

# Computational Linguistics, Natural Language Understanding, Natural Language Processing, Large Language Models and a Proposal

Presentations @ FCIM of UTM  
May 2026, Chisinau, Moldova\*

Nikitas N. Karanikolas

Dept. of Informatics and Computer Engineering

University of West Attica

<https://users.uniwa.gr/nnk/>

[nnk@uniwa.gr](mailto:nnk@uniwa.gr)

\* Erasmus+ mobility, 4-8 May, 2026

# Layout

- Short Definitions
- Linguistics
- Computational Linguistics
- Natural Language Generation
- Natural Language Understanding
- Natural Language Processing
- Large Language Models
- Proposal
- Closing

# Short Definitions

# Short Definitions of Linguistics and CL

- **Linguistics** is the scientific study of language. It spans both theoretical and practical applications. **Theoretical linguistics** is concerned with understanding the universal and fundamental nature of language and developing a general theoretical framework for describing it. **Applied linguistics** seeks to utilize the scientific findings of the study of language for practical purposes, such as developing methods of improving language education and literacy.
- **Computational Linguistics (CL)** is an interdisciplinary field concerned with the computational modelling of natural language, as well as the study of appropriate computational approaches to linguistic questions. Computational linguistics draws upon linguistics, computer science, artificial intelligence, logic, philosophy, cognitive science.

# Short Definitions of NLU, NLG, NLP, and LLMs

- **Natural Language Understanding (NLU)** is the effort, methods, tools and techniques to transform the human/natural language text to some structured form (like Predicate Calculus, FRL, KL-ONE, Case Grammars, etc) and handle it by computers.
- **Natural Language Generation (NLG)** is the reverse. It is a software process that produces natural written or spoken text from structured and unstructured data.
- **Natural Language Processing (NLP)** is a broader term. It is a branch of Artificial Intelligence (AI) which enables computers to analyze and synthesize natural (human) language and speech. It includes both NLU and NLG and more others.
- **Large Language Models (LLMs)** is a software process that uses deep learning techniques and massively large data sets to learn how to understand and generate natural text

# Linguistics

# Detailed presentation of Linguistics

- Linguistics emerged from the field of philology. Linguistics is also related to the philosophy of language, stylistics, rhetoric, semiotics, lexicography, and translation.
- The areas of linguistic analysis are **syntax** (how the words are ordered to form sentences), **semantics** (what are the meanings, the communicated ideas, facts and intentions of writer of sentences, documents or speaker(s) for oral communications), and **pragmatics** (how the context and the overall knowledge contributes to get the meaning).
- Two other areas of linguistic analysis are **morphology** (how morphemes – smaller than a word parts – are combined to form words), and **phonology** (the subset of sounds that can be created by the human mouth and which can be combined to express the sound effect of the words of some human language).

# Detailed presentation of Linguistics

- Relevant International Lecture of the instructor about **Phonology**:  
Machine Learning of Phonetic Transcription Rules for Greek, March 2019,  
Novi Sad, Serbia.  
[https://users.uniwa.gr/nnk/InternationalLectures/Machine\\_Learning\\_of\\_Phonetic\\_Transcription\\_Rules\\_for\\_Greek\\_2019.pdf](https://users.uniwa.gr/nnk/InternationalLectures/Machine_Learning_of_Phonetic_Transcription_Rules_for_Greek_2019.pdf)
- Regarding **Morphology**:  
the interest reader can study the appendix A of this presentation.
- Regarding **Pragmatics** we are giving an example:  
If some affirmative sentence is “Mary and Helen are sisters.” then pragmatics are who gives the negative answer to the question “Is Mary mother of Helen?”. The negative answer is based on the belief that a mother and her daughter can not be sisters.

# Computational Linguistics - CL

# Detailed presentation of CL

- The CL field overlapped with artificial intelligence since the efforts in the United States in the 1950s to use computers to automatically translate texts from foreign languages, particularly Russian scientific journals, into English. Since rule-based approaches were able to make arithmetic (systematic) calculations much faster and more accurately than humans, it was expected that lexicon, morphology, syntax and semantics can be learned using explicit rules, as well. After the failure of rule-based approaches, the term Computational Linguistics is coined in order to distinguish the field from AI.
- CL is a very active disciplinary supported by the Association of Computational Linguistics (ACL) and its more well known events are the COLING Conferences

# Natural Language Generation – NLG

# Detailed Presentation of NLG

- Systems, models and applications that can produce discourse (written text or oral speech) with meaning not recognized as artificial
- Stages: Planning, Microplanning, Realizing, Presenting (optional)
- Planning - decide what is:
  - the most interesting parts of input data (structured data),
  - the order of communicated ideas/facts,
  - the hidden rhetorical relations [cause, sequence, etc]
- Microplanning:
  - aggregation,
  - anaphora generation,
  - selection of lexical items,
  - decide the syntactic

# Detailed Presentation of NLG

- Realizing morphologically and orthographically correct discourse:
  - inflection,
  - orthography,
  - ordering of adjectives.
- Presenting (optional stage):
  - Written text: Titles, Emphasis (e.g. bold), punctuation marks
  - Oral speeches: Intonation, Sentence Type (Affirmative, Negative, Imperative, Exclamatory, Interrogative)
- Prerequisite for NLG is NLU

# Natural Language Understanding - NLU

# Overall NLU process

- NLU is the translation of human/natural language text to some structured form (Predicate Calculus, FRL, KL-ONE, Case Grammars, etc)
- Natural Language Understanding follows (usually) the following tasks:
  - Tokenization,
  - Part of Speech Tagging,
  - Syntactic Analysis,
  - Structural Disambiguation (Resolution of Syntactic Ambiguity),
  - Word Sense Disambiguation,
  - Semantic Representation,
  - Anaphora Resolution,
  - Optional Tasks: Affective Computing, Discourse Analysis, Pragmatics.

# Syntax

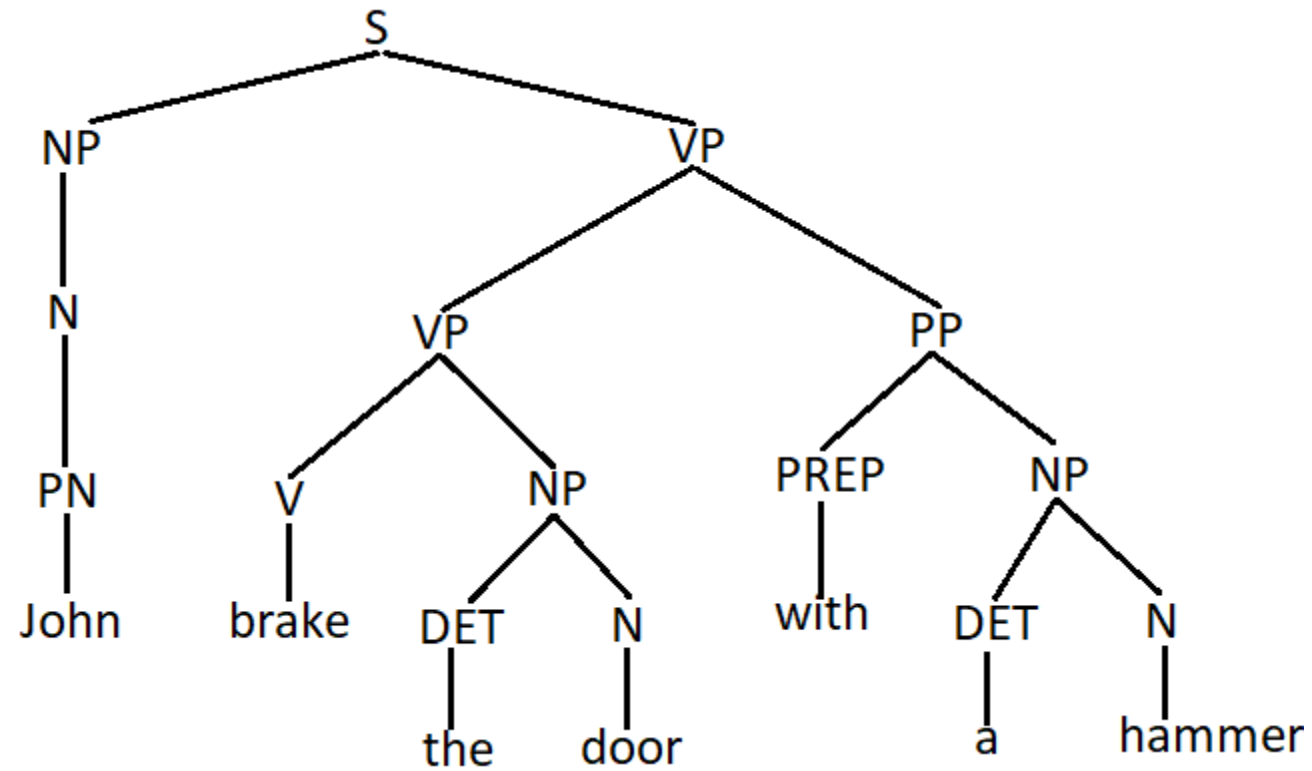
- A **syntactic analyzer** will check if a sentence is well formed and will return the **syntax tree**.
- The parts of this tree will then analyzed for representing the meaning of the sentence. There are **restrictions** for the allowed **syntactic sub-structures** that can correspond to **semantic structures**.
- Without syntactic analysis, we can not check these constraints.

# A syntax tree (phrase structure tree)

- John broke the door with a hammer

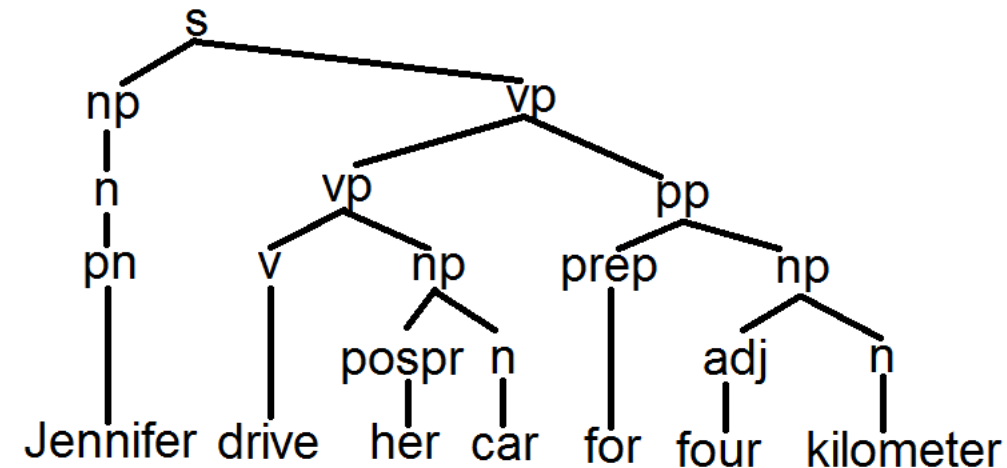
```
s(  np(n(pn(John))),
    vp(  vp(  v(broke,past).
          np(det(the),n(door)))
        pp(  prep(with),
            np(det(a), n(hammer)))
          )
    )
)
```

# Phrase Structure / Syntax / Derivation - Tree



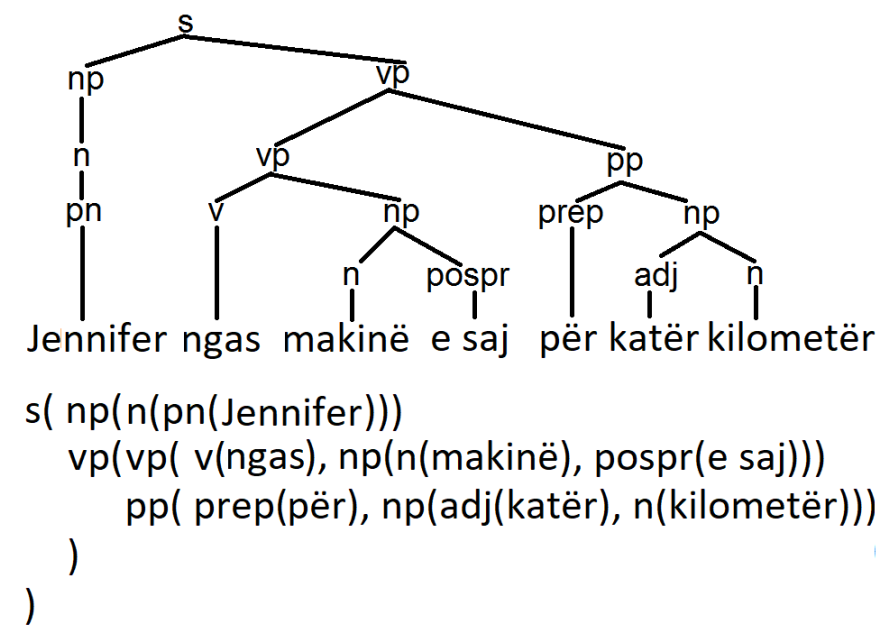
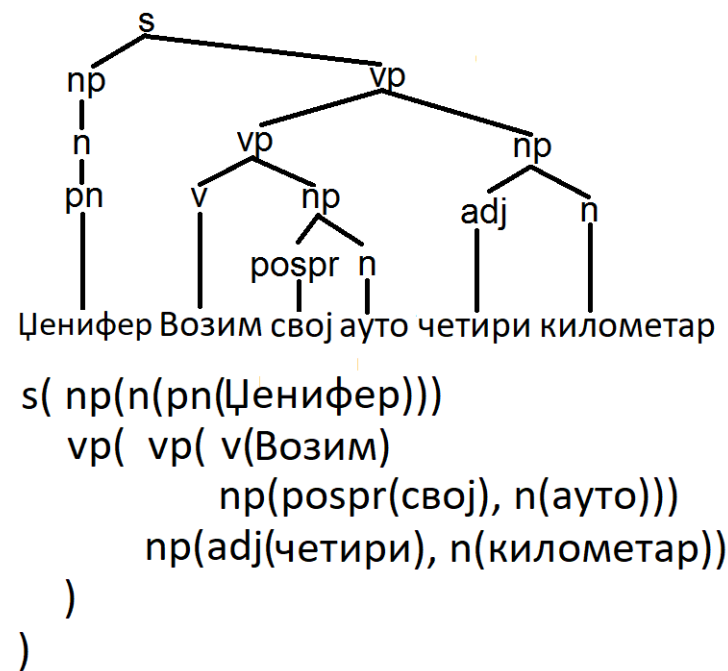
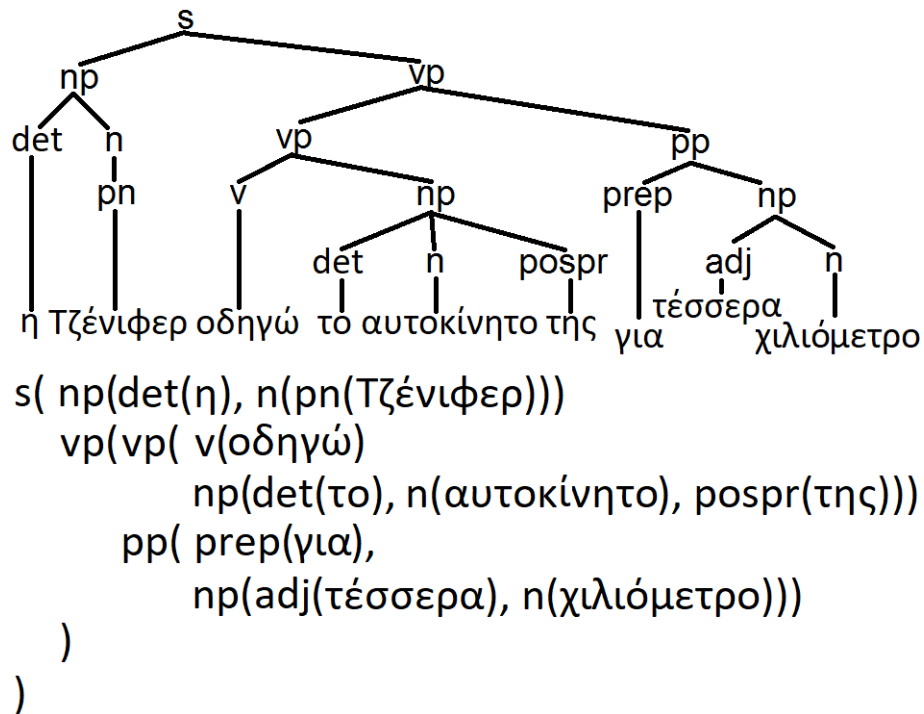
# Syntactic Analysis – Balkan Languages

- Jennifer drove her car, for four kilometers
- (in Greek) “Η Τζένιφερ οδήγησε το αυτοκίνητό της για τέσσερα χιλιόμετρα”,
- (in Serbian) “Џенифер је возила свој ауто четири километра”,
- (in Albanian) “Jennifer e ngau makinën e saj për katër kilometra”,
- (in Bulgarian) “Дженифър кара колата си четири километра”.



```
s( np(n(pn(Jennifer))),
    vp( vp( v(drive),
            np(pospr(her), n(car)))
        prep(for),
        np(adj(four), n(kilometer)))
    )
)
```

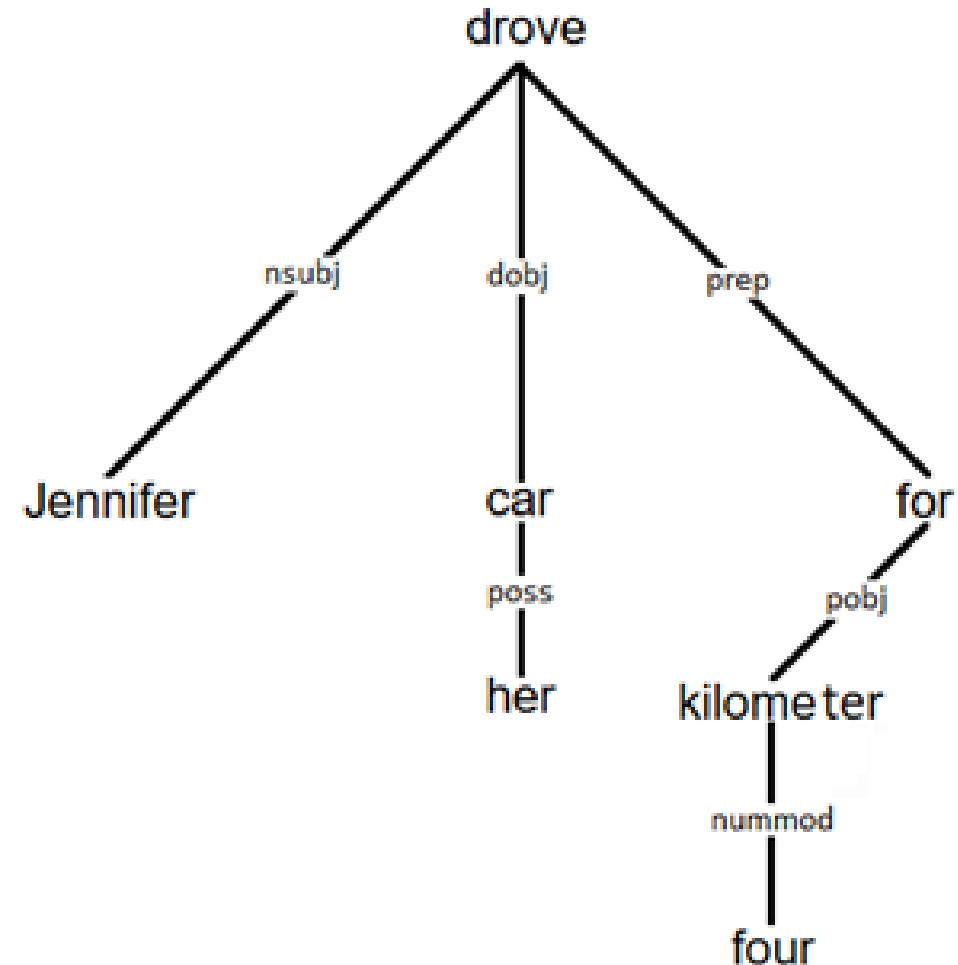
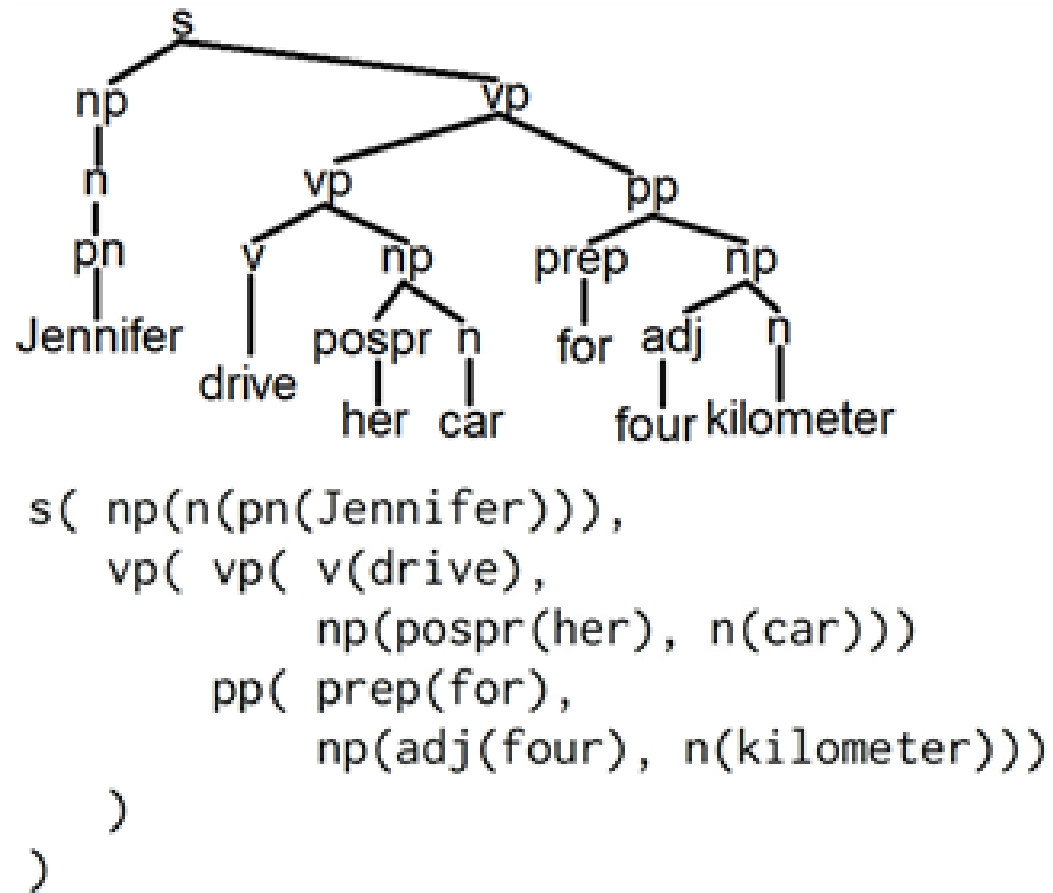
# Syntactic Analysis: Greek, Serbian, Albanian



- We don't present the syntactic analysis results for the Bulgarian example sentence, because of size limitations.
- We do not see prepositions in Serbian and Bulgarian sentence examples.
- In the examples in the other 3 languages (English, Greek, and Albanian) we see.
- The possessive pronoun precedes the noun in English and Serbian. In Greek, Albanian, and Bulgarian the possessive pronoun follows the noun.

# Constituent Parsing versus Dependency Parsing

"Jennifer drove her car, for four kilometers"



# Grammars

- Phrase Structure Grammars and Rules
  - *constituency-based*  
*binary division of the clause into **subject** (noun phrase NP) and **predicate** (verb phrase VP)*
  - a one-to-one-or-more correspondence
  - $A \rightarrow B C$ ;  $A \rightarrow (B) C$ ;  $A \rightarrow \{B, C\}$
  - $S \rightarrow NP VP$ ;  $NP \rightarrow (Det) N1$ ;  $N1 \rightarrow (AP) N1 (PP)$
  - Can be context free or context sensitive
- Dependency Grammars
  - dependency relation  
understanding of sentence logic in terms of predicates and their arguments
  - a one-to-one relation

# More for Grammars and Parsing

- Augmented Transition Networks (ATNs)
- Generalized Phrase Structure Grammar (GPSG)
- Lexical-Function grammar (LFG)
- Head-Driven Phrase Structure Grammar (HDPGs)
- Categorical Grammar (CG)
- Shallow parsing

# Structural Disambiguation (Resolution of Syntactic Ambiguity)

syntactic/parse (phrase structure) trees (in nested lists form) for the sentence  
**"Look at the dog with one eye."**

```
s( np(n(you))
  vp( v(look),
      pp (p(at),np(det(the),n(dog))),
      pp (p(with),np(adj(one),n(eye)))
    )
  )
```

```
s( np(n(you))
  vp( v(look),
      pp ( p(at),
          np( det(the),
              n(dog),
              pp(p(with), np(adj(one),n(eye)))
            )
        )
    )
  )
```

# Semantics

- A well formed syntactically sentence is not always correct
- “John drunk 3 liters gasoline” is syntactically correct but **gasoline** is a liquid that is not suitable for drinking by humans and John is a human.
- The semantics are what recognizes that John is a proper name and consequently it refers to a human and that gasoline is a liquid that is **not a kind of food or beverage** (that a selectional restriction imposes)
- There is also some semantic restriction that the consumed item in some verb of feeding should be food or beverage

# Representation

- One of the possible representations is the case grammar.
- The system was created by the American linguist **Charles J. Fillmore** in (1968). This theory analyzes the surface syntactic structure of sentences by studying the combination of **deep cases** (i.e. semantic roles) required by a specific **verb**.
- Deep cases can be: Agent, Object, Benefactor, Location, Instrument, etc
- See more in Bertram Bruce, DEEP CASE SYSTEMS FOR LANGUAGE UNDERSTANDING

# Case Grammars

- For instance, the verb "give" in English requires an Agent (A) and Object (O), and a Beneficiary (B); e.g. "Jones (A) gave money (O) to the school (B)."
- According to Fillmore, each verb selects a certain number of deep cases which form its **case frame**.
- Case frames are subject to certain constraints, such as that a deep case can occur only once per sentence.
- Some of the cases are **obligatory** and others are **optional**.
- Obligatory cases may not be deleted, at the risk of producing ungrammatical sentences. For example, *Mary gave the apples* is ungrammatical in this sense.

# Semantic representation

for sentence “Jennifer drove her car, for four kilometers”

## DRIVE

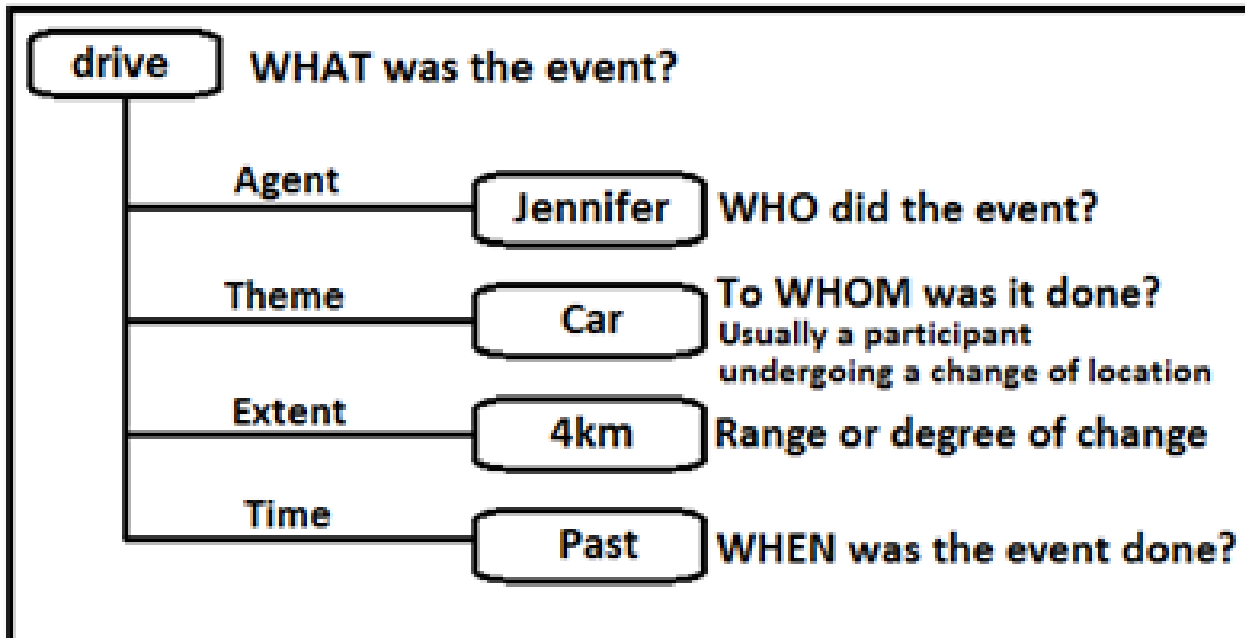
### Acceptable Thematic Roles:

Agent, Theme, Source, Destination, Location, Extent, Duration, Time

### Selectional Restrictions:

\*\*\*\*\*

*Jennifer drove her car for four kilometers*



## Key terms / ideas

- Frame Representation
- It is based on Verbs
- **Verbs** are having **Semantic/Thematic roles** or Slots (see Case grammars)
- We have to fill each Semantic role with **words** (or **phrases**) of the sentence
- **Selectional Restrictions** are restrictions to select the correct word or phrase as a slot filler

# Selectional Restrictions Types

- Based on the use of **specific prepositions** (introducing noun phrases)
- Based on the **position of (noun) phrases** in sentence (eg. Direct Object, Indirect Object) or **type sub-sentence** (clause)
- Based on the **meaning** of the word / phrase. What is the **position of the meaning in some hierarchy / ontology** of physical and abstract objects existing in the world?

# Thematic Roles according to Verbnet

Table 2: Thematic roles and example classes that use them

<b>Actor:</b>	used for some communication classes (e.g., Chitchat-37.6, Marry-36.2, Meet-36.2) when both arguments can be considered symmetrical (pseudo-agents).
<b>Agent:</b>	generally a human or an animate subject. Used mostly as a volitional agent, but also used in VerbNet for internally controlled subjects such as forces and machines.
<b>Asset:</b>	used for the Sum of Money Alternation, present in classes such as Build-26.1, Get-13.5.1, and Obtain-13.5.2 with `currency' as a selectional restriction.
<b>Attribute:</b>	attribute of Patient/Theme refers to a quality of something that is being changed, as in (The price)att of oil soared. At the moment, we have only one class using this role Calibratable cos-45.6 to capture the Possessor Subject Possessor-Attribute Factoring Alternation. The selectional restriction `scalar' (defined as a quantity, such as mass, length, time, or temperature, which is completely specified by a number on an appropriate scale) ensures the nature of Attribute.
<b>Beneficiary:</b>	the entity that benefits from some action. Used by such classes as Build-26.1, Get-13.5.1, Performance-26.7, Preparing-26.3, and Steal-10.5. Generally introduced by the preposition `for', or double object variant in the benefactive alternation.
<b>Cause:</b>	used mostly by classes involving Psychological Verbs and Verbs Involving the Body.
<b>Location, Destination, Source:</b>	used for spatial locations.
<b>Destination:</b>	end point of the motion, or direction towards which the motion is directed. Used with a `to' prepositional phrase by classes of change of location, such as Banish-10.2, and Verbs of Sending and Carrying. Also used as location direct objects in classes where the concept of destination is implicit (and location could not be Source), such as Butter-9.9, or Image impression-25.1.
<b>Source:</b>	start point of the motion. Usually introduced by a source prepositional phrase (mostly headed by `from' or `out of'). It is also used as a direct object in such classes as Clear-10.3, Leave-51.2, and Wipe instr-10.4.2.
<b>Location:</b>	underspecified destination, source, or place, in general introduced by a locative or path prepositional phrase.
<b>Experiencer:</b>	used for a participant that is aware or experiencing something. In VerbNet it is used by classes involving Psychological Verbs, Verbs of Perception, Touch, and Verbs Involving the Body.
<b>Extent:</b>	used only in the Calibratable-45.6 class, to specify the range or degree of change, as in The price of oil soared (10%)ext. This role may be added to other classes.
<b>Instrument:</b>	used for objects (or forces) that come in contact with an object and cause some change in them. Generally introduced by a `with' prepositional phrase. Also used as a subject in the Instrument Subject Alternation and as a direct object in the Poke-19 class for the Through/With Alternation and in the Hit-18.1 class for the With/Against Alternation.
<b>Material and Product:</b>	used in the Build and Grow classes to capture the key semantic components of the arguments. Used by classes from Verbs of Creation and Transformation that allow for the Material/Product Alternation.
<b>Material:</b>	start point of transformation.
<b>Product:</b>	end result of transformation.
<b>Patient:</b>	used for participants that are undergoing a process or that have been affected in some way. Verbs that explicitly (or implicitly) express changes of state have Patient as their usual direct object. We also use Patient1 and Patient2 for some classes of Verbs of Combining and Attaching and Verbs of Separating and Disassembling, where there are two roles that undergo some change with no clear distinction between them.
<b>Predicate:</b>	used for classes with a predicative complement.
<b>Recipient:</b>	target of the transfer. Used by some classes of Verbs of Change of Possession, Verbs of Communication, and Verbs Involving the Body. The selection restrictions on this role always allow for animate and sometimes for organization recipients.
<b>Stimulus:</b>	used by Verbs of Perception for events or objects that elicit some response from an experiencer. This role usually imposes no restrictions.
<b>Theme:</b>	used for participants in a location or undergoing a change of location. Also, Theme1 and Theme2 are used for a few classes where there seems to be no distinction between the arguments, such as Differ-23.4 and Exchange-13.6 classes.
<b>Time:</b>	class-specific role, used in Begin-55.1 class to express time.
<b>Topic:</b>	topic of communication verbs to handle theme/topic of the conversation or transfer of message. In some cases, like the verbs in the Say-37.7 class, it would seem better to have `Message' instead of `Topic', but we decided not to proliferate the number of roles.

# Frames

- A **frame language** is a technology used for **knowledge representation** in artificial intelligence. Frames are stored as ontologies of sets and subsets of the frame concepts.
- They are similar to class hierarchies in object-oriented languages although their fundamental design goals are different. Frames are focused on explicit and intuitive representation of knowledge whereas objects focus on encapsulation and information hiding
- Frames originated in AI research and objects primarily in software engineering. However, in practice the techniques and capabilities of frame and object-oriented languages overlap significantly.
- Implementations:
  - KL-ONE,
  - LOOM,
  - PowerLoom,
  - OWL
- Semantic editors
  - Protégé
- Semantic Reasoners (A **semantic reasoner**, **rules engine**, or simply a **reasoner**, is a piece of software able to infer logical consequences from a set of asserted facts or axioms.)
  - Pellet
  - RacerPro
  - FaCT++
  - Hermit

# More representations

- Case Grammar
- Frames
- First Order Logic
- Semantic Networks
- Conceptual Dependency
- Rule-Based
- Conceptual Graphs

# Semantically Ambiguous

- word *slug* meanings:
  - Coin
  - Bullet
  - Loafer
  - Gastropod without shell
- word *bass* meanings:
  - a type of fish
  - tones of low frequency
  - a type of instrument
- Example sentences with ambiguity – word *bass*:
  - I went fishing for some sea *bass*
  - The *bass* line of the song is too weak

# Word Sense Disambiguation (WSD)

- **word-sense disambiguation (WSD)** is an open problem for natural language processing and ontology
- WSD is identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings
- The solution to this problem impacts other computer-related writing, such as discourse, improving relevance of search engines, anaphora resolution, coherence, inference, *etc.*

# Word Sense Disambiguation (WSD)

- As in all NLP there two main approaches for WSD
  - deep approaches
  - shallow approaches.
- **Deep approaches** presume access to a comprehensive body of world knowledge. Knowledge, such as "you can go fishing for a type of fish, but not for low frequency sounds" and "songs have low frequency sounds as parts, but not types of fish", is then used to determine in which sense the word bass is used. These approaches are not very successful in practice, mainly because such a body of knowledge does not exist in a computer-readable format, outside very limited domains.

# WSD – Shallow approaches

- **Shallow approaches** don't try to understand the text. They just consider the surrounding words, using information such as "if bass has words sea or fishing nearby, it probably is in the fish sense; if bass has the words music or song nearby, it is probably in the music sense."
- **Rules for Shallow approaches** can be automatically derived by the computer, using a training corpus of words tagged with their word senses.
- However, shallow approaches **can be confused** by sentences like *The dogs bark at the tree* which contains the word bark near both tree and dogs. **Bark meanings: γαρ (γαυγίζω), shell (φλοιός).**

# Anaphora resolution

- John is going to visit Nick. **He** is a good man.  
Meanings:  
John (who is a good man) is going to visit Nick.  
John is going to visit Nick (who is a good man).
- **He** refers to:  
John (in the first case)  
Nick (in the second case)
- **Definition**: the problem of resolving what a pronoun, or a noun phrase refers to

# Anaphora Types

- Reflexive pronoun
  - Mary and John had dinner together. **Mary** cooked a wonderful roast beef by **herself**.
- Reciprocal pronoun
  - **Mary and Aleksandra** are friends of **each other**.
- Pronominal
  - **John** works hard. **He** wants to buy a new car.
- Lexical
  - **Engineers** from many companies attended the conference. **The participants** found the topics very attractive.

# Anaphora Types

- One anaphora
  - If you can not attend **a tutorial** in the morning, you can go for an afternoon **one**.
- Intrasentential
  - Antecedent (referenced) and anaphor are in the same sentence
- Intersentential
  - Antecedent (referenced) and anaphor are in different sentences

# Anaphora Resolution approaches

## Mitkov 1999

- Approaches to anaphora resolution usually rely on a set of "anaphora resolution factors".
- Factors used frequently in the resolution process include gender and number agreement, c-command constraints, semantic consistency, syntactic parallelism, semantic parallelism, salience, proximity etc.
- These factors can be
  - "eliminating" i.e. discounting certain noun phrases from the set of possible candidates (such as gender and number constraints, c-command constraints, semantic consistency)
  - "preferential", giving more preference to certain candidates and less to others (such as parallelism, salience).

# Anaphora resolution approaches

## Karanikolas 1993

- Chomsky's Binding Conditions imply that the antecedent of an anaphor can not be the antecedent of a pronominal.
- There exist examples in the literature where the above implications does not hold
- Two different definitions of the governed category can resolve the problem
- We have implemented a system based on these (Binding) conditions and the two different definitions of the governed category which is compliant with all examples in the literature

# Weaknesses of NLG and NLU

- NLU/NLG systems **require skilled experts** to create rules
- There exist **numerous human languages** with varying grammar and syntax
- **Rules** are **numerous, complex, overlapping,** or **contradicting** and significant effort is needed for their maintenance
- **R&D** groups working on NLU/NLG worldwide are **isolated** and not sharing their techniques and accomplishments, due to interests / competition
- Since languages evolve, NLU/NLG systems **may become outdated**
- Handling **ambiguity** and **variability** of natural languages is **hard**
- NLU/NLG systems may suffer from **bias**, originating on opinions of their developers

# Natural Language Processing - NLP

# NLP Applications

- Summarization
- Text Classification
- Machine Translation
- Question Answering
- Text Generation
- Computer Assisted Assessment
- Sentiment Analysis
- Opinion Mining
- Subjectivity Analysis
- Corpus Building
- Named Entity Recognition
- Chatbots – Virtual Assistants
- Image Captioning

# Summarization

- Can be Comprehensive (semantic oriented) or Extractive (shallow processing)
- **Extractive** is based on the selection of the most prominent sentences to convey the meaning. It is based on:
  - Weight of words (TF-IDF, TF-ISF, TF-RIDF)
  - Sentence Location (Baxentale, News Articles)
  - Title Words
- See also my previous lecture “Extractive summarization”, June 2017, Belgrade

<https://users.uniwa.gr/nnk/InternationalLectures/Extractive%20summarization%20June%202017.pdf>

# Text Summarization Applications

- **News Organizations:** Automatic generation of brief news article summaries for quick comprehension.
- **Researchers:** Condensing academic papers for rapid identification of key findings.
- **Content Aggregation Platforms and Websites:** Creating digests aiding user article selection based on interests.
- **Finance Sector:** Summarizing financial reports for swift evaluation by investors and analysts.
- **Healthcare Professionals:** Summarizing medical conversations, records, and research papers for streamlined tasks.
- **Social Media Insights:** Summarizing discussions and reviews for understanding public sentiment and reactions.
- **Market Research:** Distilling consumer feedback and reviews for prompt trend identification.
- **Educational Support:** Creating concise educational material summaries for student comprehension.

# Text Classification

- Assign to a documents a **class label** (the category that the document belongs)
- Can be based on the **existence of words or phrases**
- The method needs **training** and **training data** (pre-classified documents)
- It is Critical to create an Authority List of **words or phrases** that will be **appropriate to discriminate between classes**
- See also:

<https://users.uniwa.gr/nnk/InternationalLectures/Text%20classification%20based%20on%20phrases%20June%202017.pdf>

# Text Classification Applications

- **Email Sorting:** Distinguishing spam from genuine emails by evaluating content and sender details.
- **E-commerce Taxonomy:** Organizing products and multi-level taxonomy trees.
- **Social Media for Sentiment Analysis:** Assessing user-generated content for sentiments - positive, negative, or neutral.
- **Personalized Recommendations:** Tailored suggestions in e-commerce platforms.
- **Healthcare Data:** Classifying medical records and patient data to enhance organization and support research.
- **Legal Document Classification:** Simplifying document management for legal firms by classifying contracts and case files.
- **Fraud Detection in Finance:** Identifying suspicious transactions and user behavior for fraud detection.

# Machine Translation Applications

- **Translation Services:** Text, documents, websites, and spoken language.
- **Content Creation and Marketing:** Translating content into multiple languages.
- **Government and Diplomatic Use:** Translating official documents, treaties, and diplomatic communications.
- **News Agencies:** Swiftly translating news articles to provide timely and accurate information in multiple languages.
- **E-commerce and Retail:** Translating product descriptions, reviews, and checkout processes.
- **International Business Communications:** Translating contracts, emails, and documentation for negotiations and collaborations.
- **Travel Industry Integration:** Providing multilingual information for tourists through apps and websites.
- **Customer Support:** Facilitating communication between international customers and support teams.
- **Language Learning Platforms:** Offering language courses and translation assistance to enhance language skills.

# Question Answering – Applications

- **Company Chatbots:** Addressing customer inquiries, providing information, and troubleshooting.
- **Voice-Activated Assistants:** assist users and execute tasks.
- **E-commerce:** Enabling proactive information seeking for informed purchasing decisions.
- **E-Learning Platforms:** Addressing student queries, providing explanations and solutions.
- **Healthcare:** Assisting healthcare professionals with clinical queries, symptoms, and generating reports.
- **Legal Sector:** Providing insights into case law, statutes, and regulations for legal research.
- **Technical Support:** Offering guidance and solutions for software and hardware queries.

# Text Generation – Applications

- **Revolutionizing Industries:** Automating processes across diverse sectors.
- **Content Creation:** product descriptions, ads, promotions, news articles and blogs.
- **Financial Reporting Automation:** Analyzing market data to automatically generate financial reports.
- **Educational Contribution:**
  - Facilitating e-learning by producing quizzes, study materials, and explanations.
  - Enhancing engagement and personalized learning experiences.
- **Creative Writing Support:** Contributing to narratives, short stories, and poetry.
- **Chatbot & Virtual Assistant Roles:**
  - Empowering chatbots with natural language responses for customer inquiries.
  - Assisting users in tasks, providing information, and troubleshooting issues.

# Computer Assisted Assessment

- Mechanically assign a grade to an answer with respect to the expected (correct) answer.
- There is a need for positive training (correct answers and textbook) data and negative training data (erroneous answers)
- Can be based on phrases

# Sentiment Analysis, Opinion Mining, Subjectivity Analysis

- Subjectivity Analysis  
classify a given text as subjective or objective
- Sentiment Analysis or Polarity Analysis  
Once a text is subjective  
Assign a score Positive or Negative
- Affective Computing  
Attempt to identify emotional charge
  - Happiness
  - Sadness
  - Fear
  - Anger - Passion

# Sentiment Analysis Applications

- **Social Media and Customer Feedback:** Assessing sentiments to manage reputation and enhance product quality.
- **Diverse Data Sources:** Processing surveys, forms, and reviews to gauge customer satisfaction.
- **Financial Sector:** Analyzing news, reports, and social media for market sentiment.
- **Polling and Political Forecasting:** Gauging public sentiment for informed decision-making in politics.
- **Entertainment Industry:** Evaluating audience reactions to movies and TV shows.
- **Product Launches:** Assessing public sentiment to evaluate market reception.
- **Healthcare:** Understanding patient sentiment to drive improvements in patient care.
- **Social Media:** Measuring campaign effectiveness and tracking brand trends.
- **News Content Tailoring:** Assessing reader reactions to tailor preferences.
- **Customer Support Insights:** Identifying dissatisfied customers.
- **Travel Services:** Analyzing reviews to enhance services and increase bookings.

# Corpus Building

- Dialectal lexicon building  
See my previous lecture “Dialectal lexicon building: requirements and technical specifications”, Belgrade June 2017  
<https://users.uniwa.gr/nnk/InternationalLectures/Dialectal%20lexicon%20building%20June%202017.pdf>
- Dialectal corpora building  
See my previous lecture “Dialectal Corpora Building (for oral and written sources)”, Belgrade June 2017  
<https://users.uniwa.gr/nnk/InternationalLectures/Dialectal%20Corpora%20Building%20June%202017.pdf>

# Named Entity Recognition

- to be expanded
- See the presentaion by Cvetana Krstev, Department of Library and Information Sciences, Faculty of Philology, University of Belgrade, Serbia  
in the following link:

<https://users.uniwa.gr/nnk/InternationalLectures/NamedEntities.pdf>

# Chatbots & Virtual Assistants Applications

- **E-commerce & Online Services:** Chatbots can assist customers, provide product details, and resolve issues instantly.
- **Banking & Finance:** Virtual assistants aid with inquiries and financial guidance.
- **Travel:** Chatbots facilitate bookings, offer travel information.
- **Healthcare:** Virtual assistants provide medical information and schedule appointments.
- **Language Learning:** Virtual tutors assist in practicing languages.
- **Human Resources:** Chatbots aid with HR-related inquiries and company policies.
- **Online Retail:** Chatbots offer personalized product suggestions.
- **Business Reception:** Virtual receptionists manage calls and provide information.
- **Education:** Chatbots can assist with course-related queries and assignments.
- **News Updates:** Chatbots deliver tailored news content and updates to users.
- **Smart Home Devices:** Voice-activated assistants are used for inquiries, smart device control, and information provision.

# Image Captioning Applications

- **Social Media Platforms:** Use NLP to create image captions, benefiting user engagement and accessibility, particularly for the visually impaired.
- **E-commerce:** Image captioning in catalogs aids product management and improves search experiences.
- **Education:** Educational platforms use image captions to explain visual content, assisting students in understanding complex visuals.
- **Healthcare:** Image captioning assists in describing medical images, aiding in diagnostics and conveying findings among medical professionals.
- **Image Search and Retrieval:** Image captioning in search engines helps users find images by describing them in text queries.
- **Content Moderation:** Social media platforms use image captioning to identify inappropriate or harmful content, ensuring a safer online environment.
- **News Content:** News agencies utilize image captioning to enhance the accessibility of news articles for readers.

# Large Language Models – LLMs

- The process of evolution of language models has gone under four main stages:
  - The first stage was in 1990s, where statistical models on n-gram languages were used both in NLP and Information Retrieval tasks
  - The next stage involved the use of neural networks, such as RNNs, to be replaced later by the more efficient bi-LSTM models, leading to the class of Pre-trained Language Models (PLMs)
  - However, those models' performance excelled when the model's size or the data size was significantly increased
  - This was the opening of today's Large Language Models (LLMs)

# Detailed Presentation of LLMs with usage examples 1/9

Μπορείς να μου δώσεις τους αρχικούς χρόνους του ρήματος πλέκω στην ενεργητική φωνή στα νέα ελληνικά, διακρίνοντας τον Μέλλοντα σε στιγμιαίο, εξακολουθητικό και συντελεσμένο;



Φυσικά! Στα Νέα Ελληνικά, οι αρχικοί χρόνοι του ρήματος πλέκω στην ενεργητική φωνή, με διάκριση του Μέλλοντα, παρουσιάζονται ως εξής:

- **Ενεστώτας:** πλέκω
- **Παρατατικός:** έπλεκα
- **Μέλλοντας:**
  - **Στιγμιαίος:** θα πλέξω
  - **Εξακολουθητικός:** θα πλέκω
  - **Συντελεσμένος:** θα έχω πλέξει
- **Αόριστος:** έπλεξα
- **Παρακείμενος:** έχω πλέξει
- **Υπερσυντέλικος:** είχα πλέξει



Αν χρειάζεσαι κάτι επιπλέον, πες μου! 😊

# Detailed Presentation of LLMs with usage examples 2/9

Μπορείς να μου δώσεις τους αρχικούς χρόνους του ρήματος βλέπω στην ενεργητική φωνή στα νέα ελληνικά;



Βεβαίως! Οι αρχικοί χρόνοι του ρήματος **βλέπω** στην ενεργητική φωνή στα Νέα Ελληνικά είναι οι εξής:

- **Ενεστώτας:** βλέπω
- **Παρατατικός:** έβλεπα
- **Μέλλοντας:**
  - **Εξακολουθητικός:** θα βλέπω
  - **Στιγμιαίος:** θα δω
  - **Συντελεσμένος:** θα έχω δει
- **Αόριστος:** είδα
- **Παρακείμενος:** έχω δει
- **Υπερσυντέλικος:** είχα δει



# Detailed Presentation of LLMs with usage examples 3/9

Μπορείς να μου βρεις τους αρχικούς χρόνους στα νέα ελληνικά του ρήματος τρώω;



Βεβαίως! Οι αρχικοί χρόνοι του ρήματος **τρώω** στην ενεργητική φωνή στα Νέα Ελληνικά είναι οι εξής:

- **Ενεστώτας:** τρώω
- **Παρατατικός:** έτρωγα
- **Μέλλοντας:**
  - **Εξακολουθητικός:** θα τρώω
  - **Στιγμιαίος:** θα φάω
  - **Συντελεσμένος:** θα έχω φάει
- **Αόριστος:** έφαγα
- **Παρακείμενος:** έχω φάει
- **Υπερσυντέλικος:** είχα φάει

# Detailed Presentation of LLMs with usage examples 4/9

Failure  
'τρωθώ' =  
corrothion

Μπορείς να μου δώσεις τους αρχικούς χρόνους στα νέα ελληνικά του ρήματος τρώγομαι στην παθητική φωνή, διακρίνοντας το μέλλοντα σε εξακολουθητικό, στιγμιαίο και συντελεσμένο;



Βεβαίως! Το ρήμα "τρώγομαι" στην παθητική φωνή, με τη διάκριση του μέλλοντα σε εξακολουθητικό, στιγμιαίο και συντελεσμένο, έχει τους εξής αρχικούς χρόνους:

- Ενεστώτας: τρώγομαι
- Παρατατικός: τρωγόμεν
- Αόριστος: τραφώ
- Παρακείμενος: έχω τραφεί
- Εξακολουθητικός μέλλοντας: θα τρώγομαι
- Στιγμιαίος μέλλοντας: θα τρωθώ
- Συντελεσμένος μέλλοντας: θα έχω τραφεί
- Υπερσυντέλικος: είχα τραφεί



# Detailed Presentation of LLMs with usage examples 5/9

Καλησπέρα, θα ήθελα να μετατρέψουμε κάποιες προτάσεις από την ενεργητική σύνταξη στην παθητική σύνταξη.

Καλησπέρα! Βεβαίως, γράψτε μου τις προτάσεις που θέλετε να μετατρέψουμε. 😊

Ο Ιορδάνης μετέφερε το γάλα στο ψυγείο.

Το γάλα μεταφέρθηκε στο ψυγείο από τον Ιορδάνη.

# Detailed Presentation of LLMs with usage examples 6/9

Εχθές το βράδυ, ο Παναγιώτης έκανε τριανταπέντε τσουρέκια.

Εχθές το βράδυ, τριανταπέντε τσουρέκια έγιναν από τον Παναγιώτη.



# Detailed Presentation of LLMs with usage examples 7/9

Μπορείς να μου μετατρέψεις την παρακάτω πρόταση από ενεργητική σύνταξη σε παθητική; Ο Παναγιώτης είχε δει τις σημειώσεις του καθηγητή όταν πήγε να τον συναντήσει.

Οι σημειώσεις του καθηγητή είχαν ειπωθεί από τον Παναγιώτη όταν πήγε να τον συναντήσει.

Υπάρχει κι άλλος τρόπος;

Ναι, μπορεί να διατυπωθεί και ως εξής:

Οι σημειώσεις του καθηγητή είχαν ήδη γίνει ορατές από τον Παναγιώτη όταν πήγε να τον συναντήσει.

Ή, αν θέλουμε να διατηρήσουμε τη φυσικότητα του λόγου:

Οι σημειώσεις του καθηγητή είχαν ήδη διαβαστεί από τον Παναγιώτη όταν πήγε να τον συναντήσει.

# Detailed Presentation of LLMs with usage examples 8/9

Κλείδωσα την πόρτα πριν φύγω.

Η Μαρία έπλεξε ένα όμορφο κασκόλ για τη φίλη της.

Είδα ένα ενδιαφέρον ντοκιμαντέρ στην τηλεόραση.

Η Μαρία έφαγε το μεσημεριανό της γεύμα.

Ο δάσκαλος λέει μια ιστορία στους μαθητές του.

Ο πατέρας βγάζει την καρέκλα από την αποθήκη.

Η μητέρα βάζει τα ρούχα στο πλυντήριο.

Ο πεζός βρέχει τα πόδια του περπατώντας στη βροχή.

Ο στρατηγός στρέφει τις δυνάμεις του προς τη νέα κατεύθυνση.



# Detailed Presentation of LLMs with usage examples 9/9

Ορίστε οι προτάσεις μετατρεπμένες σε παθητική σύνταξη:

1. Η πόρτα κλείδωσε πριν φύγω.
2. Ένα όμορφο κασκόλ πλέχτηκε από τη Μαρία για τη φίλη της.
3. Ένα ενδιαφέρον ντοκιμαντέρ είδε στην τηλεόραση.
4. Το μεσημεριανό γεύμα φαγώθηκε από τη Μαρία.
5. Μια ιστορία λέγεται από τον δάσκαλο στους μαθητές του.
6. Η καρέκλα βγαίνει από την αποθήκη από τον πατέρα.
7. Τα ρούχα μπαίνουν στο πλυντήριο από τη μητέρα.
8. Τα πόδια του πεζού βρέχονται περπατώντας στη βροχή.
9. Οι δυνάμεις του στρατηγού στρέφονται προς τη νέα κατεύθυνση.

Σωστό 1: Η πόρτα κλειδώθηκε [από εμένα] πριν φύγω.

Σωστό 3: Ένα ενδιαφέρον ντοκιμαντέρ ειδώθηκε [από εμένα] στην τηλεόραση.

# Weaknesses of LLMs

- are prone to produce **misinformation**, **bias**, and **impolite** language resulting from their training data
- can generate fake, harmful, misleading, content, or propaganda, raising **ethical concerns**
- tend to ‘**hallucinate**’ and produce, with full conviction, high quality text which contains factually incorrect information
- since they rely on training data, **use potentially obsolete information** and may generate responses that are no longer accurate or relevant
- operate like **black boxes**, without the possibility to interpret / explain why they produced specific output
- are **not easily controllable** and when mistaken answers are identified, it is hard, if not impossible to diagnose and fix the error
- may handle sensitive and personal data and measures for **data protection** are needed
- are **costly to train**, since massive amounts of data, powerful computing systems and significant energy consumption are needed

# Proposal

# Resolving Weaknesses

- Life attitude / Στάση ζωής
  - Are you planning to **accept passively whatever changes that the exciting technology (and behind it the Companies) imposes**. Young people prefer to be “?smart?” and use the technology to gain whatever they want, without recognizing the drawbacks and without asking why this is this way and not the other way.
  - There is no money any more. They are **starting to exchange the plastic bank (credit, debit, prepaid) cards with other digital means** (already done with smart mobiles). This way, they will remove / cancel the bank cards. **Next, they will be possibly decide to cancel our fingers!**
- Resolving Weaknesses of NLPs – Make rule-based NLP scientists to work together and exchange technologies – Share their data resources – Spread the knowledge (and the rules) humans create.
- Resolving Weaknesses of LLMs – Set the rule-based NLPs to correct the weaknesses of LLMs – NLPs in the role of mentor for the LLMs – NLPs auditing the LLMs
- mutually beneficial synergy of the two technologies (in Greek: κοινή, επωφελία αμφοτέρων, συνλειτουργία των δύο τεχνολογιών)

# Deponent Verbs to passive voice by LLMs - Weakness

original	LLM result	correct conversion
in English	comment	in English
<p>Ο Δημήτρης αντιλαμβάνεται τις εξισώσεις γιατί είναι έξυπνος</p> <p>[Ο Dimitris <u>an-</u> <u>tilamvanetai</u> <u>tis</u> <u>exisoseis</u> <u>giati</u> <u>einai</u> <u>exypnos</u>]</p>	<p>Οι εξισώσεις αντιλαμβάνονται από τον Δημήτρη γιατί είναι έξυπνος</p> <p>[Οι <u>exisoseis</u> <u>an-</u> <u>tilamvanontai</u> <u>apo</u> <u>ton</u> <u>Dimitri</u> <u>giati</u> <u>einai</u> <u>ex-</u> <u>ypnos</u>]</p>	<p>Οι εξισώσεις γίνονται αντιληπτές από το Δημήτρη γιατί είναι έξυπνος</p> <p>[Οι <u>exisoseis</u> <u>ginontai</u> <u>antiliptes</u> <u>apo</u> <u>to</u> <u>Dimitri</u> <u>giati</u> <u>einai</u> <u>exypnos</u>]</p>
<p>Dimitris understands equations because he is smart</p>	<p>It is an unnatural / strange sentence</p>	<p>The equations are understood by Dimitris because he is smart</p>

# Deponent Verbs to passive voice by LLMs - Weakness

<b>original</b>	<b>LLM result</b>	<b>correct conversion</b>
<b>in English</b>	<b>comment</b>	<b>in English</b>
Ο Παναγιώτης βαριέται τα πάρτι [Ο Panagiotis <u>varietai ta</u> <u>parti</u> ]	Τα πάρτι βαριούνται από τον Παναγιώτη [Τα <u>parti variountai</u> <u>apo ton Panagioti</u> ]	Τα πάρτι γίνονται βαρετά στον Πα- ναγιώτη [Τα <u>parti ginontai vareta ston</u> <u>Panagioti</u> ]
Panagiotis is bored of parties	It is completely not acceptable by native speakers	Parties are getting boring for Panagiotis

# Deponent Verbs to passive voice by LLMs - Weakness

original	LLM result	correct conversion
in English	comment	in English
Εκμεταλλεύονται τους φυσικούς πόρους [ <u>Ekmetallevontai</u> <u>tous</u> <u>fysikous</u> <u>porous</u> ]	Οι φυσικοί πόροι εκμεταλλεύονται [ <u>Oi</u> <u>fysikoi</u> <u>poroi</u> <u>ek-</u> <u>metallevontai</u> ]	Οι φυσικοί πόροι είναι εκμεταλλεύσιμοι OR Οι φυσικοί πόροι γίνονται αντικείμενο εκμετάλλευσης [ <u>Oi</u> <u>fysikoi</u> <u>poroi</u> <u>einai</u> <u>ekmet-</u> <u>allefsimoi</u> OR <u>Oi</u> <u>fysikoi</u> <u>poroi</u> <u>ginontai</u> <u>an-</u> <u>tikeimeno</u> <u>ekmetallefsis</u> ]
<u>They exploit natural resources</u>	It is an unnatural / strange sentence	Natural resources are exploitable OR Natural resources become object of exploitation

# Resolving Weaknesses by NLPs auditing LLMs

## Deponent Verbs to passive voice

- Rule-based NLP/NLUs are the solution. There are specific rules for converting sentences from active syntax to passive syntax when the verb of sentence is an ‘αποθετικό’ [apothetiko] (deponent) verb. LLMs are not able (as our examples prove) to learn such rules but Rule-based (or True) NLP/NLUs have this knowledge (Rules existing in the literature which have been encoded to computer programs by humans). We can mention one such (of the available rules):
  - Auxiliary verb (είμαι [eimai] / γίνομαι [ginomai]) + adjective form of the verb in passive
  - Examples are:
    - αντιλαμβάνομαι → γίνομαι αντιληπτός [antilamvanomai → ginomai antiliptos]
    - αντιλαμβάνονται → γίνονται αντιληπτές [antilamvanontai → ginontai antiliptes]
    - (Greek verbs have person and number but Greek adjectives have case, number and sex)
- Consequently, in its simplest way, Rule-based (or True) NLP/NLUs can suggest (send prompts) to LLMs of what auxiliary verb (είμαι / γίνομαι) and which adjectival form of the verb to be used for the passive syntax conversion of a sentence having ‘αποθετικό’ [apothetiko] (deponent) verb.

# Rule-based approaches (including NLU) for Auditing Generative AI (Including LLMs)

- Digitize your/their data
- Transform your/their explicit and tacit/implicit knowledge into rule-based knowledge
- Impose your rules (rule-based knowledge) for Auditing Generative AI systems

Closing

# Most relevant work of the instructor

- “Large Language Models versus Natural Language Understanding and Generation”, PCI’2023 conference. ACM ICPS, February 2024.  
<https://doi.org/10.1145/3635059.3635104>
- “Strengths and Weaknesses of LLM-Based and Rule-Based NLP Technologies and Their Potential Synergies”, MDPI Electronics *14*(15), 2025.  
<https://doi.org/10.3390/electronics14153064>
- “Understanding and Establishing Natural Language Understanding & Natural Language Generation Technologies for reliable, ethical AI systems that don't impoverish the human languages”, BCI’2025 conference. Springer CCIS 2900, July 2026.  
DOI: 10.1007/978-3-032-24513-7\_20  
<https://users.uniwa.gr/nnk/papers/EstablNLUandNLGTechForSysNotImpoverLang.pdf>

# THANK YOU!

- ANY QUESTIONS

- CONTACT

Nikitas N. Karanikolas

University of West Attica (UniWA)

Department of Informatics and Computer Engineering

[nnk@uniwa.gr](mailto:nnk@uniwa.gr)

<http://users.uniwa.gr/nnk/>

# Appendix A: Morphology

# Morphology

- Morphology – the internal structure of words
- Morphology is the study of the internal structure of words and forms a core part of linguistic study today.
- The term morphology is Greek and is a makeup of morph- meaning 'shape, form', and -ology which means 'the study of something'.

# Word

- Words are the smallest independent units of language
  - do not depend on other words.
  - can be separated from other units
  - can change position.
- Example: The man looked at the horses.
  - s is the plural (morphology) marker, dependent on the noun horse to receive meaning.
  - Horses is a word. It is inside of a prepositional phrase (at the horses). It can occur in other positions or stand on its own

# Words and Morphemes

- Other position of word Horses:  
The horses looked at the man. (It is at the Subject position.)
- On its own:  
What is the man looking at? – Horses.
- **Morphemes are the building blocks of morphology**
  - Words have internal structure: built of even smaller pieces
- SIMPLE WORDS: Don't have internal structure (only consist of one morpheme) eg work, build, run. They can't be split into smaller parts which carry meaning or function.
- COMPLEX WORDS: Have internal structure (consist of two or more morphemes) eg worker: affix -er added to the root work to form a noun.

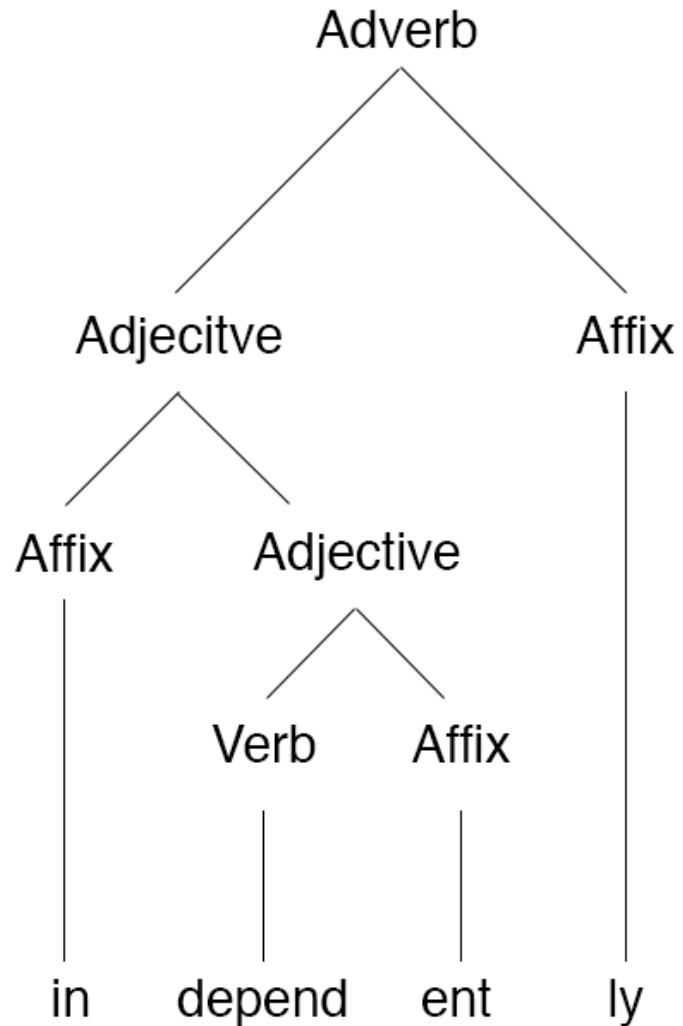
# Word, Lexeme and Word form

- The term **word** has no well-defined meaning. Instead, two related terms are used in morphology: **lexeme** and **word-form**.
- Generally, a lexeme is a set of inflected word-forms that is often represented with the **citation form** in small capitals. For instance, the lexeme eat contains the word-forms *eat*, *eats*, *eaten*, and *ate*. *Eat* and *eats* are thus considered different words-forms belonging to the same lexeme eat.
- *Eat* and *Eater*, on the other hand, are different lexemes, as they refer to two different concepts. Thus, there are three rather different notions of term word.

# Inflection vs. word formation

- Given the notion of a lexeme, it is possible to distinguish two kinds of morphological rules. Some morphological rules relate to different forms of the same lexeme; while other rules relate to different lexemes.
- Rules of the first kind are **inflectional rules**, while those of the second kind are rules of **word formation**.
- The generation of the English plural *dogs* from *dog* is an inflectional rule, while compound phrases and words like *dog catcher* or *dishwasher* are examples of word formation.
- Informally, word formation rules form new words (more accurately, new lexemes), while inflection rules yield variant forms of the same word (same lexeme).

# A morphology tree



- In, ent and ly are morphemes
- Depend (adj),  
Independ (adj),  
Independent (adj),  
and  
Independently (adverb)  
are lexemes

# Why Morphology is needed for NLU?

- Part of speech tagging:  
Noun (N),  
Verb (V),  
Adjective (Adj),  
Adverb (Adv).
- Reducing the resources (lexicon entries) needed:  
For instance, we keep only the word-form retrieve and the system is able to conclude the other word-forms (retrieves, retrieved, retrieving) that belong to the same lexeme.