Introduction au Traitement Automatique des Langues Naturelles

François Yvon

École Nationale Supérieure des télécommunications
Département Informatique et Réseaux

19 février 2006

Quelques questions à élucider

- À quoi sert le TALN ?
  - modélisation de la compétence langagière
  - reproduction de la performance langagière, avec de nombreux enjeux industriels
- Quelles sont les difficultés du TALN ?
  - production/perception d'un signal *linéaire* à valeurs *continues* ; de manière sous-jacente (interne ?) : *symbolique*, *structuré*.
  - l'*ambiguïté* (apparente) des unités linguistiques
  - l'*implicite* dans les énoncés naturels
- Quelles techniques pour le TALN ? Les *grammaires formelles* et les *grammaires d'unification* ; les outils de Représentation des connaissances ; les logiques ; l'apprentissage automatique

Un exemple introductif

(1) *Le président des antialcooliques mangeait une pomme avec un couteau*

Les niveaux de traitement :

- segmentation du texte en unités lexicales ;
- identification des composants lexicaux, de leurs propriétés : le traitement **lexical** ;
- identification des constituants de plus haut niveau, et des relations (de dominance) qu'ils entretiennent : le traitement **syntaxique** ;
- construction de la représentation du sens (i.e. des assertions que contient l'énoncé) : le traitement **sémantique** ;
- identification de la fonction de l'énoncé dans le contexte particulier de l'interaction : le traitement **pragmatique** ;

Attention à l'illusion du pipe-line !

Difficultés de la Segmentation / Normalisation

- Les écritures sans segmentation (chinois, thaï...)
- S'accomoder des ambiguités typographiques :
  - . : dans *etc.*, dans *20.3*, dans *enst.com*, dans *...*, dans *TF.1*...
  - ' : dans *jusqu'à*, dans *aujourd'hui*, dans *3'4*, dans *Sotheby's* ou *Floc'h* ...
  - - : dans *Jean-Michel*, dans *donne-t-il*, dans *06-04-62-26-16-23*, dans *1914-1918*, dans *-1.2 %*...
  - sans parler de l'espace lui-même
- Détecter et normaliser les variantes typographiques : *France-Inter* *France-inter* et *France Inter* ; *États-Unis* et *Etats-unis* et *Etats-Unis*...
- "Reconnaître" les chiffres, dates, durées, nombres, montants, numéros (de téléphone, de carte bleue), les scores...
- "Faire avec" les mots inconnus, les emprunts, les coquilles...

Le niveau lexical

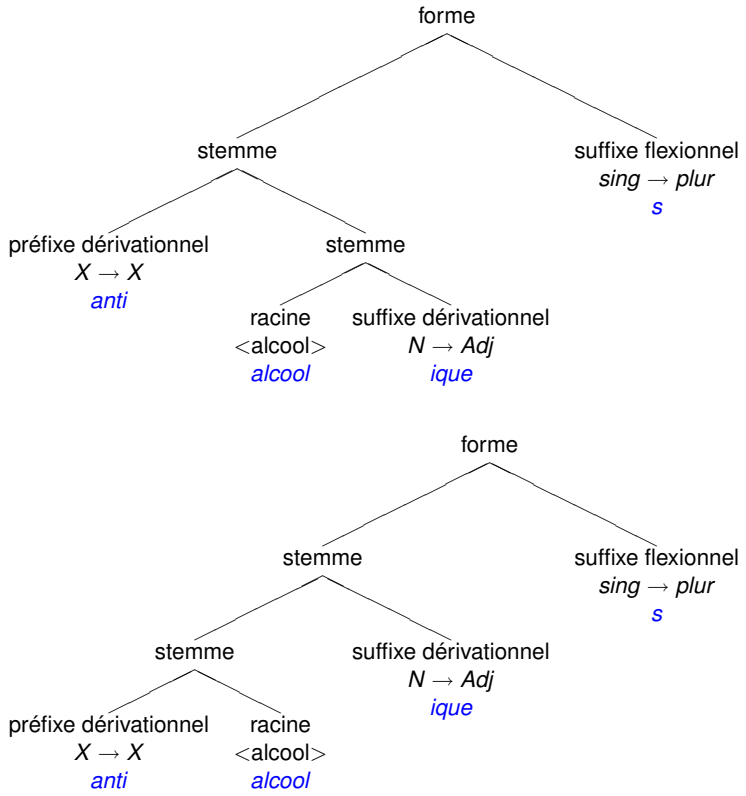
- **But** : identifier les éléments lexicaux, leur structure et leurs caractéristiques ; regrouper les formes d'une même famille.
- **Moyen** : accès lexical direct, analyse morphologique (i.e. décomposition en *morphèmes*, à partir desquels les propriétés d'une forme sont calculées).
- **Outils** : un lexique, une description des morphèmes et des procédures de décomposition/recomposition associées.
- **Difficultés** : taille du lexique, vitesse d'accès et d'analyse, représentation du lexique, traitement des mots composés.
- **Résultat** : une représentation linéaire ou arborescente du mot, ses caractéristiques morpho-syntaxiques, une représentation de sa signification, un représentant de sa famille.

Le traitement lexical : résultat

- *le* - det. masc. sing., /lə/ ; pron. pers. masc. sing., /lə/
- *président* - vrb 3pers. plur. prés. ind./ subjonctif [présid+ent], <présider(X), présider(X,Y)>, /pʁezid+ət/ ; nom masc. sing., ←présider : action de X, <president(X)>, /pʁezidɑ̃/
- *des* - det. masc./fem. plur., /dɛ+z/ ; prep. contr. *de les*. ...
- *antialcooliques* - adj. masc./fem. plur. [anti+alcool+ique+s], ← alcoolique : s'opposer à X, antialcoolique(X), /ɑ̃tialkɔlikə+z/ ; nom. masc. sing. [anti+alcool+ique+s], ← antialcoolique (adj) : être X, antialcoolique(X), /ɑ̃tialkɔlikə+z/
- *mangeait* - vrb (1,3) pers. sing. imp. ind., [mang+e+ait], <manger(X),manger(X,Y)>, /mɑ̃ʒɛ+t/
- *pomme* - nom fem. sing., [pomme], <pomme(X),fruit(X),golden(X)...>, /pɔmə/
- ...

Décomposition arborescente et linéaire

antialcooliques



anti+alcool+ique+s

La syntaxe du pauvre : étiquetage et *chunking*

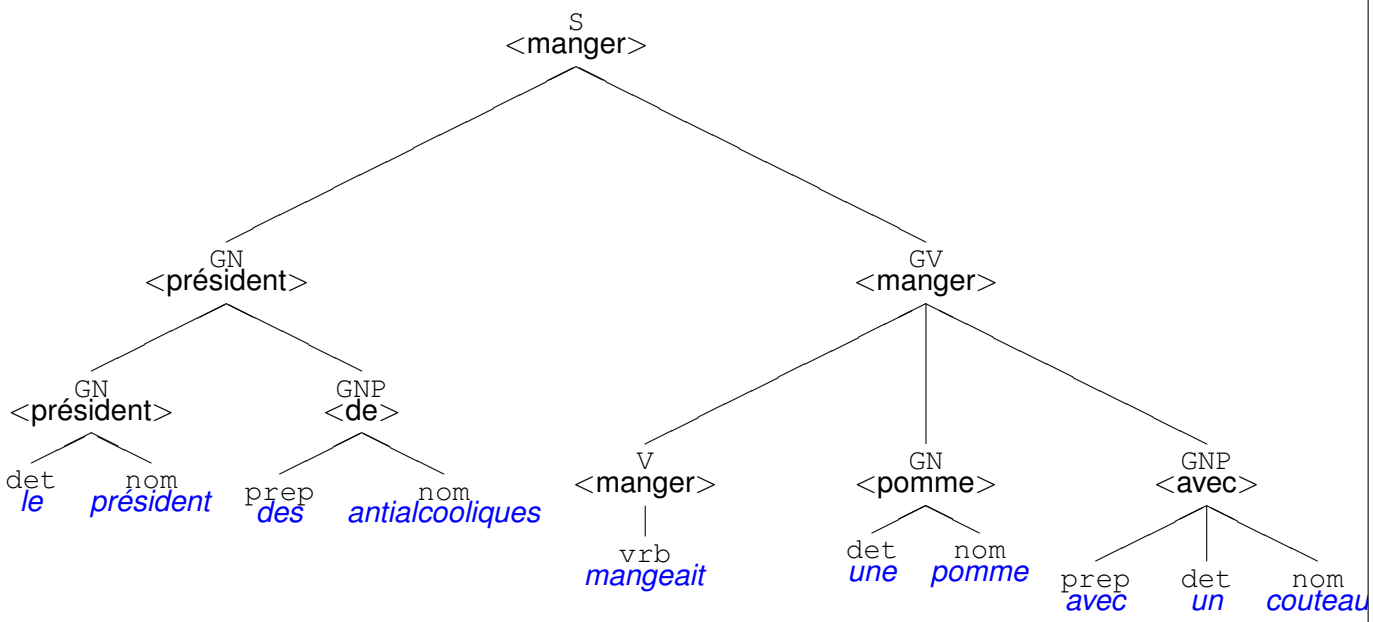
- **But** : *désambiguïser* les étiquettes morpho-syntaxiques ambiguës (*POS tagging ou étiquettage morpho-syntaxique*) ; *identifier les frontières* de groupes (mais pas leur structure interne ni les relations de dépendances) : *chunking*
- **Moyen** : règles (patrons) de désambiguïsation ; modèles statistiques (Modèles de Markov cachés, Champs conditionnels aléatoires) ; apprentissage de règles de désambiguïsation
- **Outils** : règles, patrons, corpus annotés manuellement (pour l'apprentissage)
- **Difficultés** : les mots inconnus ; combinaison de connaissances symboliques et de règles de décision numériques
- **Résultat** : l'identification des étiquettes morphosyntaxiques (tagging) ; les frontières de groupe (chunking).

[Le/Admp président/Vpi3p] [des/Prep antialcooliques/Ncmp] [mange/Vpi3s] [une/Aifs pomme/Ncfs]...

Le niveau syntaxique

- **But** : identifier les composants syntaxiques (syntagmes), leur fonction, et les relations qu'ils entretiennent entre eux.
- **Moyen** : analyse syntaxique, qui fournit une représentation arborescente des composants de l'énoncé.
- **Outils** : un analyseur syntaxique, c'est-à-dire un formalisme de description des règles syntaxiques, des règles valides pour un (sous)-langage donné, et un système d'analyse (un parseur) capable d'exploiter ces règles.
- **Difficultés** : compromis entre richesse de description, vitesse d'analyse, et prolifération des ambiguïtés, complexité des phénomènes à décrire, robustesse aux entrées "bruitées" (coquilles, casse...).
- **Résultat** : un (ou des) arbres syntaxiques représentant la phrase.

Le traitement syntaxique : résultat



L'ambiguïté lexicale

Un des principaux problèmes de l'analyse syntaxique est l'ambiguïté.
Ambiguïté lexicale :

- *souris* : formes verbales de *sourir*, nom féminin singulier et pluriel ;
- *petit* : adjectif ou nom masculin singulier ;
- *la* : déterminant ou pronom personnel féminin singulier, nom masculin ;
- *mousse* : formes verbales de *mousser*, nom masculin, nom féminin ;

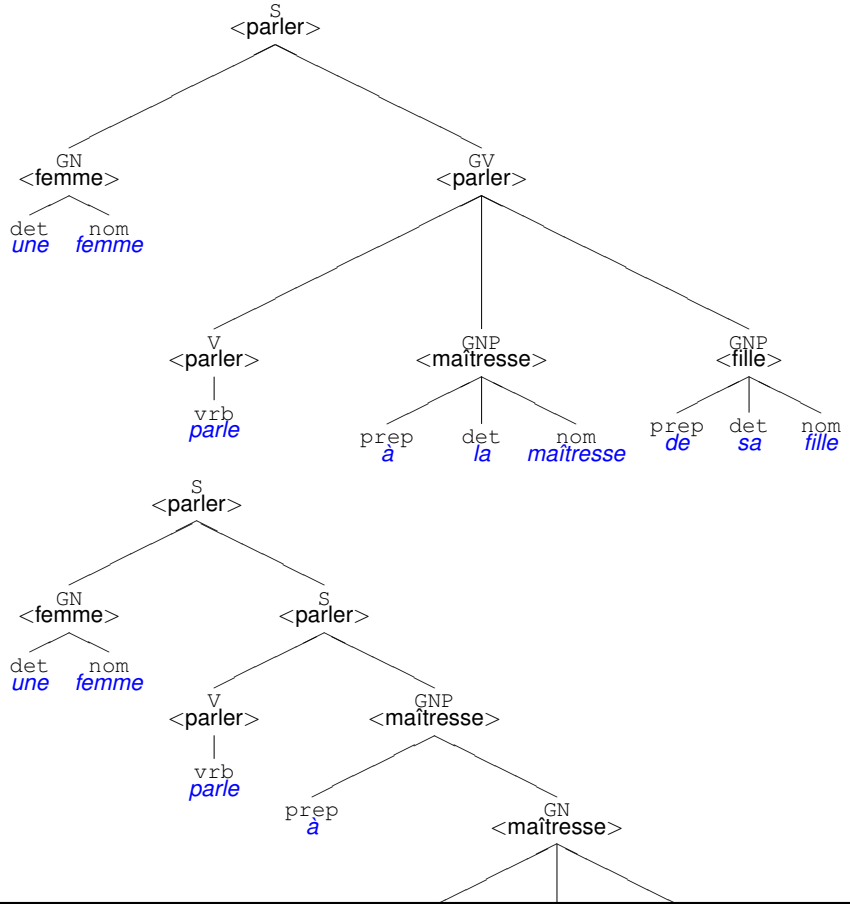
Plus la description lexicale est précise, plus l'ambiguïté est grande : *monter* (*monter un escalier*, *monter un cheval*, *monter une pièce*, ...).
Cette ambiguïté n'est pas seulement statique, mais également *dynamique* : les phénomènes syntaxiques de *translation* rendent ambigus adjectifs et participes passés (emploi nominal) : *ces affreux se sont enfuis*

L'ambiguïté syntaxique

- *La petite brise la glace* ;
- *La troupe monte Molière* vs *Le jockey monte Belino* ;
- *Elle mange une pomme avec les doigts* vs *Elle mange une pomme avec la peau* ;
- *Elle mange une glace à la fraise* vs *Elle mange une glace à la plage* ;
- *C'est la fille du cousin qui boit* ;
- *Il a parlé de déjeuner avec Paul* ;

La désambiguïsation est possible au niveau sémantique ou pragmatique ; chaque raffinement de la grammaire accroit l'ambiguïté.

L'ambiguïté syntaxique : le rattachement prépositionnel



Le niveau sémantique

- **But** : résoudre les problèmes de *référence* ; obtenir une *représentation conceptuelle* de l'énoncé dans un langage formel (formules de la logique du premier ordre, graphes conceptuels) ; *articuler* cette représentation conceptuelle avec le monde « physique » de la scène ;
- **Moyen** : calcul sémantique couplé à l'analyse syntaxique ou traduction ex-post de la représentation arborée dans un langage formel
- **Outils** : une description sémantique au niveau lexical (relations de synonymie, méronymie,hyper/hyponymie, etc), des règles de composition, des outils de représentation du monde physique ;
- **Difficultés** : explicitation partielle de l'implicite (problèmes de co-référence) ; ambiguïtés sémantiques (portée des quantifieurs) ; taille et précision de la connaissance nécessaire ; choix du formalisme de représentation (temporalité, croyances, etc).
- **Résultat** : un ensemble de représentations formelles de la scène dans lesquelles les objets et les relations qu'ils entretiennent sont identifiés ;

Le traitement sémantique : résultat

L'arbre syntaxique permet directement d'extraire les propositions (1) à (5), dont on peut déduire, compte-tenu d'une représentation du sens commun, (6), (7), (8) et (9) :

- $\exists X, president(X)$: il existe une entité X qui est président (et dont le référent est déja connu) ;
- $\exists Y, pomme(Y)$: il existe une entité Y qui est une pomme ;
- $\exists Z, couteau(Z)$: il existe une entité Z qui est un couteau ;
- $manger(X, Y)$: cette entité X mange Y ;
- $moyen(manger(X, Y), Z)$: l'opération de manger s'effectue au moyen de Z ;
- $president(X) \Rightarrow humain(X) \Rightarrow ...$;
- $(pomme(Y) \wedge manger(X, Y)) \Rightarrow aliment(Y)$
- $(pomme(Y) \wedge manger(X, Y)) \Rightarrow (golden(X)|granny(X)|...)$;
- $manger(X, Y) \Rightarrow manger(X), est\_ingere(Y)$;

Chez l'humain, ces déduction se font de manière inconsciente et quasi-réflexe.

Le niveau pragmatique

- **But** : achever la *désambiguïsation* de l'énoncé en prenant en compte la dynamique de l'interaction (ou narration) en y intégrant ce qui est implicite ; comprendre la *fonction argumentative* de l'énoncé dans le contexte plus général de l'interaction (ou de la narration) : quelle information nouvelle apporte-t-il, au sujet de quoi dit-il quelque chose, sous quel mode...
- **Moyen** : une théorie des activités humaines ; une théorie des interactions langagières (la pertinence, les conditions de félicité) ; une théorie des structures discursives...
- **Outils** : représentation des actions humaines (scripts), « grammaire » des interactions, logique
- **Difficultés** : taille de la connaissance à représenter, spécification de la « grammaire » des interactions
- **Résultat** : une représentation formelle contextualisée de l'énoncé, une connaissance de sa fonction argumentative, des connaissances nouvelles...

Applications du TALN : le traitement documentaire

Les applications les plus directes du TALN sont celles qui visent à faciliter le traitement par l'humain des immenses ressources disponibles en langage naturel, comme par exemple :

- La traduction automatique (ou l'aide à la traduction automatique) (voir `http://www.systransoft.com/`) ;
- La recherche de documents « intéressants » dans des bases documentaires ;
- Le classement ou l'indexation automatique de documents ; (eg. le SpamBuster).
- La lecture automatisée de documents, par exemple pour les stocker dans des structures formelles de données, ou pour en extraire des résumés ; (voir eg. `http://swesum.nada.kth.se/index-eng.html`)
- L'analyse d'un corpus de documents relatifs à un thème donné (histoire, stylométrie, veille technologique, etc).

Applications du TALN : la production de documents

Le TALN trouve également des applications directes dans le domaine de l'aide à la production de documents, telles que :

- les claviers « auto-correcteurs » (par exemple pour les handicapés) ;
- les correcteurs d'orthographe ou de syntaxe (voir le "Réaccentueur")
- les correcteurs « stylistiques », ou les aides intelligentes à la rédaction (thésaurus, etc) ;
- la génération automatique de documents à partir de spécifications formelles (par exemple les documentations techniques) ;
- la reconnaissance optique de caractères ;
- l'apprentissage assisté par ordinateur des langues naturelles ;

Applications du TALN : les interfaces naturelles

Le TALN trouve des applications directes dans de nombreux systèmes d'interfaces naturelles :

- interrogation en langage naturel de bases de données (traduction langage naturel $\leftrightarrow$ SQL)
- synthèse de la parole (désambiguïsation morpho-syntaxique, calcul de la prosodie) (voir eg. `http://www.elanspeech.com/demos/demos.html`) ;
- reconnaissance de la parole (filtrage *a posteriori* des multiples phrases reconnues durant l'étape de traitement du signal) ;
- interfaces vocales (reconnaissance, synthèse, génération de dialogue, gestion du dialogue, accès aux bases de connaissance, etc) ;

Tendances en TALN : l'utilisation de corpus

- Limites des approches à base de règles :
  - les experts sont rares (et chers)
  - les connaissances ne sont pas toujours transportables
  - certaines connaissances sont difficile à exprimer
- Apport des outils d'apprentissage :
  - modèles (probilistes) naturels de l'ambiguïté
  - indépendance (au moins partielle) par rapport à la langue ;
  - plasticité plus grande (l'adaptation à un nouveau domaine est facilitée)
- Réalisations :
  - étiqueteurs morpho-syntaxiques, analyseurs stochastiques
  - système de traduction automatique
  - recherche et extraction d'information