

# NLP

Natural Language Processing  
Spracovanie prirodzeného jazyka

ZS 2019/2020

Ing. Miroslav Blšták, PhD.

UISI FIIT STU

# Spracovanie prirodzeného jazyka vs. spracovanie textu

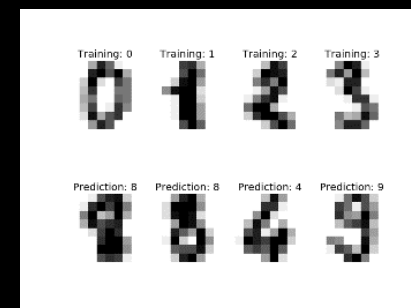
- **text** = reprezentácia (prirodzeného) jazyka
- iné reprezentácie: zvuk (reč), vizuálne (signály, posunková reč)
- prevod medzi reprezentáciami
  - image > text: OCR (Optical Character Recognition)
  - audio > text: SR (Speech recognition)
  - text > audio: TTS (text-to-speech)



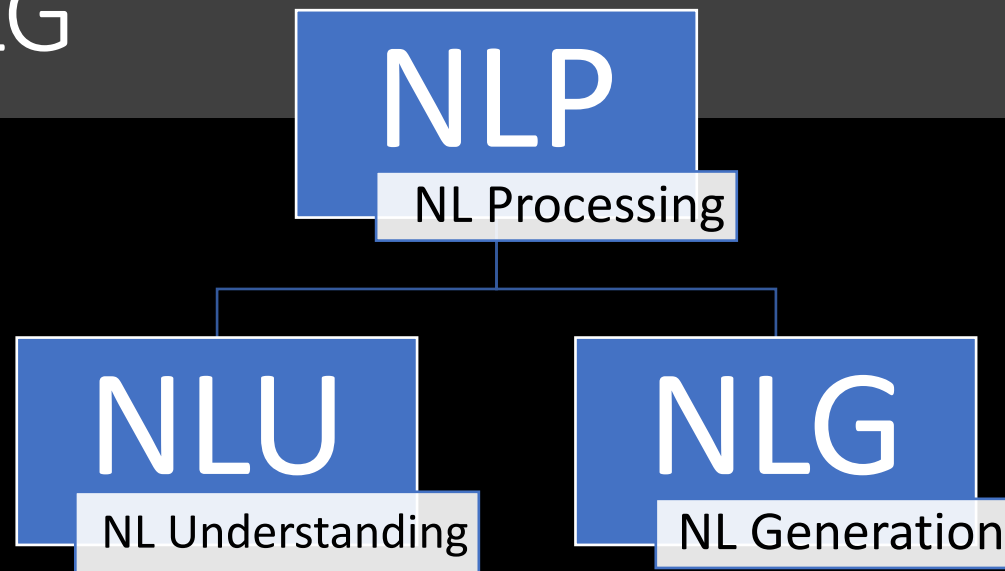
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ma est augue egestas justo, ac sagittis dolor nulla at ante. Null dolor vehicula libero, eget sollicitudin turpis arcu ut augue. et, scelerisque dolor. Pellentesque imperdiet dui at accums scelerisque ante, scelerisque gravida ligula. Fusce vitae pe sem convallis feugiat quis eu augue. Duis commodo a justo eu semper neque. Vivamus ullamcorper risus at commodo consectetur ante lobortis, tristique augue. Sed lacinia volut condimentum lacinia congue, Mauris ac tellus vel odio sus. Pellentesque sodales placerat turpis, id fermentum elit sag

## Štruktúra textu

- znaky
- tokeny (slovo/koncept)
- n-gramy
- vety
- odseky
- dokumenty (články)
- korpusy

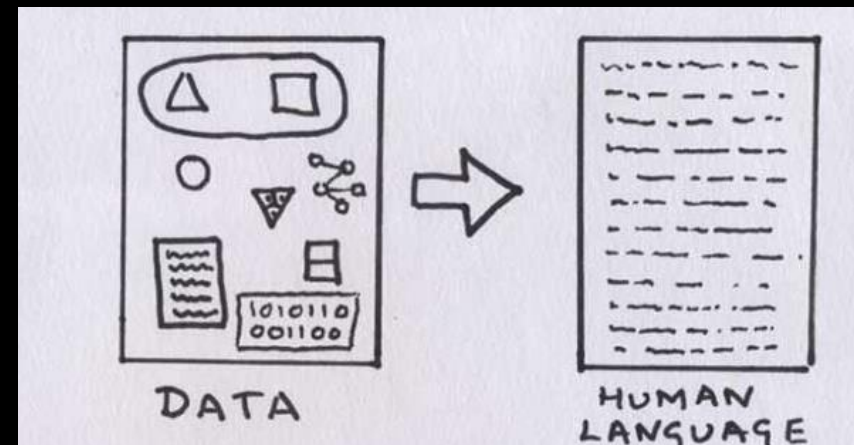


$$\text{NLP} = \text{NLU} + \text{NLG}$$

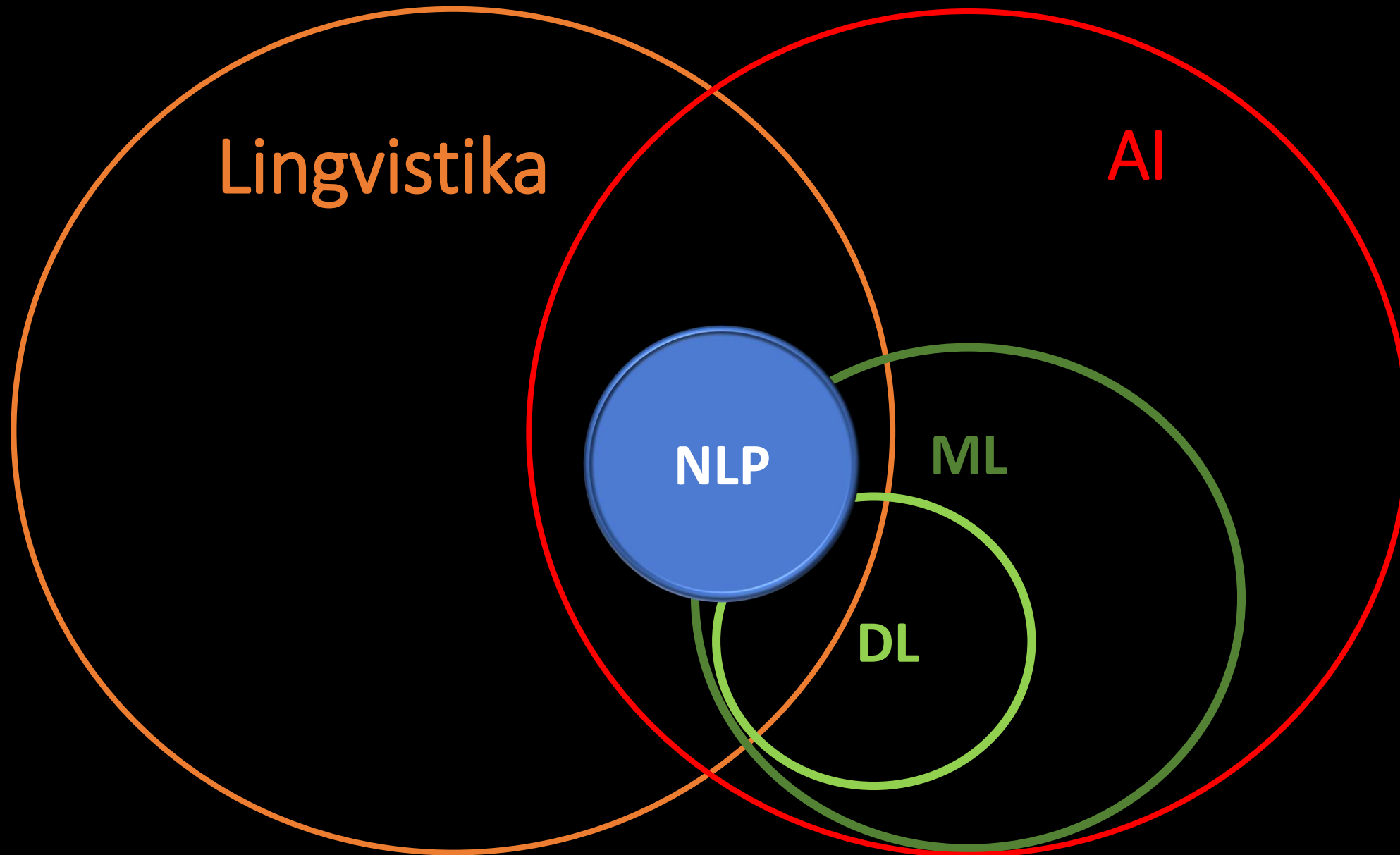


- information extraction
- sentiment analysis
- ...

- text generation
- text summarization
- ...

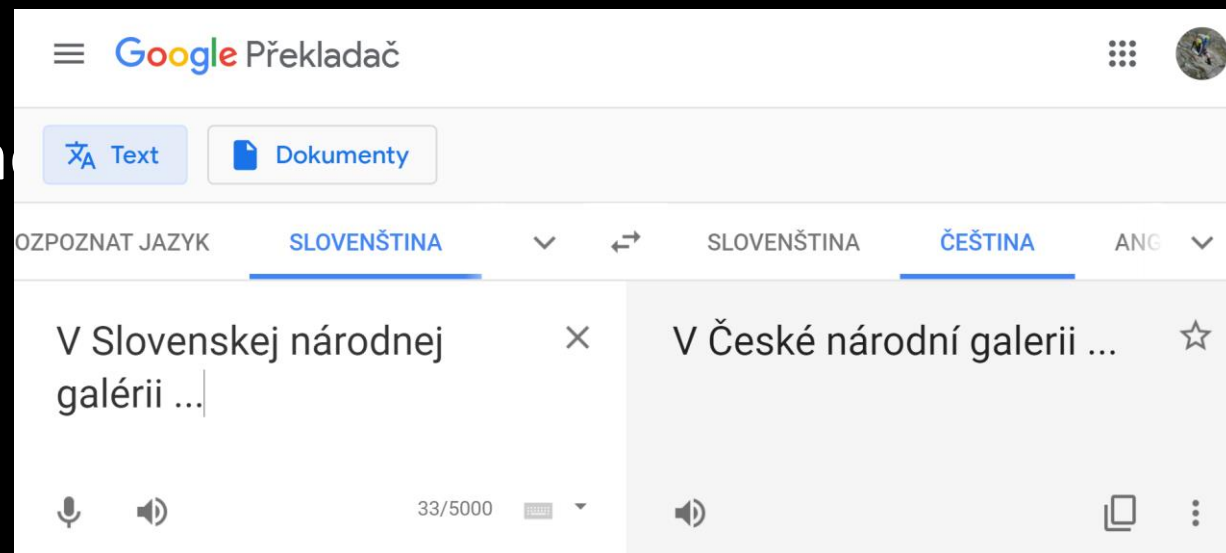


# NLP vs. AI, ML, DL



# NLP v bežnom živote

- predikcia slova, návrhy slov pri hľadani
- korekcia textu (spell-checker, “did you m
- klasifikácia textu
  - označenia mailu ako SPAM
  - zaradenie článku do kategórie
- strojový preklad
- extrakcia informácií z textu (osoby, kľúčové slová, dátum
- rozpoznávanie reči, pesničiek, ľudí, ŠPZ ...
- zaujímavé smery:
  - ML Solutions for Forensic Investigations (crime detection from text) [K. Veselovská]
  - Russian Deception Bank: A Corpus for Automated Deception Detection in Text
  - Generovanie otázok z textu



# Úlohy a ich možné zatriedenie

## Text → značka (label) klasifikácia 1/n, m/n, % [NLU]

spam detection	language identification	fake news detection	deception detection	coreference resolution
opinion mining	sentiment analysis	emotion detection	relationship extraction	information extraction
topic modeling	authorship identification	named entity recognition	negation detection	POS tagging
text similarity	paraphrase identification	word sense disambiguation	semantic role labeling	

## text → text [NLU+NLG]

summarization	sentence simplification	machine translation	chatbots	question answering	q. generation
true casing	spell-checking				

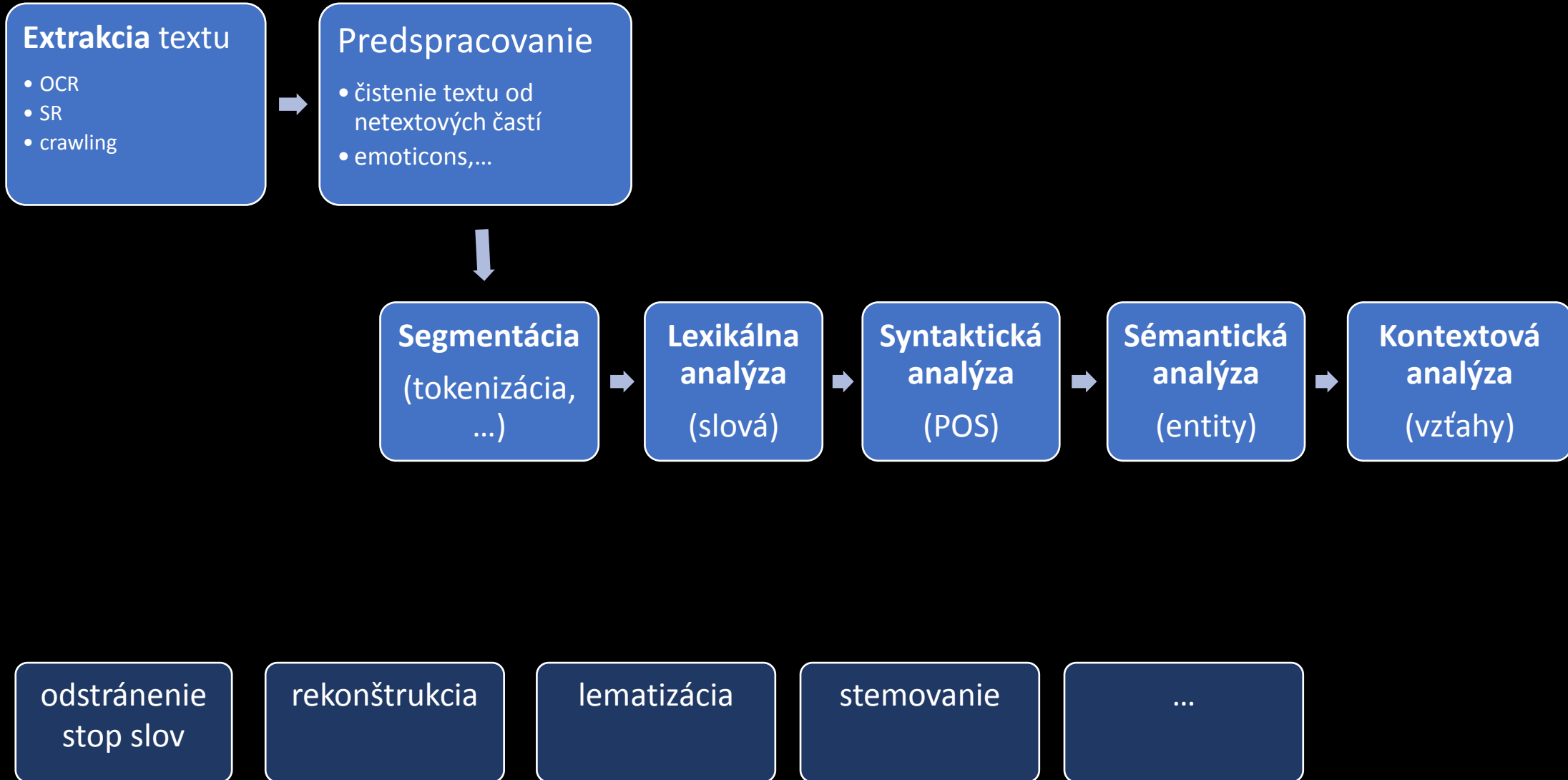
## text → tokeny (dekompozícia) [NLU]

tokenization	sentence boundary identification	phrase chunking	next-word-prediction	keywords extraction
--------------	----------------------------------	-----------------	----------------------	---------------------

## \* → text [NLG]

text generation
-----------------

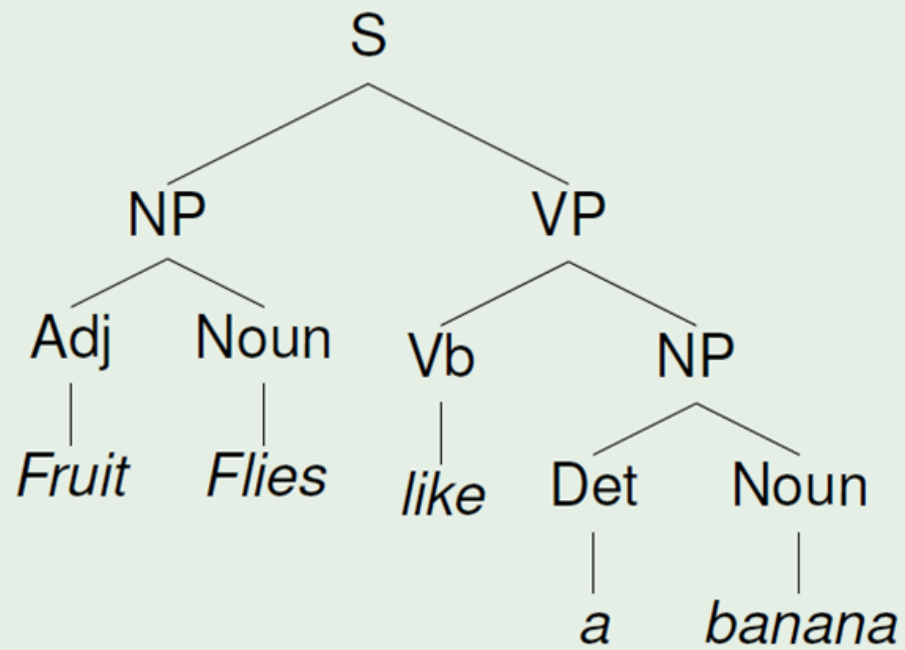
# Spracovanie textu



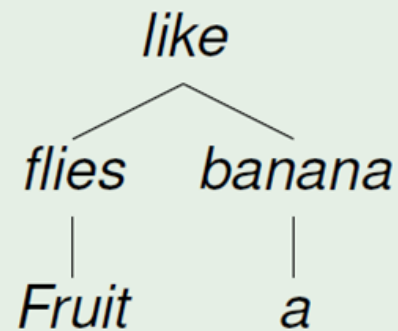
# Príklady: reprezentácia viet

Fruit flies like a banana

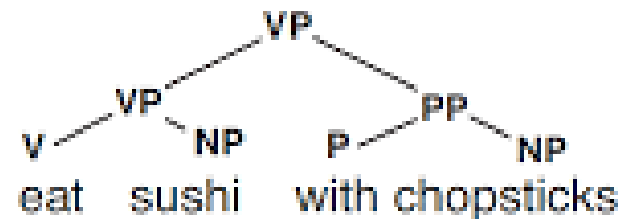
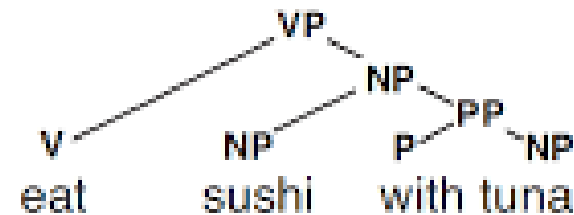
## Constituency Structure



## Dependency Structure



## Phrase structure trees



## Dependency trees





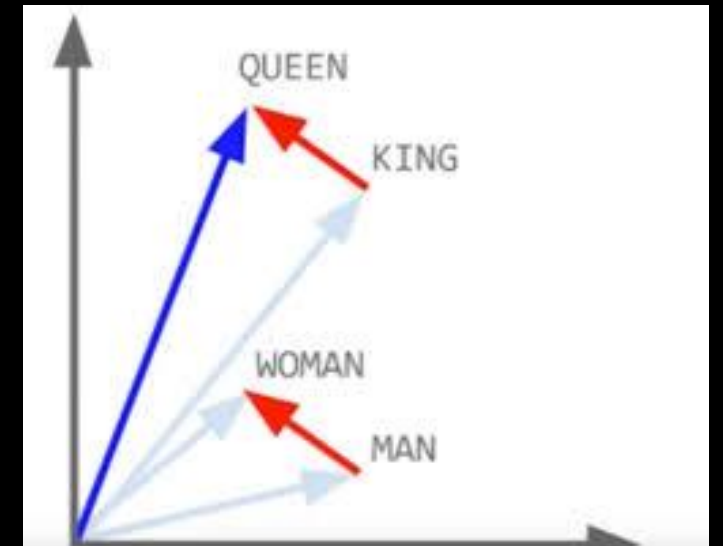
# Reprezentácia textu

- **požiadavky**

- rovnaký text = rovnaká reprezentácia (maximalizácia podobnosti)
- možnosť porovnať a usporiadať podľa podobnosti
- čo najväčší dôraz na sémantiku, čo najmenší dôraz na dĺžku

- **príklady (pojmy)**

- bag of words (BOW)
- n-gramy
- vector-space model (VSM) (prevod na čísla, matice)
  - one-hot-vector
  - LSA, SVD (Latentné črty, Singular Value Decomposition)
  - Word2Vec
- Tf-Idf, BM25 ...



# NLP přístupy

lingvistika vs. štatistika

# Klasifikácia NLP prístupov

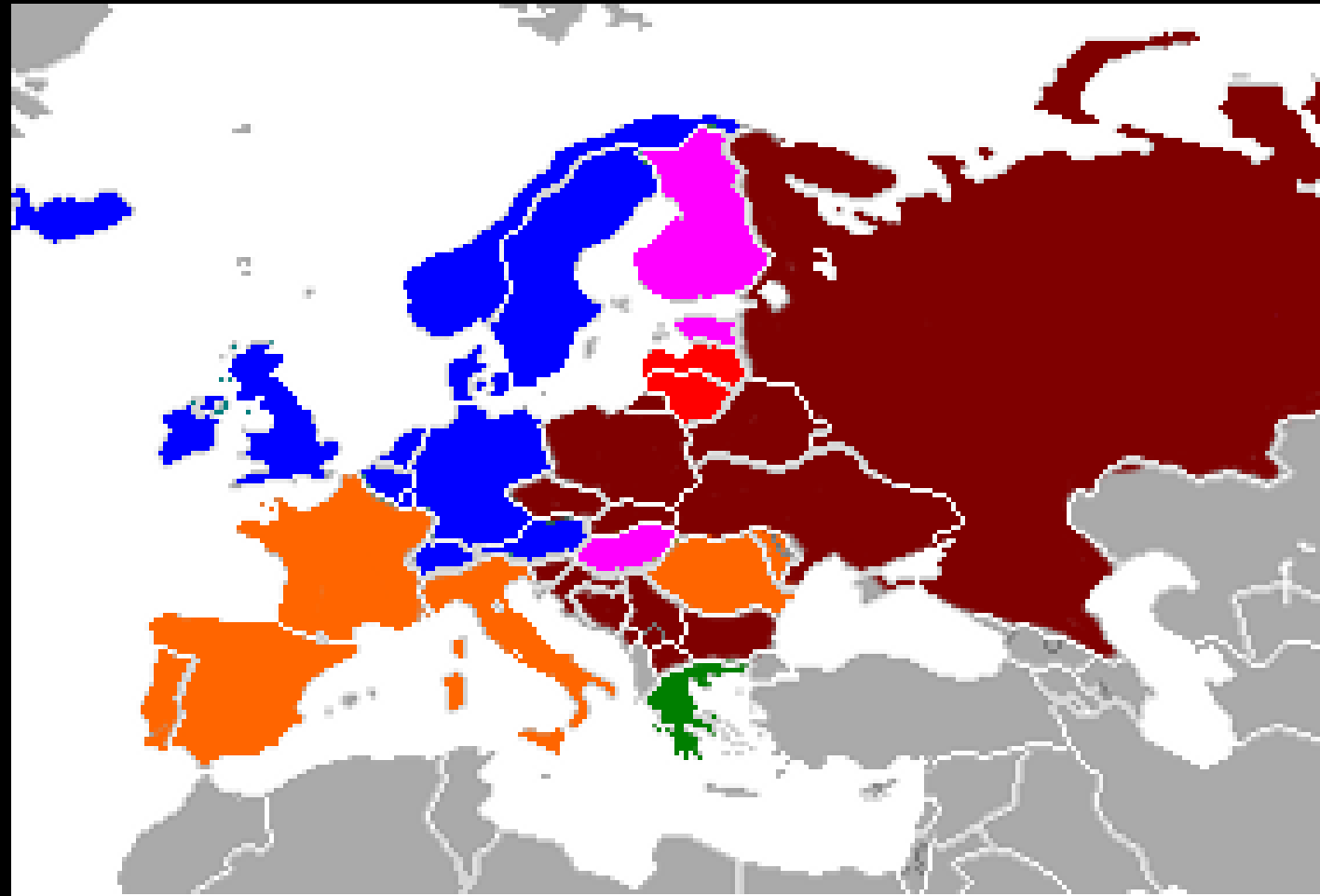
- **lingvistické vs. štatistické (AI)** (kombinácia = hybrid)
  - **lingvistika**: znalosti z pravidiel jazyka (syntax, gramatika...)
  - **štatistika**: znalosti z dát (pravdepodobnosť výskytu)
- **iné zaužívané klasifikácie**
  - rule-based vs. machine learning
  - expert-based vs. data-based (data-driven)
  - shallow learning vs. deep learning
  - ...

# Štatistika vs. lingvistika

Štatistický prístup založený na dátach (statistical approach, data-driven/data-based) metódy: štatistické, strojové učenie, hlboké učenie		Lingvistický prístup založený na expertoch (linguistic approach, expert-driven/expert-based) metódy: založené na pravidlách a vzoroch, slovníkové
dátový vedec + korpusy	<b>Vstup pre učenie</b>	jazykovedec + slovníky
používanie knižníc a funkcií	<b>IT znalosti</b>	skripty, regulárne výrazy
štatistika, dátová veda, IT	<b>znalosti</b>	jazykoveda, IT
reálne dáta, ľahšie dostupné	<b>Vstupné dáta</b>	korektné dáta, ťažšie dostupné
automaticky dopočítané	<b>Chýbajúce dáta</b>	treba doplniť
zložité	<b>Hľadanie chýb</b>	jednoduché

# Lingvistika: jazykové rodiny

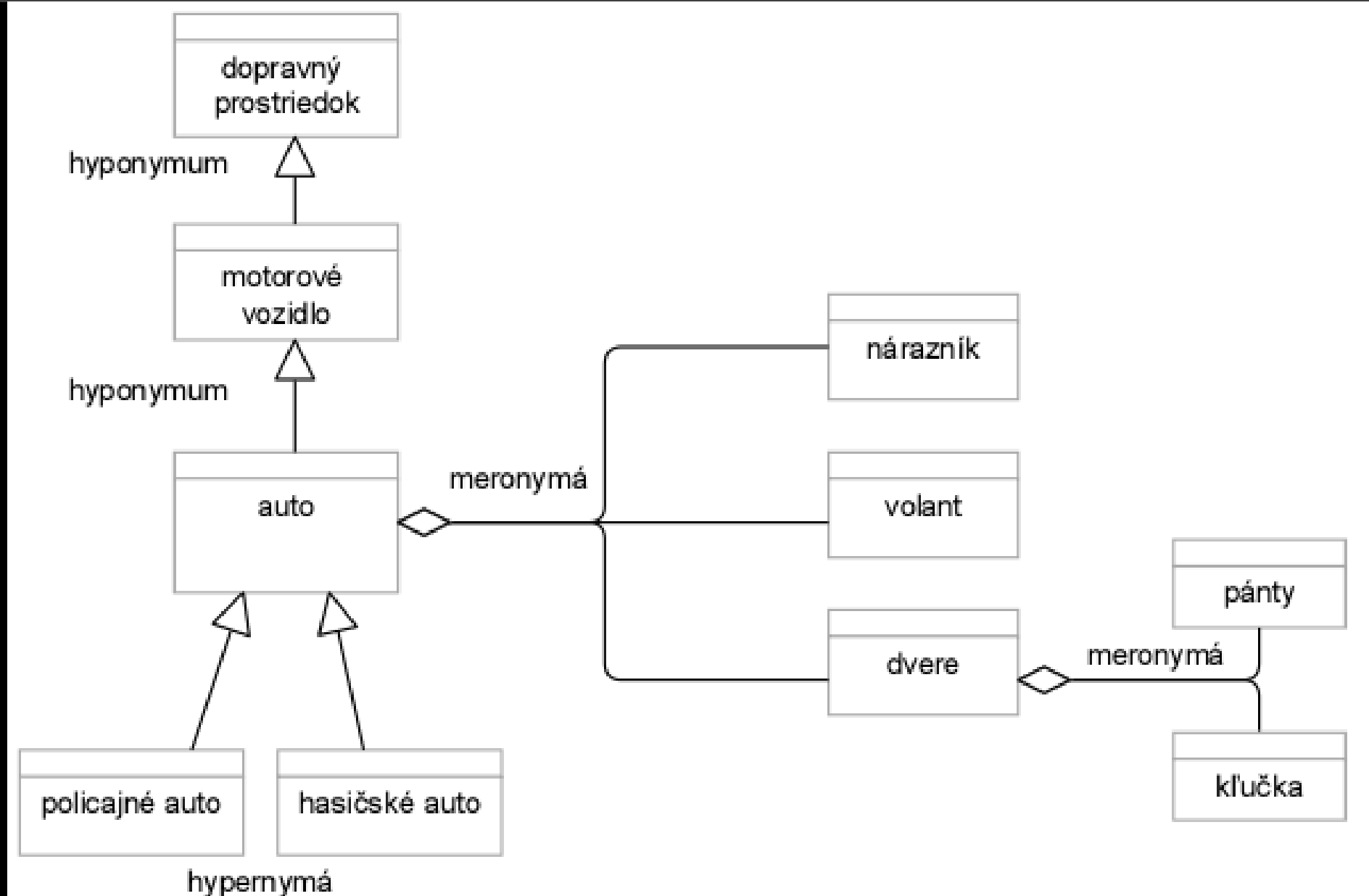
- základná slovná zásoba
- morfológia (bohatosť tvarov)
- abeceda (znaky)
- slovné druhy
- slovosled (voľný, pevný)
- spájanie slov, tvorenie viet



# Anglický jazyk - databázy slov / pravidiel / entít ...

- Vzťahy medzi slovami:
  - WordNet (EuroWordNet, SKWordNet, ...)
- Syntax a sémantika slov:
  - ConceptNet, VerbNet, FrameNet, PropBank, ...
- Word embeddings:
  - Word2Vec, Glove, ...
- Entity:
  - Viaf, GeoNames, DBpedia, Namespedia, Knowledge graph, Wikidata, ...

# Slovníky (vztahy mezi slovy) - Wordnet



# Anglický jazyk - nástroje

- Stanford CoreNLP
  - <https://stanfordnlp.github.io/CoreNLP/>
- IBM Watson (AlchemyAPI)
  - <https://www.ibm.com/watson/services/natural-language-understanding/>
- Text Razor
  - <https://www.textrazor.com/>
- Python NLTK
  - tutoriály: <https://www.guru99.com/nltk-tutorial.html>



# Slovenčina a slovné druhy

- 10 slovných druhov podľa pravidiel jazyka
- Tagset pre NLP zvyčajne obsahuje aj ďalšie:
  - participium (G): pracujúci, platiaci, nezvestný, ...
  - značky (abbreviation) (W): NATO, OSN, ...
  - ...
- príklad: <https://korpus.sk/morpho.html>
- charakteristiky jazyka:
  - voľný slovosled, bohatá morfológia
  - počet písmen: 26 vs. 46 **(1,5x viac)**
  - počet tvarov slov: cca 50 vs. cca 1000 **(20x viac)**
    - počet tvarov pre jedno prídavné meno: 3 vs. 168
  - ? o koľko dát a výpočtov viac potrebujeme na dosiahnutie úrovne spracovania angličtiny?



NLP nástroje pre slovenčinu

# NLP pre slovenčinu

- SAV <http://try.ui.sav.sk> (stav: ?)
  - FIIT <http://text.fiit.stuba.sk> (stav: aktívny)
  - TUKE <https://nlp.web.tuke.sk/pages/index> (stav: ?)
  - Filip Bednárík <http://nlp.bednarik.top> (stav: aktívny)
  - NLP4SK (stav: aktívny)
- 
- ďalšie zaujímavé odkazy na nástroje: <https://github.com/essential-data/nlp-sk-interesting-links>
  - Karlova univerzita <http://lindat.mff.cuni.cz/services/morphodita/>
  - alternatívne riešenie: použitie multi-jazykových nástrojov



[text.fiit.stuba.sk](http://text.fiit.stuba.sk)

doc. Ing. Marián Šimko, PhD.  
študentské projekty (BP, DP, OP)

# text.fiit.stuba.sk – prehľad nástrojov

- samostatné nástroje (voľne prístupné REST API)
  - tokenizátor (pravidlový)
  - segmentátor viet (pravidlový, štatistický)
  - stop slová
  - diakritikovač (štatistický)
  - korektor (štatistický)
  - lematizácia
  - stemmer
  - POS tagger (korpusový, pravidlový, štatistický CRF, štatistický ME)
  - ~~NER~~
  - ~~závislostný parser~~

text-api.fiit.stuba.sk:9000

Ing. Samuel Pecár  
doktorand na FIIT

# text-api.fiit.stuba.sk:9000

- POST <http://text-api.fiit.stuba.sk:9000/api>
- využíva nástroje z text.fiit.stuba.sk
- anotátory
  - diakritikovač
  - lematizátor
  - tokenizér
  - POS

<http://nlp.bednarik.top>

Ing. Filip Bednárík  
(absolvent FIIT)



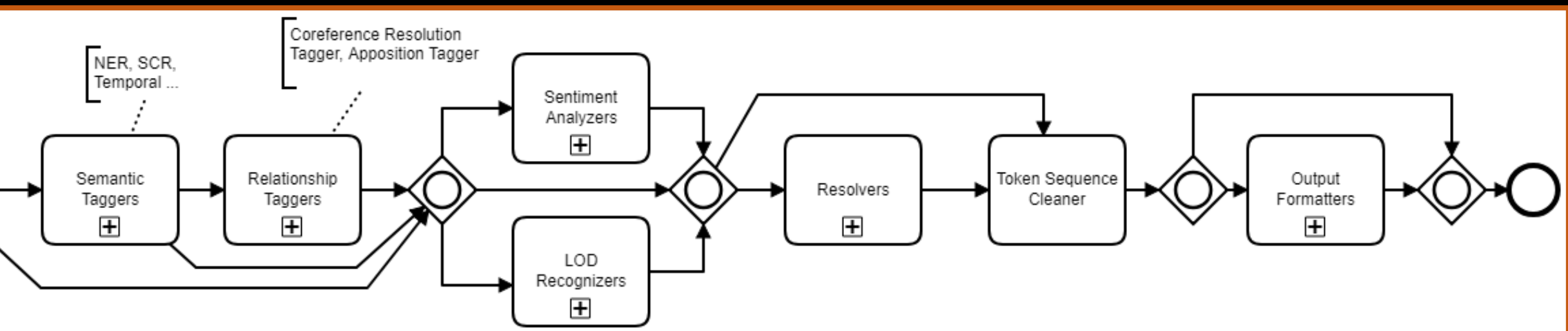
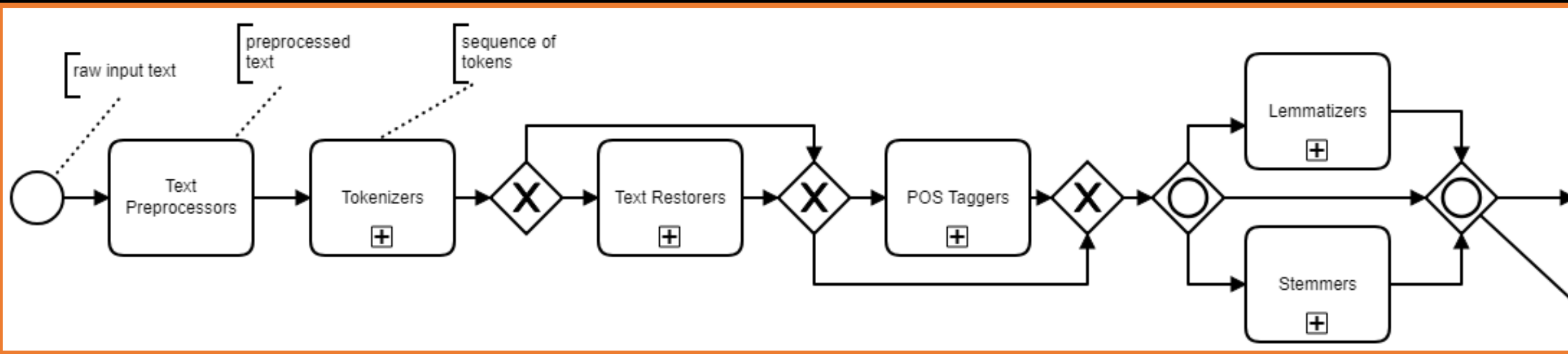
# nlp.bednarik.top – prehľad nástrojov

- rôzne nástroje (voľne prístupné REST API)
- <https://github.com/drndos/nlp-tools>
  - tokenizácia
  - segmentácia viet
  - lematizácia
  - stemovanie
  - POS tagger
  - NER
  - **[+]** syntaktická analýza (parser natrénovaný pomocou Stanford CoreNLP)
  - **[+]** anonymizácia súdnych rozhodnutí

NLP4SK

Ing. Miroslav Blšták, PhD.  
+ partnerské projekty

# NLP4SK – architektúra nástroja



# NLP4SK – prehľad nástrojov

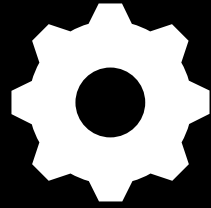
- **[+]** **Predspracovanie textu**: vyčistenie od netextových častí (napr. referencií), validácia znakov...
- **[+]** **Tokenizácia a identifikácia viet**: (**token**, **meta**)
- **Rekonštrukcia textu**: napríklad diakritiky, skratiek alebo nespisovných slov (**word\_original**)
- **Lexikálna analýza**: identifikácia slov v slovníku (**word**)
- **Lematizácia** (angl. lemmatization, **lemma**): určenie základných tvarov slov
- **Stemovanie** (angl. stemming, **stem**): určenie koreňov slov
- **POS tagger** (angl. **pos** tagging) slovný druh a gramatické kategórie
  - **[+]** **probabilistic POS tagger**
- **[+]** **Sémantická analýza**:
  - názvoslovných entít (angl. Named Entity Recognition, **ner**),
  - dátumov, časov, číselných údajov a pod.
  - identifikácia a extrakcia sémantických kategórií (**scr**)
- **Identifikácia sentimentu** (angl. **sentiment** analysis)
- **[+]?** **Identifikácia vzťahov medzi slovami v texte** (**ent**, **coref**)
- **LOD** (**lod**)

# NLP4SK – Response (nážorný príklad)

```
[{  
  "word"      : "škola",  
  "meta"      : "{S}",  
  "pos"       : ["SSns1", "SSns4"],  
  "lemma"     : ["škola"],  
  "stem"      : ["ško1"],  
  "ner"       : ["ORG/school"],  
  "scr"       : [],  
  "word_original" : "skola"  
}, {  
  ...  
}]
```

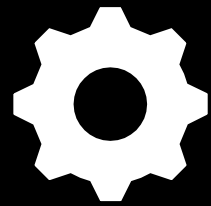
# NLP4SK – použitie

- POST: <http://arl6.library.sk/nlp4sk/api>
- obmedzenia APIKEY
  - použitie modulov
  - počet dopytov za sekundu
  - veľkosť vstupného textu (v znakoch)
  - IP filtrovanie
  - logovanie
- ARL6 webapi <http://arl6.library.sk/nlp4sk>
  - online dokumentácie
  - demo



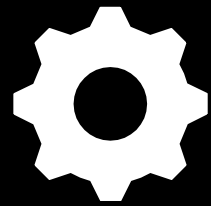
## úvod do NLP

- pojmy a zaradenie NLP
- úlohy NLP
- reprezentácia textu



## NLP prístupy

- lingvistické
- štatistické
- zdroje dát



## nástroje pre slovenčinu

# Zaujímavé odkazy

- ML Solutions for Forensic Investigations (crime detection from text) [Kateřina Veselovská]  
[https://www.youtube.com/watch?v=NGVbmvMfR\\_w](https://www.youtube.com/watch?v=NGVbmvMfR_w)
- <https://medium.com/sciforce/a-comprehensive-guide-to-natural-language-generation-dd63a4b6e548>
- <http://www.butleranalytics.com/natural-language-generation-explained/>
- <https://devopedia.org/natural-language-processing>
- **NLP Tutorial AI with Python** <https://medium.com/@rinu.gour123/nlp-tutorial-ai-with-python-natural-language-processing-ed81fdb3f0a3>
- **Russian Deception Bank: A Corpus for Automated Deception Detection in Text**  
<https://www.tsdconference.org/tsd2016/download/cbblr16-725.pdf>
- Nettle D. (2012). Social scale and structural complexity in human languages. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 367(1597), 1829–1836. doi:10.1098/rstb.2011.0216
- <https://whispertrouble.com/fascinating-map-shows-how-long-itll-take-you-to-learn-a-foreign-language/>