TD – Langue et Informatique - Segmentation et tokenization automatiques

P. Mercuriali – L2 – Centre Tesnière – CRIT – UMLP

Année 2025-2026

Objectifs pédagogiques
- Identifier les ambiguïtés de segmentation liées aux séparateurs
- Comprendre que la sémantique d'un séparateur varie selon l'usage et la langue
- Tester et comparer des outils de segmentation automatique
- Évaluer les erreurs typiques des outils de TAL
- Relier segmentation et enjeux en traduction, lexicographie, NLP

1 Rappels théoriques

Questions/discussion
- Qu'est-ce qui sépare les mots dans une phrase ?
- Un espace sépare-t-il toujours les mots ?
- Qu'est-ce qui sépare les phrases ?
- Un point sépare-t-il toujours les phrases ?
- Peut-on traduire correctement sans bien segmenter ?

La segmentation n'est pas une découpe mécanique.

2 Activité manuelle avec séparateurs ambigus

M. Jacques a déclaré : « La victoire 3-0 de l'équipe franco-allemande n'est pas due à un hasard : en effet, la S.N.C.F. a enregistré une hausse de 2,5 % en 2026 ». M. Jacques a refusé d'expliquer la corrélation. D'ailleurs, son frère, O'Hara, a couru le cent mètres en 12'3. Il commente : « I'm also competing for the 5,000 meters. » Ce projet, bien que complexe, est re-pensé aujourd'hui, après la découverte de ce qu'il a qualifié de 'super-tired-syndrome'.

Instructions
1. Repérer tous les séparateurs
2. Pour chaque séparateur, indiquer
   - sa fonction
   - son contexte (langue ?)
   - s'il sépare ou non une unité lexicale
3. Lister les problèmes possibles pour une machine

Remarque : noter l'inconsistance des marques de citation.

Tableau à compléter :

Séparateur	Communauté	Exemple	Fonction	Difficulté de segmentation
.	FR, EN	S.N.C.F.	sigle	faux découpage en mots ou en phrases
etc.	⋮	⋮	⋮	⋮

3 Comparaison d'outils en ligne

Choisissez deux outils parmi la liste suivante. Choisir comme langue "français" lorsque l'option est proposée.
- Regex `https://pierremercuriali.github.io/interactive/tokenizer.html` (tokenizer par règles et expressions régulières ("regex"))
- UDPipe `https://lindat.mff.cuni.cz/services/udpipe/`

— Treetagger `http://corpora.lancs.ac.uk/tree-tagger/`
— GATE (Tweet Tokenizer) `https://cloud.gate.ac.uk/shopfront/displayItem/french-tweet-tokenizer`
— Stanza de Stanford `https://stanza.stanford.edu/` (analyse de dépendances, mais avec tokenization)

Instructions

1. Copier-coller le texte
2. Observer
   — où l'outil segmente
   — où l'outil échoue à segmenter
3. Repérer, là où l'outil a échoué :
   — les découpages excessifs
   — le manque de découpage
   — les unités mal reconnues
   — les nombres, signes, noms propres mal traités

4 Comparaison des outils

1. Quel outil découpe le plus ?
2. Quel outil respecte le mieux
   — les nombres ?
   — les noms propres ?
   — les sigles ?
3. L'outil tient-il compte
   — du contexte sémantique ?
   — de la langue ?

5 Comparaisons interlinguistiques

EN vs FR
— Virgule décimale (FR 2,5 ; EN 2.5)
— Apostrophe (FR élision ; EN possession)
— Tiret (FR mots composés ; EN hyphen et dash)
Conclusion : une segmentation correcte dépend de la langue et du domaine