



Master 2 ATAL

Applications Multilingues

---

# Segmentation en mots du Japonais

---

*Étudiant :*  
Coraline MARIE

*Encadrant :*  
Florian BOUDIN

3 novembre 2014



UNIVERSITÉ DE NANTES

# Table des matières

Introduction	2
Première approche	3
Implémentation des trigrammes	4
Reconnaissance des alphabets	5
Conclusion	6

# Introduction

Le Japonais est une langue très particulière, dont l'écriture diffère totalement des langues latines. L'utilisation d'alphabets différents et l'absence de marqueurs explicites, tel que les espaces, la rend plus délicate à traiter, notamment dans le cadre de la Recherche d'Information.

Afin de faciliter le traitement de textes en langue Japonaise, il faut au préalable appliquer un processus de segmentation en mots, appelée *tokenisation*. Ce rapport présente donc différentes méthodes combinables permettant de segmenter en mots, des textes en japonais. Il y aura d'abord une description de l'implémentation de la méthode des Modèles de Markov de Constantine P. Papageorgiou. Il y aura également une présentation de l'amélioration de cette méthode par l'implémentation de ngrammes, avant de finir par l'explication d'une méthode reposant sur l'analyse des différents alphabets utilisés par la langue Japonaise.

# Première approche

# Implémentation des trigrammes

# Reconnaissance des alphabets

Comme il l'a été dit précédemment, le Japonais est composé de plusieurs alphabets :

## Conclusion