

# Analyse statistique de textes

[max.silberztein@univ-fcomte.fr](mailto:max.silberztein@univ-fcomte.fr)

# Structure du cours

## 1. Approches méthodologiques

- Méthodes empiriques et formelles

## 2. Analyse statistique de texte

- les mesures, ex. score standard
- Etude de termes
- Etude d'un thème

## 3. Le projet personnel

- constituer un corpus
- maîtriser un outil d'analyse statistique de texte
- étudier trois thèmes
- présenter les résultats

# 1. Approches méthodologiques

# Outils informatiques pour traiter des textes : deux approches

Les outils informatiques utilisés pour analyser des corpus de textes écrits en langue naturelle suivent deux approches :

- L'approche **formelle** : on décrit précisément les expressions en jeu, en utilisant des dictionnaires et des grammaires construites par des spécialistes
- L'approche **empirique** : on utilise des corpus d'apprentissage pré-analysés que l'on compare avec le texte traité, et on copie les analyses préconstruites.

# Succès de l'approche empirique pour l'intelligence artificielle

## AlphaGo versus Lee Sedol

---

[Article](#) [Talk](#)

---

From Wikipedia, the free encyclopedia

**AlphaGo versus Lee Sedol**, also known as the **DeepMind Challenge Match**, was a five-game [Go](#) match between top Go player [Lee Sedol](#) and [AlphaGo](#), a [computer Go](#) program developed by [DeepMind](#), played in [Seoul](#), South Korea between 9 and 15 March 2016. AlphaGo won all but the fourth game;<sup>[1]</sup> all games were won by resignation.<sup>[2]</sup> The match has been compared with the historic chess match between [Deep Blue and Garry Kasparov](#) in 1997.

## ChatGPT and the spectacular evolution of natural language processing

25 APR. 2023

The buzz around ChatGPT raised interest on recent advances in natural language processing based on pre-trained artificial intelligence models and on the principle of transfer learning. Michalis Vazirgiannis and Moussa Kamal Eddine are specialists of this field in the Data Science and Mining group at the Computer Science Laboratory of the Ecole Polytechnique. They explain how ChatGPT work, the latest developments and the issues at stake.

# Outils **spectaculaires** en 2024

- ChatGPT, Microsoft Copilot, Google Gemini
- Gamma (produit des présentations Powerpoint)
- Copy.AI, Jasper, Madgicx (produit des campagnes de marketing)
- Midjourney, Dall-E3 (produit des images à parti de leur description)
- Lex (produit des réseaux d'idées « brainstorming »)
- Speechify (Text to Speech)
- Grammarly Business (corrige fautes d'orthographe et style)
- Fathom (gère les idées produites lors de meetings)
- Elicit (produit des résumés d'articles scientifiques)
- Perplexity AI (moteur de recherche intelligent)
- Sybill (produit des emails de relance)
- Tesla (voitures autonomes)
- ...

C'est magique !



Siri. Quelqu'un à qui parler.

Siri est l'assistant intelligent qui comprend ce que vous dites, voit ce que vous voulez dire et trouve les réponses à vos questions. Et comme il maîtrise désormais davantage de langues, il se rend indispensable dans plus de pays<sup>2</sup>.

# Promesses,

## promesses



Wired

<https://www.wired.com> > ... > Electric Vehicles

### Elon Musk Says Every New Tesla Can Drive Itself

Oct 19, 2016 — Tonight, at a press conference, he announced that every new Tesla will be fully capable of **driving** itself. After being upgraded with a suite of ...



← Tweet



**Elon Musk** ✓  
@elonmusk

In ~2 years, summon should work anywhere connected by land & not blocked by borders, eg you're in LA and the car is in NY

9:11 PM · Jan 10, 2016

4,208 Retweets 534 Quote Tweets 7,414 Likes



17:27

5G



electrek



### Tesla's self-driving is still not working in Boring Company's one-way tunnels

Fred Lambert | Aug. 27th 2024 12:12 pm PT

Tesla's self-driving system is still not working in The Boring Company's one-way tunnels under Las Vegas despite reportedly working on it for years.

But self-driving in Tesla vehicles on public roads is around the corner?

Tesla uberbulls often like to say that Tesla is the leader in self-driving because while it doesn't have a commercially available autonomous ride-hailing service like Waymo, it doesn't rely on geo-fencing and manning like Waymo.



# mais avec quelques bémols tout de même (hallucinations)



How many countries begin with the letter "Y"?



As of my last knowledge update in January 2022, there are three countries whose names start with the letter "Y." These countries are:

1. Yemen
2. Zambia
3. Zimbabwe

Please note that geopolitical changes can occur, and the number of countries in the world can change over time, so it's a good practice to verify this information with up-to-date sources for the most current and accurate count.



Lindsay McGregor

@LindsMcGregor · Jul 5

↻ 46   ❤️ 330   ▾

Me: Alexa, turn on family room lights.  
Alexa: Okay, playing "Livin' on a Prayer, lullaby version."  
[#alexafail](#) [#robottakeover](#) 🤪



© Twitter

**Free-styling! Alexa decided she'd rather turn on a lullaby than the lights... surely making for a rather dark musical experience**



If 1 woman can make 1 baby in 9 months, how many months does it take 9 women to make 1 baby? Explain each step you used to arrive at your answer.



First, we need to determine how many babies each woman makes per month. We can do this by dividing the number of babies each woman makes (1) by the number of months it takes for each woman to make a baby (9). This gives us a rate of  $1/9$  babies per month for each woman.

Next, we need to determine how many babies 9 women make per month. We can do this by multiplying the number of women (9) by the rate at which each woman makes babies ( $1/9$  babies per month). This gives us a total rate of  $9 * (1/9) = 1$  baby per month.

Finally, we need to determine how many months it takes for 9 women to make 1 baby. We can do this by dividing the number of babies we want (1) by the rate at which the women are making babies (1 baby per month). This gives us a total of  $1 / 1 = 1$  month.

Therefore, it takes 9 women 1 month to make 1 baby.

# ... et quelques problèmes éthiques

07:37 98

PC GAMER

## OpenAI says it's 'impossible' to create ChatGPT without copyrighted content, as if that's somehow a good excuse

In the face of a growing number of lawsuits, OpenAI insists that the use of copyrighted content to train LLMs is fair use.



(Image Credit: Getty Images)

By Andy Chalk, published January 08, 2024

Just a couple weeks after being **sued by the New York Times** over allegations that it copied and used "millions" of copyrighted news articles to train its large-language models, OpenAI has told the UK's House of Lords communications and digital select committee (via **The Guardian**) that it has to

Today News+ Sports Audio Following

News+

THE NEW YORKER

## More people are being arrested for crimes they have no connection to — and AI is to blame

Play Now 40 min

Today News+ Sports Audio Following

# Approche empirique pour le traitement automatique de textes en langue naturelle (TAL)

*Automatic part of speech tagging is an area of natural language processing where statistical techniques have been more successful than rule-based methods. (Eric Brill, 1992)*

Pas de linguistique à la conférence *Computational Linguistics* ???

*The availability of large text corpora has changed the scientific approach to language in linguistics and cognitive science. (Manning, Schütze 2001)*

## THE 29<sup>TH</sup> INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS

TITLE
A Coarse-to-fine Cascaded Evidence-Distillation Neural Network for Explainable Fake News Detection
Finding Influential Instances for Distantly Supervised Relation Extraction
Stability of Syntactic Dialect Classification Over Space and Time
Hierarchical Attention Network for Explainable Depression Detection on Twitter Aided by Metaphor Concept Mappings
CMQA: A Dataset of Conditional Question Answering with Multiple-Span Answers
TSAM: A Two-Stream Attention Model for Causal Emotion Entailment
Few-shot Named Entity Recognition with Entity-level Prototypical Network Enhanced by Dispersedly Distributed Prototypes
Two Languages Are Better Than One: Bilingual Enhancement For Chinese Named Entity Recognition
DCT-Centered Temporal Relation Extraction
Debiasing Word Embeddings with Nonlinear Geometry
CofeNet: Context and Former-Label Enhanced Net for Complicated Quotation Extraction

Screenshot

# Les applications de TAL

Aujourd'hui, le grand public, et les chercheurs en SHS (Histoire, Linguistique de corpus, Littérature, Socio-linguistique, Psychologie, etc.) disposent de nombreux outils pour traiter leur corpus, mais ces outils traitent des formes graphiques et non des unités linguistiques :

*Alceste, Hyperbase, IRaMuTeQ, Lexico 3, Mallet, Sketch Engine, Le Trameur, TXM, Word2Vec, etc.*

# Les principaux outils utilisés aujourd'hui en H.N.

- ATISHS <http://atish.univ-fcomte.fr>
- Hyperbase : <http://hyperbase.unice.fr/hyperbase/>
- IRaMuTeQ : <http://www.iramuteq.org/>
- Lexico : <http://www.tal.univ-paris3.fr/lexico/lex3-10pas/index.htm>
- NooJ : <http://www.nooj4nlp.org> , <http://webnooj.univ-fcomte.fr>,
- Sketch Engine : <https://www.sketchengine.eu/>
- TXM : <http://portal.textometrie.org/demo/>

# 1. Le vocabulaire

## Formes graphiques vs. Unités linguistiques

- **Formes graphiques**, mots informatiques : séquences de lettres délimitées par des espaces ou des ponctuations.

*... parce que Luc mange une pomme de terre ...*

8 formes graphiques  $\neq$  5 unités linguistiques

- **Unités linguistiques** : la plupart des formes graphiques correspondent à différentes unités linguistiques

# Le nom *carte* a plus de 14 sens

Voir aussi : **carté**

→ **Français** [ modifier le wikicode ]

Etymologie [ modifier le wikicode ]

*(Date à préciser)* Du latin *charta* (« papier, écrit, livre »), lui-même du grec ancien *χάρτης*, *khártēs* (« feuille de papyrus ou de papier »).

Nom commun [ modifier le wikicode ]

- carte** \kaʁt\ *féminin*
- (Rare)* Matière obtenue par assemblage de plusieurs feuilles de papier collées l'une sur l'autre.
  - (Commerce)* *(Vieilli)* Petit présentoir cartonné sur lequel étaient placés de menus objets en vente
    - D'un œil vague, Bert entrevoyait ces phrases, par un intervalle vide, dans le carreau de la porte, au-dessus d'une **carte** sur laquelle étaient fixées des valves neuves. — (H. G. Wells, *La Guerre dans les airs*, 1908, traduction d'Henry-D. Davray et B. Kozakiewicz, Mercure de France, Paris, 1910, page 66 de l'édition de 1921)
  - (Cartographie, Géographie)* Représentation géométrique conventionnelle, en positions relatives, de phénomènes concrets ou abstraits (pays, région, ville, astres dans le ciel, etc.), localisables dans l'espace, à échelle réduite<sup>[1]</sup>.
    - Rencontré le lieutenant Bertet. Il n'en revient pas d'être déjà en Alsace. Il n'avait pris que des **cartes** de Prusse et de Bavière, comme s'il s'agissait dans cette guerre de délivrer la Pologne [...]. — (Jean Giraudoux, *Retour d'Alsace* - Août 1914, 1916)
    - Je savais par mes **cartes** qu'un étroit et tortueux chenal, d'environ 150 mètres de largeur, menait au port et que rien ne l'indiquait à la surface de la mer. — (Alain Gerbault, *À la poursuite* Screenshot : *De New-York à Tahiti*, 1929)

La France se regarde . Le Florentin de la nature, Librairie Hachette, Paris, 1931

- Comme l'histoire des plantes fossiles est encore bien incomplète, on ne met pas facilement en relation ces flores avec les **cartes** paléogéographiques. — (Henri Gaussen, *Géographie des Plantes*, Armand Colin, 1933, page 76)
- Il possédait si bien la **carte** des fossiles du département qu'il ne donnait jamais un coup de piochon sans exhumer un spécimen rare. — (Jean Rogissart, *Passantes d'Octobre*, Librairie Arthème Fayard, Paris, 1958, page 70)
- Je pense à ces **cartes** d'autrefois, où l'on semblait moins chercher à reproduire une contrée qu'à évoquer un paysage, où le dessinateur ne se bornait pas à situer les villes, les bourgs, les routes et les cours d'eau, mais où il les figurait en relief, avec de petites maisons, des ponts, des personnages, et des attelages courant à grand train. — (José Cabanis, *Les cartes du temps*, Gallimard, 1962, Le Livre de Poche, pages 45-46)
- (Par extension)* Géographie d'un pays, d'une région, etc.
  - L'État [...] : ses structures et la légitimité de ses souverains sont attaqués du Levant à l'Asie du Sud-Ouest. Un conflit intérieur et son corollaire d'instabilité régionale peuvent donc faire craindre la balkanisation de la région et, par suite, une refonte de sa **carte**. — (Panayiotis Jerasimof Vattikiotis, *L'Islam et l'État*, 1987, traduction d'Odette Guillard, 1992, page 87)



Carte (8) informatique.



La carte d'un restaurant. (9)

- (Cartes à jouer)* Rectangle de carton qui porte une figure, un symbole ou des instructions, et qui sert à jouer.
  - Le papier filigrané destiné à former le devant des **cartes** désignées ci-dessus, sera fabriqué et fourni par la régie ; les fabricar ne pourront point en employer d'autre. — (Arrêté du Directoire exécutif qui détermine le mode de perception et fixe le montant du droit de timbre sur les cartes à jouer, 3 pluviôse an VI, dans *Collection général des lois, décrets, arrêtés, sénatus-consultes*, Paris : Imprimerie Royale, juin 1818, volume 6, page 727)
  - Pour ma part, je n'aime guère les **cartes** et les accuse d'avoir tué la conversation en France et ailleurs. D'aucuns prétendront que c'est là leur principal avantage. Après tout, c'est bien possible... principalement dans les salons « mondains ». — (Samivel, *L'amateur d'abîmes*, 1940, réédition Le Livre de Poche, pages 147-148)
  - Les ailes du ventilateur, ce hanneton, chassent comme des feuilles les **cartes** du mort étalées sur la table, [...]. — (Paul Nizan, *Aden Arabie*, Rieder, 1932 ; Maspéro, 1960, chapitre V, page 76)
  - Le donneur distribue 7 **cartes** à chaque joueur et retourne la première **carte** de la pioche, qui constitue l'entame. — (Yann Caudal, *Le grand guide des jeux de cartes*, Éditions Eyrolles, 2011, page 71)
- Petit carton comportant divers types de renseignements.
  - J'ai besoin de ma **carte** d'identité.

- Petit rectangle de plastique comportant une puce électronique ou une bande magnétique pour réaliser diverses opérations.
  - Les distributeurs automatiques de billets sont légion au Japon mais la plupart refusent les **cartes** étrangères. Même lorsqu'un DAB porte les logos Visa et MasterCard, il n'accepte souvent que les **cartes** japonaises attachées à ces deux réseaux. — (Japon - Nord de Honshu (Tohoku), *Lonely Planet*, 2016)
- (Électronique)* Plaque faite de diverses couches de matériaux sur laquelle se trouve le circuit imprimé qui relie des composants électroniques entre eux.
  - Ainsi, pour le contrôle de moteurs à commutation électronique ([...]), une **carte** contient généralement un pont de transistors de puissance alimenté par une tension continue et commandé par un circuit de modulation de largeur d'impulsions. — (Sylvain Allano, *Petits moteurs électriques*, Éditions Techniques de l'Ingénieur, n° D 3 720, juin 1995, page 13)
- (Cuisine)* Au restaurant, liste des plats et des boissons parmi lesquels on peut choisir.
  - Les deux hommes consultèrent la **carte** : il commandèrent chacun une douzaine d'escargots qu'ils arrosèrent de muscadet. — (Francis Carco, *L'Homme de minuit*, Éditions Albin Michel, Paris, 1938)
  - Nous étions, en cet instant, comme des badauds qui lisent à l'extérieur la **carte** du restaurant et s'esbaudissent parce que tout alors paraît bon et possible. — (Pierre Sansot, *Cahiers d'enfrance*, Éditions Champ Vallon, 1989, page 32)
- (Reliure)* Carton mince et fort servant de faux dos.
  - Morceau de papier fort ou de carton mince placé entre la peau et le dos du livre dans les reliures à dos brisé. La **carte** est libre ou fixe selon qu'elle est collée ou non sur le dos du livre. — (Albert Maire, *Manuel pratique du bibliothécaire, Lexique des termes du livre*, Alphonse Picard et fils, 1896, page 302)
- Carton au format d'une carte à jouer, mais représentant une vedette (généralement sportive) avec parfois des renseignements sur elle.
  - La **carte**, identifiée comme T206, n'aurait été imprimée que 50 à 200 fois après que Wagner, des Pirates de Pittsburgh, eut demandé qu'elle soit retirée. — (Agence QMI, *Une demi-carte vendue à près de 500 000 \$*, site tvnouvelles.ca, 7 février 2022)

- avoir les cartes en main
- avoir toutes les cartes en main
- avoir une carte à jouer
- abattre sa dernière carte
- abattre ses cartes
- bicarte
- bien jouer ses cartes
- brouiller les cartes
- brûler une carte
- bureau à la carte
- carte à gratter
- carte à jouer
- carte à payer
- carte à perforations marginales
- carte à puce
- carte à vol d'oiseau
- carte aéronautique
- carte aéronautique de navigation à petite échelle
- carte aéronautique du monde
- carte anaglyptique
- carte astrale
- **carte aveugle**
- carte bancaire
- carte bathymétrique
- carte blanche
- carte bleue
- carte cadeau, carte-cadeau
- carte chorographique
- carte cognitive
- **carte d'abonnement**
- carte d'achat
- carte d'alimentation
- carte d'appel
- carte d'approche à vue

Screenshot

- carte interzones
- carte itinéraire
- carte-lettre
- carte magnétique
- carte maillée
- carte maximum
- carte mémoire
- carte mentale
- carte-mère, carte mère
- carte Michelin
- carte muette
- carte murale
- carte nationale
- carte nautique
- carte non régulière
- carte numérique
- carte numérisée
- carte océanique
- carte orographique
- carte par mailles
- carte par points
- carte particulière
- carte perforée
- **carte-photo**
- carte photographique
- carte planimétrique
- carte postale
- carte prévisionnelle
- carte qualitative
- carte quantitative
- carte radar
- carte régulière
- carte réseau
- carte routière

- carte d'assurance
- carte d'atterrissage
- carte d'échantillons
- carte d'embarquement
- carte d'entrée
- carte d'entreprise
- carte d'état-major
- carte d'étudiant
- carte d'identité
- carte d'impact
- carte d'inventaire
- carte d'itinéraire
- carte d'obstacle d'aérodrome
- carte de cabotage
- carte de crédit
- carte de débit
- carte de détresse
- carte de France
- carte de membre
- carte de navigation
- carte de paiement
- carte de pilotage côtier
- carte de pilotage hauturier
- carte de reconnaissance
- carte de reconstitution
- carte de séjour
- carte de situation
- carte de synthèse
- carte de téléphone
- carte de tracé de navigation
- carte de traversée
- carte de visite
- carte de vœux
- carte déclassée
- carte des idées
- carte du Tendre

Screenshot

- carte SD
- carte SIM
- carte SIS
- carte soleil
- carte-soleil
- carte spectrale
- carte stéréoscopique
- carte sur fond d'image
- carte téléphonique
- carte topographique
- carte tuner TV
- carte vectorielle
- carte vermeil
- carte verte
- carte vidéo
- carte vitale
- carte-vue
- carter
- carterie
- cartomancien
- carton
- cartonner
- cartophile
- cartophilie
- château de cartes
- corps de carte
- dessous des cartes
- donner carte blanche
- en carte
- encarté
- encarter
- encarteuse
- être en carte
- iconocarte
- jeu de cartes
- jouer cartes sur table

# Multiples sens du mot anglais *shit*

ISMO

<https://www.youtube.com/watch?v=igh9iO5BxBo&t=69s>

Quel est l'intérêt de savoir que le mot *carte* apparaît de façon exceptionnellement fréquente chez tel auteur, dans telle oeuvre, dans les journaux en février 2024 ?

Quel est l'intérêt de dire que le mot *abriter* est caractéristique des discours de telle personnalité politique, si ce mot a plusieurs sens, dont certains positifs et d'autres négatifs ?



## 2. Morphologie

- Les unités linguistiques peuvent être fléchies ou dérivées en conservant leur sens
- Les flexions et dérivations ne sont pas régulières
- Une centaine de paradigmes morphologiques en français

Les sous-  
fréquences dans les  
textes sont  
réparties en  
binomes, dont le  
premier élément  
indique le numéro  
d'ordre du texte ,  
le second la  
fréquence du mot

Cliquer sur un  
mot pour l'activer



Index hiérarchique

Chercher un mot



Graphique



Éditer



Quitter



Retour



Sommaire



## Formes

111 **amour**  
1 amour»  
1 amouracher  
26 amoureuse  
8 amoureuxment  
2 amoureuses  
55 amoureux  
49 amours  
1 amphigourique  
2 amphithéâtre  
1 ample  
5 ampleur  
1 amplifia  
13 amusa  
5 amusaient  
25 amusait  
16 amusant  
2 amusante  
1 amusantes

## Lemmes

Che

N° 1 TEX1 36  
N° 3 TEX3 10  
TOUS LES TEXTES

N° 2 TE  
N° 4 TE

**amour** fréq  
CLIQUEZ SUR UN TEXTE (ou sur TOUS LES TEXTES)  
pour afficher les contextes du mot "amour"  
Cliquez AILLEURS dans cette fenêtre

0 and\_2  
3 and\_3  
8 andouille\_2  
1 andré\_2

T C:\HY

Formes	Lemmes
111 <b>amour</b>	N° 1 TEX1 36
1 amour»	N° 2 TE
1 amouracher	N° 3 TEX3 10
26 amoureuse	N° 4 TE
8 amoureuxment	TOUS LES TEXTES
2 amoureuses	
55 amoureux	
1 amours	
1 amphigourique	<b>amour</b> fréq
2 amphithéâtre	CLIQUEZ SUR UN TEXTE (ou sur T
1 ample	contextes du mot "amour"
5 ampleur	Cliquer AILLEURS dans cette fenêtre
1 amplifia	
13 amusa	
5 amusaient	
25 amusait	
16 amusant	
2 amusante	
1 amusantes	

Les sous-  
fréquences dans les  
textes sont  
réparties en  
binomes, dont le  
premier élément  
indique le numéro  
d'ordre du texte ,  
le second la  
fréquence du mot

Cliquer sur un  
mot pour l'activer

Index hiérarchique

Chercher un mot

Graphique

Éditer

Quitter

Retour

Sommaire

3 and\_2  
3 and\_3  
8 andouille\_2  
1 andré 2

Lorsqu'on recherche *amour*, le logiciel nous propose des formes de *amuser*,  
mais aucune forme de *aimer*

# CONCORDANCE

French Web 2017 (frT...

Obtenez plus d'espace +

lemme **aimer** 1871115 > filtre **Arranet 1, 2 23** inférieur à 0.01 X

phrase

Détails

phrase

1  [doc#15462](#) **<s>** C' est d' ailleurs au cours d' un   
PPSCNN0/ce VMIP3S0/être SP/de RG/ailleurs SP/à NCFP000/cour SP/de DI0MS0/un  
voyage à Versailles qu' il rencontre **Aimée**  
NCMS000/voyage SP/à NPFP000/versailles PR00000/que PP3MSN0/il VMIP1S0/rencontrer VMP00SF/aimer  
d' **Arranet** , dont il tombe éperdument  
SP/de NPMS000/arranet Fc/, PR00000/dont PP3MSN0/il VMIP1S0/tomber RG/éperdument  
amoureux dès le premier regard . **</s>**  
AQ0MN00/amoureux SP/dès DA0MS0/le AQ0MS00/premier NCMS000/regard Fp/.

2  [doc#15462](#) **<s>** Il doute aussi de son amour pour **Aimée**   
PP3MSN0/il VMIP1S0/douter RG/aussi SP/de DP3CSS/son NCMS000/amour SP/pour VMP00SF/aimer  
d' **Arranet** et pense de plus en plus à la  
SP/de NPMS000/arranet CC/et VMIP1S0/penser SP/de RG/plus SP/en RG/plus SP/à DA0FS0/le  
Satin : " A l' amour , feu couvant  
NPMS000/satin Fd/: Fe" VMIP3S0/avoir DA0CS0/le NCMS000/amour Fc/, NCMS000/feu VMG0000/couver  
, qu' il continuait à lui porter , s'  
Fc/, CS/que PP3MSN0/il VMII3S0/continuer SP/à PP3CS00/lui VMN0000/porter Fc/, PP3CN00/se  
ajoutaient une estime , un respect et une

Today

AB Antonio, 15:01  
 15:02  
hello

AB Antonio, 15:02  
bonjour Max, on peut commencer le rdv 15:02  
ok 15:02

15:02  
Call started

15:53

Type a message

lemme **aimer** 1871115

filtre **Arranet 1, 2** 23 inférieur à 0.01

phrase

Détails

phrase

1  [doc#15462](#) <s> C' est d' ailleurs au cours d' un voyage à Versailles qu' il rencontre **Aimée** d' **Arranet** , dont il tombe éperdument amoureux dès le premier regard . </s>

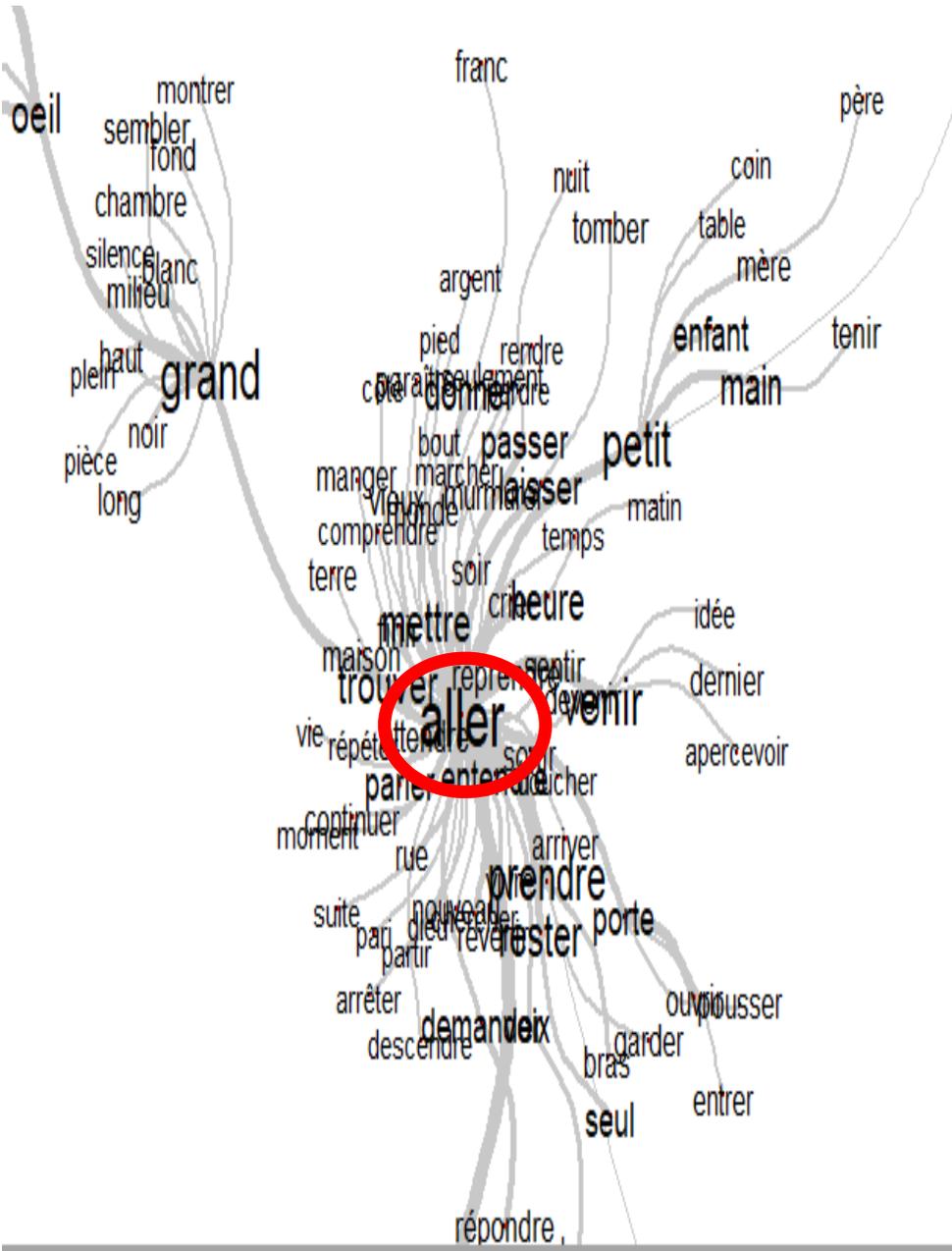
2  [doc#15462](#) <s> Il doute aussi de son amour pour **Aimée** d' **Arranet** et pense de plus en plus à la Satin : " A l' amour , feu couvant , qu' il continuait à lui porter , s' ajoutaient une estime , un respect et une

# 3. Les mots grammaticaux

- De nombreux mots sont des mots grammaticaux
- Les mots grammaticaux ne portent pas de sens
- Par ex., il y a une centaine de verbes auxiliaires en français

# Les mots grammaticaux dizaines de sens... ou pas de sens ?

## *aller*



1. Joe va manger [Joe **is going to** eat]
2. Joe va à la maison [Joe **is going** home]
3. Comment allez-vous ? [How **are** you?]
4. Ces couleurs vont bien ensemble [these colors **match**]
5. Ce bois va de la route jusqu'à la rivière
6. Leur engagement va croissant
7. Tous ses voeux vont à la paix
8. Vous allez vite en besogne
9. Rien ne va plus
10. Sa santé va en déclinant
11. Il en va de même pour le prix du gaz
12. Allez, nous en viendrons à bout !
13. Il y va de notre santé
- 14...

# Une centaine de verbes auxiliaires

AutoSave OFF | Home | Insert | Draw | Page Layout | Formulas | Data | Review | View | Tell me | Share | Comments

B3 | fx | -

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	
	N0=Nhum	N0=Nnr	Ppv0		aux=avoir	aux=être	N0estVppW	N0u		Prép	N1="que P"	N1="que Psubj"	Tp=lc	Tc=passé	Tc=présent	Tc=futur	Vc=devoir	Vc=pouvoir	Vc=savoir	Ppv	
1	+	+	<E>	devoir	+	-	-	-	<E>	-	-	-	+	+	-	-	+	+	-	-	<i>Il doit neiger en ce moment à Gap</i>
2	+	-	<E>	pouvoir	+	-	-	-	<E>	-	-	-	-	+	+	-	-	-	-	-	<i>Il peut toujours pleuvoir (&lt;E&gt;+j'ai mon parapluie)</i>
3	+	+	<E>	devoir	+	-	-	-	<E>	-	-	-	+	+	+	+	+	+	+	-	<i>Max doit partir</i>
4	+	+	<E>	être	+	-	-	-	<E>	-	-	-	-	-	+	-	-	-	-	-	<i>Max était pour partir quand il s'est mis à pleuvoir</i>
5	+	+	<E>	pouvoir	+	-	-	-	<E>	-	-	-	+	+	+	+	+	+	+	+	<i>(Max+le livre) peut passer entre les barreaux</i>
6	+	-	<E>	pouvoir	+	-	-	-	<E>	-	-	+	-	-	-	-	-	-	-	+	<i>Max peut réussir</i>
7	+	-	se	prendre	-	+	-	-	à	-	-	+	-	-	-	-	-	-	-	+	<i>Max se prend à (rêver+espérer)</i>
8	+	-	<E>	aller	-	+	-	-	jusqu'à	-	-	+	-	-	-	-	-	-	-	-	<i>Luc est allé jusqu'à insulter Léa</i>
9	+	+	<E>	aller	-	-	-	-	<E>	-	-	-	-	-	-	+	+	+	+	-	<i>La chaise va tomber</i>
10	+	+	ne	aller Nég	-	+	-	-	sans	-	+	+	-	-	-	+	+	+	+	-	<i>Cette mesure n'ira pas sans créer des troubles</i>
11	+	-	<E>	passer	+	-	-	-	pour	-	-	-	+	+	-	+	+	+	+	-	<i>Ida en passera par faire la vaisselle</i>
12	+	+	<E>	daigner	+	-	-	-	<E>	-	-	-	-	-	-	-	-	-	-	-	<i>Il a daigné pleuvoir</i>
13	+	+	<E>	manquer	+	-	-	-	de	-	-	+	-	-	-	-	+	+	+	-	<i>Il a manqué (&lt;E&gt;+de) pleuvoir</i>
14	+	+	<E>	partir	-	+	-	+	pour	-	+	-	-	-	+	+	-	+	+	-	<i>La pluie est partie pour durer</i>
15	+	-	se	presser	-	+	+	+	de	-	-	+	-	-	-	-	-	-	-	-	<i>La pluie ne se presse pas de tomber</i>
16	+	+	<E>	menacer	+	-	-	-	de	-	-	+	-	-	-	-	+	+	+	-	<i>La pluie menace (&lt;E&gt;+de tomber)</i>
17	+	-	ne se	gêner Nég	-	+	+	-	pour	-	+	-	-	+	+	-	-	-	-	-	<i>La pluie ne se gêne pas pour tomber</i>
18	+	+	n'en	finir Nég	+	-	-	-	de	-	-	+	-	-	-	-	-	-	-	-	<i>La pluie n'en finit pas (&lt;E&gt;+de tomber)</i>
19	-	+	<E>	refuser	+	-	-	-	de	-	-	+	-	-	-	-	-	-	-	-	<i>La pluie refuse de tomber</i>
20	+	+	<E>	risquer	+	-	-	-	de	-	-	-	+	+	+	+	+	+	+	-	<i>La pluie risque de tomber</i>
21	+	+	<E>	arrêter	+	-	-	+	de	-	-	+	-	-	-	-	-	-	-	-	<i>Luc arrête de travailler</i>
22	+	+	<E>	cesser	+	-	-	+	de	-	-	+	-	-	-	+	+	+	+	-	<i>Max a cessé de dormir</i>
23	+	+	<E>	finir	+	-	-	+	de	-	-	+	-	-	-	-	-	-	-	-	<i>Max a fini de ranger</i>
24	+	+	<E>	finir	+	-	-	-	par	-	-	+	-	-	-	+	+	+	+	-	<i>Max a fini par ranger</i>
25	+	-	<E>	négliger	+	-	-	-	de	-	-	+	-	-	-	-	-	-	-	-	<i>Max a négligé de se laver</i>
26	+	-	<E>	omettre	+	-	-	-	de	-	-	+	-	-	-	-	-	-	-	-	<i>Max a omis de fermer la porte</i>
27	+	-	<E>	oser	+	-	-	+	<E>	-	-	-	-	-	-	-	-	-	+	+	<i>Max a osé venir</i>
28	+	-	<E>	oublier	+	-	-	-	de	-	-	+	-	-	-	-	-	-	-	-	<i>Max a oublié de venir</i>
29	+	+	<E>	stopper	+	-	-	+	de	-	-	+	-	-	-	-	-	-	-	-	<i>Max a stoppé de faire des bêtises</i>

Ready | cm1 | + | 159%

## 2. Analyse statistique

# Analyse statistique

- Quantifier l'importance, l'intérêt d'une information ?

# Analyse statistique, score standard

- Quantifier *l'importance, l'intérêt* d'une information ?
- Pas par la fréquence, ex. *le* est très fréquent.

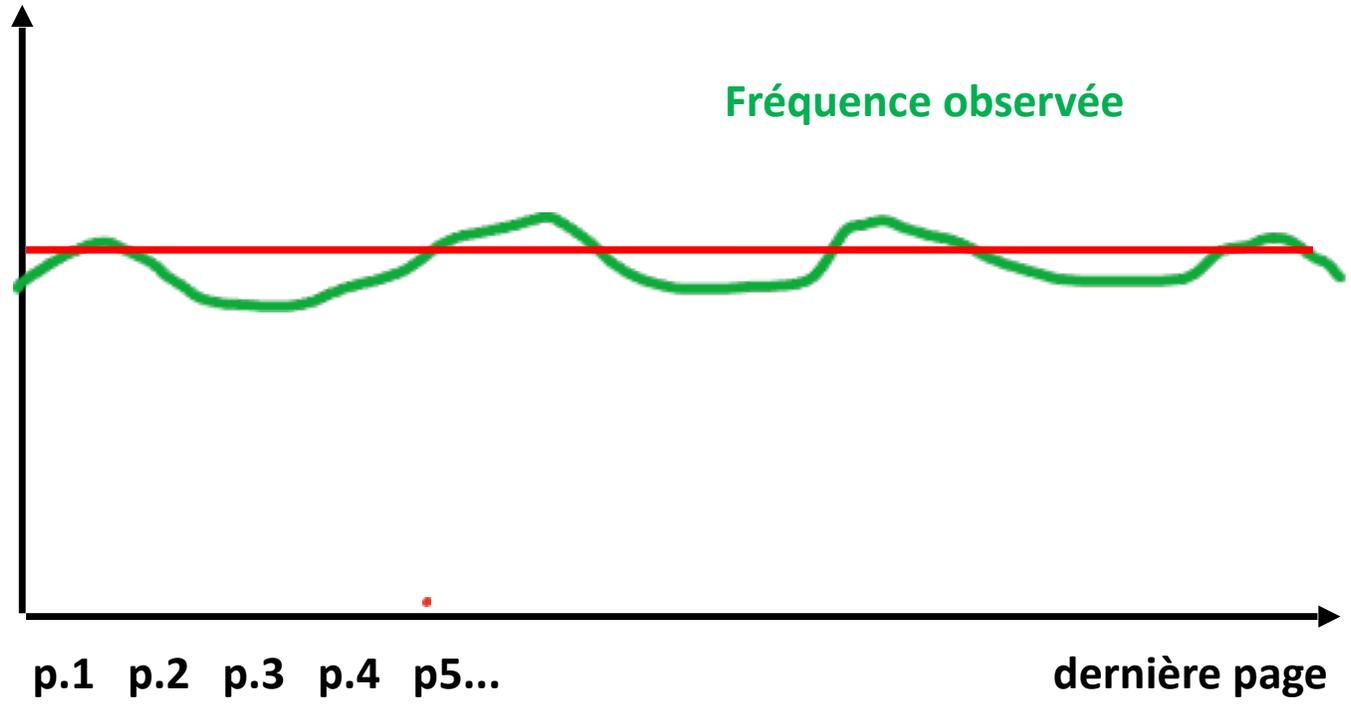
# Analyse statistique, score standard

- Quantifier **l'importance** d'une information ?
- Pas par la fréquence, ex. **le** est très fréquent ; aucun intérêt.
- Une seule occurrence du terme **COVID** dans un texte du XIXe siècle serait extrêmement intéressant !

# Analyse statistique, score standard

- Quantifier **l'importance** d'une information ?
- Pas par la fréquence, ex. **le** est très fréquent ; aucun intérêt.
- Une seule occurrence du terme **COVID** dans un texte du XIXe siècle serait extrêmement intéressant !
- On va donc chercher des accidents de fréquence, i.e. des différences anormales entre la fréquence attendue et la fréquence observée

**Fréquence  
attendue**

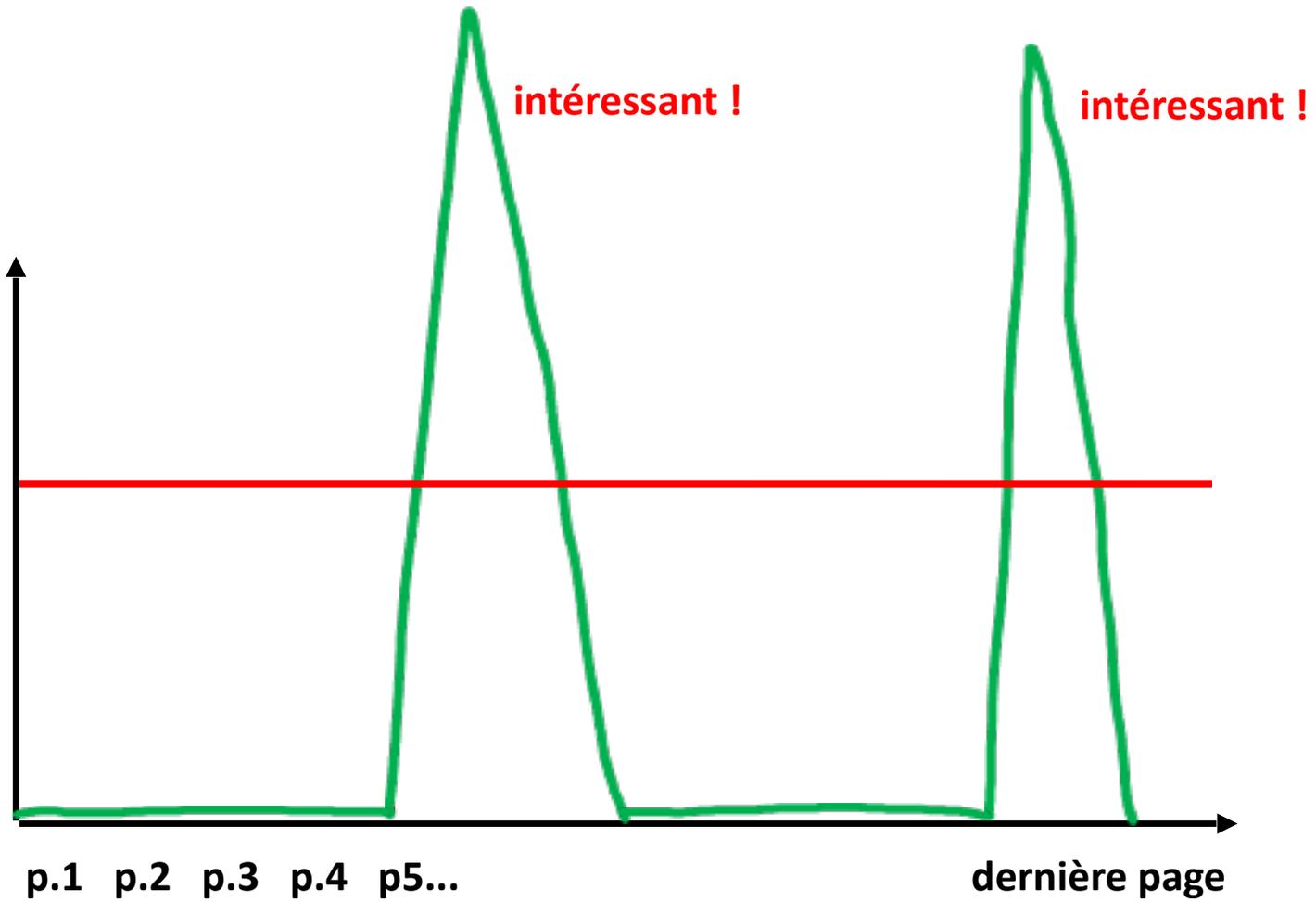


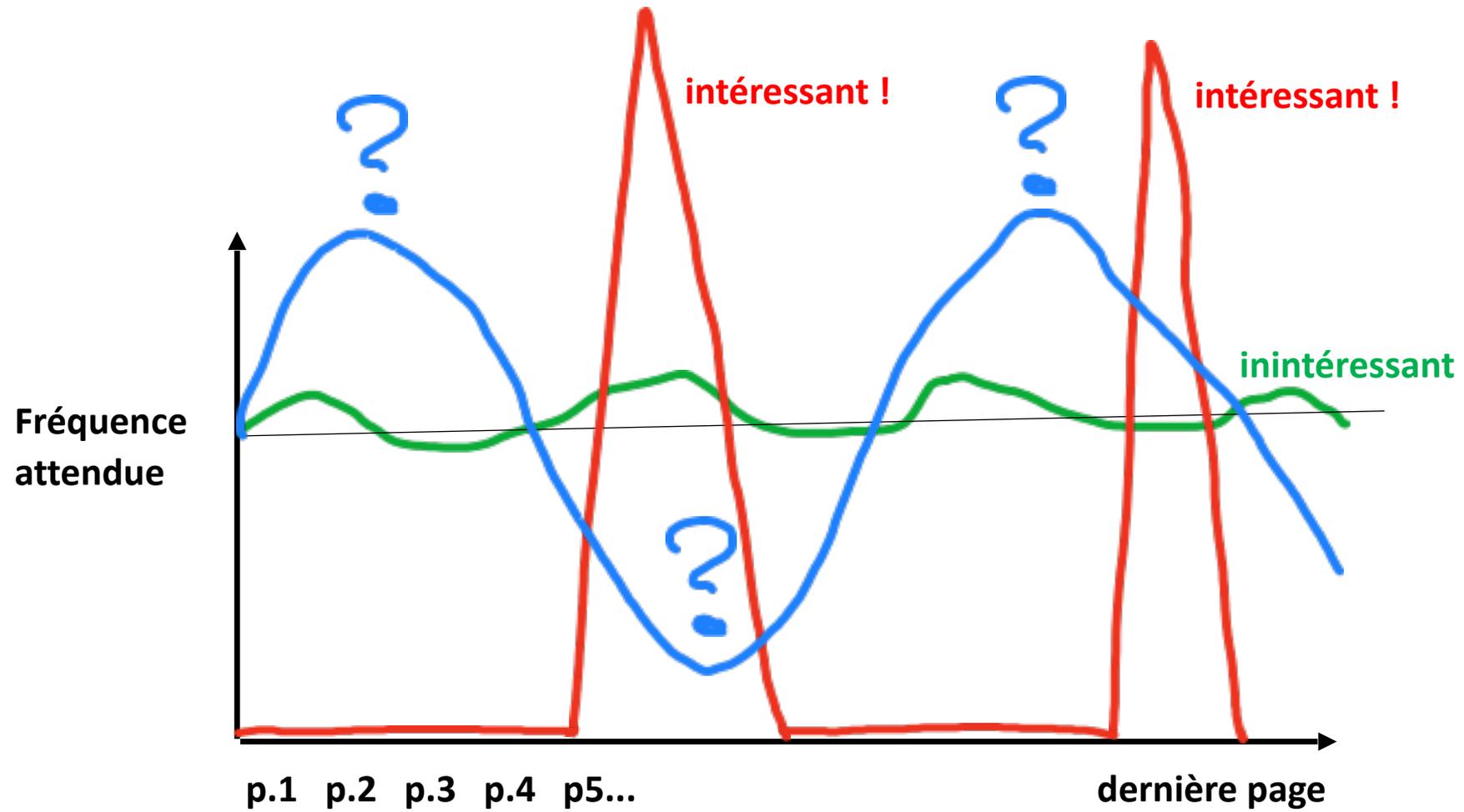
**Fréquence observée**

p.1 p.2 p.3 p.4 p5...

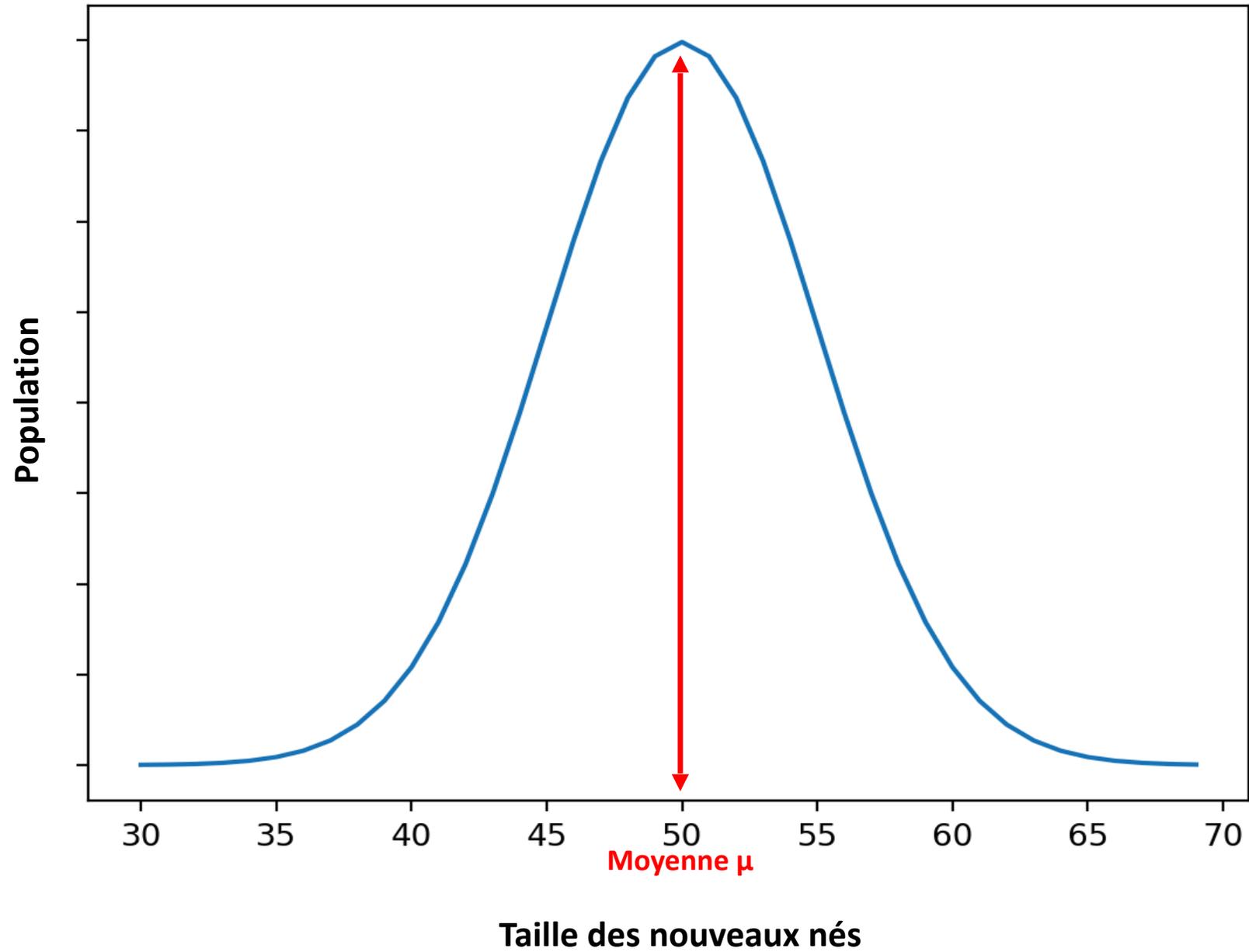
dernière page

Fréquence  
attendue

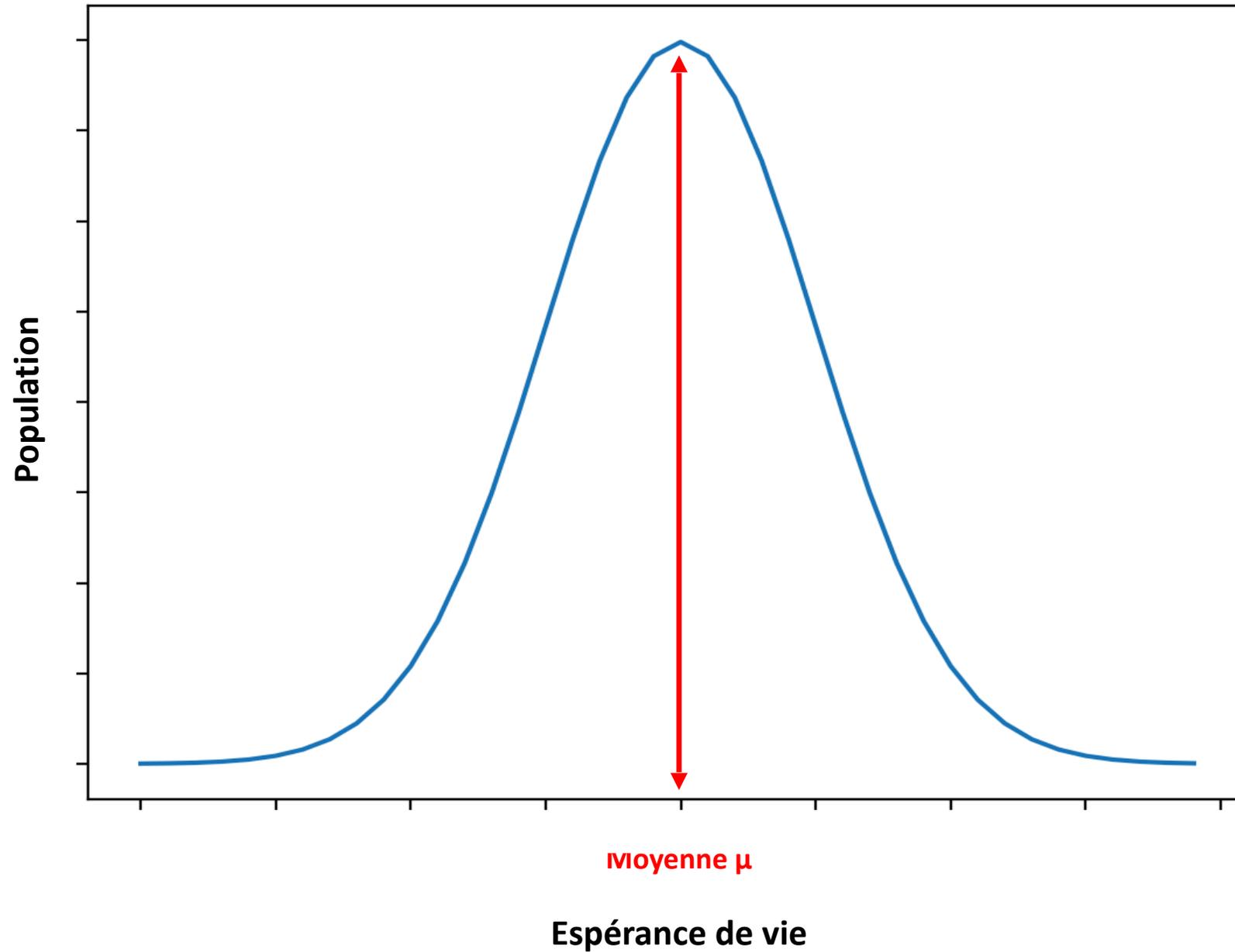




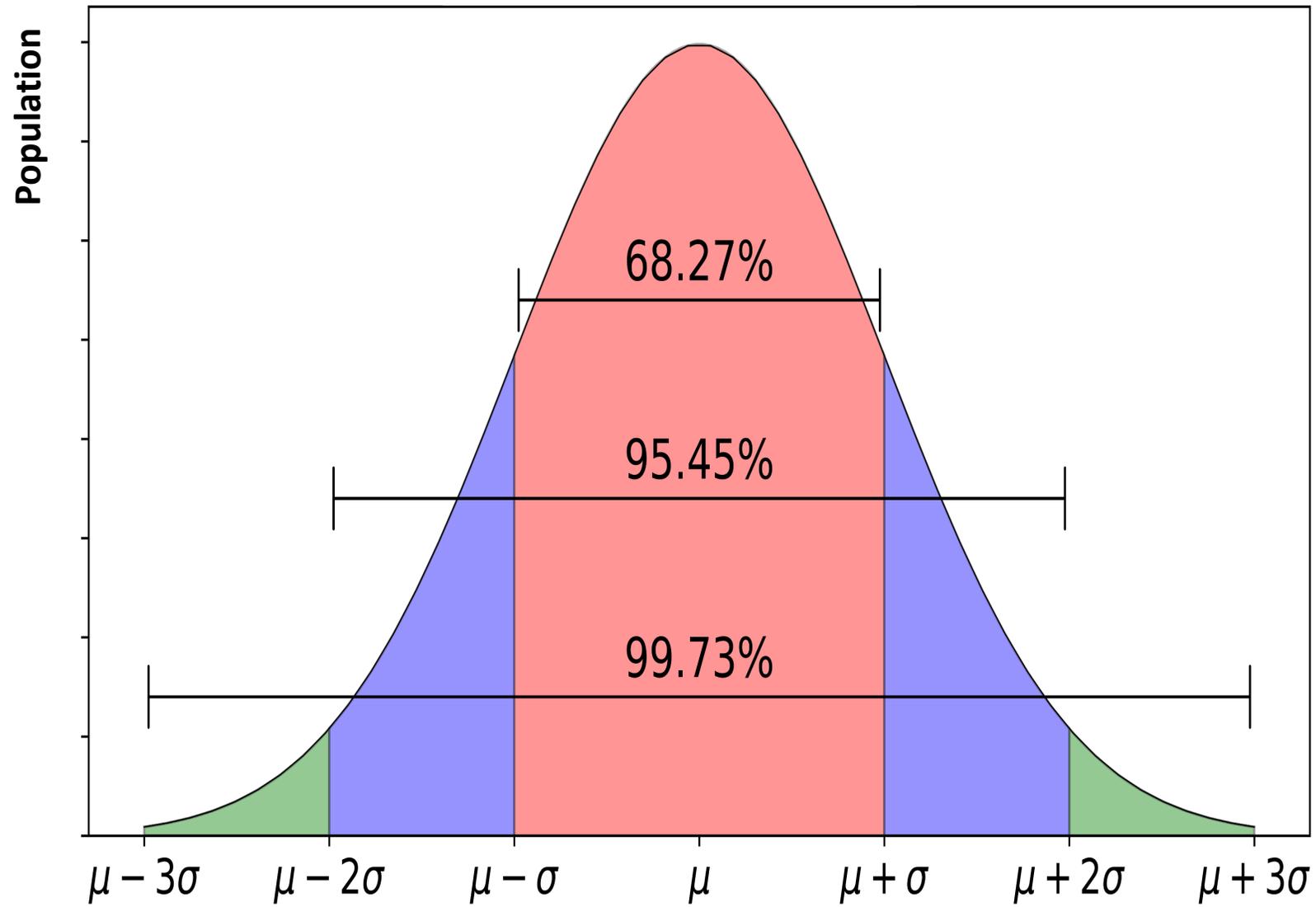
# La distribution normale de Gauss



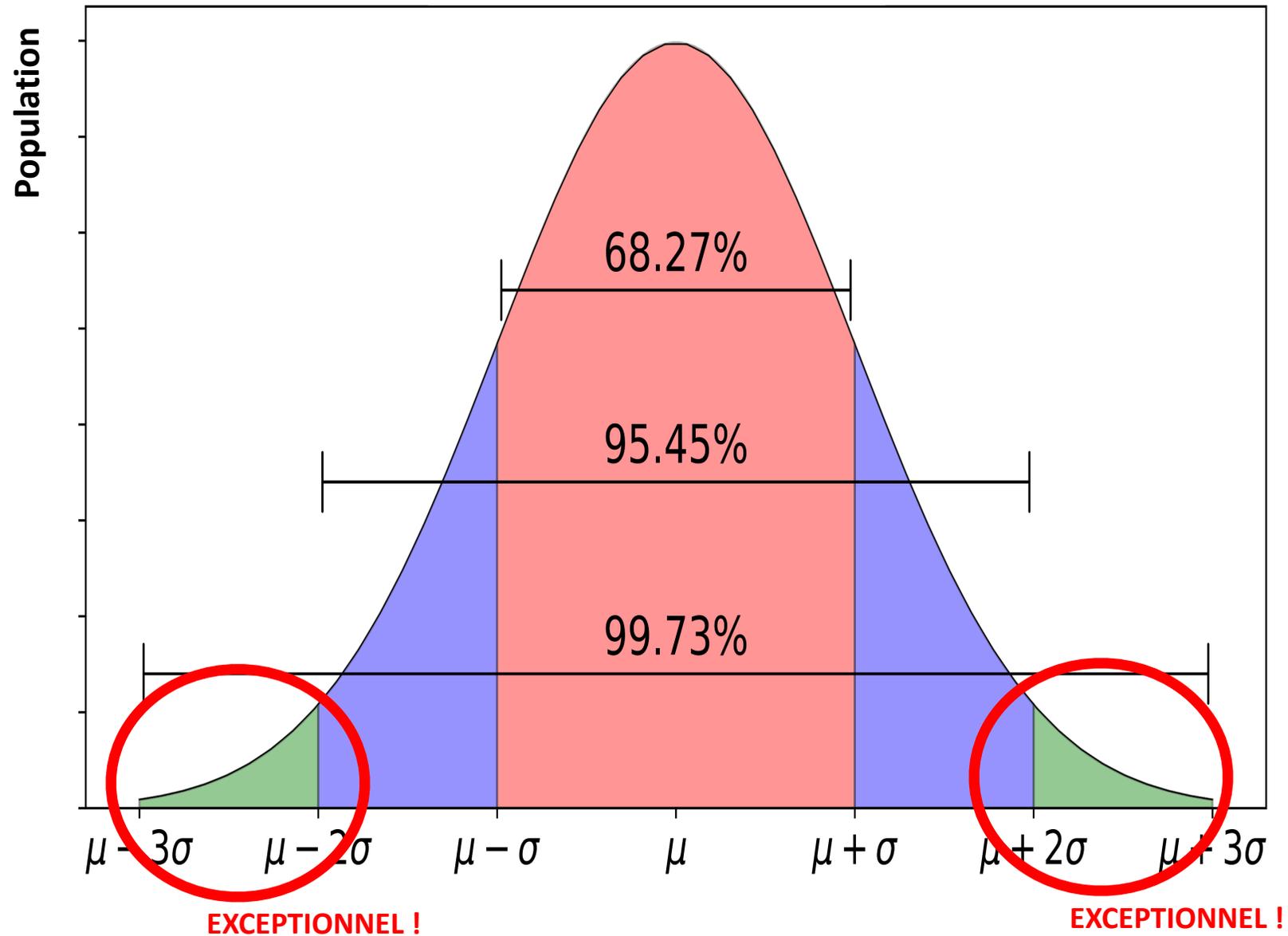
# La distribution normale de Gauss



# La distribution normale de Gauss

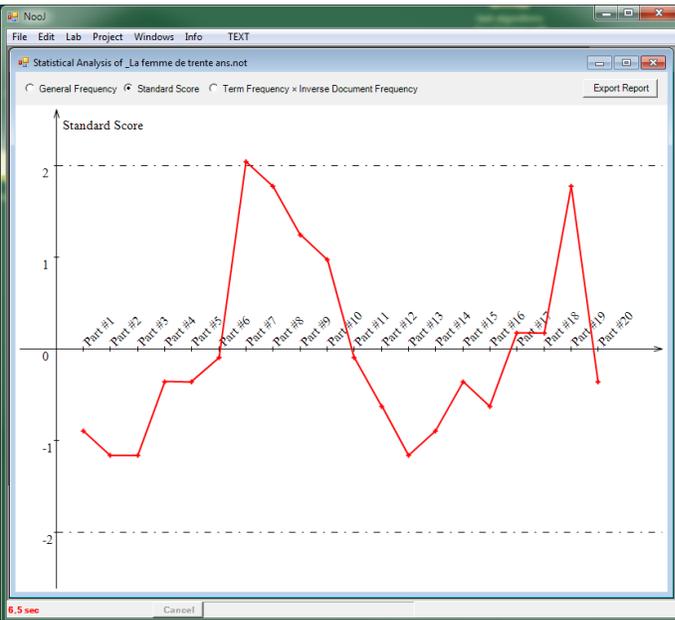


# La distribution normale de Gauss



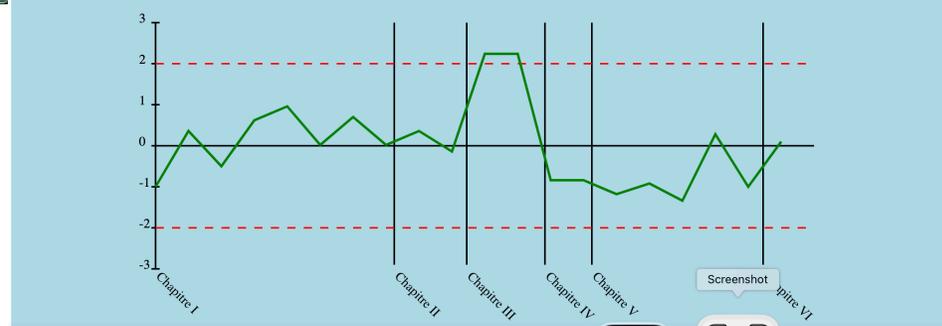
# Exemple d'analyse statistique d'une concordance par score standard

## NooJ



## webnooj

The screenshot shows the webnooj website at 'webnooj.univ-fcomte.fr'. The page title is 'La femme de trente ans (Honoré de Balzac, 1842)'. Below the title, it shows 'Chapitre I : Premières fautes' and a snippet of text: 'Au commencement du mois d'avril 1813, il y eut un dimanche dont la matinée promettait un de ces beaux jours où les Parisiens voient pour la première fois de l'année leurs pavés sans boue et leur ciel sans nuages. Avant midi, un cabriolet à pompe attelé de deux chevaux fringants déboucha dans la rue de Rivoli par la rue Castiglione, et s'arrêta derrière'. A search query 'Amour' is entered, and the 'Appliquer la requête' button is visible.



## ATISHS

The screenshot shows the ATISHS software interface. The main window displays a concordance analysis for 'La tentation de Saint Antoine'. The graph shows the standard score across the text. The y-axis ranges from -2 to 2. The x-axis is labeled '1874 - La tentation de Saint Antoine'. The score starts at -1.2, peaks at 2.1 for '1874 - La tentation de Saint Antoine', and ends at -1.2 for '1874 - La tentation de Saint Antoine'.

# Exemple d'analyse par score standard

- Logiciels NooJ, webnooj et ATISHS
- CONCORDANCE > Statistical Analysis > Standard Score
- Si le score standard est  $> 2$ , alors la mesure est anormalement élevée => très intéressant
- Si le score standard est  $< 2$ , alors la mesure est anormalement faible => très intéressant
- Entre les deux, la mesure est "normale", i.e. inintéressant

# Etude statistique de l'emploi de termes

- Etudier les occurrences des termes *femme* et *fille*
- Etudier les occurrences des termes *homme* et *garçon*
- Comparer les deux

- Etudier les occurrences des termes *femme* et *fille*

The screenshot shows the NooJ web interface. The browser address bar displays 'webnooj.univ-fcomte.fr/language/français'. The page title is 'NooJ' and the language is set to 'français'. The main content area displays the title 'La femme de trente ans (Honoré de Balzac, 1842)' and the chapter 'Chapitre I : Premières fautes'. The text of the chapter begins with 'Au commencement du mois d'avril 1813, il y eut un dimanche dont la matinée promettait un de ces beaux jours où les Parisiens voient pour la première fois de l'année leurs pavés sans boue et leur ciel sans nuages. Avant midi, ...'. The interface includes a search bar with the query '<femme> | <fille>' and a button 'Appliquer la requête'. Below the text, there is a navigation menu with options: 'Concordance', 'Collocations', 'Fréquences', 'Évolution', 'Score standard', 'Test d'indépendance', and 'Analyse factorielle'. The 'Évolution' option is selected, showing a line graph with the y-axis ranging from -3 to 3 and the x-axis labeled with chapters 'Chapitre I' through 'Chapitre VI'. The graph shows a fluctuating line with a peak around Chapitre III and a trough around Chapitre V.

Chapitre	Score standard
Chapitre I	-1.0
Chapitre II	-1.0
Chapitre III	2.5
Chapitre IV	0.2
Chapitre V	-1.5
Chapitre VI	1.2

- Etudier les occurrences des termes *homme* et *garçon*

NooJ

français Outil des humanités numériques - Linguistique de corpus - Environnement de développement linguistique - Moteur linguistique pour les applications de TAL.

Connexion Pour ajouter votre corpus ou des requêtes, contactez max.silberstein@univ-fcomte.fr

### La femme de trente ans (Honoré de Balzac, 1842)

Chapitre I : Premières fautes

Au commencement du mois d'avril 1813, il y eut un dimanche dont la matinée promettait un de ces beaux jours où les Parisiens voient pour la première fois de l'année leurs pavés sans boue et leur ciel sans nuages. Avant midi,

Sélectionner une requête prédéfinie : Amour-Catégories

Entrer une requête : ?

<homme> | <garçon>

Appliquer la requête

Concordance Collocations Fréquences Évolution **Score standard** Test d'indépendance Analyse factorielle

Chapitre	Score standard
Chapitre I	0
Chapitre II	-1.2
Chapitre III	0.8
Chapitre IV	-1.5
Chapitre V	2.5
Chapitre VI	-1.5

- Comparer les occurrences, test d'indépendance  $\chi^2$

NooJ

français

Outil des humanités numériques - Linguistique de corpus - Environnement de développement linguistique - Moteur linguistique pour les applications de TAL.

Connexion Pour ajouter votre corpus ou des requêtes, contactez max.silberstein@univ-fcomte.fr

Sélectionner une requête prédéfinie : Amour-Catégories

Entrer une requête : ?

<homme> | <garçon>)/M | (femme | fille)/F

### La femme de trente ans (Honoré de Balzac, 1842)

Chapitre I : Premières fautes

Au commencement du mois d'avril 1813, il y eut un dimanche dont la matinée promettait un de ces beaux jours où les Parisiens voient pour la première fois de l'année leurs pavés sans boue et leur ciel sans nuages. Avant midi,

Appliquer la requête

Concordance Collocations Fréquences Évolution **Score standard** Test d'indépendance Analyse factorielle

NooJ

français

Outil des humanités numériques - Linguistique de corpus - Environnement de développement linguistique - Moteur linguistique pour les applications de TAL.

Connexion Pour ajouter votre corpus ou des requêtes, contactez max.silberstein@univ-fcomte.fr

Sélectionner une requête prédéfinie : Amour-Catégories

Entrer une requête : ?

<homme> | <garçon>)/M | (femme | fille)/F

### La femme de trente ans (Honoré de Balzac, 1842)

Chapitre I : Premières fautes

Au commencement du mois d'avril 1813, il y eut un dimanche dont la matinée promettait un de ces beaux jours où les Parisiens voient pour la première fois de l'année leurs pavés sans boue et leur ciel sans nuages. Avant midi,

Appliquer la requête

Concordance Collocations Fréquences Évolution Score standard **Test d'indépendance** Analyse factorielle

		$\chi^2$	
		F	M
F	---	51.20	
M	51.20	---	
Degré de liberté = 20			

Probabilité d'indépendance

Degré de liberté	Probabilité d'indépendance										
	1%	2.50%	5%	10%	25%	50%	75%	90%	95%	97.5%	99%
20	37.566	34.170	31.410	28.412	23.828	19.337	15.452	12.443	10.851	9.591	8.260
40	63.691	59.342	55.758	51.805	45.616	39.335	33.660	29.051	26.509	24.433	22.164
60	88.379	83.298	79.082	74.397	66.981	59.335	52.294	46.459	43.188	40.482	37.485
80	112.329	106.629	101.879	96.578	88.130	79.334	71.145	64.278	60.391	57.153	53.540
100	135.807	129.561	124.342	118.498	109.141	99.334	90.133	82.358	77.929	74.222	70.065

# Etude d'un thème

- Chercher les occurrences du thème de la *mort* dans le roman "La femme de trente ans".

# Etude d'un thème

- Chercher les occurrences du thème de la *mort* dans le roman "La femme de trente ans".

<assassiner> | <mourir> | <périr> | <tuer> |  
<veuf> | funérailles | <cadavre> | <meurtre> |  
<fatal> | <mortel> |  
fatalement | mortellement

# Etudier un thème, ex. *la mort*

Locate a pattern in *\_La femme de trente ans*

Pattern is:

- a string of characters:
- a PERL regular expression:
- a NooJ regular expression:

décès | <décéder> | deuil | <fatal> | <mort> | <mortel> | <mourir> | <périr> | funérailles

Concordance for Text *\_La femme de trente ans.not*

Reset Display: 10 characters before, and 10 after. Display:  Matches  Outputs  word forms

Text	Before	Seq.	After
treizième de l'année 1813 . Le surlendemain, Napoléon partait pour cette	fatale	campagne pendant laquelle il allait perdre successivement Bessières et Duroc	
D'ici je puis encore apercevoir l'Empereur ; s'il	périssait	pendant la campagne, je ne l'aurais jamais vu." Le	
Wagram au milieu du feu, à la Moscowa parmi les	morts	, il est toujours tranquille comme Baptiste lui !" Cette réponse à	
allez donc encore m'affliger en me parlant de votre	mort	. J'étais si gaie ! Voulez-vous bien chasser vos vilaines	
te voir fidèle à ton vieux père jusqu'à sa	mort	, j'espérais te conserver près de moi heureuse et brillante	
un médaillon suspendu à son cou par une chaîne de	deuil	, et le portrait de son père lui apparut soudain. A	
sa tristesse dans le chagrin que lui avait causé la	mort	de son père, de qui elle portait encore le deuil	
mort de son père, de qui elle portait encore le	deuil	. Au bout de huit jours, la douairière admira la douceur	
'est mis à étudier sa maladie, que l'on croyait	mortelle	. Insensiblement, il a pris goût à l'anatomie, à la	
ne savent aimer; ils croient que d'aller à la	mort	le lendemain les dispense d'avoir, la veille, des égards	
pour nous. Autrefois, l'on savait aussi bien aimer que	mourir	à propos. Ma nièce, je vous le formerai. Je mettrai	
à souhaiter un divorce, si toutefois vous n'étiez pas	morte	avant d'en venir au désespoir." Julie écoutait sa tante	
elle perdit la comtesse de Listomère-Landon. La vieille dame	mourut	de joie et d'une goutte remontée au coeur, en	
la femme et du mari plus parfait, cette personne était	morte	. Julie sentit toute l'étendue de cette perte. Il n	
à son maître imbécile, parer de fleurs une maison de	deuil	, et afficher le bonheur sur un visage pâli par de	
attendit avec résignation la fin de ses peines en espérant	mourir	jeune. Depuis son retour de Touraine, sa santé s'était	

Query 136/136

5.8 sec Cancel

# Applications : la veille informationnelle

- Veille technologique, ex. Thalès
- Veille épidémiologique, ex. CHU de Marseille
- Analyse historique, ex. les « ennemis » d'alors
- Analyse littéraire, ex. les « ennemis » d'alors
- Analyse psychologique, ex. entretiens d'enfants

# 3. Le projet

# Le projet

## 1. Construire un corpus

- Récupérer un texte, un roman (ou un extrait de journaux)
- Le nettoyer, le diviser en chapitres (ou en article)

## 2. Choisir un logiciel d'analyse statistique ; apprendre à le maîtriser

## 3. Etudier trois thèmes, par exemple :

- *alcoolisme, amour, argent, mort, pauvreté, religion* pour un corpus littéraire
- *emploi, immigration, inflation, insécurité, pandémie* pour un corpus journalistique

## 4. Petite présentation PPT

# Les principaux outils utilisés aujourd'hui

- ATISHS <http://atish.univ-fcomte.fr>
- Hyperbase : <http://hyperbase.unice.fr/hyperbase/>
- IRaMuTeQ : <http://www.iramuteq.org/>
- Lexico : <http://www.tal.univ-paris3.fr/lexico/lex3-10pas/index.htm>
- NooJ : <http://www.nooj4nlp.org> , <http://webnooj.univ-fcomte.fr>,
- Sketch Engine : <https://www.sketchengine.eu/>
- TXM : <http://portal.textometrie.org/demo/>