# Slide 1 — Data Understanding (EN & FR)

**Define simple and complex?**

1. Domain knowledge – Reader’s background
2. Sentence structure
3. Is a long sentence must be a complex sentence?
4. Vikidia is shorter/simpler?
5. Wikipedia varies more?
6. French dataset

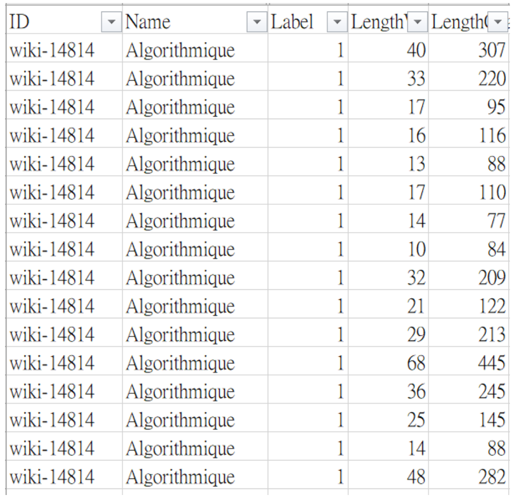
## 6.

Differences between Prepositions and Articles:

Double or Triple Preposition and Article Combinations

English: the man's book (4 words)

French: le livre de l'homme (5 words)



| ID | Name | Label | Length' | Length( |
|---|---|---|---|---|
| wiki-14814 | Algorithmique | 1 | 40 | 307 |
| wiki-14814 | Algorithmique | 1 | 33 | 220 |
| wiki-14814 | Algorithmique | 1 | 17 | 95 |
| wiki-14814 | Algorithmique | 1 | 16 | 116 |
| wiki-14814 | Algorithmique | 1 | 13 | 88 |
| wiki-14814 | Algorithmique | 1 | 17 | 110 |
| wiki-14814 | Algorithmique | 1 | 14 | 77 |
| wiki-14814 | Algorithmique | 1 | 10 | 84 |
| wiki-14814 | Algorithmique | 1 | 32 | 209 |
| wiki-14814 | Algorithmique | 1 | 21 | 122 |
| wiki-14814 | Algorithmique | 1 | 29 | 213 |
| wiki-14814 | Algorithmique | 1 | 68 | 445 |
| wiki-14814 | Algorithmique | 1 | 36 | 245 |
| wiki-14814 | Algorithmique | 1 | 25 | 145 |
| wiki-14814 | Algorithmique | 1 | 14 | 88 |
| wiki-14814 | Algorithmique | 1 | 48 | 282 |

Gender and Number Agreement

English: a small table

**French:** une petite table (same number of words, but French has more letters)

Verbal Structure:

Conjugation often adds extra letters or particles.

English verb forms are relatively stable (e.g., past tense usually only adds -ed)

## 1.2.3.4.5.

- viki-100 Scotland The majority voted to remain as part of the United Kingdom.
- viki-1004 Amsterdam Many people describe Amsterdam as a English speaking city because a person who visit Amsterdam hears a lot of English and not so much Dutch and sometimes only English.
- viki-1 Bulbasaur In Pokemon Red, Pokemon Blue, Pokemon Green, Pokemon FireRed, Pokemon WaterBlue and Pokemon LeafGreen, you can only receive it at the start, from Professor Oak.
- wiki-1033 Sony Sony acquired Ericsson's share of the venture in 2012 for over US$1 billion.
- wiki-1033 Sony Sony later pulled the ads, suspended Manning's creator and his supervisor and paid fines to the state of Connecticut and to fans who saw the reviewed films in the US.
- wiki-1016 Computer memory In 1967, Dawon Kahng and Simon Sze of Bell Labs proposed that the floating gate of a MOS semiconductor device could be used for the cell of a reprogrammable read-only memory  (ROM), which led to Dov Frohman of Intel inventing EPROM  (erasable PROM) in 1971.

# Slide 2 — Task 0: Quick Overview

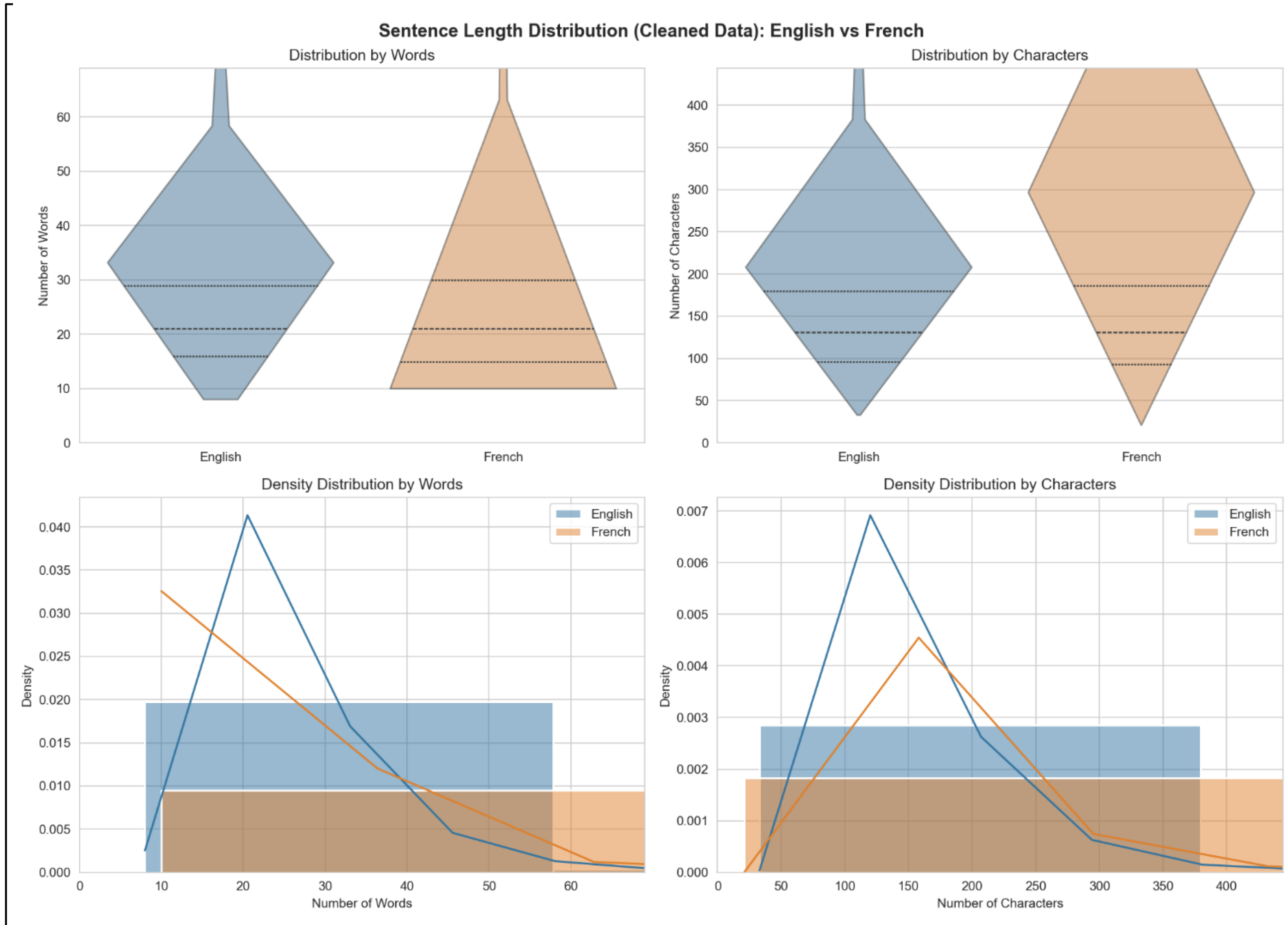
- Dataset balance?
- Why Quartile?
- Length Words or Length Chars?
- Duplicate, Nan and 0?
- Label noise? Leakage?

**Official Label Definition (from README):**

- Label = 0: Simple (sentence annotated as simple, from Vikidia)
- Label = 1: Complex (sentence annotated as complex, from Wikipedia)

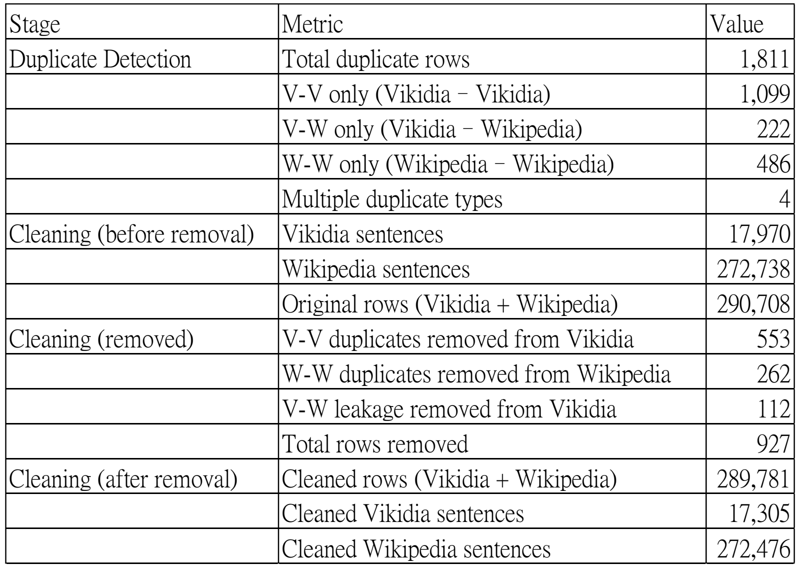
## Dataset Summary Statistics (Cleaned Data)

| Dataset | Total Sentences | Simple (Label=0) | Complex (Label=1) | IQR (all) | IQR (Simple) | IQR (Complex) |
|---|---|---|---|---|---|---|
| English | 289,781 | 17,305 (5.97%) | 272,476 (94.03%) | Q1=16, Q3=29 | Q1=13, Q3=22 | Q1=16, Q3=29 |
| French | 1,653,175 | 188,856 (11.42%) | 1,464,319 (88.58%) | Q1=15, Q3=30 | Q1=14, Q3=25 | Q1=15, Q3=30 |


Sentence Length Distribution (Cleaned Data): English vs French

wiki-968 Belabo Upazila Here the 2,500-year-old civilisation of Wari Bateshwari has been discovered. 1-Complex

viki-1263 Christianity Jesus lived a perfect life without sin, and taught his followers many things about right and wrong, and also performed miracles like healing certain people who were blind, deaf, or sick in other ways. 0-Simple



| Stage | Metric | Value |
|---|---|---|
| Duplicate Detection | Total duplicate rows | 1,811 |
| | V-V only (Vikidia – Vikidia) | 1,099 |
| | V-W only (Vikidia – Wikipedia) | 222 |
| | W-W only (Wikipedia – Wikipedia) | 486 |
| | Multiple duplicate types | 4 |
| Cleaning (before removal) | Vikidia sentences | 17,970 |
| | Wikipedia sentences | 272,738 |
| | Original rows (Vikidia + Wikipedia) | 290,708 |
| Cleaning (removed) | V-V duplicates removed from Vikidia | 553 |
| | W-W duplicates removed from Wikipedia | 262 |
| | V-W leakage removed from Vikidia | 112 |
| | Total rows removed | 927 |
| Cleaning (after removal) | Cleaned rows (Vikidia + Wikipedia) | 289,781 |
| | Cleaned Vikidia sentences | 17,305 |
| | Cleaned Wikipedia sentences | 272,476 |

# Slide 3 — Task 1: Estimating True Simple Proportion

## 1. Naïve Estimate

| Language | Total sentences | Simple (Label=0) | Naïve simple proportion |
|---|---|---|---|
| EN | 289,781 | 17,305 | 0.0597 (5.97%) |
| FR | 1,653,175 | 188,856 | 0.1142 (11.42%) |

## 2. Anchor-Based Training:

Simple anchors: Vikidia sentences (Label=0), LengthWords ≤ Q1;
Complex anchors: Wikipedia sentences (Label=1), LengthWords ≥ Q3.
Trains LR and RF classifiers on these "clean" extremes.
Full dataset features, e.g.:words_chars_ratio, cos_simi, avg_word_len, long_word_ratio, ttr(Type–Token Ratio), punct_density, comma_density, digit_ratio, upper_ratio, has_parens, n_tokens ,max_depth, avg_depth, avg_dependency_distance, func_word_ratio, n_clauses, clause_ratio, noun_ratio,verb_ratio

3. **ACC Calibration:** Uses cross-validated TPR/FPR to correct classifier bias Formula: **p_true = (p_pred - FPR) / (TPR - FPR)**

Estimate prevalence with ACC (Adjusted Classify & Count): Apply the trained model to all sentences to get P(simple | x).Compute the raw predicted simple proportion p_pred = mean(1[P ≥ 0.5]).Correct for classifier bias with the ACC formula:
p_true = (p_pred - FPR) / (TPR - FPR)
This gives an **adjusted estimate of the true simple proportion** in each dataset.

## Wikipedia internal:

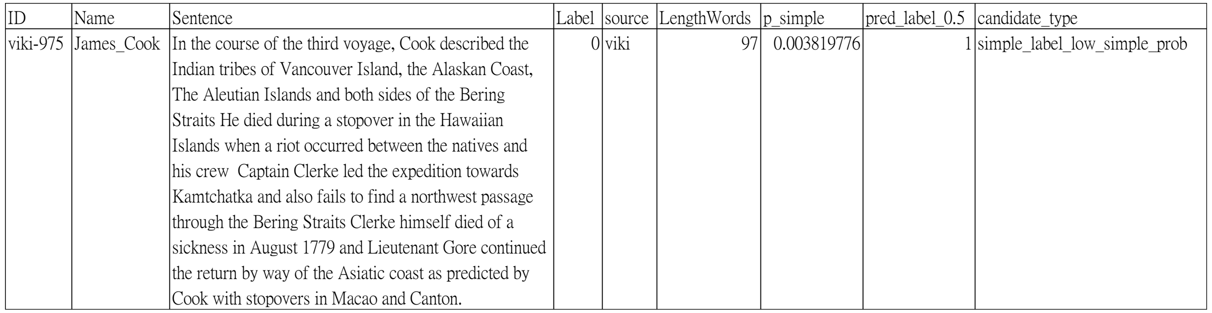
| Metric | English | French |
|---|---|---|
| Naïve simple proportion (Label=0 / Total) | 0.0597 | 0.1142 |
| ACC-corrected true simple proportion | **0.6336** (63.36%) | **0.6264** (62.64%) |

## Vikidia-style sentences inside Wikipedia subset:

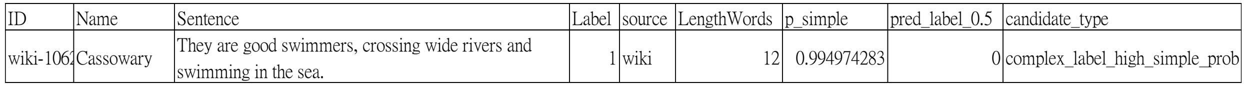
For each, use P(simple | x) from the model as a “Vikidia-likeness” score.Two estimates:
Soft estimate = mean P(simple) over Wikipedia sentences.
Hard estimate = proportion of Wikipedia sentences with P(simple) ≥ 0.8.

| Metric | English | French |
|---|---|---|
| Wikipedia internal (soft estimate) | 0.5723 | 0.5668 |
| Wikipedia internal (hard, P(simple) ≥ 0.8) | 0.4894 | 0.4923 |

Mislabel candidates_en:



| ID | Name | Sentence | Label | source | LengthWords | p_simple | pred_label_0.5 | candidate_type |
|---|---|---|---|---|---|---|---|---|
| viki-975 | James_Cook | In the course of the third voyage, Cook described the Indian tribes of Vancouver Island, the Alaskan Coast, The Aleutian Islands and both sides of the Bering Straits He died during a stopover in the Hawaiian Islands when a riot occurred between the natives and his crew  Captain Clerke led the expedition towards Kamtchatka and also fails to find a northwest passage through the Bering Straits Clerke himself died of a sickness in August 1779 and Lieutenant Gore continued the return by way of the Asiatic coast as predicted by Cook with stopovers in Macao and Canton. | 0 | viki | 97 | 0.003819776 | 1 | simple_label_low_simple_prob |



| ID | Name | Sentence | Label | source | LengthWords | p_simple | pred_label_0.5 | candidate_type |
|---|---|---|---|---|---|---|---|---|
| wiki-106 | Cassowary | They are good swimmers, crossing wide rivers and swimming in the sea. | 1 | wiki | 12 | 0.994974283 | 0 | complex_label_high_simple_prob |

# Slide 4 — Task 2: Additional Analysis

| Configuration | Value |
|---|---|
| Model | DistilBERT (distilbert-base-uncased) |
| GPU | Tesla P100-PCIE-16GB |
| Dataset | English (289,781 sentences) |
| Training | Anchor-based + Downsampling |
| Dataset Spilit | Train 80%, Val 10%, Test 10% |
| Anchor test size | 2,632 sentences (878 simple, 1,754 complex). |

| Anchor Test Set Performance | |
|---|---|
| **Metric** | **Score** |
| Accuracy | 1.0000 (100%) |
| F1 (weighted) | 1 |
| ROC-AUC | 1 |
| TPR | 1 |
| FPR | 0 |

| Confusion matrix (anchors): | Pred Simple (0) | Pred Complex (1) |
|---|---|---|
| **True Simple 0** | 878 | 0 |
| **True Complex 1** | 0 | 1,754 |

| Full Corpus Error Analysis | |
|---|---|
| **Metric** | **Value** |
| Total Errors | 154,526 / 289,781 (53.33%) |
| False Positives (Simple→Complex) | 3,372 (100% from Vikidia) |
| False Negatives (Complex→Simple) | 151,154 (100% from Wikipedia) |

Simple anchors: short Vikidia sentences (Label=0, LengthWords ≤ 16, 8,773 samples).Complex anchors: long Wikipedia sentences (Label=1, LengthWords ≥ 29, 73,032 → downsampled to 17,546).

Final anchor set: 26,319 sentences (ratio ≈ 2:1 complex:simple).Split: 80% train / 10% val / 10% test (stratified).

| Global Sentence Complexity | | |
|---|---|---|
| Estimate Method | Simple (Label=0) | Complex (Label=1) |
| Naïve (from Label) | 5.97% | 94.03% |
| Predicted (P ≥ 0.5) | 56.97% | 43.03% |
| ACC-corrected | 56.97% | 43.03% |

| Vikidia Internal Analysis (Source=viki) | |
|---|---|
| Metric | Value |
| Total sentences | 17,305 |
| Soft estimate (mean P(Complex)) | 19.49% |
| Predicted Complex (P ≥ 0.5) | 19.49% |
| High confidence Complex (P ≥ 0.9) | 18.99% |

| Wikipedia Internal Analysis (Source=wiki) | |
|---|---|
| Metric | Value |
| Total sentences | 272,476 |
| Soft estimate (mean P(Simple)) | 55.47% |
| Predicted Simple (P ≥ 0.5) | 55.47% |
| High confidence Simple (P ≥ 0.9) | 54.82% |

| Key Insight | |
|---|---|
| **Finding** | **Interpretation** |
| Naïve Simple = 5.97% → Predicted Simple = 56.97% | Model finds ~57% of sentences have "simple" characteristics |
| ~55% Wikipedia predicted as Simple | Many Wikipedia sentences are linguistically simple |
| ~19% Vikidia predicted as Complex | Some Vikidia sentences are actually complex |

## ERROR ANALYSIS ON FULL CORPUS
## Total errors: 154,526 / 289,781 (53.33%)

**[1] FALSE POSITIVES (pred=Complex, true=Simple)**

Total FP: 3,372
High confidence FP (P(Complex) >= 0.9): 3,287
These Simple sentences were predicted as Complex
By source: Wikipedia=0, Vikidia=3,372

FP-1 [P(Complex)=1.000] [Source=viki] [Len=39]:
The secondary waves are recorded by the seismograph for the second because they have lower speed to the first and also the transverse, as they vibrate...

FP-2 [P(Complex)=1.000] [Source=viki] [Len=34]:
However, rather that settle into a legal career he became more involved in (talk or information that tries to change people's minds) efforts, and the ...

**[2] FALSE NEGATIVES (pred=Simple, true=Complex)**

Total FN: 151,154
High confidence FN (P(Simple) >= 0.9): 149,382
These Complex sentences were predicted as Simple
By source: Wikipedia=151,154, Vikidia=0

FN-1 [P(Simple)=1.000] [Source=wiki] [Len=14]:
The famous painting of Juan Luna, the Spoliarium, can be found in the complex....

FN-2 [P(Simple)=1.000] [Source=wiki] [Len=14]:
Beginning in 1834, it visited the American Rendezvous to buy furs at low prices....

**[3] BOUNDARY CASES (P(Simple) in (0.45 c, 0.55 s)) Total boundary: 330**

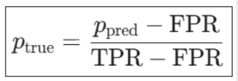
B-1 [P(Simple)=0.500] [CORRECT] [True=Complex]: Anguish and grief, like darkness and rain, may be depicted; but gladness and joy, like the rainbow, defy the skill of pen or pencil....

B-2 [P(Simple)=0.500] [CORRECT] [True=Complex]: Italian immigrants brought New York-style pizza and Italian cuisine into the city, while Jewish immigrants and Irish immigrants brought pastrami and c...

# Slide 5 — Takeaways

## Learnings:

**1. Prevalence / estimation concepts**



$$p_{\text{true}} = \frac{p_{\text{pred}} - \text{FPR}}{\text{TPR} - \text{FPR}}$$

- Naïve estimate
- True prevalence / proportion
- Soft / Hard estimation
- ACC (Adjusted Classify and Count) calibration

**2. Learning / modelling strategy**

- Anchor-Based Semi-Supervised Fine Tuning (Inductive Semi-Supervised Learning (SSL))

**3. Data / label quality**

- Label noise / Label bias

**4. Uncertainty & decision boundary analysis**

- Boundary cases (P(Simple) near 0.5)

**5. Feature extraction & linguistic tooling**

- TextDescriptives

**6. Technique**

- Github release

## Limitations:

**1. Dataset & annotation**

- Data and label assumptions.
- No human ground truth complexity.
- Cross-lingual comparability.

**2. Preprocessing & corpus cleaning**

- Duplicates and leakages (V–V, W–W, V–W) are detected by exact string match.
- Cleaning and basic statistics: spelling errors, broken markup, strange tokenisation, non-sentential fragments, outliers.

**3. Representation / feature**

- Features focus on surface and syntactic properties; semantic difficulty, idiomaticity, or conceptual density are not explicitly encoded.
- DistilBERT is trained on general language (domain mismatch for “simplicity”/pedagogical style).

**4. Modelling / training-regime**

- Anchors are drawn from extremes.
- Less reliable for medium-length, mixed signals (the majority of the corpus).

**5. Estimation & calibration**

- ACC Calibration Assumptions Violated: Stable TPR/FPR.

**6. Evaluation & external validity**

- No comparison with established readability metrics (Flesch-Kincaid, Lix).