



# Functionality and areas of operation of Natural Language Processing

---

November 2021

## Executive Summary

**Natural language processing, a sub-area of artificial intelligence, is gaining more and more importance - NLP solutions are increasingly used in popular and everyday social applications as well as in business contexts.**

The best known, for example, are solutions/applications/services such as voice-controlled assistants - be it the driving assistant in a car or Alexa, Siri and the like. Chatbots used on websites are also based on NLP. Explicit NLP applications such as intelligent document and contract management help companies to automate and optimise processes, identify business risks and tap into new revenue potential.

"NLP" is often used as a buzzword or generic term. But just as not all AI is the same, the same applies to NLP.

Basically, one can say that NLP enables computers to recognise, understand, analyse and interpret human language. Be it spoken language or written words. This can also include intentions, moods of the speaker or author. The basis for this is the combination of statistical machine learning and computational linguistics, i.e. the rule-based modelling of human language.

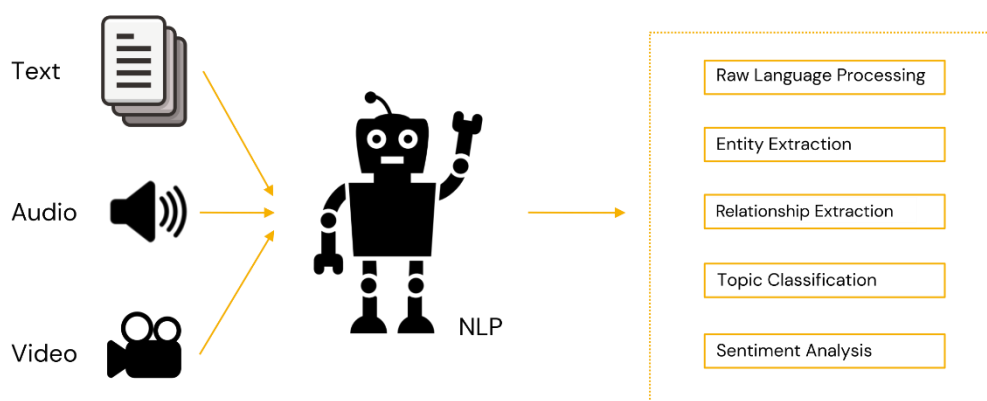
Since human language is highly complex, this makes NLP a complex field in which precise distinctions must be made. Processing human language is not new, but the methods are developing rapidly. And with it, the possibilities to process a lot of data in a very short time - be it a voice command to the navigation device or the analysis of thousands of documents.

With the continuously growing number of documents and contracts, as well as the ever-increasing expectations of customers and consumers for intelligent solutions for everyday objects, the demands on NLP solutions are also increasing, and they must be further developed in response.

## Challenge: structuring partly highly unstructured data

People need years to learn how to use language. Nuances and countless irregularities make it difficult to recognise and distinguish. This makes it difficult for NLP solutions to convert unstructured language into structured data for use. Now we are not just talking about sarcasm, metaphors and synonyms. There are also exceptions in grammar, syntax and idioms, but above all context. Example:

In German, for example, "I'm going to my bank" can mean both going to a branch of a bank to deposit money there and the intention of leisurely reading a book on a bench in a park at one's favourite spot.



### Specific challenges with spoken language:

Dialects, slang, loan words, mumbled, bad connection, reeled, voice pitches, amplitude variations, rhetorical questions, half-sentences, single words, etc.. This makes recognition and thus the creation of a context, which is essential for processing, more difficult.

### Specific challenges with written language:

Typing errors, different spellings and sentence structure as well as grammatical errors, missing or incorrect punctuation marks and abbreviations create major - but solvable - challenges for NLP solutions, as many entities have to be recognised and correctly assigned. Homonyms are also among them: Words that can have different meanings. For example, in German "Band" - this can be a book of a multi-part book series, as well as a narrow textile strip or a music group.

NLP solutions recognise the context, constantly improve, they learn and thus optimise themselves automatically. The more training data there is at the beginning of an NLP project, the better.

## The solution: train to extract and extract to analyse

Nothing works without data. There are ready-made NLP solutions that are trained with synthetic data. "Training" in this context means that they make assumptions, test them, analyse them and then optimise themselves accordingly, make new refined assumptions, cast them in new models, test and analyse them again - a continuous and ongoing process. In order to adapt individual NLP solutions to the respective requirements in order to achieve defined goals, they have to be created, ideally with one's own data.

Generalised, one can say that it is about clustering NLP language into individual components. Then to elicit the relationships of these components to each other and to create a context from this.

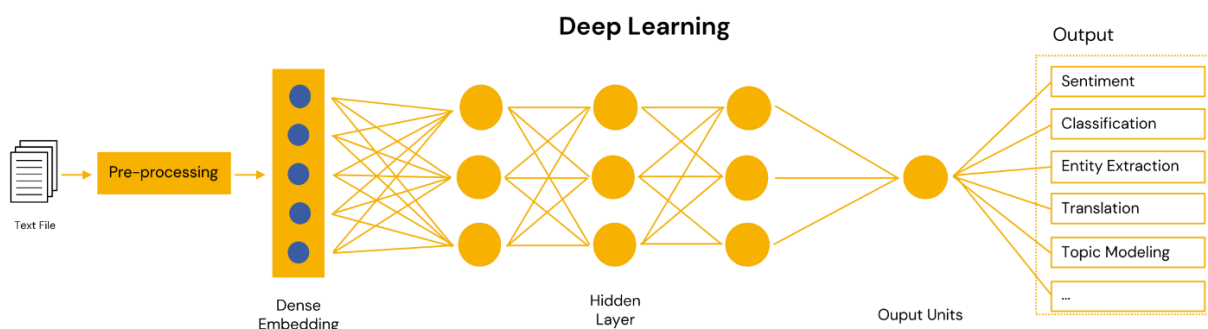
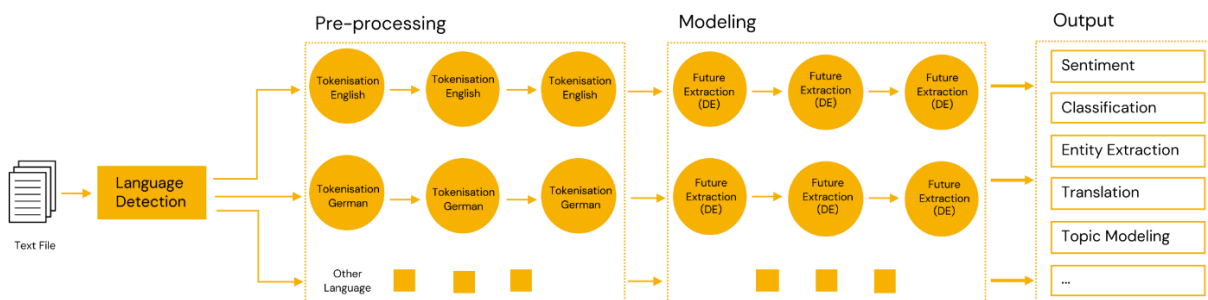
### Integral part of NLP: syntax and semantic analysis

In **syntactic analysis**, the arrangement of words in a sentence is checked to see if it makes sense. Syntactic analysis evaluates whether the extracted language conforms to grammatical rules and derives rules to create appropriate algorithms. Syntax analysis can be divided into, among others:

- Lemmatisation - a kind of reduction or grouping of different almost meaningless words into groups. For example, "joy", "enthusiasm", "cheerfulness" and "pleasure".
- Word segmentation - the division of a long text document into individual units.
- Stemming - recognising a word in an inflected form and classifying it accordingly.
- Tagging - word types are recognised, i.e. whether verb, noun or adjective etc.

**Semantic analysis** examines the exact content and meaning of a text and places it in context. For this purpose, the already mentioned example "bank" - bank for withdrawing money or bank for resting. In addition, semantic analysis forms the basis for the creation of AI-generated texts and voice output - be it short news articles in print or online or the voice assistance systems already mentioned: "Alexa, play me my *favourite playlist* from *last month*!"

Modern Transformer models perform these steps automatically, so to speak; they are integrated into the developed model. The advantage of this is that you don't have to do it manually. The disadvantage, however, is that it requires a complex model with a lot of training data to learn all this once.



### What NLP can do specifically:

- **Speech-to-text and text-to-speech conversion** - the conversion of voice commands into computer actions or written text and vice versa. Dictation programmes or voice-controlled systems are well-known examples.
- **Machine translation** - Google Translate is probably the best-known example. The automatic translation of written or spoken text from one language to another. Translation apps on smartphones are also based on this, as are apps that automatically translate text objects photographed by smartphone (menus in other languages as an example).
- **Sentiment analysis:** This is the recognition of moods or subjective opinions in texts. For example, emails can be immediately identified as complaints or support requests, categorised and processed accordingly.
- **Content categorisation:** a linguistically based document summary. This includes content search and indexing, content warnings and duplicate detection.
- **Contextual extraction:** automatic extraction of structured information from text-based sources - for example, in contracts, different entities such as name, contract duration, notice period, conditions, special clauses, etc.
- **Document summarisation:** extensive texts are automatically summarised.
- **Forecasting:** based on analysed documents such as contracts, customer communication via email or data captured via OCR, predictions can be made for upcoming or changing customer behaviour patterns.

## Conclusion:

Natural Language Processing is not new and not a dream of the future. NLP is already used in many areas today. Whether it is "Alexa, I like this song", a corresponding algorithm is adapted for the creation of title suggestions, the navigation assistant is started in the car by voice command or thousands of contracts are analysed and evaluated in seconds.

The applications and possible uses are extremely diverse, and the technology for them is developing rapidly. The basis for everything is and will remain: Data. First, language must be transformed into data, whether spoken or written. This unstructured data is structured, processed and edited using machine learning. The more data there is, the more precise the results, because the NLP solution can learn better.

## About tetrel

At tetrel, autonomous data science teams work for companies and government organisations on turnkey document analysis solutions to extract structured data from complex documents. These teams are assembled individually according to requirements.

tetrel is an independent and self-funded company. Together with our partners, our mission is to identify strategic use cases for Artificial Intelligence, evaluate them and turn them into productive solutions. In doing so, we understand AI as a tool to support highly qualified employees and a lever to efficiently increase productivity.

## Your contact



Johannes Humbert  
+49 176 83 33 51 46  
johannes.humbert@tetrel.ai