Vers une simplification automatique de la parole en français

Les enjeux de l'extraction des données d'apprentissage pour la simplification linguistique

Lucía Ormaechea
Doctorante, assistante
Université de Genève et
Université Grenoble-Alpes

Colloque International l'Association for French Language Studies (AFLS)
8 septembre de 2023, Université de Lille, Villeneuve-d'Ascq

Plan

1. Introduction

La **S**implification **A**utomatique de **T**extes **(SAT)** est un domaine du TAL qui vise à **réduire** automatiquement la **complexité linguistique** des textes, **sans pour autant perdre** les **informations** et la **signification originales** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	La Ligue nationale de hockey ou LNH est une ***association sportive*** professionnelle ~~***nord-américaine regroupant des franchises***~~ de hockey sur glace ***du Canada et des États-Unis***.
SIMPLIFIÉ	La Ligue nationale de hockey ou ***tout simplement*** LNH est une ***ligue sportive*** professionnelle de hockey sur glace ***d'Amérique du Nord***.

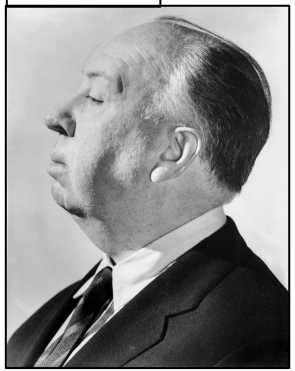
1. Introduction

La **S**implification **A**utomatique de **T**extes **(SAT)** est un domaine du TAL qui vise à **réduire** automatiquement la **complexité linguistique** des textes, **sans pour autant perdre** les **informations** et la **signification originales** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	La Ligue nationale de hockey ou LNH est une ***association sportive*** professionnelle ~~***nord-américaine regroupant des franchises***~~ de hockey sur glace ***du Canada et des États-Unis***.
SIMPLIFIÉ	La Ligue nationale de hockey ou ***tout simplement*** LNH est une ***ligue sportive*** professionnelle de hockey sur glace ***d'Amérique du Nord***.

SAT : une analogie avec la relation *forme-substance* :

Complexe



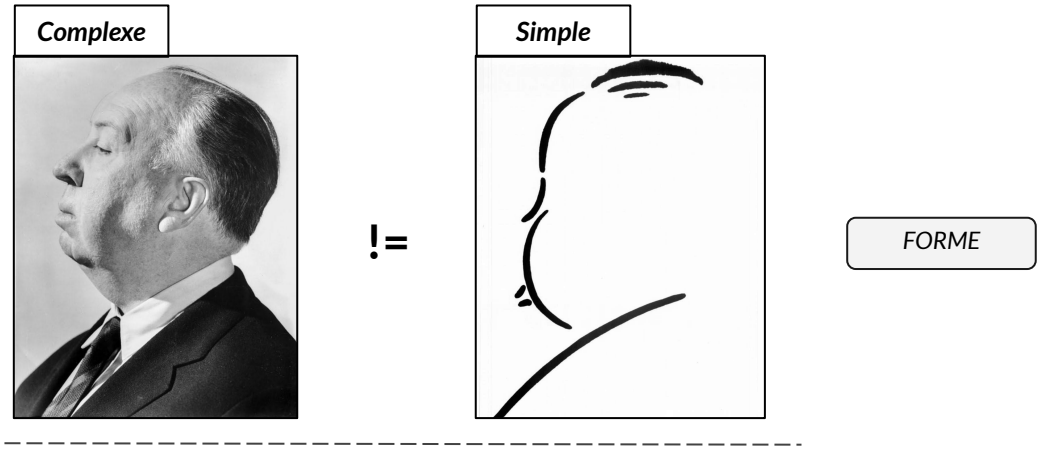
FORME

1. Introduction

La **S**implification **A**utomatique de **T**extes **(SAT)** est un domaine du TAL qui vise à **réduire** automatiquement la **complexité linguistique** des textes, **sans pour autant perdre** les **informations** et la **signification originales** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	La Ligue nationale de hockey ou LNH est une ***association sportive*** professionnelle ~~***nord-américaine regroupant des franchises***~~ de hockey sur glace ***du Canada et des États-Unis***.
SIMPLIFIÉ	La Ligue nationale de hockey ou ***tout simplement*** LNH est une ***ligue sportive*** professionnelle de hockey sur glace ***d'Amérique du Nord***.

SAT : une analogie avec la relation *forme-substance* :

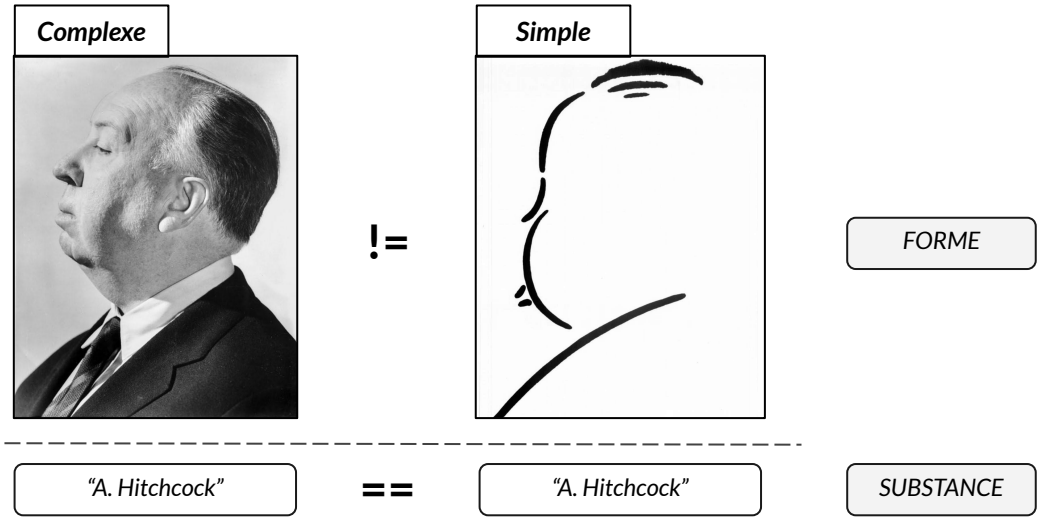


1. Introduction

La **S**implification **A**utomatique de **T**extes **(SAT)** est un domaine du TAL qui vise à **réduire** automatiquement la **complexité linguistique** des textes, **sans pour autant perdre** les **informations** et la **signification originales** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	La Ligue nationale de hockey ou LNH est une ***association sportive*** professionnelle ~~***nord-américaine regroupant des franchises***~~ de hockey sur glace ***du Canada et des États-Unis***.
SIMPLIFIÉ	La Ligue nationale de hockey ou ***tout simplement*** LNH est une ***ligue sportive*** professionnelle de hockey sur glace ***d'Amérique du Nord***.

SAT : une analogie avec la relation *forme-substance* :



1. Introduction

1.1. Qu'est-ce que la simplification automatique de textes ?

1.2. SAT en français : une tâche à faibles ressources

1.3. *Bridging the gap* : une nouvelle méthode pour extraire des paires *complexes-simples*

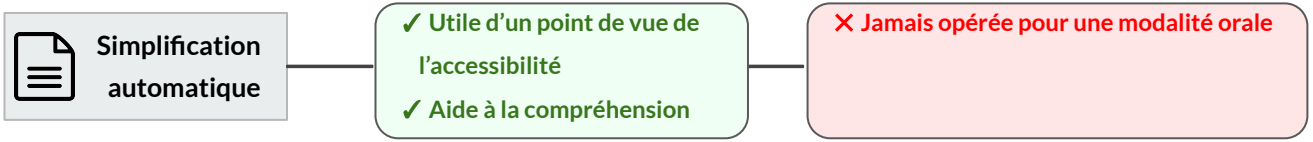
ORIGINAL	La Ligue nationale de hockey ou LNH est une ***association sportive*** professionnelle ~~***nord-américaine regroupant des franchises***~~ de hockey sur glace ***du Canada et des États-Unis***.
SIMPLIFIÉ	La Ligue nationale de hockey ou ***tout simplement*** LNH est une ***ligue sportive*** professionnelle de hockey sur glace ***d'Amérique du Nord***.

1. Introduction

1.1. Qu'est-ce que la simplification automatique de textes ?

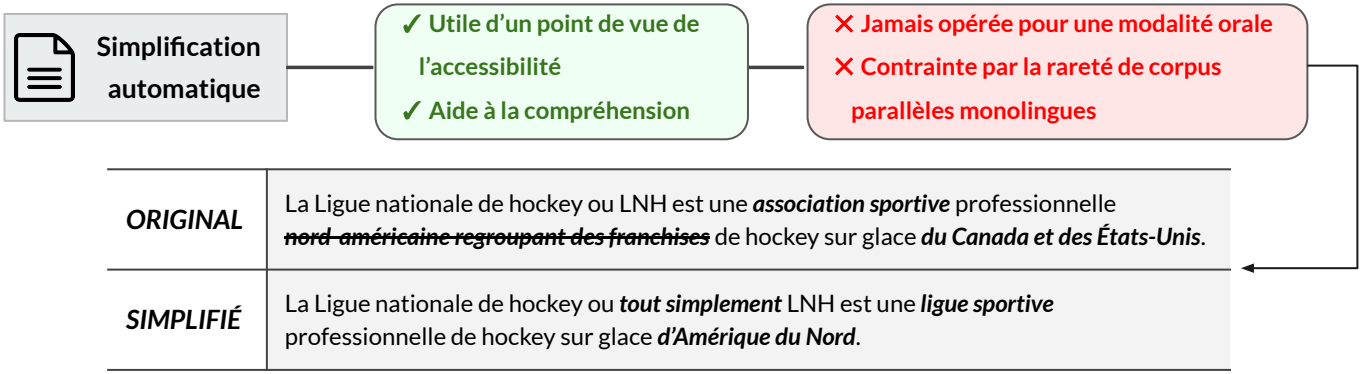
1.2. SAT en français : une tâche à faibles ressources

1.3. *Bridging the gap* : une nouvelle méthode pour extraire des paires *complexes-simples*



ORIGINAL	La Ligue nationale de hockey ou LNH est une ***association sportive*** professionnelle ~~***nord-américaine regroupant des franchises***~~ de hockey sur glace ***du Canada et des États-Unis***.
SIMPLIFIÉ	La Ligue nationale de hockey ou ***tout simplement*** LNH est une ***ligue sportive*** professionnelle de hockey sur glace ***d'Amérique du Nord***.

1. Introduction

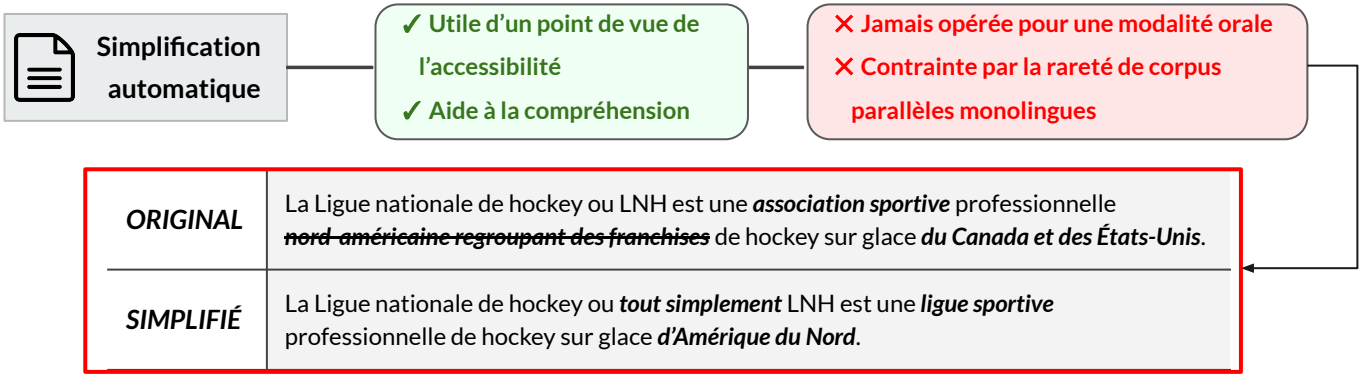


1. Introduction

1.1. Qu'est-ce que la simplification automatique de textes ?

1.2. SAT en français : une tâche à faibles ressources

1.3. *Bridging the gap* : une nouvelle méthode pour extraire des paires *complexes-simples*



Dans des langues moins dotées que l'anglais : *admettons... le français*

Le seul jeu de données monolingue disponible → **ALECTOR** [Gala *et al.*, 2020]:

- **Créé manuellement** par des éditeurs professionnels.
- Comprenant des textes scolaires, **simplifiés** pour un **public enfant**.

1. Introduction

Simplification automatique

✓ Utile d'un point de vue de l'accessibilité

✓ Aide à la compréhension

✕ Jamais opérée pour une modalité orale

✕ Contrainte par la rareté de corpus parallèles monolingues

ORIGINAL	La Ligue nationale de hockey ou LNH est une ***association sportive*** professionnelle ~~***nord-américaine regroupant des franchises***~~ de hockey sur glace ***du Canada et des États-Unis***.
SIMPLIFIÉ	La Ligue nationale de hockey ou ***tout simplement*** LNH est une ***ligue sportive*** professionnelle de hockey sur glace ***d'Amérique du Nord***.

Dans des langues moins dotées que l'anglais :

admettons... le français

Le seul jeu de données monolingue disponible → **ALECTOR** [Gala *et al.*, 2020]:

- **Créé manuellement** par des éditeurs professionnels.
- Comprenant des textes scolaires, **simplifiés** pour un **public enfant**.

Jeux de données créés manuellement

✓ Simplifications fiables

✕ Création coûteuse

✕ Très petite taille

1. Introduction

1.1. Qu'est-ce que la simplification automatique de textes ?

1.2. SAT en français : une tâche à faibles ressources

1.3. *Bridging the gap* : une nouvelle méthode pour extraire des paires *complexes-simples*

Simplification automatique

✓ Utile d'un point de vue de l'accessibilité
✓ Aide à la compréhension

✕ Jamais opérée pour une modalité orale
✕ Contrainte par la rareté de corpus parallèles monolingues

ORIGINAL	La Ligue nationale de hockey ou LNH est une ***association sportive*** professionnelle ~~***nord-américaine regroupant des franchises***~~ de hockey sur glace ***du Canada et des États-Unis***.
SIMPLIFIÉ	La Ligue nationale de hockey ou ***tout simplement*** LNH est une ***ligue sportive*** professionnelle de hockey sur glace ***d'Amérique du Nord***.

Dans des langues moins dotées que l'anglais :

admettons... le français

Le seul jeu de données monolingue disponible → **ALECTOR** [Gala *et al.*, 2020]:

- **Créé manuellement** par des éditeurs professionnels.
- Comprenant des textes scolaires, **simplifiés** pour un **public enfant**.

Jeux de données créés manuellement

✓ Simplifications fiables

✕ Création coûteuse

✕ Très petite taille

Et si l'on explorait des méthodes automatiques d'extraction de données ?

1. Introduction

1.1. Qu'est-ce que la simplification automatique de textes ?

1.2. SAT en français : une tâche à faibles ressources

1.3. *Bridging the gap* : une nouvelle méthode pour extraire des paires *complexes-simples*

PRINCIPALE CONTRIBUTION	• Présentation d'une **nouvelle méthode** pour **extraire** des paires de **phrases *complexes-simples*** à partir des **corpus comparables**. • **Conçue** spécifiquement pour la **tâche en aval**.

"Création d'un corpus parallèle en français de paires de phrases complexes-simples"

1. Introduction

PRINCIPALE CONTRIBUTION	• Présentation d'une **nouvelle méthode** pour **extraire** des paires de **phrases *complexes-simples*** à partir des **corpus comparables**. • **Conçue** spécifiquement pour la **tâche en aval**.

"Création d'un corpus parallèle en français de paires de phrases complexes-simples"

Approches automatiques pour la compilation des données en SAT :

- Basées sur des **corpus comparables** ayant des **registres distincts** :
  - **Wikipédia** → Édition française.
  - **Vikidia** → Version adaptée et simplifiée de la première.

1. Introduction

PRINCIPALE CONTRIBUTION	• Présentation d'une **nouvelle méthode** pour **extraire** des paires de **phrases *complexes-simples*** à partir des **corpus comparables**. • **Conçue** spécifiquement pour la **tâche en aval**.

"Création d'un corpus parallèle en français de paires de phrases complexes-simples"

Approches automatiques pour la compilation des données en SAT :

- Basées sur des **corpus comparables** ayant des **registres distincts** :
  - **Wikipédia** → Édition française.
  - **Vikidia** → Version adaptée et simplifiée de la première.
- **Normalement** fondés sur des **mesures de similarité sémantique,** mais ignorent si **la phrase cible** constitue effectivement une **simplification.**

1. Introduction

PRINCIPALE CONTRIBUTION	• Présentation d'une **nouvelle méthode** pour **extraire** des paires de **phrases *complexes-simples*** à partir des **corpus comparables**. • **Conçue** spécifiquement pour la **tâche en aval**.

"Création d'un corpus parallèle en français de paires de phrases complexes-simples"

Approches automatiques pour la compilation des données en SAT :

- Basées sur des **corpus comparables** ayant des **registres distincts** :
  - **Wikipédia** → Édition française.
  - **Vikidia** → Version adaptée et simplifiée de la première.
- **Normalement** fondés sur des **mesures de similarité sémantique,** mais **ignorent** si **la phrase cible** constitue effectivement une **simplification.**

C'est pourquoi, nous proposons :

Étapes	Deux conditions à respecter	Examinées de manière séquentielle
1	Préservation du sens original	Filtrage par similarité sémantique
2	Gain en simplicité linguistique	Filtrage par gain de simplicité

2. Acquisition de données : Wikipédia et Vikidia

“Obtenir des données différenciées par registre de langue”

Extraction du contenu textuel des deux **encyclopédies** :

- Récupération des textes **standards** (Wikipédia) et leurs versions **simplifiées** (Vikidia).
- Accent mis sur les ***lead sections*** (résumés introductifs).

“Obtenir des données différenciées par registre de langue”

Extraction du contenu textuel des deux **encyclopédies** :

- Récupération des textes **standards** (Wikipédia) et leurs versions **simplifiées** (Vikidia).
- Accent mis sur les ***lead sections*** (résumés introductifs).
- Sous l’hypothèse que → **Plus de chances de trouver des phrases alignées :**

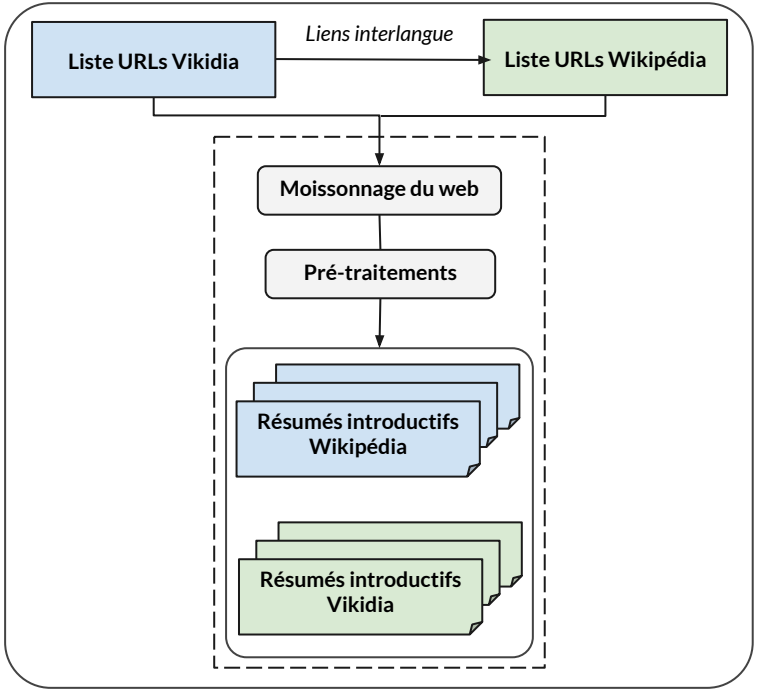
1. Tous deux partagent un **style définitionnel**.
2. Les **extraits communs** sont **plus susceptibles** d’apparaître **au début**.

"Obtenir des données différenciées par registre de langue"

Extraction du contenu textuel des deux **encyclopédies** :

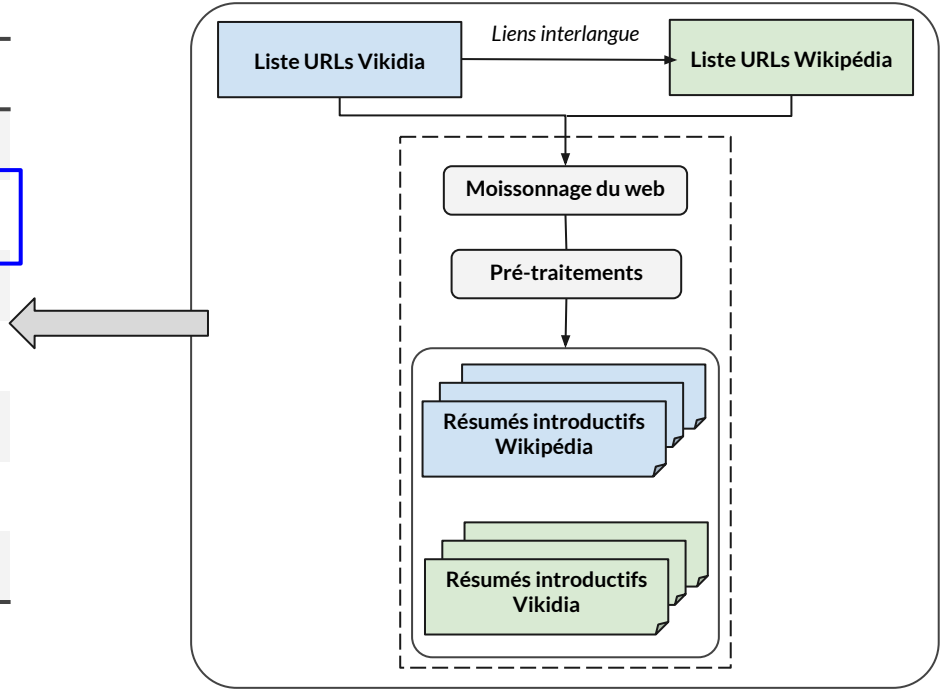
- Récupération des textes **standards** (Wikipédia) et leurs versions **simplifiées** (Vikidia).
- Accent mis sur les ***lead sections*** (résumés introductifs).
- Sous l'hypothèse que → **Plus de chances de trouver des phrases alignées :**

1. Tous deux partagent un **style définitionnel**.
2. Les **extraits communs** sont **plus susceptibles** d'apparaître **au début**.



"Obtenir des données différenciées par registre de langue"

Acquisition de données	Wiki-textes	Viki-textes
# documents	34,806	
# phrases	**165,806**	**134,348**
# tokens	4,030,148	2,373,045
# types	294,979	195,791
Type/token ratio	7.32	8.25
Longueur moyenne mot	5.32	5.08
Longueur moyenne phrase	25.22	18.16



3. Méthode de filtrage en deux étapes pour la SAT

3. Méthode de filtrage en deux étapes pour la SAT

3. Méthode de filtrage en deux étapes pour la SAT

1.1. Étape I : Préservation du sens original

1.2. Étape II : Gain en simplicité linguistique

"Identifier les paires complexes-simples qui sont appropriées pour la SAT"

Acquisition de données	Wiki-textes	Viki-textes
# documents	34,806	
# phrases	165,806	134,348
# tokens	4,030,148	2,373,045
	...	

Alignement automatique des phrases :

Utilisation de ***multilingual sentence transformers*** **(SBERT)**[*]:

- Générant des **vecteurs de sens** à taille fixe, au niveau de la **phrase**.
- **Similarité cosinus** entre les phrases ***Wiki:Viki*** [-1,1].

[*] *https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1*

"Identifier les paires complexes-simples qui sont appropriées pour la SAT"

*Mais quelles paires sont **suffisamment** cohérentes d'un point de vue sémantique ?*

Acquisition de données	Wiki-textes	Viki-textes
# documents	34,806	
# phrases	165,806	134,348
# tokens	4,030,148	2,373,045
	...	

Alignement automatique des phrases :

Utilisation de ***multilingual sentence transformers*** **(SBERT)**[*] :

- Générant des **vecteurs de sens** à taille fixe, au niveau de la **phrase**.
- **Similarité cosinus** entre les phrases ***Wiki*:*Viki*** [-1,1].

[*] *https://huggingface.co/sentence-transformers/paraphrase-xlm-r-multilingual-v1*

"Extraire les paires Wiki et Viki qui présentent un chevauchement sémantique élevé"

3 exemples de paires de phrases *Wiki*:*Viki* :

Label	Wiki-texte	Viki-texte
`Valable`	*L'expression « Maison-Blanche » est souvent employée pour désigner, par métonymie, l'administration du président.*	*Par métonymie, la Maison-Blanche désigne aussi le gouvernement américain et son entourage.*
`Partiellement valable`	*Neal McDonough est un acteur et producteur américain né le 13 février 1966 à Dorchester (Massachusetts).*	*Neal McDonough est un acteur américain.*
`Non valable`	*L'information désigne à la fois le message à communiquer et les symboles utilisés pour l'écrire.*	*Les écrits, les sons, les images, les odeurs ou les goûts contiennent de l'information.*

"Extraire les paires Wiki et Viki qui présentent un chevauchement sémantique élevé"

3 exemples de paires de phrases *Wiki*:*Viki* :

Label	Wiki-texte	Viki-texte
`Valable`	*L'expression « Maison-Blanche » est souvent employée pour désigner, par métonymie, l'administration du président.*	*Par métonymie, la Maison-Blanche désigne aussi le gouvernement américain et son entourage.*
`Partiellement valable`	*Neal McDonough est un acteur **et producteur** américain **né le 13 février 1966 à Dorchester (Massachusetts).***	*Neal McDonough est un acteur américain.*
`Non valable`	*L'information désigne à la fois le message à communiquer et les symboles utilisés pour l'écrire.*	*Les écrits, les sons, les images, les odeurs ou les goûts contiennent de l'information.*

3. Méthode de filtrage en deux étapes pour la SAT

"Extraire les paires Wiki et Viki qui présentent un chevauchement sémantique élevé"

3 exemples de paires de phrases *Wiki*:*Viki* :

Label	Wiki-texte	Viki-texte
`Valable`	*L'expression « Maison-Blanche » est souvent employée pour désigner, par métonymie, l'administration du président.*	*Par métonymie, la Maison-Blanche désigne aussi le gouvernement américain et son entourage.*
`Partiellement valable`	*Neal McDonough est un acteur et producteur américain né le 13 février 1966 à Dorchester (Massachusetts).*	*Neal McDonough est un acteur américain.*
`Non valable`	*L'information **désigne à la fois le message à communiquer et les symboles utilisés pour l'écrire.***	*Les écrits, les sons, les images, les odeurs ou les goûts contiennent de l'information.*

"Extraire les paires Wiki et Viki qui présentent un chevauchement sémantique élevé"

Recours à une annotation manuelle (1/2) :

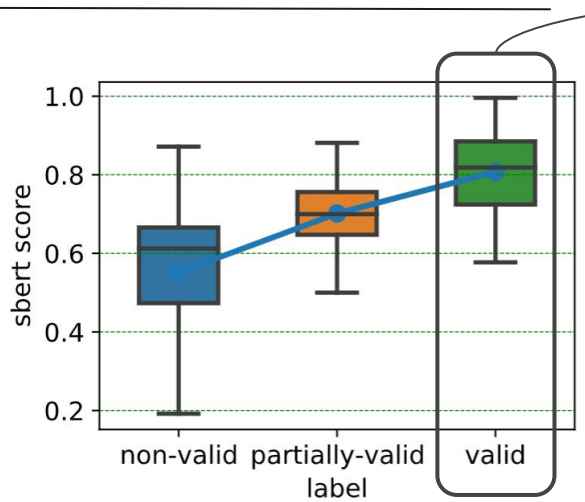
- Sélection aléatoire de **500 paires de phrases** de l'ensemble de données initial.
- **Deux annotateurs** ont évalué dans quelle mesure chaque paire de phrases ***Wiki***:***Viki*** véhiculait le même sens :

  1. `Valable` → L'information et le sens véhiculés par ***Wiki*** et ***Viki*** sont équivalents.
  2. `Partiellement valable` → Des informations sont partiellement perdues de ***Wiki*** à ***Viki*** ou *vice versa*.
  3. `Non-valable` → Les informations entre ***Wiki*** et ***Viki*** sont divergentes.

- **Accord inter-annotateur élevé** (Cohen's kappa = **0.87**).

“Extraire les paires Wiki et Viki qui présentent un chevauchement sémantique élevé”

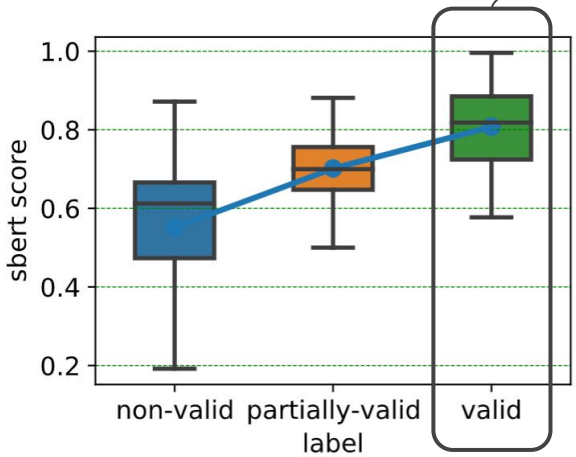
Recours à une annotation manuelle (2/2) :



Score moyen pour le label `Valable` = 0.81

"Extraire les paires Wiki et Viki qui présentent un chevauchement sémantique élevé"

Recours à une annotation manuelle (2/2) :

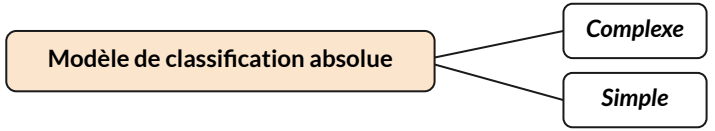


Score moyen pour le label `Valable` = 0.81

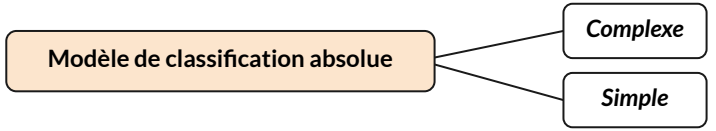
Seuil appliqué pour pour le filtrage sémantique des paires Wiki:Viki


Acquisition de données
Articles Wikipédia
Articles Vikidia
Filtrage par préservation du sens
Filtrage par gain de simplicité

Très souvent, la simplicité est évaluée automatiquement avec…



Très souvent, la simplicité est évaluée automatiquement avec…



✕ Approche insuffisante pour la SAT

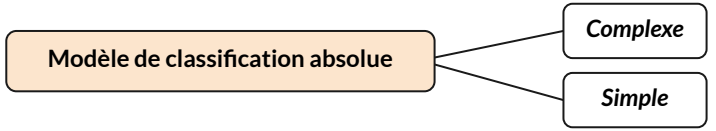
La **simplification** est une **opération intrinsèquement relative** → On transforme une phrase en une version **relativement plus simple** (ce qui n'équivaut pas nécessairement à **simple**).

3. Méthode de filtrage en deux étapes pour la SAT

1.1. Étape I : Préservation du sens original

1.2. Étape II : Gain en simplicité linguistique

Très souvent, la simplicité est évaluée automatiquement avec…



✕ Approche insuffisante pour la SAT

La **simplification** est une **opération intrinsèquement relative** → On transforme une phrase en une version **relativement plus simple** (ce qui n'équivaut pas nécessairement à **simple**).

C'est pourquoi :

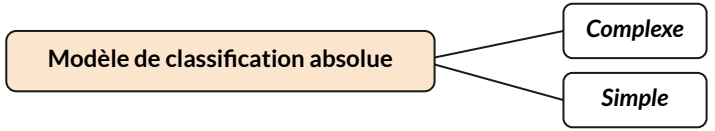
- **Complémenter** deux modèles au classification absolu ; basés sur **l'affinage** des modèles **FLAUBERT** [Le *et al.*, 2020].
- **Objectif** : proposer une **méthode plus fine** pour trouver des **paires** de **phrases** pertinentes pour la **SAT**.

3. Méthode de filtrage en deux étapes pour la SAT

1.1. Étape I : Préservation du sens original

1.2. Étape II : Gain en simplicité linguistique

Très souvent, la simplicité est évaluée automatiquement avec…

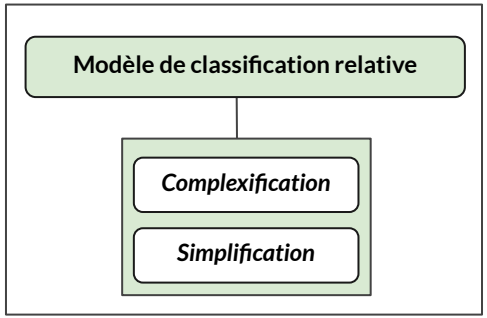


✕ Approche insuffisante pour la SAT

La **simplification** est une **opération intrinsèquement relative** → On transforme une phrase en une version **relativement plus simple** (ce qui n'équivaut pas nécessairement à **simple**).

C'est pourquoi :

- **Complémenter** deux modèles au classification absolu ; basés sur **l'affinage** des modèles **FLAUBERT** [Le *et al.*, 2020].
- **Objectif** : proposer une **méthode plus fine** pour trouver des **paires** de **phrases** pertinentes pour la **SAT**.

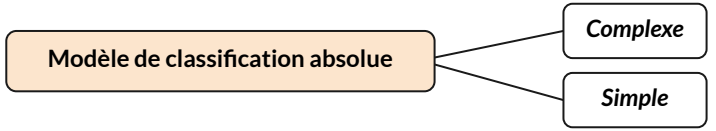


3. Méthode de filtrage en deux étapes pour la SAT

1.1. Étape I : Préservation du sens original

1.2. Étape II : Gain en simplicité linguistique

Très souvent, la simplicité est évaluée automatiquement avec…

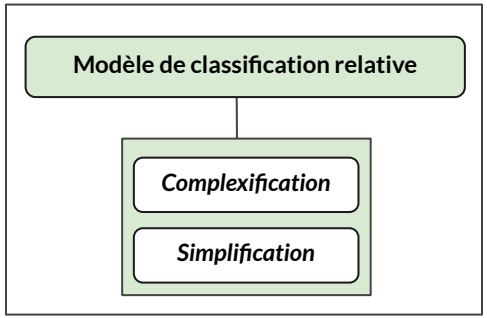


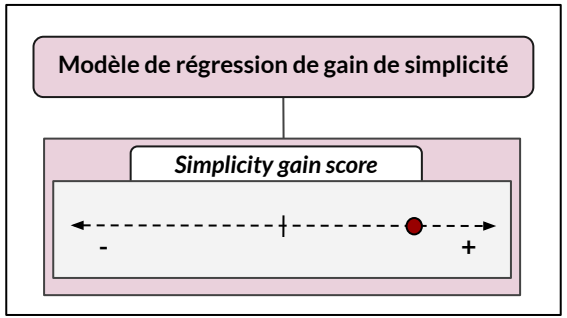
✕ Approche insuffisante pour la SAT

La **simplification** est une **opération intrinsèquement relative** → On transforme une phrase en une version **relativement plus simple** (ce qui n'équivaut pas nécessairement à **simple**).

C'est pourquoi :

- **Complémenter** deux modèles au classification absolu ; basés sur **l'affinage** des modèles **FLAUBERT** [Le *et al.*, 2020].
- **Objectif** : proposer une **méthode plus fine** pour trouver des **paires** de **phrases** pertinentes pour la **SAT**.



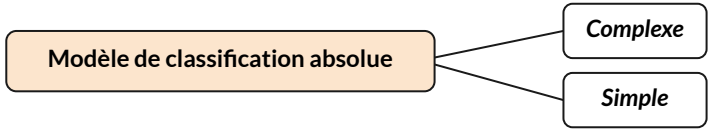


3. Méthode de filtrage en deux étapes pour la SAT

1.1. Étape I : Préservation du sens original

1.2. Étape II : Gain en simplicité linguistique

Très souvent, la simplicité est évaluée automatiquement avec…

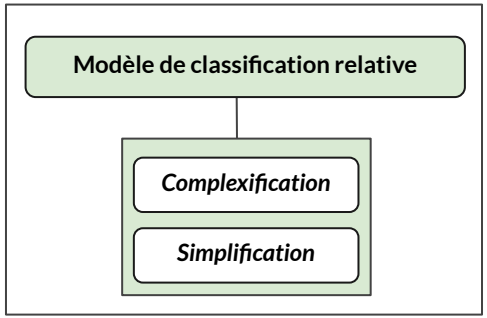


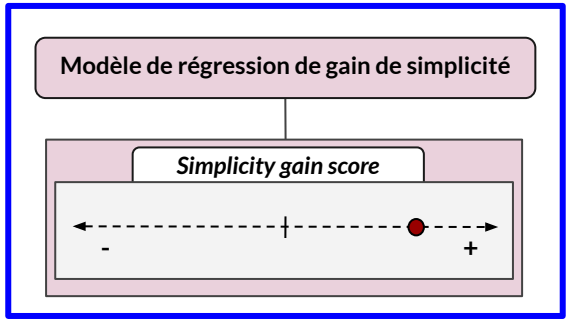
✕ Approche insuffisante pour la SAT

La **simplification** est une **opération intrinsèquement relative** → On transforme une phrase en une version **relativement plus simple** (ce qui n'équivaut pas nécessairement à **simple**).

C'est pourquoi :

- **Complémenter** deux modèles au classification absolu ; basés sur **l'affinage** des modèles **FLAUBERT** [Le *et al.*, 2020].
- **Objectif** : proposer une **méthode plus fine** pour trouver des **paires** de **phrases** pertinentes pour la **SAT**.





"Déterminer quelles paires constituent des simplifications valables"

GAIN → **Différence absolue** entre la valeur d'un trait dans la phrase source et la phrase cible, y compris sa **polarité**.

- **Ensemble** de **traits**, divisés en trois groupes : **structurels**, **lexicaux** et **syntaxiques**, suivant la littérature précédente [Tanguy & Tulechki, 2009 ; Brunato *et al.*, 2022].
- Définissant un **score** de gain de simplicité.

"Déterminer quelles paires constituent des simplifications valables"

modèle de régression

pairs	feature gains			
wv_1	g_{11}	g_{12}	g_{13}	…
wv_2	g_{21}	g_{22}	g_{23}	…
…				

GAIN → **Différence absolue** entre la valeur d'un trait dans la phrase source et la phrase cible, y compris sa **polarité**.

- **Ensemble** de **traits**, divisés en trois groupes : **structurels**, **lexicaux** et **syntaxiques**, suivant la littérature précédente [Tanguy & Tulechki, 2009 ; Brunato *et al.*, 2022].
- Définissant un **score** de gain de simplicité.

Trait
Nombre de mots

8 mots → *6 mots* = *gain = +2*

Wiki ***Viki***

"Déterminer quelles paires constituent des simplifications valables"

Tâche	**Modèle de classification absolue**		**Modèle de classification relative**		**Modèle de régression**	
Eval. dataset	*Corpus test* (précision %)		*Corpus test* (précision %)		*Corpus test* (score MSE)	
Modèle *transformer*	*untuned*	*tuned*	*untuned*	*tuned*	*untuned*	*tuned*
`flaubert-small`	49,54	**70,11**	49,78	92,99	1,89	0,39
`flaubert-base`	50,97	69,82	49,88	93,82	1,18	0,35
`flaubert-large`	52,29	69,19	52,18	**94,16**	4,59	**0,23**

"Déterminer quelles paires constituent des simplifications valables"

	Wiki	*Viki*
	En France, ce lézard est strictement protégé par la loi.	*En France, il est protégé par la loi.*
Modèle de classification absolue		
Modèle de classification relative		
Modèle de régression de gain de simplicité		

"Déterminer quelles paires constituent des simplifications valables"

	Wiki	*Viki*
	En France, ce lézard est strictement protégé par la loi.	*En France, il est protégé par la loi.*
Modèle de classification absolue	`Complexe`	`Simple`
Modèle de classification relative		
Modèle de régression de gain de simplicité		

"Déterminer quelles paires constituent des simplifications valables"

	Wiki	*Viki*
	En France, ce lézard est strictement protégé par la loi.	*En France, il est protégé par la loi.*
Modèle de classification absolue	`Complexe`	`Simple`
Modèle de classification relative	`Simplification`	
Modèle de régression de gain de simplicité		

"Déterminer quelles paires constituent des simplifications valables"

	Wiki	*Viki*
	En France, ce lézard est strictement protégé par la loi.	*En France, il est protégé par la loi.*
Modèle de classification absolue	`Complexe`	`Simple`
Modèle de classification relative	`Simplification`	
Modèle de régression de gain de simplicité	`+0,84`	

"Déterminer quelles paires constituent des simplifications valables"

	Wiki	***Viki***
	Praticien précoce et représentant éminent du concept français de la haute gastronomie, il est considéré comme le fondateur de ce style grandiose, recherché à la fois par les cours royales et les nouveaux riches de Paris	*Il est considéré comme l'un des pionniers, sinon le fondateur, de la gastronomie française.*
Modèle de classification absolue		
Modèle de classification relative		
Modèle de régression de gain de simplicité		

"Déterminer quelles paires constituent des simplifications valables"

	Wiki	*Viki*
	Praticien précoce et représentant éminent du concept français de la haute gastronomie, il est considéré comme le fondateur de ce style grandiose, recherché à la fois par les cours royales et les nouveaux riches de Paris	*Il est considéré comme l'un des pionniers, sinon le fondateur, de la gastronomie française.*
Modèle de classification absolue	`Complexe`	`Complexe`
Modèle de classification relative		
Modèle de régression de gain de simplicité		

"Déterminer quelles paires constituent des simplifications valables"

	Wiki	*Viki*
	Praticien précoce et représentant éminent du concept français de la haute gastronomie, il est considéré comme le fondateur de ce style grandiose, recherché à la fois par les cours royales et les nouveaux riches de Paris	*Il est considéré comme l'un des pionniers, sinon le fondateur, de la gastronomie française.*
Modèle de classification absolue	`Complexe`	`Complexe`
Modèle de classification relative	`Simplification`	
Modèle de régression de gain de simplicité	`+1,95`	

"Déterminer quelles paires constituent des simplifications valables"

	Wiki	*Viki*
	Makassar ou Macassar est une ville d'Indonésie et la capitale de la province de Sulawesi du Sud.	*Macassar ou Makassar est une ville d'Indonésie, située sur l'île de Sulawesi (ou Célèbes), en bordure du détroit du même nom.*
Modèle de classification absolue		
Modèle de classification relative		
Modèle de régression de gain de simplicité		

"Déterminer quelles paires constituent des simplifications valables"

	Wiki	*Viki*
	Makassar ou Macassar est une ville d'Indonésie et la capitale de la province de Sulawesi du Sud.	*Macassar ou Makassar est une ville d'Indonésie, située sur l'île de Sulawesi (ou Célèbes), en bordure du détroit du même nom.*
Modèle de classification absolue	`Simple`	`Complexe`
Modèle de classification relative		
Modèle de régression de gain de simplicité		

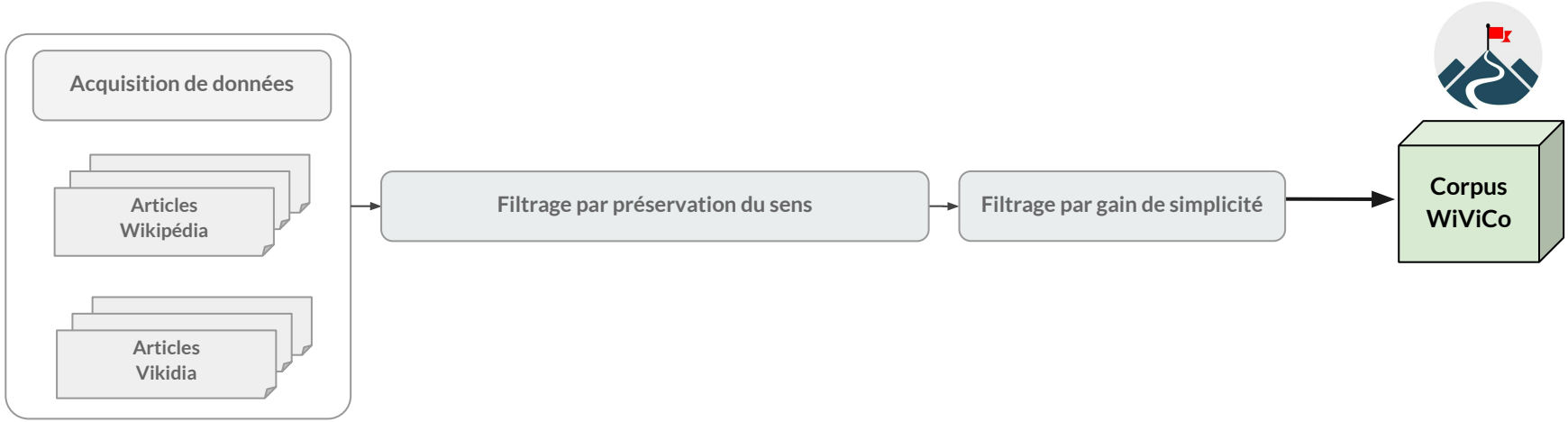
"Déterminer quelles paires constituent des simplifications valables"

	Wiki	*Viki*
	Makassar ou Macassar est une ville d'Indonésie et la capitale de la province de Sulawesi du Sud.	*Macassar ou Makassar est une ville d'Indonésie, située sur l'île de Sulawesi (ou Célèbes), en bordure du détroit du même nom.*
Modèle de classification absolue	`Simple`	`Complexe`
Modèle de classification relative	`Complexification`	
Modèle de régression de gain de simplicité		

"Déterminer quelles paires constituent des simplifications valables"

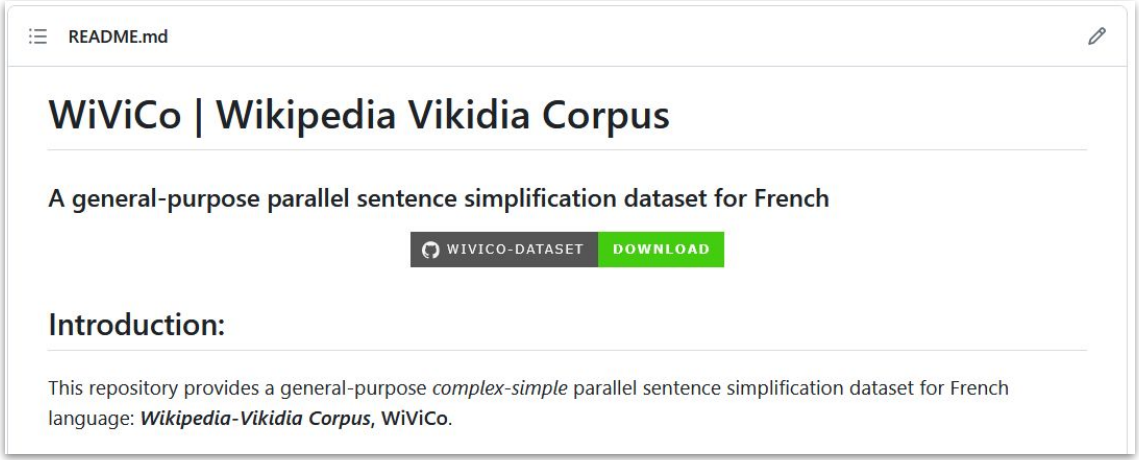
	Wiki	*Viki*
	Makassar ou Macassar est une ville d'Indonésie et la capitale de la province de Sulawesi du Sud.	*Macassar ou Makassar est une ville d'Indonésie, située sur l'île de Sulawesi (ou Célèbes), en bordure du détroit du même nom.*
Modèle de classification absolue	`Simple`	`Complexe`
Modèle de classification relative	`Complexification`	
Modèle de régression de gain de simplicité	`-2,65`	

Suite à l'application de ce filtrage en deux étapes…



4. Résultats : *Wikipedia-Vikidia Corpus* (WiViCo)

À l'issue de l'application de cette méthode de filtrage :


README.md

WiViCo | Wikipedia Vikidia Corpus

A general-purpose parallel sentence simplification dataset for French

WIVICO-DATASET DOWNLOAD

Introduction:

This repository provides a general-purpose *complex-simple* parallel sentence simplification dataset for French language: ***Wikipedia-Vikidia Corpus***, WiViCo.

https://github.com/lormaechea/wivico

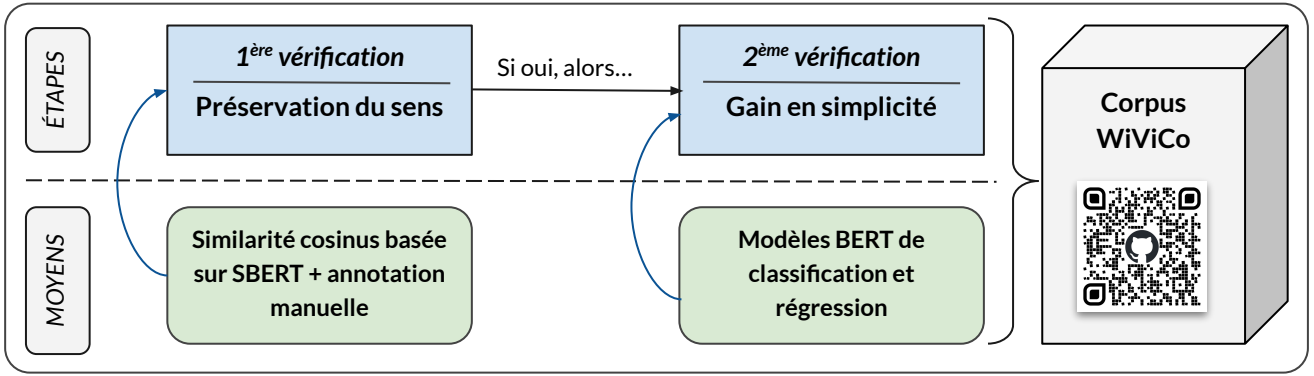
- Jeu de données en français associant des paires de **phrases *complexes-simples*.**
- Version actuelle : presque **65*k* phrases parallèles.**

5. Conclusions

Principale contribution :

Introduction d'une nouvelle **méthode** pour **exploiter des corpus comparables** :

- Spécifiquement **ciblée** pour la **tâche** de **SAT**.
- **Approche séquentielle** pour satisfaire les **2 conditions principales** qu'un **texte simplifié** doit remplir pour être considéré comme étant **valable** :



5. Conclusions

Poursuites :

RÉGLAGE

- **Amélioration** de **l'interprétabilité** du **score** de **gain** de **simplicité**.
- **Besoin** d'un contraste avec des **jugements humains.**

5. Conclusions

Poursuites :

RÉGLAGE

- **Amélioration** de **l'interprétabilité** du **score** de **gain** de **simplicité**.
- **Besoin** d'un contraste avec des **jugements humains.**

GÉNÉRATION

- **Entraînement** des **modèles** de **SAT** en français**.**
- **Affinage** d'un **grand modèle de langage pré-entraîné** → Pour notre **tâche en aval.**

5. Conclusions

Poursuites :

RÉGLAGE

- **Amélioration** de **l'interprétabilité** du **score** de **gain** de **simplicité**.
- **Besoin** d'un contraste avec des **jugements humains.**

GÉNÉRATION

- **Entraînement** des **modèles** de **SAT** en français**.**
- **Affinage** d'un **grand modèle de langage pré-entraîné** → Pour notre **tâche en aval.**

ADAPTATION

- **Étendre** l'application de modèles de **SAT** à une **modalité orale.**
- **Examiner** comment les **adapter** pour les **rendre robustes** au **bruit** induit par les **caractéristiques de la langue parlée**.

Merci de votre attention !

Lucia.OrmaecheaGrijalba@unige.ch

https://luciaormaechea.com/fr/

Références

[Par ordre d'apparition]

- **C. Horn, C. Manduca, and D. Kauchak**. Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463. Association for Computational Linguistics, **2014**. URL http://aclweb.org/anthology/P14-2075

- **S. Stajner**. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics*: ACL-IJCNLP 2021, pages 2637–2652. Association for Computational Linguistics, **2021**. URL https://aclanthology.org/2021.findings-acl.233

- **N. Gala, A. Tack, L. Javourey-Drevet, T. François, and J. C. Ziegler**. Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association, **2020**. URL https://aclanthology.org/2020.lrec-1.169

- **X. Zhang and M. Lapata**. Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics, **2017**. URL https://aclanthology.org/D17-1062/