## Grammar Versus Semantics

(Original version by Carlo A. Trugenberger, published by Zeno R.R. Davatz yweesee GmbH, abridged by Paul Wälti on 23. February 2023)

*Current mainstream methods focus on grammar*

While there are many excellent tools for analyzing structured, numerical data, unstructured text is the holy grail of modern artificial intelligence (**AI**) research. In the past, the mainstream focus has been on natural language processing (**NLP**), which aims to teach machines how to correctly linking of words. Essentially, this amounts to teaching machines the grammar of a language. Generative language models like **ChatGPT** are perfectly trained on the grammar of a language and write excellent prose, but cannot understand the content of a text.

NLP and ChatGPT require extensive training, and both deal only with the relative positioning of groups of a few words. It's like studying individual trees and completely missing the forest. They can give nonsensical answers and have an extremely large bias (distortion of reality, false conclusions). This is because grammar has very little to do with the meaning of a text.

*InfoCodex's semantic technology focuses on content recognition*

We teach machines the meaning of cross-linguistic phrases grouped into semantic clouds or synonym groups. These semantic clouds are mapped to nodes in a hierarchical taxonomy structure. These nodes have a unique meaning, for example, mustang → pony → horse → mammal → etc., but the synonym group of Ford Mustang refers to the node → passenger car. The semantic structure is represented in a universal, machine-readable linguistic database. The knowledge summarized in this database comes from about 100 renowned international knowledge repositories such as the wordnet of the Princeton University.

For a given document collection, information-theoretic algorithms are used to build a mathematical model of the text in a 100-dimensional content space of the most relevant high-level concepts, and self-organizing neural networks are then used to formulate a meaning similarity of these high-level topics. This combination of **AI** and **semantics** enables applications beyond the reach of NLP and generative language models such as ChatGPT.

First and foremost, semantic technology requires no training with large document collections and can be applied immediately in new and unfamiliar situations. This is because the knowledge is already contained in the large universal linguistic database.

Second, NLP and knowledge graphs are good at retrieving known information, but fail at knowledge discovery: they can only recognize facts explicitly stated by an author. IBM Watson, for example, excels at answering questions about known topics (Jeopardy), but fails to discover new knowledge in drug discovery. Similarly, generative language models such as ChatGPT can only combine existing ideas but never capture new ones. Semantic InfoCodex technology, on the other hand, is able to discover new relationships (hidden facts) by analyzing large amounts of literature, i.e., by analyzing the "whole forest." And the researcher's goal is to discover new knowledge that is not already explicitly contained in natural language documents.

Third, InfoCodex Document Summarizer can instantly summarize even very long documents, such as research, legal, or medical reports, without requiring delaying individual training. Another example of how semantics, or meaning, is outpacing grammar.