Introduction au Traitement Automatique de la Langue

Master SITIS

3 octobre 2017

C'est quoi ?

Traitement du texte libre (courriers, rapports, résultats de bio ...)

Tirer des informations d'un texte ou d'une collection de textes semi ou pas structurés

C'est quoi ?

Problèmes que ce cours vous aidera à résoudre

- Vous cherchez "cancer du sein" dans un article mais ne trouvez pas la mention de "carcinome du sein"
- Un de vos documents s'est retrouvÃ© plein de problÃ¨mes ressemblant Ã Ã§a...
- Vous cherchez "déprimer" dans un texte mais vous ne trouvez pas "déprimait"
- Vous voulez trouver toutes les dates dans un texte

Applications du TAL

- correction orthographique
- traduction automatique
- moteurs de recherche (indexation)
- résumé automatique
- génération de texte (chatbot)
- filtre anti-spam
- détection du plagiat
- ...

Quelques notions de linguistique

- phonétique et phonologie
- morphologie
- syntaxe
- sémantique

Morphologie

- composition des mots à partir d'unités de sens
- unité : morphème

Morphologie

radical	soign-	faire	porte - monnaie
affixes	▶ -er ▶ -er -ai ▶ -ant ▶ -ant -e	▶ re- ▶ dé-	

Morphologie

forme lemmatisée	forme fléchie
soigner	soigne, soignerai, soignons..
beau	belle, belles, beaux

Morphologie

- mots lexicaux
  - noms
  - adjectifs
  - verbes
  - adverbes
- mots grammaticaux (ou mots vides, mots outils, "petits mots")
  - déterminants
  - pronoms
  - conjonctions
  - prépositions
  - ...

Morphologie

- mots lexicaux
  - désignent des objets de la réalité, des concepts
  - on en crée régulièrement de nouveaux, impossible d'en faire une liste exhaustive
- mots grammaticaux
  - servent à construire la phrase, ne réfèrent pas à un concept, n'ont pas de contenu lexical
  - on en crée très rarement de nouveaux, il est possible de les lister, mais la liste exacte dépend de vos besoins

Morphologie

Lexiques généraux, avec informations morphologiques et autres

- Lefff http://www.labri.fr/perso/clement/lefff/
- lexique.org
- glàff http://redac.univ-tlse2.fr/lexiques/glaff.html

Etiquetage morphologique

J	ai	eu	une	crise	d'	angoisse
pronom	verbe	verbe	det	nom	prep	nom

Etiquetage morphologique

J	ai	eu	une	crise	d'	angoisse
pronom	verbe	verbe	det	nom	prep	nom

Elle	aura	une	migraine	avec	aura
pronom	verbe	det	nom	prep	nom

Etiquetage morphologique + lemmatisation

outil : Treetagger

étiquetage morphologique : Utile pour désambiguiser, ou si vous voulez supprimer tous les mots grammaticaux, par exemple pour construire un index.

lemmatisation : Utile pour rechercher un mot dans le texte, construire un index...

Syntaxe

- composition de la phrase à partir des mots
- unité : syntagme

Syntaxe

Approche sans tenir compte de la syntaxe, "sac de mots"

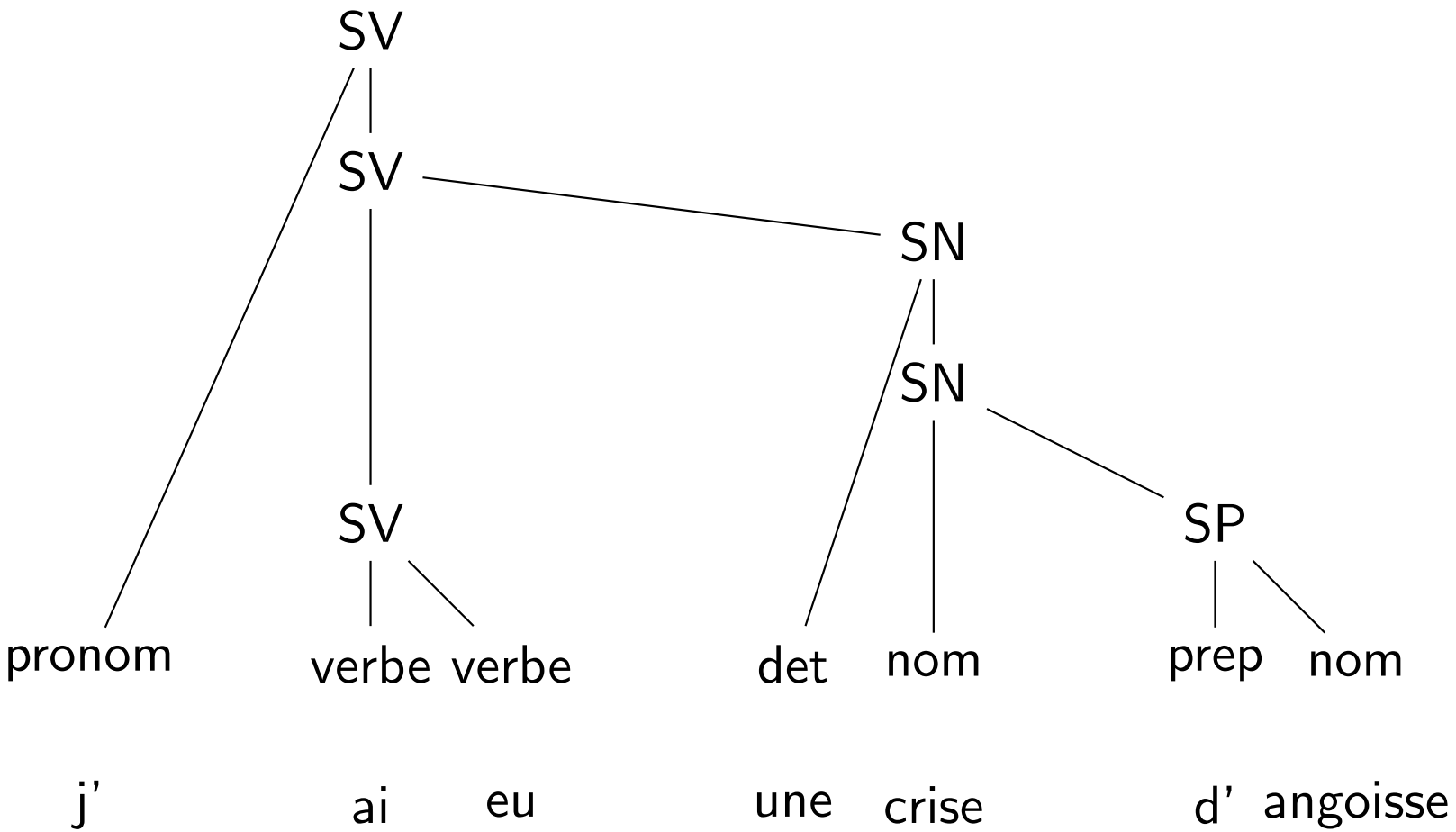
Je prends du valium pour dormir.

Syntaxe

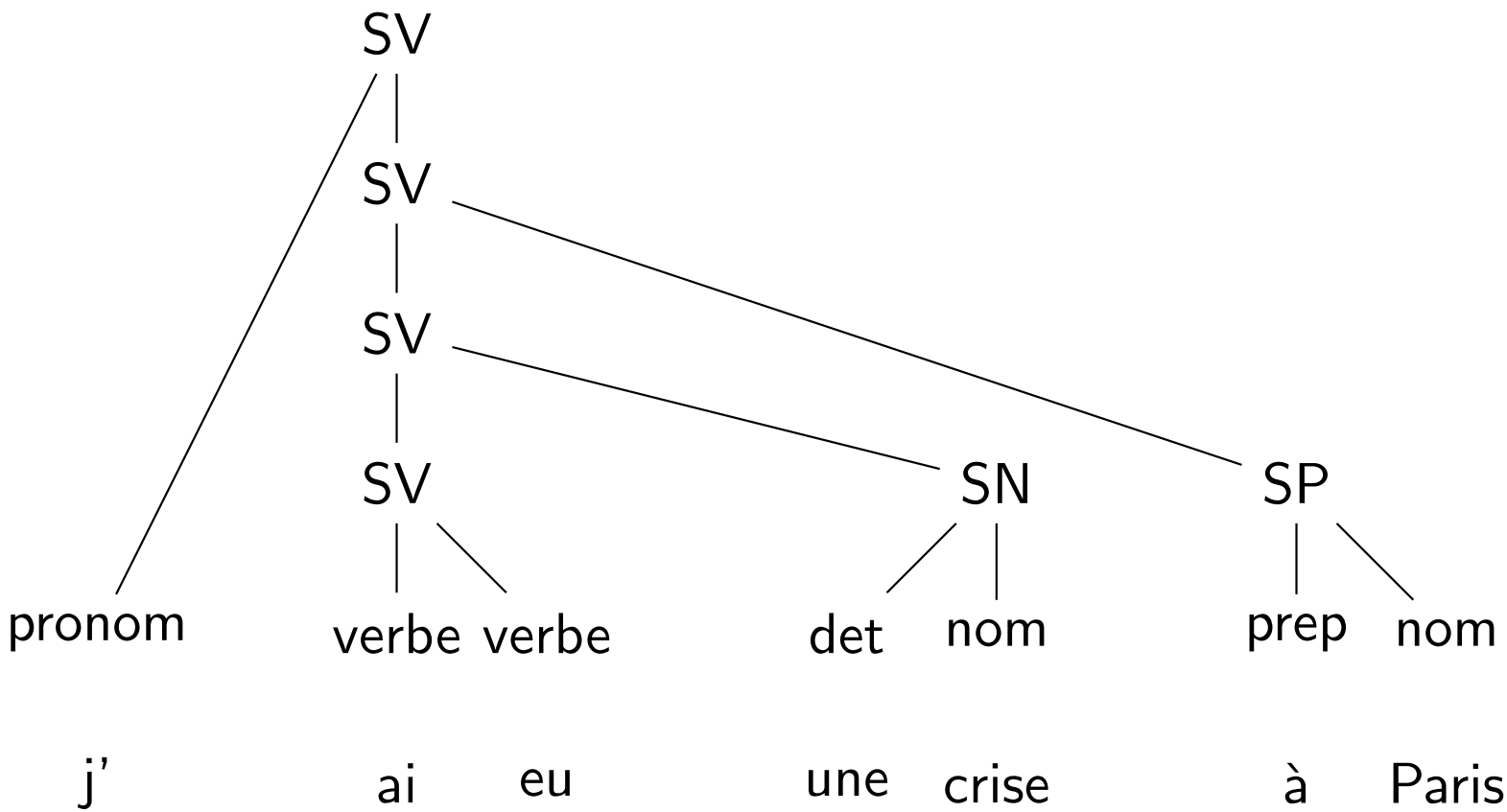
Approche sans tenir compte de la syntaxe, "sac de mots"

Je prends du valium pour dormir.
Je ne prends pas de valium pour dormir car j'ai peur de la dépendance.
J'ai pris une aspirine avant de dormir.
Je prends du café parce que sinon j'arrête pas de dormir pendant les cours.

Dépendances syntaxiques



Dépendances syntaxiques



Dépendances syntaxiques

outil : bonsai parser
https://alpage.inria.fr/statgram/frdep/fr_stat_dep_bky.html (ne marche pas très bien)

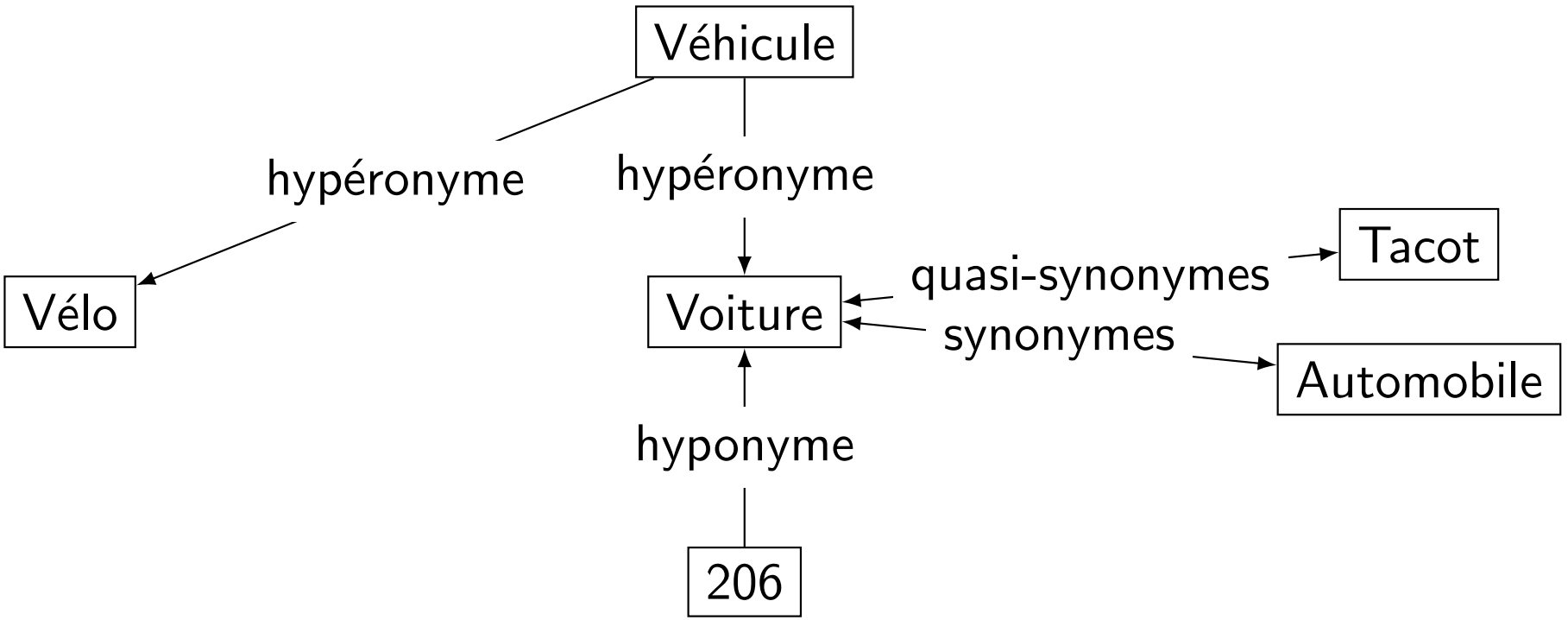
Sémantique

- définit le sens des mots et de leurs relations
- unité : trait sémantique
- permet de créer des ontologies

Traits sémantiques

	peut rouler	peut voler	est motorisé	plusieurs personnes
voiture	+	-	+	+
vélo	+	-	-	-
avion	+	+	+	+
rollers	+	-	-	-

Relations Sémantiques



Relations Sémantiques

- polysème : 1 mot qui a plusieurs sens
  - fil = fil de la discussion — fil à coudre — fil de forum
  - espace = espace typographique — cosmos — place vide

Relations Sémantiques

- polysème : 1 mot qui a plusieurs sens
  - fil = fil de la discussion — fil à coudre — fil de forum
  - espace = espace typographique — cosmos — place vide
- homonyme : 2 mots différents qui s'écrivent et se prononcent pareil
  - avocat = le fruit — la profession

Relations Sémantiques

- polysème : 1 mot qui a plusieurs sens
  - fil = fil de la discussion — fil à coudre — fil de forum
  - espace = espace typographique — cosmos — place vide
- homonyme : 2 mots différents qui s'écrivent et se prononcent pareil
  - avocat = le fruit — la profession
- homographe : même écriture, prononciation différente
  - couvent = le bâtiment — le verbe

Relations Sémantiques

- polysème : 1 mot qui a plusieurs sens
  - fil = fil de la discussion — fil à coudre — fil de forum
  - espace = espace typographique — cosmos — place vide
- homonyme : 2 mots différents qui s'écrivent et se prononcent pareil
  - avocat = le fruit — la profession
- homographe : même écriture, prononciation différente
  - couvent = le bâtiment — le verbe
- homophone : même prononciation, écriture différente
  - balai — balet
  - auteur — hauteur

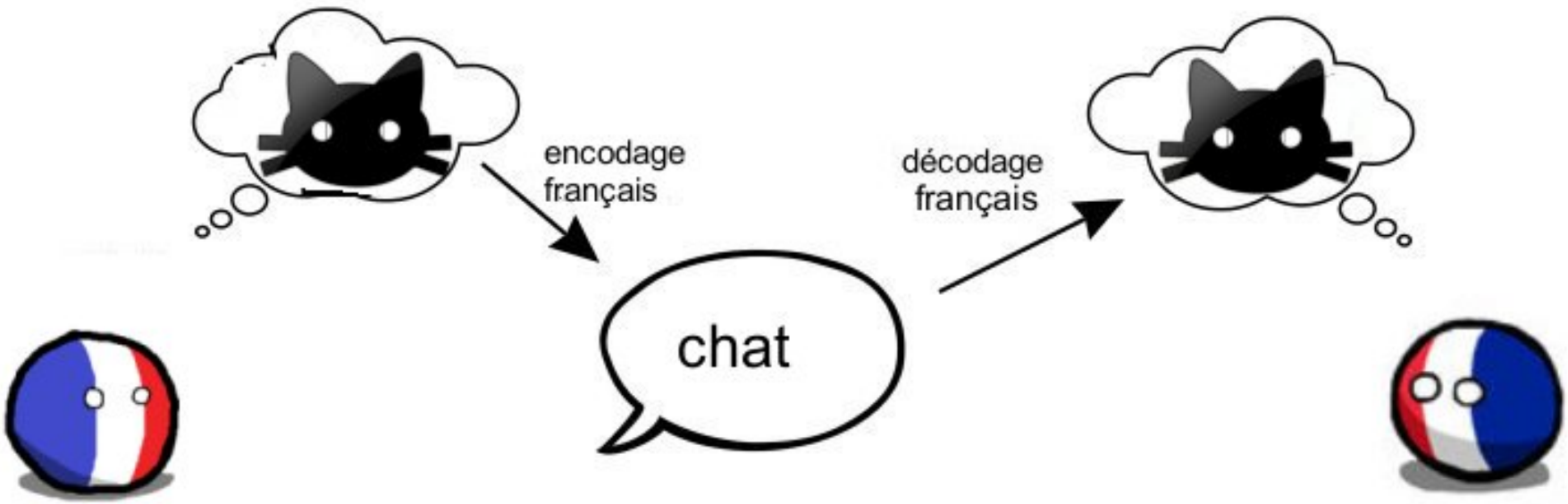
Pour aller plus loin

http://www.lattice.cnrs.fr/sites/itellier/enseignement.html
"introduction au TAL et à l'ingénierie linguistique"

Encodage

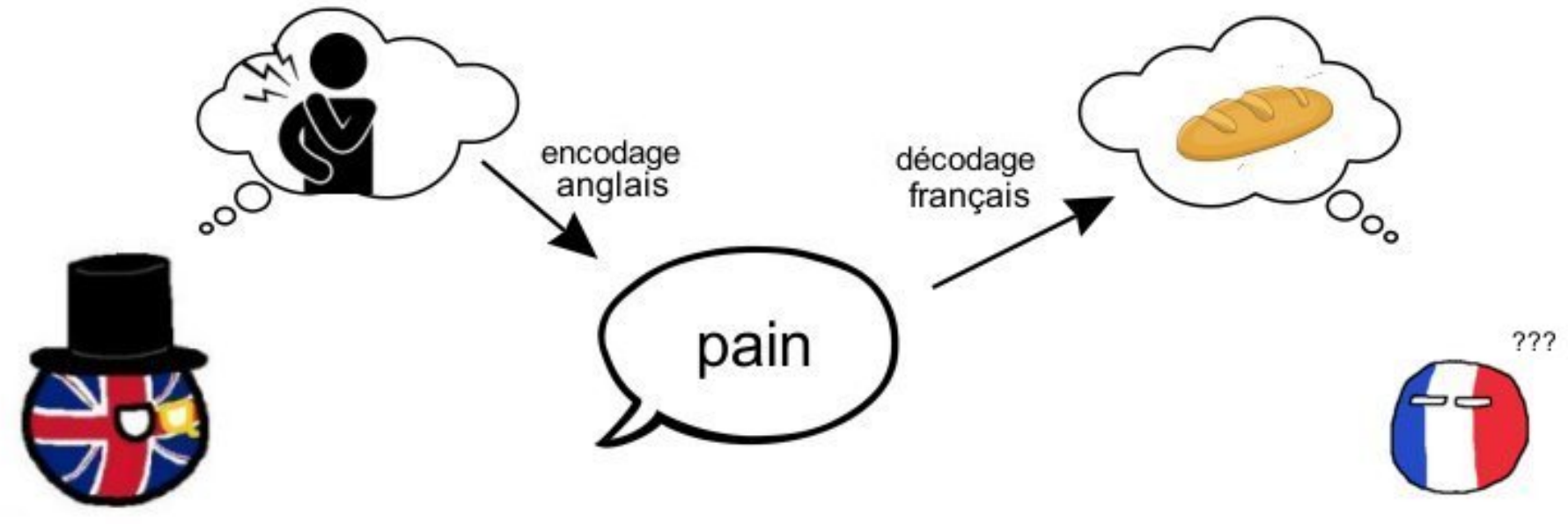
Une chaine mal dÃ©codÃ©e

Encodage



information → encodage → message → décodage → information

Problème d'encodage



information → encodage → message → décodage → information

Problème d'encodage

Si l'expéditeur et le destinataire n'utilisent pas le même encodage, le message devient erroné, ou ne veut plus rien dire

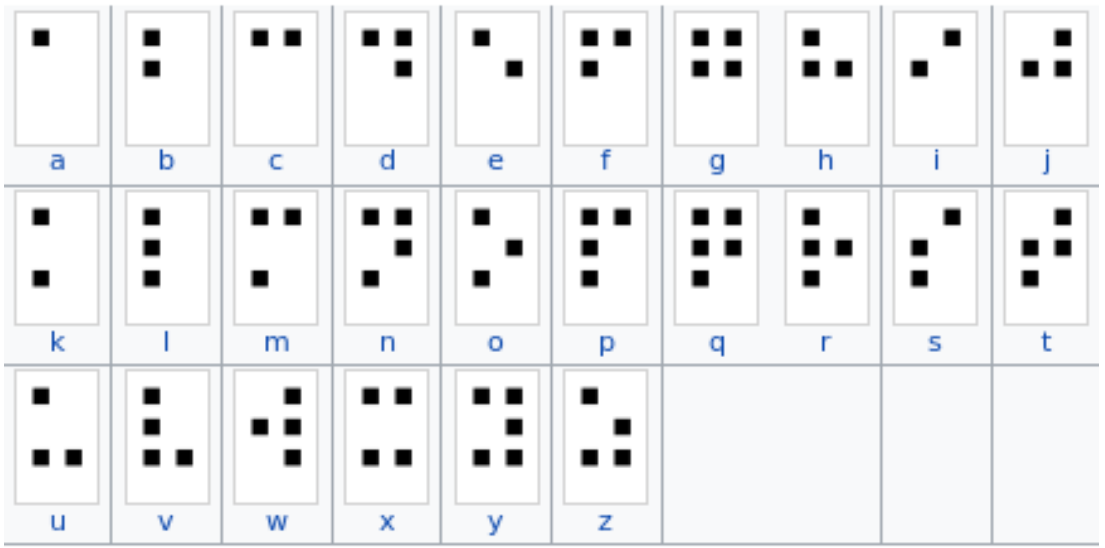
Encodage

Un encodage ça peut être...

- une langue parlée
- une langue écrite
- le morse
- l'écriture braille
- ...

Encodage

Un encodage a besoin d'au minimum 2 symboles
Un encodage peut être organisé en blocs de longueur fixe
Le braille utilise 2 symboles (point et vide) organisés en blocs de 6



Encodages informatiques

Les encodages informatiques sont binaires, ils utilisent 2 symboles (0 et 1)
Ils sont généralement organisés en blocs de 8 (un octet)
Par exemple 01000101 est un octet

Encodages informatiques

Un octet en binaire n'est pas très facile à lire...
Au lieu de les écrire en binaire, on les écrit le plus souvent en hexadécimal : on utilise 16 symboles
0 1 2 3 4 5 6 7 8 9 A B C D E F

Encodages informatiques

décimal	binaire	hexadécimal
0	0	0
1	1	1
2	10	2
3	11	3
4	100	4
...		
9	1001	9
10	1010	A
11	1011	B
...		
15	1111	F
16	10000	10
17	10001	11
...		

Encodages compatibles ASCII

L'encodage le plus simple est l'ASCII qui encode du texte
A = 41 ou 01000001
B = 42 ou 01000010
C = 43 ou 01000011
J = 4A ou 1001010
K = 4B ou 1001011
etc

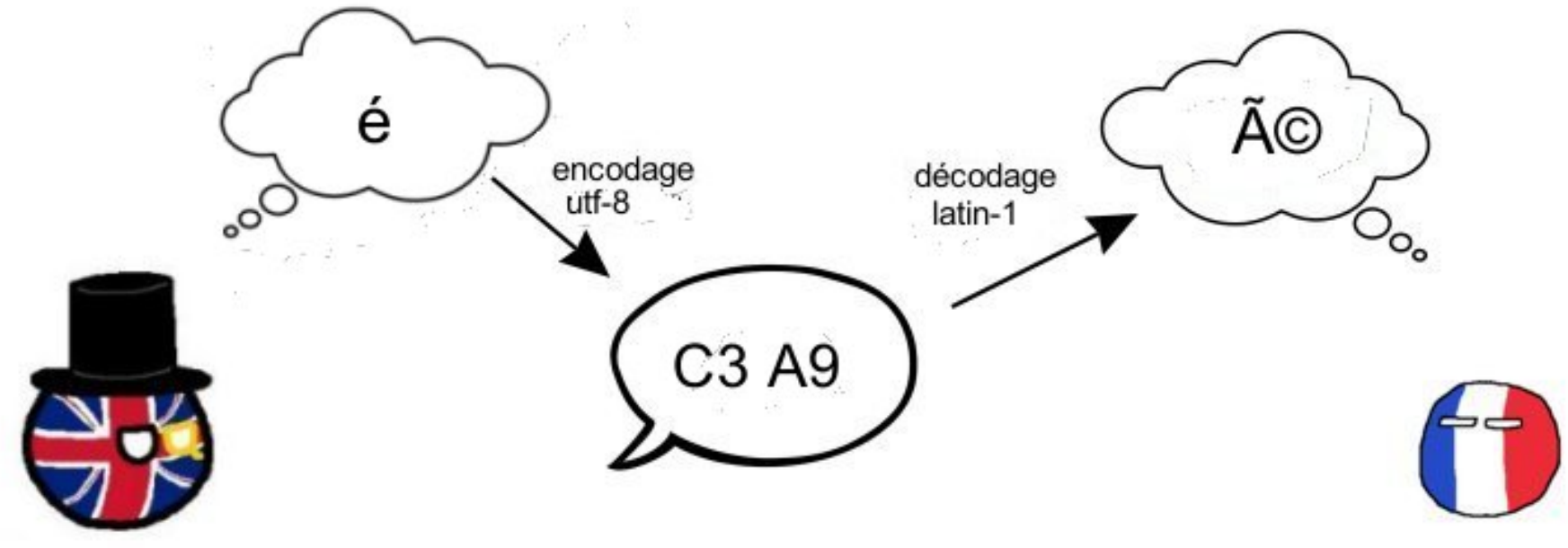
Encodages compatibles ASCII

L'ascii original ne comprend pas tous les caractères dont on peut avoir besoin pour écrire du texte !
L'ascii original était écrit sur seulement 7 bits, au lieu de 8 pour un octet. Il va de 00 à 7F. Si on encode sur un octet, il nous reste 80 à FF qui sont inutilisés...

Encodages compatibles ASCII

Il existe de nombreux encodages qui encodent 1 caractère sur 1 octet, utilisent l'ASCII de 00 à 7F et font à leur sauce de 80 à FF

Problème d'encodage



information → encodage → message → décodage → information

Problème d'encodage

"j'ai écrit un mail" → encodage → envoi du message → décodage → "j'ai Ã©crit un email"

Toutes les lettres sont correctement décodées, sauf le "é", pourquoi ?

Problème d'encodage

L'expéditeur et le destinataire utilisent 2 encodages différents, mais qui sont tous les deux dérivés de l'ascii

Tous les caractères sur la plage 00-7F sont correctement décodés, mais le caractère "é" se trouve sur la plage 80-FF et les deux encodages ne le placent pas au même endroit

Problème d'encodage

Comment résoudre le problème ?

Problème d'encodage

Comment résoudre le problème ?

Il faut changer l'encodage qu'on utilise pour décoder, pour le même encodage que l'expéditeur a utilisé

Problème d'encodage

Le plus simple est d'essayer les encodages les plus courants

- utf-8 (le standard, veillez à l'utiliser autant que possible)
- latin 1 (iso-8859-1) (très courant)
- latin 15 (iso-8859-15) (peu courant)
- latin autre (rare, peut être croisé si vous travaillez avec des fichiers édité par quelqu'un ailleurs en europe. un polonais par exemple.)
- windows 1252 (cp1252 ou ansi) (utilisé par certains programmes windows)
- mac roman (utilisé par les mac)

Astuces pour deviner l'encodage

- 1 caractère s'est changé en Ã + quelque chose : encodé en utf-8, décodé en latin-quelque chose
- 1 caractère s'est changé en √ + quelque chose : encodé en utf-8, décodé en mac-roman
- votre programme vous dit que le fichier est invalide, corrompu, ou contient des caractères invalides : encodé en quelque chose autre que utf-8, décodé en utf-8
- 1 caractère s'est changé en 1 autre caractère : encodé et décodé en quelque chose autre que utf-8

Astuces pour deviner l'encodage

Les éditeurs de texte Kate (linux) et Notepad++ (windows) peuvent changer facilement l'encodage d'un fichier pour faire des tests

Bless Hex Editor permet de visualiser les octets composant un fichier

http://www.fileformat.info pour des informations sur les caractères unicode, ou trouver le caractère à partir des octets

Unicode

Pas un encodage, un répertoire de caractères

Identifie un caractère

Contient des informations sur les caractères

Extraction du texte

- HTML : toujours utiliser un outil fait pour (jsoup en java)
- CSV : peut être parsé "à la main", mais n'oubliez pas qu'il peut y avoir une virgule dans un champ texte
- doc, docx, odt... $\rightarrow$ catdoc
- PDF $\rightarrow$ pdftotext
- images $\rightarrow$ OCR

Encodage

Assurez vous que vous connaissez l'encodage du texte. Si vous avez plusieurs sources de données, assurez vous qu'elles utilisent le même encodage, convertissez avant toute chose si besoin

Diacritiques

Si vous voulez supprimer les diacritiques (accents, ç, etc) voir bout de code dans diacritics.txt

Tokenisation

Découpage du texte en mots

Sur quels caractères découper ?
Possible de faire l'inventaire des caractères séparateurs ?

Tokenisation

caractères blancs (espace, tabulation, espace insécable, tabulation verticale...)

caractères qui ne sont pas des lettres (les caractères accentués ne sont pas dans la liste des lettres de A à Z, mais sont des lettres)

signes de ponctuation sur le clavier (il y aussi les signes utilisés par d'autres langues, les puces, les smileys...)

Tokenisation

Cas problématiques

- aujourd'hui
- peut-être, est-il, porte-monnaie, porte monnaie
- au fur et à mesure
- mangerai, vais manger

Tokenisation

Solutions

- solution idéale : utiliser un outil fait exprès, qui contient un dictionnaire (treetagger...)
- solution réaliste : tokeniser sur les caractères qu'unicode considère comme des séparateurs (\b ou \W avec le flag Pattern.UNICODE_CHARACTER_CLASS en java)
- solution si vous n'avez pas accès aux infos unicode : demandez à google comment accéder aux infos unicode

Lemmatisation

supprimer accord, conjugaison, mettre les mots dans leur forme dictionnairique

forme fléchie	forme lemmatisée
déprime	déprimer
déprimé	déprimer
dépressive	dépressif
antidépresseurs	antidépresseur

Lemmatisation

cas ambigus : couvent, président, aura

Stemmatisation

neutraliser tous les affixes

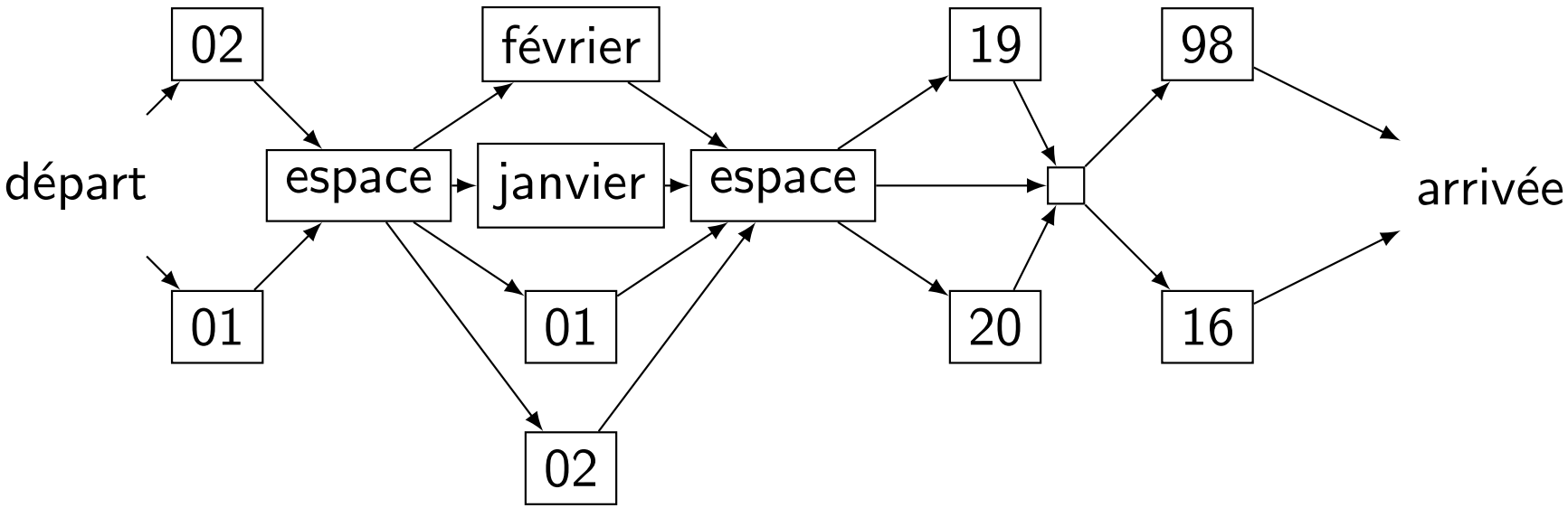
forme fléchie	forme lemmatisée	forme stemmée
déprime	déprimer	déprim-
déprimé	déprimer	déprim-
dépressive	dépressif	dépress-
antidépresseurs	antidépresseur	dépress-

Stemmatisation

outil : snowball stemmer

marche pas très bien (déprim != dépress)

Expressions régulières



Expressions régulières

\d\d (janvier|février|\d\d) (19|20)\d \d

Expressions régulières

https://regex101.com/

Expressions régulières

Pour aller plus loin : https://regexcrossword.com/

Pour aller plus loin

www.lattice.cnrs.fr/sites/itellier/enseignement.html
"introduction au TALN"