



# SD TSIA 214 – Machine learning for natural language processing

Chloé Clavel,  
[chloe.clavel@telecom-paristech.fr](mailto:chloe.clavel@telecom-paristech.fr),

Telecom ParisTech, France



# Natural Language processing ?

In French : Traitement Automatique du Langage Naturel (TALN)

## Crossroads of :

- ▶ Artificial Intelligence
- ▶ Linguistics
- ▶ Machine learning

# Natural Language processing ?

## Objectives :

- ▶ Extract meaning from textual data
- ▶ Speech synthesis, natural language generation

[https://www.youtube.com/watch?v=Ea\\_ytY0UDs0](https://www.youtube.com/watch?v=Ea_ytY0UDs0) Luc Steels -  
BREAKING THE WALL TO LIVING ROBOTS. How Artificial  
Intelligence Research Tries to Build Intelligent Autonomous  
Systems - 1 min 52



# NLP applications

?

# NLP applications

- ▶ Automatic translation (Google translate)
- ▶ Textual data mining/ document classification / information extraction
- ▶ Spell-checkers
- ▶ Automatic summary
- ▶ Human-Computer interactions
- ▶ Speech recognition
- ▶ Speech synthesis
- ▶ Opinion analysis (from social media)

# The textual Data and its challenges

Challenges : Moving away from the academic writing to spontaneous expressions (abbreviations, hashtags, acronyms, typos/mistakes, oral transcript)



The screenshot displays a Twitter thread. The first tweet is from Agence France-Presse (@afpr) dated 5 Sept, reporting a chemical incident in Fessenheim where two people were slightly burned, according to EDF, with a link and hashtag. The second tweet is from Stéphane GRAND (@Stephane\_Grand) also dated 5 Sept, mentioning the incident and two slightly injured people, also according to EDF. The third tweet is from Mediapart (@mediapart) dated 4 Sept, reporting on the nuclear industry's failures with EDF and its EPR, with a link. Below the tweets are two replies: the first starts with a breathing icon and the text '{breath} bonjour Madame . C'est bon madame , vous n'y êtes pour rien , mais je vais passer ma colère sur vous .', and the second starts with a fire icon and the text 'D'accord .'.

**Agence France-Presse** @afpr 5 Sept  
Incident chimique à Fessenheim: 2 personnes légèrement brûlées, selon EDF [bit.ly/TmN97R](https://bit.ly/TmN97R) #AFP  
Ouvrir

**Stéphane GRAND** @Stephane\_Grand 5 Sept  
Incident à #Fessenheim : deux personnes ont été légèrement brûlées, selon EDF  
Ouvrir

**Mediapart** @mediapart 4 Sept  
Nucléaire: les déboires anglo-saxons d'EDF avec son EPR  
[bit.ly/R3z2EY](https://bit.ly/R3z2EY)  
Ouvrir

{breath} bonjour Madame . C'est bon madame , vous n'y êtes pour rien , mais je vais passer ma colère sur vous .

D'accord .

## Lectures and pedagogical team

1. Introduction to Natural Language Processing **CLAVEL Chloe**
2. Natural Language Preprocessing and resources **CLAVEL Chloe**
3. Syntax and Parsing **Jean-Louis Dessalles**
4. Text clustering and text categorization **CLAVEL Chloe**
5. Deep learning for NLP **CLAVEL Chloe**
6. Hidden Markov Models **Laurence Likforman**
7. Non-negative Matrix Factorization **Slim ESSID**
8. ML for opinion analysis **Chloé Clavel**

## Lab sessions

1. Syntax and Parsing **Jean-Louis Dessalles**
2. Naive Bayes for opinion categorization **Chloé Clavel**
3. Text segmentation using Markov Models **Laurence Likforman**
4. NMF For Topic Modelling **Laurence Likforman**





# Evaluation

- ▶ Multiple Choice Questions Test
- ▶ Labs

## At the end of the course...

- ▶ You will be able to describe and implement the different methods for text representation into vectors
- ▶ You will master the main linguistic issues for NLP
- ▶ You will be able to build a text classification framework
- ▶ you will master more involved machine learning methods



# Prerequisites

TSIA-SD 210 : supervised machine learning methods including neural networks



# Materials

See pedagogical website