Dossier TAL pour l’UE [MISH504T - Traitement Automatique des Langues](https://iris.univ-tlse2.fr/course/view.php?id=212)

2023-2024

---------------------------------------------------------------------

Nom du groupe (tel qu’attribué sur IRIS) :

Nom des étudiants :

Date :

**Table des matières**

[I. Évaluation du module Stanza 1](#__RefHeading___Toc1119_16753595961)

[I.1. Message qui a été attribué au groupe et son analyse Stanza 2](#__RefHeading___Toc1191_339405809)

[I.2. La segmentation en phrases 3](#__RefHeading___Toc1193_339405809)

[I.3. La segmentation en tokens 3](#__RefHeading___Toc1121_1675359596)

[I.4. L’étiquetage morpho-syntaxique 3](#__RefHeading___Toc1193_3005927150)

[a. Evaluation globale 3](#__RefHeading___Toc68500_2644616758)

[b. Etiquetage morpho-syntaxique des VERBES : 4](#__RefHeading___Toc68502_2644616758)

[I.5. La lemmatisation 4](#__RefHeading___Toc975_33940580911)

[I.6. L’analyse syntaxique en relations de dépendance 5](#__RefHeading___Toc1458_339405809)

[a. Evaluation globale 5](#__RefHeading___Toc68504_2644616758)

[b. Analyse syntaxique des relations Sujet : 5](#__RefHeading___Toc68506_2644616758)

[II. Bilan de l’évaluation 5](#__RefHeading___Toc1460_339405809)

[III. Extraction d’information dans le corpus Cocorep-velo.txt 6](#__RefHeading___Toc1462_339405809)

[III.1. Description des données globales 6](#__RefHeading___Toc1071_1675359596)

[III.2. Extraction d’information 6](#__RefHeading___Toc78411_2644616758)

[a. Mots utilisés pour désigner un vélo : 7](#__RefHeading___Toc68508_2644616758)

[b. Types de vélo mentionnés 7](#__RefHeading___Toc1657_339405809)

[c. Script conçu pour réaliser l’extraction 7](#__RefHeading___Toc1195_3005927150)

[d. Propositions d’amélioration 7](#__RefHeading___Toc1659_339405809)

Index des tableaux

[Tableau 1: Version verticalisée et corrigée du message <#https :://hgkjgkjg> 2](#Tableau!12|sequence)

[Tableau 2: xx 3](#Tableau!0|sequence)

[Tableau 3: xx 3](#Tableau!1|sequence)

[*Tableau 4: xx* 3](#Tableau!2|sequence)

[Tableau 5: xx 4](#Tableau!3|sequence)

[Tableau 6: xx 4](#Tableau!4|sequence)

[Tableau 7: xx 4](#Tableau!5|sequence)

[*Tableau 8: xx* 4](#Tableau!6|sequence)

[*Tableau 9: xx* 4](#Tableau!7|sequence)

[Tableau 10: xx 5](#Tableau!8|sequence)

[Tableau 11: Aperçu quantitatif du corpus 6](#Tableau!9|sequence)

[Tableau 12: mots utilisés pour désigner un vélo et leur fréquence 7](#Tableau!10|sequence)

[Tableau 13: Termes désignant les types de vélo mentionnés dans le corpus 7](#Tableau!11|sequence)

# Évaluation du module Stanza

Cette partie concerne l’évaluation de l’efficacité des modules de traitements automatique du langage proposés par Stanza.

***L’évaluation portera sur un seul texte*** *du corpus dont l’identifiant a été attribué à votre groupe (voir l’attribution dans le tableur disponible dans les fichiers déposés sur IRIS).*

*Pour chaque niveau d’analyse, l’évaluation consiste à*

1. *identifier les erreurs commises par l’analyseur*
2. *calculer les mesures de précision et, dans certains cas, les mesures de rappel (pour certains modules, seule la précision est calculable).*
3. *commenter les erreurs observées et les classer par catégories, en vous appuyant autant que possible sur des caractéristiques linguistiques telles que vues au cours de votre formation (parties du discours, relations syntaxiques, etc.) et énoncées en cours : ambiguïté, variation, évolutivité et implicite du langage naturel. Si pour certains niveaux vous n’observez aucune erreur, l’indiquer explicitement.*
4. *conclure en indiquant quel est selon vous le meilleur outil (en fonction des scores de précision et rappel ET de la « gravité » des erreurs observées).*

## Message qui a été attribué au groupe et son analyse Stanza

Identifiant du message : <#https :://hgkjgkjg>

Texte du message : J ai achete le nouvel album de paul k. qu' est sorti hier. Pas fou fou par rapport a berlin calling."

*Insérez ici un tableau correspondant à la version verticalisée du message dans lequel toutes cellules présentant une erreur d’analyse sera colorée en rouge avec, dans la dernière colonne, votre proposition de correction (merci de suivre le modèle donné en exemple ci-dessous)*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **num** | **token** | **lemme** | **POS** | **Num (lemme) gouverneur** | **Relation syntaxique** |  | **Erreur(s) détectée(s)** | | | | | | |
| **phr** | **tok** | **lem** | | **POS** | **gouv** | **rel** |
| #PHR1 : J ai achete le nouvel album de paul k. | | | | | | |  | | | | | | |
| 0 | J | je | PRON | 2 (acheter) | nsubj |  |  |  |  | |  |  |  |
| 1 | ai | avoir | AUX | 2 (acheter) | aux:tense |  |  |  |  | |  |  |  |
| 2 | achete | acheter | VERB | 2 (acheter) | root |  |  |  |  | |  |  |  |
| 3 | le | le | DET | 5 (album) | det |  |  |  |  | |  |  |  |
| 4 | nouvel | nouveau | ADJ | 5 (album) | amod |  |  |  |  | |  |  |  |
| 5 | album | album | NOUN | 2 (acheter) | obj |  |  |  |  | |  |  |  |
| 6 | de | de | ADP | 7 (paul) | case |  |  |  |  | |  |  |  |
| 7 | paul | paul | X | 5 (album) | nmod |  |  |  |  | | PROPN |  |  |
| 8 | k. | k. | X | 7 (paul) | flat:foreign |  |  |  |  | | PROPN |  | nmod |
| #PHR2 : qu' est sorti hier. | | | | | | | Même phrase, pas de segmentation | | | | | | |
| 9 | qu' | que | PRON | 11 (sortir) | obj |  |  |  |  |  | |  | nsubj |
| 10 | est | être | AUX | 11 (sortir) | aux:pass |  |  |  |  |  | |  |  |
| 11 | sorti | sortir | VERB | 11 (sortir) | root |  |  |  |  |  | |  |  |
| 12 | hier | hier | ADV | 11 (sortir) | advmod |  |  |  |  |  | |  |  |
| 13 | . | . | PUNCT | 11 (sortir) | punct |  |  |  |  |  | |  |  |
| #PHR3 : Pas fou fou par rapport a berlin calling. | | | | | | |  | | | | | | |
| 14 | Pas | pas | ADV | 15 (fou) | advmod |  |  |  |  |  | |  |  |
| 15 | fou | fou | ADJ | 15 (fou) | root |  |  |  |  |  | |  |  |
| 16 | fou | fou | ADJ | 15 (fou) | amod |  |  |  |  |  | |  |  |
| 17 | par | par | ADP | 18 (rapport) | case |  |  |  |  |  | |  |  |
| 18 | rapport | rapport | NOUN | 15 (fou) | obl:mod |  |  |  |  |  | |  |  |
| 19 | a | à | ADP | 20 (berlin) | case |  |  |  |  |  | |  |  |
| 20 | berlin | berlin | X | 18 (rapport) | nmod |  |  |  |  | PROPN | |  |  |
| 21 | calling | calling | X | 20 (berlin) | flat:foreign |  |  |  |  | PROPN | |  | nmod |
| 22 | . | . | PUNCT | 15 (fou) | punct |  |  |  |  |  | |  |  |

Tableau 1: Version verticalisée et corrigée du message <#https :://hgkjgkjg>

## La segmentation en phrases

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Stanza | |  |
|  | Nb. de caractères... | ...de fin de phrase | ...de non fin de phrase | Total |
| Référence  (vous) | ...de fin de phrase | *nb. de caractères de fin de phrase selon vous ET Stanza* | *nb. de caractères de fin de phrase selon vous MAIS PAS Stanza* | *nb. de caractères de fin de phrase selon vous* |
| ...de non fin de phrase | *nb. de caractères de fin de phrase selon Stanza MAIS PAS vous* | *nb. de caractères de NON fin de phrase selon vous ET Stanza* | *nb. de caractères de NON fin de phrase selon vous* |
|  | *Total* | *nb. de caractères de fin de phrase selon Stanza* | *nb. de caractères de NON fin de phrase selon Stanza* | *nb. de caractères* |

Tableau 2: xx

Calcul de la précision et du rappel de la segmentation en phrase :

**P =**

**R =**

Commentaires sur les erreurs observées :

*Commentez les erreurs observées en indiquant le numéro de phrase indiqué dans la version verticalisée présentée dans le tableau 1. Si aucune erreur n’est observée, indiquez-le et passez à la suite.*

## La segmentation en tokens

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Stanza | |  |
|  | Nb. de caractères... | ...de fin de token | ...de non fin de token | Total |
| Référence  (vous) | ...de fin de token |  |  |  |
| ...de non fin de token |  |  |  |
|  | *Total* |  |  |  |

Tableau 3: xx

Calcul de la précision et du rappel de la segmentation en tokens:

**P =**

**R =**

Commentaires sur les erreurs observées :

*Commentez les erreurs observées en indiquant le numéro du token indiqué dans la version verticalisée présentée dans le tableau 1. Si aucune erreur n’est observée, indiquez-le et passez à la suite.*

## L’étiquetage morpho-syntaxique

### Evaluation globale

|  |  |  |  |
| --- | --- | --- | --- |
|  | nb. de mots | nb. de mots correctement étiquetés morphosyntaxiquement | Précision globale |
| Stanza |  |  |  |

*Tableau 4: xx*

*Commentez les erreurs observées en indiquant le numéro de token indiqué dans la version verticalisée présentée dans le tableau ci-dessus. Si aucune erreur n’est observée, indiquez-le et passez à la suite.*

Commentaires sur les erreurs observées :

### Etiquetage morpho-syntaxique des VERBES :

Pour cette évaluation, on considère que les VERBES sont les mots étiquetés VERB et AUX.

Liste des verbes selon vous : \_\_\_\_\_

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Stanza | |  |
|  | Nb. de mots ... | où pos = VERB|AUX | où pos ≠ VERB|AUX | Total |
| Référence  (vous) | où pos = VERB|AUX | *nb. de mots étiquetés VERBE selon vous ET Stanza* | *nb. de mots étiquetés VERBE selon vous MAIS PAS Stanza* | *nb. de mots étiquetés VERBE selon vous* |
| où pos ≠ VERB|AUX | *nb. de mots étiquetés VERBE selon Stanza MAIS PAS vous* | *nb. de mots avec une autre étiquette que VERBE selon vous ET Stanza* | *nb. de mots avec une autre étiquette que VERBE selon vous* |
|  | *Total* | *nb. de mots étiquetés VERBE selon Stanza* | *nb. de mots avec une autre étiquette que VERBE selon Stanza* | *nb. total de mots* |

Tableau 5: xx

Calcul de la précision et du rappel de l’étiquetage des verbes :

**P =**

**R =**

Commentaires sur les erreurs observées :

*Commentez les erreurs observées en indiquant le numéro du token indiqué dans la version verticalisée présentée dans le tableau 1. Si aucune erreur n’est observée, indiquez-le et passez à la suite.*

## La lemmatisation

|  |  |  |  |
| --- | --- | --- | --- |
|  | nb. de mots | nb. de mots  correctement lemmatisés | Précision |
| Stanza |  |  |  |

Tableau 6: xx

*Commentez les erreurs observées en indiquant le numéro de token indiqué dans la version verticalisée présentée dans le tableau ci-dessus. Si aucune erreur n’est observée, indiquez-le et passez à la suite.*

Commentaires sur les erreurs observées :

## L’analyse syntaxique en relations de dépendance

### Evaluation globale

|  |  |  |  |
| --- | --- | --- | --- |
|  | nb. de mots | nb. de mots avec le bon gouverneur | Précision |
| Stanza |  |  |  |

*Tableau 7: xx*

*Commentez les erreurs observées en indiquant le numéro de token indiqué dans la version verticalisée présentée dans le tableau ci-dessus. Si aucune erreur n’est observée, indiquez-le et passez à la suite.*

Commentaires sur les erreurs observées :

### Analyse syntaxique des relations Sujet :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Stanza | |  |
|  | Nb. de mots ... | où dep = nsubj | où dep ≠ nsubj | Total |
| Référence  (vous) | où dep = nsubj |  |  |  |
| où dep ≠ nsubj |  |  |  |
|  | *Total* |  |  |  |

Tableau 8: xx

Calcul de la précision et du rappel de l’étiquetage des relations sujets :

**P =**

**R =**

Commentaires sur les erreurs observées :

*Commentez les erreurs observées en indiquant le numéro du token indiqué dans la version verticalisée présentée dans le tableau 1. Si aucune erreur n’est observée, indiquez-le et passez à la suite.*

# Bilan de l’évaluation

|  |  |  |
| --- | --- | --- |
|  | Stanza | |
|  | Précision (P) | Rappel (R) |
| Segmentation en phrases |  |  |
| Segmentation en tokens |  |  |
| Étiquetage morphosyntaxique |  |  |
| Étiquetage des verbes |  |  |
| Lemmatisation |  |  |
| Analyse des relations de dépendance syntaxique |  |  |
| Analyse des relations sujet |  |  |

Tableau : Bilan de l'évaluation Stanza

Selon vos observations des erreurs et les scores obtenus, que pensez-vous des performances du module Stanza ?

# Extraction d’information dans le corpus Cocorep-velo.txt

## Description des données globales

Nombre de textes : \_\_\_\_\_

Cette partie consiste à donner un aperçu quantitatif des unités linguistiques contenues dans le corpus selon l’analyse proposée par Stanza.

|  |  |
| --- | --- |
|  | Analyse Stanza |
| Nombre de phrases |  |
| Nombre moyen de phrases par message |  |
| Nombre de mots |  |
| Nombre moyen de mots par phrase |  |
| Nombre de noms communs (pos = NOUN) |  |
| Lemmes des 5 noms communs les plus fréquents | *lemme (fréquence)*  *lemme (fréquence)*  *lemme (fréquence)*  *lemme (fréquence)*  *lemme (fréquence)* |
| Nombre d’adjectifs (pos = ADJ) |  |
| Lemmes des 5 adjectifs les plus fréquents | *lemme (fréquence)*  *lemme (fréquence)*  *lemme (fréquence)*  *lemme (fréquence)*  *lemme (fréquence)* |
| Nombre de verbes (pos = VERB|AUX) |  |
| Lemmes des 5 verbes les plus fréquents | *lemme (fréquence)*  *lemme (fréquence)*  *lemme (fréquence)*  *lemme (fréquence)*  *lemme (fréquence)* |
| Couple NOUN-VERB le plus fréquent tel que : | *NOUN ← nsubj ← VERB (fréquence)* |

Tableau 9: Aperçu quantitatif du corpus

## Extraction d’information

*Cette partie du travail consiste à mettre en place une fonction qui permet d’extraire le type de vélo à propos duquel les locuteurs demandent une aide et de compter ces types. Pour commencer, vous devez extraire les mots référant à un vélo (sous toutes ses formes y compris les mots mal orthographiés e.g. Veyllo ou les termes équivalents byciclette, etc.), puis extraire les noms ou adjectifs qui viennent compléter ces mots en suivant les relations de dépendance syntaxique. Exemples : «*vélo elliptique techness se400 », « velo d'appartement domyos vm 630 »...

*Vous êtes ensuite libres de peaufiner votre script pour avoir une extraction plus fine, des contextes plus complets, plus diversifiés, ou fournir la fréquence des couples extraits.*

*Cette extraction doit se faire sur l’ensemble du corpus.*

### Mots utilisés pour désigner un vélo :

|  |  |  |
| --- | --- | --- |
| **Mot** | **Catégorie motphosyntaxique (POS)** | **Fréquence** |
|  |  |  |
|  |  |  |
|  |  |  |

Tableau 10: mots utilisés pour désigner un vélo et leur fréquence

### Types de vélo mentionnés

|  |  |
| --- | --- |
| **Terme  *(ex de phrase ou extrait de phrase)*** | **Fréquence** |
|  |  |
|  |  |
|  |  |

Tableau 11: Termes désignant les types de vélo mentionnés dans le corpus

### Script conçu pour réaliser l’extraction

*Copier-coller ci-dessous le script ou la fonction python*

### Propositions d’amélioration