Natural Language Processing
Unit-I

Unit-I

1.Finding the Structure of Words

This section deals with words, its structure and its models

1.1 Words and Their Components

- 1.1.1 Tokens
- 1.1.2 Lexemes
- 1.1.3 Morphemes
- 1.1.4 Typology

1.2 Issues and Challenges

- 1.2.1 Irregularity
- 1.2.2 Ambiguity
- 1.2.3 Productivity

1.3 Morphological Models

- 1.3.1 Dictionary Lookup
- 1.3.2 Finite-State Morphology
- 1.3.3 Unification-Based Morphology
- 1.3.4 Functional Morphology
- 1.3.5 Morphology Induction

2.**Finding the Structure of Documents**

This chapter mainly deals with Sentence and topic detection or segmentation.

2.1 Introduction

- 2.1.1 Sentence Boundary Detection
- 2.1.2 Topic Boundary Detection

2.2 Methods

This section deals with statistical classical approaches (Generative and Discriminative approaches)

- 2.2.1 Generative Sequence Classification Methods
- 2.2.2 Discriminative Local Classification Methods
- 2.3.3 Discriminative Sequence Classification Methods
- 2.2.4 Hybrid Approaches
- 2.2.5 Extensions for Global Modelling for Sentence Segmentation

2.3 Complexity of the Approaches

2.4 Performance of the Approaches

NATURAL LANGUAGE PROCESSING(NLP)

UNIT - I

i.Finding the Structure of Words:

- Words and Their Components
- Issues and Challenges
- Morphological Models

ii.Finding the Structure of Documents:

- Introduction
- Methods
- Complexity of the Approaches
- Performances of the Approaches

Natural Language Processing

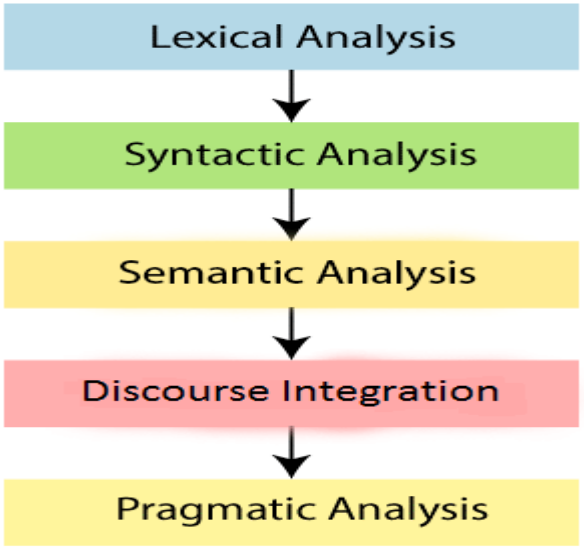
- Humans communicate through some form of language either by text or speech.
- To make interactions between computers and humans, computers need to understand natural languages used by humans.
- Natural language processing is all about making computers learn, understand, analyse, manipulate and interpret natural(human) languages.
- NLP stands for **Natural Language Processing**, which is a part of **Computer Science, Human language,** and **Artificial Intelligence**.
- Processing of Natural Language is required when you want an intelligent system like robot to perform as per your instructions, when you want to hear decision from a dialogue based clinical expert system, etc.
- The ability of machines to interpret human language is now at the core of many applications that we use every day - chatbots, Email classification and spam filters, search engines, grammar checkers, voice assistants, and social language translators.
- The input and output of an NLP system can be Speech or Written Text

NLP Terminology

- **Phonology** − It is study of organizing sound systematically.
- ***Morphology***: The study of the formation and internal structure of words.
- **Morpheme** − It is primitive unit of meaning in a language.
- ***Syntax***: The study of the formation and internal structure of sentences.
- ***Semantics***: The study of the meaning of sentences.
- **Pragmatics** − It deals with using and understanding sentences in different situations and how the interpretation of the sentence is affected.
- **Discourse** − It deals with how the immediately preceding sentence can affect the interpretation of the next sentence.
- **World Knowledge** − It includes the general knowledge about the world.

Steps in NLP

- There are general five steps :
    1. Lexical Analysis
    2. Syntactic Analysis (Parsing)
    3. Semantic Analysis
    4. Discourse Integration
    5. Pragmatic Analysis



Lexical Analysis –

- The first phase of NLP is the Lexical Analysis.
- This phase scans the source code as a stream of characters and converts it into meaningful lexemes.
- It divides the whole text into paragraphs, sentences, and words.

Syntactic Analysis (Parsing) –

- Syntactic Analysis is used to check grammar, word arrangements, and shows the relationship among the words.
- The sentence such as “The school goes to boy” is rejected by English syntactic analyzer.

Semantic Analysis –

- Semantic analysis is concerned with the meaning representation.
- It mainly focuses on the literal meaning of words, phrases, and sentences.
- The semantic analyzer disregards sentence such as “hot ice-cream”.

Discourse Integration –

- Discourse Integration depends upon the sentences that proceeds it and also invokes the meaning of the sentences that follow it.

Pragmatic Analysis –

- During this, what was said is re-interpreted on what it actually meant.
- It involves deriving those aspects of language which require real world knowledge.
- **Example:** "Open the door" is interpreted as a request instead of an order.

Finding the Structure of Words

- Human language is a complicated thing.
- We use it to express our thoughts, and through language, we receive information and infer its meaning.
- Trying to understand language all together is not a viable approach.
- Linguists have developed whole disciplines that look at language from different perspectives and at different levels of detail.
- The point of **morphology**, for instance, is to study the variable forms and functions of words,
- The syntax is concerned with the arrangement of words into phrases, clauses, and sentences.
- Word structure constraints due to pronunciation are described by **phonology**,
- The conventions for writing constitute the **orthography** of a language.
- The meaning of a linguistic expression is its semantics, and etymology and lexicology cover especially the evolution of words and explain the semantic, morphological, and other links among them.
- Words are perhaps the most intuitive units of language, yet they are in general tricky to define.
- Knowing how to work with them allows, in particular, the development of **syntactic** and **semantic** abstractions and simplifies other advanced views on language.
- Here, first we explore how to identify words of distinct types in human languages, and how the internal structure of words can be modelled in connection with the grammatical properties and lexical concepts the words should represent.

- The discovery of word structure is **morphological parsing**.
- In many languages, words are delimited in the orthography by whitespace and punctuation.
- But in many other languages, the writing system leaves it up to the reader to tell words apart or determine their exact phonological forms.

Words and Their Components

- Words are defined in most languages as the smallest linguistic units that can form a complete utterance by themselves.
- The minimal parts of words that deliver aspects of meaning to them are called **morphemes**.

Tokens

- Suppose, for a moment, that words in English are delimited only by whitespace and punctuation (the marks, such as full stop, comma, and brackets)
- Example: Will you read the newspaper? Will you read it? I won't read it.

- If we confront our assumption with insights from syntax, we notice two here: words ***newspaper*** and ***won't***.
- Being a compound word, ***newspaper*** has an interesting **derivational structure**.
- In writing, *newspaper* and the associated concept is distinguished from the isolated *news* and *paper*.
- For reasons of generality, linguists prefer to analyze *won't* as two syntactic words, or tokens, each of which has its independent role and can be reverted to its normalized form.
- The structure of ***won't*** could be parsed as ***will*** followed by ***not***.
- In English, this kind of tokenization and **normalization** may apply to just a limited set of cases, but in other languages, these phenomena have to be treated in a less trivial manner.
- In Arabic or Hebrew, certain tokens are concatenated in writing with the preceding or the following ones, possibly changing their forms as well.
- The underlying lexical or syntactic units are thereby blurred into one compact string of letters and no longer appear as distinct words.
- Tokens behaving in this way can be found in various languages and are often called **clitics**.
- In the writing systems of Chinese, Japanese, and Thai, whitespace is not used to separate words.

Lexemes

- By the term word, we often denote not just the one **linguistic form** in the given context but also the **concept behind the form** and the **set of alternative forms** that can express it.
- Such **sets** are called **lexemes or lexical items**, and they constitute the **lexicon** of a language.
- Lexemes can be divided by their behaviour into the lexical categories of verbs, nouns, adjectives, conjunctions, particles, or other parts of speech.
- The citation **form of a lexeme**, by which it is commonly identified, is also called its **lemma**.
- When we convert a word into its other forms, such as turning the **singular *mouse*** into the **plural *mice* or *mouses***, we say we **inflect** the lexeme.
- When we transform a lexeme into another one that is morphologically related, regardless of its lexical category, we say we derive the lexeme: for instance, the nouns ***receiver* and *reception*** are derived from the verb *to* ***receive***.
- Example: Did you see him? I **didn’t** see him. I didn’t see **anyone**.
- Example presents the problem of tokenization of ***didn’t*** and the investigation of the internal structure of ***anyone***.

- In the paraphrase ***I saw no one***, the lexeme ***to see*** would be inflected into the form ***saw*** to reflect its grammatical function of expressing **positive past tense**.
- Likewise, ***him*** is the oblique case form of ***he*** or even of a more abstract lexeme representing all personal pronouns.
- In the paraphrase, ***no one*** can be perceived as the minimal word synonymous with ***nobody***.
- The difficulty with the definition of what counts as a word need not pose a problem for the syntactic description if we understand ***no one*** **as two closely connected tokens treated as one fixed element**.

Morphemes

- Morphological theories differ on whether and how to associate the properties of word forms with their structural components.
- These components are usually called **segments** or **morphs**.
- The morphs that by themselves represent some aspect of the meaning of a word are called **morphemes** of some function.
- Human languages employ a variety of devices by which morphs and morphemes are combined into word forms.

Morphology

- Morphology is the domain of linguistics that analyses the internal structure of words.
- Morphological analysis – exploring the structure of words
- Words are built up of minimal meaningful elements called morphemes:
  - played = play-ed
  - cats = cat-s
  - unfriendly = un-friend-ly
- Two types of morphemes:
  - i Stems: play, cat, friend
  - ii Affixes: -ed, -s, un-, -ly
- Two main types of affixes:
  - i Prefixes precede the stem: un-
  - ii Suffixes follow the stem: -ed, -s, un-, -ly
- Stemming = find the stem by stripping off affixes
  - play = play
  - replayed = re-play-ed
  - computerized = comput-er-ize-d

Problems in morphological processing

- Inflectional morphology: inflected forms are constructed from base forms and inflectional affixes.
- Inflection relates different forms of the same word

Lemma	Singular	Plural
cat	cat	cats
dog	dog	dogs
knife	knife	knives
sheep	sheep	sheep
mouse	mouse	mice

- Derivational morphology: words are constructed from roots (or stems) and derivational affixes:

  inter+national = international

  international+ize = internationalize

  internationalize+ation = internationalization

- The simplest morphological process concatenates morphs one by one, as in *dis-agree-ment-s*, where *agree* is a free lexical morpheme and the other elements are bound grammatical morphemes contributing some partial meaning to the whole word.
- in a more complex scheme, morphs can interact with each other, and their forms may become subject to additional phonological and orthographic changes denoted as morphophonemic.
- The alternative forms of a morpheme are termed **allomorphs**.

Typology

- Morphological typology divides languages into groups by characterizing the prevalent morphological phenomena in those languages.
- It can consider various criteria, and during the history of linguistics, different classifications have been proposed.
- Let us outline the typology that is based on quantitative relations between words, their morphemes, and their features:
- **Isolating**, or **analytic**, languages include no or relatively few words that would comprise more than one morpheme (typical members are Chinese, Vietnamese, and Thai; analytic tendencies are also found in English).

- **Synthetic** languages can combine more morphemes in one word and are further divided into agglutinative and fusional languages.
- **Agglutinative** languages have morphemes associated with only a single function at a time (as in Korean, Japanese, Finnish, and Tamil, etc.)
- **Fusional** languages are defined by their feature-per-morpheme ratio higher than one (as in Arabic, Czech, Latin, Sanskrit, German, etc.).
- In accordance with the notions about word formation processes mentioned earlier, we can also find out using concatenative and nonlinear:
- **Concatenative** languages linking morphs and morphemes one after another.
- **Nonlinear** languages allowing structural components to merge nonsequentially to apply tonal morphemes or change the consonantal or vocalic templates of words.

Morphological Typology

- **Morphological typology** is a way of classifying the languages of the world that groups languages according to their common morphological structures.
- The field organizes languages on the basis of how those languages form words by combining morphemes.
- The morphological typology classifies languages into **two broad classes** of **synthetic languages** and **analytical languages.**
- The **synthetic class** is then further sub classified as either **agglutinative languages** or **fusional languages.**
- Analytic languages contain very little inflection, instead relying on features like word order and auxiliary words to convey meaning.
- Synthetic languages, ones that are not analytic, are divided into two categories: agglutinative and fusional languages.
- Agglutinative languages rely primarily on discrete particles(prefixes, suffixes, and infixes) for inflection, ex: inter+national = international, international+ize = internationalize.
- While fusional languages "fuse" inflectional categories together, often allowing one word ending to contain several categories, such that the original root can be difficult to extract (anybody, newspaper).

Issues and Challenges

- **Irregularity**: word forms are not described by a prototypical linguistic model.
- **Ambiguity**: word forms be understood in multiple ways out of the context of their discourse.
- **Productivity**: is the inventory of words in a language finite, or is it unlimited?
- Morphological parsing tries to eliminate the variability of word forms to provide higher-level linguistic units whose lexical and morphological properties are explicit and well defined.
- It attempts to remove unnecessary irregularity and give limits to ambiguity, both of which are present inherently in human language.
- By irregularity, we mean existence of such forms and structures that are not described appropriately by a prototypical linguistic model.
- Some irregularities can be understood by redesigning the model and improving its rules, but other lexically dependent irregularities often cannot be generalized

- Ambiguity is indeterminacy (not being interpreted) in interpretation of expressions of language.
- Morphological modelling also faces the problem of productivity and creativity in language, by which unconventional but perfectly meaningful new words or new senses are coined.

Irregularity

- Morphological parsing is motivated by the quest for generalization and abstraction in the world of words.
- Immediate descriptions of given linguistic data may not be the ultimate ones, due to either their inadequate accuracy or inappropriate complexity, and better formulations may be needed.
- The design principles of the morphological model are therefore very important.
- In Arabic, the deeper study of the morphological processes that are in effect during inflection and derivation, even for the so-called irregular words, is essential for mastering the whole morphological and phonological system.
- With the proper abstractions made, irregular morphology can be seen as merely enforcing some extended rules, the nature of which is phonological, over the underlying or prototypical regular word forms.

P-STEM	P—3MS	P—2FS	P—3MP	II2MS	IS1—S	IJ1—S	I-STEM	
qaraʾ	*qaraʾa*	*qaraʾti*	*qaraʾū*	*taqraʾu*	*ʾaqraʾa*	*ʾaqraʾ*	*qraʾ*	S
faʿal	*faʿal-a*	*faʿal-ti*	*faʿal-ū*	*ta-fʿal-u*	*ʾa-fʿal-a*	*ʾa-fʿal*	*fʿal*	I
faʿal	*faʿal-a*	*faʿal-ti*	*faʿal-ū*	*ta-fʿal-u*	*ʾa-fʿal-a*	*ʾa-fʿal-*	*fʿal*	M
...	...-*a*	...-*ti*	...-*ū*	*ta-...-u*	*ʾa-...-a*	*ʾa-...-*	...	
faʿā	*faʿā-a*	*faʿā-ti*	*faʿā-ū*	*ta-fā-u*	*ʾa-fā-a*	*ʾa-fā-*	*fā*	M
faʿā	*faʿā*	*faʿal-ti*	*faʿ-aw*	*ta-fā*	*ʾa-fā*	*ʾa-fa*	*fā*	I
raʾā	*raʾā*	*raʾayti*	*raʾaw*	*tarā*	*ʾarā*	*ʾara*	*rā*	S

Table: Discovering the regularity of Arabic morphology using morphophonemic templates, where uniform structural operations apply to different kinds of stems.

In rows, surface forms S of *qara_* ‘to read’ and *ra_ ‿a* ‘to see’ and their inflections are analyzed into immediate I and morphophonemic M templates, in which dashes mark the structural boundaries where merge rules are enforced.

The outer columns of the table correspond to P perfective and I imperfective stems declared in the lexicon; the inner columns treat active verb forms of the following morphosyntactic properties: I indicative, S subjunctive, J jussive mood; 1 first, 2 second, 3 third person; M masculine, F feminine gender; S singular, P plural number.

- Table illustrates differences between a naive model of word structure in Arabic and the model proposed in Smrˇz and Smrˇz and Bielick´y where morphophonemic merge rules and templates are involved.

- Morphophonemic templates capture morphological processes by just organizing stem patterns and generic affixes without any context-dependent variation of the affixes or ad hoc modification of the stems.
- The merge rules, indeed very neatly or effectively concise, then ensure that such structured representations can be converted into exactly the surface forms, both orthographic and phonological, used in the natural language.
- Applying the merge rules is independent of and irrespective of any grammatical parameters or information other than that contained in a template.
- Most morphological irregularities are thus successfully removed.

Ambiguity

- Morphological ambiguity is the possibility that word forms be understood in multiple ways out of the context of their discourse (communication in speech or writing).
- Words forms that look the same but have distinct functions or meaning are called homonyms.
- Ambiguity is present in all aspects of morphological processing and language processing at large.

- Table arranges homonyms on the basis of their behaviour with different endings.

Systematic homonyms arise as verbs combined with endings in Korean

(-ko)		(-e)		(-un)		Meaning
묻고	*mwut.ko*	묻어	*mwut.e*	묻은	*mwut.un*	'bury'
묻고	*mwut.ko*	물어	*mwul.e*	물은	*mwul.un*	'ask'
물고	*mwul.ko*	물어	*mwul.e*	문	*mwun*	'bite'
걷고	*ket.ko*	걷어	*ket.e*	걷은	*ket.un*	'roll up'
걷고	*ket.ko*	걸어	*kel.e*	걸은	*kel.un*	'walk'
걸고	*kel.ko*	걸어	*kel.e*	건	*ken*	'hang'
굽고	*kwup.ko*	굽어	*kwup.e*	굽은	*kwup.un*	'be bent'
굽고	*kwup.ko*	구워	*kwu.we*	구운	*kwu.wun*	'bake'
이르고	*i.lu.ko*	이르러	*i.lu.le*	이른	*i.lun*	'reach'
이르고	*i.lu.ko*	일러	*il.le*	이른	*i.lun*	'say'

- Arabic is a language of rich morphology, both derivational and inflectional.
- Because Arabic script usually does not encode short vowels and omits yet some other diacritical marks that would record the phonological form exactly, the degree of its morphological ambiguity is considerably increased.
- When inflected syntactic words are combined in an utterance, additional phonological and orthographic changes can take place, as shown in Figure.
- In Sanskrit, one such euphony rule is known as external *sandhi.*

dirāsatī	دراستي	drAsty	→	*dirāsatu ī*	دراسة ي	drAsp y
			→	*dirāsati ī*	دراسة ي	drAsp y
			→	*dirāsata ī*	دراسة ي	drAsp y
muʿallimīya	معلمي	mElmy	→	*muʿallimū ī*	معلمو ي	mElmw y
			→	*muʿallimī ī*	معلمي ي	mElmy y
katabtumūhā	كتبتموها	ktbtmwhA	→	*katabtum hā*	كتبتم ها	ktbtm hA
ʾiğrāʾuhu	إجراؤه	IjrAWh	→	*ʾiğrāʾu hu*	إجراء ه	IjrA’ h
ʾiğrāʾihi	إجرائه	IjrA}h	→	*ʾiğrāʾi hu*	إجراء ه	IjrA’ h
ʾiğrāʾahu	إجراءه	IjrA’h	→	*ʾiğrāʾa hu*	إجراء ه	IjrA’ h
li-ʾl-ʾasafi	للأسف	llOsf	→	*li ʾl-ʾasafi li*	ل الأسف	l AlOsf

- cases are expressed by the same word form with *dirāsatī* ‘my study’ and *muʿallimīya* ‘ my teachers’, but the original case endings are distinct.

Productivity

- Is the inventory of words in a language finite, or is it unlimited?
- This question leads directly to discerning two fundamental approaches to language, summarized in the distinction between *langue* and *parole*, or in the competence versus performance duality by Noam Chomsky.
- In one view, language can be seen as simply a collection of utterances (parole) actually pronounced or written (performance).
- This ideal data set can in practice be approximated by linguistic corpora, which are finite collections of linguistic data that are studied with empirical(based on) methods and can be used for comparison when linguistic models are developed.
- Yet, if we consider language as a system (langue), we discover in it structural devices like recursion, iteration, or compounding(make up; constitute)that allow to produce (competence) an infinite set of concrete linguistic utterances.
- This general potential holds for morphological processes as well and is called morphological productivity.
- We denote the set of word forms found in a corpus of a language as its vocabulary.

- The members of this set are word types, whereas every original instance of a word form is a word token.
- The distribution of words or other elements of language follows the “80/20 rule,” also known as the law of the vital few.
- It says that most of the word tokens in a given corpus can be identified with just a couple of word types in its vocabulary, and words from the rest of the vocabulary occur much less commonly if not rarely in the corpus.
- Furthermore, new, unexpected words will always appear as the collection of linguistic data is enlarged.
- In Czech, negation is a productive morphological operation. Verbs, nouns, adjectives, and adverbs can be prefixed with *ne-* to define the complementary lexical concept.

Morphological Models

- There are many possible approaches to designing and implementing morphological models.
- Over time, computational linguistics has witnessed the development of a number of formalisms and frameworks, in particular grammars of different kinds and expressive power, with which to address whole classes of problems in processing natural as well as formal languages.
- Let us now look at the most prominent types of computational approaches to morphology.

Dictionary Lookup

- Morphological parsing is a process by which word forms of a language are associated with corresponding linguistic descriptions.
- Morphological systems that specify these associations by merely enumerating(is **the act or process of making or stating a list of things one after another)** them case by case do not offer any generalization means.
- Likewise for systems in which analyzing a word form is reduced to looking it up verbatim in word lists, dictionaries, or databases, unless they are constructed by and kept in sync with more sophisticated models of the language.

- In this context, a dictionary is understood as a data structure that directly enables obtaining some precomputed results, in our case word analyses.
- The data structure can be optimized for efficient lookup, and the results can be shared. Lookup operations are relatively simple and usually quick.
- Dictionaries can be implemented, for instance, as lists, binary search trees, tries, hash tables, and so on.
- Because the set of associations between word forms and their desired descriptions is declared by plain enumeration, the coverage of the model is finite and the generative potential of the language is not exploited.
- Despite all that, an enumerative model is often sufficient for the given purpose, deals easily with exceptions, and can implement even complex morphology.
- For instance, dictionary-based *approaches* to Korean depend on a large dictionary of all possible combinations of allomorphs and morphological alternations.
- These approaches do not allow development of reusable morphological rules, though.

Finite-State Morphology

- By finite-state morphological models, we mean those in which the specifications written by human programmers are directly compiled into finite-state transducers.
- The two most popular tools supporting this approach, XFST (Xerox Finite-State Tool) and LexTools.
- Finite-state transducers are computational devices extending the power of finite-state automata.
- They consist of a finite set of nodes connected by directed edges labeled with pairs of input and output symbols.
- In such a network or graph, nodes are also called states, while edges are called arcs.
- Traversing the network from the set of initial states to the set of final states along the arcs is equivalent to reading the sequences of encountered input symbols and writing the sequences of corresponding output symbols.
- The set of possible sequences accepted by the transducer defines the input language; the set of possible sequences emitted by the transducer defines the output language.

Input	Input Morphological parsed output
Cats	cat +N +PL
Cat	cat +N +SG
Cities	city +N +PL
Geese	goose +N +PL
Goose	goose +N +SG) or (goose +V)
Gooses	goose +V +3SG
merging	merge +V +PRES-PART
Caught	(caught +V +PAST-PART) or (catch +V +PAST)

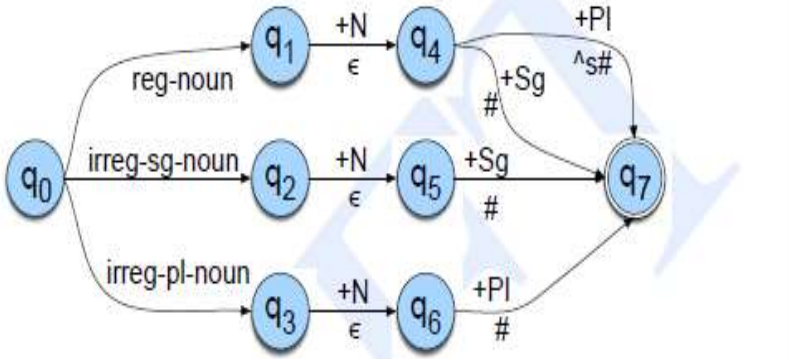


Figure 3.13 A schematic transducer for English nominal number inflection T_{num}. The symbols above each arc represent elements of the morphological parse in the lexical tape; the symbols below each arc represent the surface tape (or the intermediate tape, to be described later), using the morpheme-boundary symbol ^ and word-boundary marker #. The labels on the arcs leaving q_0 are schematic, and need to be expanded by individual words in the lexicon

- For example, a finite-state transducer could translate the infinite regular language consisting of the words *vnuk*, *pravnuk*, *prapravnuk*, *...* to the matching words in the infinite regular language defined by *grandson*, *great-grandson*, *great-great-grandson*.
- In finite-state computational morphology, it is common to refer to the input word forms as **surface strings** and to the output descriptions as **lexical strings**, if the transducer is used for morphological analysis, or vice versa, if it is used for morphological generation.

•In English, a finite-state transducer could analyze the surface string children into the lexical string child [+plural], for instance, or generate women from woman [+plural].

- Relations on languages can also be viewed as functions. Let us have a relation *R*, and let us denote by [Σ] the set of all sequences over some set of symbols Σ, so that the domain and the range of *R* are subsets of [Σ].
- We can then consider *R* as a function mapping an input string into a set of output strings, formally denoted by this type signature, where [Σ] equals *String*:

$$\mathcal{R} :: [\Sigma] \rightarrow \{[\Sigma]\} \qquad \mathcal{R} :: String \rightarrow \{String\} \qquad (1.1)$$

- A theoretical limitation of finite-state models of morphology is the problem of capturing **reduplication** of words or their elements (e.g., to express plurality) found in several human languages.
- Finite-state technology can be applied to the morphological modeling of isolating and agglutinative languages in a quite straightforward manner. Korean finite-state models are discussed by Kim, Lee and Rim, and Han, to mention a few.

Unification-Based Morphology

- The concepts and methods of these formalisms are often closely connected to those of logic programming.
- In finite-state morphological models, both surface and lexical forms are by themselves unstructured strings of atomic symbols.
- In higher-level approaches, linguistic information is expressed by more appropriate data structures that can include complex values or can be recursively nested if needed.
- Morphological parsing *P* thus associates linear forms φ with alternatives of structured content ψ, cf.

$$\mathcal{P} :: \phi \rightarrow \{\psi\} \qquad \mathcal{P} :: form \rightarrow \{content\} \qquad (1.2)$$

- Erjavec argues that for morphological modelling, word forms are best captured by regular expressions, while the linguistic content is best described through **typed feature structures**.
- Feature structures can be viewed as directed acyclic graphs.
- A node in a feature structure comprises a set of attributes whose values can be

- Nodes are associated with types, and atomic values are attributeless nodes distinguished by their type.
- Instead of unique instances of values everywhere, references can be used to establish value instance identity.
- Feature structures are usually displayed as attribute-value matrices or as nested symbolic expressions.
- Unification is the key operation by which feature structures can be merged into a more informative feature structure.
- Unification of feature structures can also fail, which means that the information in them is mutually incompatible.
- Morphological models of this kind are typically formulated as logic programs, and unification is used to solve the system of constraints imposed by the model.
- Advantages of this approach include better abstraction possibilities for developing a morphological grammar as well as elimination of redundant information from it.
- Unification-based models have been implemented for Russian, Czech, Slovene, Persian, Hebrew, Arabic, and other languages.

Functional Morphology

- Functional morphology defines its models using principles of functional programming and type theory.
- It treats morphological operations and processes as pure mathematical functions and organizes the linguistic as well as abstract elements of a model into distinct types of values and type classes.
- Though functional morphology is not limited to modelling particular types of morphologies in human languages, it is especially useful for fusional morphologies.
- Linguistic notions like paradigms, rules and exceptions, grammatical categories and parameters, lexemes, morphemes, and morphs can be represented intuitively(without conscious reasoning; instinctively) and succinctly(in a brief and clearly expressed manner) in this approach.
- Functional morphology implementations are intended to be reused as programming libraries capable of handling the complete morphology of a language and to be incorporated into various kinds of applications.

- Morphological parsing is just one usage of the system, the others being morphological generation, lexicon browsing, and so on.
- we can describe inflection *I*, derivation *D*, and lookup *L* as functions of these generic type

$$\mathcal{I} :: lexeme \rightarrow \{parameter\} \rightarrow \{form\} \quad (1.3)$$
$$\mathcal{D} :: lexeme \rightarrow \{parameter\} \rightarrow \{lexeme\} \quad (1.4)$$
$$\mathcal{L} :: content \rightarrow \{lexeme\} \quad (1.5)$$

- Many functional morphology implementations are embedded in a general-purpose programming language, which gives programmers more freedom with advanced programming techniques and allows them to develop full-featured, real-world applications for their models.
- The Zen toolkit for Sanskrit morphology is written in OCaml.
- It influenced the functional morphology framework in Haskell, with which morphologies of Latin, Swedish, Spanish, Urdu, and other languages have been implemented.
- In Haskell, in particular, developers can take advantage of its syntactic flexibility and design their own notation for the functional constructs that model the given problem.

- The notation then constitutes a so-called domain-specific embedded language, which makes programming even more fun.
- Even without the options provided by general-purpose programming languages, functional morphology models achieve high levels of abstraction.
- Morphological grammars in Grammatical Framework can be extended with descriptions of the syntax and semantics of a language.
- Grammatical Framework itself supports multilinguality, and models of more than a dozen languages are available in it as open-source software.