Initiation au traitement automatique des langues

Guy Perrier

1 - Linguistique et Traitement Automatique des Langues

1. L'essor du traitement automatique des langues dans la société
2. Les différents niveaux de la langue
3. Grammaires et lexiques
4. Langages formels et langues naturelles
5. Chaîne de traitement automatique d'une langue

1.1 - L'essor du Traitement Automatique des Langues dans la société

- Avec l'explosion d'Internet et des **Nouvelles Technologies de l'Information et de la Communication** (NTIC), notre société est devenue une société de l'information.

- La **langue naturelle** est de loin le premier vecteur d'information.

- Les NTIC permettent la production et la circulation d'une masse énorme d'informations, qui requiert l'aide d'**outils automatiques** pour son **traitement**.

1.1 - L'essor du Traitement Automatique des Langues dans la société

Les principales applications du **Traitement Automatique des Langues** (TAL) :

- la traduction assistée par ordinateur (historiquement la première application, dans les années 1950),
- la correction orthographique et grammaticale,
- la recherche d'informations textuelles, les moteurs de recherche,
- la fouille de textes, l'indexation de documents
- le résumé automatique,
- la génération automatique de textes,
- la synthèse de la parole,
- la reconnaissance vocale,
- la reconnaissance de l'écriture manuscrite,
- les agents conversationnels.

1.1 - L'essor du Traitement Automatique des Langues dans la société

Quelques références sur les industries et les métiers du TAL :

- Le livre blanc sur le TAL dans les industries de l'information :

  http://www.gfii.fr/uploads/docs/9099455310a24fb4483401e66b0de8832a843895.pdf

- L'étude de marché des technologies de la langue en Europe :

  http://www.technolangue.net/IMG/pdf/EtudeMarche-Technolangue2006.pdf

- La langue française à l'ère du numérique :

  http://www.meta-net.eu/whitepapers/e-book/french.pdf

1.1 - L'essor du Traitement Automatique des Langues dans la société

- Le **Traitement Automatique des Langues** (TAL) est l'application des mathématiques et de l'informatique à la linguistique. C'est un domaine technologique particulier mais aussi c'est aussi un domaine scientifique. Il est alors parfois désigné sous le terme de **linguistique informatique**.

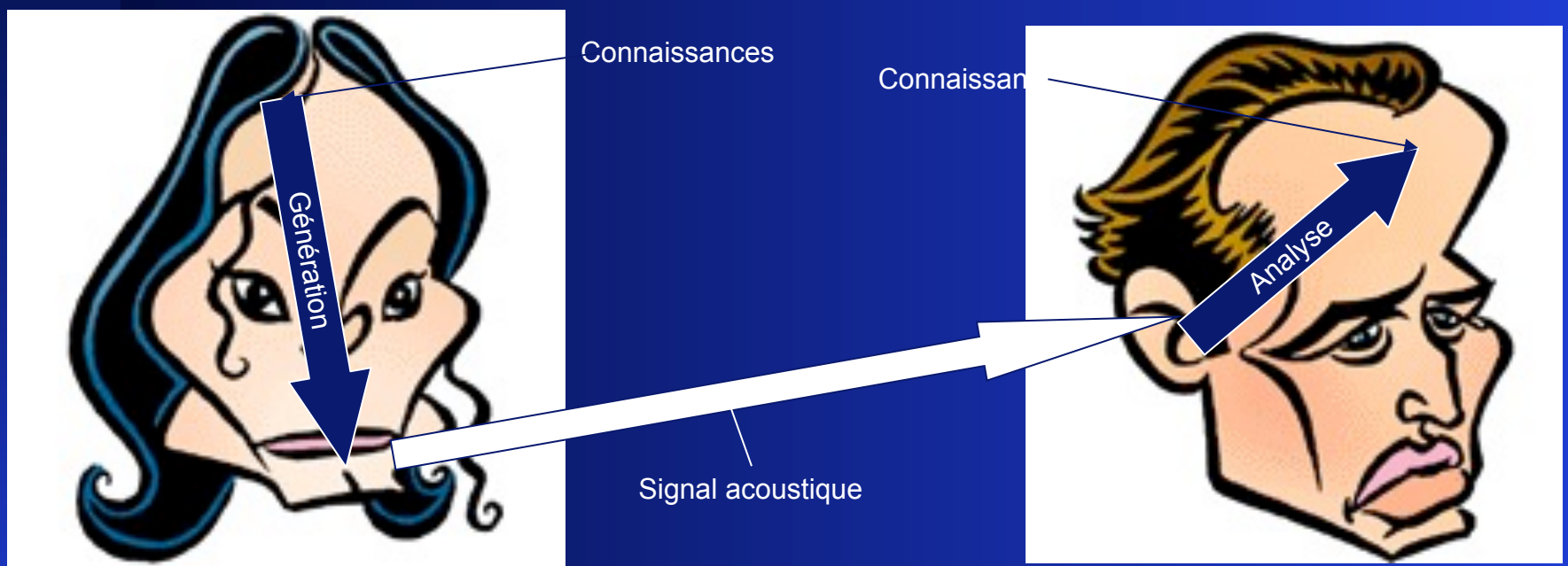
- Références sur la recherche en TAL :
  - Association for Computational Linguistics, association internationale des chercheurs en TAL : http://www.aclweb.org/
  - ATALA association française de la recherche en TAL : http://www.atala.org/
  - A Nancy, laboratoires LORIA (http://www.loria.fr/) pour l'informatique et ATILF (http://www.atilf.fr/) pour la linguistique

1.1 - L'essor du Traitement Automatique des Langues dans la société

- A Nancy, la formation universitaire en TAL est dispensée dans la spécialité TAL du master SCA de Nancy (http://talc.loria.fr/-Master-Sciences-Cognitives-et-.html).
- En première année, les étudiants ont seulement une initiation au TAL partagée avec les étudiants du master SDL.
- En seconde année, la formation est totalement orientée vers le TAL. Les étudiants ont la possibilité de personnaliser leur parcours selon qu'ils se destinent à la recherche ou à l'ingénierie.
- Cette formation est en France la seule formation de master en TAL intégrée à un programme européen Erasmus Mundus, le programme « Language and Communication Technologies (LCT) » (http://www.lct-master.org/)

1.2 - Les différents niveaux de la langue

- Le langage humain : faculté de représenter des connaissances et de communiquer



1.2 - Les différents niveaux de la langue

- Le langage humain opère dans une **société** au moyen d'un **système de signes**, une langue, pour produire et comprendre des énoncés.

- Les signes de base d'une langues sont ses **mots**. Un mot est un couple (forme phonologique - le « **signifiant** », sens - le « **signifié** ») (Ferdinand de Saussure, Cours de linguistique générale 1916).

- Du signifiant au signifié, on peut distinguer différents niveaux de la langue qui ont une relative autonomie tout en interagissant les uns avec les autres.

1.2 - Les différents niveaux de la langue

- La **phonétique** touche aux aspects physiques de la production et de la réception des sons d'une langue. Les unités élémentaires sont les **phones**.

- La **phonologie** concerne le groupement des sons pour former les mots et les énoncés d'une langue via un système de **phonèmes** et de règles phonologiques. La suite des sons réalisant un énoncé est aussi modulé par la **prosodie**.

- La **morphologie** concerne la combinaison des signes minimaux d'une langue, ses **morphèmes**, pour former des mots.

1.2 - Les différents niveaux de la langue

- **La syntaxe** touche à la combinaison des mots pour former des phrases. Selon la notion de base qui est utilisée pour expliquer la combinaison, nous avons deux points de vue : les **grammaires de dépendance** mettent en avant la notion de **dépendance** entre mots et les **grammaires syntagmatiques** mettent en avant la notion de **syntagme**, regroupement de mots.

- La **sémantique** touche au sens des énoncés indépendamment du contexte. La **logique** est habituellement le cadre utilisé pour représenter la sémantique.

- La **pragmatique** touche au sens des énoncés relativement à leur usage en contexte (discours, résolution des références, structure communicative, dialogue …)

1.3 - Grammaires et lexiques

- La **grammaire** d'une langue est un système de **catégories** et de **règles** gouvernant les niveaux phonologique, morphologique, syntaxique, sémantique et pragmatique d'une langue.

- Le **lexique** d'une langue est l'ensemble de tous ses mots avec leurs propriétés linguistiques. Un **lexème** est un élément du lexique qui contient de l'information phonologique, morphologique, syntaxique et sémantique relative à un mot de la langue.

- La grammaire et le lexique sont complémentaires : tous les deux participent à la caractérisation de la langue.

1.4 - Langages formels et langues naturelles

- Un **langage formel** L sur un alphabet fini de symboles Σ est une partie de l'ensemble Σ^* de chaînes d'éléments de Σ.
- Si un langage L est infini, il est important d'avoir une procédure de calcul pour **reconnaître** L, c'est-à-dire pour décider si une chaîne donnée de Σ^* appartient à L.
- On classe les langages formels selon la complexité des procédures permettant de les reconnaître : indécidables, récursivement énumérables, récursifs, algébriques, réguliers…

1.4 - Langages formels et langues naturelles

- Dans un langage formel, les symboles sont concaténés pour former les mots du langage d'une façon potentiellement infinie. De même, dans une langue naturelle, les mots sont concaténés pour former des énoncés d'une manière aussi potentiellement infinie.

- Dans un langage formel, les mots sont des objets à double face forme/sens. De même, dans une langue naturelle, les énoncés sont des objets à double face son/sens.

- Les travaux de Chomsky sur la formalisation des grammaires pour les langues naturelles (1956) ont joué un rôle important dans le développement de la théorie des langages formels.

1.4 - Langages formels et langues naturelles

- L'**ambiguïté** est rejetée des langages formels tandis qu'elle a une place importante dans les langues naturelles : un énoncé est ambigu s'il y a plusieurs structures linguistiques alternatives qui peuvent lui être associées. Selon la source de cette multiplicité, on distingue l'ambiguïté lexicale, phonologique, syntaxique ou sémantique.

- Les langages formels sont **figés** alors que les langues naturelles sont **évolutives**. En conséquence, la frontière entre les énoncés acceptables linguistiquement et ceux qui ne le sont pas est floue et mobile.

1.4 - Langages formels et langues naturelles : exercices

1. Chacune des phrases suivantes est ambiguë. Analyser la source de cette ambiguïté.

   a) *Jean voit l'avion qui est en train de se poser*

   b) *Jean regarde la femme sur le balcon.*

   c) *Le beau livre la porte.*

   d) *Je connais la belle ferme qui est sur la colline depuis longtemps.*

1.5 - Chaîne de traitement automatique d'une langue

- Le TAL est guidé par deux paradigmes :
  - ✓ Le **paradigme symbolique** vise à modéliser les connaissances linguistiques à l'aide de systèmes symboliques. Il amène à distinguer la notion de **compétence** de celle de **performance**.
  - ✓ Le **paradigme stochastique** vise à extraire l'information linguistique de corpus à l'aide de méthodes stochastiques. Une des techniques utilisées est l'**apprentissage statistique**.
- Le TAL est organisé selon deux sens: l'**analyse** qui va des énoncés à leur interprétation pragmatique et la **génération** qui mène des buts pragmatiques aux énoncés représentant leur réalisation linguistique.

1.5 - Chaîne de traitement automatique d'une langue

- Un exemple extrait de la thèse de François-Régis Chaumartin : indexation d'un texte par un moteur de recherche.

- Le texte de départ est un avis de consommateur : *"Je tenais à féliciter la caissiére Céline pour son accueil chaleureux et souriant du samedi 16 février malgré la foule incroyable ce jour la, elle a su faire abstraction de cela et garder le sourire et la bonne humeur. FELICITATIONS"*

1.5 - Chaîne de traitement automatique d'une langue

- Traitement des caractères (suppression des diacritiques et transformation des majuscules en minuscules) : *"je tenais à feliciter la caissiere celine pour son accueil chaleureux et souriant du samedi 16 fevrier malgre la foule incroyable ce jour la, elle a su faire abstraction de cela et garder le sourire et la bonne humeur. felicitations"*

- Découpage du texte en mots.

- Suppression des mots vides de sens : *"je tenais à feliciter la caissiere celine pour son accueil chaleureux et souriant du samedi 16 fevrier malgre la foule incroyable ce jour la, elle a su faire abstraction de cela et garder le sourire et la bonne humeur. felicitations"*

1.5 - Chaîne de traitement automatique d'une langue

- Suppression des affixes morphologiques pour ne conserver que la racine des mots : *"je tenais à feliciter la caissiere celine pour son accueil chaleureux et souriant du samedi 16 fevrier malgre la foule incroyable ce jour la, elle a su faire abstraction de cela et garder le sourire et la bonne humeur. felicitations"*

- Création d'un vecteur termes-fréquences : *{ abstract=1, accueil=1, bon=1, caiss=1, celin=1, chaleur=1, fair=1, felicit=2, fevri=1, foul=1, gard=1, humeur=1, incroy=1, jour=1, samed=1, souri=2, su=1, ten=1 }*

1.5 - Chaîne de traitement automatique d'une langue

- Représentation sémantique souhaitable du texte : *"Je tenais à féliciter la caissière Céline pour son accueil chaleureux et souriant du samedi 16 février malgré la foule incroyable ce jour la, elle a su faire abstraction de cela et garder le sourire et la bonne humeur. FELICITATIONS"*