

Linguistique computationnelle

Séance 1 - Mardi 26 avril 2005

Philippe Blache

LPL, CNRS-Univ. de Provence

pb@lpl.univ-aix.fr

Jean-Philippe Prost

Macquarie University, Sydney

et LPL, CNRS-Univ. de Provence

jpprost@ics.mq.edu.au

Linguistique computationnelle

L'étude de la linguistique au moyen d'outils
et de techniques informatiques.

Objectifs

- comprendre les bases linguistiques du domaine
- décrire l'architecture générale d'un système de TALN
- comprendre, décrire et discuter différentes techniques (symboliques) mises en œuvre dans certains systèmes de TALN
- comprendre dans quelle mesure ces techniques sont en rapport avec d'autres domaines de l'informatique (théorique)

Plan général

- Introduction
- Les grammaires syntagmatiques
- les grammaires logiques
- Grammaires syntagmatiques de haut niveau
- Les contraintes

Plan cours 1

- Introduction
- Applications
- Différents domaines
- Techniques existantes
- Les étapes de l'analyse symbolique
 - Morphologie
 - Annotation syntaxique

Plan cours 1

- **Introduction**
- Applications
- Différents domaines
- Techniques existantes
- Les étapes de l'analyse symbolique

Références et ressources

- Jurafsky & Martin, 2000, *Speech and Language Processing*, Prentice-Hall
- Copestake, 2004, *Natural Language Processing* (handouts).
<http://www.cl.cam.ac.uk/users/aac/>
- Linguist List, Corpora
- ESSLLI
- ACL, EACL, CoLing, ATALA
- Natural Language Toolkit (NLTK)

Plan cours 1

- Introduction
- **Applications**
- Différents domaines
- Techniques existantes
- Les étapes de l'analyse symbolique

Applications

- Correction orthographique et grammaticale
Spelling and grammar checking
- Communication alternative
- Recherche et extraction d'information
Information Retrieval/Extraction
 - Question-réponse
Question Answering
- Classification de documents
- Reconnaissance de la parole
Speech recognition

Applications

- Systèmes de dialogue
- Synthèse vocale
 - Text-To-Speech (TTS)
- Génération automatique
Natural Language Generation
- Interface BDD
- Résumé automatique, paraphrase
Summarisation
- Traduction automatique
Machine Translation

Plan cours 1

- Introduction
- Applications
- **Différents domaines**
- Techniques existantes
- Les étapes de l'analyse symbolique

Domaines d'analyse

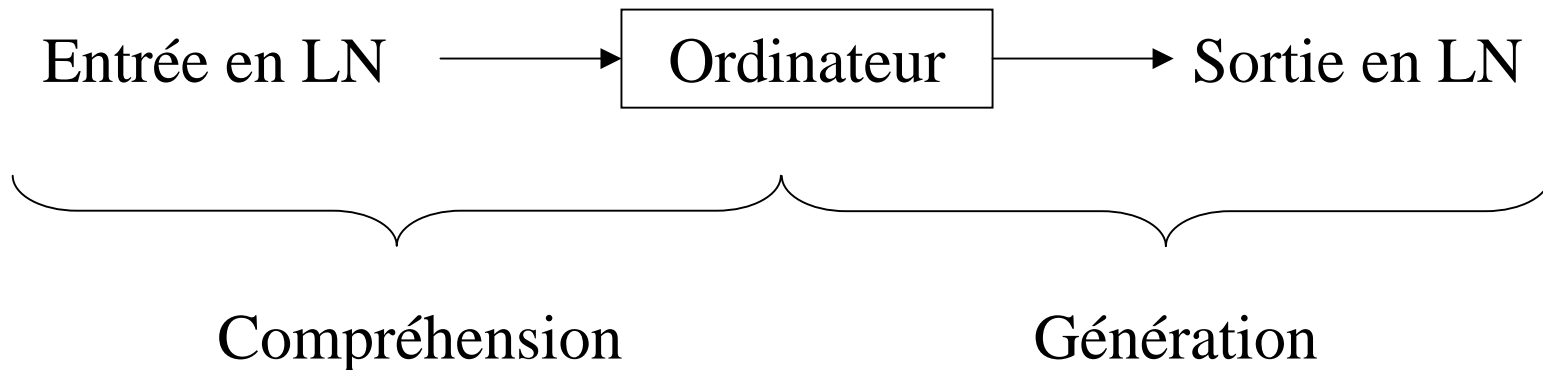
- Phonétique et phonologie
- Morphologie
- Syntaxe
- Sémantique
- Pragmatique

Plan cours 1

- Introduction
- Applications
- Différents domaines
- **Techniques existantes**
- Les étapes de l'analyse symbolique

TALN

NLP – Natural Language Processing



Architecture générique

1. input pre-processing
2. analyse morphologique
3. Annotation syntaxique
Part Of Speech (POS) tagging
4. Analyse syntaxique et sémantique (compositionnelle)
Parsing
5. Désambiguisation
6. Résolution de contexte (anaphores)
7. Planification de texte
8. Réalisation de surface
9. Post-processing (TTS, formatage)

Stochastique vs. symbolique

- Approches symboliques
 - à base de connaissance
- Approches stochastiques (as. numériques)
 - statistiques
 - probabilités
 - apprentissage
- Approches mixtes

Plan cours 1

- Introduction
- Applications
- Différents domaines
- Techniques existantes
- **Les étapes de l'analyse symbolique**
 - **Morphologie**
 - Annotation syntaxique

Morphologie

- structure des mots
- Morphème : unité d'information
 - radical
stem
 - affixe
 - préfixe, suffixe, infixe et circumfixe
- Inflexion et dérivation
- Lexique

Kurde

(Kurdistan : Turquie, Iran et Iraq)

• aaqil	SAGE	• aaqilii	PRÉVOYANCE
• diz	VOLEUR	• dizii	VOL
• draiž	LONG	• draižii	LONGUEUR
• zaanaa	SAGE	• zaanaaii	ÉRUDITION
• garm	CHAUD	• garmii	CHALEUR

1. Quel est le signifié du suffixe –ii ?
2. Si raasii signifie VÉRITÉ, quel est le signifié attendu de raas ?
3. Il y a deux formes qui sont glosées comme SAGE. Quelle différence de sens y a-t-il entre les deux?

(source : Jean-Yves Morin)

Kanuri

(Nigéria)

• <i>gana</i>	PETIT	• <i>'n'ngana</i>	PETITESSE
• <i>kura</i>	GROS	• <i>n'mkura</i>	GROSSEUR
• <i>kurugu</i>	LONG	• <i>n'mkurugu</i>	LONGUEUR
• <i>karite</i>	EXCELLENT	• <i>n'mkarite</i>	EXCELLENCE
• <i>dibi</i>	MÉCHANT	• <i>n'mdibi</i>	MÉCHANCETÉ

1. Donner la liste de tous les morphes que vous pouvez identifier dans ce corpus.
2. Si *n'mN'la* signifie BONTÉ, quel est le signifiant attendu pour BON ?

(source : Jean-Yves Morin)

Ganda

(Ouganda)

• <i>omukazi</i>	FEMME	• <i>'abakazi</i>	FEMMES
• <i>omusawa</i>	MÉDECIN	• <i>abasawa</i>	MÉDECINS
• <i>omusika</i>	HÉRITIER	• <i>abasika</i>	HÉRITIERS
• <i>omuwala</i>	FILLE	• <i>abawala</i>	FILLES
• <i>omulenzi</i>	GARÇON	• <i>abalenzi</i>	GARÇONS

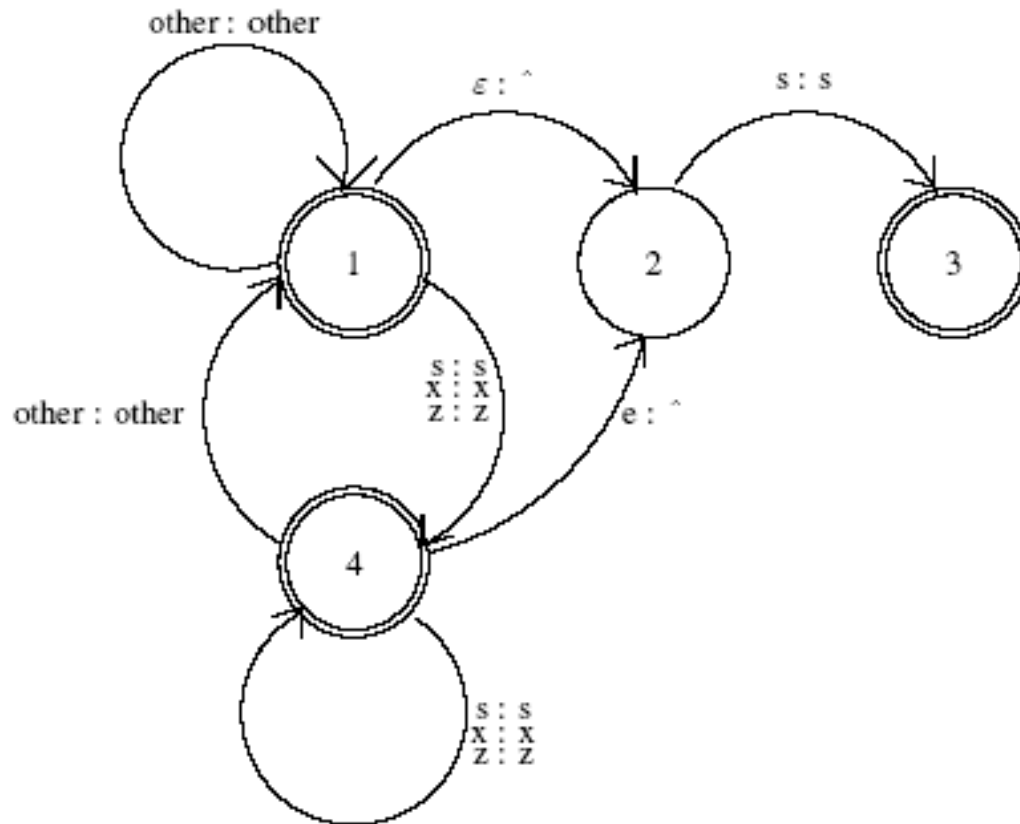
1. Donner la liste de tous les morphes que vous pouvez identifier dans ce corpus.

(source : Jean-Yves Morin)

Analyse morphologique

- Transducateur d'états finis
Finite-State Transducer (FST)
- Morphologie à deux niveaux :
 - niveau lexical
CHAT +N +PL
 - niveau de surface
CHATS

Exemple de FST



Aller plus loin

[J&M00] ch. 2, 3

Plan cours 1

- Introduction
- Applications
- Différents domaines
- Techniques existantes
- **Les étapes de l'analyse symbolique**
 - Morphologie
 - **Annotation syntaxique**

Part-Of-Speech (POS) Tagging

- Classification et annotation des mots d'un corpus

Le_DET_MASC_SG **livre**_NOM_MASC_SG **rouge**_ADJ_MASC_SG

- Approche stochastique
 - prédiction par N-gram
 - apprentissage sur corpus annoté

Plan général

- Introduction
- **Les grammaires syntagmatiques**
- les grammaires logiques
- Grammaires syntagmatiques de haut niveau
- Les contraintes