




# Module 1

Introduction to NLP – Ambiguous Nature of NLP – Morphological Analysis – Syntax Analysis – Semantic Analysis – Pragmatic Analysis – Discourse Analysis – Introduction to real-life applications of NLP – Introduction to Corpora – Corpora Analysis

# What is NLP

- Study of Interaction between Computers and Human Language
- NLP = Computer Science + AI + Computational Linguistics

## Common NLP Tasks

 Easy	 Medium	 Hard
<ul style="list-style-type: none"><li>• Chunking</li><li>• Part-of-Speech Tagging</li><li>• Named Entity Recognition</li><li>• Spam Detection</li><li>• Thesaurus</li></ul>	<ul style="list-style-type: none"><li>• Syntactic Parsing</li><li>• Word Sense Disambiguation</li><li>• Sentiment Analysis</li><li>• Topic Modeling</li><li>• Information Retrieval</li></ul>	<ul style="list-style-type: none"><li>• Machine Translation</li><li>• Text Generation</li><li>• Automatic Summarization</li><li>• Question Answering</li><li>• Conversational Interfaces</li></ul>

## More Complex Languages Than English

- **German:** Donaudampfschiffahrtsgesellschaftskapitän (5 “words”)
- **Chinese:** 50,000 different characters (2-3k to read a newspaper)
- **Japanese:** 3 writing systems
- **Thai:** Ambiguous word boundaries and sentence concepts
- **Slavic:** Different word forms depending on gender, case, tense

# Why is language processing difficult? - Ambiguity

Consider trying to build a system that would answer email sent by customers to a retailer selling laptops and accessories via the Internet. This might be expected to handle queries such as the following:

- Has my order number 4291 been shipped yet?
- Is FD5 compatible with a 505G?
- What is the speed of the 505G

Very similar strings can mean very different things, while very different strings can mean much the same thing. 1 and 2 below look very similar but mean something completely different, while 2 and 3 look very different but mean much the same thing.

1. How fast is the 505G?
2. How fast will my 505G arrive?
3. Please tell me when I can expect the 505G I ordered

# Ambiguity – Levels

## Different levels of Ambiguity

- **Morphological ambiguity**: ambiguity in parsing of a text into sentence/word/sub words. e.g. Mr. Modi is meeting with industrialist to generate jobs in machine-learning and AI field.
- **Lexical ambiguity**: It is at very primitive level such as word-level. For example, treating the word board as noun or verb? Solution: POS tagging Word Sense Disambiguation
- **Syntax Level ambiguity**: A sentence can be parsed in different ways. For example, “He lifted the beetle with red cap.” - Did he use cap to lift the beetle or he lifted a beetle that had red cap? Solution: probabilistic parsing
- **Semantic Level Ambiguity**: This occurs when the meaning of the words themselves can be misinterpreted. Even after the syntax and the meanings of the individual words have been resolved, there are two ways of interpreting the sentence. Consider the example, “Seema loves her mother and Sriya does too”.  $\Rightarrow$  Sriya loves whose mother? - her own or Seema’s mother.

# Ambiguity – Levels

## Different levels of Ambiguity

- **Discourse level ambiguity:** where the interpretation is ambiguous by the virtue of context (previous word/sentence/paragraph). e.g. Referential Ambiguity.
  - Referring to something using pronouns. For example, Rima went to Gauri. She said, “I am tired.”  $\Rightarrow$  Exactly who is tired? Solution: Co-reference resolution. Referential ambiguity is also known as Anaphoric Ambiguity.
- **Pragmatic level ambiguity:** Pragmatic ambiguity refers to a situation where the context of a phrase gives it multiple interpretation and it require real world knowledge for correct interpretation.
  - Pragmatic Ambiguity occurs when context does not provide enough information to clarify the statement.
  - The problem involves processing user intention, sentiment, belief, world, etc.- all of which are highly complex tasks.

# Lexical Ambiguities - 1

## 1. Word as different Parts of Speech(PoS)

- board as noun or verb?
- book as noun or verb?

I board on the flight.  $\Rightarrow$  Verb

The board has discussed the finances of 2022.  $\Rightarrow$  Noun

I book a flight ticket.  $\Rightarrow$  Verb

I am reading a book.  $\Rightarrow$  Noun.

**Solution: Parts-of-speech tagging.**

# Lexical Ambiguities - 2

## 2. **Word with different sense**

make as to create or engage?

- make love not war. implies do or engage in.
- make a mess in one's office. implies create.
- make a mistake implies carry out or commit.

**Solution: Word Sense Disambiguation (WSD) eg. Wordnet.**



# Syntax Level Ambiguities

- “He lifted the Beetle with Red Cap”  
Did he lift a beetle that had red cap?



Did he use cap to lift the beetle?

**Solution : Probabilistic Parsing**

# Identify the Ambiguities in the Sentences

- The Car hit the pole while it was moving
- Seema loves her mother and Sriya does too
- The man saw the man with the telescope
- Buy Books for Children
- Old men and Women were taken to safe locations
- Every man loves a woman
- Violonist Linked to JAL Crash Blossoms
- Teacher Strikes Idle Kids
- Red Tape Holds up New Bridges

# Other Difficulties in NLP

## non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

## segmentation issues

the New York-New Haven Railroad  
the New York-New Haven Railroad

## idioms

dark horse  
get cold feet  
lose face  
throw in the towel

## neologisms

unfriend  
Retweet  
bromance

## world knowledge

Mary and Sue are sisters.  
Mary and Sue are mothers.

## tricky entity names

Where is *A Bug's Life* playing .  
*Let It Be* was recorded ...  
... a mutation on the *for* gene

# Generic System Components

- **Input preprocessing** - Speech recognizer or text preprocessor (non-trivial in languages like Chinese or for highly structured text for any language) or gesture recognizer. Such system might themselves be very complex.
- **Morphological analysis** - Structure of words – Difficult for Turkish, Basque.
  - In English, “unusually” can be thought of as composed of a prefix un-, a stem usual, and an affix -ly
- **Part of Speech tagging** - Not an essential part of most deep processing systems, but sometimes used as a way of cutting down parser search space.
  - Example - “permit us.”
  - PoS Tagging - [('permit', 'VB'), ('us', 'PRP')]

# Generic System Components

- **Parsing** - Includes syntax and compositional semantics, which are sometimes treated as separate components
  - Syntax – Checking formal and Computational aspects
  - Compositional semantics is the construction of meaning based on syntax
- **Disambiguation** - Done as part of parsing, or (partially) left to a later phase
  - Example - For instance, consider the word “bank.” It can mean a financial institution (“I deposited money in the bank”) or the side of a river (“The boat went down the river bank”). NLP systems use techniques like Word Sense Disambiguation (WSD)

# Generic System Components

- **Context module** - Maintains information about the context, for anaphora resolution, for instance.
  - **Named Entity Recognition (NER)**
    - “Microsoft Corporation was founded by Bill Gates and Paul Allen in 1975,” NER would recognize “Microsoft Corporation,” “Bill Gates,” and “Paul Allen” as named entities.
  - **Dependency Parsing**
    - Dependency parsing analyzes the grammatical structure of sentences.
    - “The cat chased the mouse,” dependency parsing reveals that “chased” depends on “cat” and “mouse” depends on “chased”.
  - **Coreference Resolution**
    - Coreference resolution links pronouns (like “he,” “she,” “it”) to their corresponding nouns.
    - For example, in “John loves his dog. He takes it for walks,” coreference resolution connects “he” to “John” and “it” to “dog”.
  - **Sentiment Analysis**
    - Sentiment analysis determines the emotional tone of text (positive, negative, neutral).
    - For instance, “The movie was fantastic!” would be classified as positive sentiment.

# Generic System Components

- **Text planning** - The part of language generation that's concerned with deciding what meaning to convey
- **Tactical generation** - Converts meaning representations to strings.
  - The process of deciding how to convey a specific meaning in a natural language utterance. It focuses on the strategic choices made during language generation, such as selecting appropriate words, structures, and expressions to effectively communicate a message.
- **Morphological generation** - As with morphological analysis, this is relatively straightforward for English.
  - Example : Run -> Runs, ran, running  
Happy -> Unhappy
- **Output processing** - Text-to-speech, Text formatter, etc. As with input processing, this may be complex.
  - NLP Models

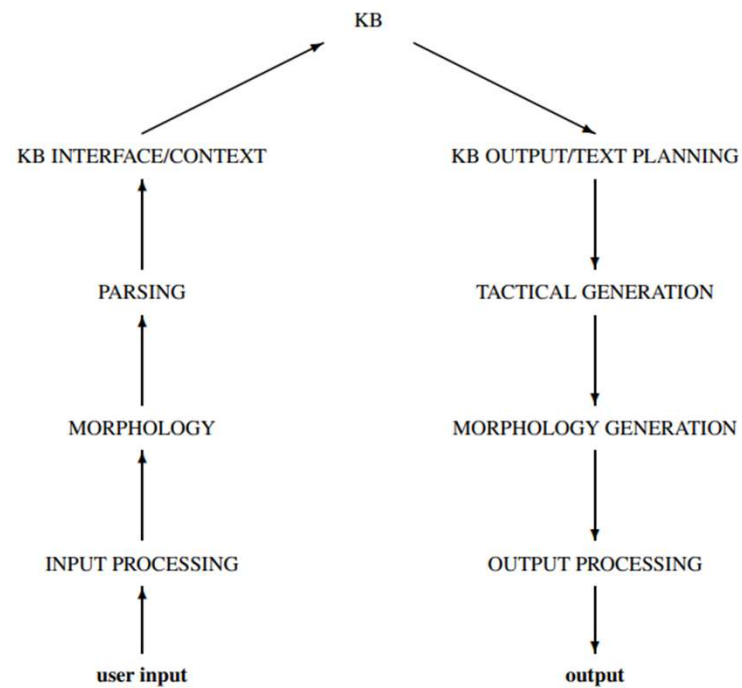


Figure – Generic System Components



# Text Pre-Processing

- Text preprocessing
  - Converting a raw text file - > sequence of digital bits, into a well-defined sequence of linguistically meaningful units
  - At the lowest level characters representing the individual graphemes in a language's written system, words consisting of one or more characters, and sentences consisting of one or more words.
  - Text preprocessing can be divided into two stages:
    - Document triage
    - Text segmentation

# Challenges in Text Preprocessing

- Depends on the underlying Language system
- Writing systems - > logographic, syllabic and alphabetic
  - Example -> English is a alphabetic language system
- Character Encoding Identification and Its Impact on Tokenization
  - Example -> English coded in 8-Bit Encoding of Latin-1
- Language Dependence
  - In English, “ , .” marks word and sentence boundaries
  - Need to have a generic system that applies to different writing systems
- Corpus Dependence
  - Robust text segmentation algorithms designed for use with a large corpora must have the capability to handle the range of irregularities (Web pages)

# Text Processing – Document Triage

- Document triage is the process of converting a set of digital files into well-defined text documents.
- Early - > Current Corpora
- Steps
  - Character encoding identification
    - To know 1/more bytes to form the character
  - Language identification
    - To apply language specific algorithms
  - Text sectioning
    - Identify actual content within a file -> discarding undesirable contents

# Text Processing – Text Segmentation

- Text segmentation is the process of converting a well-defined text corpus into its component words and sentences.
- Steps
  - Word Segmentation / Tokenization
    - Locating word boundaries
  - Text Normalization
    - involves merging different written forms of a token into a canonical normalized form.
    - “Mr.”, “Mr”, “mister”, and “Mister” to a normalized form.
  - Sentence Segmentation
    - Locating sentence Boundaries - > Longer processing units
    - Most written languages have punctuation marks that occur at sentence boundaries
    - Also called as , Sentence boundary Detection, Sentence boundary disambiguation, Sentence boundary recognition.

# Tokenization

- Factors affecting Tokenization
  - Space-Delimited / Unsegmented Languages
  - Word Structure Types - > isolating, agglutinating, Inflectional
    - Isolating, where words do not divide into smaller units;
    - Agglutinating (or agglutinative), where words divide into smaller units (morphemes) with clear boundaries between the morphemes;
      - Consider the word “unapproachable.” We can break it down into:
        - “un-” (meaning “not”)
        - “approach” (the base verb)
        - “-able” (forming an adjective)
    - Inflectional, where the boundaries between morphemes are not clear and where the component morphemes can express more than one grammatical meaning.
      - Big - > Bigger, biggest
    - Examples : Mandarin Chinese is predominantly isolating, Japanese is strongly agglutinative, and Latin is largely inflectional

# Sentence Segmentation

- Sentence Boundary Punctuation
- use of certain punctuation marks to separate sentences

Here is a sentence. Here is another.

Here is a sentence; here is another.
- Contextual factors assisting sentence segmentation
  - Case distinctions, Part of speech, Word Length, Lexical endings, Prefixes and suffixes, Abbreviation classes, Internal punctuation, and proper nouns.

# Morphological Analysis

- Morphology concerns the structure of words
- Words are assumed to be made up of morphemes
- Inflectional vs derivational morphology
  - Inflectional morphology concerns properties such as tense, aspect, number, person, gender, and case
  - Derivational affixes, such as un-, re-, anti- etc, have a broader range of semantic possibilities

# Morphological Analysis

- The problem of recognizing that *foxes* breaks down into the two morphemes *fox* and *-es* is called ***morphological parsing***.
- Similar problem in the information retrieval domain: ***stemming***
- Given the **surface** or **input form** *going*, we might want to produce the parsed form: VERB-go + GERUND-ing
- In this
  - morphological knowledge and
  - The **finite-state transducer**
- It is quite inefficient to list all forms of noun and verb in the dictionary because the productivity of the forms.
- Morphological parsing is necessary more than just IR, but also
  - Machine translation
  - Spelling checking



# Morphological Analysis

- Morphology is the study of the way words are built up from smaller meaning-bearing units, **morphemes**.
- Two broad classes of morphemes:
  - **The stems:** the “main” morpheme of the word, supplying the main meaning, while
  - **The affixes:** add “additional” meaning of various kinds.
- Affixes are further divided into **prefixes, suffixes, infixes, and circumfixes**.
  - Suffix: *eat-s*
  - Prefix: *un-buckle*
  - Circumfix: *ge-sag-t* (said) *sagen* (to say) (in German)
  - Infix: *hingi* (borrow) *humingi* (the agent of an action) )in Philippine language Tagalog)

# Morphological Analysis

Two broad classes of ways to form words from morphemes:

- **Inflection:** 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, and
- **Derivation:** 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.

- Verbal inflection is more complicated than nominal inflection.
  - English has three kinds of verbs:
    - **Main verbs**, *eat, sleep, impeach*
    - **Modal verbs**, *can will, should*
    - **Primary verbs**, *be, have, do*
  - Morphological forms of regular verbs

stem	walk	merge	try	map
-s form	walks	merges	tries	maps
-ing principle	walking	merging	trying	mapping
Past form or <i>-ed</i> participle	walked	merged	tried	mapped

- These regular verbs and forms are significant in the morphology of English because of their *majority* and being *productive*.

– Morphological forms of irregular verbs

stem	eat	catch	cut
-s form	eats	catches	cuts
-ing principle	eating	catching	cutting
Past form	ate	caught	cut
-ed participle	eaten	caught	cut

- **Nominalization** in English:

- The formation of new nouns, often from verbs or adjectives

Suffix	Base Verb/Adjective	Derived Noun
-action	computerize (V)	computerization
-ee	appoint (V)	appointee
-er	kill (V)	killer
-ness	fuzzy (A)	fuzziness

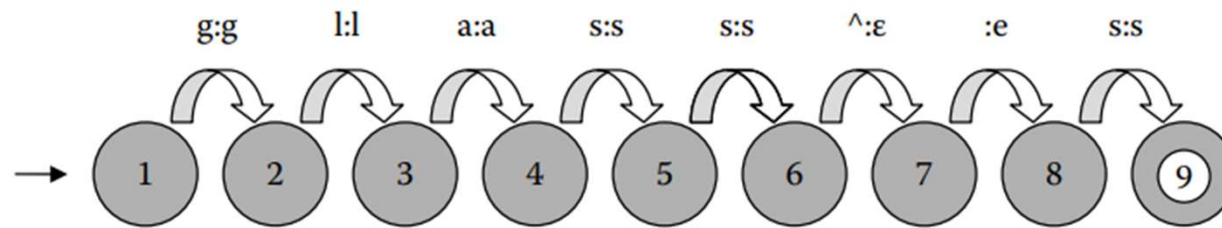
- Adjectives derived from nouns or verbs

Suffix	Base Noun/Verb	Derived Adjective
-al	computation (N)	computational
-able	embrace (V)	embraceable
-less	clue (A)	clueless

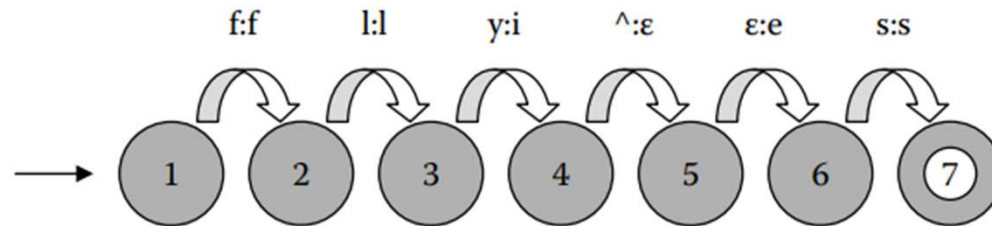


# Finite-State Morphological Parsing

- We need at least the following to build a morphological parser:
  1. **Lexicon**: the list of stems and affixes, together with basic information about them (Noun stem or Verb stem, etc.)
  2. **Morphotactics**: the model of morpheme ordering that explains which classes of morphemes can follow other classes of morphemes inside a word. E.g., the rule that English plural morpheme follows the noun rather than preceding it.
  3. **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 changes *city* + *-s* to *cities*).



A spelling rule FST for *glasses*.



A spelling rule FST for *flies*.

# Syntax Analysis

- Syntactic Analysis – Check the grammar, arrangement of words, and the relationship between words.
- Parts of Speech(PoS) Tagging, Dependency Parser, Grammar Checking
- Incorrect Syntax: Rise in sun the east.
- Correct Syntax: Sun rises in the east.



# Semantic Analysis

- Understanding the utterance (Analysing the meaning of the word)
- Depends on the result of parsing, lexical information
- Semantic Representation
  - Formal vs Cognitive
    - Formal - Capturing the meaning of language using structured methods.
      - Example : A word can be represented as a vector , logical rules defined to explain relation
    - Cognitive - understanding and generating human language by connecting it to real-world experiences and concepts.
      - Example : “cat” might be connected to “animal,” “pet,” and “mammal.”
  - Compositional vs Lexical
    - Lexical - This involves understanding the meaning of individual words in a text. It’s like fetching the dictionary definition for each word.
    - Compositional - Analyzes how combinations of words form the context and convey specific meanings.

# Pragmatic Analysis

- Focuses on interpreting the inferred meaning of a text beyond its literal content.
- It delves into how context influences meaning, including how statements are perceived in various contexts.
- It is a complex phase where machines should have knowledge not only about the provided text but also about the real world.
- There can be multiple scenarios where the intent of a sentence can be misunderstood if the machine doesn't have real world knowledge.

"Thank you for coming so late, we have wrapped up the meeting" (Contains sarcasm).

"Can you share your screen?" (here the context is about computer's screen share during a remote meeting).

# Models for NLP

- **State Machines:** Deterministic and non-deterministic finite state automata, finite state transducers, weighted automata. e.g. Markov models and Hidden Markov models which combines States machines with probabilistic model.
- **Formal rule systems:** regular grammars, regular relations, Context Free Grammars (CFG), feature-augmented grammars. It is used while dealing with phonology, morphology and syntax. e.g. Compiler to check the syntax.
- **Logic-based models:** first order logic, predicate calculus. It helps in dealing with semantics, pragmatics, and discourse.

# Models for NLP

- **Probabilistic Models:** helps to resolve ambiguities. It is one class of Machine learning based models. The algorithm involves both state-machine and formal rule system and makes use of a search space representing hypothesis about input. Dynamic Programming is used to optimize this search process.
- **Deep-Learning based models:** word vectors(word2vec, GLoVE, BERT, fastText), LSTM, BiLSTM, Transformer-based models, BERT, Large Language Models (LLM)

# Corpora

- Machine readable authentic text
- Transcripts of spoken data

## Need for Corpus

- A corpus is made for the study of language in a broad sense
  - To test existing linguistic theory and hypotheses
  - To generate and verify new linguistic hypotheses
  - Beyond linguistics, to provide textual evidence in text-based humanities and social sciences subjects
- The purpose is reflected in a well-designed corpus

# Why Corpora

- Two main reasons:
  - To evaluate our systems on
    - Good science requires controlled experimentation.
    - Good engineering requires benchmarks.
  - To help our systems work well (data-driven methods/machine learning)
    - When a system's behavior is determined solely by manual rules or databases, it is said to be rule-based, symbolic, or knowledge-driven (early days of computational linguistics)
    - Learning: collecting statistics or patterns automatically from corpora to govern the system's behavior (dominant in most areas of contemporary NLP)
      - – supervised learning: the data provides example input/output pairs (main focus in this course)
      - – core behavior: training; refining behavior: tuning

# Corpora

## Benefits of Corpus

- Corpus data is more reliable
  - A corpus pools together linguistic intuitions of a range of language speakers, which offsets the potential biases in intuitions of individual speakers.
- Corpus data is more natural.
  - It is used in real communications instead of being invented specifically for linguistic analysis.
- Corpus data is contextualized
  - Attested language use which has already occurred in real linguistic context
- Corpus data is quantitative
  - Corpora can provide frequencies and statistics readily 5 Corpus data can find differences that intuitions alone cannot perceive.
  - Ex: synonyms: totally, absolutely, utterly, completely, entirely

# Corpora

- **Features to be considered while designing**
  - Support for multiple languages
  - Support for variations of a language
    - AAE (African American English)
    - AAVE(African American Vernacular English)
  - Code switching with translation and transliteration
  - Associate a Datasheet to the corpus
    - Motivation – why / whom / who funded
    - Situation – Spoken conversation, Edited text, Social media communication, Monologue / Dialogue
    - Speaker Demographics – Age / Gender of the author
    - Collection Process – How big? , Data collected with consent, Data preprocessed? Metadata information?
    - Markup and Annotation Process - Demographics of Annotation
    - Distribution – Copyright and IP restrictions



## Some of English Corpora

- News on Web(NOW)
- iWeb(Intelligent Web based Corpus)
- GloWbE(Global Web Based English)
- Corpus of Contemporary American English(COCA)

# Benefits of Corpus Data

- Corpus data is **more reliable**
  - A corpus pools together linguistic intuitions of a range of language speakers, which offsets the potential biases in intuitions of individual speakers.
- Corpus data is **more natural**.
  - It is used in real communications instead of being invented specifically for linguistic analysis.
- Corpus data is **contextualized**
  - Attested language use which has already occurred in real linguistic context.
- Corpus data is **quantitative**
  - Corpora can provide frequencies and statistics readily
- Corpus data **can find differences** that intuitions alone cannot perceive.
  - e.g. synonyms: totally, absolutely, utterly, completely, entirely

# Why use Corpus Data

- A corpus can be more **comprehensive and balanced**
  - Even expert speakers have only a partial knowledge of a language
- A corpus can show us what is **common and typical**.
  - Even expert speakers tend to notice the unusual and think of what is possible
- A corpus can readily give us **accurate statistics**
  - Even expert speakers cannot quantify their knowledge of language
- A corpus can **store and recall** all the information that has been stored in it.
  - Even expert speakers cannot remember everything they know.
- A corpus can provide us with a **vast number of examples** in real communication context.
  - Even experts speakers cannot make up vast number of natural examples.
- A corpus can give you more **objective evidence**
  - Even expert speakers have prejudices and preferences and every language has cultural connotations and underlying ideology

# Why use Corpus Data

- A corpus can be made **permanently accessible** to all.
  - Even expert speakers are not always available to be consulted.
- A constantly updated corpus can **reflect even recent changes** in the language.
  - Even expert speakers cannot keep up with language change
- A corpus can **encompass the actual language** use of many expert speakers
  - Even expert speakers lack authority: they can be challenged by other expert speakers

# References

- Daniel Jurafsky and James H Martin “Speech and Language Processing”, Prentice Hall 2017.
- Nitin Indurkha, Fred J Damerau , “ Handbook of Natural language Processing”