Traitement Automatique des Langues pour les Systèmes d'Information

Introduction

Damien Nouvel
Nathalie Friburger
Jean-Yves Antoine

Introduction
Organisation du cours

- Introduction : 1h cours (D. Nouvel)
- Morphologie, terminologie et lexiques : 3h cours, 2h TP (D. Nouvel)
- Etiquetage morpho-syntaxique : 2h cours, 2h TP (D. Nouvel)
- Entité nommées : 2h cours, 4h TP (N. Friburger)
- Syntaxe et sémantique : 4h cours, 2h TP (J.Y. Antoine)

Introduction
Plan

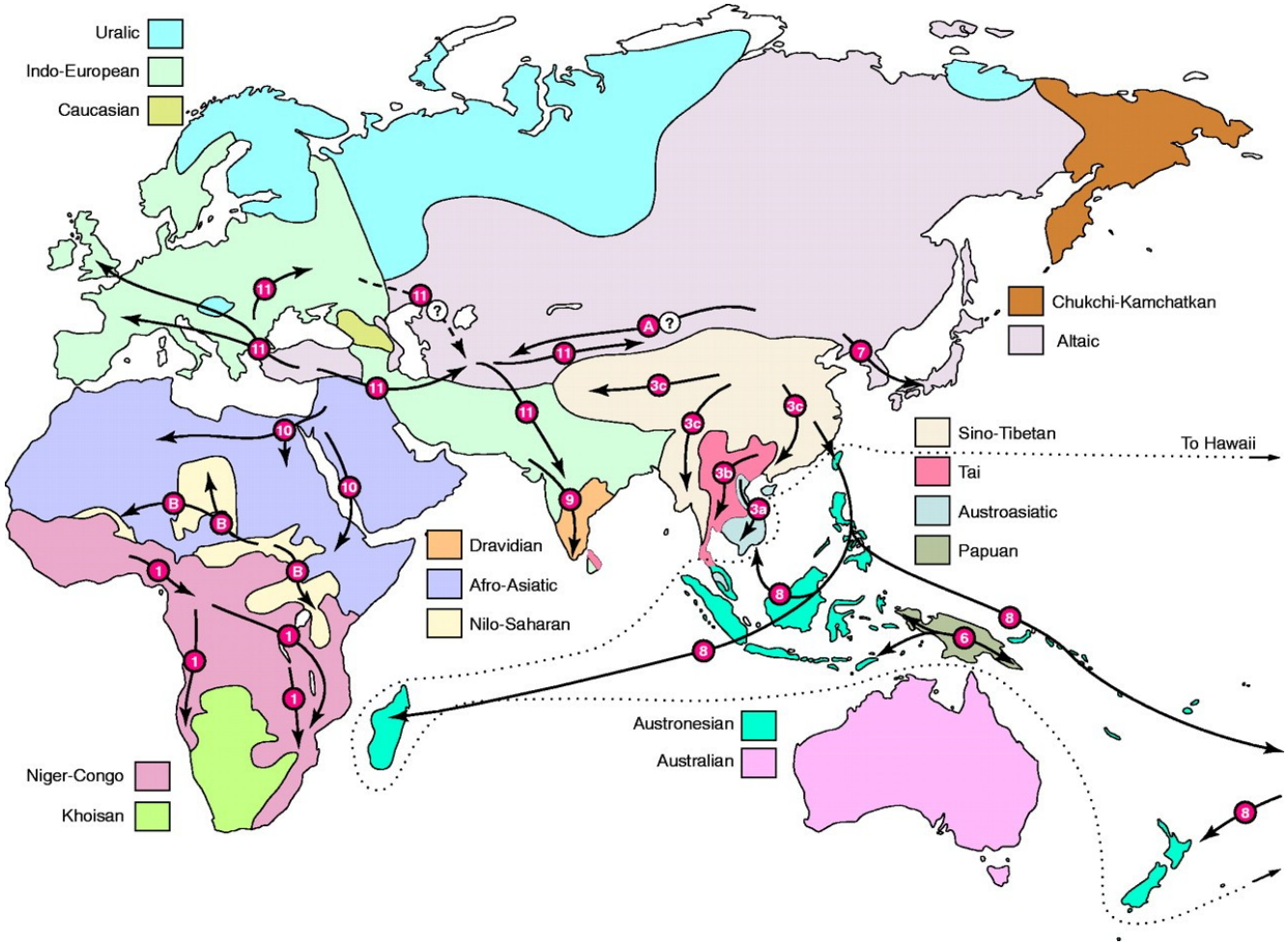
- Le langage et l'informatique
- Représentation du langage
- Le TAL, pour quoi faire ?
- Les enjeux du TAL pour les SI

Introduction
Plan

- Le langage et l'informatique
- Représentation du langage
- Le TAL, pour quoi faire ?
- Les enjeux du TAL pour les SI

- Le langage comme aptitude à la communication / transmission d'informations :
  - **Émission** des messages (gestes, cordes vocales, stylo, clavier, braille) de manière **intelligible** par le locuteur
  - **Réception** des messages (vision, ouïe, toucher, écrans) **interprétation et compréhension** par l'interlocuteur
- Origine du langage :
  - **Parlé (≈3000)** : ≈ -200 000 ans, homo sapiens ?
  - **Ecrit (≈350)** : ≈ -6 000 ans, Uruk ?
  - Informatique : 1951, A0 (Hopper, 1er compilé)
  - 1865 : la Société Linguistique de Paris indique ne plus recevoir aucune communication concernant l'origine du langage

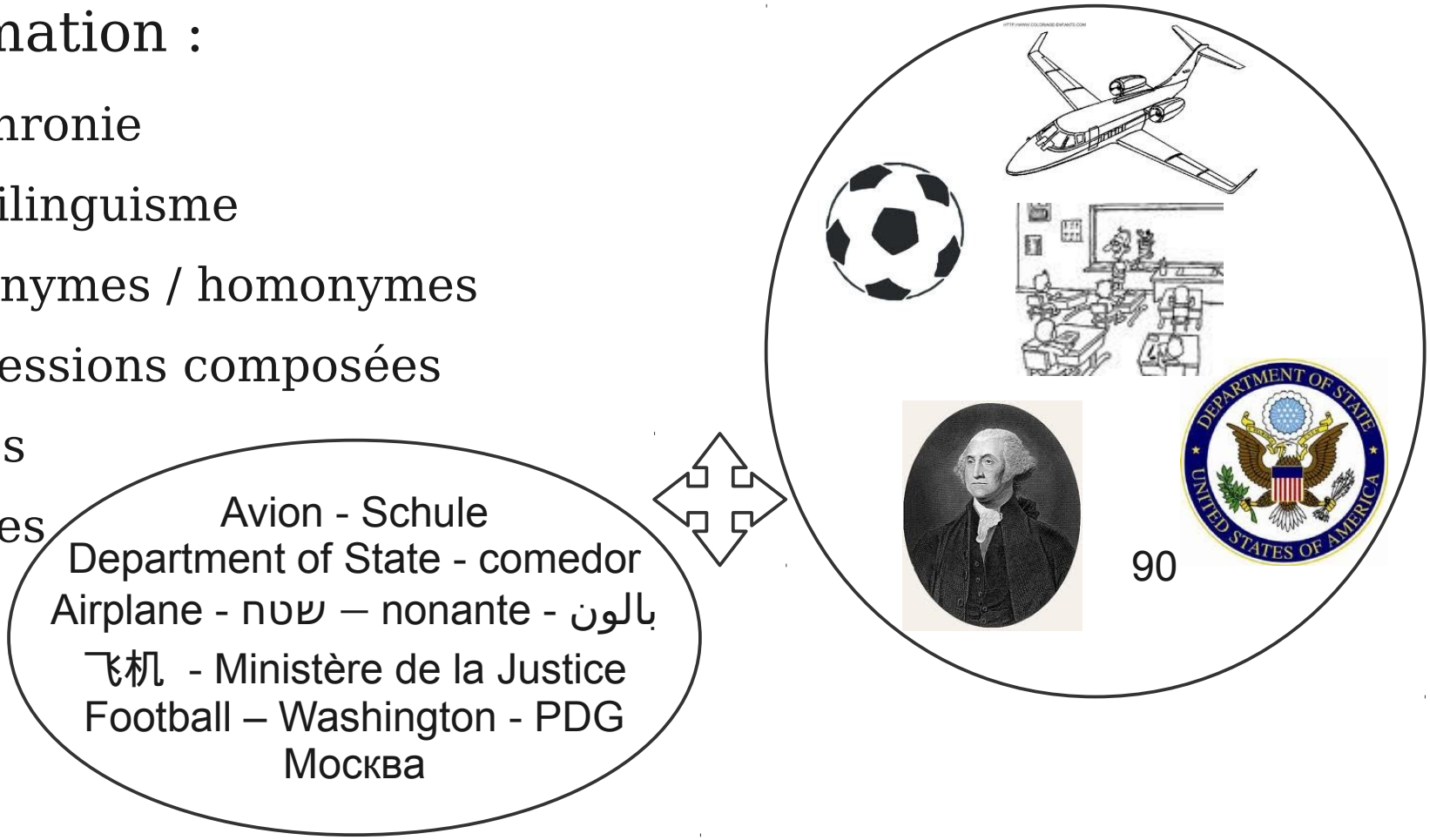
- Les grandes familles de langues :



- Le langage permet :
  - La **concision** grâce à l'**implicite** et la déduction
  - La **souplesse** et la **correction d'erreur** : « Même lorquse l'on dégarde faurtemnt un mesaje, il retse comrpéhansib par l'inlertocuteur »
  - La **générativité** une phrase est compréhensible grâce aux éléments connus dont elle est composée (compositionnalité)
  - La **multiplicité des modes d'expression** (langues, écrit / oral, signes) et l'**évolutivité** (convention sociale)
- Difficultés pour traiter le langage automatiquement :
  - La concision, les erreurs (bruit), très large espace des expressions possibles et des modes d'expression...

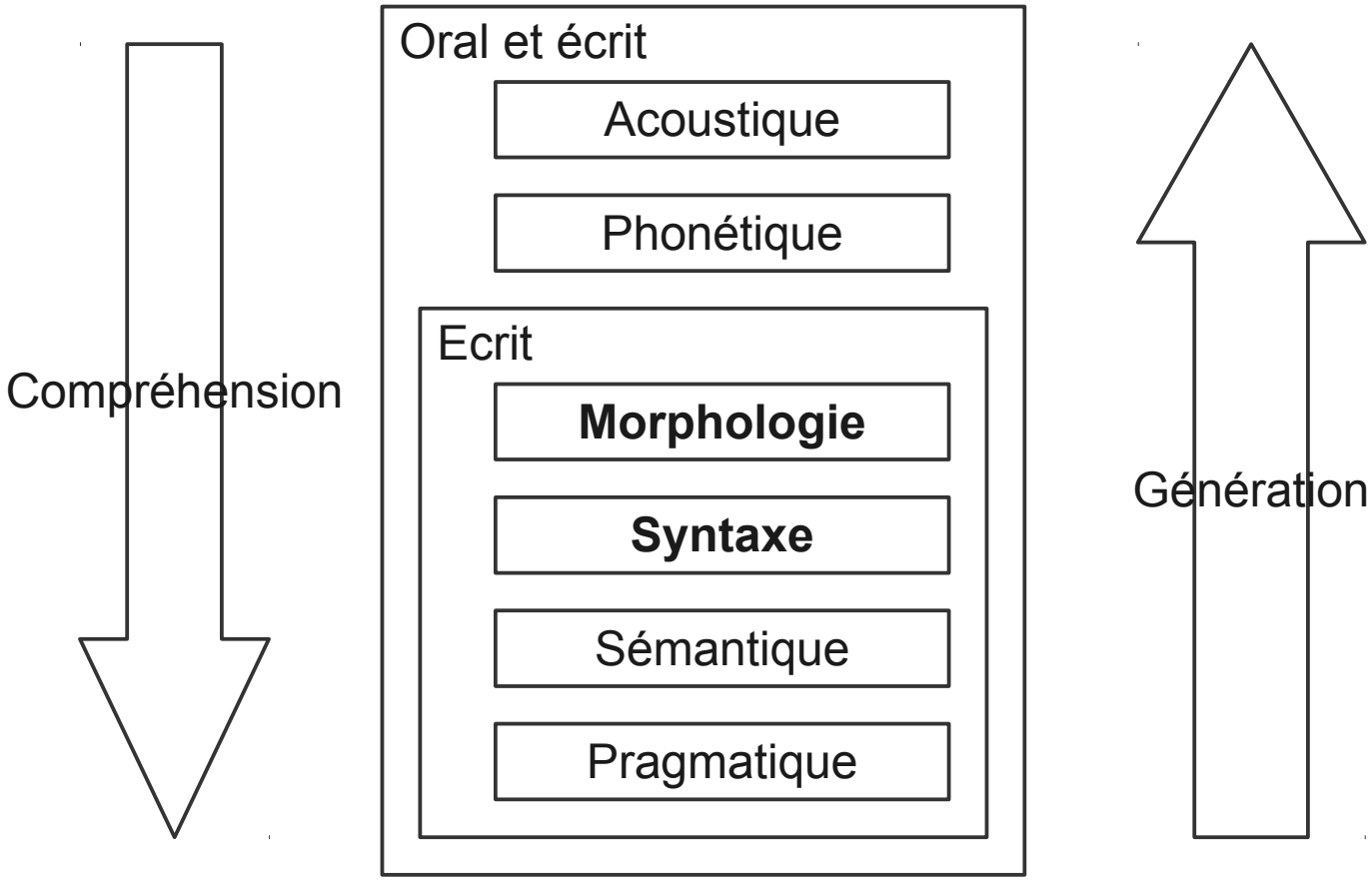
- **De plus en plus de langage** dans l'informatique :
  - Compilation de programmes
  - Communication (emails, Internet, chat, présentation)
  - Gestion documentaire
  - Encyclopédies (Wikipedia, dictionnaires)
  - Traitement de documents multimédias (audio, vidéo)
  - Social networking (Facebook, Twitter, Viadeo, LinkedIn)
  - … ?
- L'informatique évolue : depuis le « calculateur » les machines deviennent des **outils de communication** :
  - Réseaux, smartphones, gps, serveur vocal, traduction, etc.

- Comprendre le langage : un challenge en traitement de l'information :
  - Diachronie
  - Multilinguisme
  - Synonymes / homonymes
  - Expressions composées
  - Sigles
  - Tropes
  - ...



- Le langage et l'informatique
- Représentation du langage
- Le TAL, pour quoi faire ?
- Les enjeux du TAL pour les SI

- Les différents niveaux pour représenter le langage :

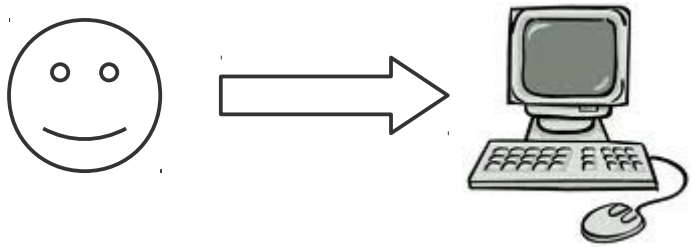


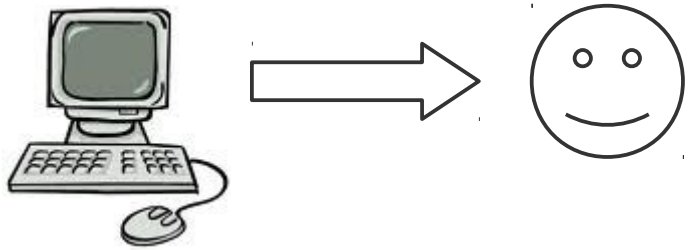
- Formats informatique pour le langage :
  - Document audio / vidéo :
    - mp3, wav, aac, avi, flv...
  - Documents textes, encodage des caractères :
    - ASCII, ISO (8859-1), **UTF (8)**, KOI8-R, EBCDIC, Windows 1250...
  - Documents textes, mise en forme :
    - HTML, Word, Latex, PDF, RDF, OWL...
  - Documents textes, structuration et annotation :
    - XML, Word, Latex, OWL...
  - Bases de documents :
    - Fichiers, CMS, DB2, Lotus Notes, Documentum, Jahia, Alfresco...

- Le langage et l'informatique
- Représentation du langage
- Le TAL, pour quoi faire ?
- Les enjeux du TAL pour les SI

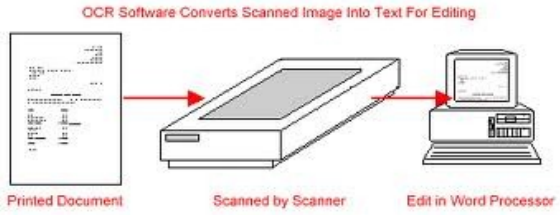
Introduction
Le TAL, pour quoi faire ?

- Traitement Automatique des Langues :
  - Compréhension :
    - Encodage, stockage de documents
    - **Classification** de documents
    - **Extraction d'information**
    - **Recherche d'information**
  - Génération :
    - Restitution
    - Visualisation
    - Résumé
    - Synthèse





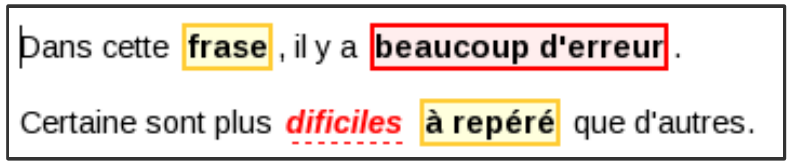
Introduction
Le TAL, pour quoi faire ?



- Reconnaissance de caractères (OCR) :
  - Principe général :
    - Numérisation de documents écrits (scanner) en images
    - Application de techniques de reconnaissance de formes (lettres) à l'aide d'apprentissage (réseaux de neurones, HMM)
    - Exploitation d'un **modèle de langage** (dont des ressources : dictionnaires, grammaires, etc.) pour déterminer l'hypothèse la plus probable
  - Applications pratiques : dématérialisation de documents (bibliothèques), formulaires (chèques, administration), adresses pour le tri postal, identification d'immatriculation
  - Industriels du domaine : Nuance (ScanSoft / Xerox), ABBYY, IRIS, NovoDynamics, Datacap, EDT, Ligature

Introduction
Le TAL, pour quoi faire ?

- Correction orthographique / grammaticale :
  - Principe général :
    - Identifier les mots (tokenization)
    - Correction orthographique : mots qui n'appartiennent pas au **dictionnaire** et qui ne sont pas en langue étrangère, ni des noms propres, ni des chiffres, ni des sigles...
    - Correction grammaticale : déterminer la fonction des mots au sein de la phrase (déterminant, nom, verbe, adverbe, etc.) puis réaliser une analyse syntaxique à l'aide de **grammaires**



  - Applications pratiques : correxion de document rédiger par des étudiants dont le nivau de français ai (parfoi) aflijant
  - Industriels du domaine (en français) : Synapse, Druide, Microsoft, Diagonal, Machinæ Sapiens

Introduction
Le TAL, pour quoi faire ?

- Traduction automatique :
  - Principe général :
    - Sélection des langues source et cible
    - Deux stratégies (actuellement) :
      - Utilisation d'un **modèle de langage** pour la source et pour la cible, éventuellement d'un modèle « pivot »
      - Recherche des traductions possibles et probables à travers des **corpus** « comparables » et « alignés »
  - Applications pratiques : traduction de documents (notamment au sein de l'UE), dictionnaires bilingues, recherche d'informations multilingue
  - Industriels du domaine : Systran, Softissimo, Avanquest, Prompt, WorldLingo

Introduction
Le TAL, pour quoi faire ?

- Extraction et recherche d'informations :
  - Principe général :
    - Enregistrer des documents (ou leurs adresses) et déterminer un ensemble de **caractéristiques** selon leur analyse
    - Construire des **indexs** accessibles et régulièrement mis à jour
    - Répondre à la demande aux **requêtes** par sélection des documents les plus **pertinents**
  - Applications pratiques : recherche en ligne, veille, surveillance, résumé automatique, classification de documents
  - Industriels du domaine : Google, Yahoo, Baidu, Microsoft, Exalead, Altavista, Ask, WolframAlpha, Lucene...

- Reconnaissance de la parole :
  - Principe général :
    - Traitement **acoustique** du flux audio
    - Analyse du signal (transformée de Fourier)
    - Reconnaissance par modèles (appris : HMM ou réseaux de neurones), avec implémentation de **modèle de langage** (phonétique, N-grammes) qui donne la séquence la plus probable

  - Applications pratiques : dictaphones (smartphones), serveurs vocaux (hotline), transcriptions automatiques (sous-titres, notamment pour les malentendants)
  - Industriels du domaine : IBM, Dragon, Microsoft

- Synthèse vocale :
  - Principe général :
    - Sélection de la langue cible
    - **Transcription** phonétique du texte
    - Modélisation de l'intonation et de la prosodie
    - Production du signal audio

  - Applications pratiques : transports, serveurs vocaux (hotlines), systèmes de navigation GPS, vocalisation (notamment pour malvoyants), personnages de jeux
  - Industriels du domaine : Xerox, ElanSpeech (et beaucoup d'autres)

- Le langage et l'informatique
- Représentation du langage
- Le TAL, pour quoi faire ?
- Les enjeux du TAL pour les SI

- Les systèmes d'informations doivent satisfaire aux exigences suivantes :
  - Gestion de bases de données **volumineuses, hétérogènes** (formats, langues, présentation) et **pas nécessairement structurées** (flux de textes, de sons, d'image, de vidéos, etc.)
  - Répondre **rapidement**, de manière **pertinente** et visuellement bien présentée à des requêtes
- Les modèles de langages permettent de réaliser (semi) automatiquement certaines tâches sur des documents qui vont structurer le système d'information

- Quelques applications en vue :
  - Publicité et marketing ciblé
  - Reconnaissance vocale (émissions, téléphone)
  - Traduction automatique en temps-réel
  - Renseignement (terrorisme, espionnage)
  - Aide à la décision (Rue89 : B-6)
  - Fouille d'opinion (politique, économie)
  - Détection d'émotions (robots-compagnons)