Traitement Automatique des Langues

Jean-Yves Antoine

Université François Rabelais de Tours

www.info.univ-tours.fr/~antoine

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

Traitement Automatique des Langues

INTRODUCTION : TECHNOLOGIES LANGAGIERES

Technologies langagières

- **Technologies langagières**

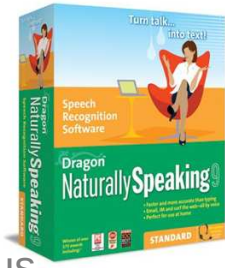
  informatique appliquée à un type particulier de données, le **langage naturel** : parole, document écrit.

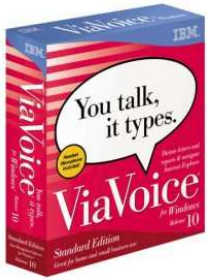
- **Pourquoi travailler sur le langage naturel ?**

  – information textuelle à traiter (traduction, recherche documentaire, SMS…)
  – seule modalité d ’accès à l ’information (téléphonie, contrôle vocal de systèmes)
  – multimodalité : modalité complémentaire permettant soit de faciliter l’utilisation d’un système (*monitoring vocal* : limitation charge cognitive) ou de l ’enrichir.

Quelques applications : parole

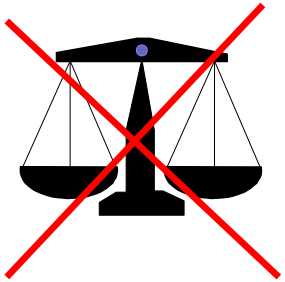
- **Reconnaissance de la parole**
  - ASR : *Automatic Speech Recognition*
  - parole interactive (dialogue) ou non (dictée vocale)





Exemples d'application : saisie automatique de compte rendus médicaux, aide à la communication pour personnes handicapées, sous-titrage automatique d'émissions radiodiffusées

- **Reconnaissance du locuteur**
  - vérification ou identification
  - dépendante ou indépendante du texte
  - **Remarque** : non infaillible à 100 %

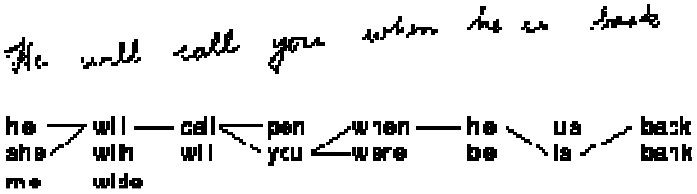


- **Synthèse de parole**
  - à partir du texte (*TTS* : *text-to-speech synthesis*)
  - synthèse par règle ⇨ par concaténation ⇨ par sélection d'unités
  - **démo** : LIMSI 68 & CNET 76 ⇨ CNET 93 ⇨ FT R&D 2003

Démonstrateur FT R&D : http://tts.elibel.tm.fr

Quelques applications : écrit

- **Reconnaissance de caractères**
  - OCR : *Optical Character Recognition*
  - Texte imprimé ou manuscrit
  - en contexte (texte) ou hors-contexte (caractères isolés)



Exemple d'application : numérisation du patrimoine de la BNF

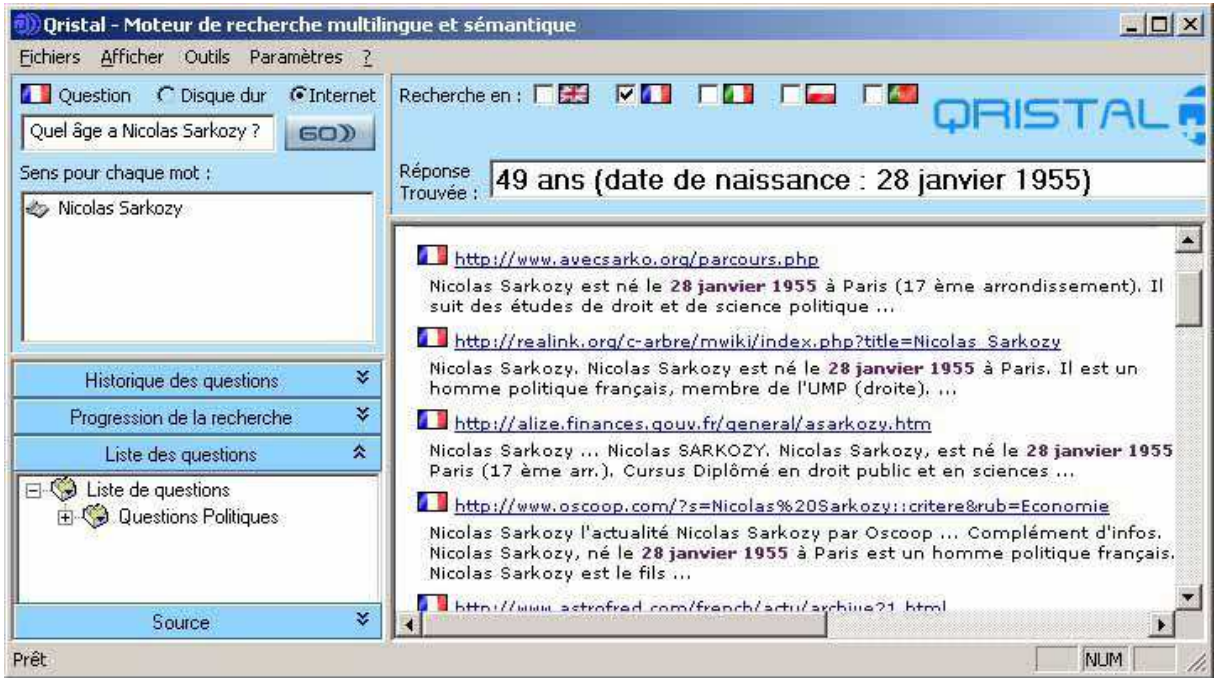
- **Traduction automatique**
  - traduction « off-line » ou assistée par ordinateur (TAO)
  - traduction rapide de textes (grand public), de documentations techniques, localisation (adaptation à l'environnement culturel)

Quelques applications : document

Recherche documentaire et indexation

- **recherche documentaire** : donner une liste de documents pertinents
- **recherche d'information** : extraire une information précise d 'un ensemble de documents.
- **dimension interactive** : question/réponse (*question/answering*)
- **Remarque** : indexation préalable (manuelle ou automatique)

Exemple : indexation et interrogation des journaux en ligne (Ouest-France, Le Monde)



Quelques applications : document

- **Ressources terminologiques**
  - **outils d'aide à la construction** de ressources terminologiques
  - aide à la traduction : bases terminologiques multilingues
  - terminologies de référence pour la documentation technique (extraction automatique de termes complexes)

- **Fouille de texte**
  - Veille technologique, recherche d'information

  Exemple : SAS Miner

Quelques applications : dialogue

- **Dialogue Homme-Machine**
  - **Exemple** : réservation de billets, renseignement touristique par téléphone
  - **Dialogue médié par l'ordinateur**
    - Traduction parole – parole
    - Aide la communication pour personnes handicapées

- **Autres applications**
  - Fouille de texte et veille technologique
  - Résumé automatique
  - …

Acteurs économiques

- **Acteurs économiques**
  - commercialisation et/ou R&D
  - **grandes entreprises généralistes** en télécommunications, informatique et production/contrôle de l ’information et des médias.
  - **PME** spécialisées dans le domaine

- **Reconnaissance Automatique de la Parole**

PHILIPS sense and simplicity — IBM — NUANCE

Microsoft — telisma your voice in action — Lucent Technologies Bell Labs Innovations

- **Synthèse de la parole**

acapela group — Microsoft — france telecom

Acteurs économiques

- Traduction Automatique

- Recherche d'information textuelle

- Ressources terminologiques

- Reconnaissance de caractères

Objectifs du cours

- **Comprendre un domaine technologique en essor**
  - Comprendre les objets manipulés (linguistique) et leur traitement (notions de base)
  - Cours entre théorie et pratique : des notions sur la linguistique et le traitement automatique, et une partie pratique consacrée à l'utilisation d'outils

- **Techniques en Traitement Automatique des Langues (TAL)**
  - 22 h de cours seulement
  - Fondements Linguistiques du TAL : 12 heures
  - Quelques outils et applications du TAL : 10 heures

- **Enseignants**
  - Nathalie FRIBURGER, Agata SAVARY, Jean-Yves ANTOINE

Bibliographie

- **Pierrel J.-M.** (2000) Ingénierie des langues. Hermès, Paris, France

- **Cole R.A., Mariani J., Uszkoreit H., Zaenen A., Zue V.** (Eds.) *Survey of the state of the art in Human language technology*. CSLU, Oregon. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

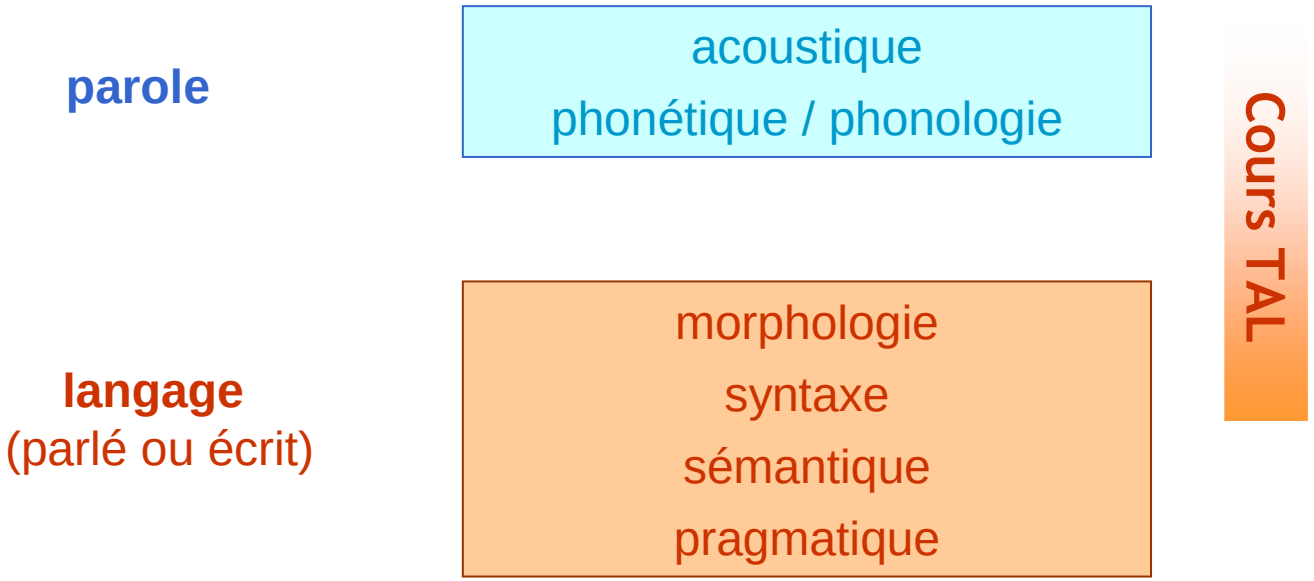
- **Mitkov R.** (Eds.) (2003) The Oxford handbook of computational linguistics. Oxford University Press Inc., Mew-York, NJ.

Traitement Automatique des Langues

THEORIE LINGUISTIQUE
VUE D'ENSEMBLE

Niveaux d'analyse du langage

- Niveaux de description



- **Vision restrictive :** interdépendances entre niveaux

Niveaux infra-linguistiques : parole

- **Acoustique**
  - caractéristiques du signal de parole : **traitement du signal**
- **Phonétique**
  - étude de la production de la parole par l ’appareil phonatoire
  - étude de la perception de la parole par le système auditif
  - ⇨ classification des sons en parole en **classes phonétiques**

    exemple *:* voyelles vs. consonnes
- **Phonologie**
  - interface entre les sons de parole (phonétique) et les niveaux linguistiques
  - **phonème** : son « abstrait » à plusieurs réalisations phonétiques suivant le contexte

Morphologie

- étude de la construction des mots (formes lexicales) à partir d'unités minimales de sens : **morphèmes**

  étude ⇨ *études, étudier, étudiées, étudiant, …*
  lasser ⇨ *délasser ;**porte-avions, porte-feuille***

- la catégorie grammaticale d'un mot (*partie du discours* ou *catégorie morpho-syntaxique*) voire son rôle syntaxique peut être expliquée par sa composition morphologique (application : *guesser*)

Nominatif (sujet ou attribut)	*der gute Mann*
Génitif (complément du nom)	*des guten Mannes*
Datif (COI et comlément d'attribution)	*dem guten Mann*
Accusatif (COD)	*den guten Mann*

- influence sur la prononciation

  Les poules du couvent couvent

Lexique : paradigmes, entités nommées, termes de spécialités …

Syntaxe

- Étude des énoncés grammaticalement corrects sous la forme des relations structurales entre les mots qui les composent.
- Analyse basée le plus souvent sur les catégories morpho-syntaxiques (POS : *Part-of-Speech* ou parties du discours)
- **Jugement de grammaticalité**

  Cette phrase est grammaticalement correcte

  Cette phrase pas grammaticalement correcte être

  Ce phrase ne sont pas grammaticalement correctes

  Cette herbe est anticonstitutionnellement rôtie

- **Structure syntaxique** essentielle à la compréhension

  Jean aime Marie *Je vois la fenêtre de ma chambre*

- [Chomsky 1956] grammaires formelles ⇨ langages informatiques

Sémantique

Étude **hors-contexte** du sens des mots (sémantique lexicale) et de leur combinaison pour former la signification **littérale** de l 'énoncé.

- **Sémantique lexicale**
  - caractérisation de classes sémantiques à l'aide de traits (*sèmes*)

    chat, matou, minou : /animé/ + /animal/ + /félin/ + /domestique/
  - liens entre la significations des mots

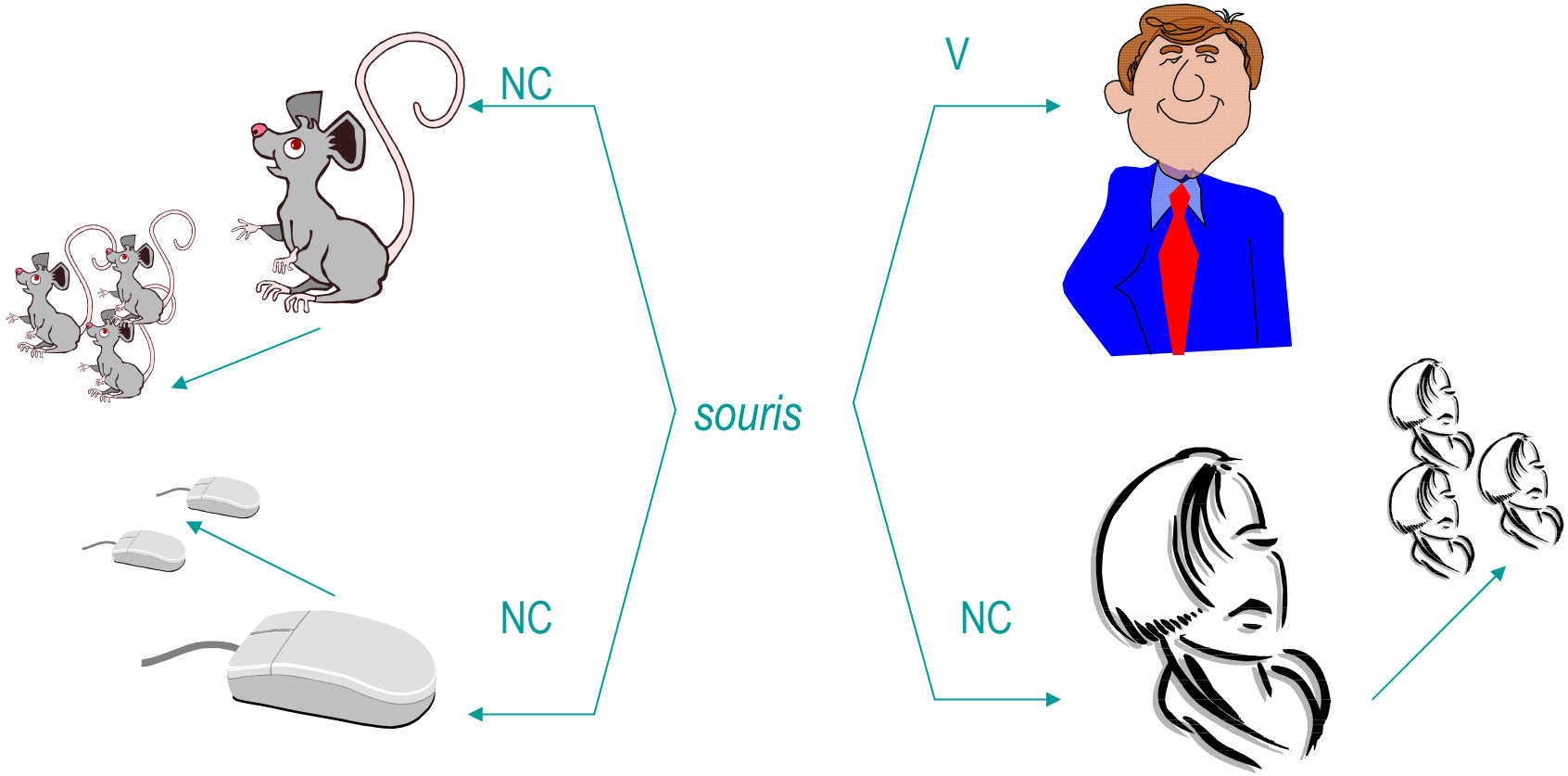
    - **Wordnet** *:* synonymie/antinomie, hyponymie/hyperonimie (est_un), partie_de
    - **LSA** (*Latent Semantic Analysis*) : *co-occurrences traduisant une affinité de sens*
- **Sémantique de l'énoncé**
  - **structure sémantique** : sens = ensemble des dépendances sémantiques (prédicat/argument) entre les éléments de l'énoncé

    Jacques donne le sac à Bernadette

    donner(AGENT = Jacques ; OBJET = sac ; PATIENT = Bernadette)
  - la syntaxe reflète en partie le sens mais ne le fonde pas : relations entre structures syntaxique et sémantique

Sémantique

- **Polysémie** : plusieurs sens pour une même entité lexicale



problème plus général : **ambiguïté** intrinsèque du langage naturel

Exemples *Un avocat véreux*
Je vois le toit de ma ferme

Pragmatique

- **Interprétation en contexte**
  - **référence : contexte de la tâche** — déterminer l'objet de la tâche associé à un élément du discours

    *Pouvez-vous me sortir **le dossier de Monsieur Durand***

    *Je cherche un hôtel **au sud de la rocade*** ⇨référence spatiale
  - **co-référence : contexte du discours** (ou du **dialogue** en DHM)

    *Jacques voit le sac de Bernadette. Il **le lui** donne. **Elle le** remercie puis **le** sert sert vigoureusement contre **elle**.* ⇨anaphores
  - **contexte de l 'univers du discours**(*world knowledge*)

    J 'ai réservé deux classes affaires sur le AF2031 d 'aujourd'hui ⇨ellipses
- **Dialogue**
  - **intention de l'interlocuteur**: actes de dialogue (*speech acts* [Searle 69])

    informer, demander, ordonner, accepter...
  - **gestion du dialogue** : grammaires de dialogue, gestion par plan

Bibliographie

- **Allen J.** (1995) Natural Language Understanding. Benjamin / Cummings Publ. Comp. Redwood City, CA. (chap. 1)
- **Huang X., Acero A., Hon H-W.** (2001) Spoken Language Processing : a guide to theory, algorithm and system development. Pretice Hall, Upper Saddle River, NJ. (chap. 2)
- **Fuchs C.** (Dir.) (1993) *Linguistique et Traitement Automatique des Langues*. Coll. Hachette Supérieur. Hachette, Paris.