Extracting sentence simplification pairs from French comparable corpora using a two-step filtering method

Swiss Text Analytics Conference 2023

June 13th, University of Applied Sciences and Arts of Western Switzerland, Neuchâtel

Lucía **Ormaechea** and Nikos **Tsourakis**

Department of Translation Technology

UNIVERSITÉ DE GENÈVE

FACULTÉ DE TRADUCTION ET D'INTERPRÉTATION

Plan

1. Introduction

Automatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.

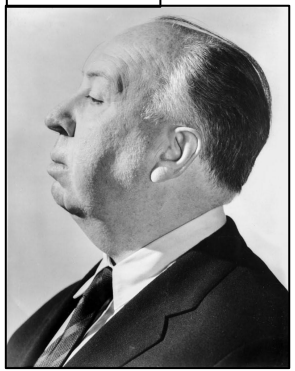
1. Introduction

Automatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.

ATS: an analogy with the form-substance relation:

Complex



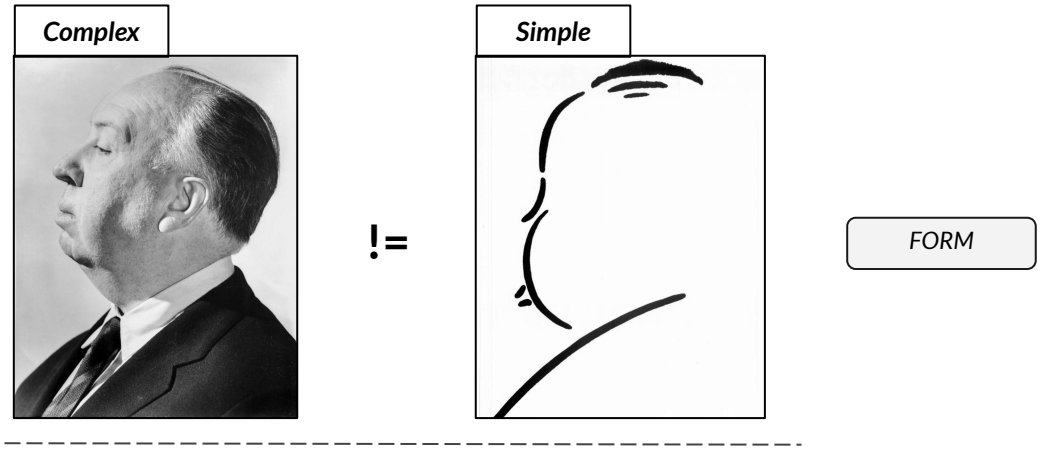
FORM

1. Introduction

Automatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.

ATS: an analogy with the form-substance relation:

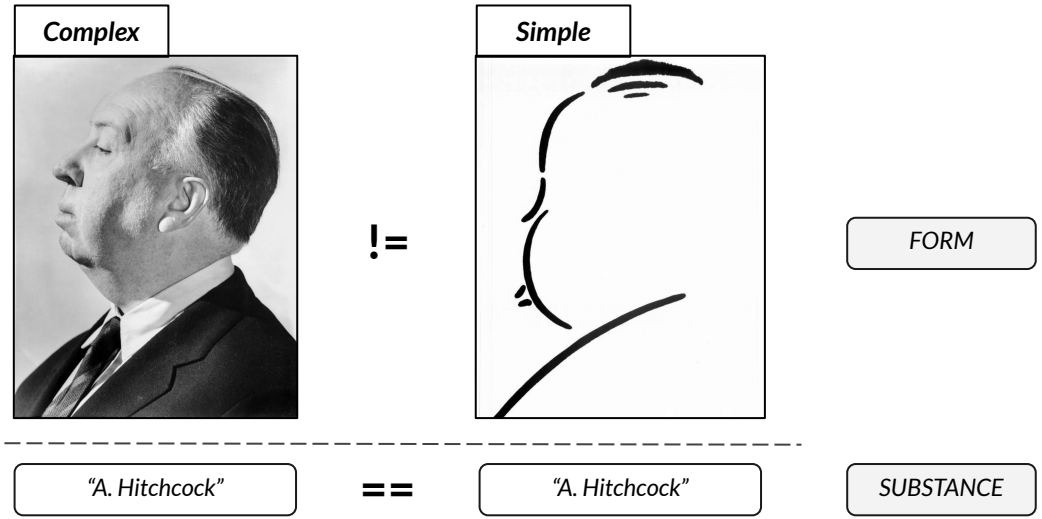


1. Introduction

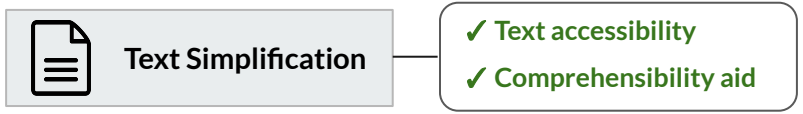
Automatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** [Horn *et al.*, 2014; Stajner, 2021].

ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.

ATS: an analogy with the form-substance relation:



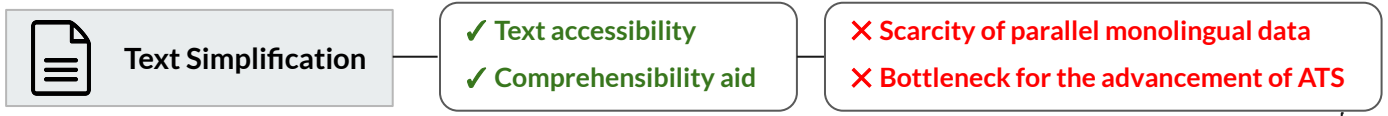
1. Introduction



ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.

1. Introduction

1. Introduction



ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.

In less resource-rich languages than English: *say.... French*

Only available monolingual parallel dataset → **ALECTOR** [Gala *et al.*, 2020]:

- **Manually created** by professional editors.
- Comprising schoolbook texts, simplified for child audiences.

1. Introduction

Text Simplification

✓ Text accessibility
✓ Comprehensibility aid

✕ Scarcity of parallel monolingual data
✕ Bottleneck for the advancement of ATS

ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.

In less resource-rich languages than English: *say.... French*

Only available monolingual parallel dataset → **ALECTOR** [Gala *et al.*, 2020]:

- **Manually created** by professional editors.
- Comprising schoolbook texts, simplified for child audiences.

Manually crafted datasets

✓ High-quality and reliable simplifications

✕ Costly compilation, both economically and time-wise

✕ Compact size, preventing the implementation of ML algorithms

1. Introduction

Text Simplification

✓ Text accessibility
✓ Comprehensibility aid

✕ Scarcity of parallel monolingual data
✕ Bottleneck for the advancement of ATS

ORIGINAL	The second ***largest*** city of China ~~***and one of the world's major cities***~~, Beijing has played ***a vital*** role in Chinese history.
SIMPLIFIED	***Beijing*** is the second ***biggest*** city of China. Beijing has played ***an important*** role in Chinese history.

In less resource-rich languages than English: *say.... French*

Only available monolingual parallel dataset → **ALECTOR** [Gala *et al.*, 2020]:

- **Manually created** by professional editors.
- Comprising schoolbook texts, simplified for child audiences.

Manually crafted datasets

✓ High-quality and reliable simplifications

✕ Costly compilation, both economically and time-wise

✕ Compact size, preventing the implementation of ML algorithms

What about exploring automatic methods?

1. Introduction

MAIN CONTRIBUTION	• Introduce a **new method** to **mine comparable corpora** to **extract *complex-simple* pairs**. • Properly **adapted** to the **ATS** task.

"Create a parallel French corpus of complex-simple pairs for ATS"

1. Introduction

MAIN CONTRIBUTION	• Introduce a **new method** to **mine comparable corpora** to **extract *complex-simple* pairs**. • Properly **adapted** to the **ATS** task.

"Create a parallel French corpus of complex-simple pairs for ATS"

Automatic approaches for ATS data compilation:

- **Typically** based on **register-diversified comparable corpora**:
  - **Wikipedia** → French edition.
  - **Vikidia** → Simplified version of the former.

1. Introduction

MAIN CONTRIBUTION	• Introduce a **new method** to **mine comparable corpora** to **extract *complex-simple* pairs**. • Properly **adapted** to the **ATS** task.

"Create a parallel French corpus of complex-simple pairs for ATS"

Automatic approaches for ATS data compilation:

- **Typically** based on **register-diversified comparable corpora**:
  - **Wikipedia** → French edition.
  - **Vikidia** → Simplified version of the former.
- **Usually** rely on **semantic similarity measures**, but are **oblivious** of whether the **target sentence constitutes indeed a simplification**.

1. Introduction

MAIN CONTRIBUTION	• Introduce a **new method** to **mine comparable corpora** to **extract *complex-simple* pairs**. • Properly **adapted** to the **ATS** task.

"Create a parallel French corpus of complex-simple pairs for ATS"

Automatic approaches for ATS data compilation:

- **Typically** based on **register-diversified comparable corpora**:
  - **Wikipedia** → French edition.
  - **Vikidia** → Simplified version of the former.
- **Usually** rely on **semantic similarity measures**, but are **oblivious** of whether the **target sentence constitutes indeed a simplification**.

That's why we propose…

Stages	Two primary conditions within ATS	Tackled in a sequential manner
1	Retention of the original meaning and information	Sentence semantic similarity filtering
2	Linguistic simplicity gain with respect to the reference	Simplicity gain filtering

2. Data acquisition: scraping Wikipedia and Vikidia

"Obtain register-differentiated data"

Extracting textual content from both **encyclopedias**:

- Scrape standard texts (Wikipedia) and their simplified versions (Vikidia).
- Special focus on ***lead sections*** (initial summaries).

"Obtain register-differentiated data"

Extracting textual content from both **encyclopedias**:

- Scrape standard texts (Wikipedia) and their simplified versions (Vikidia).
- Special focus on ***lead sections*** (initial summaries).
- Assumption that → **Greater chances of finding aligned sentences**:

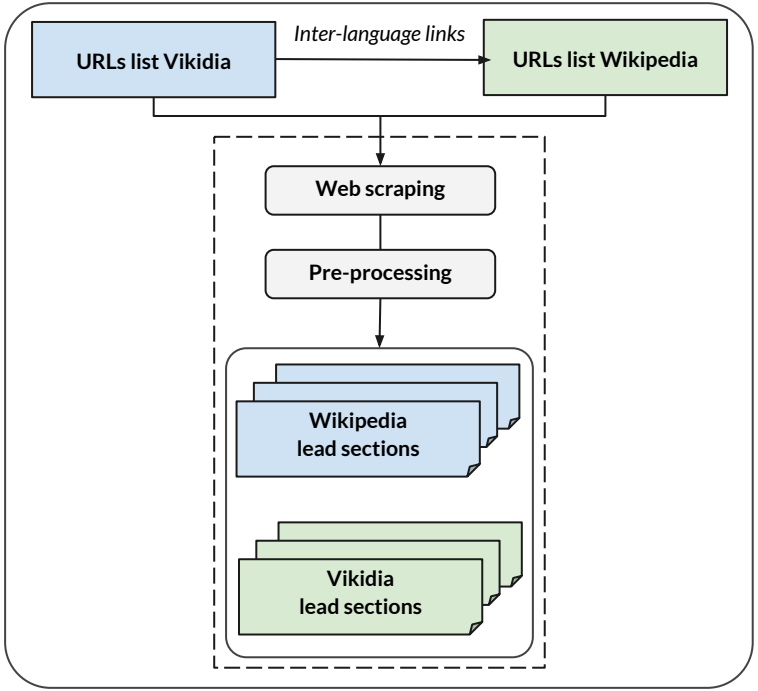
1. Both share a typically **definitional style**.
2. **Common excerpts** are potentially **more likely** to appear in the **beginning**.

"Obtain register-differentiated data"

Extracting textual content from both **encyclopedias**:

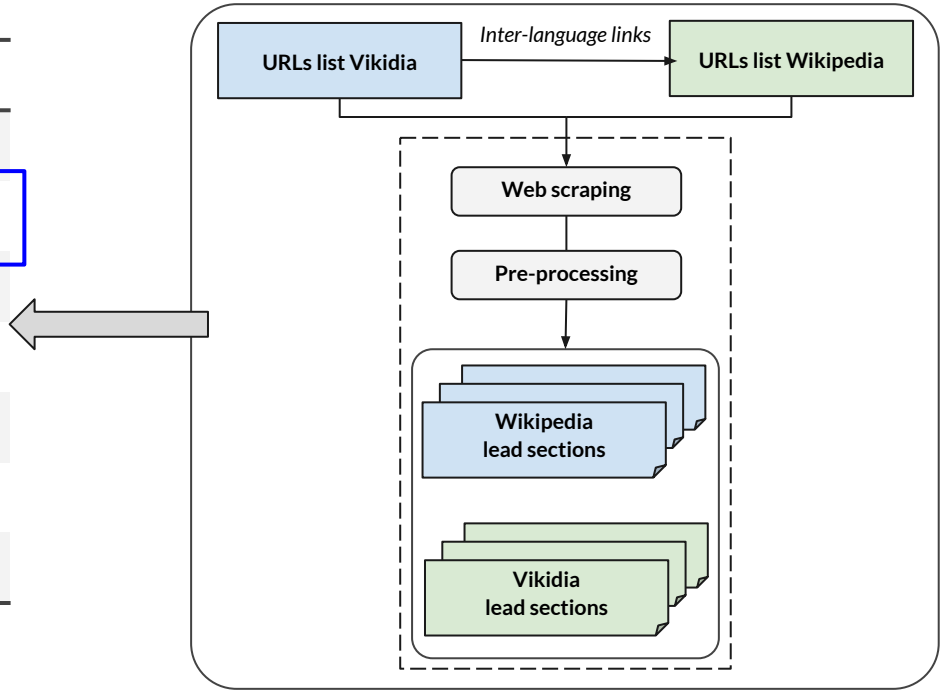
- Scrape standard texts (Wikipedia) and their simplified versions (Vikidia).
- Special focus on ***lead sections*** (initial summaries).
- Assumption that → **Greater chances of finding aligned sentences**:

1. Both share a typically **definitional style**.
2. **Common excerpts** are potentially **more likely** to appear in the **beginning**.



"Obtain register-differentiated data"

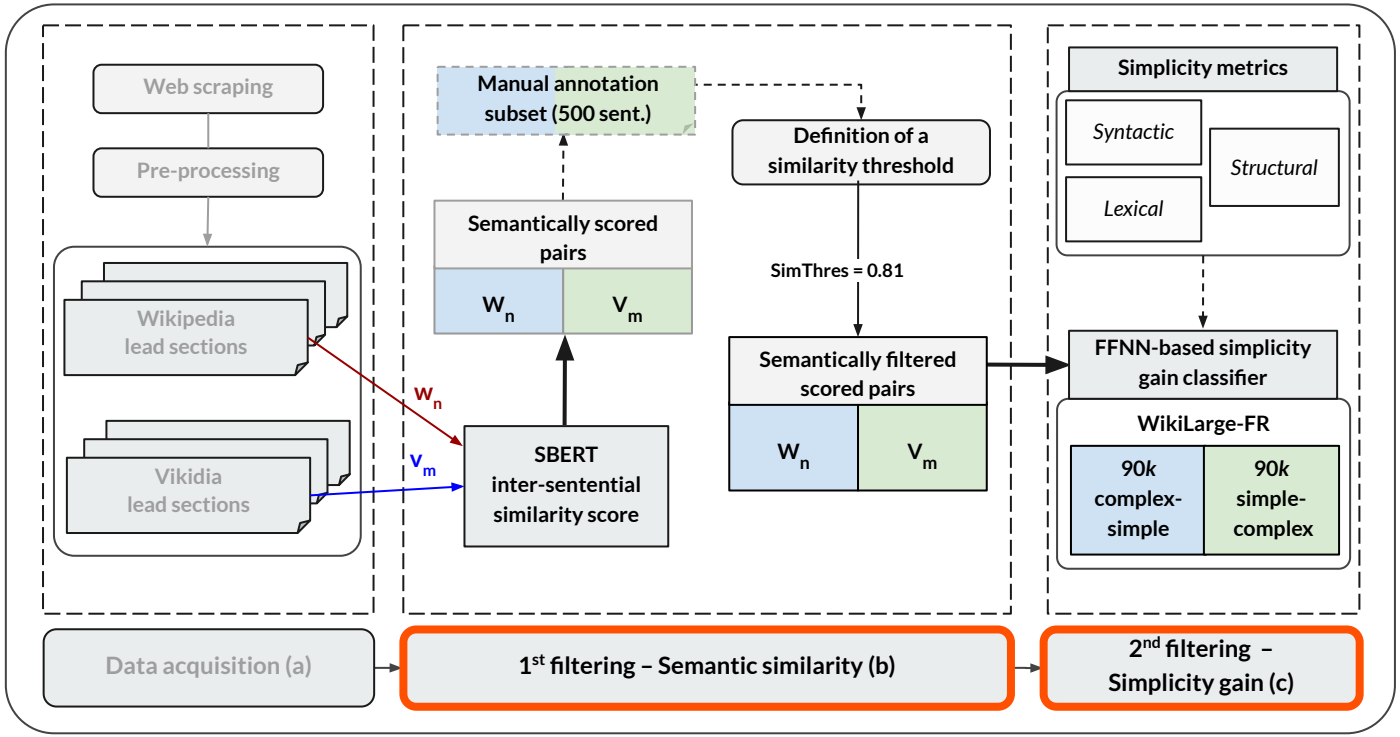
Data acquisition	Wiki-texts	Viki-texts
# documents	34,806	
# sentences	**165,806**	**134,348**
# tokens	4,030,148	2,373,045
# types	294,979	195,791
Type/token ratio	7.32	8.25
Avg. word length	5.32	5.08
Avg. sent. length	25.22	18.16



3. Two-step ATS-targeted filtering method

1.1. Stage I: Sentence semantic similarity filtering

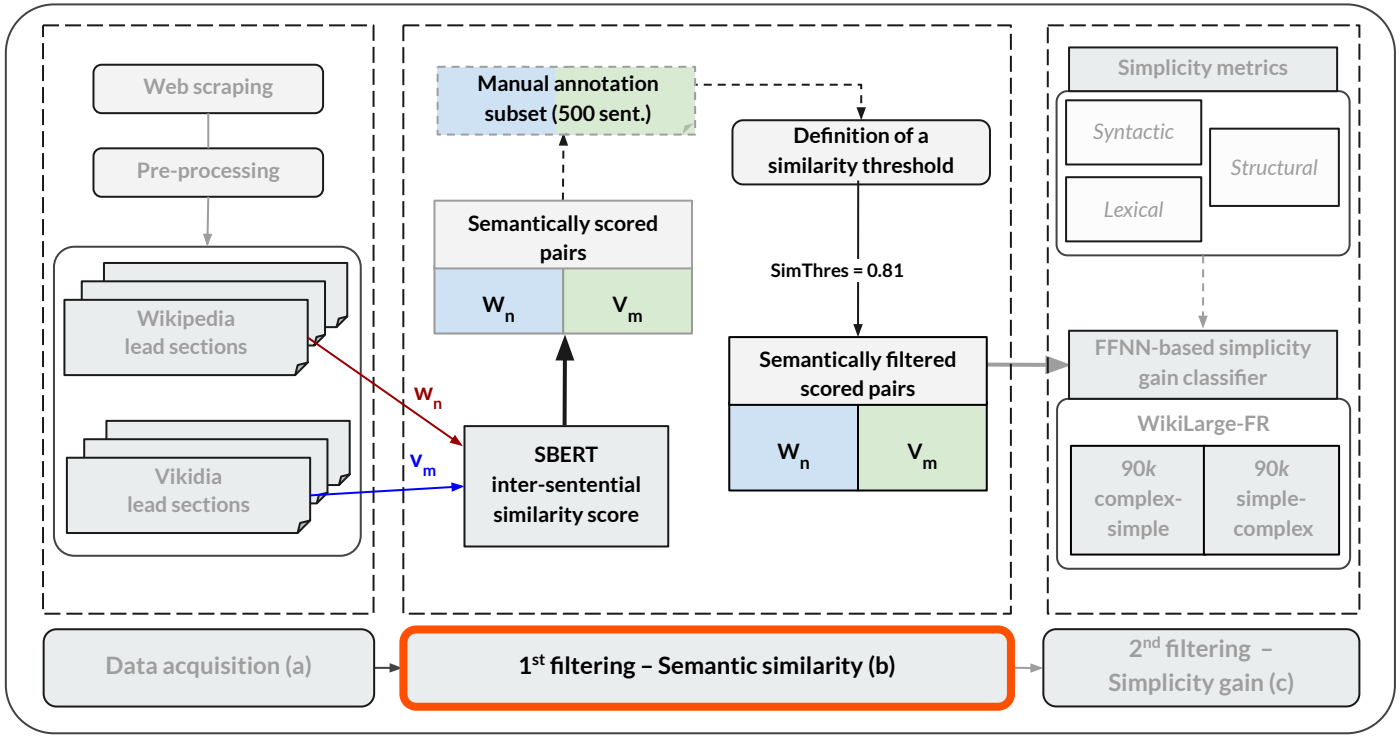
1.2. Stage II: Simplicity gain filtering



3. Two-step ATS-targeted filtering method

1.1. Stage I: Sentence semantic similarity filtering

1.2. Stage II: Simplicity gain filtering



"Determine which complex-simple pairs are suitable for ATS"

But, which pairs are sufficiently semantically consistent?

Data acquisition	Wiki-texts	Viki-texts
# documents	34,806	
# sentences	165,806	134,348
# tokens	4,030,148	2,373,045
...		

Automatic sentence alignment:

- **SBERT** (Sentence-BERT) using **multilingual** sentence transformers*:
  - Generate fixed-length sentence **embeddings**.
  - The output is a **768-dimensional** dense vector representation.
  - **Cosine similarity** between complex and simple sentences.

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

"Extract the Wiki- and Viki-pairs that exhibit a high semantic overlap"

Three ***Wiki***:***Viki*** example sentences

Label	*Wikipedia sentence*	*Vikidia sentence*
Valid	*The term "White House" is often used as a metonym for the president's administration.*	*By metonymy, the White House also refers to the US government and its entourage.*
Partially valid	*Neal McDonough is an American actor and producer born on February 13, 1966 in Dorchester, Massachusetts.*	*Neal McDonough is an American actor.*
Non-valid	*The information refers both to the message to be communicated and the symbols used to write it.*	*The written words, sounds, images, smells or tastes contain information.*

"Extract the Wiki- and Viki-pairs that exhibit a high semantic overlap"

Three *Wiki*:*Viki* example sentences

Label	*Wikipedia sentence*	*Vikidia sentence*
Valid	*The term "White House" is often used as a metonym for the president's administration.*	*By metonymy, the White House also refers to the US government and its entourage.*
Partially valid	*Neal McDonough is an American actor and producer born on February 13, 1966 in Dorchester, Massachusetts.*	*Neal McDonough is an American actor.*
Non-valid	*The information refers both to the message to be communicated and the symbols used to write it.*	*The written words, sounds, images, smells or tastes contain information.*

"Extract the Wiki- and Viki-pairs that exhibit a high semantic overlap"

Three *Wiki*:*Viki* example sentences

Label	*Wikipedia sentence*	*Vikidia sentence*
Valid	*The term "White House" is often used as a metonym for the president's administration.*	*By metonymy, the White House also refers to the US government and its entourage.*
Partially valid	*Neal McDonough is an American actor and producer born on February 13, 1966 in Dorchester, Massachusetts.*	*Neal McDonough is an American actor.*
Non-valid	*The information refers both to the message to be communicated and the symbols used to write it.*	*The written words, sounds, images, smells or tastes contain information.*

"Extract the Wiki- and Viki-pairs that exhibit a high semantic overlap"

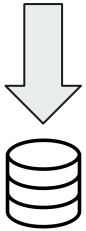
Manual annotation:

- Randomly selecting **500 samples** from the initial dataset.
- **Two annotators** determined to which extent each pair of ***Wiki***:***Viki*** sentences conveyed the same meaning using three judgement labels:
  - **Valid**, where the meaning and information from ***Wiki*** to ***Viki*** is fully preserved.
  - **Partially valid**, where information is partially lost from ***Wiki*** to ***Viki*** or vice versa.
  - **Non-valid**, where information between ***Wiki*** to ***Viki*** is dissimilar.
- High agreement (Cohen's kappa = **0.87).**

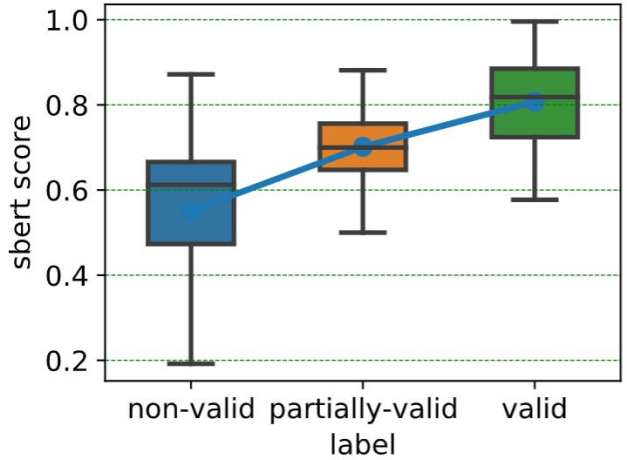
"Extract the Wiki- and Viki-pairs that exhibit a high semantic overlap"

Manual annotation:

The mean score for the valid case was equal to 0.81, which we consider as the cutoff threshold for the semantic filtering of Wiki to Viki pairs.



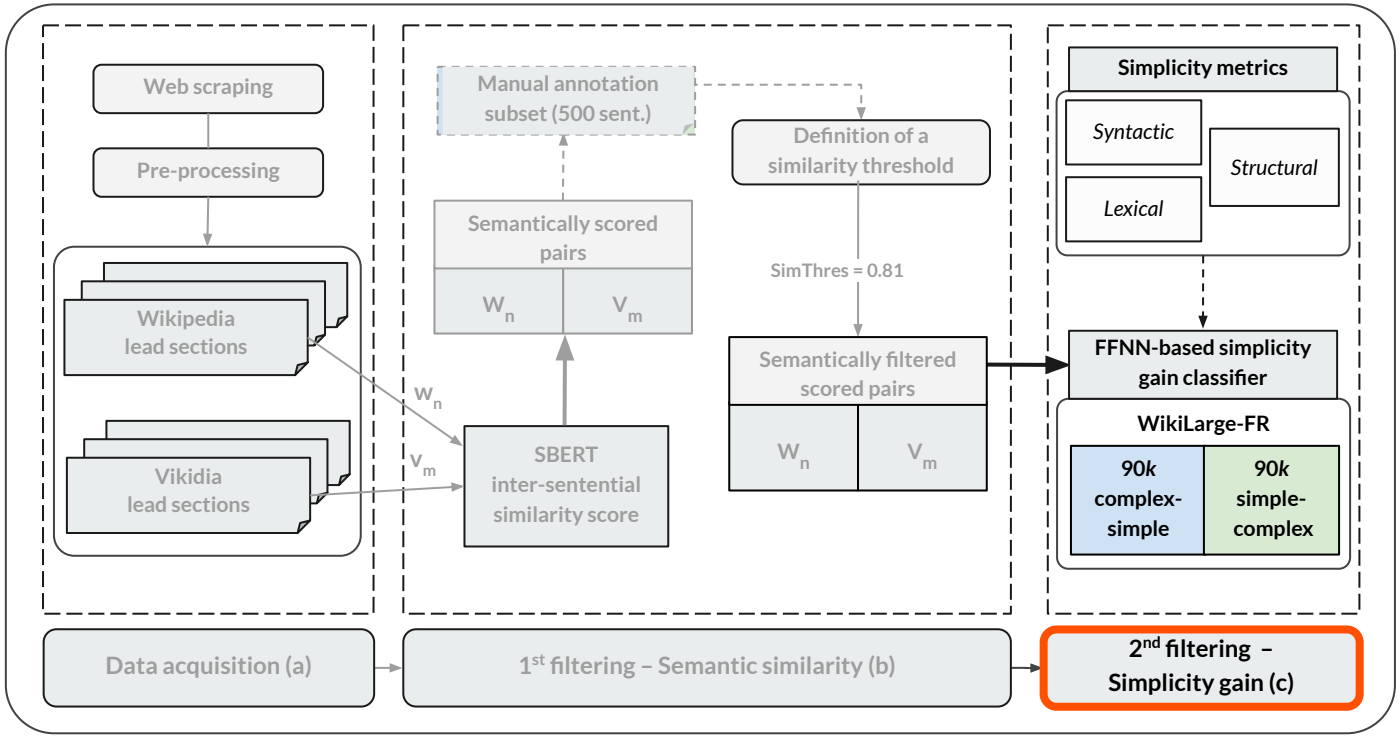
Semantically filtered pairs



3. Two-step ATS-targeted filtering method

1.1. Stage I: Sentence semantic similarity filtering

1.2. Stage II: Simplicity gain filtering



"Determine which pairs constitute valid simplifications"

binary classifier

Gain: Signifies the **absolute difference** between a feature value in the source and target sentence, including polarity.

Feature
Number of words

8 words → *6 words* = *gain = -2*

Wiki *Viki*

pairs	feature gains			
wv_1	g_{11}	g_{12}	g_{13}	…
wv_2	g_{21}	g_{22}	g_{23}	…
wv_3	g_{31}	g_{32}	g_{33}	…
wv_4	g_{41}	g_{42}	g_{43}	…

…

Feature group	Feature	Description
Structural	Sentence Length (SL)	Difference in the number of characters between the target and source sentences.
	Number of Words (NW)	Difference in the number of words between the target and source sentences.
	Word Error Rate (WER)	Word-based similarity between the source and target sentences.
	BLEU score	*n*-gram overlap via precision of the target sentence with its corresponding source sentence.
Lexical	Number of Named Entities (NNE)	Difference in the number of named entities (organizations, people, places, *etc.*) between the target and source sentences.
	CEFR score	Within a sentence, sum of the frequencies of CEFR levels of all non-stop words multiplied by their complexity weight value.
Syntactic	Maximum Depth Tree (MDT)	Difference in the maximum depth of the dependency tree between the target and source sentences.
	Incomplete Dependency Theory (IDT)	Within a phrase, the average number of incomplete dependencies between the current and next token.
	Dependency Locality Theory (DLT)	For every head token in a sentence, the number of discourse referents starting from the current token and ending to its longest leftmost dependent [32]. Values are then combined using an average function.
	Combined IDT+DLT	Sum of IDT+DLT metrics for all tokens in a sentence. Resulting values are then combined using an average function.
	Left Embeddedness (LE)	Within a sentence, the number of tokens on the left-hand-side of the root verb that are not verbs.
	Noun Nested Distance (NND)	The average nested distance of all nouns within a phrase that have as ancestor another noun in the dependency tree.

"Determine which pairs constitute valid simplifications"

- Trained a simplicity gain classifier on these features.
  - Feed-Forward Neural Network (FFNN) - Four hidden layers of 256 nodes each.
- Exploited ~180k example pairs from **WikiLarge** (EN).
  - Resorted to Google Translate (FR).
- Extracted values of features independently for each article and combined them for a single pair.
- Split the dataset into **two halves** interchanging the source/target.
  - Acquired simplification/complexification pairs.
  - Calculated the gain values for each pair.

3. Two-step ATS-targeted filtering method

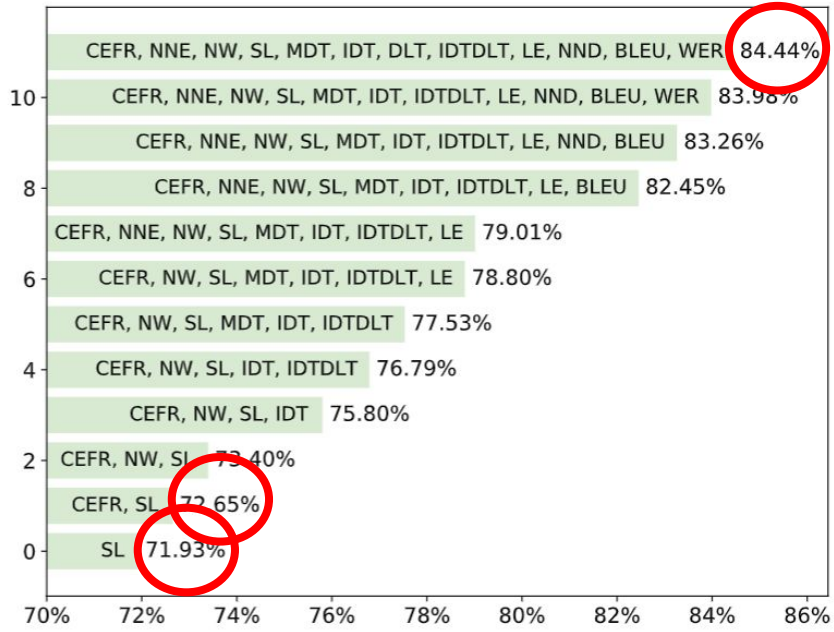
1.1. Stage I: Sentence semantic similarity filtering

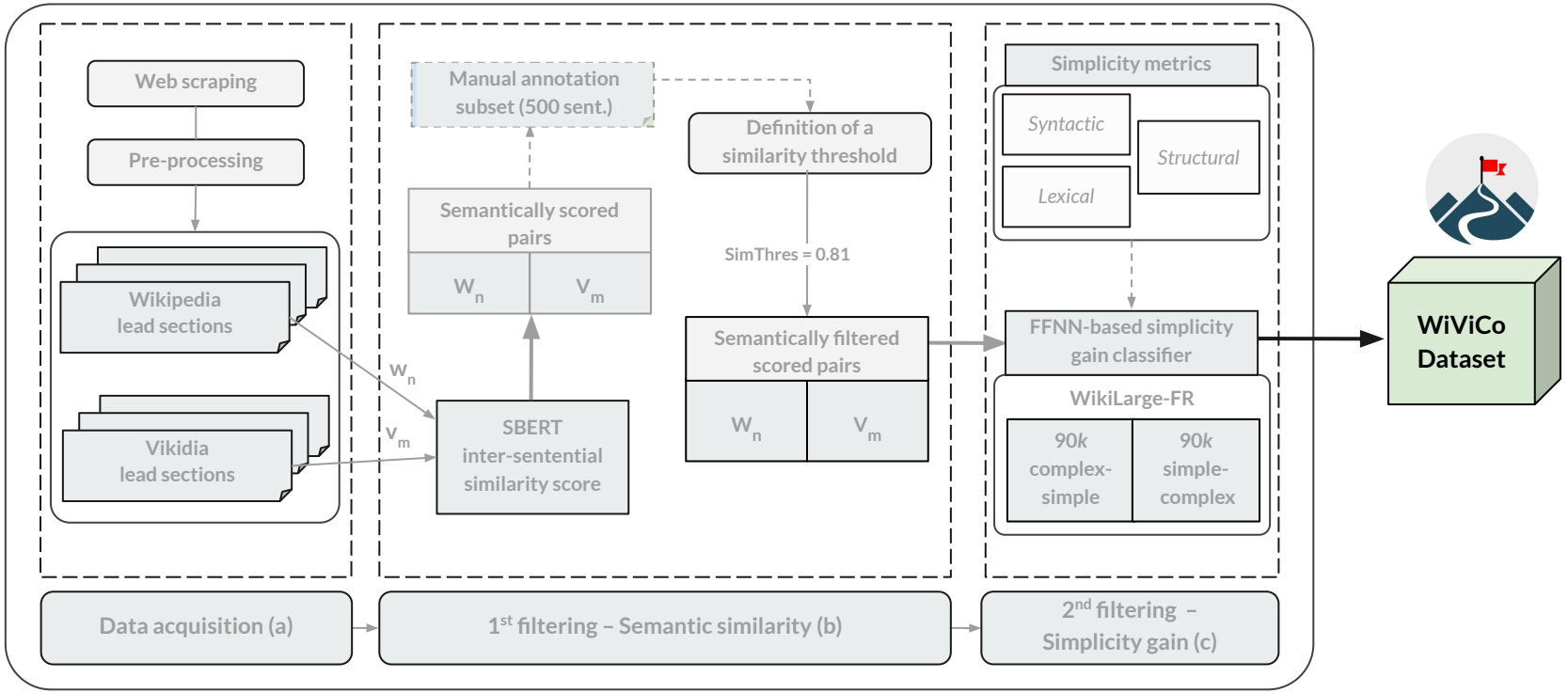
1.2. Stage II: Simplicity gain filtering

Feature Selection

Feature	Description
SL	Sentence Length
CEFR	Common European Framework of Reference score
NW	Number of Words
IDT	Incomplete Dependency Theory
IDTDLT	Combined IDT+DLT
MDT	Maximum Depth Tree
LE	Left Embeddedness
NNE	Number of Named Entities
BLEU	BLEU score
WER	Word Error Rate
NND	Noun Nested Distance

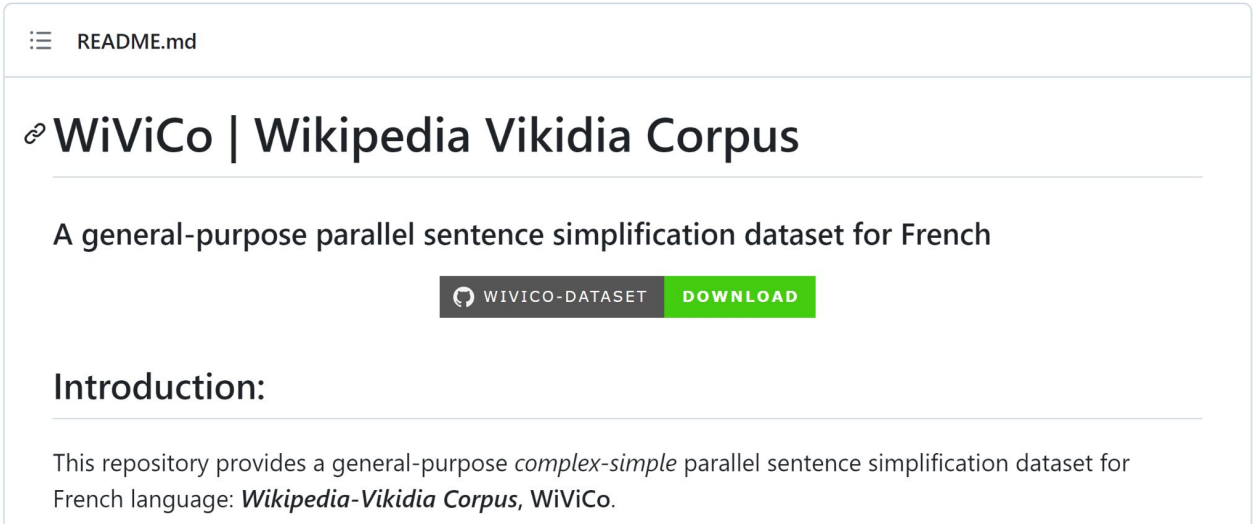
Classification Performance




Web scraping
Pre-processing
Wikipedia lead sections
Vikidia lead sections
Manual annotation subset (500 sent.)
Semantically scored pairs
W n
V m
SBERT inter-sentential similarity score
Definition of a similarity threshold
SimThres = 0.81
Semantically filtered scored pairs
W n
V m
Simplicity metrics
Syntactic
Structural
Lexical
FFNN-based simplicity gain classifier
WikiLarge-FR
90k complex-simple
90k simple-complex
WiViCo Dataset
Data acquisition (a)
1st filtering – Semantic similarity (b)
2nd filtering – Simplicity gain (c)

- With the implementation of the proposed two-step filtering method we created the **WiViCo monolingual parallel corpus.**
- Utilizing the sigmoid output layer of the classifier we provide simplification pairs based on **lenient** or **stricter thresholds**.
- Based on different **cutoff probability thresholds**, we enumerate all Wiki:Viki samples in each class:

	Label	
Probability	**0** (*non-simplified*)	**1** (*simplified*)
>0.9	44,049	20,692
>0.8	22,556	42,185
>0.7	18,642	46,099
>0.6	11,087	53,654
>0.5	7,302	57,439



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.

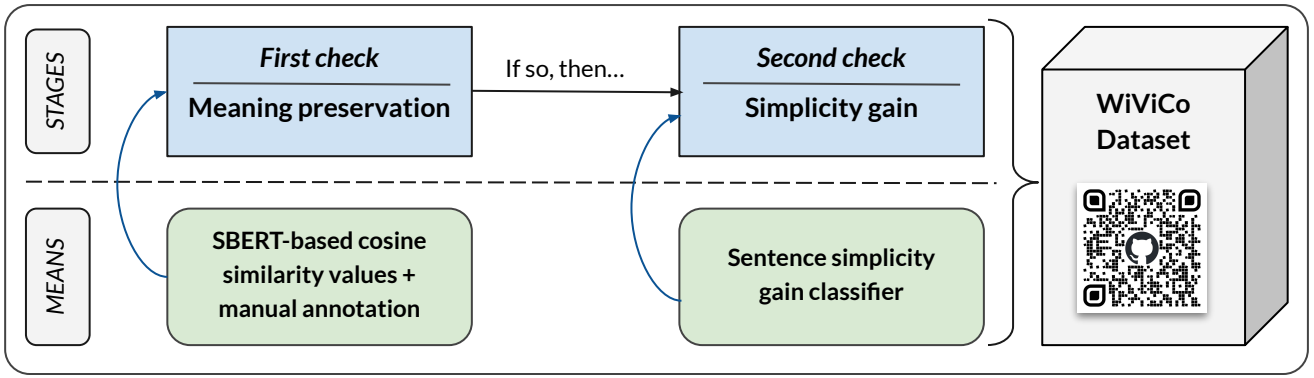
https://github.com/lormaechea/wivico

5. Conclusions

Main contribution:

Introduction of a **new method** to **mine comparable corpora**:

- Specifically **targeted** to **ATS**.
- **Sequential approach** to satisfy the **2 main conditions** for a **simplified text** to be **valid**:



5. Conclusions

Goals to be explored:

Once the data compilation is done, move on to the **generation stage**, by:

- **Training** general-domain ***seq2seq*** **ATS models** for **French**.
- **Fine-tuning LLMs** for our **downstream** task.

5. Conclusions

Goals to be explored:

Once the data compilation is done, move on to the **generation stage**, by:

- **Training** general-domain ***seq2seq*** **ATS models** for **French**.
- **Fine-tuning LLMs** for our **downstream** task.

Goals being currently explored:

- Introducing a **comprehensive approach** for **assessing sentence simplicity**.

SIMPLER **! =** SIMPLE

Simplification is **inherently** a **relative process** → A given text is transformed into a relatively **simpler** version, which does not necessarily equate to **simple**.

- Examining a **BERT-based fine-tuning approach** to **qualify** and **quantify sentence simplicity**.

Thanks for your attention!

Any questions?

 Lucia.OrmaecheaGrijalba@unige.ch

 Nikolaos.Tsourakis@unige.ch

References

[In order of appearance]

- **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