Morphology

- Introduction
- Morphology
- Morphological analysis
- Using finite state techniques in morphological analysis

Introduction (I)

Morphology is the study of the way words are built from smaller units: morphemes ***un-believe-able-ly***

Two broad classes of morphemes: **stems** (main meaning) and **affixes** (additional).

Introduction (II)

Affixes

- **Prefixes:** precede the stem: ***un-certain***, ***un-chain***

- **Sufixes: *eat-s***

- **Circumfixes:** prefixes and sufixes: ***sagen – ge-sag-t***

- **Infixes:** Inserted in the middle of the word: tagalog language, not in formal English (but in dialects: bl**dy,f**king, ***abso-bl**dy-lutely***).

Introduction (III)

Morphemes

- 1 morpheme:

  Evitar (verb to avoid)

- 2 morphemes:
  - evitable = evitar + able (adj: can be avoided)
- 3 morphemes:
  - inevitable = in + evitar + able

    (adj: cannot be avoided)
- 4 morphemes:
  - inevitabilidad = in + evitar + able + idad

    (noun: cannot be avoided)

Introduction (IV)

Agglutinative languages tend to string affixes together

- Turkish, ten or more affixes
- English no more than five

Different ways to combine morphemes:

Inflection: stem + grammatical morpheme

syntactic function: plural and gender in nouns
tense on verbs

Derivation: stem + grammatical morpheme

different class, different meaning

Computerize-computerization

Introduction (V)

Different ways to combine morphemes:

Compounding.Combination of multiple stems: ***doghouse***

Cliticization: stem+ clitic (reduced in form): ***I’ve***

Inflection in English is simple.

- Suffixes: -s,-ed,-ing

Derivation is more complex.

- Suffixes: –ation,-ness, -able
- Prefixes: co-,re-

Introduction (VI)

Morphological parsing is the process of finding the constituent morphemes in a word

cat +N+ pl for cats

To build a morphological parser we need:

A *lexicon,* the list of stems and affixed and basic information about them.

Morphotactics is the model of morpheme ordering that explains the allowable morpheme sequences.

Orthografics rules: spelling rules to model the changes when combining morphemes: city- cities

Introduction (VII)

Result of morphologic analysis

- Morphosyntactic categorization (**POS**)
  - e.g. Parole tagset, more than 150 categories for Spanish
  - e.g. Penn Treebank tagset , about 30 categories for English
- Morphological features
  - Number, case, gender, lexical functions

Introduction (VIII)

Many constraints on morphotactics can be represented by finite automata

Finite state transducers are an extension of finite-state automata that can generate output symbols.

Finite state transducers are used for: morphology representation, parsing, spelling error detection:

Lexicon and spelling rules can be represented by composing and intersecting transducers

Introduction (IX)

Problems

- Detect the affixes
  - Suffixes, prefixes, infixes, interfixes
- Inflectional affixes different from derivational affixes
- Derivation implies sometimes a semantic change not always predictible
- Inflection does not change POS, sometimes derivation does
- Inflection affects other words in the sentence

  agreement

– A derivativational suffix can be followed by an inflectional one   love => lover => lovers

Morphology (I)

- Morphology studies the sructure of a word as a composition of morphemes
- Morphotactics studies the word formation rules

  Valid combinations between morphemes

  Simple concatenation

  Complex models root/pattern
- Phonological alterations (Morphophonology)
  - Changes when concatenating morphemes
  - Source: Phonology, morphology, orthography
  - variable in number and complexity
  - e.g. vocalic harmony

Morphology (II)

Inflectional Morphology

	Regular Nouns	Irregular Noun
Singular	cat	mouse
Plural	cats	mice

Morphology (III)

Inflectional Morphology

Morphological Form Classes	Regularly	Inflected	Verbs	
stem	walk	merge	try	map
-s form	walks	merges	tries	maps
-ing participle	walking	merging	trying	mapping
Past form or -ed participle	walked	merged	tried	mapped

Morphology (III)

Inflectional Morphology

- number
  - thrush thrushes
  - cheval chevaux
  - casa casas
- verbal form
  - walk walkes walked walking
  - amo amas aman ...
- gender
  - niño niña

Morphology (IV)

Derivational Morphology

Suffix	Base Verb/Adjective	Derived Noun
-ation	computerize (V)	computerization
-ee	appoint (V)	appointee
-er	kill (V)	killer
-ness	fuzzy (A)	fuzziness

Morphology (V)

Derivational Morphology

- Form
  - Without change barcelonés
  - Prefix inevitable
  - Suffix importantísimo
- Source
  - verb => adjective tardar => tardío
  - verb => noun sufrir => sufrimiento
  - noun => noun actor => actorazo
  - noun => adjective atleta => atlético
  - adjective => adjective rojo => rojizo
  - adjective => adverb alegre => alegremente

Morphological Analysis (I)

Types of morphological analyzers

Formaries

- Dictionaries of word forms

  + efficiency

  + Languages with few variants (e.g. English)

  + extensibility

  + Possibility of building and maintenance from a morphological generator

  – Languages with high flexive variation

  – derivation, composition

- FS techniques
  - FSA
    - 1 level analyzers
  - FST
    - > 1 level analyzers

Morphological Analysis (II)

Morphological analyzers of two levels

- General model for languages with morpheme concatenation
- Independence between lingware and analyzer
- Valid for analysis and generation
- Distinction between lexical and superficial levels
- Parallel rules for morphophonology
- Simple implementation

Morphological Analysis (III)

- Morphological rules
  - Define the relations betweens characters (surface) and morphemes and map strings of characters and the morphemic structure of the word.
- Spelling rules
  - Perform at the level of the letters forming the word. Can be used to define the valid phomological alterations.
- Ritchie, Pulman, Black, Russell, 1987

Morphological Analysis (IV)

- input:
  - form
- output
  - lemma + morphological features

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)

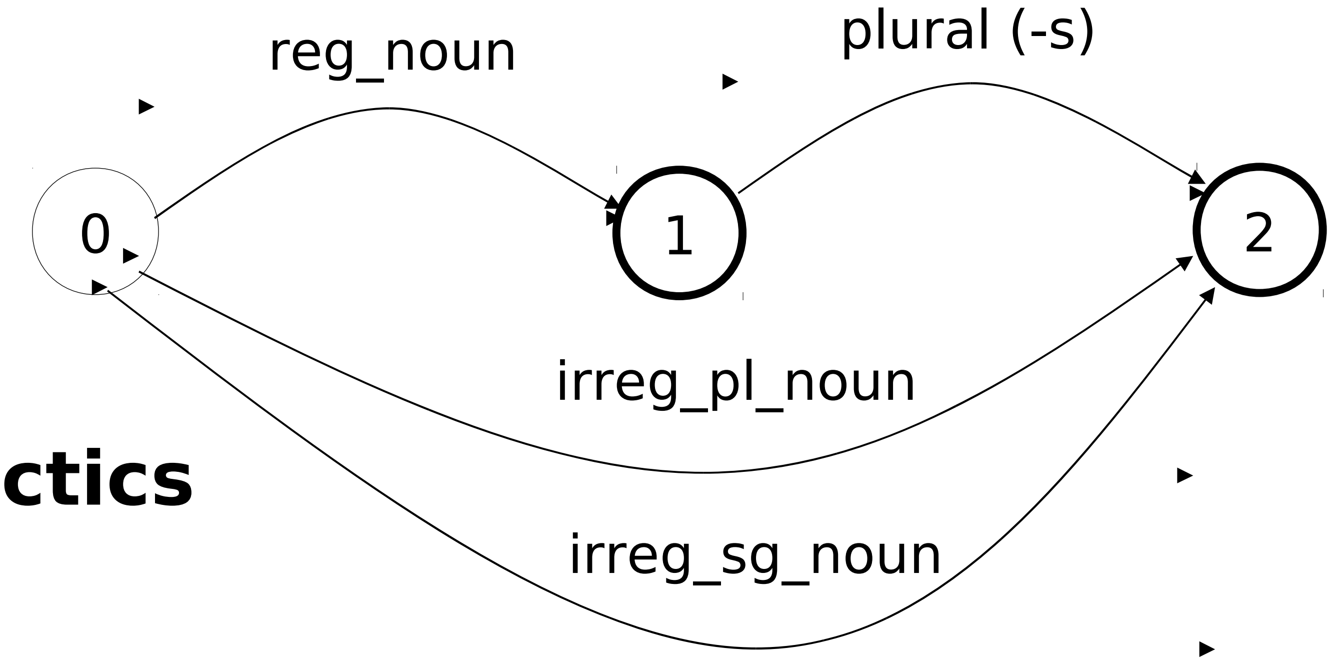
Morphological Analysis (V)

Using FST

- As a recognizer
  - From a pair of input strings (one lexical and the other superficial) determines if one is transduction of the other
- As a generator
  - Generates pairs of strings
- As a translator
  - From a superficial string generates its lexical translation

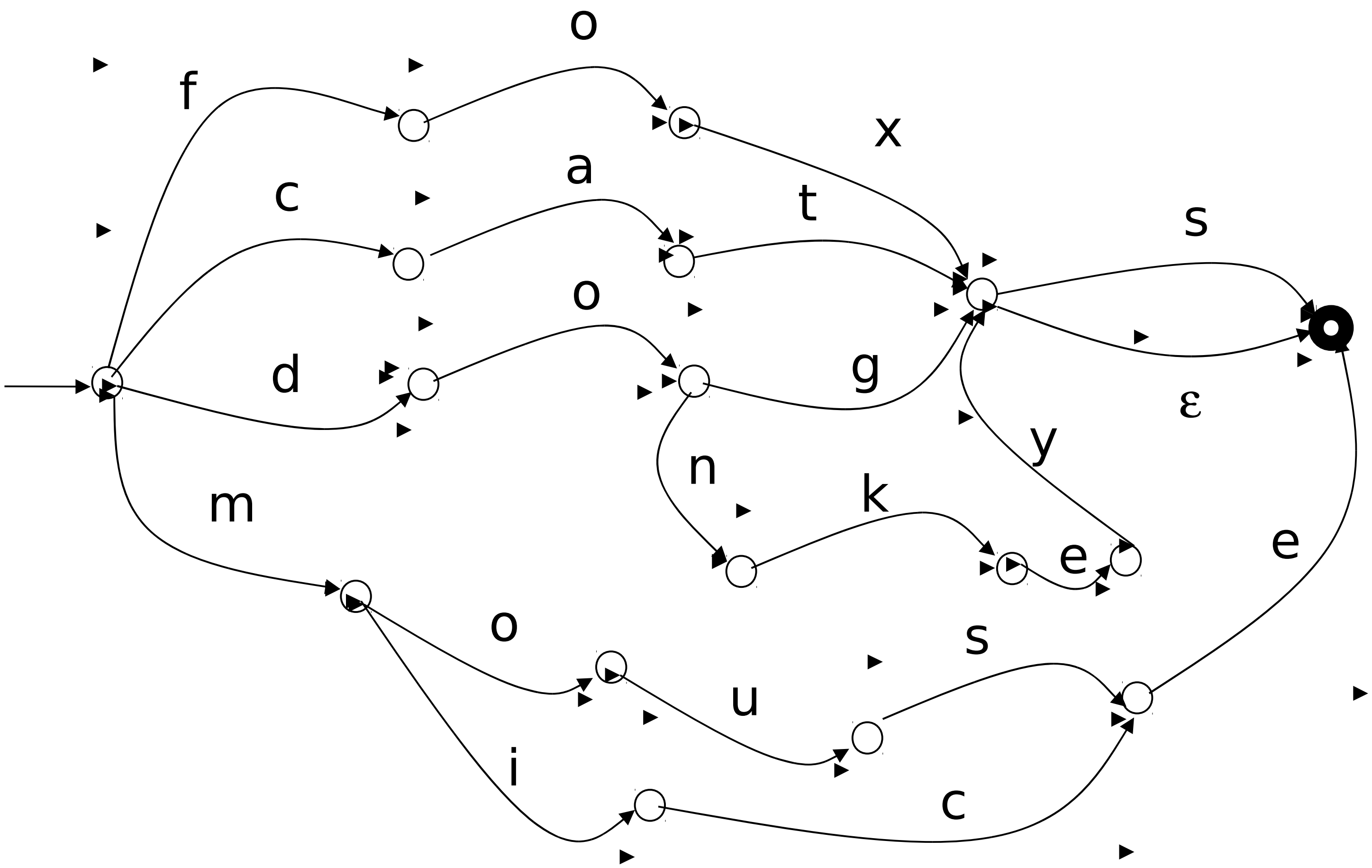
Morphological Analysis (VI)

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



Morphotactics

Morphological Analysis (VII)



fog
cat
dog
donkey
mouse
mice

Letter Transducers

Morphological Analysis (VIII)

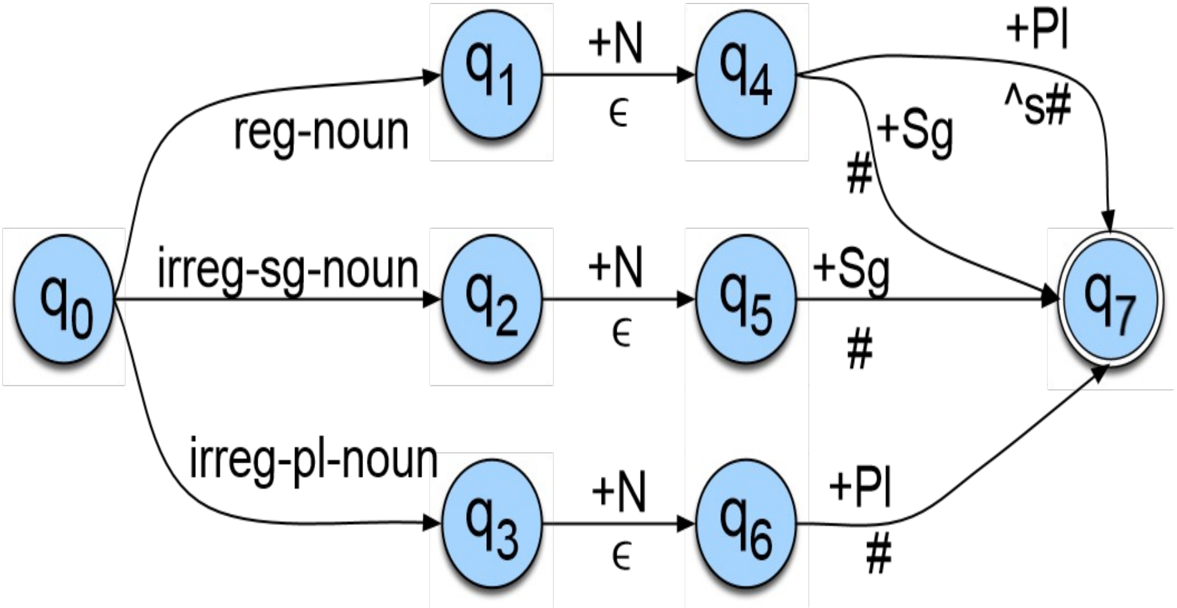
upper level	lexic	cat + N	cat + N + pl
lower level	surface	cat	cats

⇩

c:c a:a t:t +N:ε +pl:s

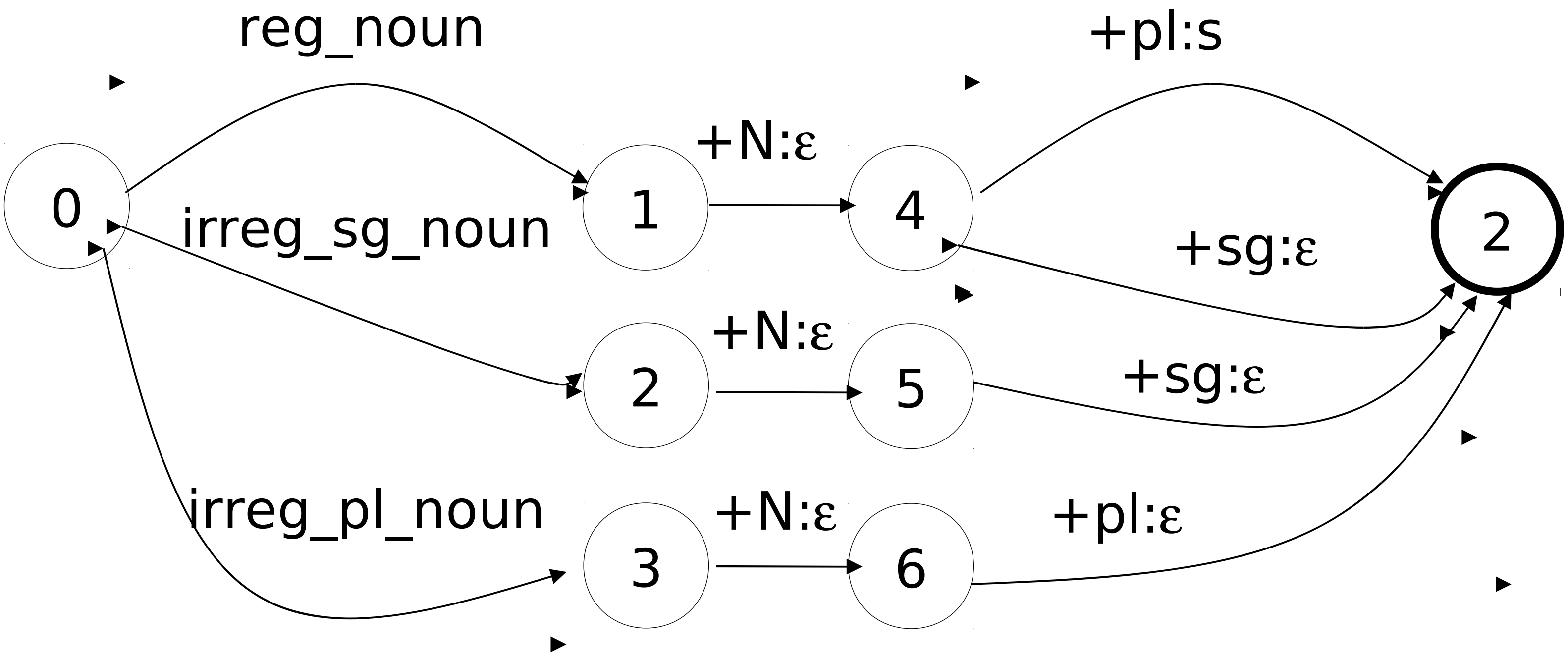
Morphological Analysis (IX)

Using FST



Morphological Analysis (X)

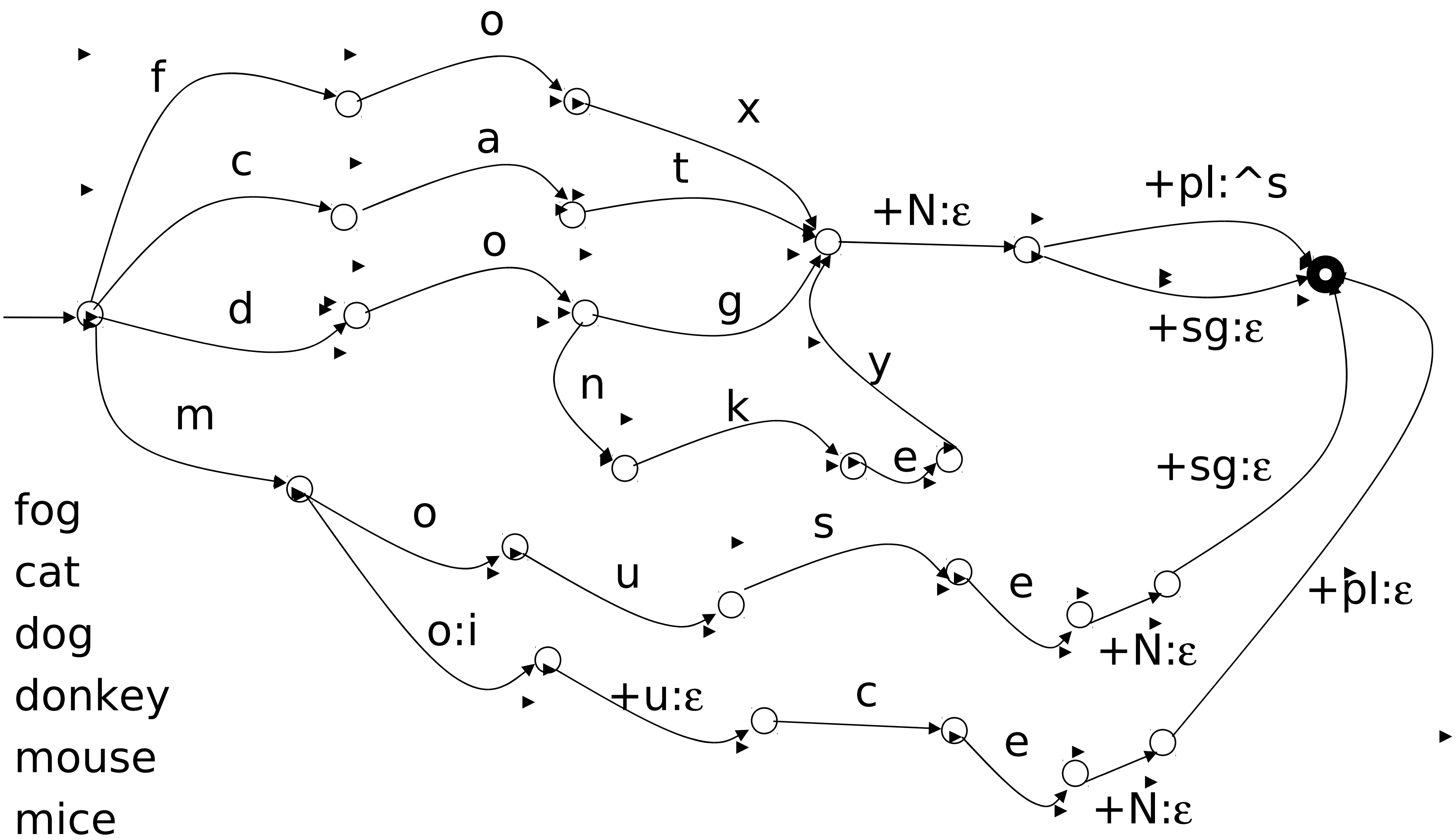
reg_noun	irreg_pl_noun	irreg_sg_noun	plural
fox	sheep	sheep	s
cat	m o:i u:ε ce	mouse	
dog	g o:e o:e se	goose	



Morphological Analysis (XI)

	lexical level	f	o	x	+N	+pl
morphotactics						
	intermediate level	f	o	x	^	s
spelling rules						
	superficial level	f	o	x	e	s

Morphological Analysis (XII)



fog
cat
dog
donkey
mouse
mice

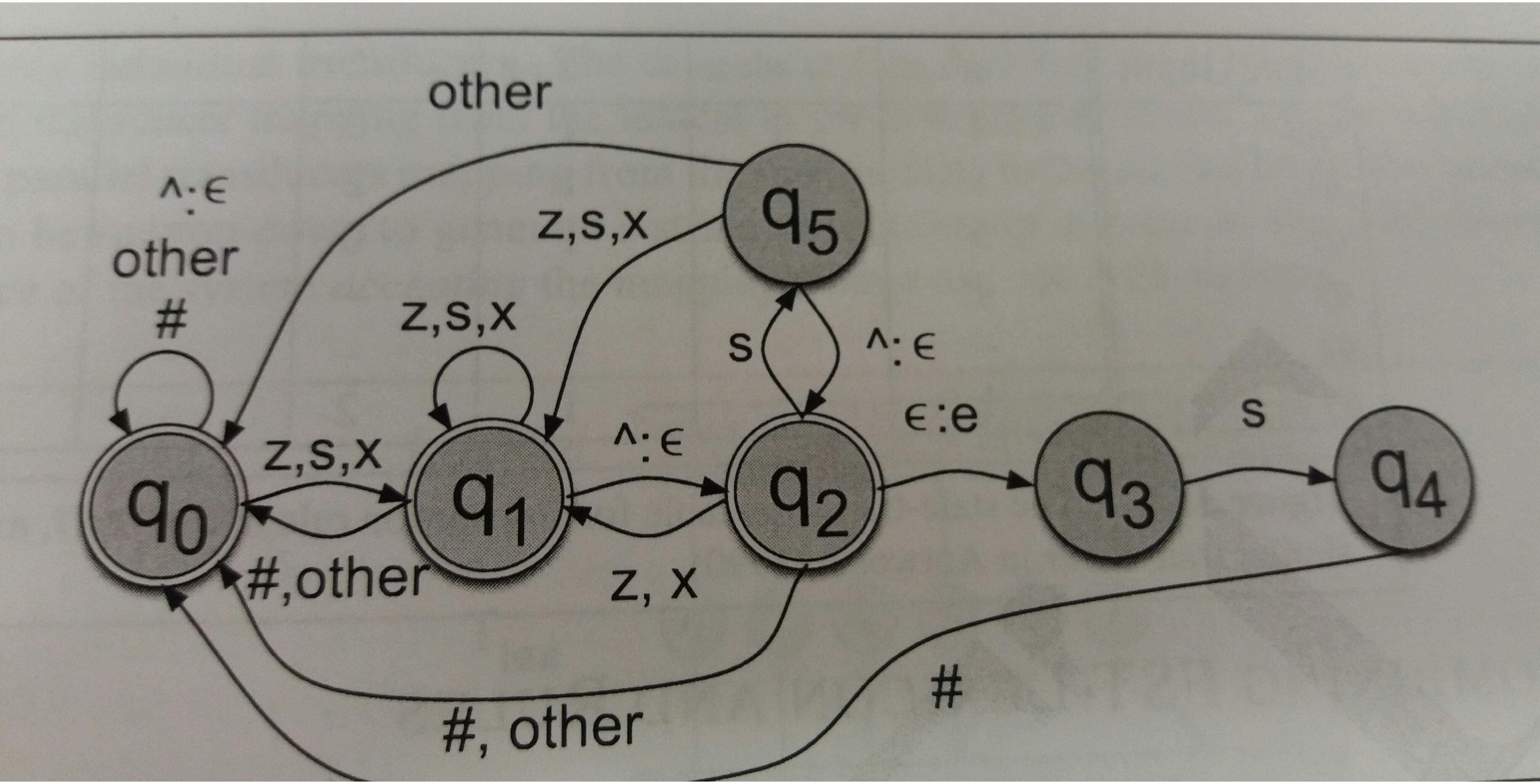
Morphological Analysis (XIII)

Spelling rules

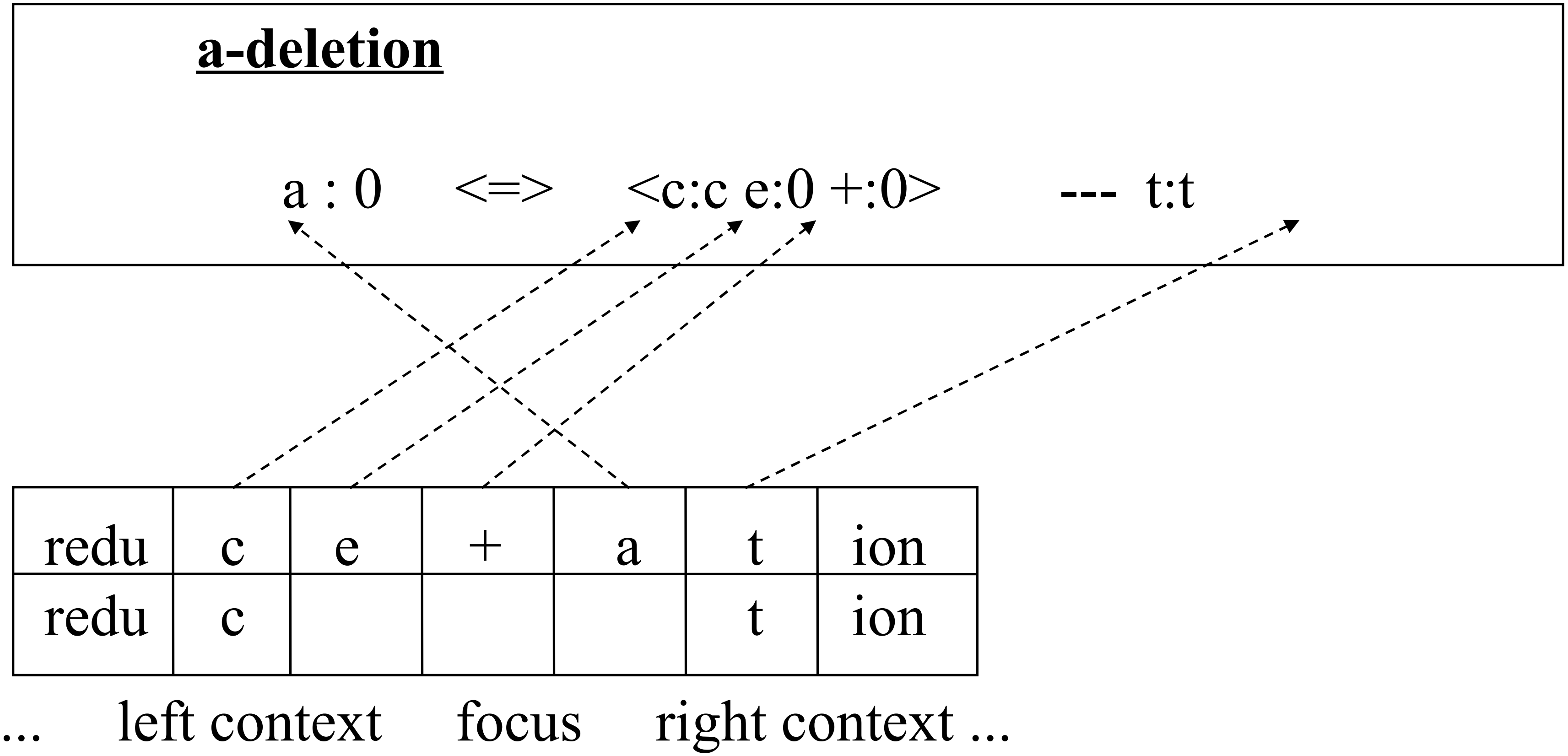
name	description	example
consonant doubling beg/begging	single letter consonant doubled before -ing/-ed	
e deletion	silent e dropped before -ing/-ed	move/moved make/making
e insertion	e added after -s,-z,-x,-ch,-sh before -s	box/boxes watch/watches
y replacement	-y changes to -ie before -s, to i before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

Morphological Analysis (XIV)

Transducer for the E-insertion rule



Morphological Analysis (XV)



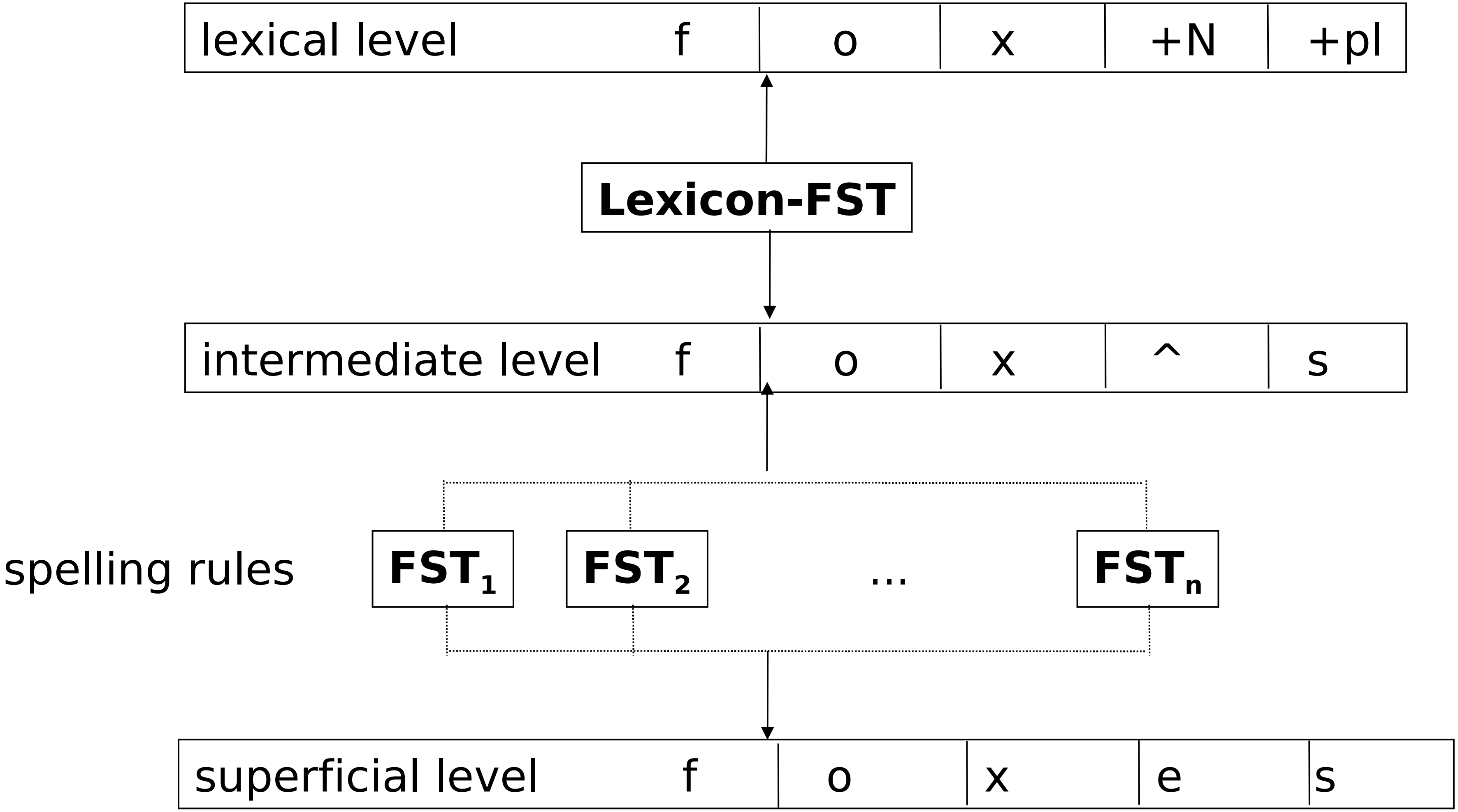
a-deletion

a : 0 <=> <c:c e:0 +:0> --- t:t

redu	c	e	+	a	t	ion
redu	c				t	ion

... left context focus right context ...

Morphological Analysis (XVI)



lexical level f | o | x | +N | +pl

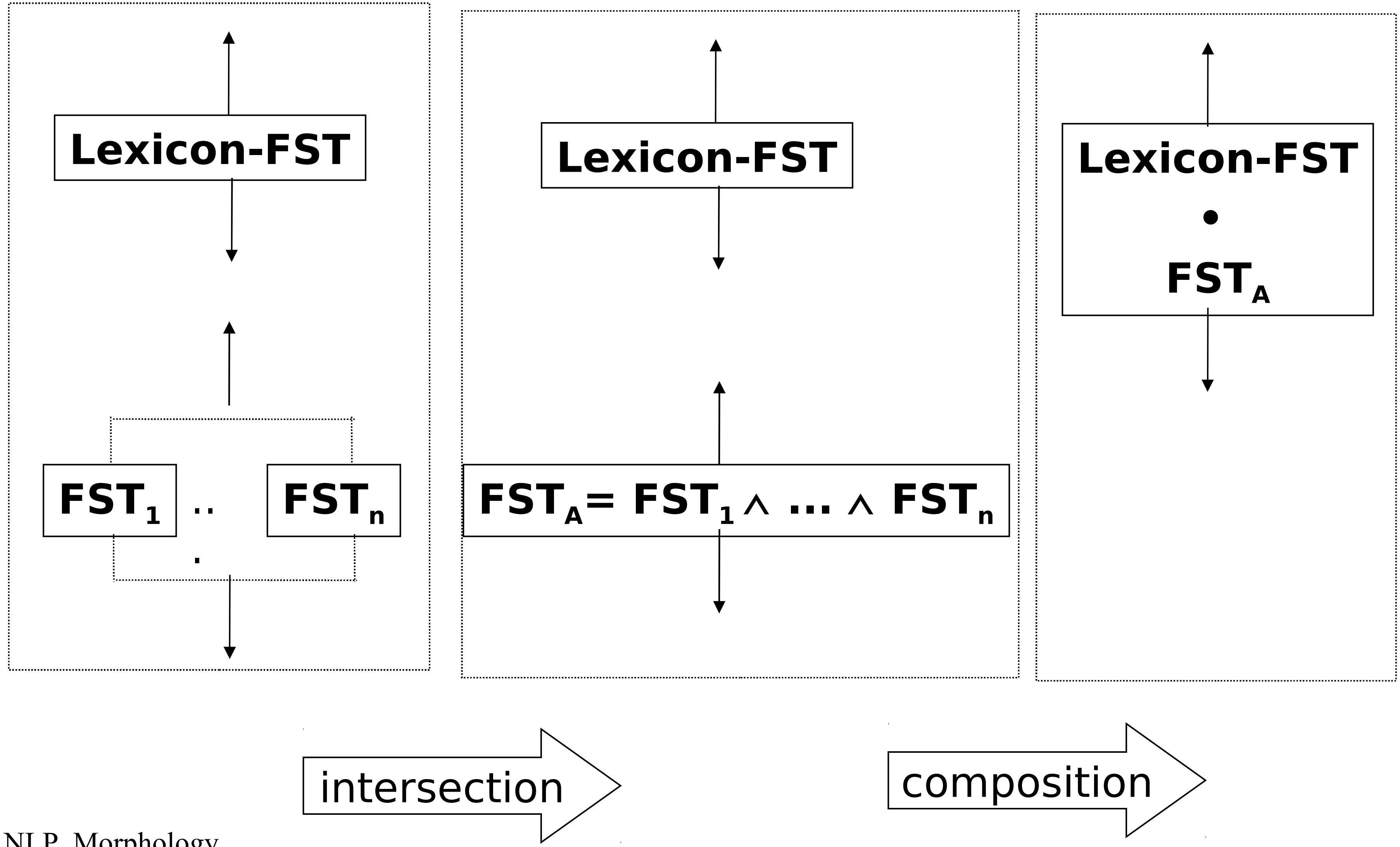
Lexicon-FST

intermediate level f | o | x | ^ | s

spelling rules FST$_1$ FST$_2$ … FST$_n$

superficial level f | o | x | e | s

Morphological Analysis (XVII)



Automatic morphology learning (I)

- Problem
  - Paradigm stem + affixea
  - Obtaining the stems
  - Classification of stems into models
  - Learning part of the morphology (e.g. derivational)
- Two approaches
  - No previous morphologic knowledge is available
    - Goldsmith, 2001
    - Brent, 1999
    - Snover, Brent, 2001, 2002
  - Morphologic knowledge can be used
    - Oliver at al, 2002

Automatic morphology learning (II)

- Automatic morphological analysis
  - Identification of borders betwen morphemes
    - Zellig Harris
      - {prefix, suffix} conditional entropy
  - bigrams and trigrams with high probability of forming a morpheme
  - Learning of patterns or rules of mapping between pairs of words
  - Global approach (top-down)
    - Golsdmith, Brent, de Marcken

Automatic morphology learning (III)

- Goldsmith’s system based on MDL (Minimum Description Length)
  - Initial Partition: word -> stem + suffix
    - split-all-words
      - A good candidate to {stem, suffix} splitting in a word has to be a good candidate in many other words
    - MI (mutual information) strategy
      - Faster convergence
  - Learning Signatures
    - {signatures, stem, suffixes}
  - MDL

Automatic morphology learning (IV)

- Semi-automatic morphological analysis
  - Oliver, 2004
  - Starts with a set of manually written morphological rules
    - TL:TF:Desc
      - lemma ending
      - form ending
      - POS
  - Lists of non flexive classes , closed classes and irregular words
  - Corpora
    - Serbo-Croatian 9 Mw
    - Russian 16 Mw

Summary (I)

- Morphology
  - Structure of a word as a composition of morphemes
  - Related to word formation rules

    Inflection

    Derivation

    Composition
- Morphotactics

  Which morphemes can be concatenated with which others

Summary (II)

Different ways to combine morphemes:

Inflection: stem + grammatical morpheme (syntactic function: plural, gender, tense)

Derivation: stem + grammatical morpheme (different class, different meaning).

Computerize-computerization

Compounding.Combination of multiple stems: **doghouse**

Cliticization: stem+ clitic (reduced in form): **I’ve**

Inflection in English is simple (-s,-ed,-ing)

Derivation is more complex (suffixes –ation,-ness,-able, prefixes co-,re-)

Summary (III)

Morphologic analysis

- Decompose a word into a concatenation of morphemes
- Usually some of the morphemes contain the meaning
  - One (root or stem) in flexion and derivation
  - More than one in composition
- The other (affixes) provide morphological features

- Problems
  - Phonological alterations in morpheme concatenation

Summary (IV)

Result of morphologic analysis

- Morphosyntactic categorization (POS)
  - e.g. Parole tagset (VMIP1S0), more than 150 categories for Spanish
  - e.g. Penn Treebank tagset (VBD), about 30 categories for English
- Morphological features
  - Number, case, gender, lexical functions