1) Une appli typique
2) Positionnement scientifique du TM
3) Tâches courantes
4) Technos utilisées
5) Gestion des formats du texte en entrée
6) Prétraitements ou « la chaîne TAL »
7) Dataset distributionnel et clustering
8) Ré-annotation

Une appli de text mining typique

1) collecte des données et gestion des formats d'entrée

2) « prétraitements linguistiques »

- bien se poser la question de leur raison d'être

3) utilisation de quelques métadonnées

- linguistiques (par ex. wordnet, listes prospero) : élaboration des données (mesures plus construites)
- externes (par ex. notices biblio) : facettes des jeux de données, croisements possibles sur les mesures

4) quelques opérations ou calculs

- indexation (RI)
- agrégation, cooccurrences, PCA...
- + prédictions ML et modèles génératifs…
- et quelques visualisations
  - nuage de tags, graphe, dimensions, groupes, évolution...

5) quelques usages

- recherche documentaire
- création de lexiques ou taxonomies
- annotation automatique
- résumé automatique
- *etc.*

  ⇒ exports de mesures ou de représentations intermédiaires

Historique text mining

- **Classement thématique et bibliométrie**
- **Conférences TREC**
  - repérer des faits dans des actus
  - entités nommées et pattern matching
  - pondération
  - résumé automatique et filtrage automatique
- **Recherche documentaire**
  - pondération relation mot <=> document
  - profils de mots clefs (=> mots voisins)
- **Ontologies expertes**
  - ex : symptômes <=> maladies <=> médicaments
  - reconnaissance du signe comme terme lié à un concept ou référent « objectif »
  - propriétés et interactions du référent stipulés par des experts
- **Analyse d'opinion**
  - retours utilisateurs : sujet évoqué, niveau d'insatisfaction
- **etc.**
  - (toutes formes de recoupement des communications)

De quelle science le TM est-il l'ingénierie ?

- **Sciences du :**
  - sens dans le texte
  - pour un usage info (données, web)
  - souvent dans de grandes bases
- **§lg : TAL, linguistique, pragmatique**
- **§info : « computer science » et ergonomie**
- **§stats : statistique et « data science »**

Structuration de données texte dans des bases

- **on parle du texte comme « non-structuré » des bases de données**
- **ex : surlignage thématique de contenus = structuration**
- **« non-structuré » ? ça dépend pour qui...**
  - lettres, linguistique et en général SHS : figures, niveaux d'articulation, form. disc., référence, etc.
  - imprimerie : styles, disposition, formes de lectures
  - bibliothécaires : métadonnées bibliographiques, index thématiques
- **structurer**
  - baliser, collecter, recenser, classer, utiliser comme clé de classement,
  - aller vers des données catégorielles ou numériques
  - comment bien utiliser la séquentialité ???
    - et ses différents niveaux d'articulation
  - expliciter des propriétés implicites des données

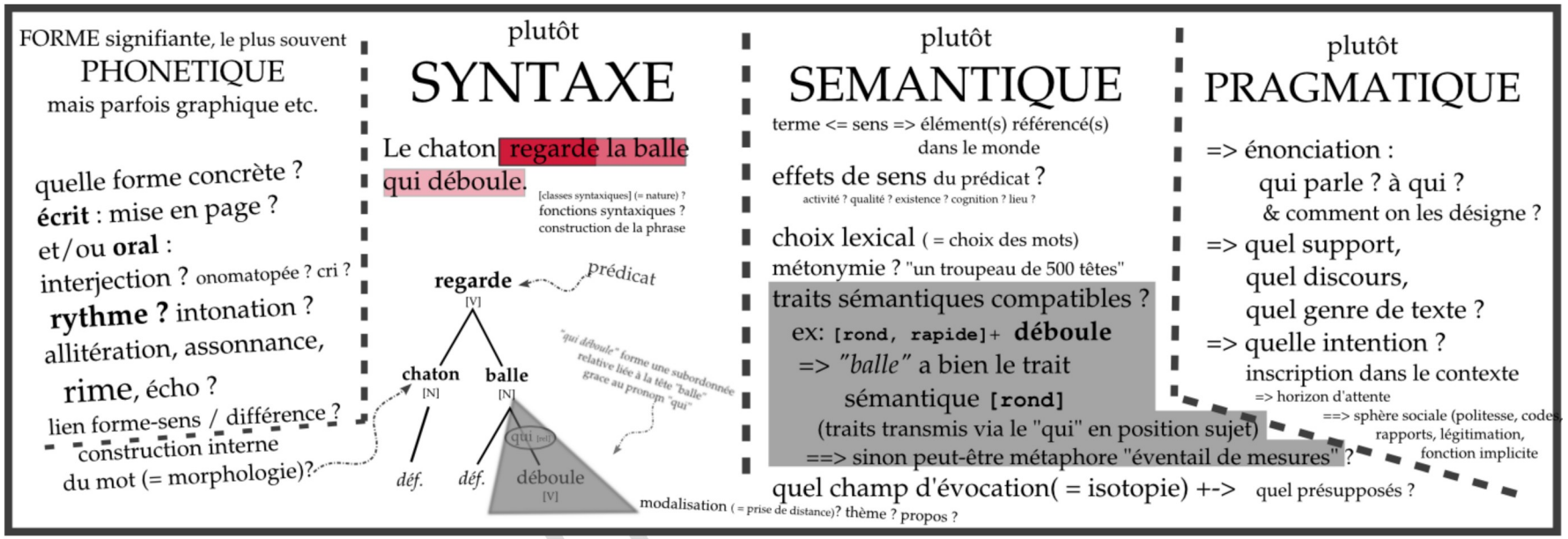
    => en tirer des représentations stats

    => ou plutôt des briques intermédiaires pour des réprésentations stats

    => et usages ML : propager des valeurs, etc

Niveaux et unités d'analyse postulés

- **en linguistique**

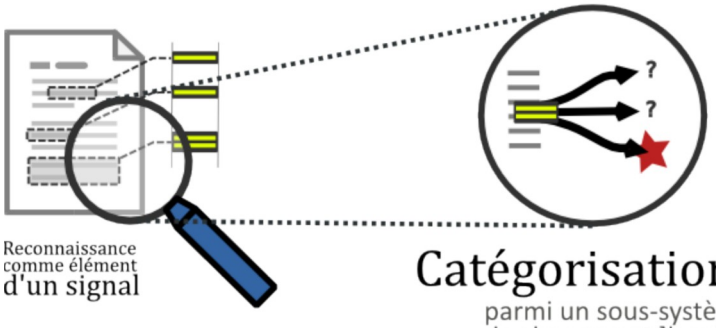


- double articulation, et en fait multiple

1) corpus > texte ou énoncé
   - séquences et/ou propositions > syntagme ou phrase (noyau préd. + constituants affixes)
     - mots > morphèmes composés
2) forme graphique ou sonore des morphèmes

- **correspondances en info**
- **correspondances en stats**

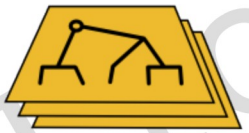
La nécessité de mécaniques locales



Catégorisation
parmi un sous-système
de classement]local[

Sous-système
de classement à
nombre fini
de classes

"La communication suppose donc que le sens de l'acte sémique soit soumis, par l'émetteur autant que par le récepteur, à un double classement ; par conséquent, si l'on admet que, "concevoir" un objet, c'est le reconnaître comme membre de l'extension d'une classe, la communication suppose que le sens de l'acte sémique soit conçu deux fois, une fois comme membre de l'extension d'une des classes composant le système d'intercompréhension et une autre fois en tant que membre du signifié d'un signal."



Ontologie
dans le cadre
d'un système culturel

aka système de
classemement
à nombre infini
de classes

Prieto, Luis J. (1975)
Pertinence et pratique

Le système d'intercompréhension sur lequel se fonde un acte de parole est toujours un système de classement comportant un nombre pratiquement infini de classes et, par conséquent, un système de clas--sement avec lequel on ne peut opérer que parce qu'il « se résout » en systèmes de classement partiels.

(p. 60, cf. aussi pp. 39 et 44-45)

---

traitement automatique du langage naturel

- **histo**
  -
  -
  -
  -
  -
  -
  -
  -
  -
  -
  -

Même si on a une perception directe du sens de l'ensemble et des types d'information qu'on cherche ce n'est pas toujours suffisant

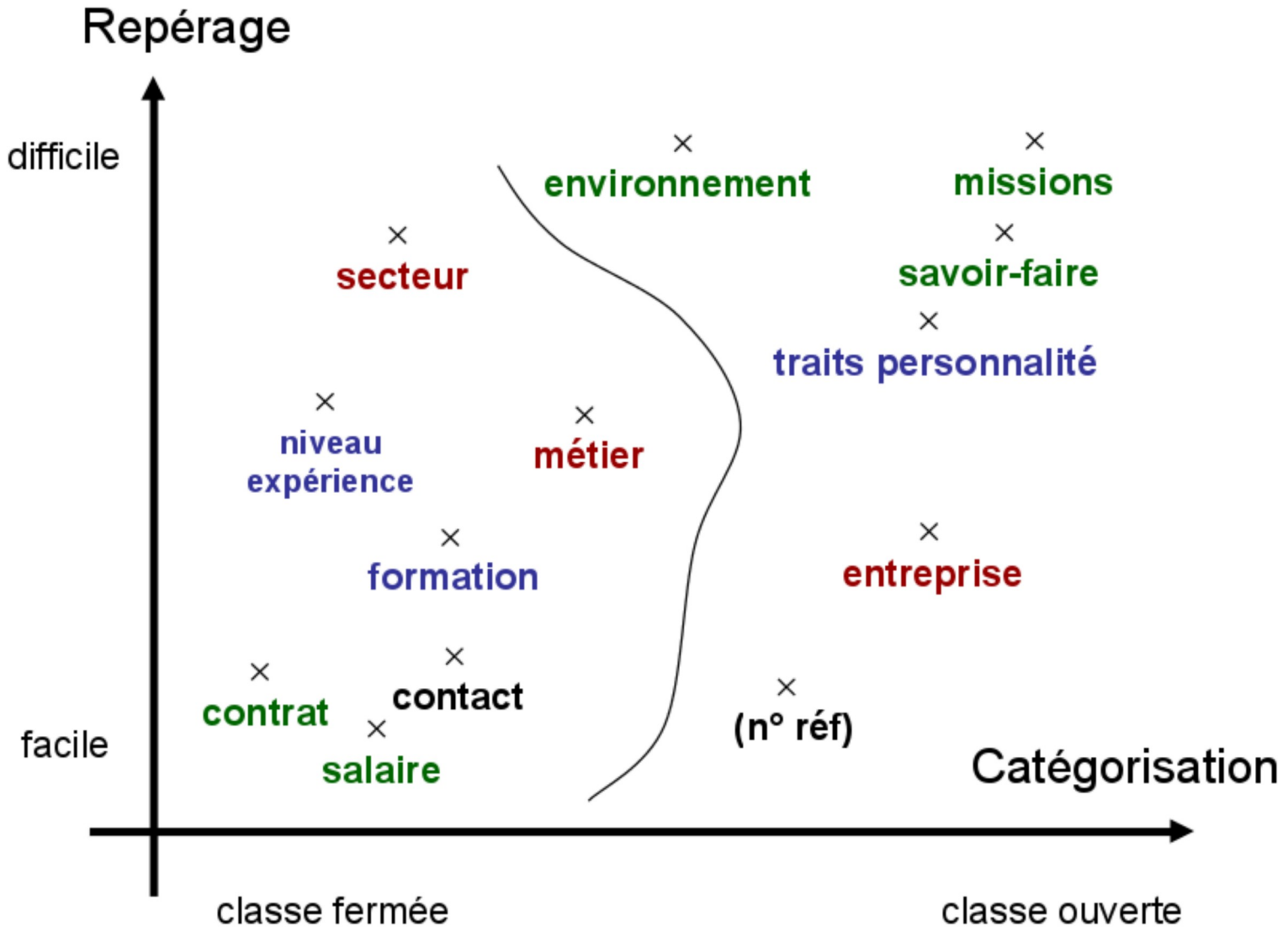
  - parfois c'est clair et facile à automatiser
  - ex solutions rapides via matching, grace à marqueur ou forme connue, comprise
  - mais pas toujours généralisable, donc un peu bricolage

- **les mécanismes du langage ou de la construction du texte**
  - sont multi facteurs et pas souvent universaux
- **on sera amenés à essayer d'en expliciter les mécanismes**
  - Difficultés nombreuses : mélanger les concepts §lgq avec les concepts §info!!!
  - Un programme est une forme de calcul opératoire et donc une semi-formalisation théorique
    - or les théories actuelles du TAL ne fournissent pas non plusune théorie d'ensemble
- **on se concentrera aujourd'hui sur les données lexicales**
  - ex : extraction de terminologies à partir de corpus
  - quelques exemples sur vocabulaires de métiers ou terminologie scientifques

proposition : les types d'information

- **un objet d'analyse hybride**
  - entre modèles linguistiques
  - et modèles informatiques
  - une notion que je propose pour faire le pont
- **on peut y raccrocher les principaux caractères linguistiques**
  - dont on dispose avant ou qu'on a observés
  - qui nous intéressent
- **ce sont des catégories qui marchent pour du design applicatif**
  - qui intègre souvent plusieurs technologies disparates
  - qui a besoin de sérialiser les données entre modules

proposition : les types d'information



Domaine	Principaux types	Fréq. moy. (par offre)	Long. moy. (en mots)
Poste	intitulé du poste, mention du domaine professionnel	2,36	3,98
Mobilité	lieu de travail, déplacements prévus, zone commerciale	0,85	2,67
Etablissement	nom, groupe, site web, données quantitatives (CA, effectifs)	2,43	3,15
Recruteur	nom, spécialisation, site web	0,23	3,27
Secteur	activité, branche, produit ou service vendu	2,06	5,02
Environnement	responsable, interlocuteurs, équipe/service, conditions de travail	2,09	4,63
Missions	fonction principale, tâches, objectifs liés au poste	**7,99**	8,06
Contrat	type de contrat, durée, horaires, salaire	1,10	3,41
Expérience	[ancienneté + nature de l'expérience] (sous forme phrastique composée)	1,29	9,45
Savoir-faire	connaissance d'un champ, compétences techniques, langues	2,04	3,74
Personnalité	*(pas d'articulation secondaire émergente)*	3,25	2,63
Formation	diplôme, niveau, filière de qualification	0,86	5,26
Contact	e-mail, personne à contacter, adresse, n° de réf., procédure à suivre	0,79	7,20

Table: Typologie retenue : domaines, types, annotations/domaine

Tâches courantes

- **suivi de news**
- **catégorisation de documents**
- **extraction de topics**
  - représentés par des listes de mots
  - ou par des catégories de docs sur « projetés » par ces listes
- **structuration de dates, horaires**
- **analyse d'opinion**

Une multitude de sous-tâches (briques applicatives)

- **segmentation de document**
  - parties, headers, fin, tables, etc
- **étiquetage morpho-syntaxique**
-

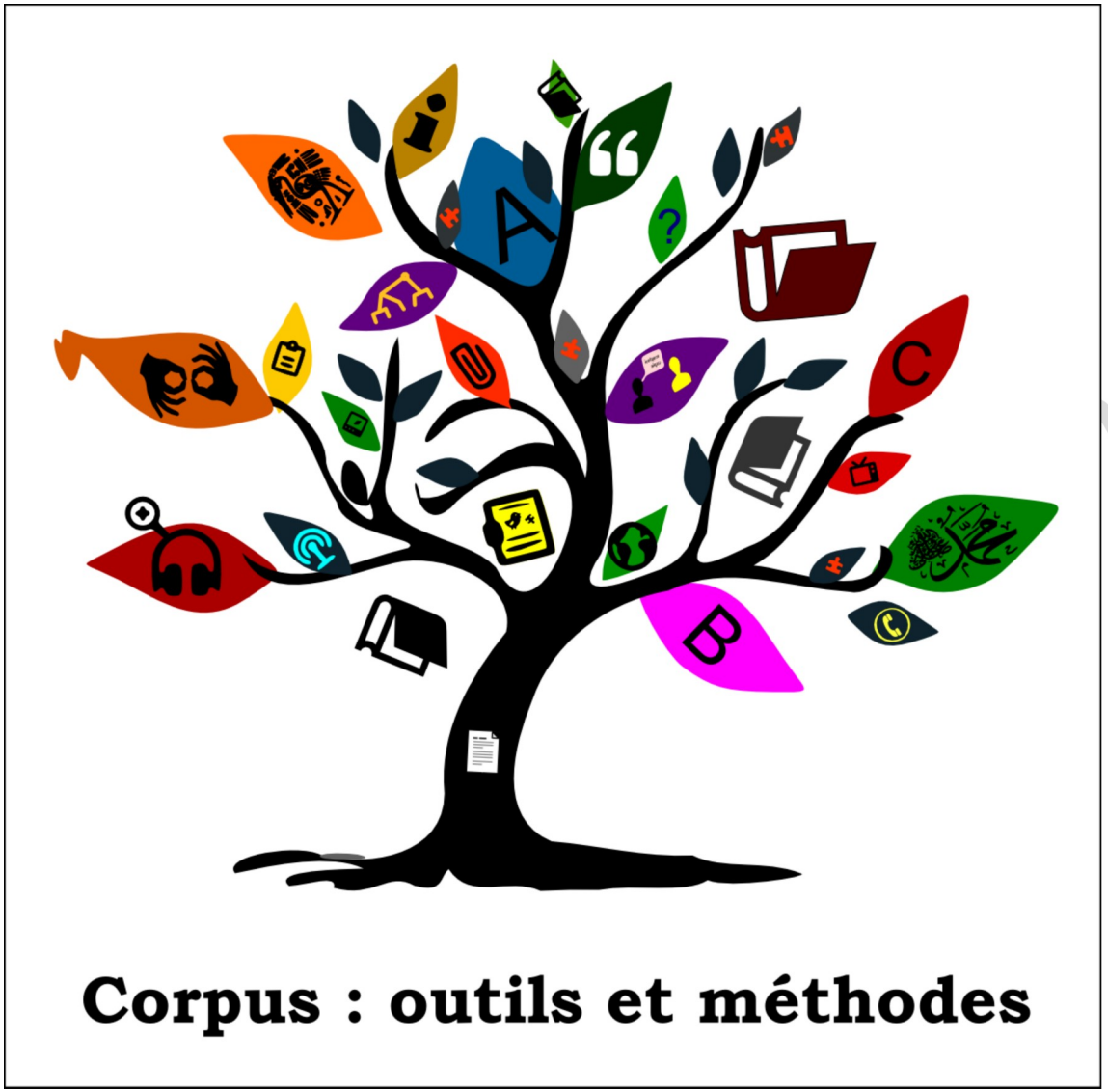
Technologies utilisées

- **baliseurs => à différents niveaux de la séq.**
  - expressions régulières
    - et en général parsing CFG
    - unitex, xrce, lilian
  - HMMs et CRFs
    - wapiti
- **classifieurs**
  - arbres de décisions
  - regression logistique
  - svms
- **réseaux neuronaux (combinent les 2)**

Types de logiciels existants

- **Annotation (ex : Glozz)**
- **Edition d'ontologie (ex : Protégé OWL)**
- **Analyse exploratoire**
  - stats (ex : Iramuteq, Reliure)
  - viz (ex :

Gestion des formats de texte



- nombreuses formes
- flux vs stable
- séq. vs set
- encodage !?
- media/support
- source, périmètre th.
- genre, style, registre
- pré-annotations
- contenu txt
- etc.

Gestion des formats de texte

- **En entrée on aurait du texte brut**
- **Pour s'entraîner il faut le refabriquer**
- **Selon les sources on devra :**
  - Automatiser les téléchargements
  - Parser du XML ou du JSON
  - Récupérer les champs qui nous intéressent
- **Exemple sur un corpus istex**
  - Repérage http://demo.istex.fr
  - suite dans ipython

Gestion des formats de texte

- **Téléchargement => quel chemin API correspondant**
  - https://api.istex.fr/document/?q=ICIREQUETE
  - => hits (avec un LONGID)
  - Contenus pour chaque document
    - https://api.istex.fr/document/LONGID/metadata/mods
- **Automatisation**
  - Lieu de stockage ?
    - base de donnée ou système de fichiers
  - Forme de stockage
    - métadonnées
    - contenus textuels

Gestion des formats de texte

- **« set, dict, list » : si on a du json c'est quasiment immédiat**
  - En python

```
from json import loads

json_dict = loads(json_string)
```

- **« séq. » : parser le XML et trouver les champs**
  - Avoir les bonnes librairies (libXML2 et lxml)
  - Prendre 2 sec pour déclarer les namespaces

    => xpath (`/ns:chemin/ns:vers/ns:mon/ns:contenu`)
- **Si vous travaillez du html au lieu du xml**
  - C'est très similaire

```
from lxml import html

dom_tree = html.fromstring(r.text)

dom_tree.xpath('//table[@class="odTable odTableAuto"]')[0]
```

Gestion des formats de texte

Si on avait plus de temps

- **Questions d'encodage**
- **Types et « calibre » des documents**
  - les types de documents sont liés aux contenus du discours (via perspectives d'interprétation)
- **Organisation interne du document**
  - « boilerplate removal » <=> dans quels contextes ont trouve « l'information » ?
    - séquences (Charolles)
      - les zones textuelles typiques d'un genre
      - « grammaire textuelle », normée par les usages et modelée par les intentions sémantiques (ce qu'on veut évoquer)
    - structure rhétorique
      - Traditionnelle : *cf.* Barthes, L'ancienne rhétorique
      - moderne (SDRT etc.) : *cf.* Mann & Thompson, Danlos, Luc & Virbel
  - automatisation de la segmentation
    - par ex :
      - Pinto etal 2003 *Table extraction using conditional random fields*
      - Kessler 2009 (E-gen, pour les offres d'emploi) (via chaînes de Markov)
      - Lopez (Grobid, pour les articles scientifiques) (via CRFs)
      - Tkaczyk (Cermine, pour les articles scientifiques) (via CRFs)
- **Recyclage des métadonnées**
  - (info sur le doc) X (corrélation co entre doc et termes | co > seuil) = info sur les termes

    par exemple une catégorie thématique, un groupe d'auteurs, etc.

Prétraitements ou « la chaîne TAL »

- **simplification de l'observable**
  - Mot, « mot », MOT, mots
  - traiter, traite, traitement, traité ?
- **phrases > unités lexicales > nature gram.**

A) segmentation en « phrases »

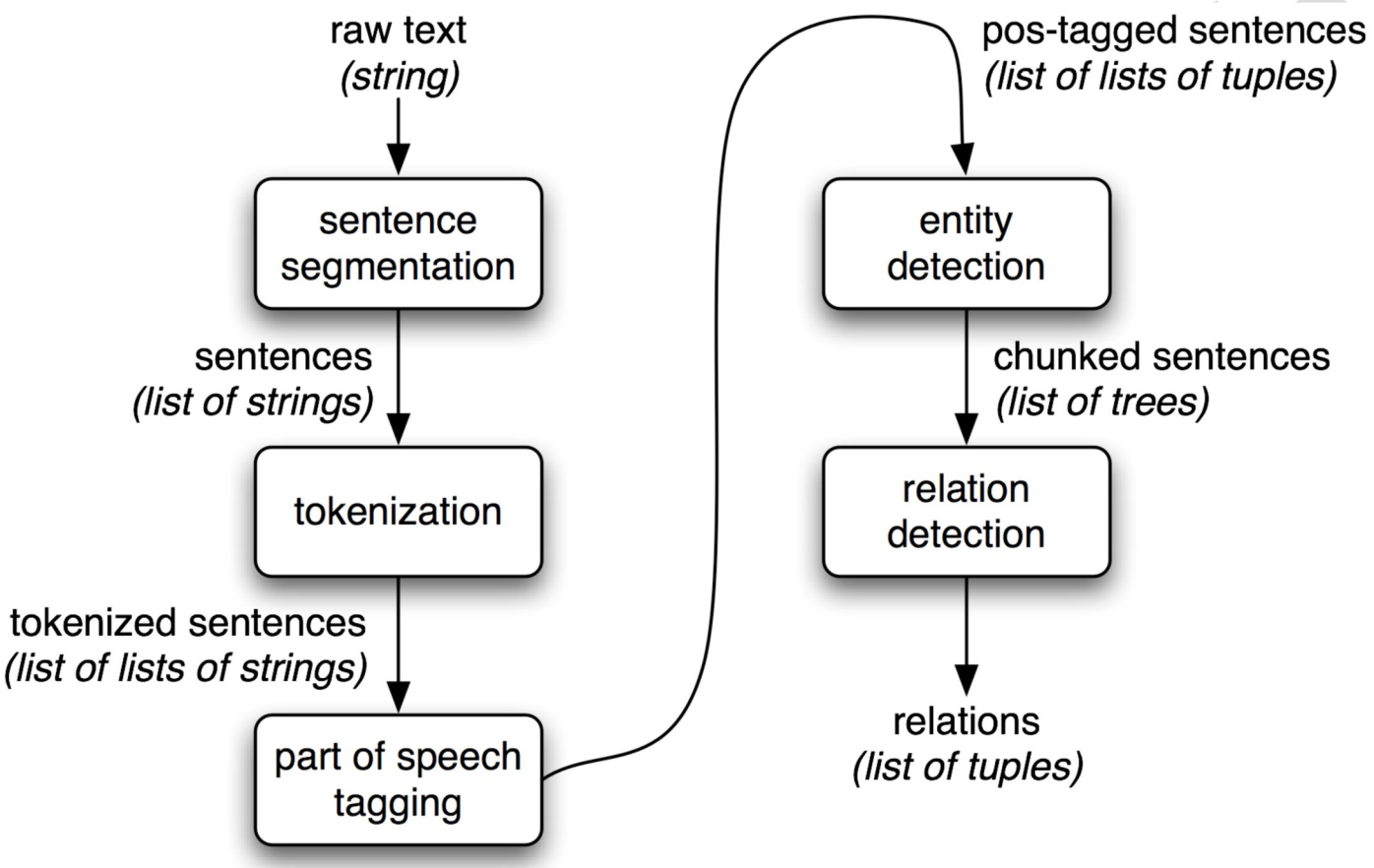
B) étiquetage (nom, adj., v., etc.)

alias « POS tagging »

C) tokénisation (lemme ou chaîne str de réf.)

Prétraitements ou « la chaîne TAL »

- cf. http://www.nltk.org/book/ch07.html



Prétraitements ou « la chaîne TAL »

- **segmentation**
  - qu'est-ce qu'une phrase
  - qu'est-ce qu'un énoncé minimal ?
  - quels délimiteurs : aucun n'est univoque
  - très complexe, pas toujours nécessaire
    - utile au POS tag
    - utile à la co-référence
    - utile à l'analyse d'opinion
      - prédication et modalités jouent dans la phrase

Prétraitements ou « la chaîne TAL »

- **segmentation en nltk**
  - http://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.punkt
  - Kiss & Strunk (2006)
  - modèle non-supervisé de détection de sigles etc. pour les filtrer des marqueurs de phrase

```
from nltk import sent_tokenize

sent_tokenize("Allez les Bleus! Aujourd'hui au stade de France c'est le
grand jour. M. Blanc, l'entraîneur de l'équipe, déclare 'ne pas savoir
quoi dire'.")


['Allez les Bleus!',
 "Aujourd'hui au stade de France c'est le grand jour.",
 "M. Blanc, l'entraîneur de l'équipe, déclare 'ne pas savoir quoi dire'."]
```

Prétraitements ou « la chaîne TAL »

- **étiquetage morphosyntaxique**
  - aka « nature » grammaticale
  - aka part of speech (POS) tagging
- **utile pour la détection de mini-séquences**
  - groupes nominaux
  - structure sujet verbe objet
  - utilisables en cascade
    - GN <= D A ? N
    - GN <= GN rel V
    - GN <= GN prep GN
    - S = GN + GV
- **et pour le typage des unités/fragments extraits**

Prétraitements ou « la chaîne TAL »

- **étiquetage morphosyntaxique en nltk**
  - nltk.pos_tag()

```
for s in sents:
    toks = nltk.word_tokenize(s)
    poss = nltk.pos_tag(toks)
```

- **tagset de l'upenn :-/**

[('We', 'PRP'), ('all', 'DT'), ('live', 'VBP'),
('in', 'IN'), ('a', 'DT'), ('yellow', 'JJ'), ('submarine', 'NN'),
('.', '.')]

Prétraitements ou « la chaîne TAL »

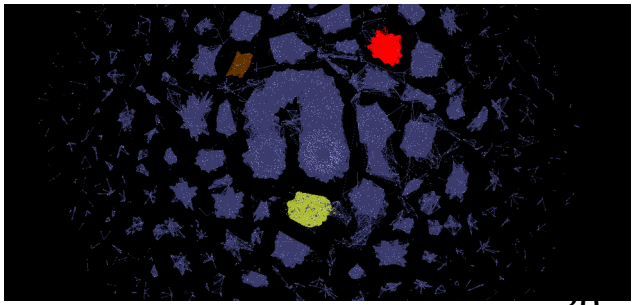
<JJ.*>*<NN.*>+(<P|IN> <DT> ? <JJ.*>*<NN.*>+)*

Du corpus au dataset

- **Lexicométrie, textométrie**
  - un pont entre ingénierie d'extraction d'infos et linguistique
- **toujours 2 axes**
  - liste : paradigmatique
    - (modélise les remplacements possibles)
  - séquence : syntagmatique
    - (modélise les enchaînements)
- **Occurrences = freq absolue**
  - Ex : Corpus offres d'emploi
  - HEREEXEMPLE
- **Figement**
  - Proba (seq « AB ») >> Proba(A) x Proba(B)
  - On peut poser AB comme unité lex
  - Souvent on compte juste les ngrammes de tokens
    - « mot »
- **Importance**
  - freq ou ln(freq)
  - taille
  - corrélation avec un facteur
  - Graphe de voisins dans un doc => centralités, etc.
  - Appli linéaire de freqs dans des fenêtres XYZ
- **Spécificité contrastive**
  - 1 autre corpus, « de référence »
  - Ecart entre freq relative de l'unité dans chaque
  - Spécificité : « poste »
- **Spécificité interne**
  - Ndocs ou fenêtres contenant le terme dans notre corpus
  - Spécificité : « atelier »
- **Corrélation avec situation, catégorie**
- **etc**
- **Aujourd'hui :**
  - matrice docs X groupes nominaux

Modèle distributionnel et clustering

- **Fenêtres de décomptes**
  - Docs
    - cf. *genre textuel*
  - Autres séquences délinéaires
    - groupes de docs § phrase chunk
    - chaque niveau a ses types et ses catégories
    - chaque niveau possible en 1 ou en n-grammes de catégories
  - Séquences réticulaires
    - /patterns pour (quelquechose)/
    - eg : tous les noms qui portent l'adj. « électrique »
    - Patrons écrits par des linguistes ou des experts d'un domaine
    - Chaînes de Markov, CRFs
- **On obtient des matrices mots X contextes**
- **Approches exploratoires : ACP, clustering**



Réflexion liminaire

le lexique comme champ empirique du sens ?

–

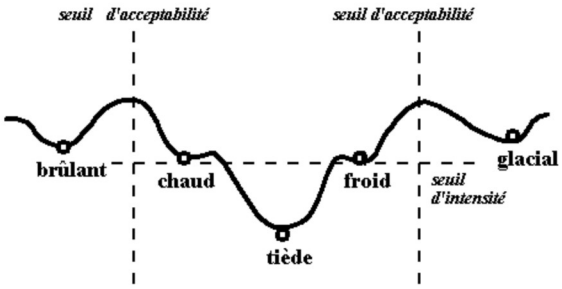
– Expressions idiomatiques, définitions « populaires »
  - usages: paraphrases attestées, compatibilités d'usage, etc. etc.

–

Réflexion liminaire
le lexique comme champ empirique du sens ?

- Un constat : variabilité infinie des emplois
  - Cependant 1 : énoncés récurrents
  - Cependant 2 : contextes caractéristiques
    - position dans l'énoncé
    - types d'énoncés contenants
      - **thèmes évoqués**
      - **situation d'interlocution**
  - Cependant 3 :
    - Sous-sytèmes ordonnés
      - mais eux-mêmes très variés, spécifiques à 1 besoin de communication
      - Ex : les dates, les molécules
      - Ex : les gradations et autres schèmes
      - Ex : les toponymes
      - Ex : les terminologies professionnelles
      - influence institutionnelle (eg l'université)

"la langue spécialisée est une langue naturelle considérée en tant que vecteur de connaissances spécialisées", (Lerat 1995 : 20)

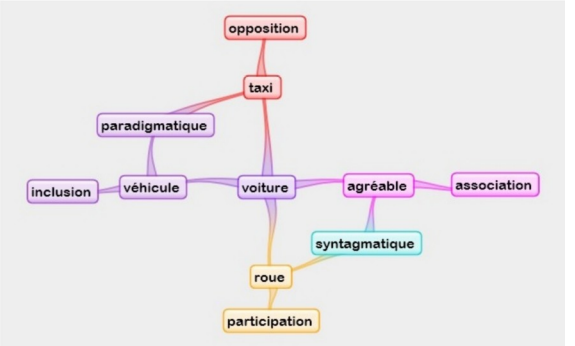


Réflexion liminaire

le lexique comme champ empirique du sens ?

- découpe interne et voisinage proche
  - morphème, unité lexicale, terme
- 2 familles de dimensions de l'observation :
  - présence dans des séquences (à différents niveaux)
  - paradigme (ensembles de « mots » partageant un trait)
- artefact :
  - interactions langagières <=> enregistrement du récurrent

système complexe (Palo Alto)



Réflexion liminaire

le lexique comme champ empirique du sens ?

- interprétation locale d'un mot ?
  - variable selon les contextes
    - « entretien »
  - un peu subjective
    - « gold »
  - mais partagée <=> situations, intentions
    - « allez »
  - mais grammaticale <=> structures/frames
  - et en constructions caractéristiques
    - patterns/motifs utilisant le mot et la struct. grammaticale
  - mais univers du discours <=> énoncés/textes
- en fait trop d'étapes pour reconstituer l'information via toutes les relations du mot
  - relations simplifiées ad hoc selon les mécanismes de texte mining
  - et unités aussi
    - ngrammes de mots
    - termes spécialisés
    - constructions (formules pré-organisées)
  - intermédiaires de ce qui est cherché
    - appelons-les entre nous « types d'informations »

Domaine	Principaux types	Fréq. moy. (par offre)	Long. moy. (en mots)
Poste	intitulé du poste, mention du domaine professionnel	2,36	3,98
Mobilité	lieu de travail, déplacements prévus, zone commerciale	0,85	2,67
Etablissement	nom, groupe, site web, données quantitatives (CA, effectifs)	2,43	3,15
Recruteur	nom, spécialisation, site web	0,23	3,27
Secteur	activité, branche, produit ou service vendu	2,06	5,02
Environnement	responsable, interlocuteurs, équipe/service, conditions de travail	2,09	4,63
Missions	fonction principale, tâches, objectifs liés au poste	**7,99**	8,06
Contrat	type de contrat, durée, horaires, salaire	1,10	3,41
Expérience	[ancienneté + nature de l'expérience] (sous forme phrastique composée)	1,29	9,45
Savoir-faire	connaissance d'un champ, compétences techniques, langues	2,04	3,74
Personnalité	*(pas d'articulation secondaire émergente)*	3,25	2,63
Formation	diplôme, niveau, filière de qualification	0,86	5,26
Contact	e-mail, personne à contacter, adresse, n° de réf., procédure à suivre	0,79	7,20

Table: Typologie retenue : domaines, types, annotations/domaine