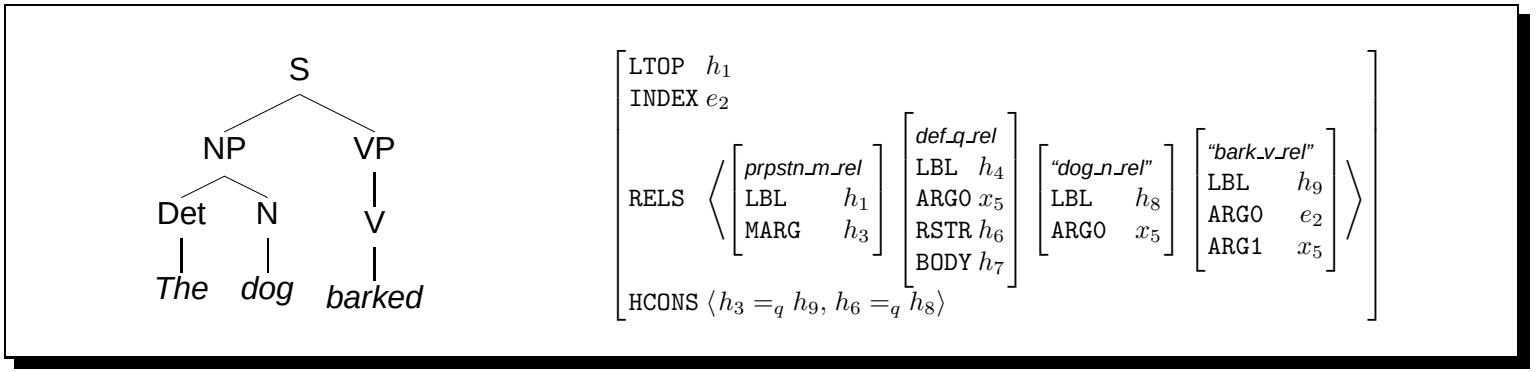


Computational Linguistics (INF2820 — Morphology)

{ eat, eats, eating, ate, eaten }

Stephan Oepen

Universitetet i Oslo

oe@ifi.uio.no

Pattern Matching on Strings: Finite-State Automata

/baa+!/

ba! — baa! — baah! — baaaa! — baaaaaaaaa!

Pattern Matching on Strings: Finite-State Automata

/baa+!/

ba! — baa! — baah! — baaaa! — baaaaaaaaa!

Recognizing Regular Languages

- Finite-State Automata (FSAs) are *very restricted* Turing machines;
- states and transitions: read one symbol at a time from input tape;

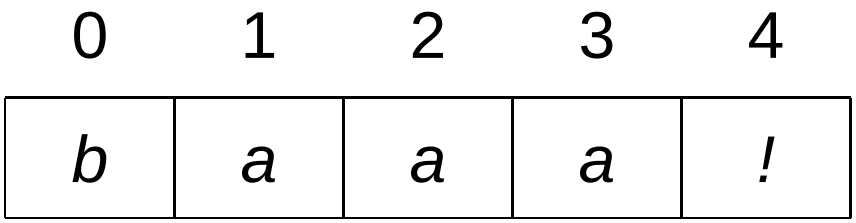
→ *accept* utterance when no more input, in a 'final' state; else *reject*.

Tracing the Recognition of a Simple Input

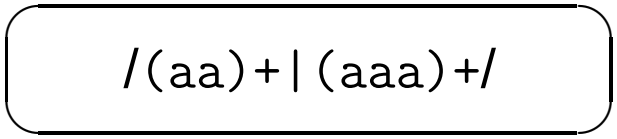
/baa+!/

ba! — baa! — baah! — baaaa! — baaaaaaaaa!

Input Tape

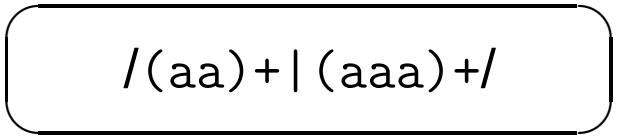


A Rather More Complex Example



aa — *aaa* — *aaaa* — *aaaaaa* — *aaaaaaaa* — *aaaaaaaaa* — ...

A Rather More Complex Example



aa — *aaa* — *aaaa* — *aaaaaa* — *aaaaaaaa* — *aaaaaaaaa* — ...

- Non-Deterministic FSAs (NFSAs): multiple transitions per symbol;

→ a *search space* of possible solutions: decisions no longer obvious.

Quite Abstractly: Three Approaches to Search

(Heuristic) Look-Ahead

- Peek at input tape one or more positions beyond the current symbol;
- try to work out (or 'guess') which branch to take for current symbol.

Parallel Computation

- Assume unlimited computational resources, i.e. any number of cpus;
- copy FSA, remaining input, and current state $\rightarrow$ multiple branches.

Backtracking (Or Back-Up)

- Keep track of possibilities (*choice points*) and remaining candidates;
- 'leave a bread crumb', go down one branch; eventually come back.

NFSA Recognition (From Jurafsky & Martin, 2008)

```
 1  procedure nd-recognize(tape , fsa) ≡
 2    agenda ← {⟨0, 0⟩};
 3    do
 4      current ← pop(agenda);
 5      state ← first(current);
 6      index ← second(current);
 7      if (index = length(tape) and state is final state) then
 8        return accept;
 9      fi
10      for(next in fsa.transitions[state, tape[index]]) do
11        agenda ← agenda ∪ {⟨next, index + 1⟩}
12      od
13      if agenda is empty then return reject; fi
14    od
15  end
```

Some Areas of Descriptive Grammar

Phonetics	***The study of speech signals.***
Phonology	***The study of sound systems.***
Morphology	***The study of word structure.***
Syntax	***The study of sentence structure.***
Semantics	***The study of language meaning.***
Pragmatics	***The study of language use.***

Some Areas of Descriptive Grammar

Phonetics	***The study of speech signals.***
Phonology	***The study of sound systems.***
Morphology	***The study of word structure.***
Syntax	***The study of sentence structure.***
Semantics	***The study of language meaning.***
Pragmatics	***The study of language use.***

Tokenization: Finding the Basic Building Blocks

```
Mr. Browne, who's prime minister, arrived.
```

Tokenization: Finding the Basic Building Blocks

Mr. Browne, who's prime minister, arrived.

He eats chocolate, candy (i.e. sugar), etc.

Morphological Categories (1 of 3)

Parts of Speech (PoS)

cat, dog, neighbours, ...	**noun (N)**
barks, chased, was, ...	**verb (V)**
fierce, angry, black, young, ...	**adjective (Adj)**
quickly, probably, not, ...	**adverb (Adv)**
a, the, my, that, ...	**determiner (D)**
of, by, on, at, under, ...	**preposition (P)**
she, mine, those, what, ...	**pronoun (Pro)**
and, neither ... nor, because, ...	**conjunction (C)**

How to discover the inventory of categories?

A Quick Tour of English Morphology

Morphological Categories (2 of 3)

Parts of Speech (PoS)

cat, dog, neighbours, ...	**noun (N)**
barks, chased, was, ...	**verb (V)**
fierce, angry, black, young, ...	**adjective (Adj)**
quickly, probably, not, ...	**adverb (Adv)**
a, the, my, that, ...	**determiner (D)**
of, by, on, at, under, ...	**preposition (P)**
she, mine, those, what, ...	**pronoun (Pro)**
and, neither ... nor, because, ...	**conjunction (C)**

- **Paradigm** set of word forms, e.g. $\{bark, barks, barking, barked\}$;
- **Unit Categories** dimensions structuring a paradigm *internally*;
- **Paradigm Categories** properties *common* to all paradigm units.

Morphological Categories (3 of 3)

Number — Person — Case — Gender

That dog barks. — Those dogs bark.
I bark. — You bark. — They bark. — Sam shaved himself.
We bark. — You bark. — Those dogs bark.
I saw her. — She saw me. — My dog barked.
...

How many distinct verb forms according to number and person?

Tense — Aspect — Mood

The dog barks. — The dog barked — The dog will bark.
The dog has barked. — The dog is barking.
If I were a carpenter, ...
...