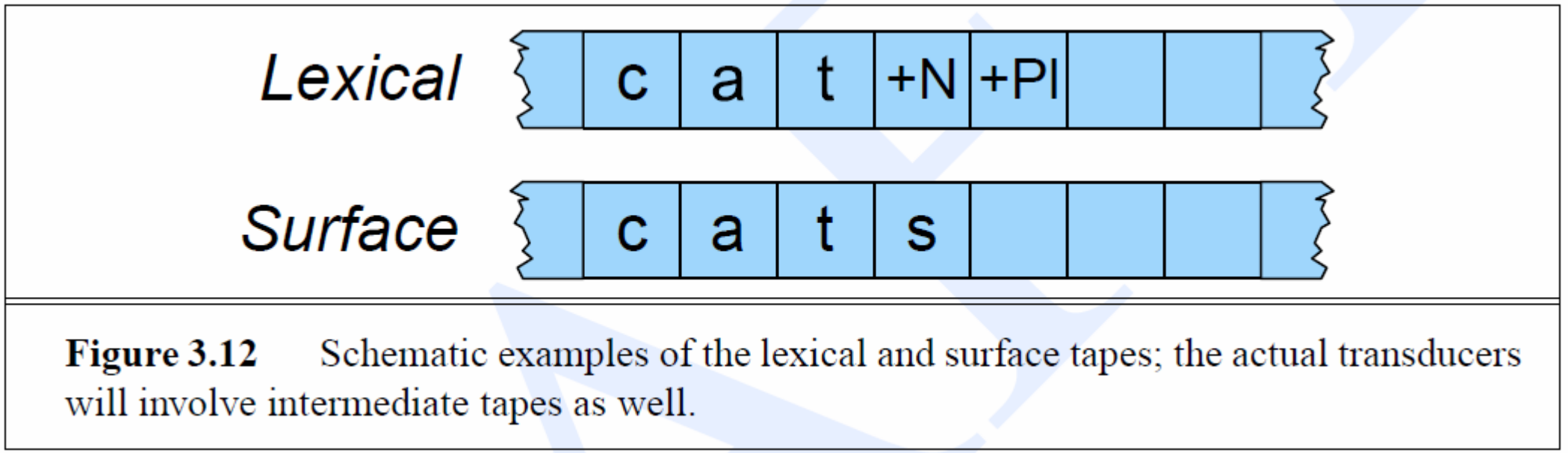


Figure 3.12 Schematic examples of the lexical and surface tapes; the actual transducers will involve intermediate tapes as well.

Transductores de estados finitos (FST)

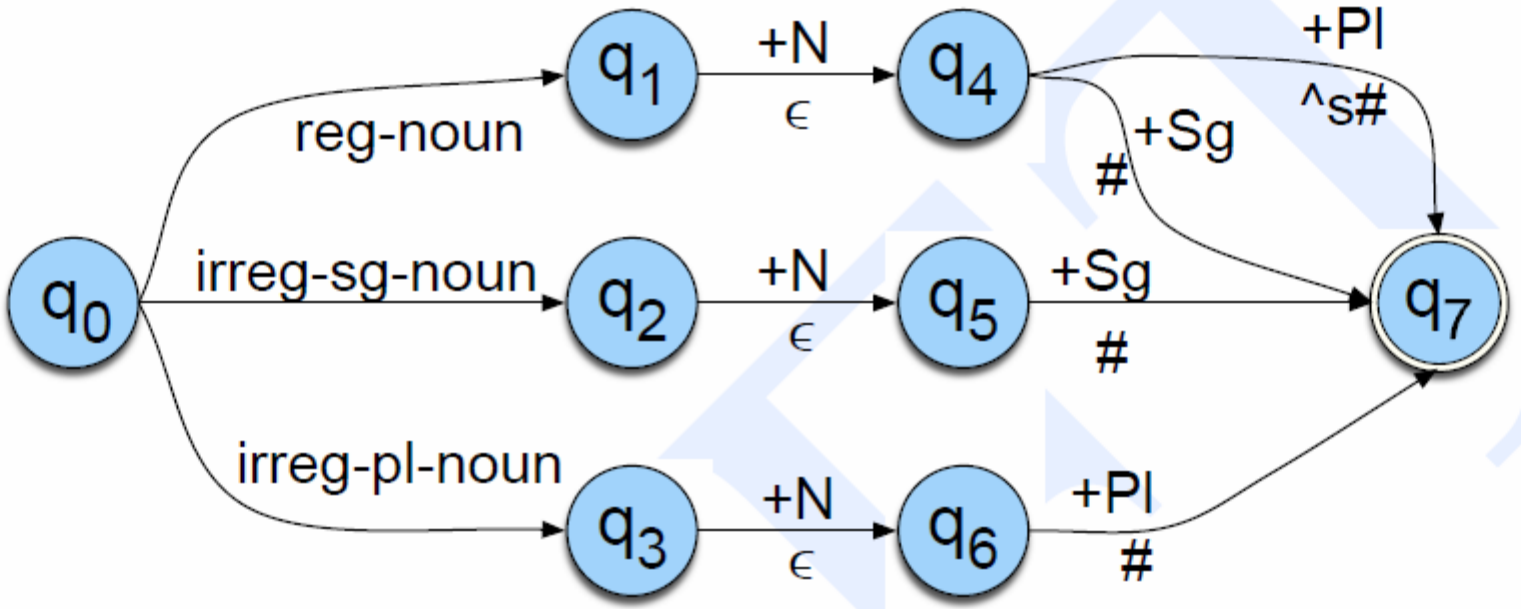


Figure 3.13 A schematic transducer for English nominal number inflection T_{num}. The symbols above each arc represent elements of the morphological parse in the lexical tape; the symbols below each arc represent the surface tape (or the intermediate tape, to be described later), using the morpheme-boundary symbol ^ and word-boundary marker #. The labels on the arcs leaving q_0 are schematic, and need to be expanded by individual words in the lexicon.

Transductores de estados finitos (FST)

reg-noun	irreg-pl-noun	irreg-sg-noun
fox	g o:e o:e s e	goose
cat	sheep	sheep
aardvark	m o:i u:ϵ s:c e	mouse

Transductores de estados finitos (FST)

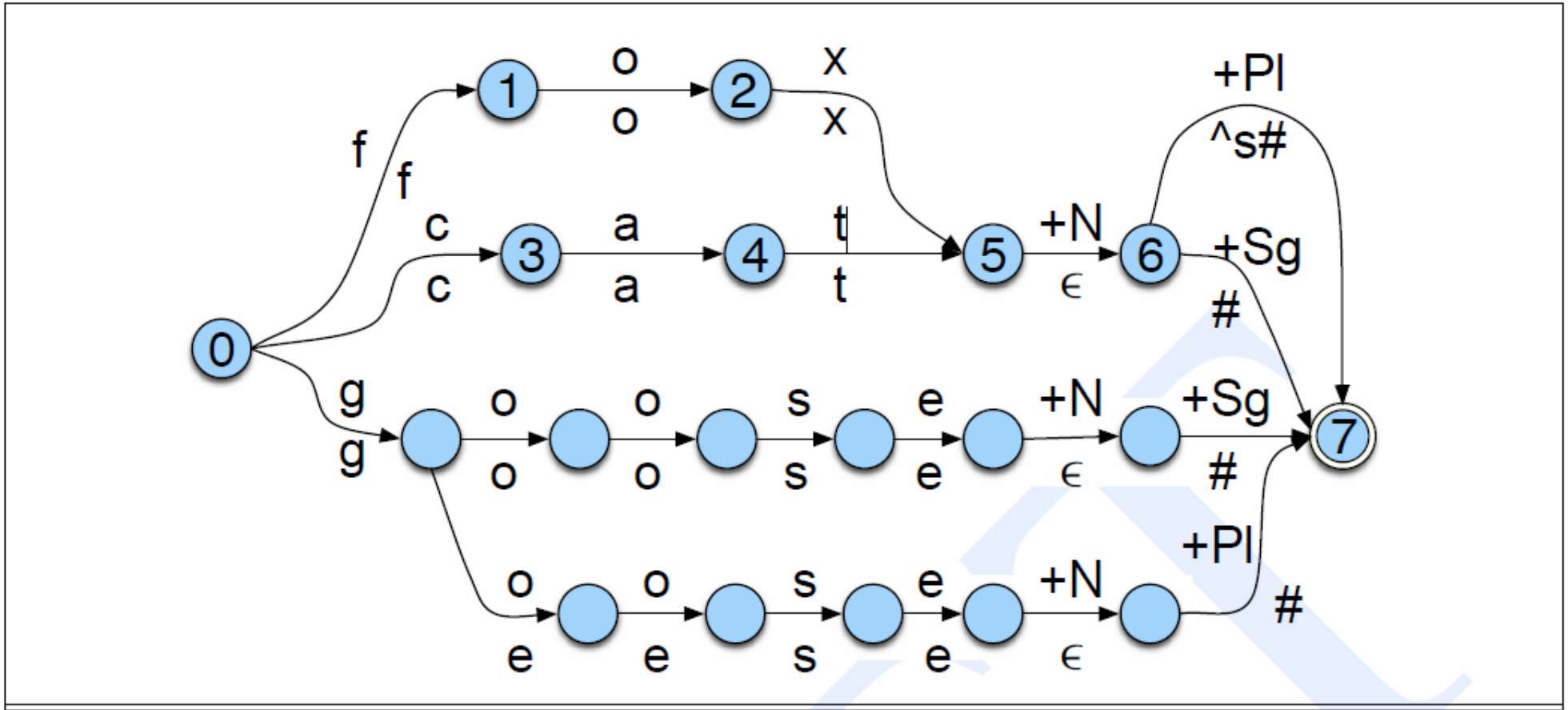


Figure 3.14 A fleshed-out English nominal inflection FST T_{lex}, expanded from T_{num} by replacing the three arcs with individual word stems (only a few sample word stems are shown).

Transductores de estados finitos (FST)

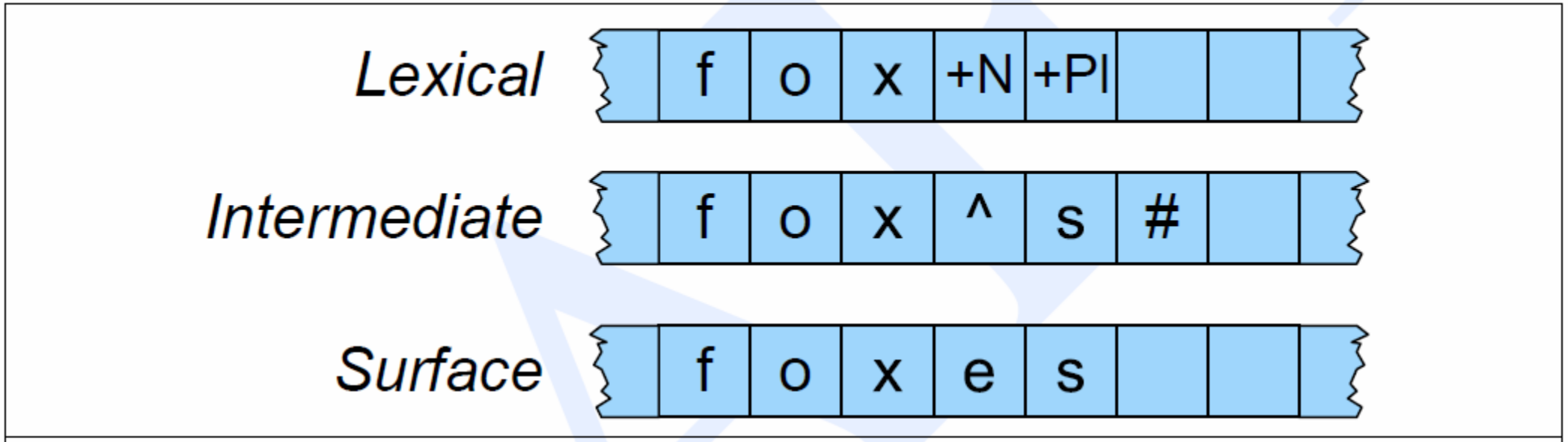


Figure 3.16 An example of the lexical, intermediate, and surface tapes. Between each pair of tapes is a two-level transducer; the lexical transducer of Fig. 3.14 between the lexical and intermediate levels, and the E-insertion spelling rule between the intermediate and surface levels. The E-insertion spelling rule inserts an *e* on the surface tape when the intermediate tape has a morpheme boundary ^ followed by the morpheme *-s*.

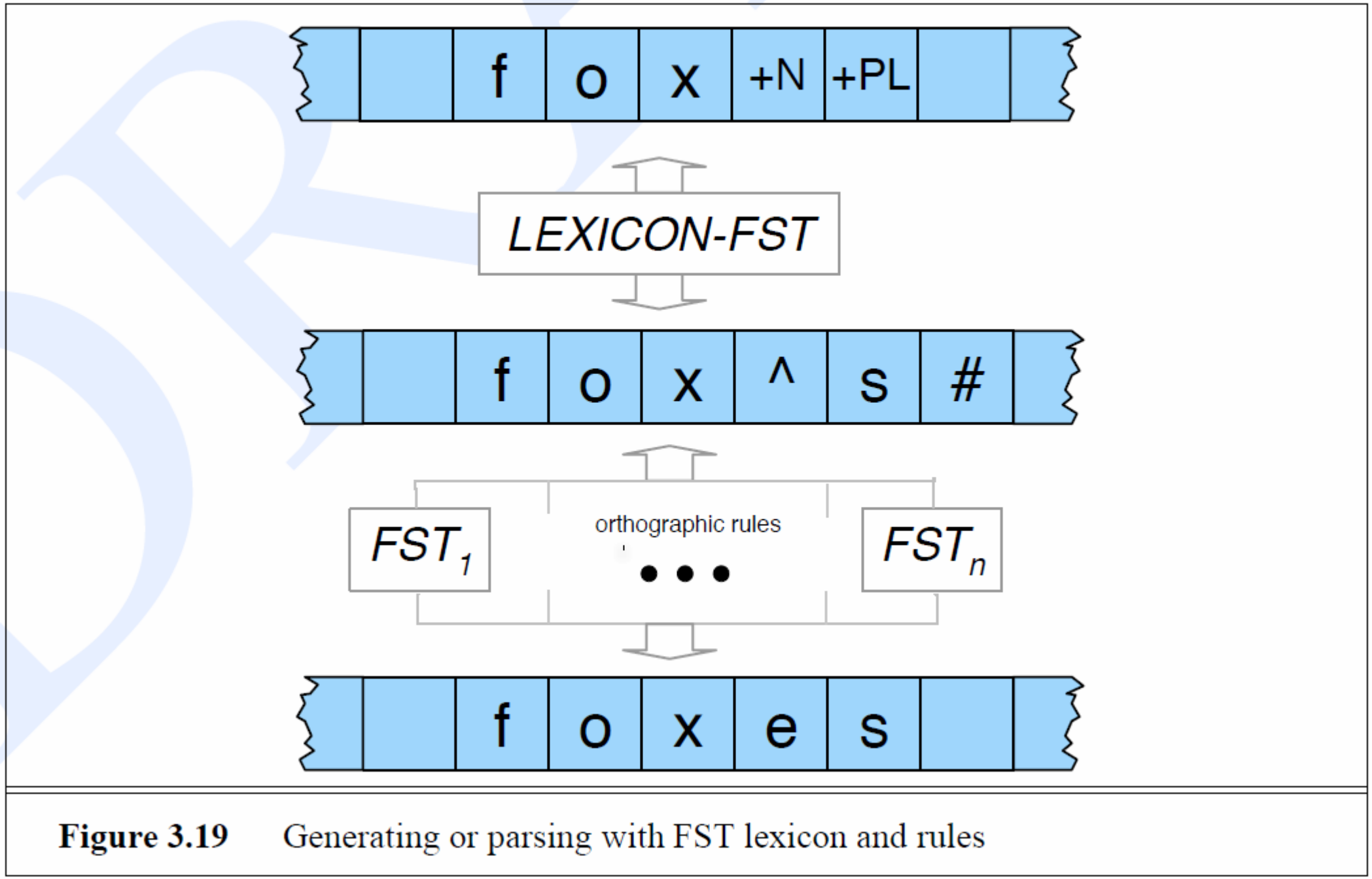


Figure 3.19 Generating or parsing with FST lexicon and rules

Descubrir morfología

- Sucesor frecuente (Harris, 1955)

gover~n, gover~ned, gover~ning, gover~nment, gover~nor, gover~ns	govern, govern~ed, govern~ing, govern~ment, govern~or, govern~s
Sucesor frecuente (gover) = 1 (n).	Sucesor frecuente (govern) = 6 (e, i, m, o, s, espacio/signo).

Descubrir morfología

- Trie: estructura de árbol que guarda cadenas. Cada cadena es un camino desde la raíz hasta la última hoja.

Descubrir morfología

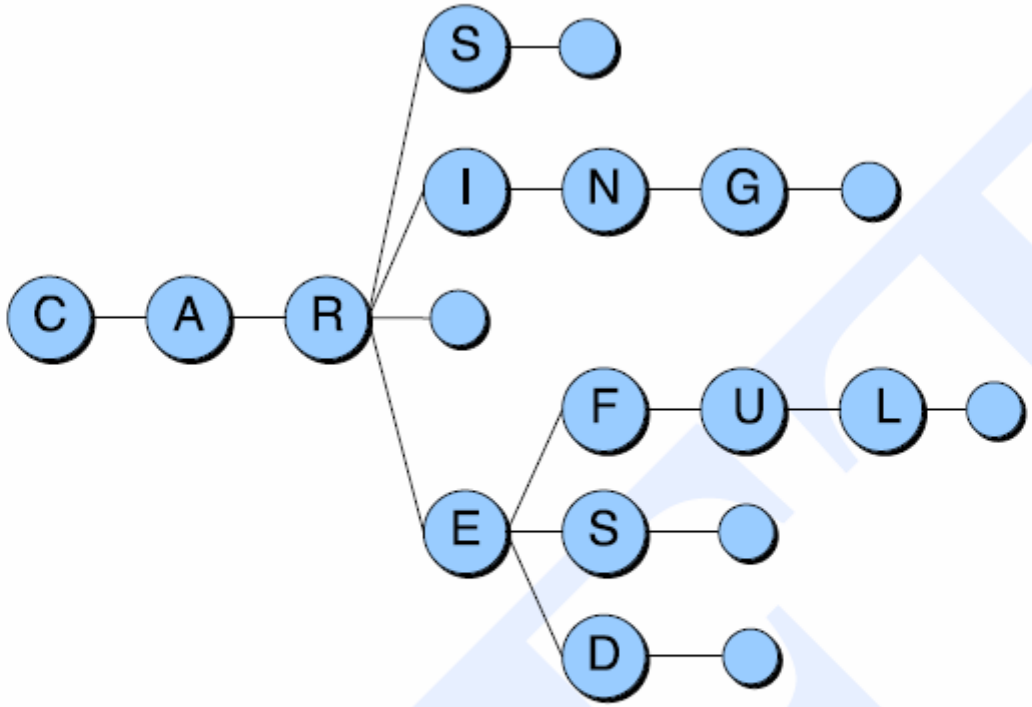


Figure 11.17 Example of a letter trie. A Harris style algorithm would insert morpheme boundaries after *car* and *care*. After Schone and Jurafsky (2000).

Ejercicio

- xooko'
- xook
- xooki'
- xooke'
- xoot
- xooko'obo'
- xooknak
- xookna'ako'on
- xooka'
- xooko'obe'

TAREA

- wiliko'
- wilike'ex
- wila'aj
- wilej
- wilike'
- wilik
- wilibe'
- wilaj
- wil
- wili'
- wile'
- wile'ex
- wilmaj
- wilajo'
- wilmaji'
- wilmi'
- wili'i
- wilo'ob
- wilike'exe'
- wilaje'
- wilme'exi'
- wilmaje'ex
- wilmaje'exi'
- wilike'exo'
- wila'e'

Fin