TAL : Traitement Automatique des Langues

Cours 10

Master LFA, 2011/2012

TAL

- ***Le Traitement automatique de la langue naturelle (TALN)*** ou ***des langues (TAL)*** est une discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle.
- Elle concerne la conception de systèmes et techniques informatiques permettant de manipuler le langage humain dans tous ses aspects.

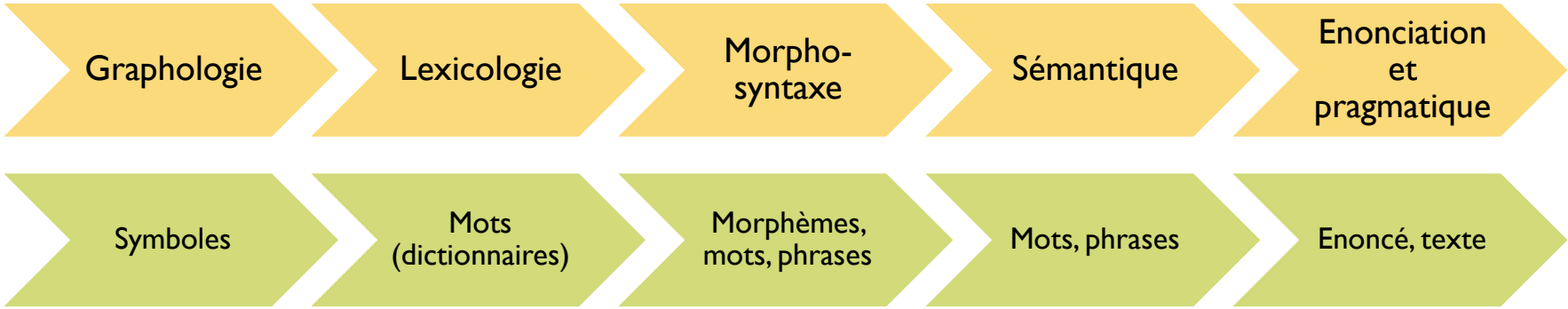
Objectifs

- Traduction automatique : historiquement la première application dès les années 1950.

	Texte écrit	Parole
Extraction / analyse	• correction orthographique • aide à la reformulation • recherche d'information fouille textuelle • reconnaissance d'entités nommées • résolution d'anaphores • classification et catégorisation de documents • reconnaissance de l'écriture manuscrite • annotation morpho-syntaxique / sémantique	reconnaissance vocale reconnaissance du locuteur
Génération	• génération automatique de textes • résumé automatique	synthèse de la parole

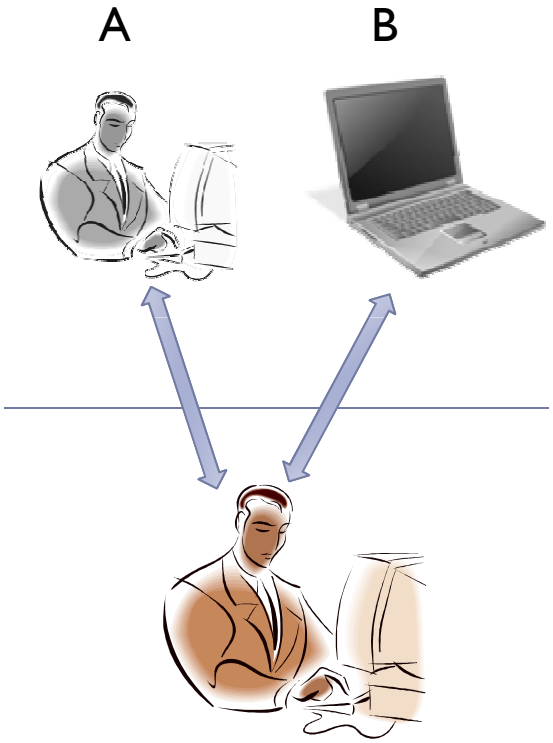
Objets d'étude

- Parole :
  - Onde sonore
  - Analyse par des méthodes statistiques afin de :
    - identifier le locuteur
    - transcrire les paroles en texte
- Texte :
  - Une suite de caractères



Test de Turing (Alain Turing 1950)

- Le test de Turing permet de déceler si un ordinateur est intelligent.
- Un humain est placé est dans une pièce et discute par clavier interposé avec une personne et un ordinateur.
- Le test est considéré comme réussi si l'humain n'arrive pas à déterminer qui est l'autre humain et qui est l'ordinateur.



Qui est l'humain, A ou B ?

Réelle intelligence vs intelligence simulée

- Le test de Turing permet de savoir si le comportement de la machine est semblable à celui d'un être humain. On évalue le comportement extérieur de la machine, ce qui ne garantit pas forcément la présence d'intelligence ou une compréhension de la langue.
- ELIZA – simulation d'entretien avec un psychothérapeute.
  - Testez-le sur : http://elizia.net/
  - Visitez le lien « la supervision » en bas de la page : ici on peut voir la liste des règles qui constituent l' « intelligence » d'Elize :
    - dans la première colonne nous avons des mots-clés ou des situations, et dans la deuxième colonne nous avons les réactions possibles d'Elize.
    - le programme suit des règles simples, sans vraiment comprendre le discours humain.

Traitement du texte

- Au-delà d'imiter un comportement humain, un système devrait être capable de « comprendre » partiellement un texte : pouvoir identifier certains éléments textuels afin de les associer à des significations.
- Pas de théorie linguistique opératoire (pour l'instant).
- Deux approches :
  1. Linguistique informatique : partir de la langue, étudier les exemples (échantillons observables), modéliser et concevoir des algorithmes pour un traitement automatique.
  2. Informatique linguistique : partir de l'informatique et des mathématiques, appliquer les modèles existants à la langue, puis observer le résultat en espérant que ça marche.

Approche statistique

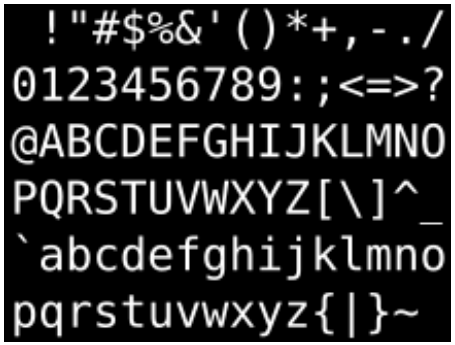
- Il s'appuient sur un ***formalisme mathématique***.
- Applicables à des corpus de très grande taille.
- Indépendantes de la langue.
- Ne nécessitent pas de connaissances linguistiques : le méthodes permettent d'observer la suite de caractères et des mots afin d'en trouver des régularités et pouvoir prédire certaines propriétés.
- Ne permettent pas de comprendre les phénomènes linguistiques.
- Méthodes : n-grammes, apprentissage automatique (voir la suite).
- Les résultats du système, notamment les erreurs, sont difficiles à expliquer et corriger.

Approche utilisant des ressources linguistiques

- Il s'agit de modéliser une certaine partie de la ***connaissance linguistique*** afin de la rendre exploitable par la machine
- Exemple : dans un analyseur morpho-syntaxique, on peut introduire la règle :
  - SI on trouve une occurrence de *« un »* ou *« une »*,
  - ALORS c'est un ***article*** et le mot suivant est un ***nom***.
- Dépendantes de la langue
- Plus difficiles à mettre en place : obligent une *conceptualisation des phénomènes linguistiques*
- Nécessitent plus de temps et plus de travail (par des linguistes)
- Les résultats du système, notamment les erreurs, peuvent être expliqués et corrigés facilement.

Codage des caractères

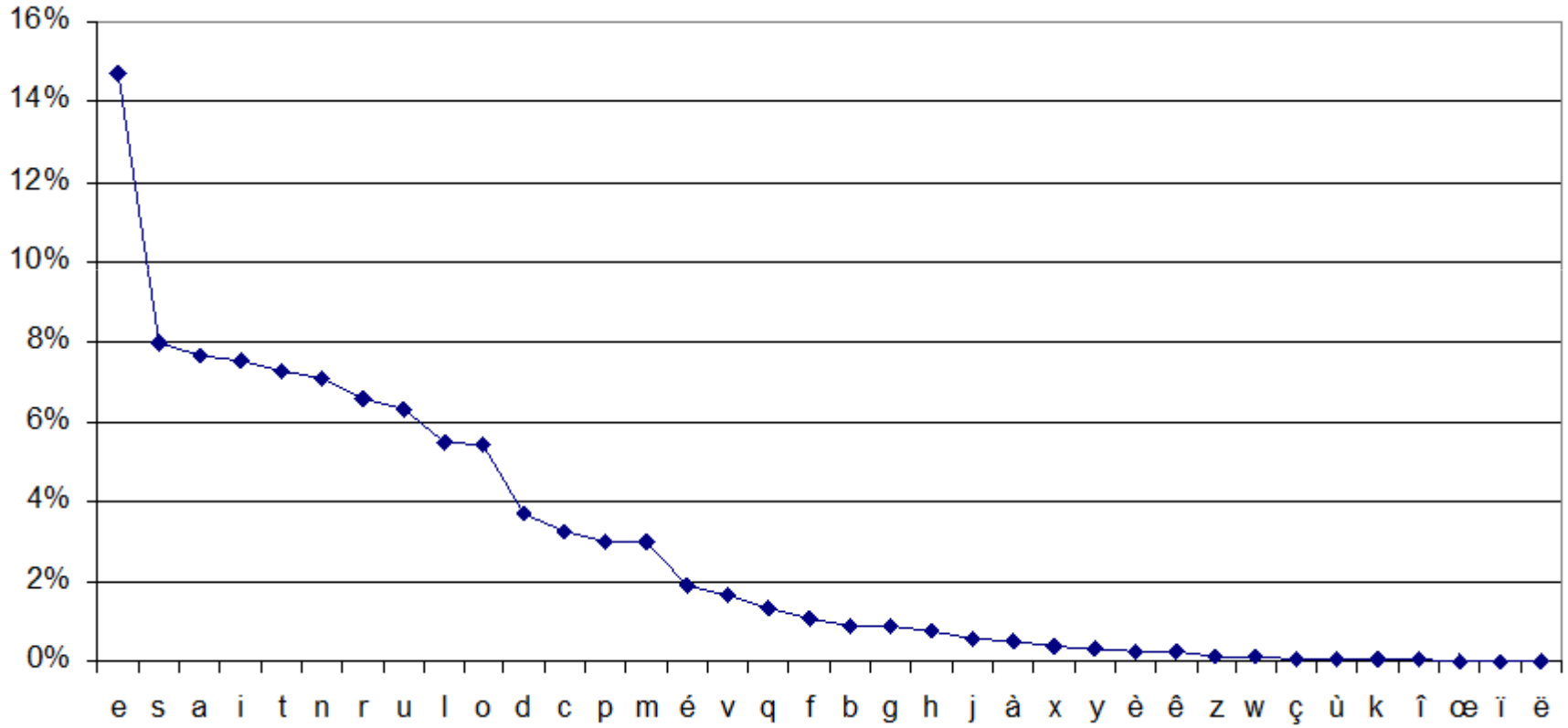
- Les caractères sont stockés suivant un code. Chaque caractère correspond à un numéro.
- Code ASCII (*American Standard Code for Information Interchange*) : comporte 128 codes (stocké sur 7 bits), dont 95 affichables.



- Windows-1252 ou CP1252 : utilisé par Windows dans les principaux langues d'Europe de l'Ouest (dont le français).
- UTF-8 :
  - chaque caractère est stocké sur 1, 2, 3 ou 4 octets
  - compatible avec ASCII
  - permet de représenter tous les alphabets, notamment les langues asiatiques et le cyrillique
  - actuellement autour de 100 000 caractères (extensible)
- UTF-16, UTF-32

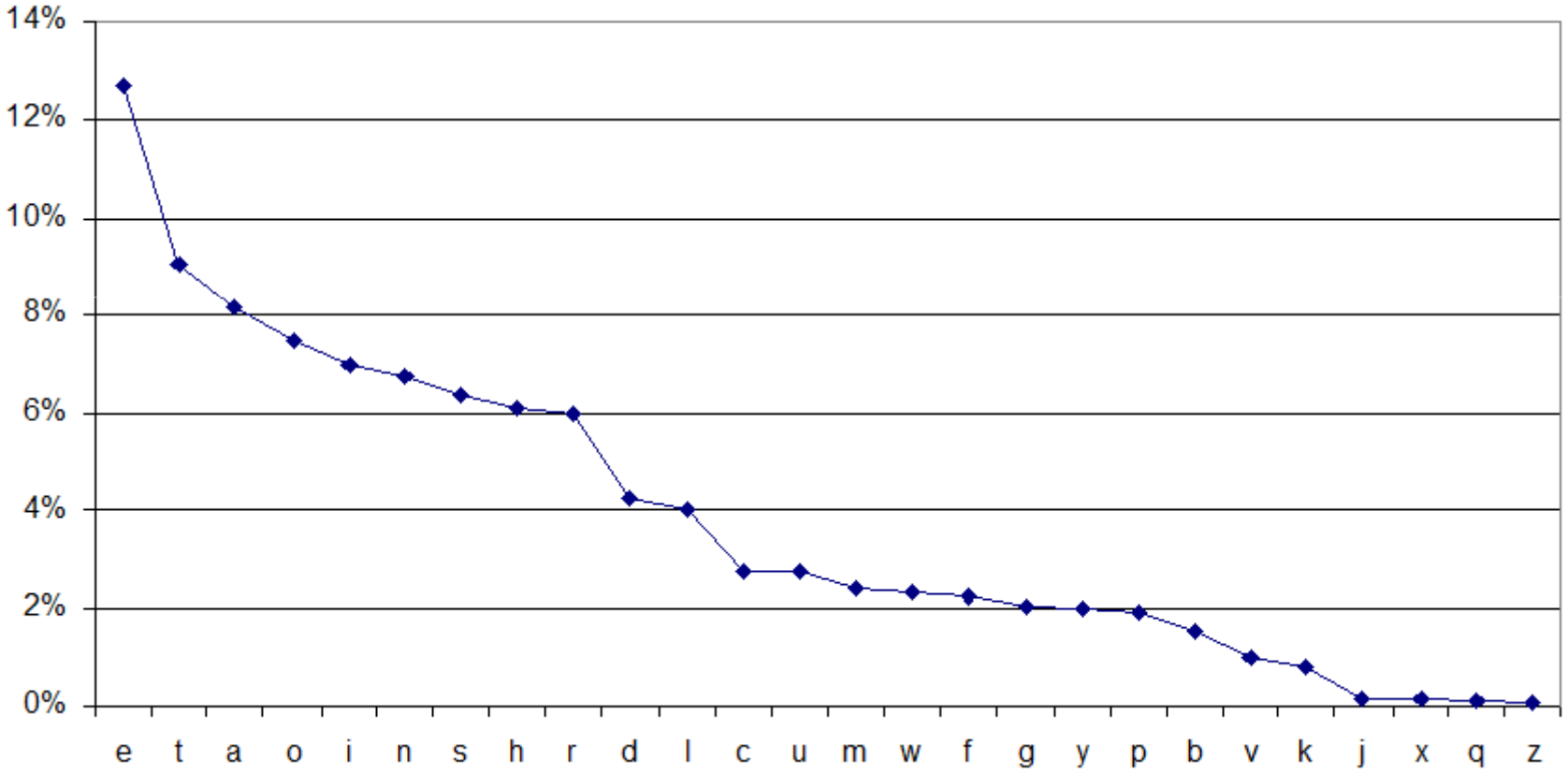
La langue comme un objet statistique

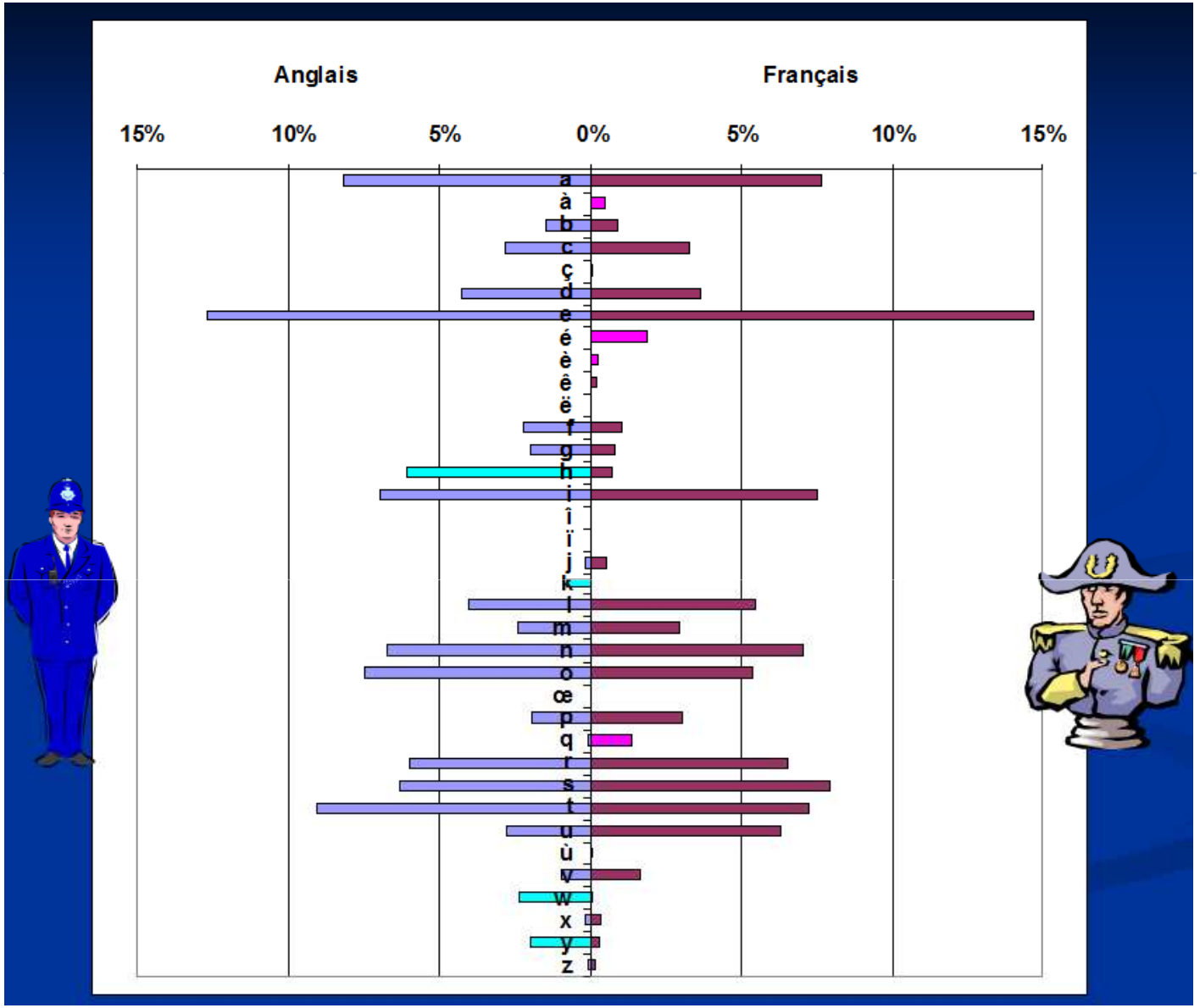
- Fréquence des caractères en français :



La langue comme un objet statistique

- Fréquence des caractères en anglais :





Source : cours de Jean Veronis

La langue comme un objet statistique

- Avec ces données, en observant les fréquences des caractères dans un texte, on peut détecter automatiquement la langue.
- Cependant, cette méthode dépend de la taille du document et du type de texte :
  - un texte avec beaucoup de verbes en 2e personne : plus de « z »
  - le menu d'un restaurant : plus de « € »
  - un texte sur la carrière de Sarkozy : plus de « z » et « y »

N-grammes

- Les caractères sont des ***uni-grammes***.
- ***Bi-grammes*** : des combinaisons de deux caractères :
  - informatique -> in nf fo or rm ma at ti iq qu ue
- ***Tri-grammes*** : des combinaisons de trois caractères :
  - informatique -> inf nfo for orm rma mat ati tiq iqu que
- …
- ***N-grammes***

- Pour détecter la langue d'un document, on va observer les fréquences des n-grammes dans le texte (plutôt que les simples uni-grammes).

Bi-grammes les plus fréquentes

Français	Anglais	Allemand	Italien	Espagnol	Portugais
on	th	en	di	de	de
es	on	er	on	en	es
de	an	ch	ri	er	to
te	he	ei	er	on	da
nt	er	un	al	ci	os
re	nd	de	to	es	re
en	in	nd	ta	re	en
le	ti	ge	ne	os	er
it	al	re	in	io	te
er	re	in	re	la	ra
et	io	ie	it	ra	nt
ti	en	te	io	na	em
ou	ri	ng	de	ec	do
io	of	he	li	al	di
la	or	ne	en	ad	it
oi	at	ht	ni	da	al
ne	it	ic	tt	to	ad
me	to	be	la	nt	co
ro	ed	it	ll	ie	ei
ns	nt	sc	el	el	as

Tri-grammes les plus fréquentes

Français	Anglais	Allemand	Italien	Espagnol	Portugais
ion	the	der	ion	ion	ent
tio	and	und	zio	cio	ito
ent	ion	ein	ell	rec	eit
oit	tio	ung	one	ere	dir
ati	ati	cht	lla	der	ire
roi	igh	ich	rit	ien	rei
dro	ght	sch	itt	cho	ção
men	rig	che	del	ent	ade
tou	ent	ech	iri	ech	dad
con	ver	die	dir	aci	men
res	one	rec	ess	ona	nte
que	all	ine	ent	nte	dos
les	eve	eit	azi	con	ess
des	ery	gen	tto	ene	con
eme	his	ver	ere	tod	tod

Source : Jean Veronis

Identification de la langue

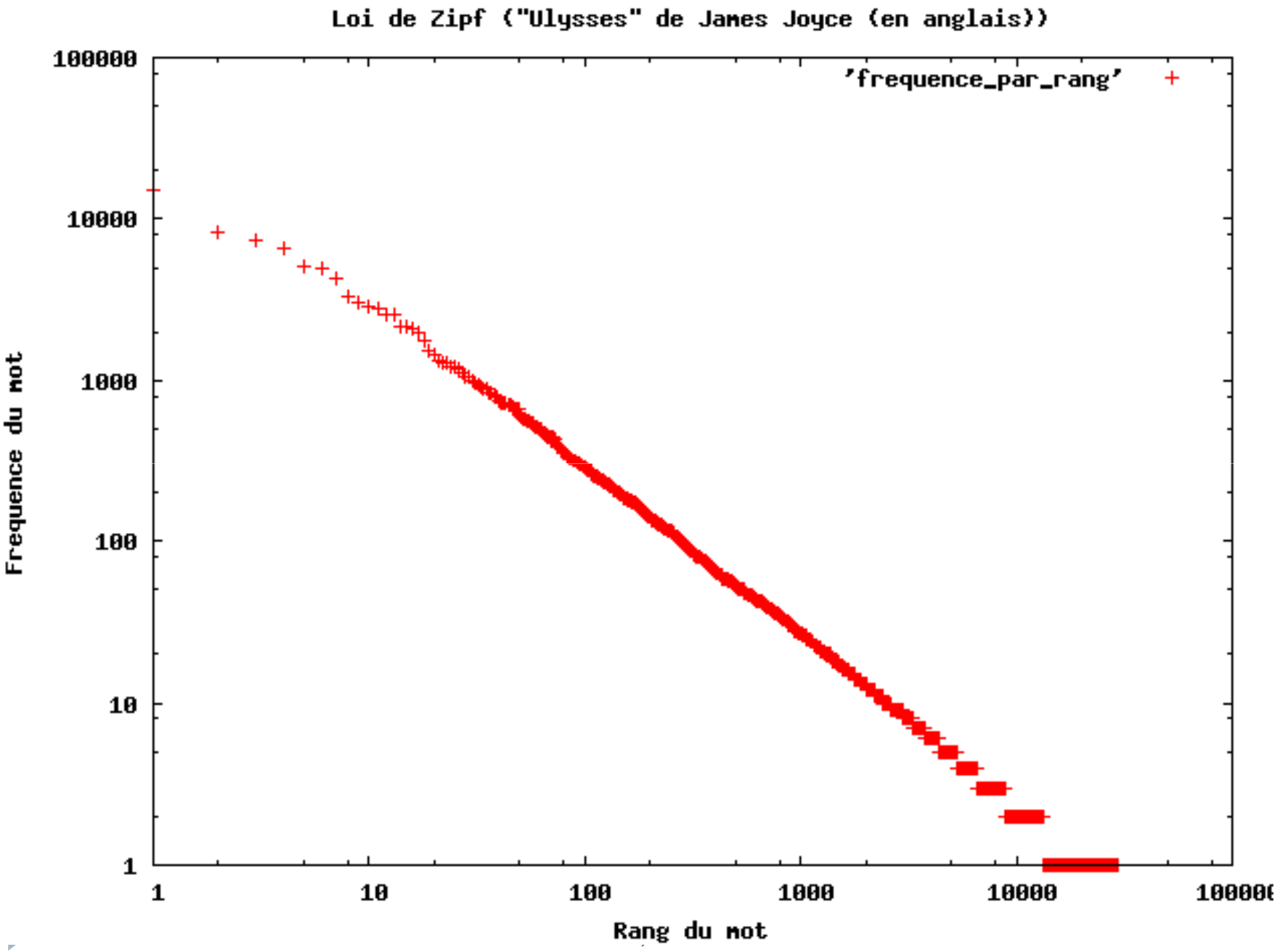
- C'est un des problèmes les plus faciles en TAL : bons résultats sur des textes longs de plusieurs phrases.
- Problème :
  - langues proches (tchèque et slovaque, anglais américain et anglais britannique)
  - la différence entre une langue et un dialecte.

- http://labs.translated.net/language-identifier/ - identification de la langue en ligne
  - pour le tester vous pouvez copier du texte à partir des différentes versions de Wikipédia.

La langue comme un objet statistique

- Nous avons vu la distribution des caractères, mais que peut-on dire sur la fréquence des mots dans un texte ?

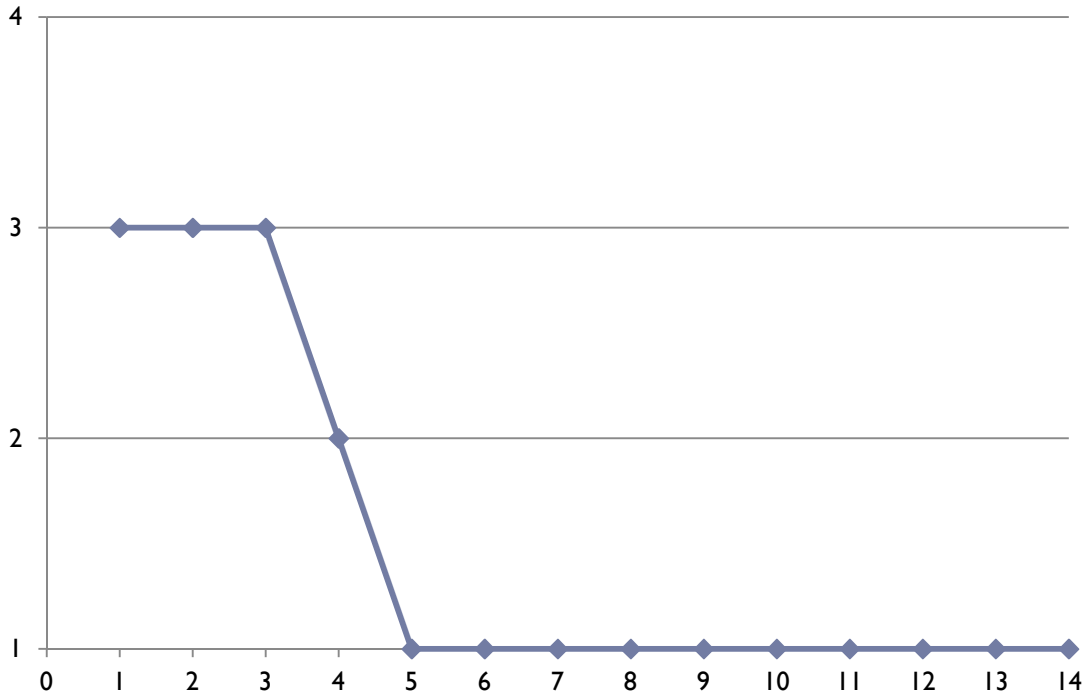
- La loi de Zipf (1949) : la fréquence d'un mot dans un texte est inversement proportionnelle à son rang.
  - C'est une loi empirique établie à partir des observations sur des textes.
  - Cas particulier de la loi de Pareto en économie.


Loi de Zipf ("Ulysses" de James Joyce (en anglais))
'frequence_par_rang'
Frequence du mot
Rang du mot
100000
10000
1000
100
10
1
1
10
100
1000
10000
100000

Loi de Zipf : exemple

- « *Les hommes ne veulent pas ce qu'ils font, mais ce en vue de quoi ils font ce qu'ils font.* » (Platon)

ce	3
ils	3
font	3
qu'	2
les	1
hommes	1
ne	1
veulent	1
pas	1
mais	1
en	1
vue	1
de	1
quoi	1

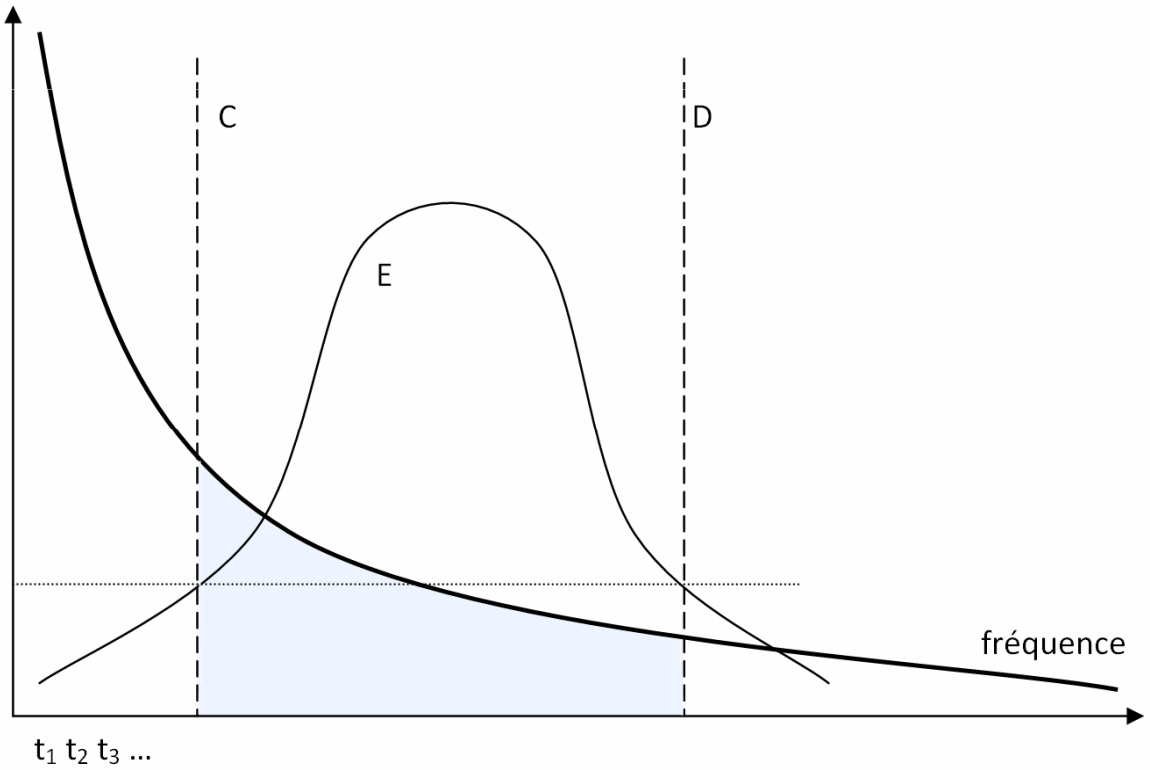


Loi de Zipf : applications

- Théorie de l'information : Claude Shanon (1948) essaye de quantifier l'information qui est transmise par chaque mot.
- Dans la recherche d'informations, pendant l'indexation on supprime tous les mots insignifiants du document pour ne garder que les mots les plus « informatifs ». On suppose alors que :
  - les mots les plus fréquents sont les moins informatifs. Ce sont généralement des mots grammaticaux ou des mots d'usage courant, qui apportent peu d'informations sur le contenu du document.
  - De même, les termes de fréquence faible ne sont pas pertinents pour décrire le contenu du document.

En recherche d'informations : Luhn 1958

- La courbe « fréquence » exprime la loi de Zipf.
- La courbe *E* exprime ***l'informativité*** des mots.
- Les mots ayant des fréquences moyennes sont les plus informatifs (en bleu ci-dessous), ils sont considérés comme représentatifs du document (suite à la théorie de Shannon).



N-grammes pour les mots

- Une ***n-gramme*** est une suite de *n* mots.
- On les utilise par exemple pour faire des analyses morpho-syntaxiques.

- Exemple : TreeTagger
  - http://web4u.setsunan.ac.jp/Website/TreeOnline.htm - version en ligne en anglais
  - C'est un analyseur morpho-syntaxique, utilisé dans de nombreux projets en TAL.
  - A partir d'une phrase (en anglais mais il existe aussi pour d'autres langues), il produit la liste des mots avec leurs catégories syntaxiques.
  - Pour connaitre la liste des étiquettes qu'il attribue : http://courses.washington.edu/hypertxt/csar-v02/penntable.html

Lexique : WordNet

- WordNet : une base de données lexicale. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise.
- Des versions de WordNet pour d'autres langues existent (EuroWordNet), mais la version anglaise est la plus complète à ce jour.
- Contient : synonymes, antonymes, hyperonymes, hyponymes, méronymes, holonymes.
- http://wordnetweb.princeton.edu/perl/webwn

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: table Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **table**, tabular array (a set of data arranged in rows and columns) *"see table 1"*
- S: (n) **table** (a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs) *"it was a sturdy table"*
- S: (n) **table** (a piece of furniture with tableware for a meal laid out on it) *"I reserved a table at my favorite restaurant"*
- S: (n) mesa, **table** (flat tableland with steep edges) *"the tribe was relatively safe on the mesa but they had to descend into the valley for water"*
- S: (n) **table** (a company of people assembled at a table for a meal or game) *"he entertained the whole table with his witty remarks"*
- S: (n) board, **table** (food or meals in general) *"she sets a fine table"; "room and board"*

Verb

- S: (v) postpone, prorogue, hold over, put over, **table**, shelve, set back, defer, remit, put off (hold back to a later time) *"let's postpone the exam"*
- S: (v) **table**, tabularize, tabularise, tabulate (arrange or enter in tabular form)

Lexique : concordance

- Permet de visualiser les différents contextes d'un mot dans un corpus.
- http://www.lextutor.ca/concordancers/concord_f.html

Home> Concordancers> French input [<Back (keep settings)] Colloc summary

Concordance for *equals* **INTELLIGENCE** sort 1 wds **left** of key Dictionnaire: Fren_Eng Get

extract [All | none | any 10 | 20 | 30 | 50] PARAMS: intelligenc equals Sort: 1 wd/s Left of key >> Change

48 hits Click keyword for more context

001. 'elle peut provoquer sont également analysés avec INTELLIGENCE et sensibilité. KARINE NAKACHE |d10 |
002. s à dire, et soient capables de les exprimer avec INTELLIGENCE et bon sens, n'avait même jamais effl
003. 'affectivité froide et superficielle. Cet homme d' INTELLIGENCE "normale" n'éprouverait, en fait, "au
004. é. En prime de ce petit livre plein de verve et d' INTELLIGENCE, Krugman nous offre une analyse origi
005. irecte dans les crimes évoqués. Impressionnante d' INTELLIGENCE et de maîtrise dans l'adversité, Winn
006. ne dramatisée d'une fiction, avec ce qu'il faut d' INTELLIGENCE, d'imagination, de courage et de savo
007. on Rachmaninov sont des modèles de réalisation, d' INTELLIGENCE et de goût, que son Schumann, en reva
008. sseur de littérature à Pise. Tous deux lumineux d' INTELLIGENCE et de raffinement complètent mutuelle
009. un bâtisseur. Il déploie dans ces rôles-là plus d' INTELLIGENCE que de brutalité, ayant tôt compris q
010. ision riche et vivante. Virtuosité, lisibilité et INTELLIGENCE du texte, souplesse et expression, ce
011. ques ECONOMIE Contre la théorie pop Avec verve et INTELLIGENCE Paul R. Krugman tord le cou à nombre
012. voirs qu'il sert non seulement son exceptionnelle INTELLIGENCE, mais aussi ce qu'il est convenu d'ap
013. onde, large dans le fortissimo, legato, fluidité, INTELLIGENCE, charme, poésie. Pour finir, Kusmin j
014. Carlos Manuel de Cespedes, un homme d'une grande INTELLIGENCE et dont les amitiés dans tous les sec
015. elle d'une personnalité hors du commun. La grande INTELLIGENCE de Thérèse, son extrême sensibilité f
016. à la fois. La triple expérience des boppers de l' INTELLIGENCE, de la joie et du risque sera dans ce

Apprentissage automatique

- C'est une méthode (mathématique, statistique) qui a été développée initialement pour la reconnaissance des images (par ex. les visages dans Picasa).
- Appliquée à la langue elle permet d'obtenir des résultats rapidement, sans devoir modéliser les phénomènes linguistiques : la machine apprend elle-même les règles de la langue en observant des échantillons fournis pour l'apprentissage (corpus d'apprentissage), à l'image d'un bébé qui apprend sa langue maternelle.
- Souvent utilisée par des informaticiens sans formation en linguistique.
- On peut obtenir des résultats plus ou moins satisfaisants (selon la tâche) qui sont très difficiles à améliorer.

Apprentissage automatique

1. Phase d'apprentissage :
   - On fournit à la machine un corpus d'apprentissage : un grand corpus avec des textes et des analyses correctes.
     - Par exemple, pour un analyseur morpho-syntaxique, c'est un grand corpus de phrases qui sont analysées (on connaît la catégorie de chaque mot).
   - L'algorithme calcule des fréquences (souvent des uni-, bi- ou tri-grammes) dans le corpus d'apprentissage.
2. Phase de test :
   - On teste le système sur un petit corpus de test. Pour ce corpus on connaît l'analyse correcte, mais la machine doit l'obtenir uniquement à partir du texte.
   - On compare la sortie du système avec l'analyse correcte.
   - Si le résultat n'est pas satisfaisant, on recommence l'apprentissage avec un nouveau corpus plus grand.

Exemple

Phrase à analyser :
« Julie ouvre la porte. »

4 mots : *« julie », « ouvre », « la », « porte »*

« julie » : ??? (ce mot n'était pas dans mon corpus d'apprentissage)

« ouvre » : dans le corpus d'apprentissage tous les occurrences de « ouvre » était des verbes, donc VERBE

« la » : dans le corpus d'apprentissage 89% des *« la »* était des articles définis, donc ARTICLE DEFINI (c'est le plus probable)

« *porte* » : dans le corpus d'apprentissage c'était tantôt un verbe tantôt un nom. Mais dans 95% des cas où « *porte* » était précédé de « *la* », c'était un nom, donc NOM.

« Julie	*ouvre*	*la*	*porte. »*
???	*VERBE*	*A. DEF*	*NOM.*

Apprentissage automatique : problèmes

- La qualité du système dépend très fortement de la taille du corpus d'apprentissage.
  - Pour un analyseur morpho-syntaxique viable, il faut un corpus d'apprentissage de plusieurs centaines de milliers de mots analysés.
- Il est difficile de produire un grand corpus d'apprentissage de qualité :
  - analyse manuelle très couteuse
  - analyse manuelle, donc erreurs humaines ou incohérences entre les analyses produites par différentes personnes.
- Pour certaines tâches, même avec un très grand corpus on obtient des résultats médiocres : le sens dans la langue est le produit d'opérations complexes. Un modèle qui considère le texte comme une suite de mots ne peut pas tenir compte de cette complexité.

Exemple

- Un système qui cherche à extraire les événements à partir d'un texte narratif, pour produire un résumé.
- Le système observe la suite de mots :

  « cinq minutes plus tard le train déraillait ».

- Verbe en imparfait : alors un événement dans le passé ? Peut-on en déduire que le train a déraillé ?

*« **Sans** l'intervention du conducteur,*

qui tenait la vitesse de sa réaction à son entrainement de commando dans sa jeunesse,

cinq minutes plus tard le train déraillait ».

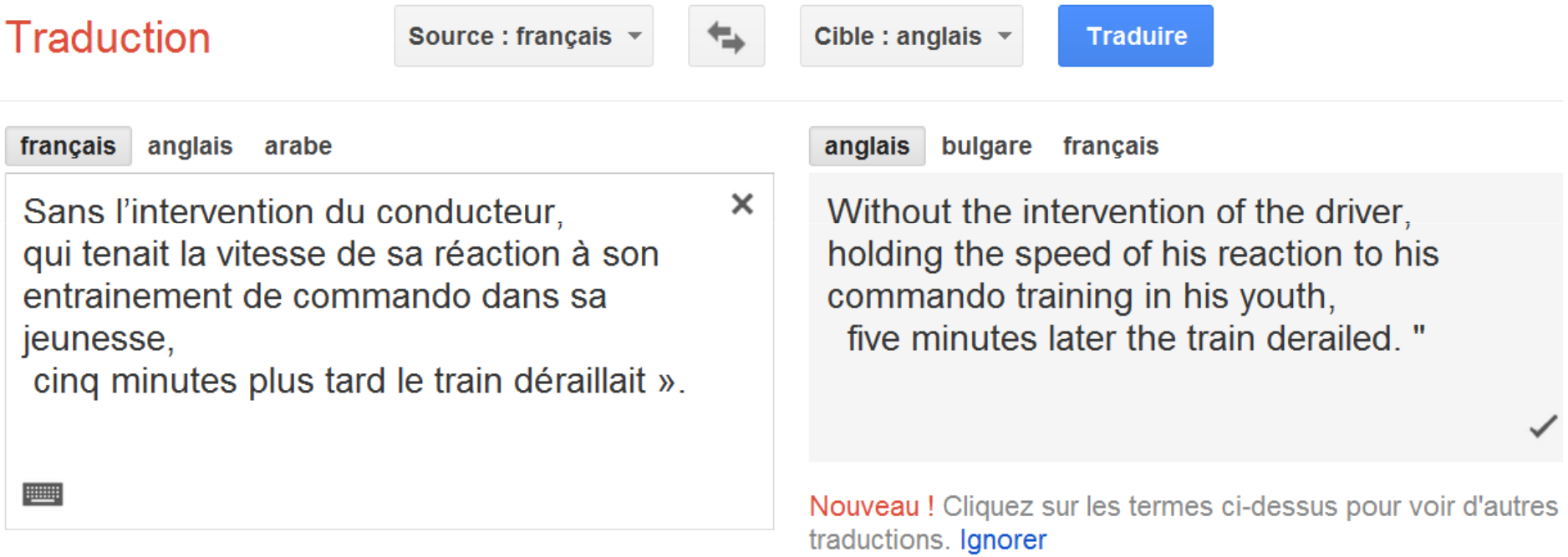
Pour analyser l'événement, aucun algorithme d'apprentissage automatique ne peut tenir compte d'un indice qui se trouve à distance 26 mots du verbe ! Surtout que l'on pourrait aussi avoir :

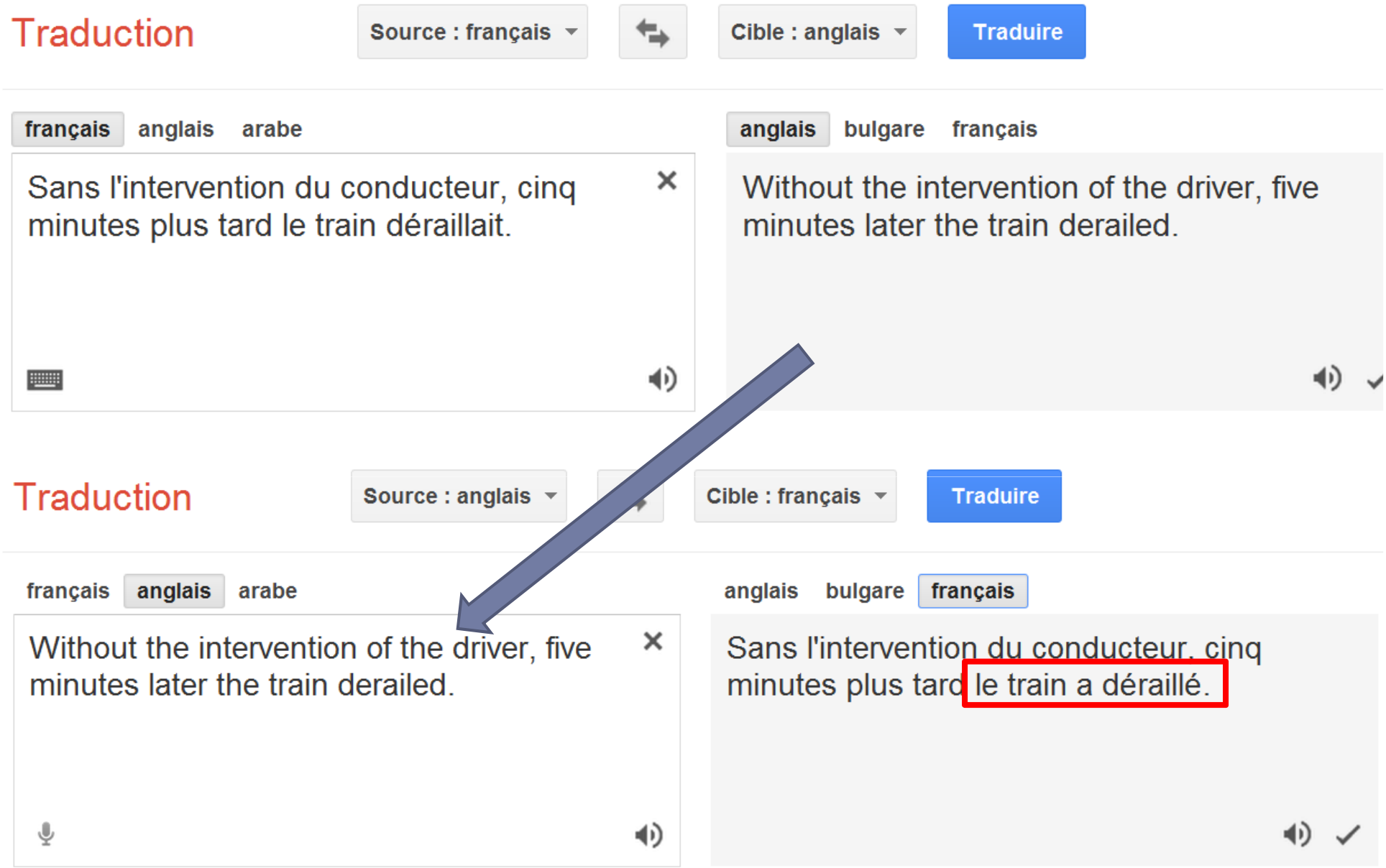
*« **Malgré** l'intervention du conducteur …,*

cinq minutes plus tard le train déraillait ».

Google Translate

- http://translate.google.fr/




Traduction
Source : français
Cible : anglais
Traduire
français anglais arabe
Sans l'intervention du conducteur, cinq minutes plus tard le train déraillait.
anglais bulgare français
Without the intervention of the driver, five minutes later the train derailed.
Traduction
Source : anglais
Cible : français
Traduire
français anglais arabe
Without the intervention of the driver, five minutes later the train derailed.
anglais bulgare français
Sans l'intervention du conducteur, cinq minutes plus tard le train a déraillé.

Google Translate

- Utilise « tout le Web » comme corpus d'apprentissage.
  - Corpus parallèles pour la traduction.
- Problème : ce corpus est « pollué » par les propres traductions de Google Translate.
- Utilise le retour des utilisateurs : possibilité donner son avis sur la traduction.

Grammaire vs modèle linguistique

- La grammaire cherche à décrire toutes les formes dans la langue. C'est une classification à des fins pédagogiques.
- Si on considère que l'imparfait est un temps de passé (comme nous l'apprend la grammaire), comment alors expliquer la phrase :

  « Qu'est-ce qu'il y avait à la télé ce soir ? »

  qui serait acceptée par tous les locuteurs français comme décrivant un événement futur (ce soir).

- Conséquence : la grammaire ne peut être un point de départ pour construire un système de TAL.

Evaluation

- Pour évaluer un système en TAL, il faut connaître la sortie souhaitée du système :
  - Ce n'est pas toujours facile. Par ex. pour une recherche d'informations sur le Web, on ne peut pas connaître tous les documents pertinents : s'il y avait un moyen de les retrouver, on aurait alors un système parfait, et donc plus besoin de l'évaluer.
  - Pour un système de résumé automatique, il existe plusieurs réponses correctes, ou chaque résumé est plus ou moins correcte.
  - Pour un système de traduction automatique, plusieurs traductions sont possibles en sachant que la traduction 100% correcte n'existe pas (toute traduction entraîne une modification du sens).

Mesures d'évaluation

- Supposons que l'on connaît la sortie souhaitée, pour un système de classification de documents.

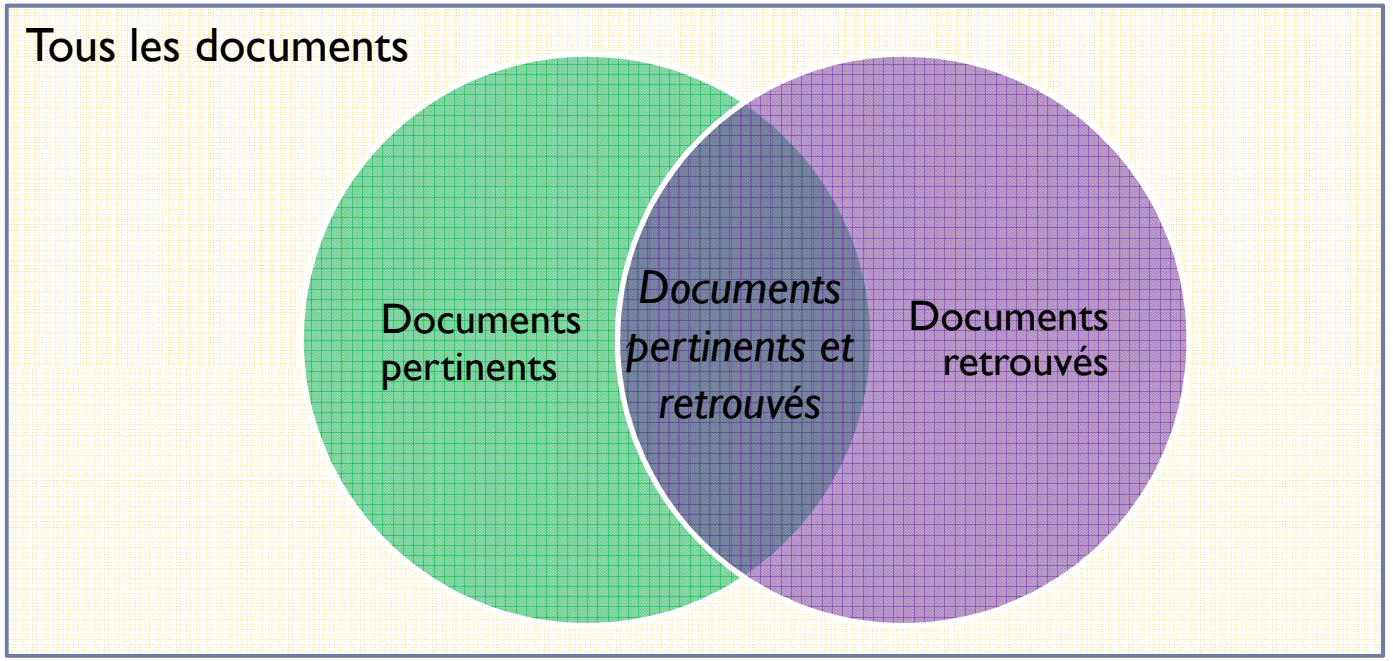
- Dans ce cas, l'évaluation s'effectue en utilisant les mesures de ***précision*** et ***rappel***.

$$Précision_i = \frac{documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ documents\ attribués\ à\ la\ classe\ i}$$

$$Rappel_i = \frac{documents\ correctement\ attribués\ à\ la\ classe\ i}{nombre\ de\ documents\ appartenant\ à\ la\ classe\ i}$$

Mesures d'évaluation

- En recherche d'informations :



$$Précision = \frac{Pertinents\ et\ retrouvés}{Retrouvés}$$

$$Rappel = \frac{Pertinents\ et\ retrouvés}{Pertinents}$$

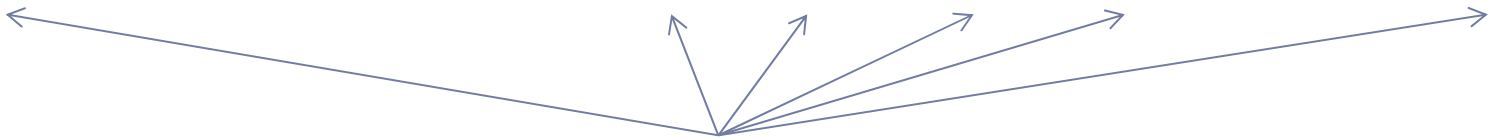
Mesures d'évaluation

- Précision et rappel :
  - ce sont des nombres réels entre 0 et 1
  - 0,5 de précision signifie que le système de trompe pour la moitié de documents qu'il a identifié.
  - 0,75 de précision signifie que le système de trompe pour 25% des documents qu'il a identifié.
  - 0,5 de rappel signifie que le système a identifié correctement la moitié de tous les documents et il n'a pas pu reconnaître l'autre moitié.
- Système parfait :
  - Précision = 1 & Rappel = 1

Précision et rappel : exemple

- Un système qui doit identifier parmi un ensemble de documents ceux qui parlent de sport.
- On le teste sur 10 documents :

Sujet réel	Sport	Sport	Autre	Autre	Autre	Sport	Autre	Sport	Autre	Sport
Document	1	2	3	4	5	6	7	8	9	10
Sortie système	Sport	Autre	Sport	Sport	Autre	Sport	Autre	Sport	Sport	Sport



Réponses correctes

Quelles sont les valeurs du rappel et de la précision ?

- La prochaine fois :
  - expressions régulières
  - annotation sémantique
  - recherche d'informations sémantique