Free Software Tools for Computational Linguistics: An Overview

Miloš D. Đurić

Faculty of Electrical Engineering, University of Belgrade

The Third PSSOH Conference: Application of Free Software and Open Hardware

Belgrade, 24th October, 2020

Introductory Remarks

- Computational Linguistics (CL)
- CL research community
- Free software tools for CL
- Towards a unitary descriptive account of free software tools for CL
- The aim: to fill the lacuna in the current scholarship on free software tools in CL
- Praat, KH Coder, NLTK

Theoretical Underpinnings

- (Transformational-)Generative Grammar ((T)GG) (Noam Chomsky, 1965)
- Relevance Theory (RT) (Dan Sperber & Deirdre Wilson,1995)
- Optimality Theory (OT) (René Kager, 1999; Alan Prince & Paul Smolensky, 1993)
- Asymmetry Theory (AT) (Anna Maria Di Sciullo, 2005)
- The Minimalist Program (Chomsky, 1995)

Relevance-Theoretic Framework

The first (cognitive) principle of relevance

Human cognitive processes are aimed at processing the most relevant information available in the most relevant way. (Sperber & Wilson, 1995: 260).

The second (communicative) principle of relevance

Every act of ostensive communication conveys a presumption of its own optimal relevance. (Sperber & Wilson, 1995: 260).

'Principe de pertinence'

Tout acte de communication ostensive communique la présomption de sa propre pertinence optimale.

A cognitive inferential account of human communication (Sperber and Wilson, 1986; Robyn Carston, 2002)

An eclectic model

GOING RADICALLY SEMANTIC

The Natural Semantic Metalanguage (NSM)

GOING RADICALLY PRAGMATIC

Relevance Theory (RT)

Semantics (Decoding) vs. Pragmatics (Inference)

Relevance – two-pronged property, a cognitive trade-off

Conceptual encodings vs. Procedural encodings

Optimality-Theoretic Framework

Markedness vs. Faithfulness

(Kager, 1999)

Corpora

- How much is enough?
- Representativeness

The Brown Corpus of Standard American English (**BCSAE**) – 1.000.000 words; defined as "the first modern, electronically readable corpus".

The Lancaster-Oslo-Bergen Corpus (**LOB**) 1.000.000 words (500 texts of 2,000 words each).

The British National Corpus (**BNC**). Some 100 million words. It contains both written and spoken material.

The Santa Barbara Corpus of Spoken American English (**SBCSAE**) 249,000 words.

Praat

- Compounds
- Complex Nominals (Judith N. Levi, 1978)
- Noun Sequences (Lucretia Vanderwende, 1994)
- Complex Constructs…

“Blackbird Pattern”

- ‘blackbird vs. black ‘bird
- Ice cream (Bloomfield, 1933)
- Ginger ale, chicken salad (Aronoff & Fudeman, 2007)

Dialectal variation

American English vs. British English

BOY scout vs. boy **SCOUT**

ICE cream vs. ice **CREAM**

Variation in similar structure cases

APPLE cake vs. apple **PIE**

LEMON cake vs. lemon **PIE**

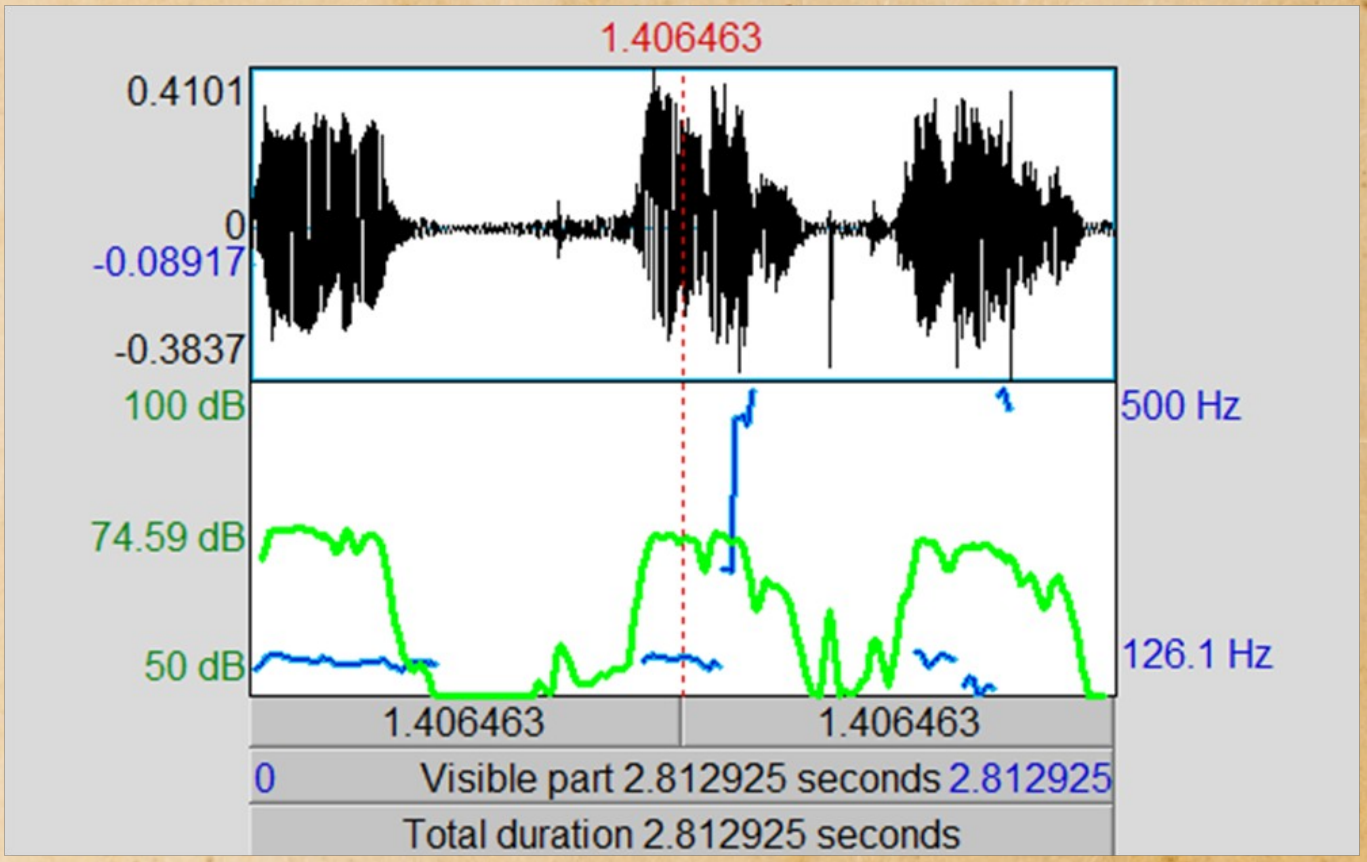
This phenomenon labelled as "family constituent bias" (Plag et al., 2006)

Multi-Constituent Constructs (MCCs) in Electrical Engineering Discourse (EED) and Computer Science Discourse (CSD)

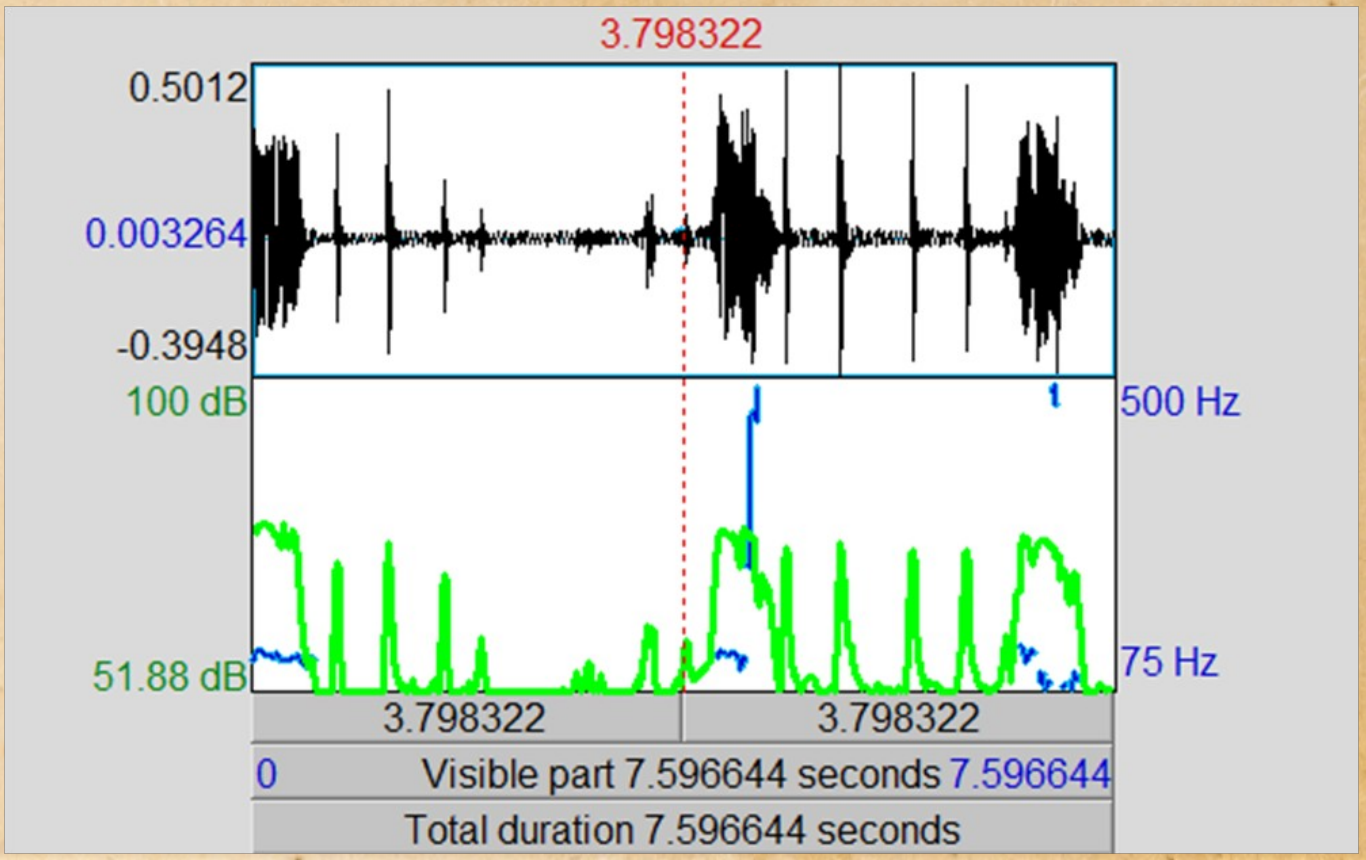
- **Bridge**
- **Diode bridge**
- **Diode bridge rectifier**
- **Three-phase diode bridge rectifier ….**

MCC “random number generator”

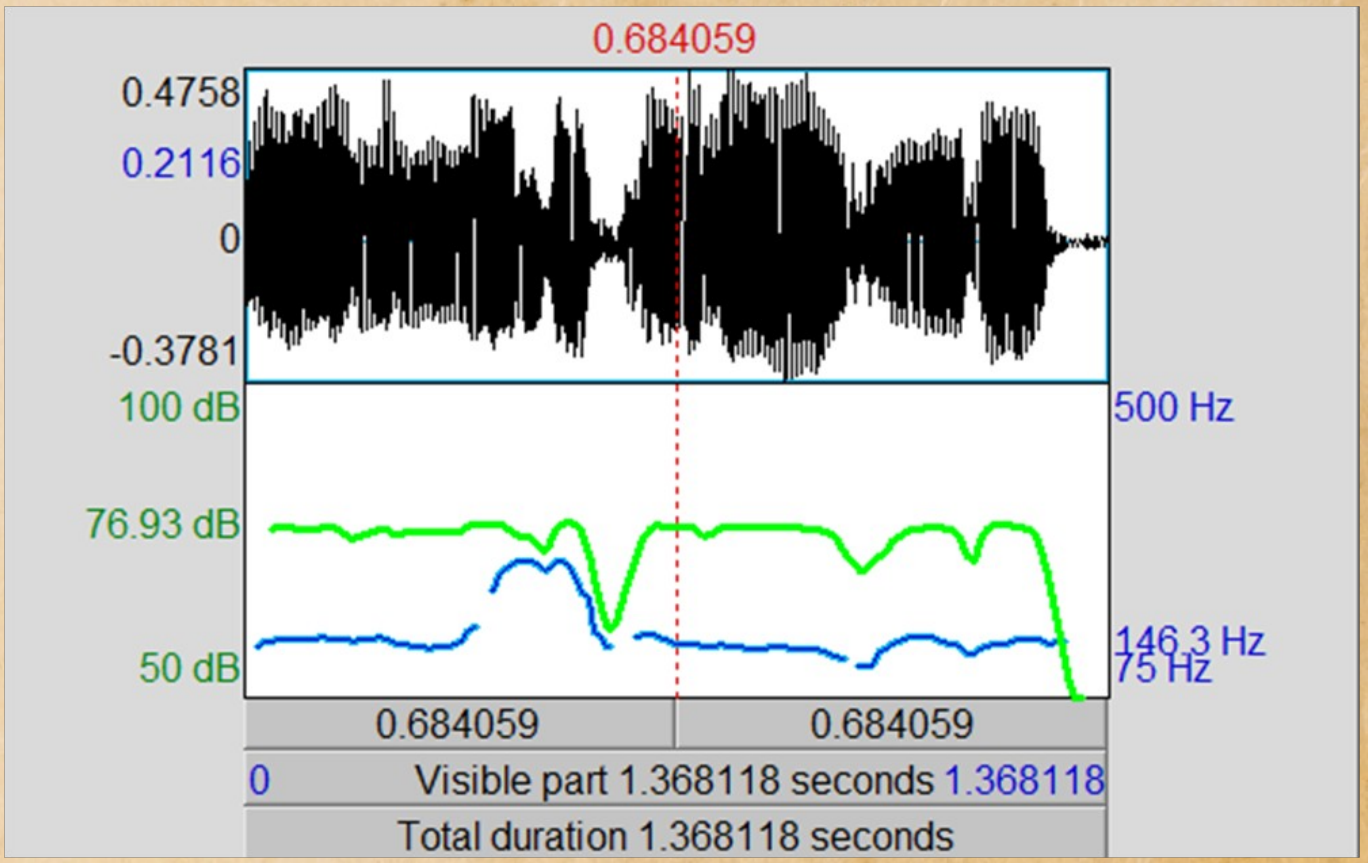
The Praat-generated token 1 of the MCC “random number generator” from my corpus.



The Praat-generated token 2 for MCC “random number generator”



The Praat-generated token 3 of the MCC "random number generator" from my corpus



The case of intra-speaker variation

- Discourse semantics
- A tentative conclusion

The MCC in conclusive, generic meaning

KH Coder

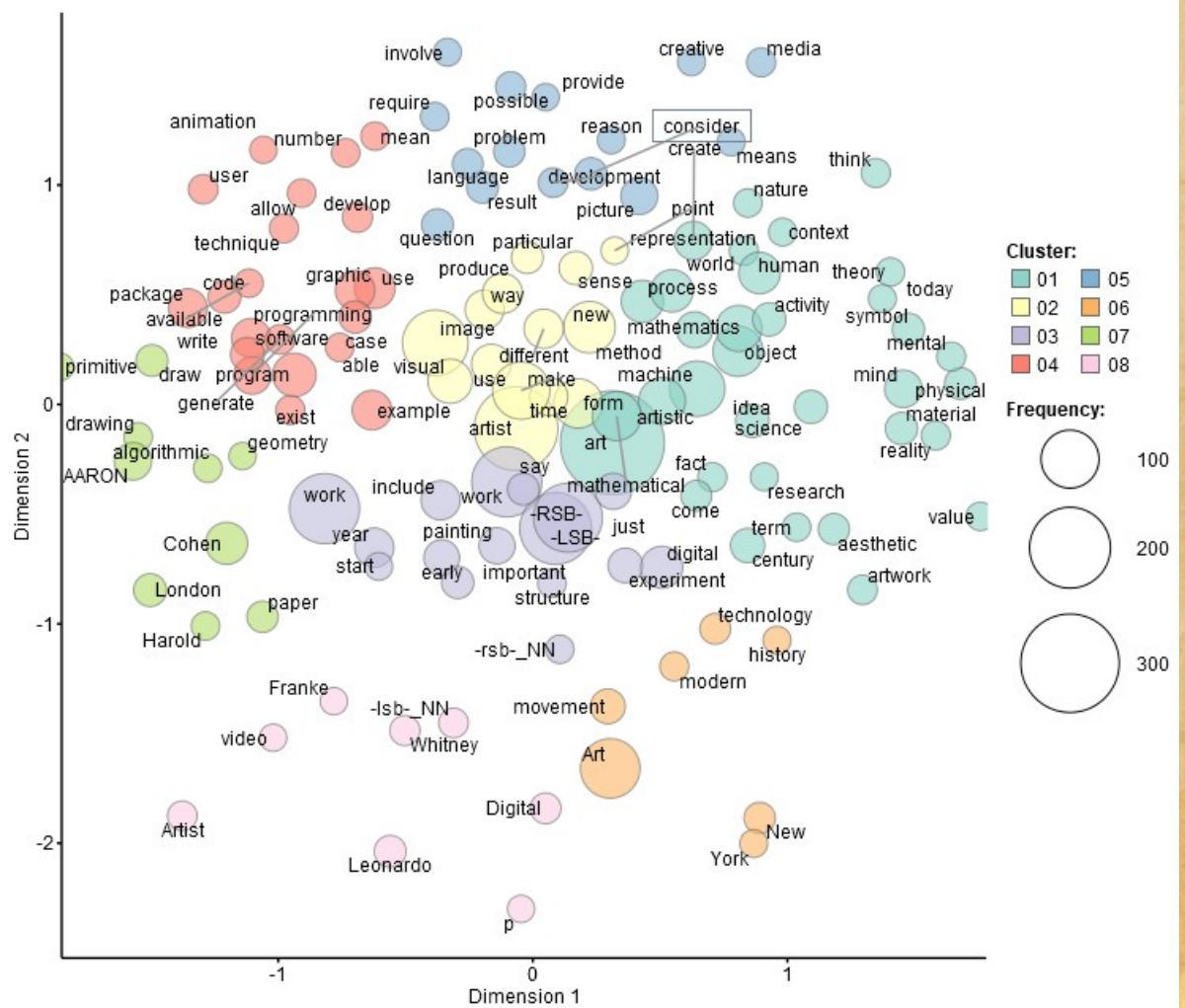
A free software tool for quantitative content analysis or text mining, and it is also utilised for computational linguistics.

Multi-dimensional scaling

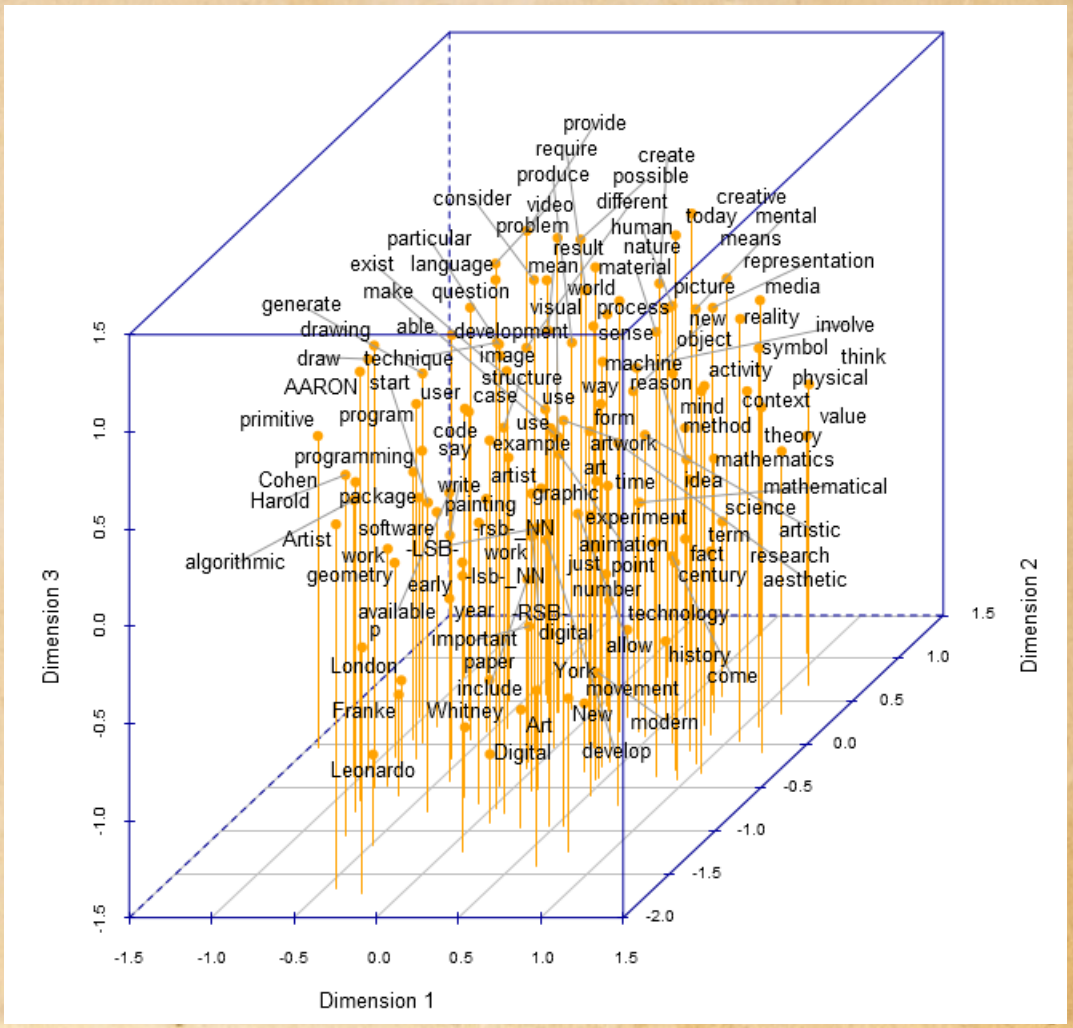
Cluster analysis

Co-occurrence network

The two-dimensional solution for non-metric multidimensional scaling (2D Cruscal) for the text excerpt from my DAM corpus.



The three-dimensional solution for non-metric multidimensional scaling (3D Cruscal) for the text excerpt from my DAM corpus.



The KH Coder-generated illustrative table for the text excerpt from my La TurboAvedon corpus.

Lexical items	Part of Speech	Frequency
AVEDON	ProperNoun	21
space	Noun	17
LATURBO	ProperNoun	16
work	Noun	16
virtual	Adj	10
New	ProperNoun	8
artist	Noun	8
live	Verb	8
avatar	Noun	7
consider	Verb	7
experience	Noun	7
media	Noun	7
paraspace	Noun	7
production	Noun	7
sculpture	Noun	7
surface	Noun	6
Sculpt	ProperNoun	5

Potential Challenges

- Orthography

  flower-bed

  flower bed
- flowerbed??
- Phone box / phone-box / phonebox

  items meaning “telephone call box” / “telephone booth”
- User-friendly (BrE) vs. user friendly (AmE)
- Letter box…

- Highly-frequent content words
- **But, what about Discourse Markers (DMs)?**

Diagnostic Tests according to Schourup (1999)

1. **Connectivity**
2. **Optionality**
3. **Non-truth-conditionality**
4. **Phonological independence**
5. **Initiality**
6. **Orality**
7. **Multi-categoriality**

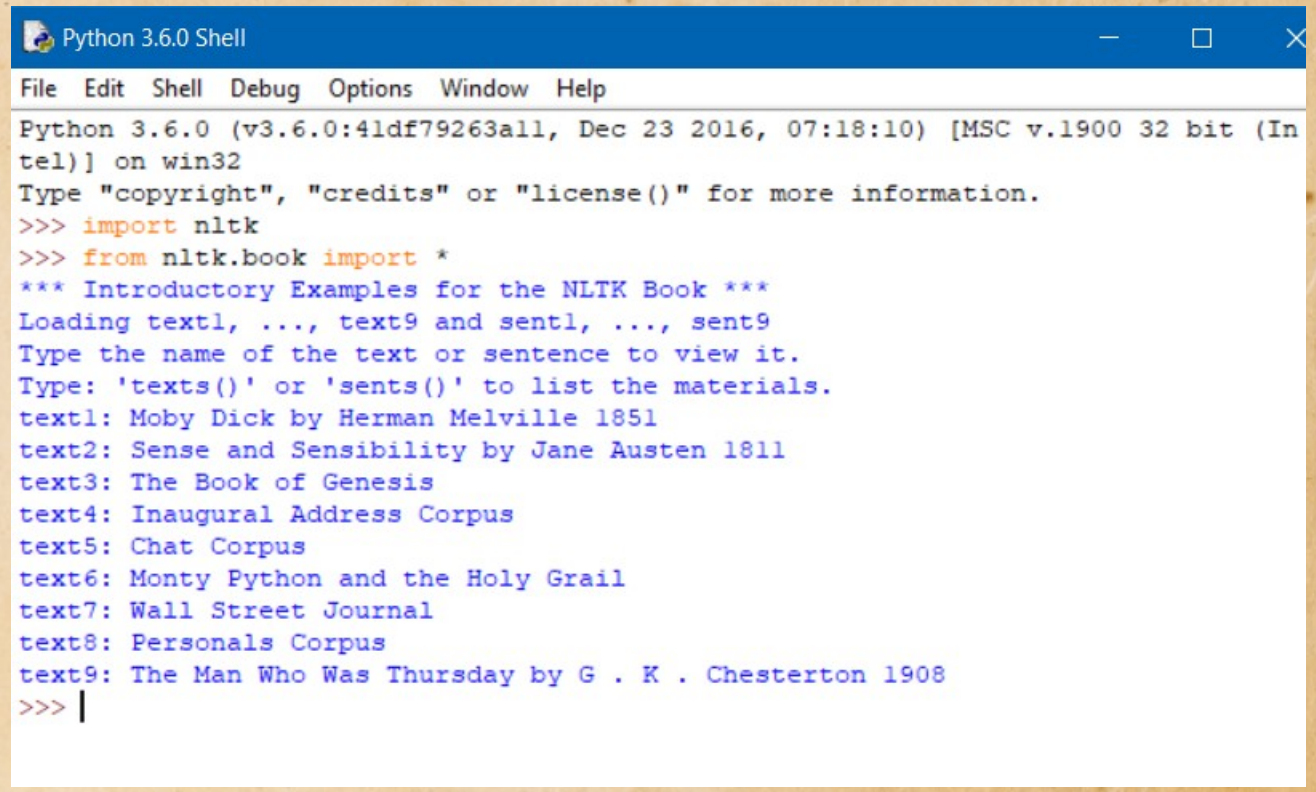
approximation, minimisation, tentativisation….

The Natural Language Toolkit (NLTK)

A collection of libraries and programs for symbolic and statistical NLP written in the Python programming language

Computational Text Analysis (CTA)

My screen capture of an illustrative example of the NLTK corpus structure.



My screen capture of the NLTK-generated concordance of the lexical item “lucky” from the first NLTK corpus.

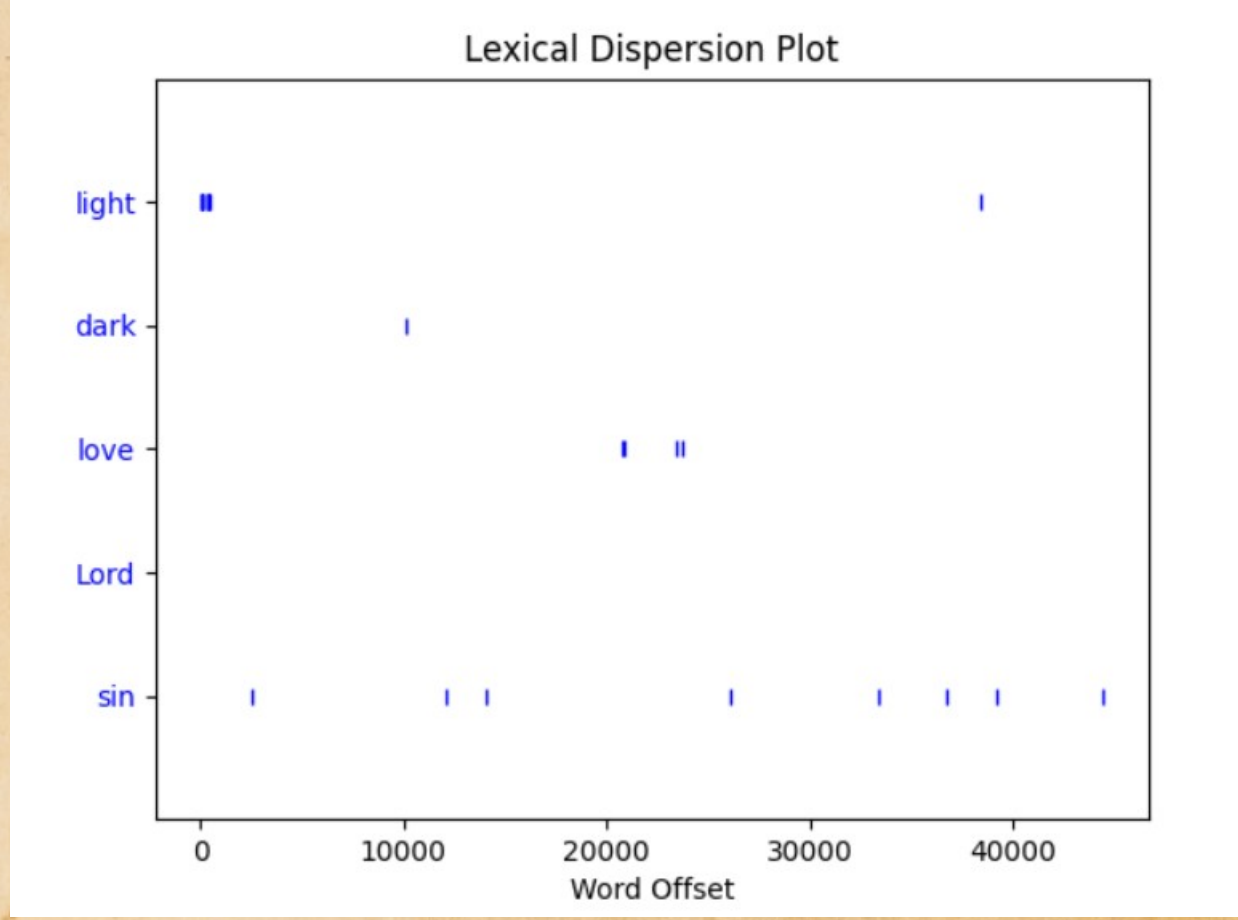
```
>>> text1.concordance("lucky")
Displaying 8 of 8 matches:
etter than nothing ; and if we had a lucky voyage , might pretty nearly pay for
 a Cape - Cod - man . A happy - go - lucky ; neither craven nor valiant ; takin
fore the wind . They are accounted a lucky omen . If you yourself can withstand
l heights ; here and there from some lucky point of view you will catch passing
olently making for one centre . This lucky salvation was cheaply purchased by t
h Sea . The voyage was a skilful and lucky one ; and returning to her berth wit
eat skull echoed -- and seizing that lucky chance , I quickly concluded my own
I ' ll be ready for them presently . Lucky now ( SNEEZES ) there ' s no knee -
>>>
```

My screen capture of an illustrative example of vocabulary counting of NLTK corpora.

```
>>> len(text2)
141576
>>> len(text3)
44764
>>> len(text4)
149797
>>>
```

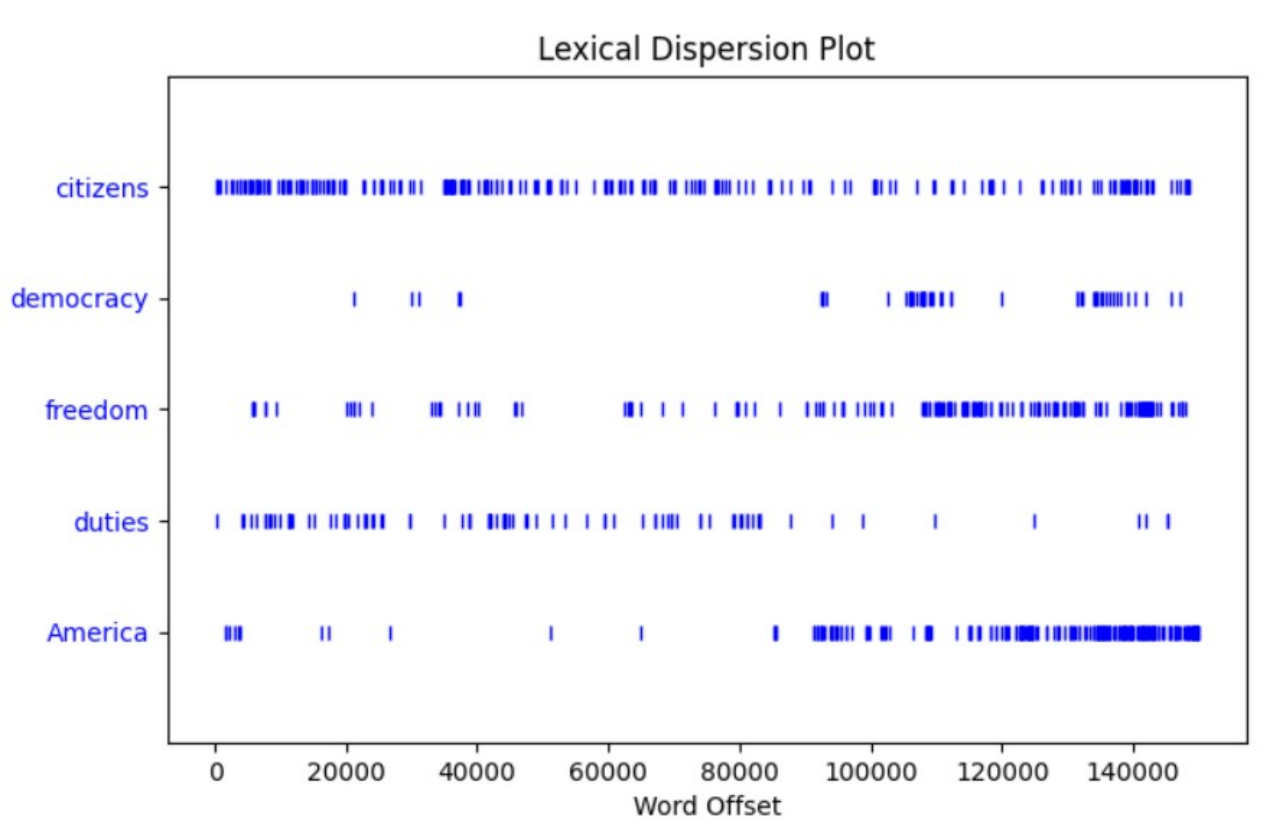
My screen capture of an illustrative example written in Python in order to obtain the lexical dispersion plot for NLTK corpus 3 (i.e. The Book of Genesis).
My screen capture of the actual lexical dispersion plot for the NLTK corpus 3 (i.e. The Book of Genesis) generated by the NLTK tool.

```
>>> text3.dispersion_plot(["light", "dark", "love", "Lord", "sin"])
```



My screen capture of the command line written in Python in order to obtain the lexical dispersion plot for the NLTK corpus 4 (i.e. Inaugural Address Corpus).
My screen capture of the lexical dispersion plot for the NLTK corpus 4 (i.e. Inaugural Address Corpus) generated by the NLTK tool.

```
>>>
>>> text4.dispersion_plot(["citizens", "democracy", "freedom", "duties", "America"])
```



My screen capture of the POS-tagger processing an illustrative utterance from my corpus (i.e. *The Ninth Gate Corpus*).

```
>>> text = word_tokenize("Andrew Telfer is writing a note at his desk in one
corner of a big, book-lined room")
>>> nltk.pos_tag(text)
[('Andrew', 'NNP'), ('Telfer', 'NNP'), ('is', 'VBZ'), ('writing', 'VBG'), ('
a', 'DT'), ('note', 'NN'), ('at', 'IN'), ('his', 'PRP$'), ('desk', 'NN'), ('
in', 'IN'), ('one', 'CD'), ('corner', 'NN'), ('of', 'IN'), ('a', 'DT'), ('bi
g', 'JJ'), (',', ','), ('book-lined', 'JJ'), ('room', 'NN')]
>>>
```

The Comparison of the Selected Free Software Tools for CL

- Visualisation
- User-friendliness
- Limitations
- Point of departure
- The ready-made language data
- The user

Concluding remarks

- Re-examination of free software tools in CL from a comparative perspective
- Utilising already available corpora
- Broadening appealing dimensions of CTA

Acknowledgements

- I should like to express my gratitude to **Professor Dr. Nadica Miljković** (Faculty of Electrical Engineering, University of Belgrade) who made me experience the excitement of going beyond the secure limits of philology and linguistics and for making my CL-Analysis pipe dreams come true. Professor Nadica Miljković offered more than valuable suggestions and ideas.
- I owe a debt of gratitude to **Professor Dr. Predrag Pejović** (Faculty of Electrical Engineering, University of Belgrade) for his patience of a saint and academic generosity whilst sharing his expertise, wisdom and academic kindness. My gratitude goes to Professor Predrag Pejović for providing me with inspiration and offering his most constructive scrutiny.
- My Python-motivated tasks have been eased by **Miss Sanja Delčev, TA** (Faculty of Electrical Engineering, University of Belgrade) who patiently answered a number of my questions and revealed the charm of the PyScripter for the Python Programming Language.

Thank you for listening!