ΓΕ77
COMPUTATIONAL LINGUISTICS

Athanasios N. Karasimos

akarasimos@gmail.com

BA in Linguistics | School of English Language and Literature
National and Kapodistrian University of Athens

Lecture 7 | Wed 9 May 2018
Lecture 8 | Wed 23 May 2018

WORDS AND TRANSDUCERS

LECTURE 5 RECAP

Language Modeling with N-Grams

FINITE-STATE AUTOMATA

- Any regular expression can be realized as a **finite state automaton** (**FSA**).
- An automaton implicitly defines a **formal language** as the set of strings the automaton **accepts**.
- An automaton can use any set of symbols for its vocabulary, including letters, words, or even graphic images.
- The behavior of a **deterministic** automaton (**DFSA**) is fully determined by the state it is in.
- A **non-deterministic** automaton (**NFSA**) sometimes has to make a choice between multiple paths to take given the same current state and next input.
- Any **NFSA** can be converted to a **DFSA.**

MORPHOLOGY PART 2

Lets talk about WORDS

THE CASE OF PLURAL

- Simple cases of plural:
  - *woodchuck* to *woodchucks*

- *But* what about fox, peccary, goose and fish?
- **Orthographic rules:** *peccary* to *peccaries*
- **Morphological rules**: *fish* with 0 plural suffix
  goose to *geese* with vowel change
- **Phonological rules**: *fox* to *foxes*

~~MORPHOLOGICAL~~ PARSING

- The task to recognize that a word (like *foxes*) breaks down into component morphemes (*fox* and *-es*) and building a structured representation of this fact is called **morphological parsing**.
- **Parsing** means taking an input and producing some sort of linguistic structure for it.
- We use the term parsing very broadly, including many kinds of structures that might be produced; morphological, syntactic, semantic, discourse; in the form of a string, or a tree, or a network.

MORPHOLOGICAL PARSING

- Morphological parsing or stemming(?) applies to many affixes other than plurals;
  - for example we might need to take any English verb form ending in *-ing* (*going*, *talking*, *congratulating*) and parse it into its verbal stem plus the *-ing* morpheme.
- So given the **surface** or **input form** *going*, we might want to produce the parsed form VERB-go + GERUND-ing.
- Morphological parsing is important throughout speech and language processing. It plays a crucial role in Web search for morphologically complex languages like Greek, Russian or German.

MORPHOLOGICAL PARSING

- Morphological parsing also plays a crucial role in part-of-speech tagging for these morphologically complex languages.
- It is important for producing the large dictionaries that are necessary for robust spell-checking.
- It is necessary in machine translation to realize for example that the French words *va* and *aller* should both translate to forms of the English verb *go*.

SURVEY OF ENGLISH MORPHOLOGY

A FAMILIAR FACE: MORPHEMES

- A **morpheme** is often defined as the minimal meaning-bearing unit in a language.
  - So for example the word *fox* consists of a single morpheme (the morpheme *fox*) while the word *cats* consists of two: the morpheme *cat* and the morpheme *-s*.
- As this example suggests, it is often distinguish two broad classes of morphemes: **stems** and **affixes**.
- Affixes: divided into **prefixes**, **suffixes**, **infixes**, and **circumfixes**. Prefixes precede the stem, suffixes follow the stem, circumfixes do both, and infixes are inserted inside the stem.
  - Circumfixes: [German] past participles (ge- and –en/-t)
- Infixes: [Tagalog] affix *um*, which marks the agent of an action, is infixed to the stem *hingi* "borrow" to produce *humingi*.

WORD FORMATION PROCESSES

- Four processes are common and play important roles in speech and language generation: **inflection**, **derivation**, **compounding**, and **cliticization.**
  - **Inflection** is the combination of a word stem with a grammatical morpheme, usually resulting in a word of the same class as the original stem, and usually filling some syntactic function like agreement.
  - **Derivation** is the combination of a word stem with a grammatical morpheme, usually resulting in a word of a *different* class, often with a meaning hard to predict exactly.
  - **Compounding** is the combination of multiple word stems together.
  - **Cliticization** is the combination of a word stem with a **clitic**. A clitic is a morpheme that acts syntactically like a word, but is reduced in form and attached (phonologically and sometimes orthographically) to another word.

MORPHOLOGICAL TASKS

- TASK I:
  - Give two examples from each word formation process
- TASK II:
  - Consider possible problematic cases for morphological parsing (inf, dev, com).
- TASK III:
  - Test these cases with a morphological parser.
  - http://nlpdotnet.com/services/Morphparser.aspx
- TASK IV:
  - Ambiguity of morphological parsing.
  - https://open.xerox.com/Services/fst-nlp-tools/Consume/Morphological%20Analysis-176
  - http://langrid.org/playground/morphological-analyzer.html

INFLECTIONAL ENGLISH

- Nominal suffixes: an affix that marks **plural** and an affix that marks p**ossessive**.
  - Regular plural suffix *-s* (also spelled *-es*), and irregular plurals:

	Regular Nouns		Irregular Nouns	
Singular	cat	thrush	mouse	ox
Plural	cats	thrushes	mice	oxen

- While the regular plural is spelled *-s* after most nouns, it is spelled *-es* after words ending in *-s* (*ibis/ibises*), *-z* (*waltz/waltzes*), *-sh* (*thrush/thrushes*), *-ch* (*finch/finches*), and sometimes *-x* (*box/boxes*). Nouns ending in *-y* preceded by a consonant change the *-y* to *-i* (*butterfly/butterflies*).
- The possessive suffix is realized by apostrophe + *-s* for regular singular nouns (*llama's*) and plural nouns not ending in *-s* (*children's*) and often by a lone apostrophe after regular plural nouns (*llamas'*) and some names ending in *-s* or *-z* (*Euripides' comedies*).

INFLECTIONAL ENGLISH

- English verbal inflection is more complicated than nominal inflection.
  - **main verbs**, (*eat, sleep, impeach*), **modal verbs** (*can, will, should*), and **primary verbs** (*be, have, do*).
- 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

- we can predict the other forms by adding one of three predictable endings and making some regular spelling.

INFLECTIONAL MORPHOLOGY

- The **irregular verbs** are those that have some more or less idiosyncratic forms of inflection. Irregular verbs in English often have five different forms, but can have as many as eight (e.g., the verb *be*) or as few as three (e.g. *cut* or *hit*).
- Morphological Form Classes

	Irregularly Inflected Verbs		
stem	eat	catch	cut
-s form	eats	catches	cuts
-ing participle	eating	catching	cutting
Past form	ate	caught	cut
-ed/-en participle	eaten	caught	cut

More complex verbal inflectional paradigm of morphologically rich languages.

DERIVATIONAL ENGLISH

- While English inflection is relatively simple compared to other languages, derivation in English is quite complex.
- A very common kind of derivation in English is the formation of new nouns, often from verbs or adjectives. This process is called **nominalization**.
  - For example, the suffix *-ation* produces nouns from verbs ending often in the suffix *-ize* (*computerize* → *computerization*).

COMPOUNDING ENGLISH

- Most English compound nouns are noun phrases (i.e. nominal phrases) that include a noun modified by adjectives or noun adjuncts.
  - The monoword forms in which two usually moderately short words appear together as one. Examples are housewife, lawsuit, wallpaper, basketball, etc.
  - The hyphenated form in which two or more words are connected by a hyphen. Compounds that contain affixes, such as house-build(er) and single-mind(ed)(ness), as well as adjective-adjective compounds and verb-verb compounds, such as blue-green and freeze-dried.
  - Loose compounds: the open or spaced form consisting of newer combinations of usually longer words, such as distance learning, player piano, lawn tennis, etc.

Modifier	Head	Compound
noun	noun	football
adjective	noun	blackboard
verb	noun	breakwater
preposition	noun	underworld
noun	adjective	snow white
adjective	adjective	blue-green
verb	adjective	tumbledown
preposition	adjective	over-ripe
noun	verb	browbeat
adjective	verb	highlight
verb	verb	freeze-dry
preposition	verb	undercut
noun	preposition	love-in
adverb	preposition	forthwith
verb	preposition	takeout
preposition	preposition	without

FINITE-STATE MORPHOLOGICAL PARSING

MORPHOLOGICAL FEATURES

some
some +*Pron+NomObl+3P+Pl*
some +*Det+SP*
features
<feature> +*Noun+Pl*
<feature> +*Verb+Pres+3sg*
that
that +*Conj+Sub*
that +*Det+Sg*
that +*Pron+NomObl+3P+Sg*
that +*Pron+Rel+NomObl+3P+SP*
<that> +*Adv*

- εργασία
  - εργασία +*Noun+Common+Fem+Sg+Acc*
  - εργασία +*Noun+Common+Fem+Sg+Voc*
  - εργασία +*Noun+Common+Fem+Sg+Nom*
- υπάρχουν
  - υπάρχω +*Verb+Indic+Pres+P3+Pl+Imperf+Active*
- σχόλια
  - σχόλιο +*Noun+Common+Neut+Pl+Acc*
  - σχόλιο +*Noun+Common+Neut+Pl+Voc*
  - σχόλιο +*Noun+Common+Neut+Pl+Nom*
- ανατροφοδότησης
  - ανατροφοδότηση +*Noun+Common+Fem+Sg+Gen*

MORPHOLOGICAL FEATURES

- The features specify additional information about the stem.
  - For example the feature +N means that the word is a noun; +Sg means it is singular, +Pl that it is plural. (check also Chapter 5 and Chapter 16); for now, consider +Sg to be a primitive unit that means "singular".
- Greek has some features that don't occur in English; for example the nouns *εργασία* and *ανατροφοδότησης* are marked +Fem (feminine).
- Note that some of the input forms will be ambiguous between different morphological parses. For now, we will consider the goal of morphological parsing merely to list all possible parses.

BUILDING A MORPHOLOGICAL PARSER

- **lexicon:** the list of stems and affixes, together with basic information about them (whether a stem is a Noun stem or a Verb stem, etc.).
- **morphotactics:** the model of morpheme ordering that explains which classes of morphemes can follow other classes of morphemes inside a word. For example, the fact that the English plural morpheme follows the noun rather than preceding it is a morphotactic fact.
- **orthographic rules:** these **spelling rules** are used to model the changes that occur in a word, usually when two morphemes combine (e.g., the *y→ie* spelling rule discussed above that changes *city* + *-s* to *cities* rather than *citys*).

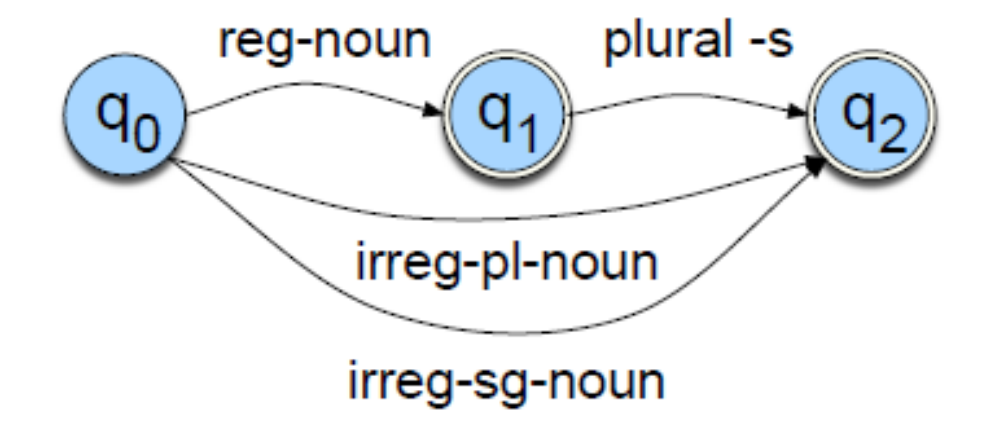
A lexicon is a repository for words. The simplest possible lexicon would consist of an explicit list of every word of the language (*every* word, i.e., including abbreviations ("AAA") and proper names ("Jane" or "Beijing")) as follows:

a, AAA, AA, Aachen, aardvark, aardwolf, aba, abaca, aback, . . .

Inconvenient or impossible to list every word in the language, computational lexicons are usually structured with a list of each of the stems and affixes of the language together with a representation of the morphotactics that tells us how they can fit together.

BUILDING A FINITE-STATE LEXICON

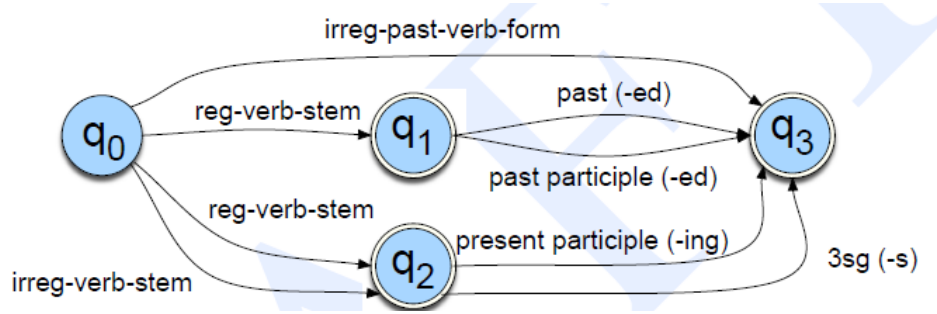
FINITE-STATE FOR NOMINAL PLURAL



How can we expand this finite-state transducer?

reg-noun	irreg-pl-noun	irreg-sg-noun	plural
fox	geese	goose	-s
cat	sheep	sheep	
aardvark	mice	mouse	

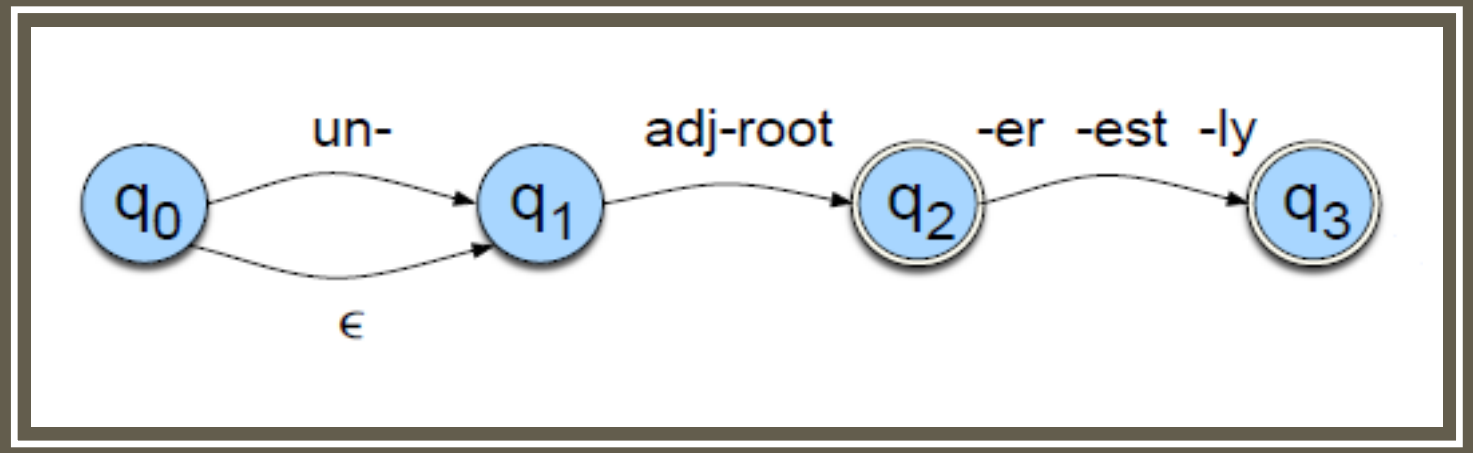
FINITE-STATE FOR VERBAL TYPES



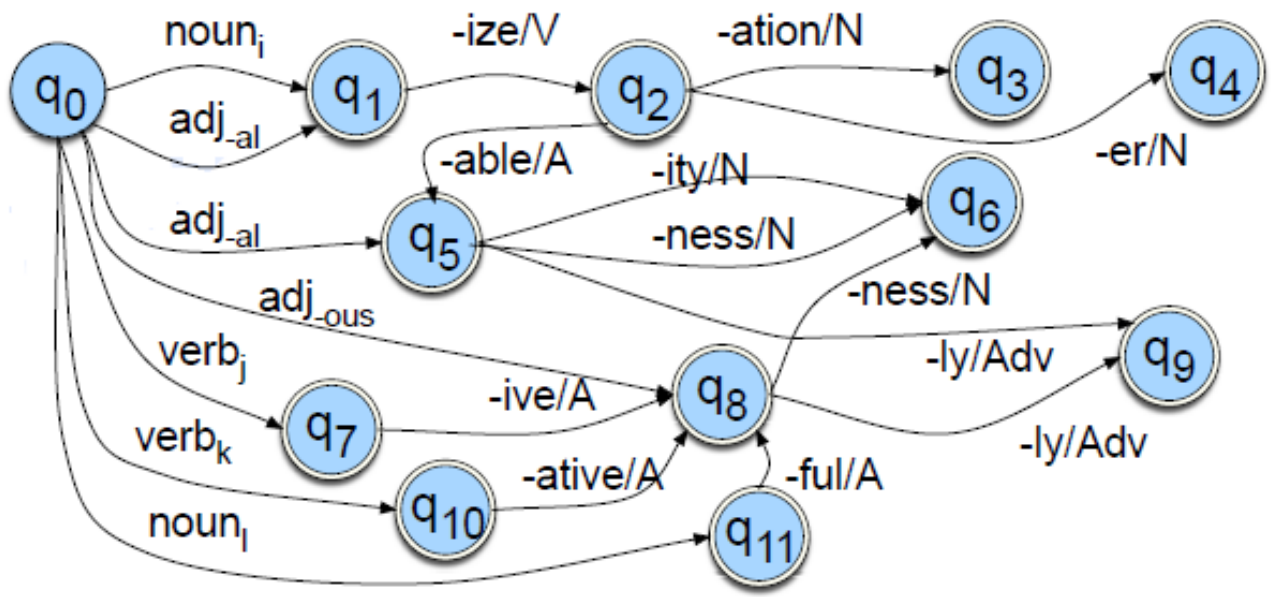
reg-verb-stem	irreg-verb-stem	irreg-past-verb	past	past-part	pres-part	3sg
walk	cut	caught	-ed	-ed	-ing	-s
fry	speak	ate				
talk	sing	eaten				
impeach		sang				

FINITE-STATE FOR ADJECTIVES

- big, bigger, biggest, cool, cooler, coolest, coolly
- happy, happier, happiest, happily red, redder, reddest
- unhappy, unhappier, unhappiest, unhappily real, unreal, really
- clear, clearer, clearest, clearly, unclear, unclearly



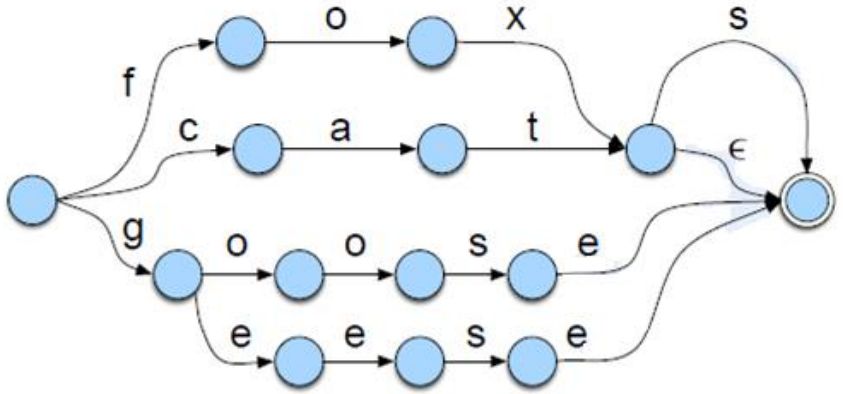
FINITE-STATE FOR DERIVATION



- i.e. *fossilize*, we can predict the word *fossilization* by following states *q*0, *q*1, and *q*2.
- Similarly, adjectives ending in *-al* or *-able* at *q*5 (*equal*, *formal*, *realizable*) can take the suffix *-ity*, or sometimes the suffix *-ness* to state *q*6 (*naturalness*, *casualness*).

MORPHOLOGICAL RECOGNITION

We can now use these FSAs to solve the problem of **morphological recognition**; that is, of determining whether an input string of letters makes up a legitimate English word or not. We do this by taking the morphotactic FSAs, and plugging in each "sublexicon" into the FSA. That is, we expand each arc (e.g., the **reg-noun-stem** arc) with all the morphemes that make up the set of **reg-noun-stem**.



FINITE-STATE TRANSDUCERS

FINITE-STATE TRANSDUCER: DEFINITION

- A transducer maps between one representation and another; a **finite-state transducer (FST)** is a type of finite automaton which maps between two sets of symbols.
- We can visualize an FST as a two-tape automaton which recognizes or generates *pairs* of strings. Intuitively, we can do this by labeling each arc in the finite-state machine with two symbol strings, one from each tape
- More general function than an FSA; where an FSA defines a formal language by defining a set of strings, an FST defines a *relation* between sets of strings.
- Another way of looking at an FST is as a machine that reads one string and generates another.

"FOUR-FOLD WAY" OF TRANSDUCERS

- **FST as recognizer:** a transducer that takes a pair of strings as input and outputs *accept* if the string-pair is in the string-pair language, and *reject* if it is not.
- **FST as generator:** a machine that outputs pairs of strings of the language. Thus the output is a yes or no, and a pair of output strings.
- **FST as translator:** a machine that reads a string and outputs another string.
- **FST as set relater:** a machine that computes relations between sets.

PARAMETERS OF FST

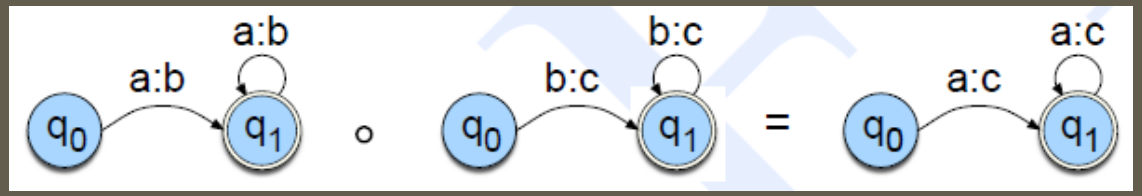
- **Q** a finite set of N states $q0, q1, \ldots, qN-1$
- **Σ** a finite set corresponding to the input alphabet
- **Δ** a finite set corresponding to the output alphabet
- $q0 \in Q$ the start state
- $F \subseteq Q$ the set of final states
- **δ(q,w)** the transition function or transition matrix between states; Given a state $q \in Q$ and a string $w \in \Sigma*$, d(q,w) returns a set of new states $Q' \in Q$. **δ** is thus a function from $Q \times \Sigma*$ to 2^Q (because there are $2Q$ possible subsets of Q). d returns a set of states rather than a single state because a given input may be ambiguous in which state it maps to.
- **σ(q,w)** the output function giving the set of possible output strings for each state and input. Given a state $q \in Q$ and a string $w \in \Sigma*$, $\sigma(q,w)$ gives a set of output strings, each a string $o \in$ D$*$. s is thus a function from $Q \times$S$*$ to $2^{\Delta}*$

REGULAR RELATIONS

- **Regular relations** are sets of pairs of strings, a natural extension of the regular languages, which are sets of strings.

FSTs have two additional closure properties that turn out to be extremely useful:

- **inversion**: The inversion of a transducer T (T^{-1}) simply switches the input and output labels. Thus if T maps from the input alphabet I to the output alphabet O, T^{-1} maps from O to I.
- **composition**: If $T1$ is a transducer from $I1$ to $O1$ and $T2$ a transducer from $O1$ to $O2$, then $T1 \circ T2$ maps from $I1$ to $O2$.



FST AS MORPHOLOGICAL PARSER

Coming soon…

READINGS

- Jurafsky D. & J. Martin (2008). SPEECH and LANGUAGE PROCESSING An introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (2nd Edition). CHAPTER 3 (pp. 1-16).

Additional References:

- Μαρκόπουλος, Γ. (1997). *Υπολογιστική Επεξεργασία του Ελληνικού Ονόματος.* Διδακτορική διατριβή (σσ. 99-106).
- Πετροπούλου, Ε. (2012). Η *Σύνθεση με Δεσμευμένο Θέμα στην Αγγλική και τη Νέα Ελληνική Θεωρητική Ανάλυση και Υπολογιστική Επεξεργασία.* Διδακτορική διατριβή ((σσ. 160-164)-172)).