UNIVERSITE LIBRE DE BRUXELLES

Faculté de Lettres, Traduction et Communication

STIC-B545 – Traitement Automatique de Corpus - 202526

RAPPORT DE TRAVAIL TP 2

ALBAN Kamga Dze

Enseignant : DE WILDE Max

Assistant : Lebailly Denis

Année académique 2025-2026

1. Resumé

Ce rapport présente les analyses menées sur le sous-corpus CAMille (Le Soir) pour l'année **1955**. J'ai procédé à l'extraction de mots-clés (bigrammes), enrichi itérativement la liste de stopwords et généré un nuage de mots, extrait les entités nommées principales (personnes, organisations, lieux) avec spaCy, et effectué une analyse de sentiment sur 10 phrases sélectionnées. Les résultats et principales limites sont présentés ci-dessous.

2. Données et traitement

Le corpus CAMille contient 7986 fichiers textes (Le Soir). Pour 1955, [100] fichiers ont été extraits et concaténés en 1955.txt. Le prétraitement a consisté en : normalisation des caractères, suppression de la ponctuation et des chiffres, tokenisation, et application d'une liste de stopwords enrichie manuellement après inspection itérative. La version propre a été sauvegardée sous 1955_clean.txt.

3. Extraction de mots-clés et nuage de mot

J'ai pour cette étape utilisé l'algorithme YAKE (configuré pour extraire des bigrammes) et enrichi les stopwords à partir des mots très fréquents non informatifs (ex. « tél », « rue », « brux »). Le nuage de mots final (figure 1) met en évidence les termes [Agence], [Pays], [Président], [Belgique], révélateurs des thèmes dominants de l'année 1955.

(Figure 1 : 1955.png)

4. Reconnaissance d'entités nommées (NER)

La reconnaissance d'entité nommé a été appliqué directement sur le texte nettoyer. Les entités ont été agrégées et normalisées sommairement.

Dans le cadre de l'année 1955, voici quelques entités extraites :

- Personnes : [Majesté le roi des animaux], [Julien sorel], [Maurice Garcon]
- Organisations : [Conseil], [C.E.C.A], [C.E.D]
- Lieux : [Bruxelles], [Congo], [France]

Limites : erreurs OCR et variantes orthographiques, fragments de mots nécessitent un post-traitement

5. Analyse de sentiment (10 phrases)

Le tableau ci-dessous recapitule l'ensemble des dix phrase que l'on a pu sélectionner ;

Phrase sélectionné	Polarité	Subjectivité	Observation/interprétation
Sa Majesté le Roi des Animaux !	Neutral	Perfectly objective	Expression factuelle, sans émotion ni jugement.
En soulevant un peu le voile, on apprit simplement que l'artiste était marié, qu'il avait deux enfants et qu'un jour le malheur s'installa à son foyer, pour ne plus le quitter.	2%positive	18%	Phrase légèrement teintée de tristesse par la mention du "malheur", mais exprimée de manière surtout descriptive. Faible subjectivité car le narrateur rapporte des faits.
Le cadet de ses fils atteint d'un mal mystérieux mourut dans ses bras et le père infortuné demeura inconsolable.	21%négative	55%	Ton dramatique et affectif : vocabulaire émotionnel Charge négative et subjectivité moyenne.
Heureusement, il lui restait son jardin.	30%positive	60%	"heureusement" traduit une émotion de soulagement. La phrase exprime une opinion personnelle, d'où une subjectivité marquée.
Cette consolation lui fut rapidement enlevée, car l'on expropria son domaine et ce fut la ruine définitive.	Neutral	Perfectly objective	Description d'un événement tragique sans marque explicite d'émotion. L'objectivité est maintenue malgré tout.
« Vireloque » a pris sa place.	Neutral	Perfectly objective	Simple information factuelle, sans connotation émotionnelle ni jugement.
L'artiste, physiquement diminué, se confina alors dans une espèce de « sauvagerie à la Rousseau »	Neutral	Perfectly objective	Narration factuelle
Et puis, il y avait, au milieu de la table, un magnifique gâteau avec vingt bougies.	100% positive	100% subjective	Description valorisante et admirative ("magnifique"). Forte implication émotionnelle du narrateur.
On peut comprendre la déception des Anversois, battus sur penalty, à 3 minutes de la fin"	30%negative	18%subjective	L'expression "déception" traduit une émotion légère, mais le ton reste explicatif et peu subjectif.
Il fallut attendre celle-ci jusqu'en 1842, et ce fut une loi de compromis ;	Neutral	Perfectly objective	Enoncé purement informatif exprimant un fait historique.