# .ch: 6

## * Introduction:

- A corpus is defined as "a collection of machine readable authentic texts including transcripts of spoken data that is sampled to be representative of a particular natural Language or language variety though "representativness" is a fluid concept."
- Corpora are crucial for NLP research and linguistic investigations.

- They are used to build & test NLP systems and contribute significantly to advancements in corpus annotation (eg: part of speach tagging, syntatic parsing).

- parallel corpora, consisting of text & its translations are essential for cross_ linguistic studies.

## * Corpus size:

- Corpus size is determined by research purpose & practical limitation such as:
  - Availability of studies data in machine - readable format.
  - Copy right restrictions.
- Small, specialized corpora can be sufficient for studying frequent linguistic features or specific domain.
- large corpora are beneficial for statistical modeling in NLP and languge engineering.

" The optimum size of corpora is determined by the research question the corpus is intended to address as well as practical considerations"

# * Balance, Representativeness & Sampling:

- Representativeness aims to capture the full range of variability in a language variety. It allows generalizing findings from the corpus to the border language.
- Two Key factors influence Represintativeness:
  - Balance: The range of genres, domains, and media included in the corpus.
  - Sampling: The method used to select text chuncks for each genre.
- Static Sample Corpora: Provide a snapshot of language at a specific time.
- Monitor Corpora, Continuously updated to track language changes over time.

- Different Corpus types have distinct criteria for representativeness:
    - General Corpora: Aim for broad representation of various text types.
    - Specialized corpora: focus on a specific domain or genre, with representativeness measured by the degree of closure or saturation (ie, how much of the vocabulary is captured).

## * Sampling Techniques:

- Defining the sampling unit (eg: book, periodical) & constructing a sampling frame (list of all units) are crucial for achieving representativeness.
- **Simple random Sampling**: Every unit has an equal chance of selection.
- **Stratified random sampling**: Dividing the population into homogenous groups (strata) and then randomly sampling each stratum ensures better representation of rare elements.

- The choice of sampling full texts rather than chunks depends on the research question & practical considerations like copyright.

## * Corpus MarKup & Annotation:

- **Corpus MarKup**: Adding codes to provide info about the text (meta data) & its structure.
- **Corpus Annotation**: Adding Interpretative linguistic info (eg: part of speach, tagging, syntatic analysis, semantic info).
- Reasons for markup annotations:
  - Preserving info about the original text.
  - Enhancing the corpus's value & enabling a wider range of research questions.
  - Facilitating linguistic analysis & info extraction.

## * Multilingual Corpora:

- Cover more than one language & are crucial for translation & contrastive studies.

- offers insights into language-specific, typological & cultural differences.
- Essential for developing NLP apps like computer-aided translation and multilangual info retrieval.
- Can be undirectional, bidirectional, or multi directional depending on the translation direction(s) included.

## * Multimodal Corpora:

- Extends beyond text to include other modes of communication like facial expressions, gestures, and pasture.
- Capture a more holistic view of spoken language, going beyond transcriptions.
- Though still in thier infancy, advancments in technology are facilitating thier development & use.

## * Conclusion:

- Corpus creation involves carful consideration of various factors, including size, balance, representativeness and the chosen sampling technique.

- Markup & annotation enrich the corpus and enable wider range of research possibilities.
- Multilangual & Multimodal corpora are emerging areas that are expanding the scope and potential of corpus linguistics.

# .ch:7

## * what is corpus Annotation ?

Corpus annotation enriches language data (text, audio, video) by adding layers of linguistic info. This can range from basic metadata (author, date) to complex linguistic features like:

- **Morphosyntactic layer**: Addressing ambiguities related to parts of speach, inflection, and morphology
- **Syntactic layer**: Captures syntactic relations, including constituent boundries, grammatical functions, and dependencies.

- **Semantic and discourse layer** : focuses on word sense disambiguation, anaphoric relations, discourse relations, and more.

## * The importance of Tree banks:

Tree banks are corpora annotated with structural info, representing syntactic, semantic, and sometim even inter-sentential relationships. The structure resembles a "tree" in graph theory with edges typically representing syntactic relations.

## * Benifits of Tree banks:

- **Enhanced linguistic Queries**: Tree banks allow for more nuanced and specific linguistic inquiries compared to raw text. For instance, researching specific grammatical constructions (eg, subject inversion) becomes feasible.

"lemmatized tagged texts are thus helpful but inquiries about subject inversion or agentless

passives are impossible to perform on corpora tagged only with part-of-speech info"

- **Linguistic Hypothesis Testing**: Tree banks provide the empirical data needed to test and validate linguistic theories, aiding in the refinement of descriptive frameworks.

"The confrontation of linguistics hypotheses with actual data leads also to checking & enriching the chosen descriptive frameworks."

- **Psycholinguistic Research**: The frequency of a specific constructions in treebanks can provide insights into psycholinguistic preferences.

"One can also check psycholinguistic preferences, in the sense that a highly frequent construction can be claimed to be preferred over a less frequent one."

## * The Penn Treebank : A Cornerstone :

The penn treebank, a widely recognized & utilized resource, offers annotated American English text with part-of-speach tags & skeletal syntactic structure. It exemplifies the value of combining automatic annotation with manual correction for achieving high-quality annotated corpora.

## * Linking Annotation and Linguistic Theory :

The choice of annotation schema is crucial & should be rooted in a well-defined linguistic theory. This ensures consistency and allows for the evaluation of different theoretical approaches against real-world data.

## * Annotation schemas :

- Constituency-based: Structures sentances into hierarchical constituents (eg: NP, VP). offers

readability and aligns with common grammatical understanding, but can introduce unnecessary complexity

- **Dependency-based**: focuses on direct grammatical relationships between words (head-dependent rlt) More flexible and captures grammatical functions but requires consistent criteria for head identification.

- **Hybrid**: Combines constituents with dependency relations, often using minimal phrases like "chunks" or "bunsetsu"

The excerpt cites examples of treebanks based on various theories, including HPSG (Head-driven Phrase Structure Grammar) and dependency-based approaches like the Prague Dependency Treebank (PDT) for Czech.

## * Impact of Treebanks on NLP :

Treebanks have fueled significant advancement in NLP technologies including :

- **Part-of-speach togging**: Achieving near-human performance levels.

- **parsing**: Analyzing sentance structure.

- **Deep parsing**: Extracting deeper linguistic info.

- **Semantic role labeling**: identifying the roles of words in a sentance.

- **Machine Translation**: Automating translation between languages.

## * Future Directions:

Ongoing research and experimentation in NLP will continue to shape future treebank annotation projects. New annotation schemes will likely emerge, driven by both technological needs and evolving linguistic theories, ultimately leading to more sophisticated and insightful language processing capabilities.

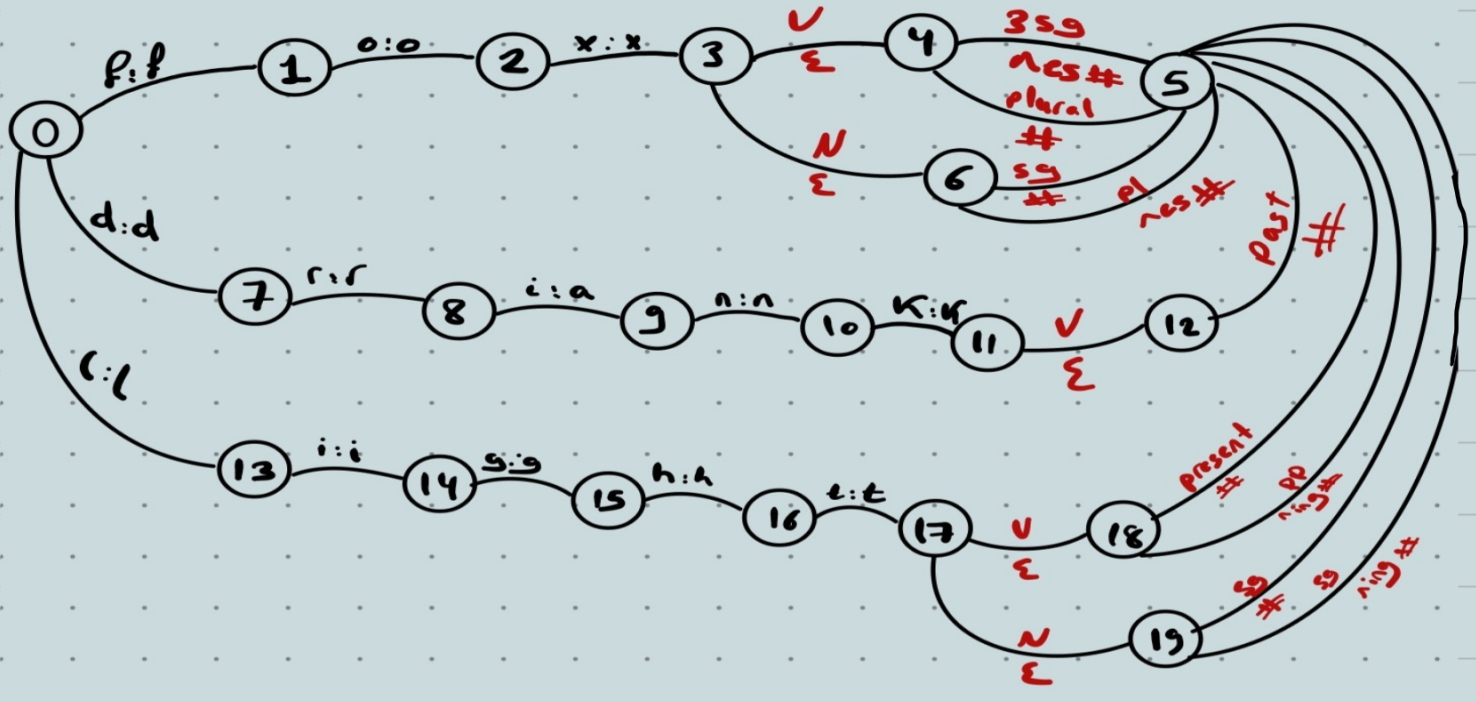
## * Conclusion:

Treebanks have become indispensable resources for NLP research and development. They enable the empirical study of language, provide training data for algorithms, and facilitate the development of advanced language technologies. The continued evolution of treebank annotation promises further breakthroughs in our understanding and processing of human language.

# FST EXAMPLE

Draw the Finite State Transducer (FST) (one graph) for the following set of words

[Fox, foxes, light, drank, lightning]



# . Ch: 8

## * Introduction

Pos tagging is a fundemental step in NLP, where each word in a sentance is labeled with its grammatical category (noun, verb, adjective, etc). It acts as a simplified form of morphological analysis and provides crucial info for subsequent NLP tasks like parsing and semantic analysis.

## * Two main challenges exist :

1. Ambiguous words : Many words can have multiple possible pos tags depending on

context. The excerpt exemplifies this with the word "can" which can function as an auxiliary verb, a main verb, or a noun.

"We can the can" ⟶ The three occurrences of the word can correspond to auxiliary, verb, and a noun categories respectively.

2. Unknown words: Words not encountered during training, pose a problem for both rule based and statistical tagging systems. Addressing unknown words is a vital for a trigger's practicality and robustness.

## * General Framework

POS taggers typically rely on the context of surrounding words and thier tags to determine

the correct tag for the target word. while long-distance dependencies can impact tagging, most approaches focus on a limited context for computational efficiency. For unknown words, morphological info like prefixes, suffixes, and capitalization is often utilized.

## * POS tagging approaches:

### 1. Rule-Based Approach:

- Early system relied on manually crafted linguistic rule, which proved laborious and lacked robustness.
- Transformation-Based learning (TBL): A significant advancement where the system learns correction rules from an initially tagged corpus. It

iteratively identifies and applies rules that minimize tagging errors. TBL demonstrates high accuracy and offers advantages over some stochastic approaches.

- Strengths: learns from data, adaptable, handles unknown words using morphological rules

- Limitations: Requires a large annotated corpus for training.

## 2. Statistical / Machine learning Approaches:

- These methods leverage statistical models trained on large datasets to predict the pos tags.

- Hidden Markov Models (HMMs):

Probabilistic models that consider tag transition probabilities (likelihood of a tag given the previous tag) and word likehood probabilities (probability of a word given its tag).

- **Strengths**: Effective for capturing sequential info, can handle unseen word sequences
- **Limitations**: Strong independence assumptions, limited context utilization.

- **Maximum Entropy (ME) Models**: Provides greater flexibility in incorporating contextual info, utilizing overlapping and interdependent features.
  - **Strengths**: More powerful context modeling, improved accuracy.

- **Limitations** : can be computationally expensive to train.

## * Conclusion :

POS tagging is a crucial step in NLP, laying the foundation for higher-level tasks like parsing & semantic Analysis. While rule-Based methods has eveloved significantly, statistical approaches, particularly HMMS, and ME models, have gained prominences due to thier data-driven nature and ability to leverage contextual info effectively. Ongoing research focuses on further enhancing accuracy and efficiency, especially for handling unKnown words and complex linguistic phenomena.

# . ch: 9

## * what are MWEs:

MWEs are expressions consisting of multiple words that function as a single lexical unit, often exhibiting unpredictable semantic or syntactic properties. Example include "take advantage", "serial number", and "by and large" while individual words like "top" and "dog" have clear meanings, thier combination in "top dog" yield a novel meaning not directly derivable from the components.

## * Significance of MWEs,

- MWEs are abundant in language, potentially rivaling the number of single words in a speaker's

lexicon.

- They contribute to language's flexibility & efficiency, enabling nuanced expression and lexical expansion.
- They pose challenges for NLP tasks like machine translation and semantic tagging, requiring specialized handling for accurate interpretation.

## * Linguistic Properties for MWEs :

MWEs are characterized by various forms of idiomaticity :

1. Lexical Idiomaticity,

   MWE components might not be stand-alone words, like "ad hoc" where neither "ad" nor "hoc" exist independently in English.

*"Lexical idiomaticity inevitably results in syntactic and semantic idiomaticity because there is no lexical knowledge associated directly with the parts from which to predict the behavior of the MWE."*

## 2. Syntactic Idiomaticity:

The MWE's syntax deviates from expected rules based on its components.

"By and large" is adverbial despite being a proposition and adjective combination.

*"Syntactic idiomaticity occurs when the syntax of the MWE is not derived directly from that of its components."*

## 3. Semantic Idiomaticity:

MWE meaning is not a straightforward composition of its parts. "Middle of the road" signifies non-extremism, not directly inferable from "middle" or "road".

*"Semantic idiomaticity is the property of the meaning of a MWE not being explicitly derivable from its parts."*

4. **Pragmatic Idiomaticity:**

MWE usage is tied to specific contexts. "Good morning" is a greetic solely for mornings.

*"Pragmatic idiomaticity is the condition of a MWE being associated with a fixed set of situations or a particular [context]."*

5. **Statistical Idiomaticity:**

Certian word combinations occur disproportionately frequently. "Black and white television" exhibits strong statistical idiomaticity compared to the less common "white and black television".

*"Statistical idiomaticity occurs when a particular combination of words occurs with markedly high frequency, relative to the component words or alternative phrasings of the same expression."*

## * Types of MWEs :

- **Nominal MWEs**: Noun compounds like "golf club" or "computer science department" including compound nominalizations like "investor hesitation".
- **Verbal MWEs**: includes:
  - Verb-particle constructions (VPCs) like "play around"
  - prepositional verbs (PVs) like "come across" and "refer to"
  - light-verb constructions (LVCs) like "take a walk" and "have a rest"
  - Verb-noun idiomatic combinations (VNICs) like "Kick the bucket"

- **Prepositional MWEs:**
  - Determinerless prepositional phrases "PP_Ds" like "on top" and "by car"
  - Complex prepositions like "on topof" and "in addition to".

## * MWE classification:

MWE can be broadly classified into:

- **Lexical phrases**: Exhibit lexical, syntactic, semantic, or pragmatic idiomaticity

  further divides into:

  - Fixed expressions: "at first"
  - semi-fixed expressions: "spill the beans"
  - Syntactically flexible expressions: "hand in the paper" vs "hand the paper in"

- **Institutionalized phrases**: primarily marked by statistical idiomaticity, like "traffic light".

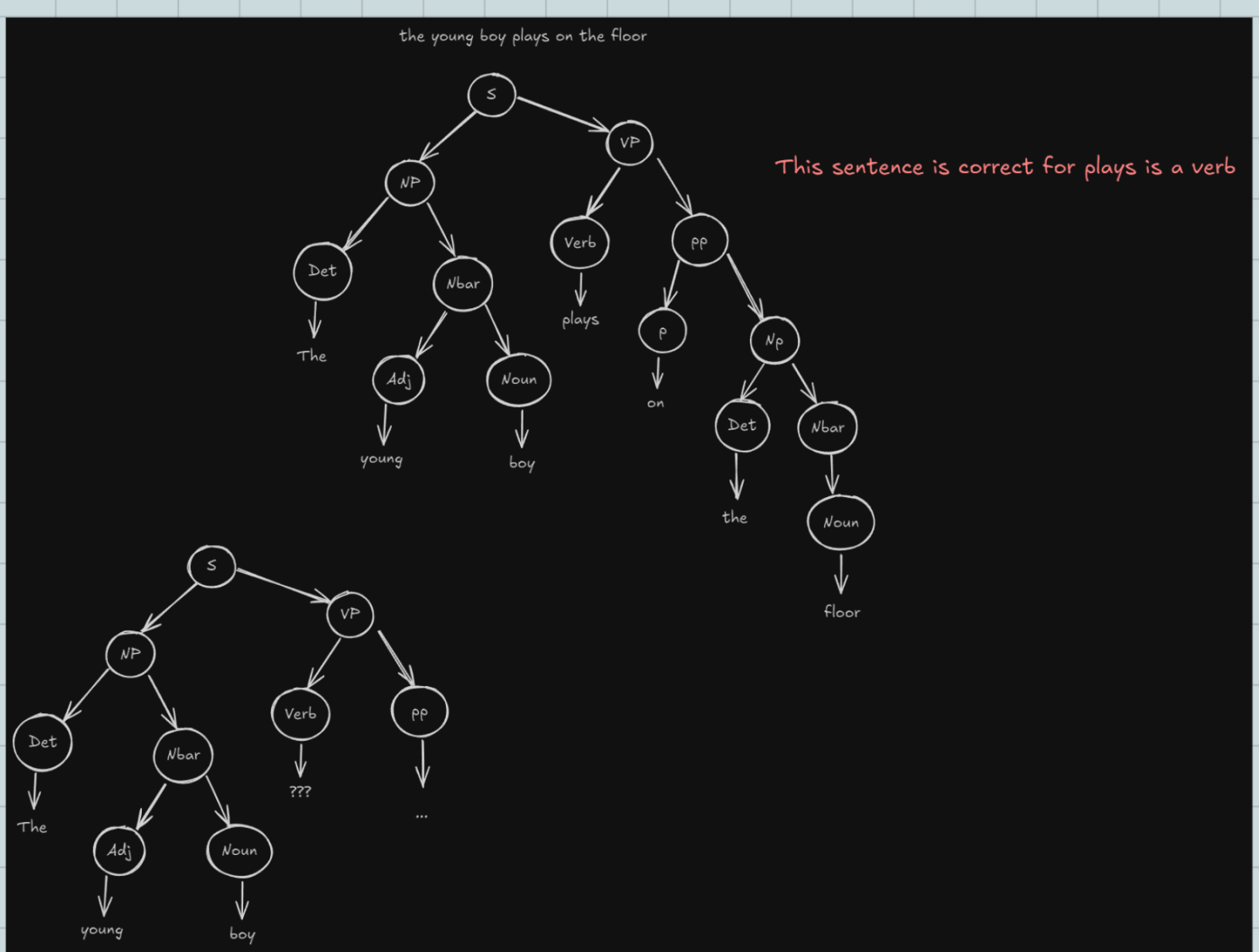
## * Challenges and application:

- Understanding MWEs requires tackling their ambiguity and underspecified semantics, especially in nominal compounds like "nut tree", "clothes tree", and "family tree".
- Syntactic disambiguation (bracketing) is crucial for interpreting MWEs like "glass window cleaner" correctly.
- MWEs are essential for accurate machine translation, capturing subtle linguistic nuances across languages.

## * Conclusion:

The study of MWEs is crucial for a comprehensive understanding for language and developing effective NLP system. Thier unique properties and diverse types necessitate specialized methods for identifications, interpretation, and application in various NLP tasks.

## Parse Tree example



# • ch: 12

## 1. Intro to Information Retrieval (IR)

- IR is a successful application of NLP, evident in the prevalence of effective search engines used by approximately 85% of the web users.

- IR deals with representing, storing, organizing, and accessing info, which can range from documents and web pages to images, videos & even spoken content.

  - IR systems face challenges related to thier imprecision and incompleteness of user queries & document descriptions. Users often provide brief and ambiguous queries, requiring a "trial-and-error" approach rather than a direct "query-response" model.

- Matching between queries and info item is probabilistic, meaning the system estimates the relevance of potential answers.

## 2. Challenges of Natural Language in IR:

- **Polysemy**: A single word can have multiple meanings. This complicates the retrievel process as the system needs to determine the intended meaning withen the query and document context. Example: the word "bank" illustrates this challeng

- **Synonymy**: Different words can refer to the same concept. Users may employ various terms to describe the same info need, requiring the system to recognize the synonymous relationships

- **Spelling Errors and Variations**: Historical variations typos, regional differences, and translation all contribute

to discrepancies in spelling, hindering accurate retrieval.

## 3. Indexing in IR

- Indexing is crucial in IR as it transforms documents and queries into representation capturing thier semantic essence. This enable efficient comparison and retrieval.

"Effective search systems do not work directly with the documents (or the queries). They use different techniques and strategies to represent the main semantic aspects of the documents and queries. This process is called indexing."

- Key concepts in indexing include:
    - Structure Analysis and tokenization: Parsing the document structure and segmenting text into word tokens.
    - Stopword Removal: Eliminating common words like "the", "and", and "is" that offers little semantic value.

- **Morphological normalization:** Standardizing word forms through stemming to group variations under a common root (eg: "running" and "run").

- **Weighting:** Assigning weights to indexing units based on factors like term frequency, document frequency, and document length to prioritize more relevant terms.

"Term weighting creates a distinction among terms and increases indexing flexibility."

## 4. Query Expansion

- To enhance matching, query expansion techniqu introduce related words or phrases into the original query. This can be achieved through the use of thesaurus or by leveraging info from the document collection itself.

"The general principle is to expand the query using words or phrases having meanings similar or related to those appearing in the original query..."

# 5. The role of NLP in IR :

- **Morphology**: Stemming, a key morphological technique, addresses variation in word forms, improving retrieval accuracy. The choice of stemming algorithm and its complexity depends on the specific language and context.

"The goal of the morphological step in IR is to conflate morphological variants into the same form."

- **Orthographic variations & Spelling Errors:**

Addressing spelling variations is crucial, Techniques involve recognizing misspellings, accounting for punctuation differences, and resolving regional variations and transliterations.

- **Syntax:** Utilizing syntactic analysis to identify relationships between words can enhance indexing by capturing phrases and long-distance dependencies.

However, challenges related to grammar coverage and parsing accuracy have limited its widespread adaption.

- **Semantics:** Semantic analysis aims to resolve words sense ambiguity and enable concept-based indexing. while word sense disambiguation can be beneficial, leveraging semantic resources like thesauri and semantic networks have proven more effective, particularly in specific domains like medicine.

## 6. Conclusion:

- IR demonstrates the successful application of NLP in managing and retrieving vast amounts of information.
- The relationship between IR and NLP is multifaceted, with morphology, spelling correction, syntax, and semantics playing important roles.
- While current IR systems achieve impressive results, ongoing challenges remain, requiring further research and development to improve the effectiveness and sophistication of information retrieval in the face of increasingly complex and voluminous data. ("the IR field is an extensive applied NLP domain that is able to cope successfully with search and retrieval in the huge volume of information stored on the Web—users are, on average, able to find what they are looking for.")

# .Ch: 13

## 1. Information Extraction (IE)

IE aims to extract structured info from text, including entities, relationships, and events. It occupies a middle ground between simple Keyword searches and full text understanding, focusing on frequent linguistic patterns without requiring deep inference.

- **Applications**:
  - Intelligence gathering (buisness, government, military)
  - Specific web searches exceeding Keyword capabilities
  - Scientific literature research
  - Database building from textual corpora
  - Biomedical article curation

- **Named Entity Recognition (NER):** a Key IE technique, identifies proper names (person, company) and specific entities like dates & times.

NER finds particular use in biomedical applications with complex terminology.

## II. Sentiment Analysis (SA) & Subjectivity

- Textual info can be broadly classified into **facts** (objective) and **opinions** (subjective)
- Traditional text processing research focused heavily on factual info mining. However, opinions are increasingly crucial for individual and organizational decision-making.
- The rise of the web, particularly user-generated content, has significantly expanded the volume & availability for opinionated text. This necessitates automated tools for opinion descovery and summarization.

- Sentiment Analysis (SA) also known as opinion mining, addresses this need. It involves:

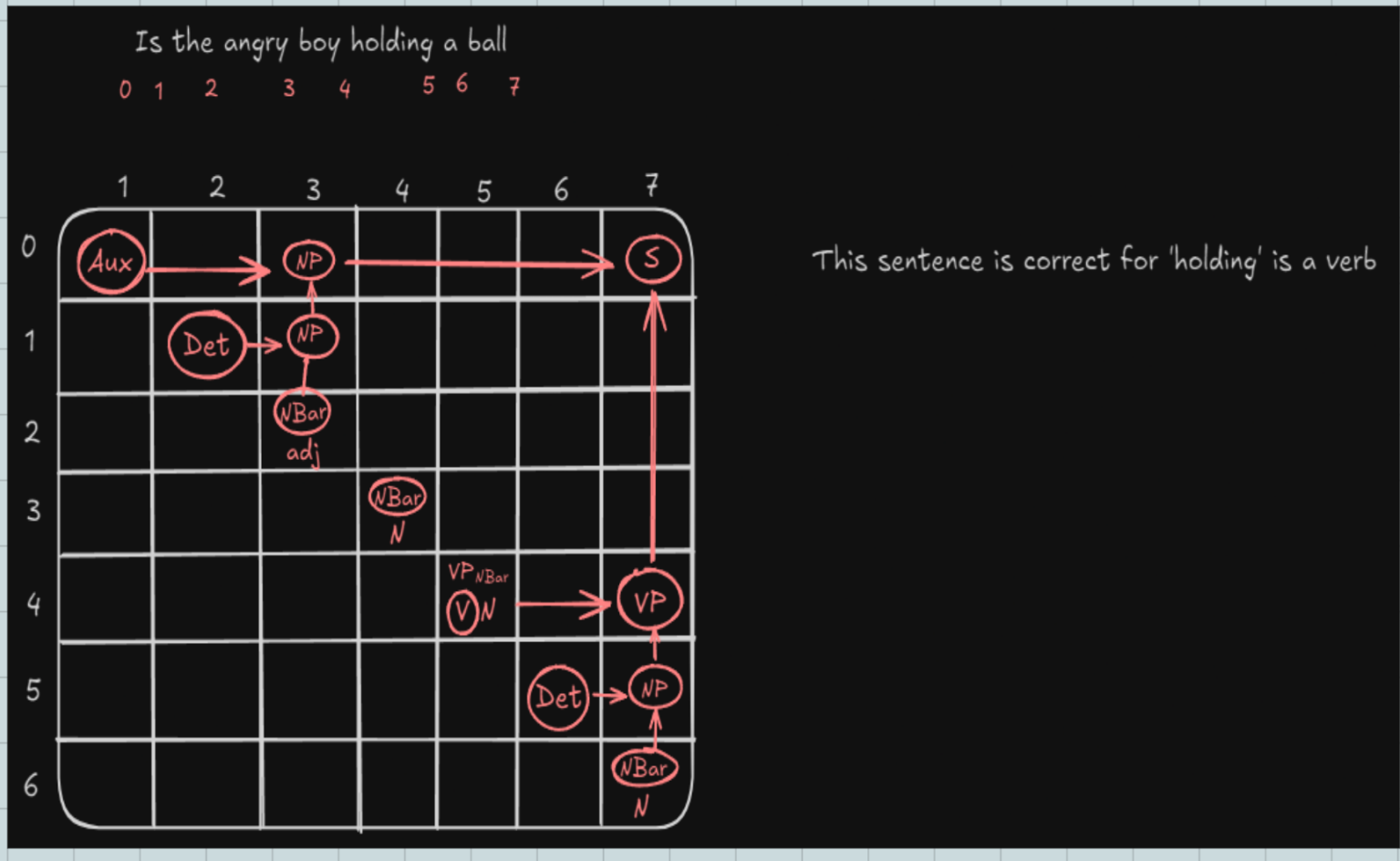
  - **Sentiment & Subjectivity Classification:** identifying sentences as subjective/objective and for subjective sentences, classifying them as positive, negative, or neutral.
  - **Feature-based SA:** identifying opinion targets and determining the sentiment expressed towards them.
  - **Sentiment analysis of comparative sentences:** understanding opinions expressed through comparisons (ex: "product A is better than product B").
  - **Opinion Search & Retrieval:** Combining info Retrieval with SA to find relevant opinions
  - **Opinion Spam detection:** identifying deceptive or manipulative opinions.

# III. Challenges & Future directions:

- SA is a complex NLP task with limited understanding and solutions. The reliance on Machine learning, despite its potential for accuracy, often yields results lacking human interpretability.

- The dynamic nature of language & the need for practical applications will continue to drive research and innovation in the field of SA.

## CKY Example



# Arabic NLP

## * Arabic Orthography:

- **Diacritics:** Arabic orthography employs diacritics (TashKiL) to indicate short vowels. However, these are often omitted in written text, leading to ambiguity: "The absence of short vowels (eg inner diacritics) prompts diverse sorts of ambiguities in Arabic. for example, the unvoweled word "كتب" can mean "wrote" (Kataba) or "books" (Kutub) depending on the diacritization.

- **Hamza Spelling:** The Hamza (ء) poses challenges due to its variable placement and rules that are "confusing even for native speakers". This

inconsistency needs to be addressed by NLP systems.

- **Defective Verb Ambiguity:** Defective verbs with long vowels undergo changes during conjugation, particularly when negated with the particle "لم" (lam), creating complexities for parsing & analysis.
- **Absence of Capital letters:** Unlike English, Arabic does not use capital letters, hindering Named Entity Recognition (NER) tasks that rely on capitalization cues to identify entities.
- **Lack of Uniformity:** Multiple transcription schemes and the absence of comprehensive lexical resources further complicate NLP development.

# * Arabic Morphology:

- **Highly Derivational & inflectional:** Arabic boasts a vast vocabulary built on a system of roots and patterns. With approx. 10,000 roots and 120 patterns, this complexity requires sophisticated morphological rules.
- **Templatic Morphological:** Words are formed by combining roots, patterns, and affixes, demanding an understanding of these structures for accurate processing.
- **Annexation:** The ability to combine words through conjunctions, though not common in traditional Arabic, adds another layer of complexity, especially in Modern standard Arabic.

# * Arabic Syntax:

- **Intricate Structure:** Automating the analysis of Arabic sentences is challenging due to the language's intricate syntax. Distinguishing between verbal & nominal sentences, with their unique structures and word order variations, requires specialized parsing techniques.
- **Multi-Word Expressions:** Multi-word expressions especially idiomatic ones, present challenges as their meaning cannot be derived from individual words. These expressions need to be treated as single units in NLP apps.
- **Anaphora Resolution:** Resolving anaphora, particularly phenominal & hidden anaphora, is a difficult task. Ambiguity arises from the lack of differentiation between human & non-human entities in pronouns.

- **Free Word Order:** Arabic allows for flexible word order, with variations like: VSO, SVO, and OVS, making sentance generation & question-answering systems more complex.

- **Agreement:** Agreement rules, sensitive to word order and encompassing number, gender, case, and definiteness, contribute to the intricacies of Arabic Syntax.

## * Conclusion:

Developing robust Arabic NLP applications demands tackling these linguistic challenges. Sophisticated algorithms are needed to handle diacritization, morphological analysis, syntactic parsing, and anaphora resolution. Building comprehensive lexical resources and addressing the lack of standardization are crucial for advancing the field. Despite the complexities, the richness and significance of the Arabic language make overcoming

these obstacles a worthwhile endeavor. As the document highlights, "Arabic as a language is both challenging and interesting." Addressing these challenges is essential for unlocking the potential of Arabic NLP and enabling its applications in various domains.

## DRT Example