Linguistics for CS

Lecture 3 - Morphology

Anca Dinu

NLP master programme

University of Bucharest, October 2025

01 Morphemes

02 Main concepts: root, stem, affix, lemma

03 FSA, FST

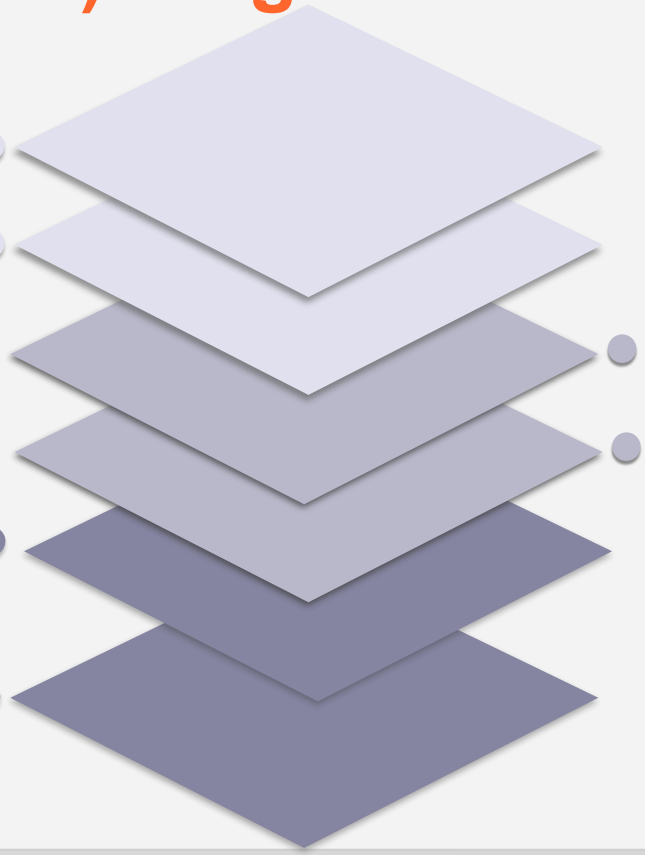
04 Computational applications in Morphology

(Isomorphic) Linguistic Levels -recap

SOUND

PHONETICS
Properties of sounds

PHONONOGY
Patterns of sounds



MORPHOLOGY
Structure of words

SYNTAX
Structure of sentence

STRUCTURE

MEANING

SEMANTICS
Meaning of words (lexical semantics), sentences and discourse (formal semantics)

PRAGMATICS
Use of language (in extra-linguistic context, intentions)

What is Morphology?(morph(Greek)=form)

- **Morphology** is the study of the structure of words.
- **Morphemes** are the smallest units that carry meaning, most fundamental units of language.
- **Examples:**
  - Inflectional morphemes (like the plural marker on nouns): friend+**s**
  - Derivational morphemes can occur either at the beginning or the end of a word: **un**+friend+**ly**

Inflectional vs. Derivational Morphemes

- An **inflectional morpheme** is a morpheme that
  - varies the word by adding grammatical contrast,
  - preserving the meaning
  - keeping the word class (pos) the same.
- verbal inflection: **conjugation**
- nominal inflection: **declension**

- Examples:
  - number: scaun - **scaun+e**
  - person: merg - **merg +i**
  - gender lup - lup + **oaică**
  - tense merg - merg +**eam**
  - diminutives: suc - suc + **uleț**
  - others: mood, voice, aspect, case

Inflectional vs. Derivational Morphemes

- A **derivational morpheme** is a morpheme that:
  - creates a new word from an existing word,
  - altering the meaning and
  - sometimes changing the word class.

- Examples:
  - Nominalization: cerceta +**re**, trada + **re**
  - Adjectivization: adora + **bil**, fenomen + **al**

Bound vs. unbound morphemes

- **Free morphemes** are the ones that are word-forms by themselves
- **Bound morphemes** are the ones which cannot be word-forms by themselves, but need to be attached to other morphs.

- Example: **Every·one live·s by sell·ing some·thing**.
  - **Every, one, live, by, sell, some** and **thing** are free morphemes
  - **-s** and **–ing** are bound morphemes.

Root, stem, lemma, affix

- **Root** is the central morpheme, which contributes the meaning of the word, common to a set of derived or inflected forms.
- **Stem** is the root of the word together with the **derivational** morphemes, to which one can add inflectional morphemes, like:
    - For the word **undone**: **do** is the **root**, **un** is a derivational morpheme, **undo** is the stem and **ne** is an inflectional morpheme.
- **Lemma**- citation of a word as it appear in a dictionary
    - **to undo**

Root, stem, lemma, affix

- **Affix** is a bound morpheme that can be attached to a root or a stem.
- Three types:
  - If an affix is attached before a base, it is called a **prefix**: **pre-**determine
  - If it is attached after a base, it is called a **suffix**: talk-**ed**, faith-**full**;
  - If it is attached in the middle of the base, it is called an **infix**. There are no infixes in English.

Wod formation processes

- Inflection: tense, mood, voice, aspect, person, number, gender, case
- Derivation
- Compounding

Extreme examples:

- German:

Rechts + schutz + ver +sicher + ungs + gesell + schaft + en

=Legal Protection Insurance Policy

- Turkish:

uygar + laş + tir + ama + dik + larimiz + dan + miş + siniz + casina

= behaving as if you are among those whom we could not cause to become civilized

Morphology and types of languges

- **isolating**: Chinese, Vietnamese
  - little or no derivation and inflection
- **analytic**: Chinese, English
  - little or no inflection
- **synthetic**:
  - **agglutinative**: Finnish, Turkish, Hungarian, Swahili
    - morphemes are concatenated with little modification
    - each affix usually encodes a single feature
  - **fusional** (inflecting): Sanskrit, Latin, Romanian, Russian, German
    - inflectional affixes often encode a feature bundle

Computational morphology

- Goals:
  - automatically decide whether a word is in a particular language or not (accepting/generating a word).
  - automatically analysing a word (decomposing a word in morphemes)
- Use:
  - Mapping tables (works only for languages with simple morphology)
  - Finite State Automaton (FSA) for the first task
  - Finite State Transducers (FST) for both tasks
  - Machine learning techniques
- Apply to:
  - spell checkers,
  - language identification,
  - speech synthesis, etc.

Ingredients of a Morphological Generator/Analyser

- List of roots with part-of-speech
- List of derivational/inflectional affixes
- morphotactic rules
- Orthographic/phonological rules

Morphotactics

- Which morphemes can be arranged in which order?
  - translat+abil+ity
  - *translat+ity+abil
  - translat+able
  - *translat+able+ity (Allomorphs able-abil)

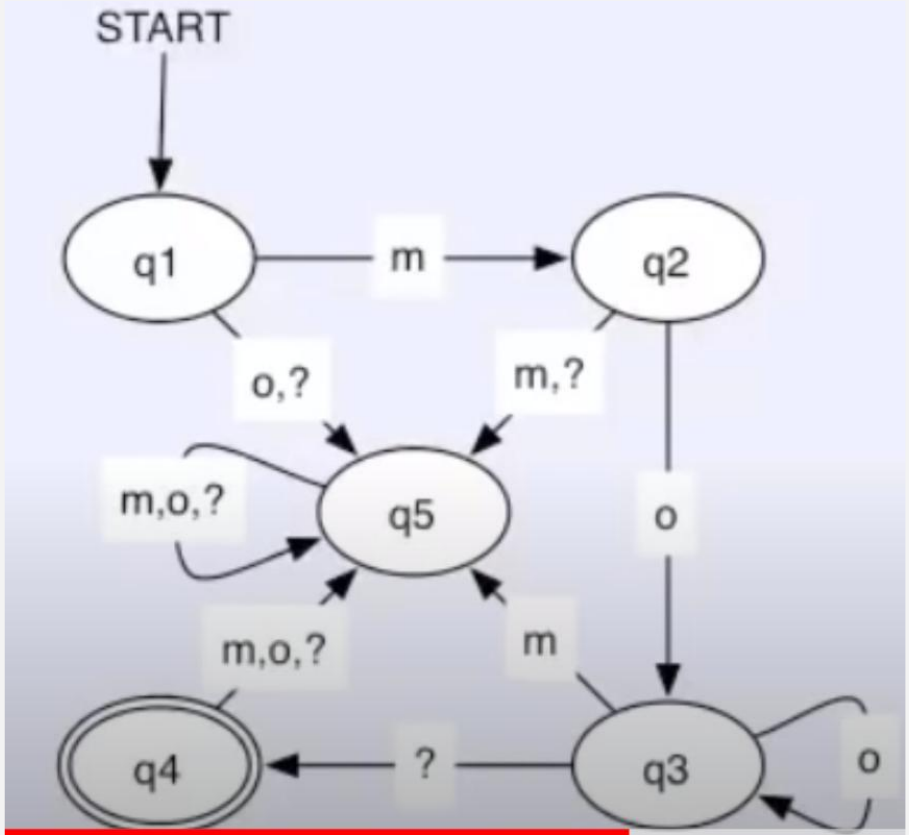
Orthographic/Phonological Rules

- How is a morpheme realised in a certain context?
  - city+s → cities
  - bake+ing → baking (e-elision)
  - crash+s → crashes (e-epenthesis)
  - beg+ing → begging (gemination)
  - ad+simil+ate → assimilate (assimilation)
  - ip+lEr → ipler kız+lEr → kızlar (vowel harmony)

Computational morphology: FSAs

- A finite state automaton is a 5-tuple (Q,Σ,q_0,A,δ) where
  - Q is a finite set of states;
  - Σ is a finite input alphabet;
  - $q_0 \in Q$ is the initial state;
  - $A \subseteq Q$ is the set of accepting states;
  - $\delta: Q \times \Sigma \rightarrow Q$ is the transition function.
- For any element q of Q and any symbol $\sigma \in \Sigma$, we interpret $\delta(q,\sigma)$ as the state to which the FSA moves, if it is in state q and receives the input σ.

FSA example: Inquisitive cow



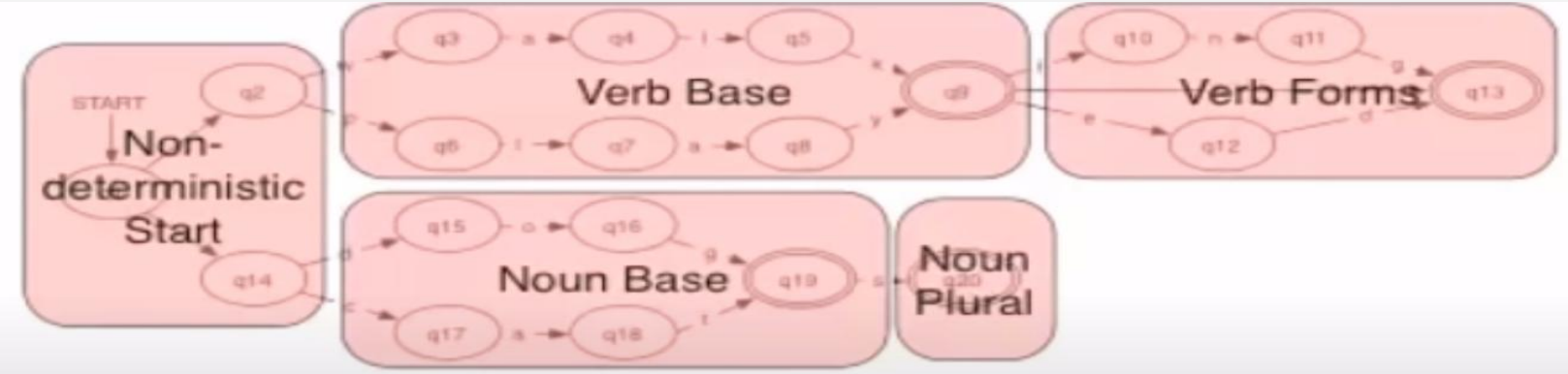
Computational morphology: FSAs

- FSA properties:
  - Equivalence with regular expressions
  - Intersection
  - Difference
  - Complementation
  - Reversal
  - Concatenation
  - Closure (infinite repetition)

- Application of FSA properties:
  - create a FSA for English stems, for nouns, verbs, adjectives, etc;
  - create a FSA that accepts inflectional forms;
  - combine them with concatenation!

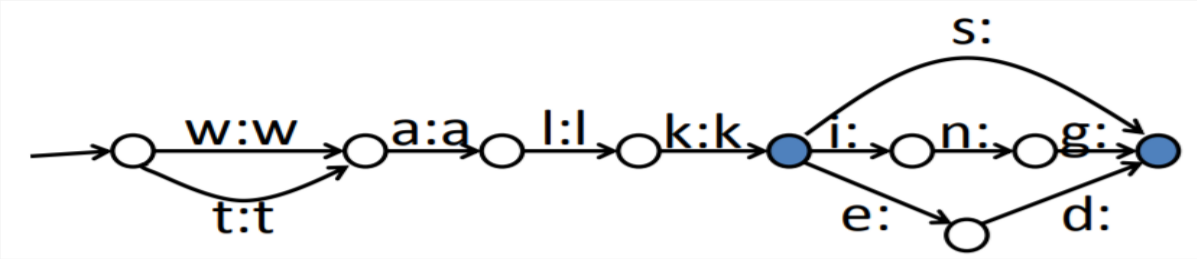
Non-deterministic FSAs

- Allows empty input
- Allows empty edges



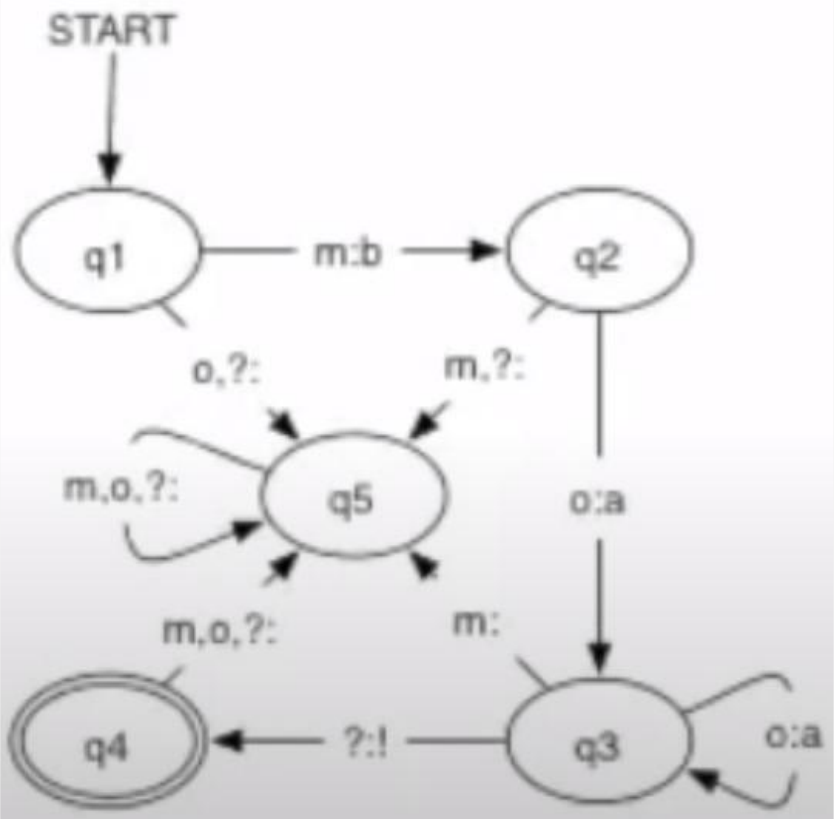
Computational morphology: FSTs

- Finite State Transducers
- FSTs are FSAs whose transitions are labelled with symbol pairs
- They map strings to (sets of) other strings



- maps walk, walks, walked, walking to walk
- and talk, talks, talked, talking to talk (in generation mode)
- can also map walk to walk, walks, walked, walking in analysis mode

FSA to FST: Inquisitive cow to emphatic sheep



Some FST toolkits

- Xerox finite-state tools xfst and lexc: well-suited for building morphological analysers
- Foma (Mans Hulden): open-source alternative to xfst/lexc
- AT&T tools: weighted transducers for tasks such as speech recognition, but little support for building morphological analysers
- openFST (Google, NYU): open-source alternative to the AT&T tools
- SFST: open-source alternative to xfst/lexc but using a more general and flexible programming language

Articles on Romanian Computational Morphology/Phonology

- Morphology:
  - Romanian verbs at RANLP (Context Sensitive Grammars) https://www.aclweb.org/anthology/R11-1075.pdf
  - Romanian Nouns at LREC, neutral detection http://www.lrec-conf.org/proceedings/lrec2012/pdf/651_Paper.pdf
  - Romanian verbs at EACL Avignon, machine learning https://www.aclweb.org/anthology/E12-1053.pdf
  - Neutral detection at Coling https://www.aclweb.org/anthology/C12-3015.pdf
  - Romanian verbs at Ranlp, with CRF https://www.aclweb.org/anthology/R13-1028.pdf
- Phonetics:
  - Automated Syllabification at TSD https://link.springer.com/chapter/10.1007/978-3-642-40585-3_57
  - Accent prediction at EACL https://www.aclweb.org/anthology/E14-4013.pdf

THANKS

You should have some questions now!

anca.dinu@lls.unibuc.ro
ancaddinu@gmail.com
+0785641041

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik

Content contributions: Jordan Boyd-Graber, Alexander Fraser & Liane Guillou, Lucia Loureiro Porto