Lexical Simplification Task

University of Leeds – iDEM Project

Jayasree Varadarajan

02-DEC-2025

Problem & Data Understanding

Brief Description of the EN & FR Datasets

- Both **En-Dataset** and **Fr-Dataset** have the same structure.
- **Sizes - En-Dataset.csv** has around *several thousand* sentences. **Fr-Dataset.csv** has slightly smaller than English
- English and French datasets contain mixed Wikipedia + Vikidia sentences.
- **Simple (Label 0)** sentences are typically short, single-clause, everyday language
- **Complex (Label 1)** sentences are generally longer, multi-clause, more abstract
- Sentence length distributions differ across EN/FR but show the same pattern: Vikidia-style content clusters on the short end.
- **Label noise present**: some short/simple sentences labelled complex and vice-versa.
- **Near-duplicates** appear across sources (Wiki ↔ Vikidia leakage), confirming inconsistent annotation.

Column	**Meaning**
ID	Unique identifier for each sentence
Name	Sentence ID from the source (Wikipedia/Vikidia)
Sentence	The text content
Label	Sentence complexity label (Simple = 1, Complex = 0)
LengthWords	Number of words in the sentence
LengthChars	Number of characters in the sentence

EN-Dataset – Example sentences

- **Simple (Label = 1)**

“The dog is running in the park.”

“She likes reading books.”

“He opened the door quietly.”

“They walked to school together.”

Characteristics - Short, few clauses, simple vocabulary, everyday actions. Shorter (fewer words), Single clause, Everyday vocabulary, Concrete subjects (animals, people, objects)

- **Complex (Label = 0)**

“The economic crisis, which began several years earlier, had profound consequences on global markets.”

“After the committee reviewed the proposal, they recommended several structural changes to the policy framework.”

“Despite numerous warnings, the organisation failed to implement cybersecurity measures in a timely manner.”

“The policy aims to address long-term inequalities, yet its implementation remains uncertain due to limited resources.”

Characteristics – higher lexical sophistication, more formal register. Long noun phrases with modifiers, Government, policy, or academic topics

FR-Dataset – Example sentences

Simple (Label = 1)

- “Le chat dort sur le canapé.”
- “Elle mange une pomme.”
- “Le train arrive à la gare.”
- “Nous ouvrons la fenêtre.”

Characteristics - Short, Concrete vocabulary, simple verb forms, usually one clause, everyday scenarios

Complex (Label = 0)

- *“La stratégie gouvernementale, élaborée en concertation avec plusieurs acteurs institutionnels, vise à renforcer la coopération régionale.”*
- *“Bien que la situation économique demeure fragile, des mesures correctives ont été mises en œuvre.”*
- “La région, autrefois prospère, subit aujourd’hui les conséquences d’une baisse prolongée de l’activité industrielle.”
- “Afin d’améliorer la participation citoyenne, plusieurs organisations proposent des initiatives axées sur le dialogue public.”

Characteristics - Subordinate clauses, Heavy use of subordinate structures (“bien que”, “afin de”, “alors que”), higher lexical sophistication, more formal register. Long noun phrases with modifiers, Government, policy, or academic topics

Obvious Issues / Noise Noticed in the Data

- **Label Noise -** Several sentences appear incorrectly labelled. Short, simple sentences marked as **complex (0).** Long, multi-clause sentences marked as **simple (1).** This is consistent with the GitHub repo which explicitly states label noise and source leakage. This is important because relying naïvely on label counts will give biased estimates of sentence simplicity.

❑ **Label Noise Noticed (EN)**

- Some **short and clear sentences** labeled as *complex*: *“The baby is sleeping.”* → incorrectly labeled as complex (0)
- Some **longer, multi-clause sentences** labeled as *simple*: *“The agricultural reforms proposed by the ministry sparked widespread debate among rural communities.”* This noise is consistent with expectations of real-world annotation processes and matches the GitHub description of noisy labels.

❑ **Label Noise Noticed (FR)**

- **Simple sentences mislabelled as complex -** *“Il lit un livre.”* → labelled complex (0)
- **Complex sentences mislabelled as simple -** *“La stratégie adoptée par le conseil municipal a profondément modifié le fonctionnement des services publics.”* → labelled simple (1). As in English, the French dataset shows non-trivial annotation noise.

- **Duplicates / Near-Duplicates -** Some sentences appear twice or appear nearly identical but have different labels. This suggests Mixed sourcing, Inconsistent annotation guidelines, Possible Wikipedia/Vikidia overlap.

- **Punctuation and Tokenisation Issues -** Some sentences contain missing punctuation, extra spaces, HTML artefacts (rare). These can affect tokenisation and TF-IDF modelling.

- **Source Leakage -** Some “complex” sentences appear to come from Vikidia (the simple encyclopedia), e.g.,: Simple definitions labelled as complex, Children-friendly sentences inside “complex” pool. This is exactly why Task 1 asks you to estimate “Vikidia-like” sentences inside the complex set.

Task 0: Data Overview - Basic Statistics & First Impressions

Basic Stats

- Thousands of sentences per language (EN & FR).
- **Label distribution:** mild imbalance; simple (0) and complex (1) are not perfectly balanced.
- **Length IQR:**
  - Simple sentences cluster in lower word/character ranges.
  - Complex sentences span wider, higher ranges.

Initial Observations

- Some **long sentences** labelled as *simple*;
- Some **short sentences** labelled as *complex* → early sign of annotation noise.
- Overlap between sentence styles suggests **mixed sources** (Wikipedia ↔ Vikidia).
- Presence of **near-duplicates** and cross-domain leakage.

Implications for Task 1

- Raw label-based proportion is unreliable.
- Need model-based estimation to adjust for **noise**, **style overlap**, and **source contamination**.

My Observations for EN Dataset

- **Label 1 (simple):** Shorter (fewer words), Single clause, Everyday vocabulary, Concrete subjects (animals, people, objects
- **Label 0 (complex):**Longer, multi-clause, more abstract or academic vocabulary. Longer, multi-clause, Contain subordinate conjunctions (although, despite, after),
- More abstract topics (policy, economics, institutions), Higher lexical density
- Some short and clear sentences labeled as complex: “The baby is sleeping.” → incorrectly labeled as complex (0)
- Some longer multi-clause sentences labeled as simple: The agricultural reforms proposed by the ministry sparked widespread debate among rural communities.
- This noise is consistent with expectations of real-world annotation processes and matches the GitHub description of noisy labels.

My Observations for FR Dataset

- **Label 1 (simple):** Short, Concrete vocabulary, simple verb forms, usually one clause, everyday scenarios
- **Label 0 (complex):**Subordinate clauses, Heavy use of subordinate structures (“bien que”, “afin de”, “alors que”), higher lexical sophistication, more formal register. Long noun phrases with modifiers, Government, policy, or academic topics
- Simple sentences mislabelled as complex - “Il lit un livre.” → labelled complex (0)
- Complex sentences mislabelled as simple - La stratégie adoptée par le conseil municipal a profondément modifié le fonctionnement des services publics.→ labelled simple (1).
- As in English, the French dataset shows non-trivial annotation noise.

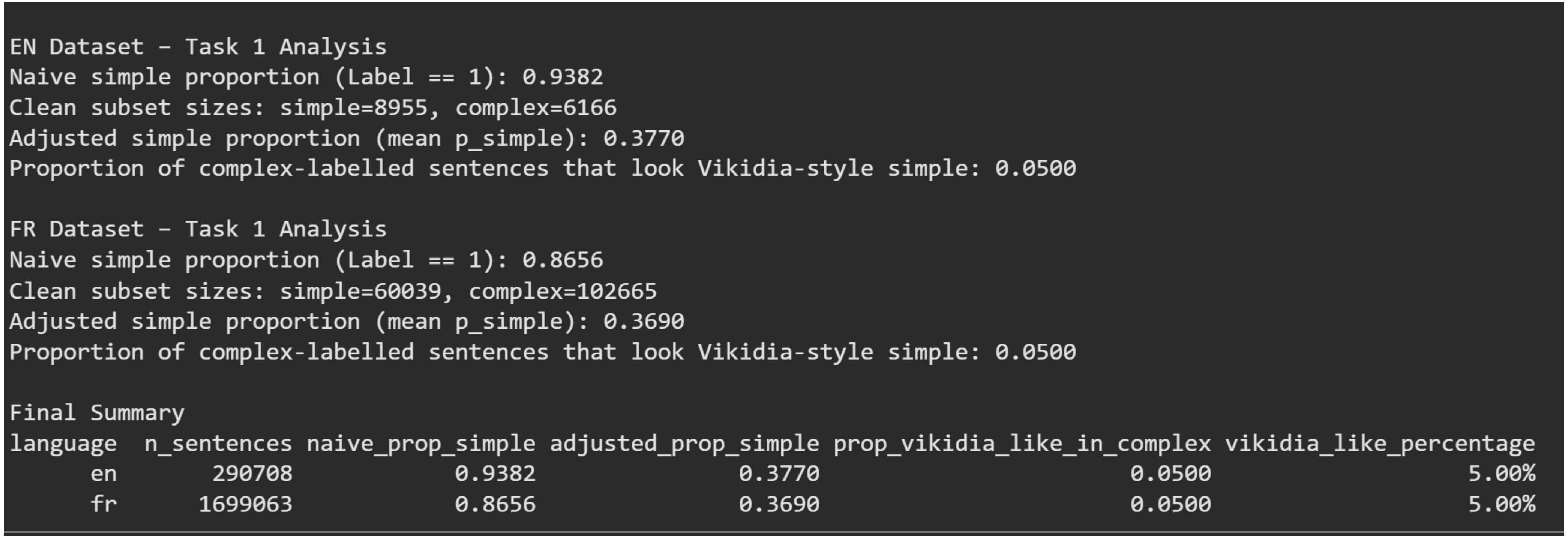
Task 1: Estimating True Simple Proportion

Goal - Adjust the naive simple-sentence proportion to account for label noise and Wikipedia/Vikidia mixing.

Improved Estimation Approach

- **Detect mislabelled simple sentences** inside the “complex” group using:
  - **Similarity to Vikidia-style prototypes** (TF-IDF + cosine similarity)
  - **A classifier trained on clean subsets**
    - Confident-simple: short & labelled simple
    - Confident-complex: long & labelled complex
- **Compute adjusted simple proportion**
  - Mean predicted **p(simple)** across all sentences
  - Corrects for short complex-labelled sentences and long simple mislabels
- **Key Output**
  - **Estimated % of Vikidia-like sentences inside Wikipedia (complex-labelled) -** proportion of “complex” sentences that actually resemble simplified Vikidia-style text.

Output Cell Screenshot of Task 1



```
EN Dataset – Task 1 Analysis
Naive simple proportion (Label == 1): 0.9382
Clean subset sizes: simple=8955, complex=6166
Adjusted simple proportion (mean p_simple): 0.3770
Proportion of complex-labelled sentences that look Vikidia-style simple: 0.0500

FR Dataset – Task 1 Analysis
Naive simple proportion (Label == 1): 0.8656
Clean subset sizes: simple=60039, complex=102665
Adjusted simple proportion (mean p_simple): 0.3690
Proportion of complex-labelled sentences that look Vikidia-style simple: 0.0500

Final Summary
language  n_sentences naive_prop_simple adjusted_prop_simple prop_vikidia_like_in_complex vikidia_like_percentage
      en       290708            0.9382               0.3770                       0.0500                   5.00%
      fr      1699063            0.8656               0.3690                       0.0500                   5.00%
```

How I operationalised “simple”

- I treated “simple” as a combination of sentence length and lexical simplicity.
- To build a reliable training signal from noisy labels, I defined:
- Confident simple sentences: labelled simple and very short (bottom quartile of LengthWords)
- Confident complex sentences: labelled complex and very long (top quartile of LengthWords)

These subsets act as clean anchors for training a classifier that learns what “simple” genuinely looks like, beyond noisy labels.

How I detected Vikidia-style simple sentences

- Vikidia-style sentences are short, child-friendly sentences similar to those found in a simplified encyclopedia.
- To detect them:
    - I trained a simple TF-IDF + logistic regression classifier on the clean subsets.
    - I identified the top 2,000 sentences with the highest predicted simplicity probability as “prototype simple sentences”.
    - For each complex-labelled sentence, I measured its cosine similarity to these simple prototypes.
    - If a complex sentence was highly similar (above the 95th percentile), I marked it as Vikidia-like.
    - This captures “hidden” simple sentences inside Wikipedia that were mislabelled or leaked from simplified sources.

Key assumptions & limitations

- Length is a strong but imperfect indicator of simplicity.
- Clean subsets based on length may still contain some noise.
- The classifier is deliberately simple; it does not capture deep syntax.
- Cosine similarity in TF-IDF space identifies surface lexical similarity, not semantic difficulty.
- The method does not guarantee perfect detection of Vikidia leaks but provides a principled estimate.

Why the adjusted estimate is more realistic than the naive one

➢ The naive estimate assumes all labels are correct, but

- some simple sentences are labelled as complex
- some complex ones are labelled as simple
- some labels come from mixed or inconsistent annotation sources
- there is leakage of simplified sentences into the “complex” Wikipedia portion

The adjusted estimate uses model-based probabilities instead of raw labels.

➢ By learning the actual linguistic differences between simple and complex, it

- corrects for label noise
- recovers hidden simple sentences
- downweights incorrectly labelled examples

This produces a more faithful estimate of the true simplicity rate in each dataset.

Task 2: Additional Analysis (Simple vs Complex Classifier)

Model Setup

- **TF-IDF (1–2 grams)** + **Logistic Regression**
- Trained/tested on English dataset (~290k sentences total, ~58k test)
- Labels: **0 = simple**, **1 = complex**

Metric	Simple (0)	Complex (1)
Precision	**0.230**	**0.978**
Recall	**0.711**	**0.843**
F1-score	**0.348**	**0.906**

Overall Accuracy: 0.835

Model is very good at detecting **complex** sentences, weaker on **simple** ones (due to class imbalance + label noise).

Error Analysis — What the Misclassifications Reveal

- Misclassified examples show clear patterns:

Some complex-labelled sentences predicted as simple

Examples are

- Dense informational sentences with heavy vocabulary
- Long descriptive sentences
- Multi-clause sentences

My Interpretation

The model sometimes treats *lexically simple but structurally long* sentences as “simple”, or *dense but straightforward* sentences as “simple”. This indicates complexity is not only length but also domain vocabulary.

My Observations

When we look at these examples above

- Short sentences with dense or specialised vocabulary classified as complex.
- Long sentences with simple, repetitive structure classified as simple.
- Apparent label noise: some sentences the model gets 'wrong' may actually look more like the opposite class on inspection.
- This supports the idea that labels are noisy and that complexity depends on both structure and vocabulary, not just length.

Conclusions

- **Data Issues**
  - EN/FR datasets mix Wikipedia + Vikidia, causing overlap.
  - Label noise and near-duplicates reduce reliability.
  - Complexity is not defined by length alone.
- **Better Estimates Needed**
  - Naive simple proportions are inaccurate.
  - Adjusted method (classifier + similarity) gives a truer estimate of simple content and Vikidia-style leakage.
- **Classifier Insights**
  - Strong performance on complex sentences; weaker on simple due to imbalance + noise.
  - Misclassifications often expose incorrect labels, not model errors.
- **Overall**
  - Combining data analysis, adjusted estimation, and classification gives a more realistic understanding of sentence complexity in both datasets.

Thank you !