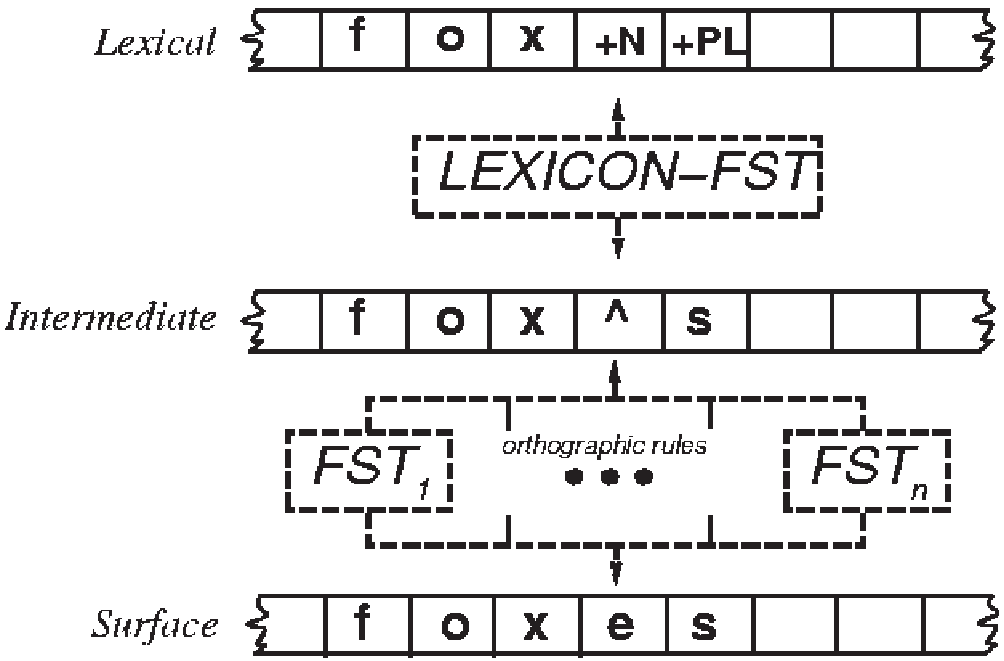
Combining FST Lexicon & Rules

- Two-level morphology used for parsing or generating:
  - The lexicon transducer maps between the lexical level (stems + morphological features) and an intermediate level (simple concat of morphemes)
  - The host of transducers, each representing a single spelling rule, all run in parallel to map between intermediate level and surface level
  - The result is a two-level cascade of transducers
  - The cascade can be:

    top-down to generate a string or

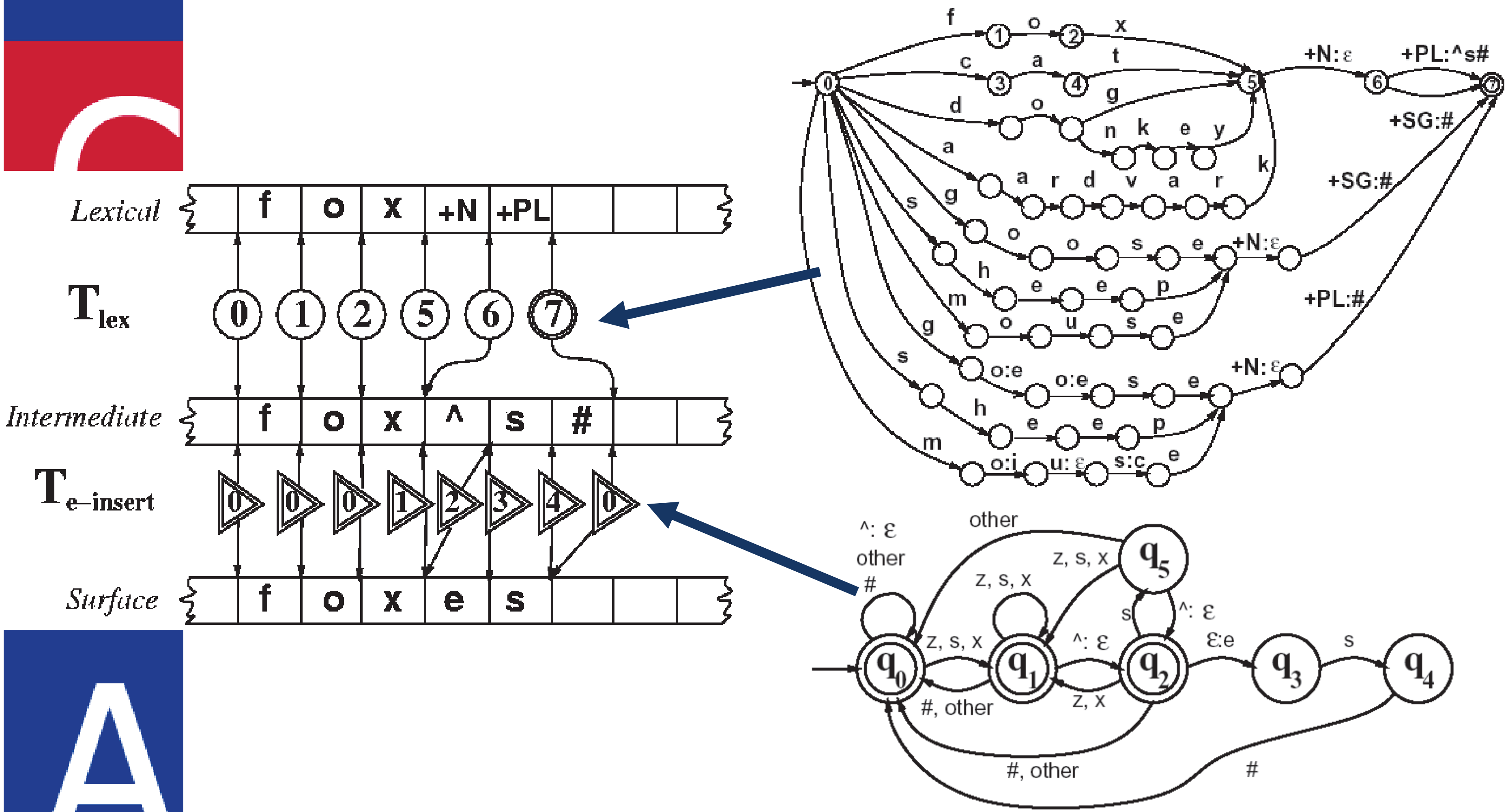
    bottom-up to parse it

Combining FST Lexicon & Rules



Combining FST Lexicon and Rules

Combining FST Lexicon & Rules



1

Combining FST Lexicon & Rules

- The power of FSTs is that the exact same cascade with the same state sequences is used
    - when machine is generating the surface tape from the lexical tape, or
    - when it is parsing the lexical tape from the surface tape.
  - Parsing can be slightly more complicated than generation, because of the problem of **ambiguity**
    - For ex: *foxes* could be `fox +V +3Sg` as well as `fox +N +PL`
    - Disambiguating requires the surrounding words
    - Noun --> *I saw two foxes yesterday*
    - Verb --> *He foxes me every time!*

The Porter Stemmer

- Information retrieval → boolean combination of relevant keywords or phrases
- In IR, morphological information is used to determine that the two words have the same stem; the suffixes are thrown away
- The mostly widely used **stemming** algorithms is the simple Porter (1980) algorithm, which is based on a series of simple cascaded rewrite rules.
    - ATIONAL → ATE (e.g., relational → relate)
    - ING → ε if stem contains vowel (e.g., motoring → motor)
  - Problem:
    - Not perfect: error of commission (organization → organ), omission (European → Europe)
  - Some improvement with smaller documents