Introduction

Damien Nouvel

Plan

Séances et modalités de contrôle

- **Séances**
  - 12 séances
  - Chaque séance : **cours** et **exercices** sur machine
- **Modalités de contrôle**
  - Un **contrôle** (S1) : 50%
  - Un **examen final** : 50%

Contenu du cours

- **Progression**
  - Introduction
  - Reconnaissance de l'écriture et de la parole
  - Extraction d'informations et indexation de documents
  - Analyse et représentation de la langue
  - Traduction automatique et assistée
  - Agents dialogiques

Plan

Considérations linguistiques

Tablette d'Uruk (-3000)

- Moyen de **communication** (oral, écrit, signé, etc.)
- Classification : groupes, langues, dialectes

- Naturelles (biolangues) vs construites (conlangs, idéolangues)

⇒ `https://www.youtube.com/watch?v=rupVq4m8a8g`

- **Structure** : discours, propositions, mots, syllabes / morphèmes, lettres / sons
- Double articulation (Martinet, 1961)
  - **Phonèmes** : unités minimales de prononciation
  - **Morphèmes** ou **monèmes** : unités minimales de sens

⇒ Comment **décomposer** le langage ?

Le langage : description vs. prescription

- **Description**
  - Observations pour l'explication (et la reproduction)
  - Langues **naturelles**

  ⇒ Comprendre le mode de fonctionnement ?
- **Prescription**
  - Contraintes lors de l'apprentissage (syntaxe, lexique)
  - Langues **artificelles** (y compris informatiques)

  ⇒ Rationnaliser, normaliser, uniformaliser la langue ...

⇒ Opposition de concepts indissociables

⇒ Langues **naturelles** et **artificielles**

Support des langues : orales, écrites, numériques

- Quelques dates
  - (?) : langues orales
  - ≈ -4000 : écriture cunéiforme (tablettes, Uruk, Irak)
  - ≈ -4000 : papyrus (rouleaux, Egypte)

Tablette d'Uruk

  - ≈ -200 : papier (Chine), parchemin et codex (Rome)
  - ≈ 1400 : imprimerie (Allemagne, Gutenberg)
  - ≈ 1800 : machine à écrire
  - ≈ 1950 : ordinateur (USA, von Neumann, Turing)
  - ≈ 1990 : téléphone portable

Langues et langages

- Quelques questions (quasi-philosophiques) à se poser
  - À quoi **sert** une langue ?
  - Comment **fonctionne** une langue ?
  - Quelles **unités** forment la langue ?
  - Comment se **combinent**-elles ?
  - Comment **représenter** la **sémantique** ?
  - …
- Fonctions du langage (Jakobson, 1963)
  - **Expressive** : expression des sentiments du locuteur
  - **Conative** : fonction relative au récepteur
  - **Phatique** : mise en place et maintien de la communication
  - **Métalinguistique** : le code devient objet du message
  - **Référentielle** : le message renvoie au monde extérieur
  - **Poétique** : la forme du texte devient l'essentiel du message

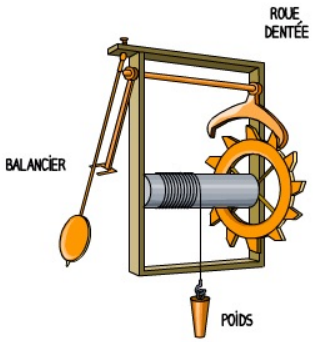
Plan

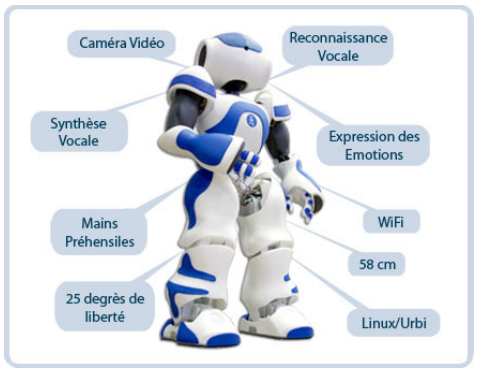
Automate, mécanique

Automate Jaquet-Droz (1770)

Qu'est-ce qu'un automate

- Du grec « automaton » : qui se meut par soi-même, qui imite les mouvements
- Pas nécessairement un « robot »





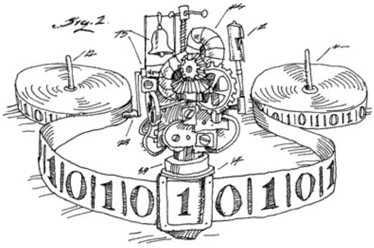
Pendule et robot Nao

- Horloges (XIII$^{\text{ème}}$) et pendules
- Pascaline (XVII$^{\text{ème}}$)
- Boites à musique (XIX$^{\text{ème}}$) et orgues de barbarie
- Ordinateurs (XX$^{\text{ème}}$)
- Nao (2010)

$\Rightarrow$ Pas d'intervention humaine, mécanisme indépendant

L'informatique comme calcul de symboles

- Dispositif purement artificiel
- Fonctionnement décrit en théorie par Turing (1936)
  - **Automate** qui **accepte** / **reconnaît** un langage
  - Machine comme : **états** et **transitions**



Machine de Turing (1936)

⇒ Formalisation du calcul, mais aussi de la **procédure**

⇒ Conçu pour le calcul, **massivement** exploité pour le **langage**

Informatique

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

Coder la langue

- Codage commme **enregistrement**
  - Des données vocales (MP3, WMA, OGG)
  - Des données images (JPG, PNG, GIF)
  - Des données textuelles (binaire, ASCII, ISO, UTF8/Unicode)
- Codage comme **représentation**
  - Lexiques, dictionnaires (listes, structures de traits)
  - Énoncés, phrases, discours
  - Documents (Word, PDF)
  - Connaissances (web sémantique)
- Codage commme **calcul** ...

Plan

Caractérisations

- 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

$\Rightarrow$ **TAL** : **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**

Capacités de traitement

- **Linguistique** : langues et langages
  - Détournement de Aspects de la théorie syntaxique (Chomsky, 1965)
    - **Compétence** : capacités génératives du *langage*
    - **Performance** : observations sur la *langue*

  ⇒ Écart entre **théorie** et **pratique**
- **Informatique** et traitements automatiques
  - **Automate** : pas d'intervention humaine
  - Utilisation de **ressources** (lexiques, grammaires)

  ⇒ Déterminisme, codes, programmes
- Capacités en terme de
  - **Temps** (processeurs)
  - **Espace** (mémoire, disques)

  ⇒ Complexité des analyses (fouille de données)

Quelques repères historiques du TAL

- Quelques dates approximatives
  - ≈ 1955 : Traduction automatique
  - ≈ 1955 : IA (Intelligence Artificielle)
  - ≈ 1957 : Grammaire générative, hiérarchie (Chomsky)
  - ≈ 1965 : Systèmes de dialogue (ELIZA, SHRDLU)
  - ≈ 1967 : Brown Corpus
  - ≈ 1970 : Reconnaissance de la parole et synthèse vocale
  - ≈ 1980 : Réseaux sémantiques (Wordnet, graphes conceptuels)
  - ≈ 1990 : Unicode
  - ≈ 1995 : Internet
  - ≈ 1995 : Recherche d'information structurée (entités nommées)
  - ≈ 1997 : FrenchTreeBank
  - ≈ 2005 : Liaison de données textuelles (Wikipedia)
  - ≈ 2012 : Deep Learning (word2vec)

Symboles, données, connaissances

- Deux types d'approches en TAL
  - **Symboliques** (automates)
  - **Numériques** (apprentissage automatique)

⇒ Elles sont **complémentaires**

- L'intelligence artificielle
  - Simuler l'humain (robotique, Nao)
  - Transhumanisme, humain augmenté, **singularité** ...

⇒ Rassurez-vous, on y est pas encore ...

- Par exemple, testons quelques agents conversationnels
  - `http://www.eliza.levillage.org`
  - `http://www.jabberwacky.com/george`
  - `http://www.mitsuku.com/`
  - `http://ec2-54-215-197-164.us-west-1.compute.amazonaws.com/speech.php`
  - et bien d'autres `https://www.chatbots.org`

TP

- Éditez un fichier texte et copiez-collez une page web
- Faites un programme qui affiche pour ce fichier
  - Le nombre d'octets
  - Le nombre de caractères Unicode
  - Le nombre de mots
  - Le nombre de phrases