Morphology and Finite-state Transducers: Part 1
ICS 482: Natural Language Processing

Lecture 5

Husni Al-Muhtaseb

بسم الله الرحمن الرحيم

Morphology and Finite-state Transducers: Part 1
ICS 482: Natural Language Processing

Lecture 5

Husni Al-Muhtaseb

NLP Credits and Acknowledgment

These slides were adapted from presentations of the Authors of the book

SPEECH and LANGUAGE PROCESSING:
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

and some modifications from presentations found in the WEB by several scholars including the following

NLP Credits and Acknowledgment

If your name is missing please contact me
muhtaseb
At
Kfupm.
Edu.
sa

NLP Credits and Acknowledgment

Husni Al-Muhtaseb
James Martin
Jim Martin
Dan Jurafsky
Sandiway Fong
Song young in
Paula Matuszek
Mary-Angela Papalaskari
Dick Crouch
Tracy Kin
L. Venkata Subramaniam
Martin Volk
Bruce R. Maxim
Jan Hajič
Srinath Srinivasa
Simeon Ntafos
Paolo Pirjanian
Ricardo Vilalta
Tom Lenaerts

Heshaam Feili
Björn Gambäck
Christian Korthals
Thomas G. Dietterich
Devika Subramanian
Duminda Wijesekera
Lee McCluskey
David J. Kriegman
Kathleen McKeown
Michael J. Ciaraldi
David Finkel
Min-Yen Kan
Andreas Geyer-Schulz
Franz J. Kurfess
Tim Finin
Nadjet Bouayad
Kathy McCoy
Hans Uszkoreit

Khurshid Ahmad
Staffan Larsson
Robert Wilensky
Feiyu Xu
Jakub Piskorski
Rohini Srihari
Mark Sanderson
Andrew Elks
Marc Davis
Ray Larson
Jimmy Lin
Marti Hearst
Andrew McCallum
Nick Kushmerick
Mark Craven
Chia-Hui Chang
Diana Maynard
James Allan

Martha Palmer
julia hirschberg
Elaine Rich
Christof Monz
Bonnie J. Dorr
Nizar Habash
Massimo Poesio
David Goss-Grubbs
Thomas K Harris
John Hutchins
Alexandros Potamianos
Mike Rosner
Latifa Al-Sulaiti
Giorgio Satta
Jerry R. Hobbs
Christopher Manning
Hinrich Schütze
Alexander Gelbukh
Gina-Anne Levow
Guitao Gao
Qing Ma
Zeynep Altan

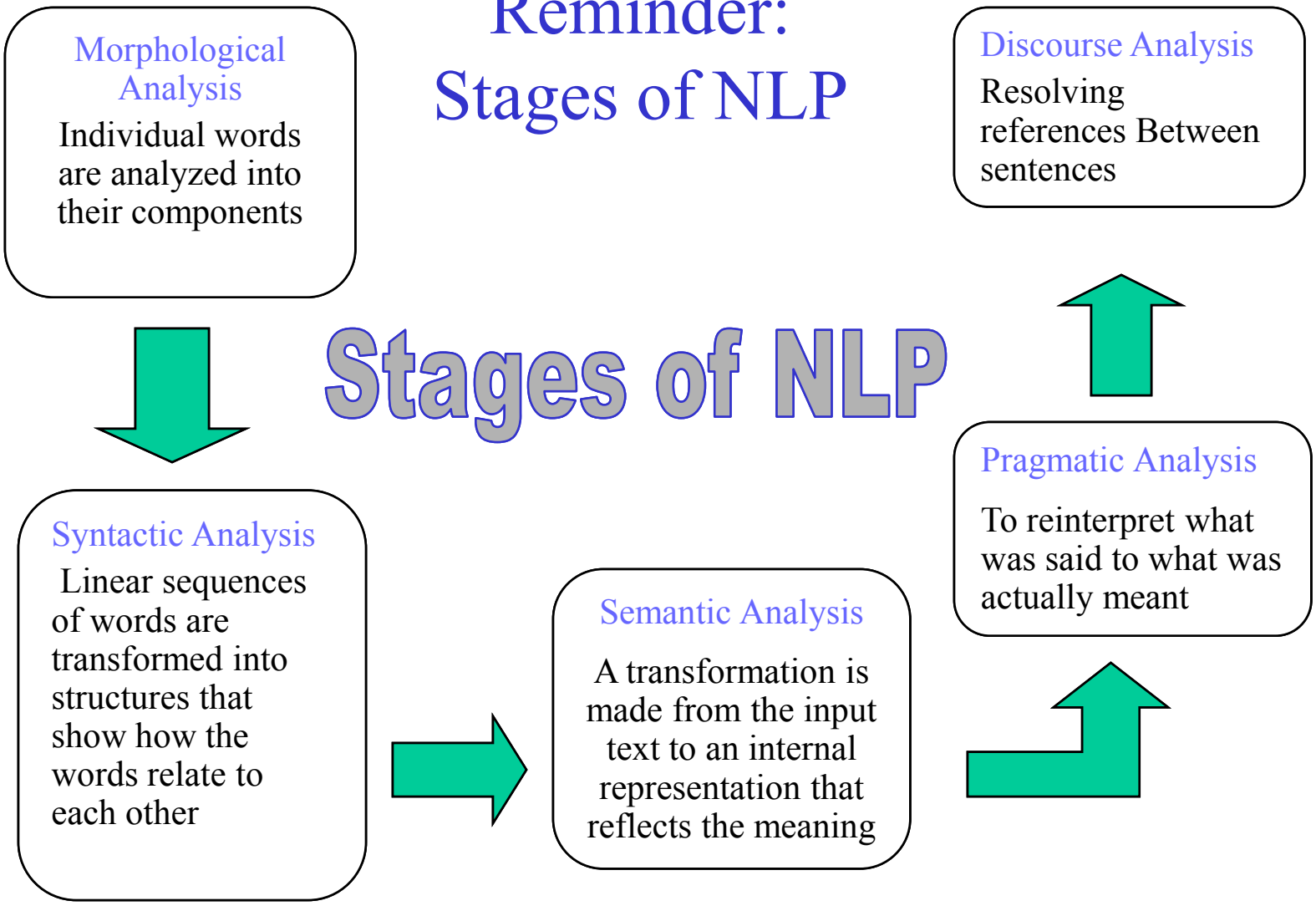
Previous Lectures

- 1 Pre-start online questionnaire
- 1 Introduce yourself
- 2 Introduction to NLP
- 2 Phases of an NLP system
- 2 NLP Applications
- 3 Chatting with Alice
- 3 Regular Expressions
- 3 Finite State Automata
- 3 Regular languages
- 3 Assignment #2
- 4 Regular Expressions & Regular languages
- 4 Deterministic & Non-deterministic FSAs
- 4 Accept, Reject, Generate terms

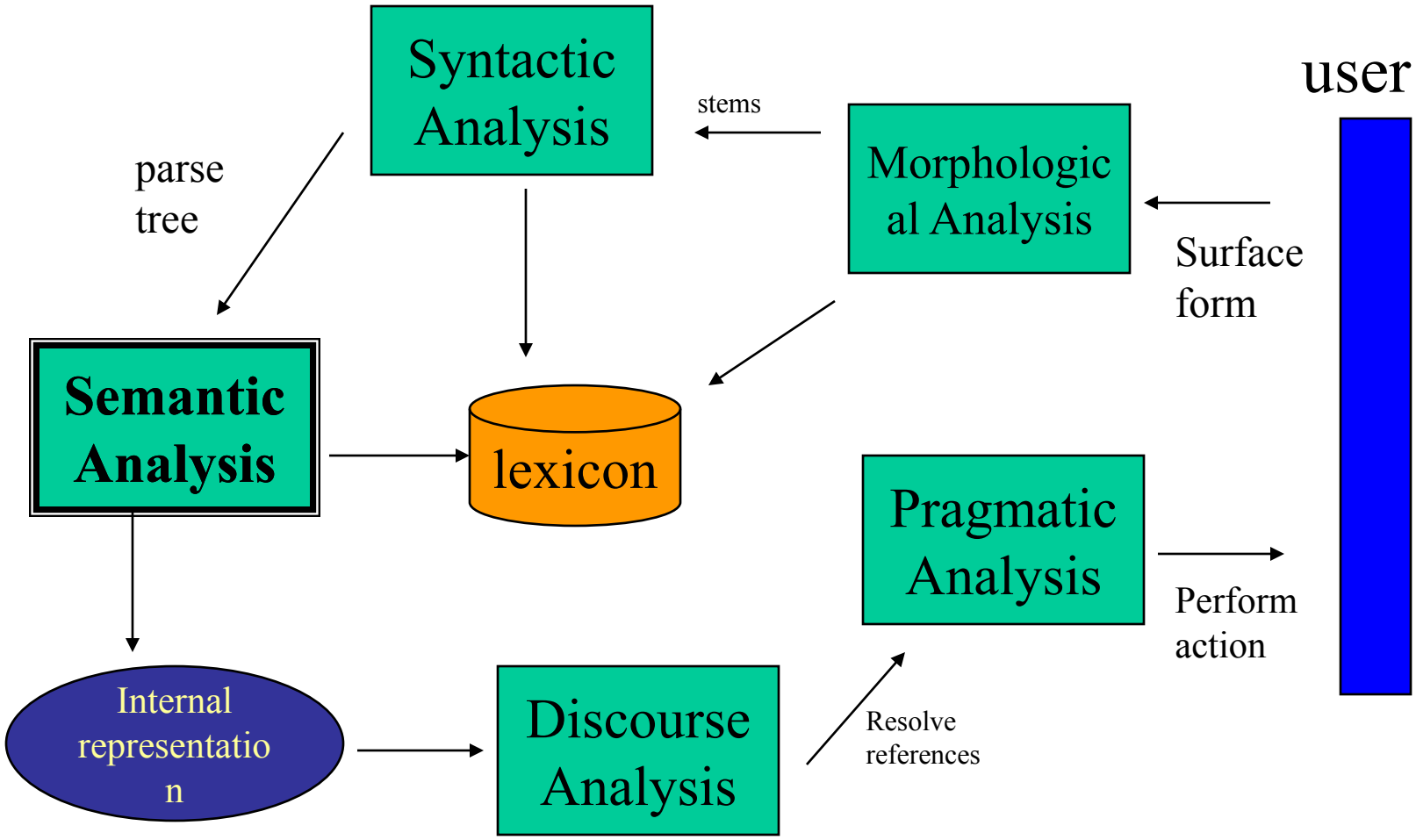
Objective of Today's Lecture

- **Morphology**
  - **Inflectional**
  - **Derivational**
  - **Compounding**
  - **Cliticization**
- **Parsing**
- **Finite State Transducers**
- **Assignment 3**

Reminder: Stages of NLP



Stages of NLP



Introduction

- **Finite-state methods are useful in dealing with the lexicon (words)**
- **Present some facts about words and computational methods**

Morphology

- **Morphology: the study of meaningful parts of words and how they are put together**

- **Morphemes: are the smallest meaningful spoken units of language**
- **Example**
  - **books: two morphemes (book and s) but one syllable**
  - **Unladylike: three morphemes, four syllables**

Morpheme Definitions

- **Root**
  - **The portion of the word that:**
    - **is common to a set of derived or inflection forms, if any, when all affixes are removed**
    - **is not further analyzable into meaningful elements**
    - **carries the principle portion of meaning of the words**
- **Stem**
  - **The root or roots of a word, together with any derivational affixes, to which inflectional affixes are added.**

Morpheme Definitions

- **Affix**
  - **A bound morpheme that is joined before, after, or within a root or stem.**
- **Clitic**
  - **a morpheme that functions syntactically like a word, but does not appear as an independent phonological word**
    - **English:** ***I've (the morpheme 've is a clitic)***

Inflectional vs. Derivational

- **Word Classes**
  - **Parts of speech: noun, verb, adjectives, etc.**
  - **Word class dictates how a word combines with morphemes to form new words**
- **Inflection:**
  - **Variation in the form of a word, typically by means of an affix, that expresses a grammatical contrast.**
    - **Doesn't change the word class**
    - **Usually produces a predictable meaning.**
- **Derivation:**
  - **The formation of a new word or inflectable stem from another word or stem.**

Inflectional Morphology

- **Adds:**
  - **tense, number, person**
- **Word class doesn't change**
- **Word serves new grammatical role**
- **Example**
  - ***come* is inflected for person and number:**

    The pizza guy comes at noon.

Derivational Morphology

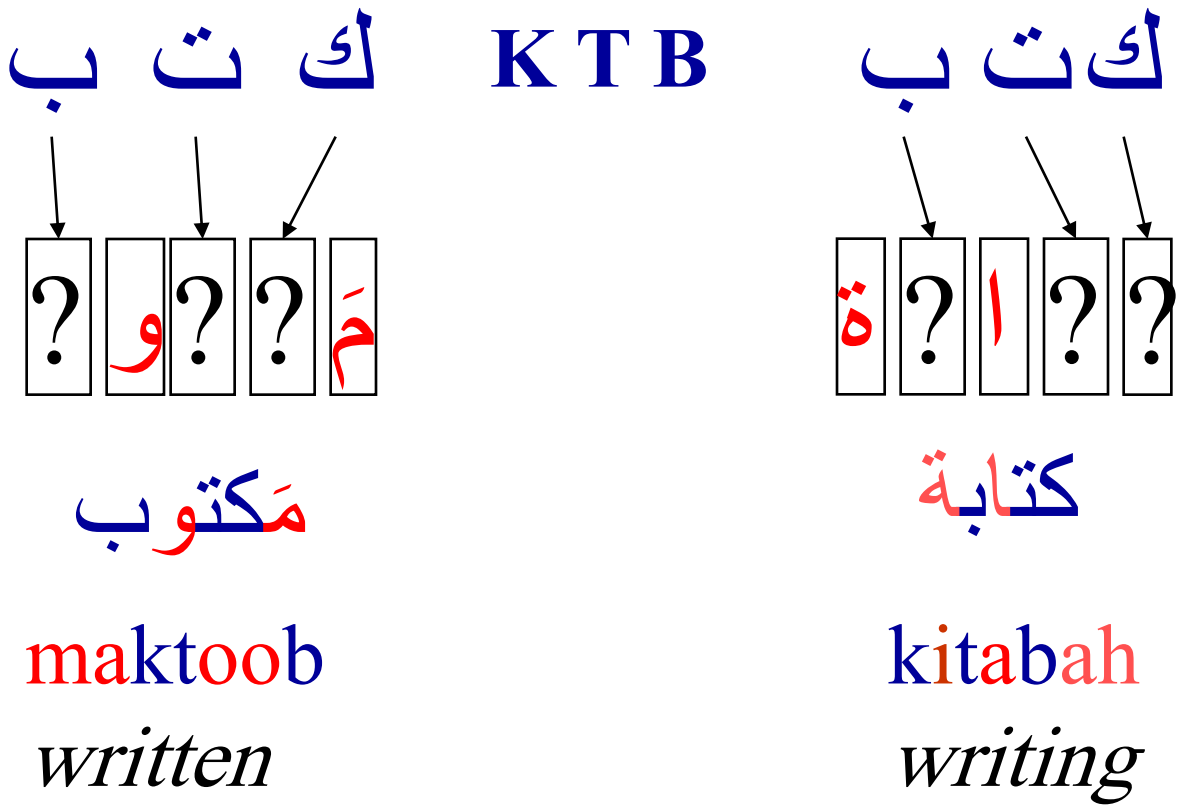
- **Nominalization (formation of nouns from other parts of speech, primarily verbs in English):**
  - **computerization**
  - **appointee**
  - **killer**
  - **fuzziness**
- **Formation of adjectives (primarily from nouns)**
  - **computational**
  - **clueless**
  - **Embraceable**

Concatinative Morphology

- ***Morpheme+Morpheme+Morpheme+…***
- **Stems: also called lemma, base form, root, lexeme**
  - **hope+ing → hoping hop → hopping**
- **Affixes**
  - **Prefixes: Antidisestablishmentarianism - وسنكتب**
  - **Suffixes: Antidisestablishmentarianism - كتبوها**
  - **Infixes: hingi (*borrow*) – humingi (*borrower*) in Tagalog - كاتب**
  - **Circumfixes: sagen (*say*) – gesagt (*said*) in German**
- **Agglutinative Languages**
  - **uygarlaştıramadıklarımızdanmışsınızcasına**
  - **uygar+laş+tır+ama+dık+lar+ımız+dan+mış+sınız+casına**
  - ***Behaving as if you are among those whom we could not cause to become civilized***

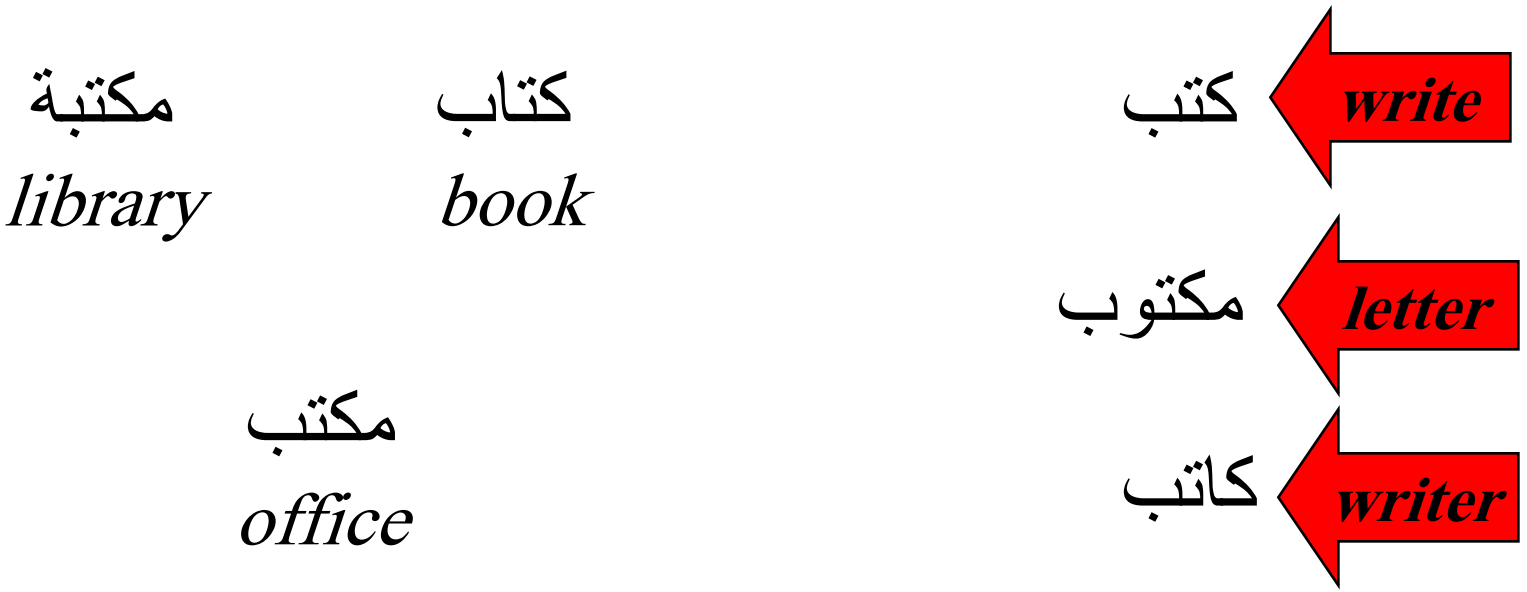
Templatic Morphology

- **Roots and Patterns**



Templatic Morphology: *Root Meaning*

- **KTB: *writing* "stuff"**



Nouns and Verbs (in English)

- **Nouns**
  - **Have simple inflectional morphology**
  - **Cat/Cats**
  - **Mouse/Mice, Ox, Oxen, Goose, Geese**
- **Verbs**
  - **More complex morphology**
  - **Walk/Walked**
  - **Go/Went, Fly/Flew**

Regular (English) Verbs

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

Irregular (English) Verbs

Morphological Form Classes	Irregularly Inflected Verbs		
Stem	**eat**	**catch**	**cut**
-s form	**eats**	**catches**	**cuts**
-ing form	**eating**	**catching**	**cutting**
Past form	**ate**	**caught**	**cut**
-ed participle	**eaten**	**caught**	**cut**

“To love” in Spanish

	Present Indicative	Imperfect Indicative	Future	Preterite	Present Subjnct.	Conditional	Imperfect Subjnct.	Future Subjnct.
1SG	amo	amaba	amaré	amé	ame	amaría	amara	amare
2SG	amas	amabas	amarás	amaste	ames	amarías	amaras	amares
3SG	ama	amaba	amará	amó	ame	amaría	amara	amáreme
1PL	amamos	amábamos	amaremos	amamos	amemos	amaríamos	amáramos	amáremos
2PL	amáis	amabais	amaréis	amasteis	améis	amaríais	amarais	amareis
3PL	aman	amaban	amarán	amaron	amen	amarían	amaran	amaren

To love in Arabic

- ?

Review: What is morphology?

- **The study of how words are composed of morphemes (the smallest meaning-bearing units of a language)**
  - **Stems**
  - **Affixes (prefixes, suffixes, circumfixes, infixes)**
    - **Immaterial**
    - **Trying**
    - **Gesagt**
    - **تكاتبا سنلزمكموها**
  - **Concatenative vs. Templatic (non-concatenative) (e.g. Arabic root-and-pattern)**

Review: What is morphology?

- **Multiple affixes**
  - **Unreadable**
    - **تكاتبا سنلزمكموها**
  - **Agglutinative languages**
    - **(e.g. Turkish, Japanese)**
  - **vs. inflectional languages**
    - **(e.g. Latin, Russian)**
  - **vs. analytic languages**
    - **(e.g. Mandarin)**

English Inflectional Morphology

- **Word stem combines with grammatical morpheme**
  - **Usually produces word of same class**
  - **Usually serves a syntactic function (e.g. agreement)**

    like → likes or liked

    bird → birds
- **Nominal morphology**
  - **Plural forms**
    - **s or es**
    - **Irregular forms**
    - **Mass vs. count nouns (email or emails)**
  - **Possessives**

Review: What is morphology?

- **Verbal inflection**
  - **Main verbs (sleep, like, fear) verbs are relatively regular**
    - **-s, ing, ed**
    - **And productive: Emailed, instant-messaged, faxed**
    - **But eat/ate/eaten, catch/caught/caught**
  - **Primary (be, have, do) and modal verbs (can, will, must) are often irregular and not productive**
    - **Be: am/is/are/were/was/been/being**
  - **Irregular verbs few (~250) but frequently occurring**
  - **English verbal inflection is much simpler than e.g. Latin**

English Derivational Morphology

- **Word stem combines with grammatical morpheme**
  - **Usually produces word of different class**
  - **More complicated than inflectional**
- **Example: nominalization**
  - **-ize verbs → -ation nouns**
  - **generalize, realize → generalization, realization**
- **Example: verbs, nouns → adjectives**
  - **embrace, pity→ embraceable, pitiable**
  - **care, wit → careless, witless**

- **Example: adjective → adverb**
  - **happy → happily**
- **More complicated to model than inflection**
  - **Less productive: *science-less, *concern-less, *go-able, *sleep-able**
  - **Meanings of derived terms harder to predict by rule**
    - **clueless, careless, nerveless**

Parsing

- **Taking a surface input and identifying its components and underlying structure**
- **Morphological parsing: parsing a word into stem and affixes and identifying the parts and their relationships**
  - **Stem and features:**
    - **goose → goose +N +SG or goose + V**
    - **geese → goose +N +PL**
    - **gooses → goose +V +3SG**
  - **Bracketing: indecipherable → [in [[de [cipher]] able]]**

Why parse words?

- **For spell-checking**
  - **Is muncheble a legal word?**
- **To identify a word's part-of-speech (POS)**
  - **For sentence parsing, for machine translation, …**
- **To identify a word's stem**
  - **For information retrieval**
- **Why not just list all word forms in a lexicon?**

What do we need to build a morphological parser?

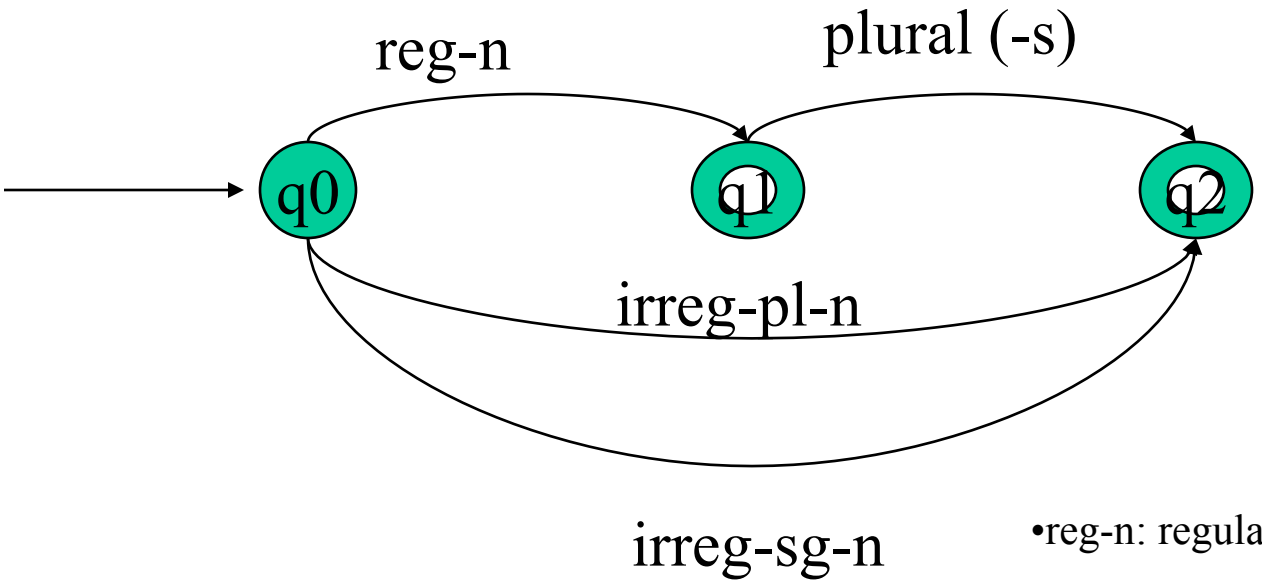
- **Lexicon: stems and affixes (w/ corresponding Part of Speech (POS))**
- **Morphotactics of the language: model of how morphemes can be affixed to a stem**
- **Orthographic rules: spelling modifications that occur when affixation occurs**
  - **in → il in context of l (in- + legal)**

Syntax and Morphology

- **Phrase-level agreement**
  - **Subject-Verb**
    - **Ali studies hard (*STUDY+3SG*)**
- **Sub-word phrasal structures**
  - ولحاجاتنا
  - و+ل+حاجات+نا
  - **and+for+need+PL+Poss:1PL**
  - ***And for our needs***

Morphotactic Models

- **English nominal inflection**



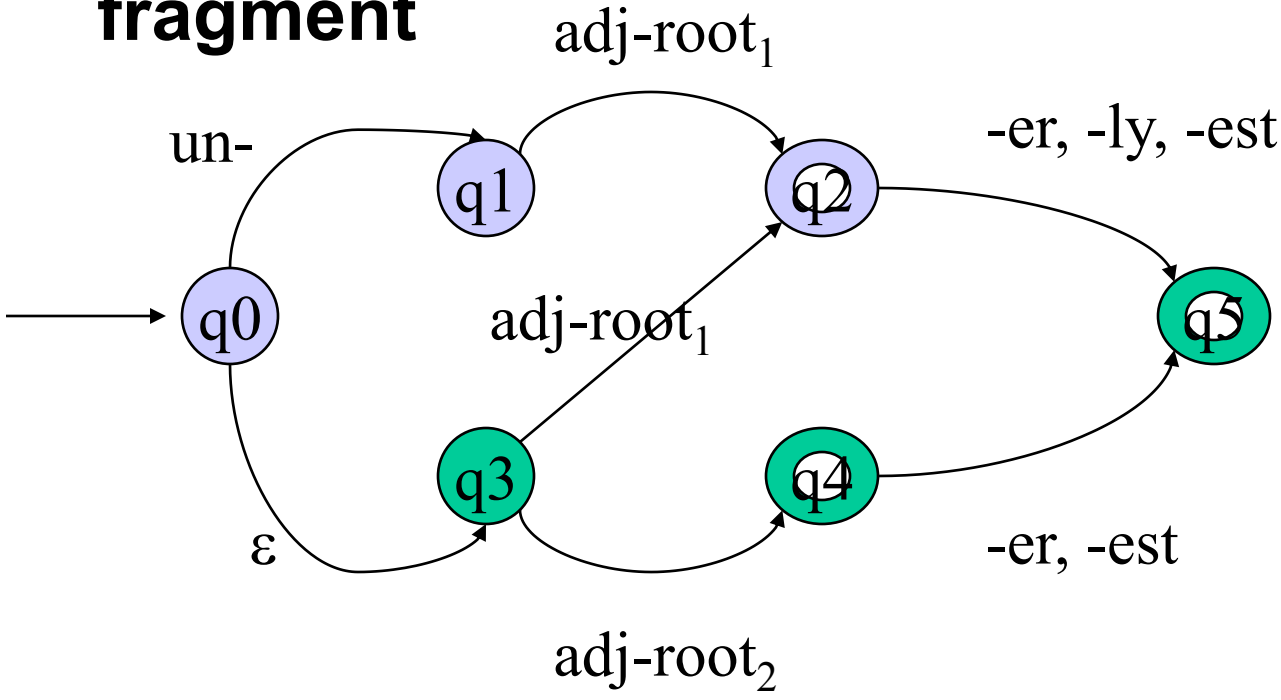
•reg-n: regular noun

•irreg-pl-n: irregular plural noun

•irreg-sg-n: irregular singular noun

•Inputs: cats, goose, geese

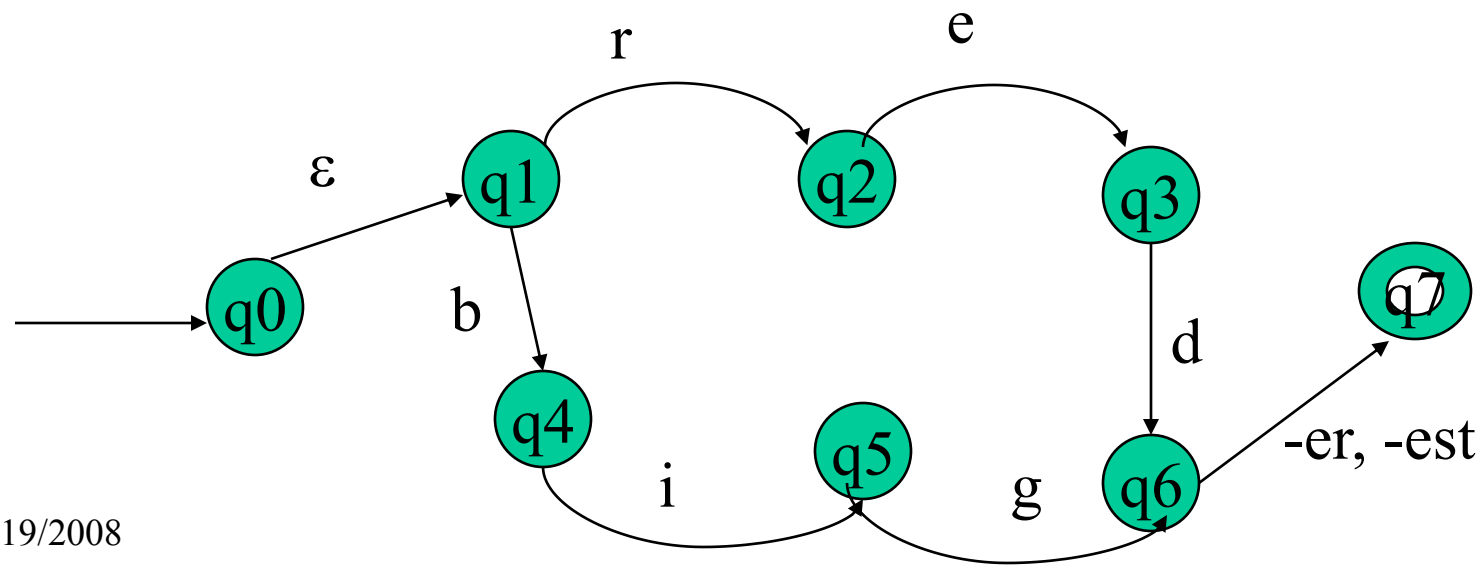
- **Derivational morphology: adjective fragment**



- Adj-root$_1$: clear, happy, real
- Adj-root$_2$: big, red

Using FSAs to Represent the Lexicon and Do Morphological Recognition

- **Lexicon: We can expand each non-terminal in our NFSA into each stem in its class (e.g. adj_root_2 = {big, red}) and expand each such stem to the letters it includes (e.g. red → r e d, big → b i g)**



Limitations

- **To cover all of English will require very large FSAs with consequent search problems**
  - **Adding new items to the lexicon means re-computing the FSA**
  - **Non-determinism**
- **FSAs can only tell us whether a word is in the language or not – what if we want to know more?**
  - **What is the stem?**
  - **What are the affixes?**
  - **We used this information to build our FSA: can we get it back?**

Parsing with Finite State Transducers

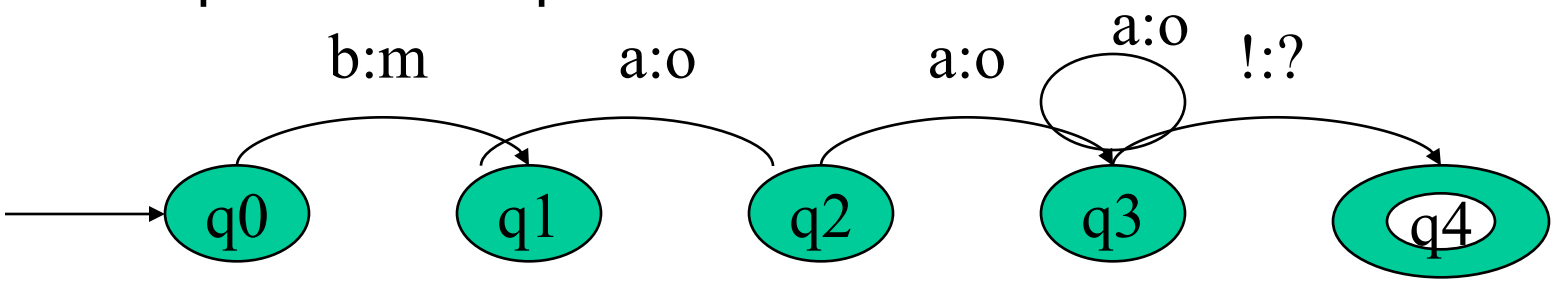
- **cats →cat +N +PL**
- **Kimmo Koskenniemi's two-level morphology**
  - **Words represented as correspondences between lexical level (the morphemes) and surface level (the orthographic word)**
  - **Morphological parsing :building mappings between the lexical and surface levels**

	c	**a**	**t**	**+N**	**+PL**	
	c	**a**	**t**	**s**		

Finite State Transducers

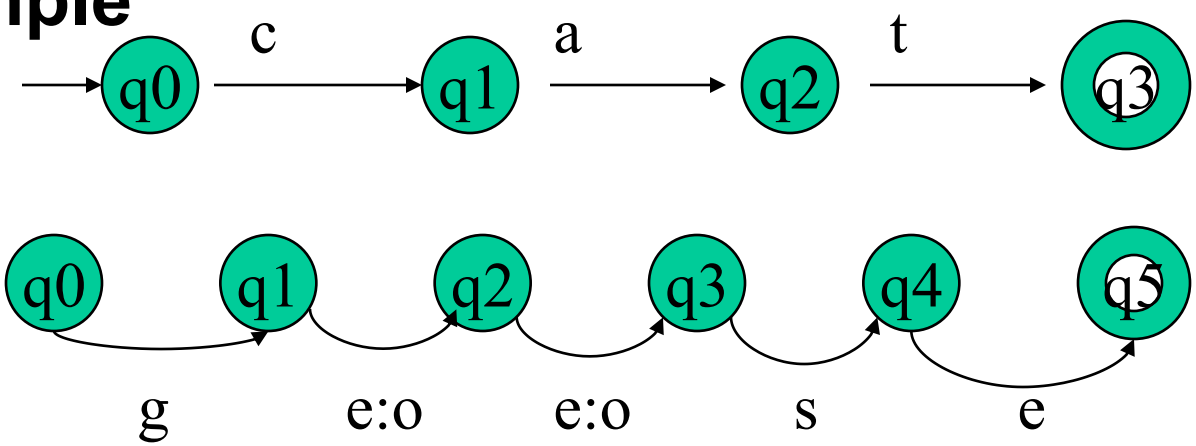
- **FSTs map between one set of symbols and another using an FSA whose alphabet Σ is composed of pairs of symbols from input and output alphabets**
- **In general, FSTs can be used for**
  - **Translator (Hello:مرحبا)**
  - **Parser/generator (Hello:How may I help you?)**
  - **To map between the lexical and surface levels of Kimmo's 2-level morphology**

- **FST is a 5-tuple consisting of**
  - **Q: set of states {q0,q1,q2,q3,q4}**
  - **Σ: an alphabet of complex symbols, each is an i/o pair such that i ∈ I (an input alphabet) and o ∈ O (an output alphabet) and Σ is in I x O**
  - **q0: a start state**
  - **F: a set of final states in Q {q4}**
  - **δ(q,i:o): a transition function mapping Q x Σ to Q**
  - Emphatic Sheep → Quizzical Cow



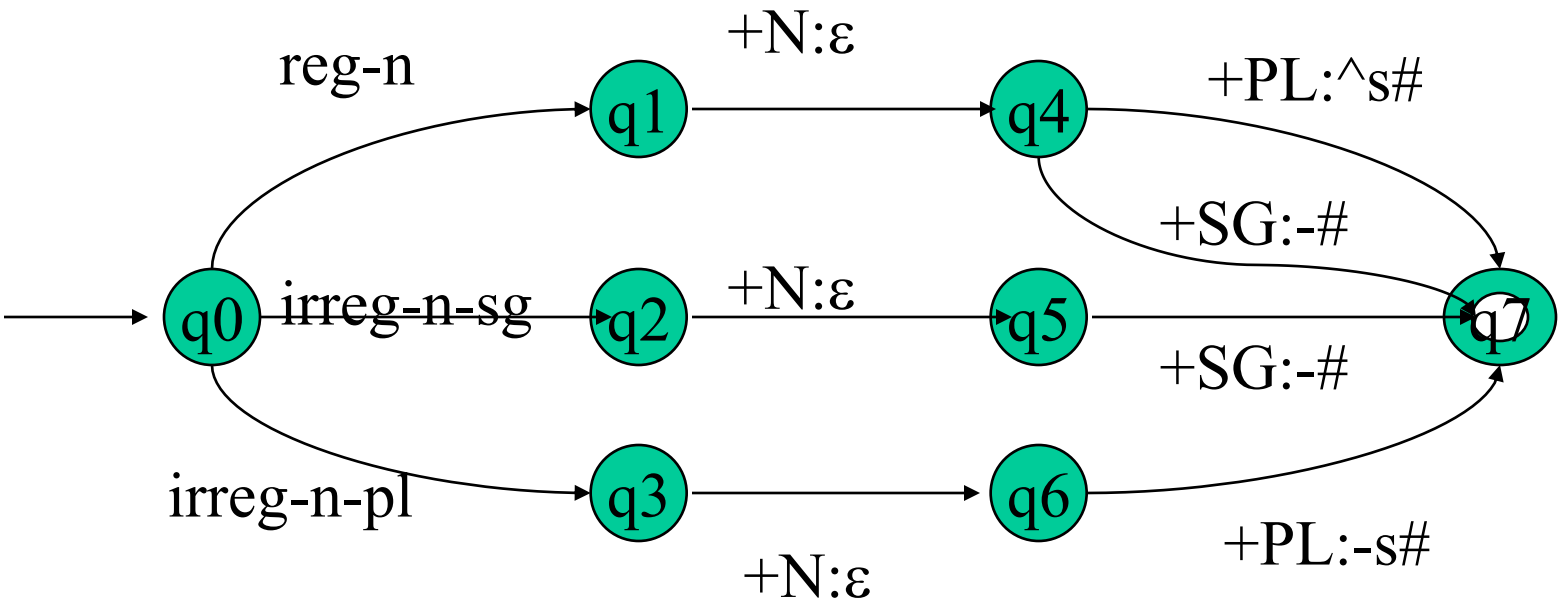
FST for a 2-level Lexicon

- **Example**



Reg-n	Irreg-pl-n	Irreg-sg-n
c a t	g o:e o:e s e	g o o s e

FST for English Nominal Inflection



Combining (cascade or composition) this FSA with FSAs for each noun type replaces e.g. reg-n with every regular noun representation in the lexicon

Orthographic Rules and FSTs

- **Define additional FSTs to implement rules such as consonant doubling (beg → begging), 'e' deletion (make → making), 'e' insertion (watch → watches), etc.**

Lexical	**f**	**o**	**x**	**+N**	**+PL**	
Intermediate	**f**	**o**	**x**	**^**	**s**	**#**
Surface	**f**	**o**	**x**	**e**	**s**	

- **Note: These FSTs can be used for generation as well as recognition by simply exchanging the input and output alphabets (e.g. ^s#:+PL)**

Administration

- Next Sunday: Quiz 1: 20 Minutes In the class
- Assignment 2: What was your findings about Python?
- New Assignment (3)

Assignment 3: Part 1
A genre for your Corpora

- Choose a Domain for your Corpora
  - Technology and Computers
  - Management
  - Weather
  - Sport
  - Economics
  - Politics
  - Education
  - Health care
  - Religion
  - History
  - Traditional Poems
  - New Poems
  - Other suggested fields

Assignment 3: Part 1
A genre for your Corpora

- Put your choice on the discussion list named 'My Corpora'.
- read other selections before
- Avoid selecting a topic that has been selected
- You might need to suggest unlisted field
  - with the arrangement of the instructor
- Collect text files and keep them in one directory as your corpora for future use
- *Suggested total size (sum of sizes of all text files)*
  - *larger than 10Mbyte of Arabic text*

Assignment 3: Part 2
List text files in a chosen directory

- Write a program that allows the user to browse and select a directory, then the program will list the names of the text files in that directory. This program is needed to be used for future assignments and the course project. You can use any language you are mastering. However, Python might be a good choice

Assignment 3: Part 3
The most used n words in your corpora

- After building your corpora, you need to find the most used 100 words in your corpora. You might do that by writing a program that let the user choose the directory of the corpora where the text files are located and find the most use n words. Where n could be 100.

Thank you

السلام عليكم ورحمة الله