Introduction

Damien Nouvel

Plan

1. Généralités

2. Contexte général

3. Corpus et statistiques

4. Visées applicatives

Séances et modalités de contrôle

- **Séances**
  - 12 séances
  - Chaque séance : **cours** et **exercices** sur machine
  - Cours S2 : *apprentissage automatique*
- **Modalités de contrôle**
  - Un **examen** final (50%)
  - Un **contrôle** (50%)
  - Une note de **contrôle continu** (+/-1 point)

Contenu du cours (prévisionnel)

- **Progression**
  - Introduction
  - Programmer en Python
  - Éléments en théorie des probabilités et en statistiques
  - Statistiques pour la linguistique
  - Théorie de l'information et mesures d'entropie
  - Statistiques pour l'évaluation
  - Classification automatique
  - Paramétrage de classifieurs (et projet)

Plan

1. Généralités

2. Contexte général

3. Corpus et statistiques

4. Visées applicatives

Informatique et langage

- **Grace Hopper** : **A0**, premier langage compilé (1951)

Traitement Automatique des Langues (TAL)

- Le langage
  - **Expressivité** (implicite, déductions)
  - **Générativité** / compositionnalité
  - **Modes d'expression** : oral, écrit, signes, etc.
  - **Evolution** par conventions sociales
- L'informatique
  - **Calcul** binaire, entiers, flottants, etc.
  - **Procédures** par séquence d'instructions
  - **Copies exactes**
  - **Télécommunications** et essor des nouvelles technologies

$\Rightarrow$ **Discipline** issue de l'essor de l'utilisation de l'informatique pour que des **humains** manipulent le **langage** par d'**autres moyens** et à une **autre échelle**

Plan

1. Généralités

2. Contexte général

3. Corpus et statistiques

4. Visées applicatives

Un corpus : quelles données, pour quoi faire

⇒ **Ensemble de documents (textuels)** ...

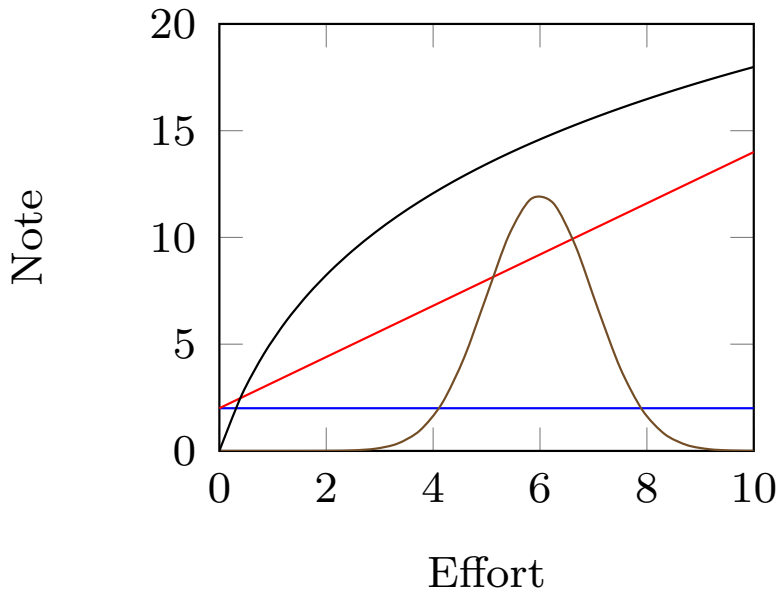
- Des *classifications* associées, par
  - auteur(s) et éditeur
  - langue
  - date de publication
  - style (roman, journal, publication, administratif, etc.)
  - thèmes (science-fiction, politique, sport, technique, etc.)

⇒ Création d'**indexs** pour trouver les documents

- **Ressources** pour
  - linguistique
  - publications scientifiques (par ex. biomédical)
  - sciences de l'homme (sociologie)
  - traitement du langage parlé
  - etc.

⇒ Données disponibles pour **traitements** et études

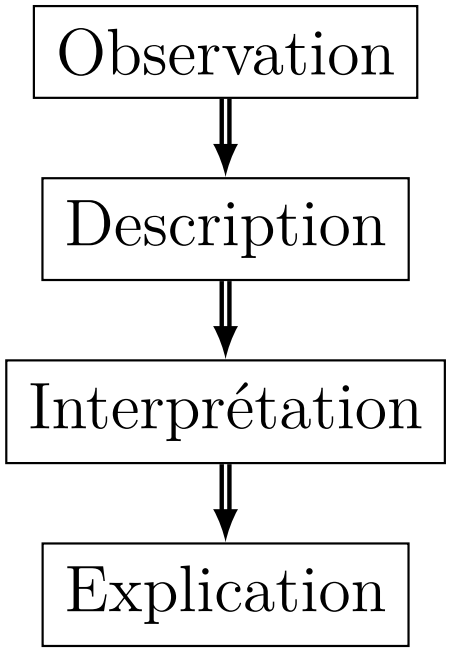
Les outils statistiques



- Les outils statistiques aident à modéliser des **corrélations** entre des données selon des **paramètres à déterminer**
- Quelques notions de base
  - **Fréquence** : nombre d'occurrences
  - **Probabilité** : nombre compris dans $[0, 1]$
  - **Loi** : distribution des probabilités

Statistiques pour les sciences

- **Découvertes scientifiques**



- **Corrélation** n'est pas **causalité**
  - Parler russe et boire de la vodka
  - Bricoler dans un garage et devenir millionaire
  - Acheter des glaces et porter des tongs

Épistémologie : règles logiques

- Expliquer un phénomène selon des **conditions**
  - **Nécessaire** : $\neg a \Rightarrow \neg b$
  - **Suffisante** : $a \Rightarrow b$
  - **Nécessaire et suffisante** : $a \Leftrightarrow b$
- **Règles logiques** pour l'inférence
  - **Déduction** :

$$\frac{(a \Rightarrow b) \wedge a}{b}$$

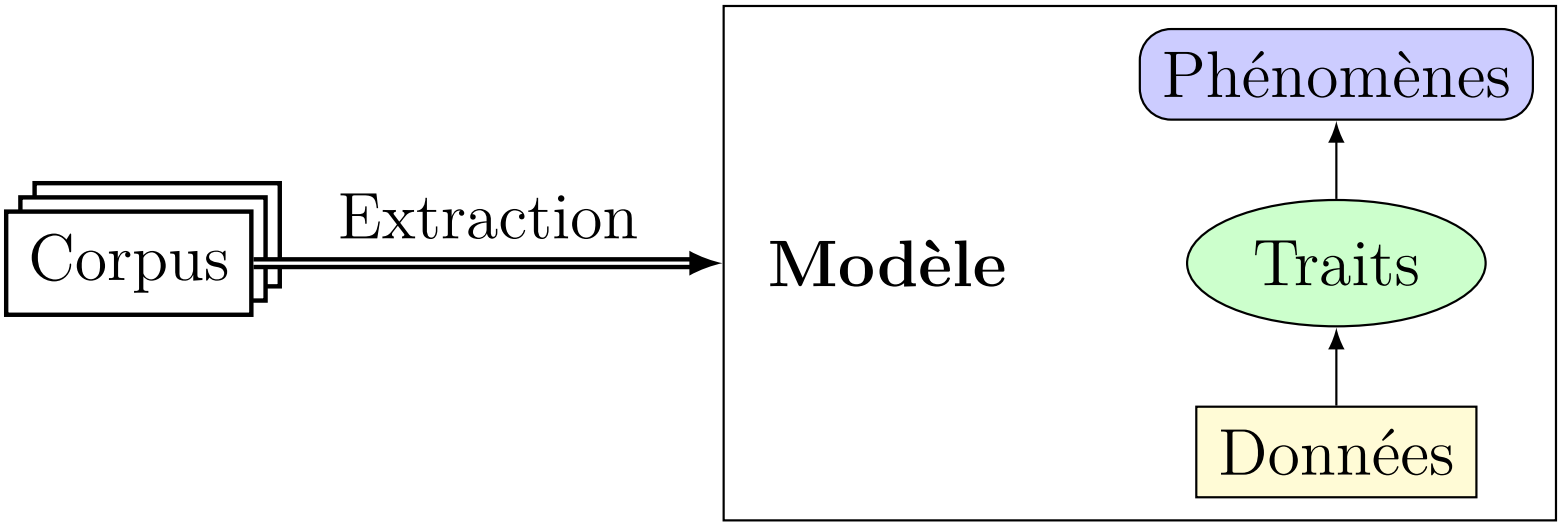
  - **Abduction** :

$$\frac{(a \Rightarrow b) \wedge b}{a}$$

  - **Induction** :

$$\frac{a \wedge b}{a \Rightarrow b}$$

Corpus, modèle, tâche



- Les corpus permettent de **décrire des phénomènes**
  - **Paramétrage** du modèle (plus ou moins *supervisé*)
  - **Evaluation** du modèle sur d'autres données
  - **Utilisation** du modèle pour d'autres tâches

⇒ Les **statistiques** permettent de **concevoir**, de **paramétrer** et d'**évaluer** l'adéquation du modèle pour une tâche donnée

Plan

1. Généralités

2. Contexte général

3. Corpus et statistiques

4. Visées applicatives

Exemples d'applications

- Niveaux de **représentation** du langage
  - Acoustique (oral)
  - Phonétique (oral)
  - Morphologie
  - Syntaxe
  - Sémantique
  - Pragmatique
- **Statistiques de corpus** pour
  - **Indexation** de documents et **recherche d'information**
  - Correction **orthographique** et **grammaticale**
  - **Reconnaissance** automatique de l'**écriture**
  - **Reconnaissance** automatique de la **parole**
  - **Traduction automatique**
  - ... (et bien d'autres)

Pourquoi les statistiques en TAL

- **Historique** rapide et incomplet des **approches TAL**
  - **Symboliques** : automates, transducteurs
  - **Heuristiques** : logique, affection manuel de poids de règles
  - **Numériques**, dont par exemple
    - Approches bayésiennes
    - Chaînes de Markov
    - Arbres de décision
    - Maximum d'entropie / régression logistique
    - SVM, CRF
    - Réseaux de neurones (deep learning, DNN, LSTM)

  ⇒ Plus de **numérique**, plus de **statistiques**
- Importance accrue des statistiques pour la linguistique
  - Mise à disposition d' **importants volumes** de textes
  - Exigence de **robustesse** des applications

  ⇒ Attention au paramétrages (baselines) !

Littérature

- Quelques références
  - Sites web
    - Michèle Jardino `http://archives.limsi.fr/Individu/jardino/coursTCAN2005.pdf`
    - Marti Hearst `http://courses.ischool.berkeley.edu/i256/f06/sched.html`
    - Christopher Manning `http://web.stanford.edu/class/cs276b/`
    - Peter Norvig `http://norvig.com/chomsky.html`
    - Revue TAL `http://atala.org/-Revue-TAL-`
    - Revue CL `http://cljournal.org/`
  - Livres
    - Initiation aux méthodes de la stat. ling. (Muller, 1993)
    - Modèles statistique pour l'accès à l'information textuelle (Gaussier, Yvon, 2011)
    - Statistique textuelle (Lebart, Salem, 1994)
    - Foundation of Statistical NLP (Manning, Schütze, 1999)
    - The Handbook of CL and NLP (Clark, Fox, Lappin, 2010)
    - Natural Language Annot. for Machine Learning (Pustejovsky, Stubbs, 2012)