SELEXINI

# Statistical measure of representativeness applied to knowledge graphs and corpora

**Valentin Nyzam and Arnaud Soulet**

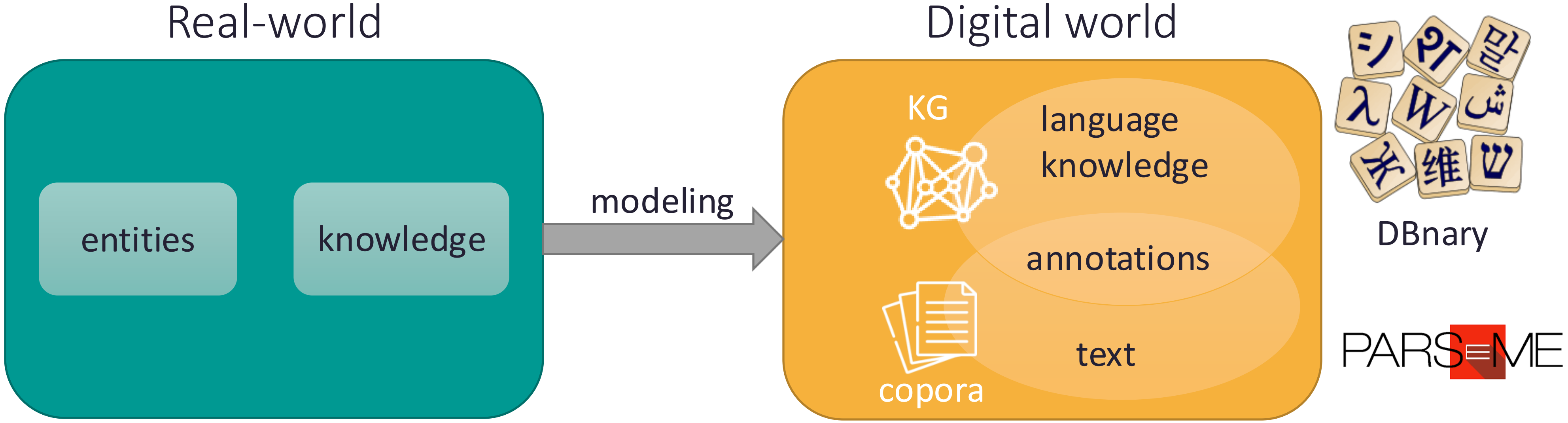
**Université de Tours**

université de TOURS

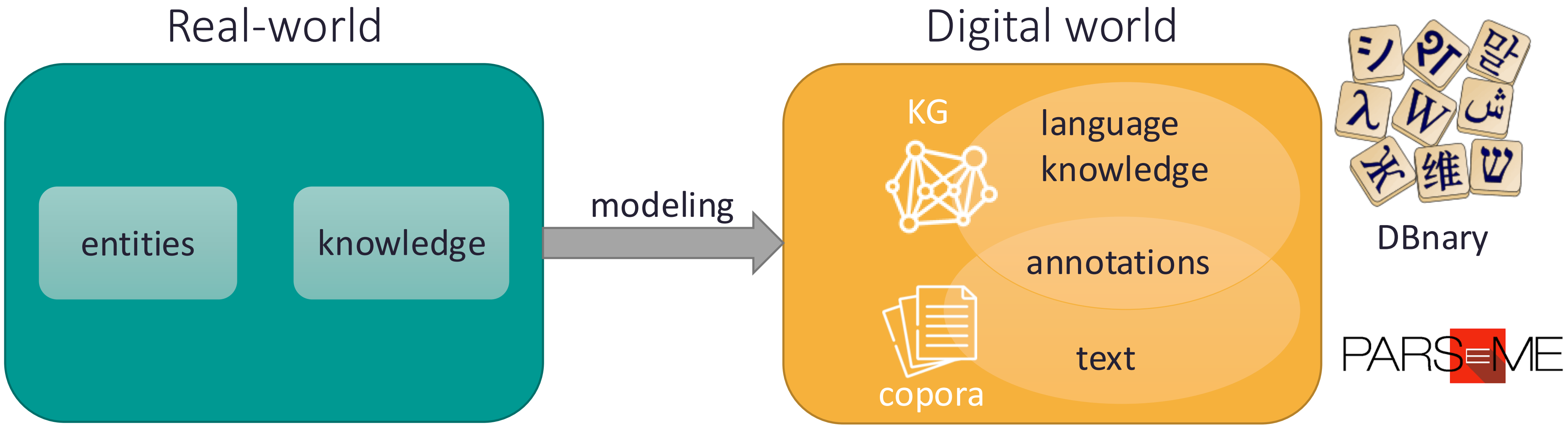
LIFAT EA 6300

**Paris – 19/06/2024**

# What's the link between knowledge graphs and corpora?



# What's the link between knowledge graphs and corpora?



**WP5: Estimating and correcting the diversity of a corpus**

# What kind of diversity?

**3 dimensions**

❑ **Variety :** 35k animals

❑ **Disparity :** 800 species

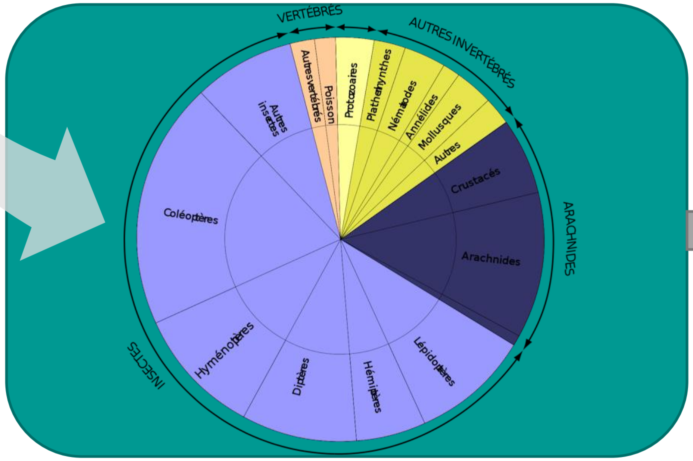
❑ **Balance :** a few individuals for each species

*[Morales et al. 2021. Measuring diversity in heterogeneous information networks. Theoretical Computer Science, 859:80–115]*

# What kind of diversity?

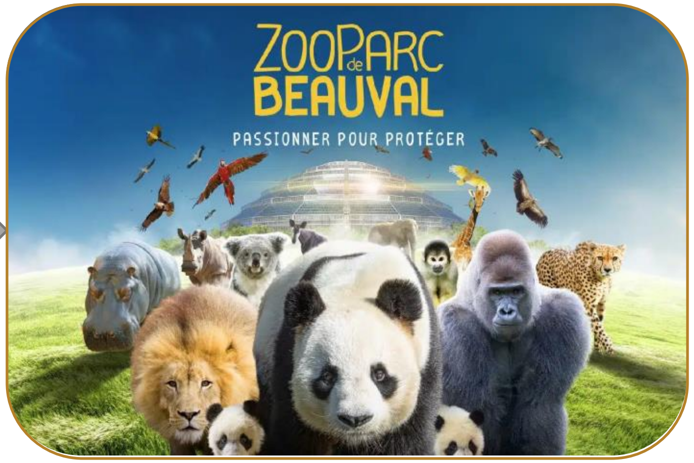
*Where have the insects gone?*

Real-world



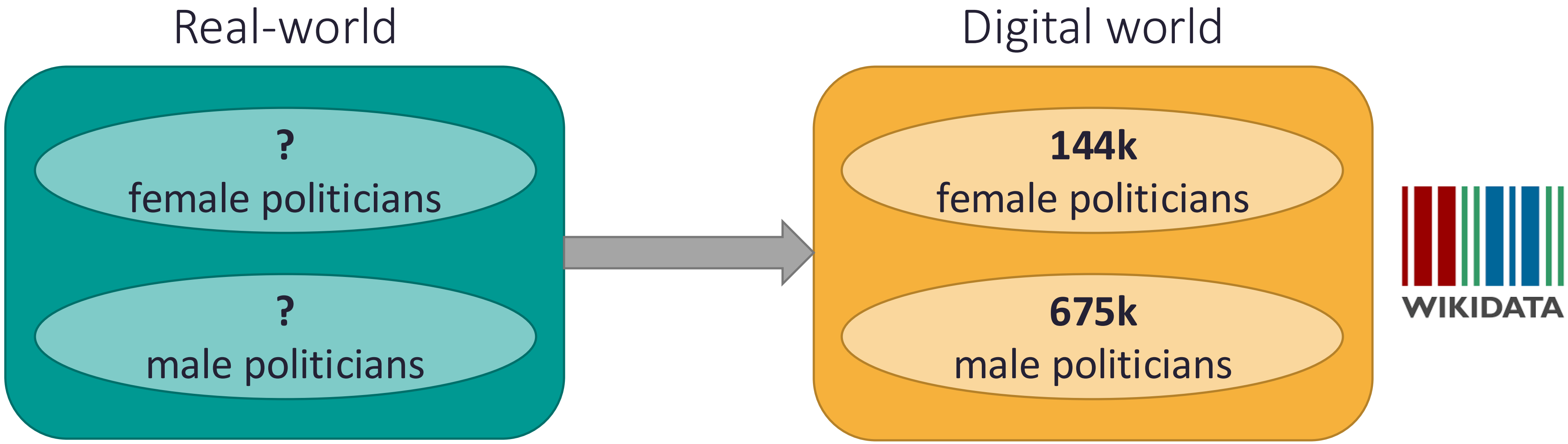
high selection bias

Digital world



**Our goal: Measuring the selection bias between the real-world and the digital world**

# Why its so important to measure the selection bias?



Real-world

?
female politicians

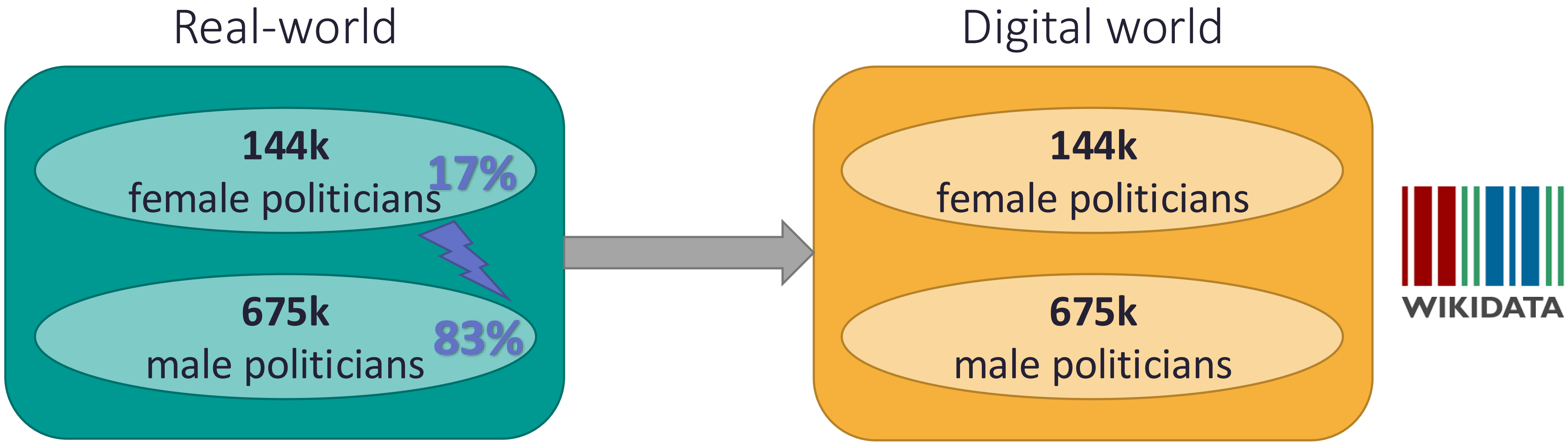
?
male politicians

Digital world

144k
female politicians

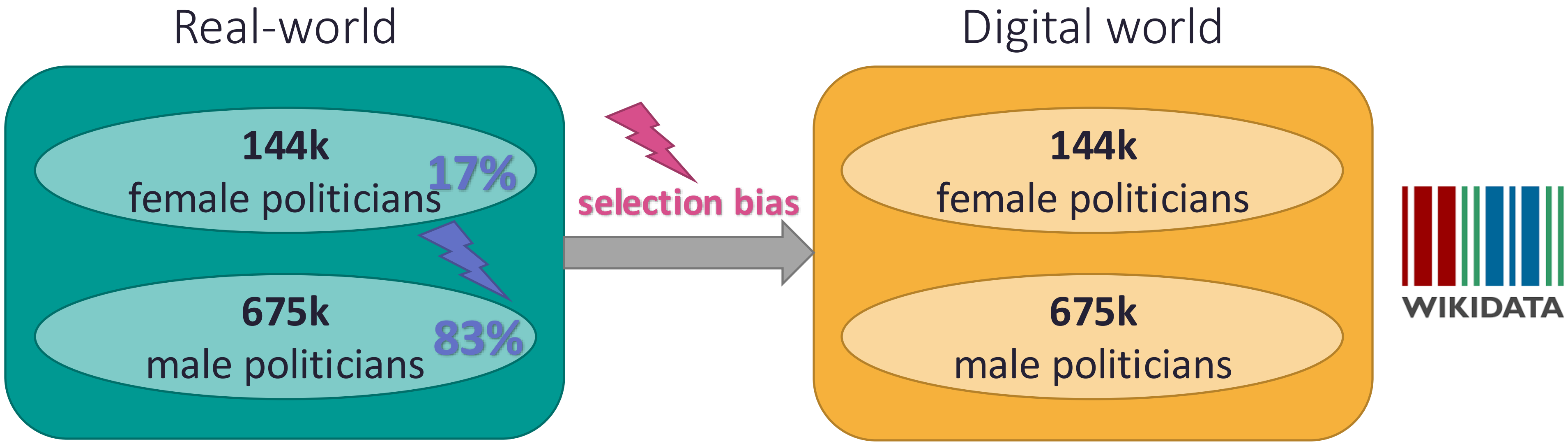
675k
male politicians

# Why its so important to measure the selection bias?



**The data are correct, there is a gender gap in reality**

# Why its so important to measure the selection bias?

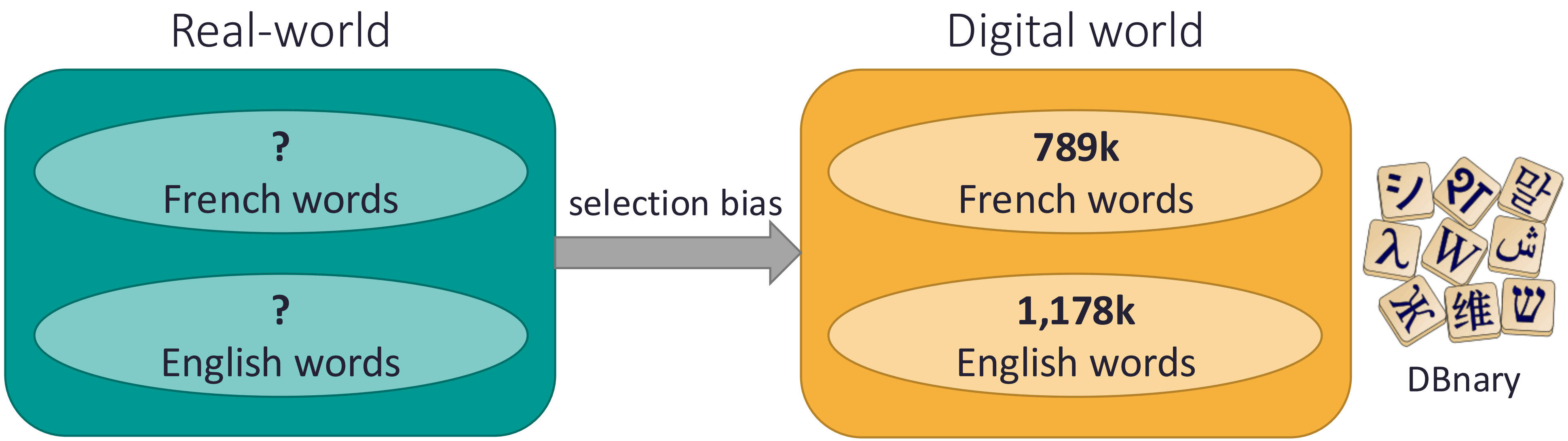


**The data are correct, there is a gender gap in  reality**

vs

**The data is wrong because of selection bias!**

# Challenge: No ground truth



Real-world

?
French words

?
English words

selection bias

Digital world

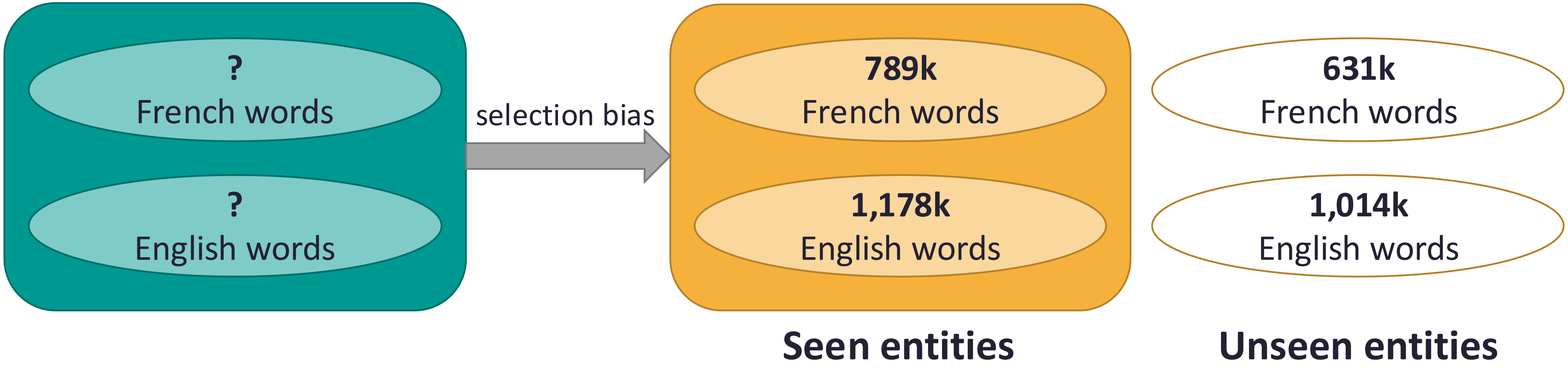
789k
French words

1,178k
English words

DBnary

**In general, there is no ground truth…**

# Principle: Comparing the proportion of unseen entities



?
French words

?
English words

selection bias

789k
French words

1,178k
English words

631k
French words

1,014k
English words

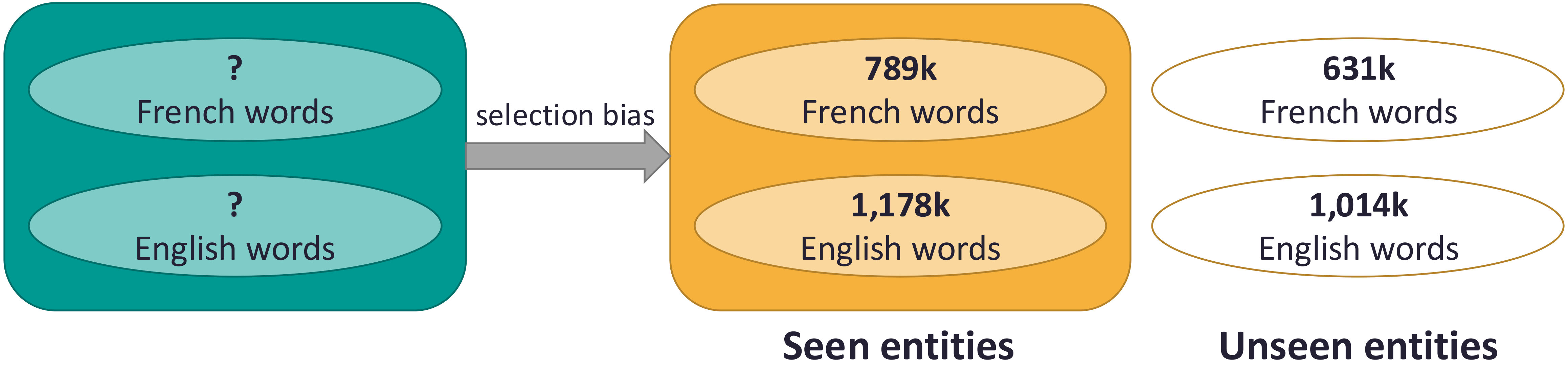
**Seen entities**

**Unseen entities**

❶ Estimating the quantity of unseen entities

# Principle: Comparing the proportion of unseen entities

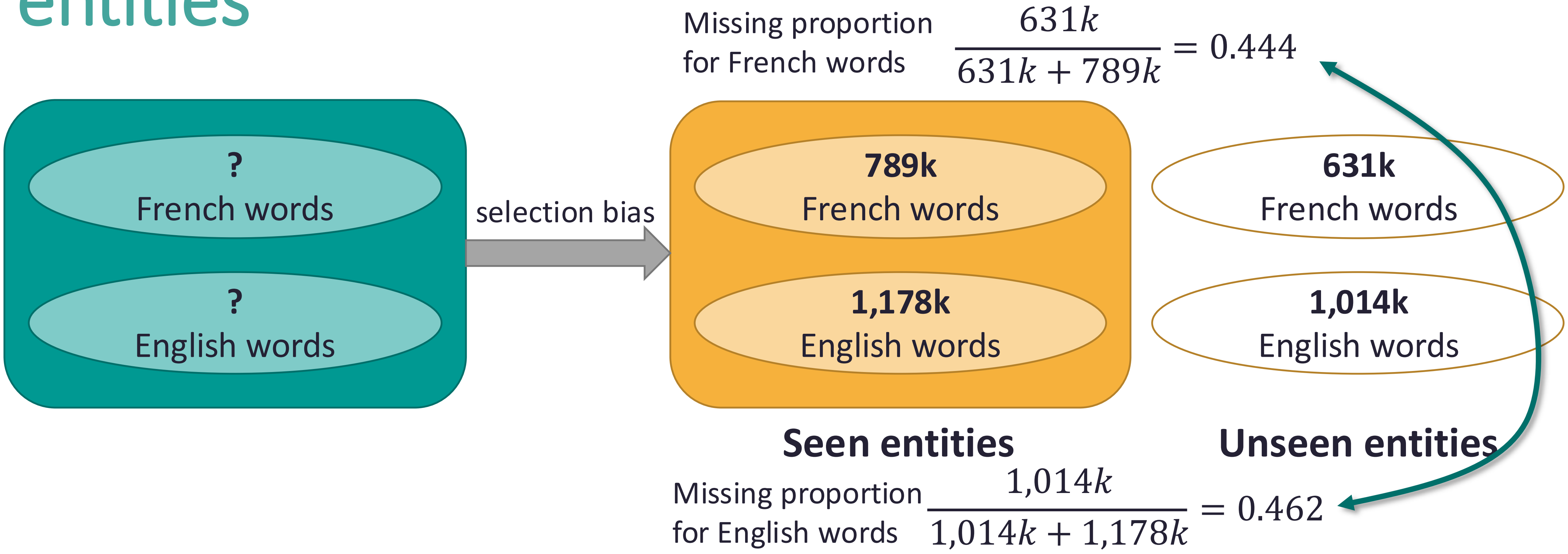
Missing proportion for French words $\frac{631k}{631k + 789k} = 0.444$



Missing proportion for English words $\frac{1{,}014k}{1{,}014k + 1{,}178k} = 0.462$

❷ Computing the missing proportion = unseen entities / total

# Principle: Comparing the proportion of unseen entities

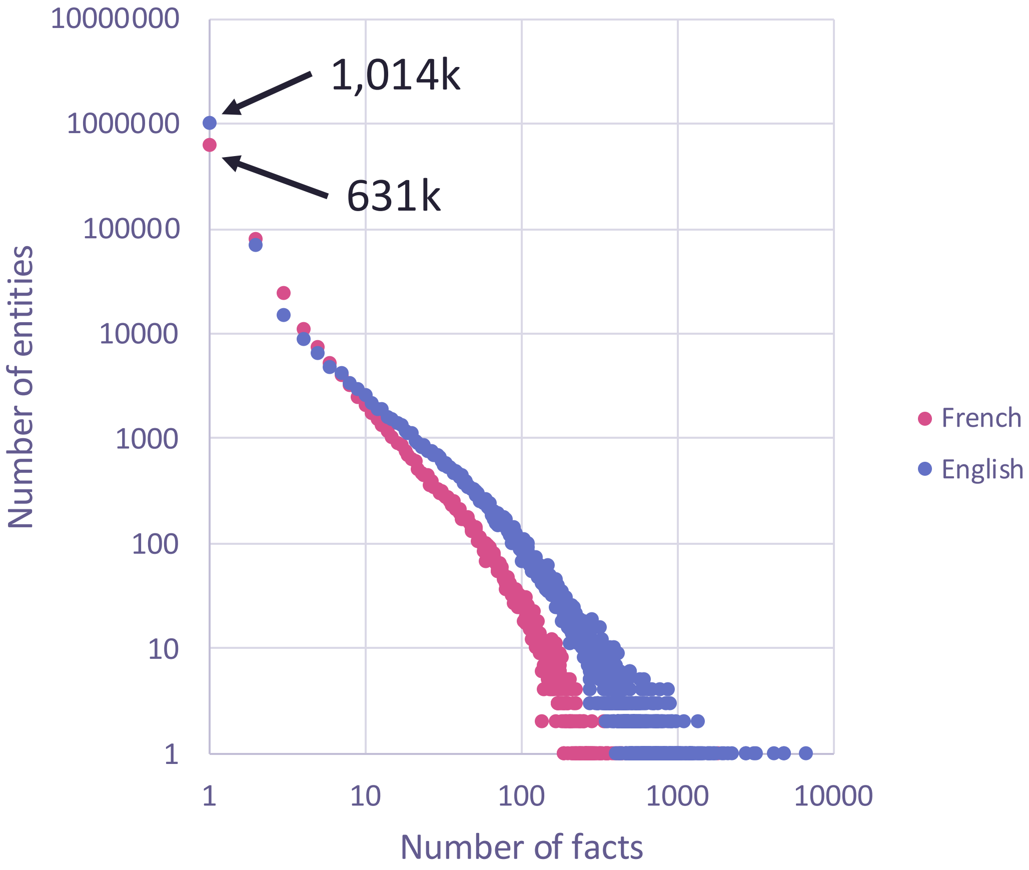


Missing proportion for French words $\frac{631k}{631k + 789k} = 0.444$

Missing proportion for English words $\frac{1{,}014k}{1{,}014k + 1{,}178k} = 0.462$

❸ selection bias = difference between missing proportions

# How to estimate the quantity of unseen entities ?



## Good-Turing frequency estimation:

**#unseen entities**

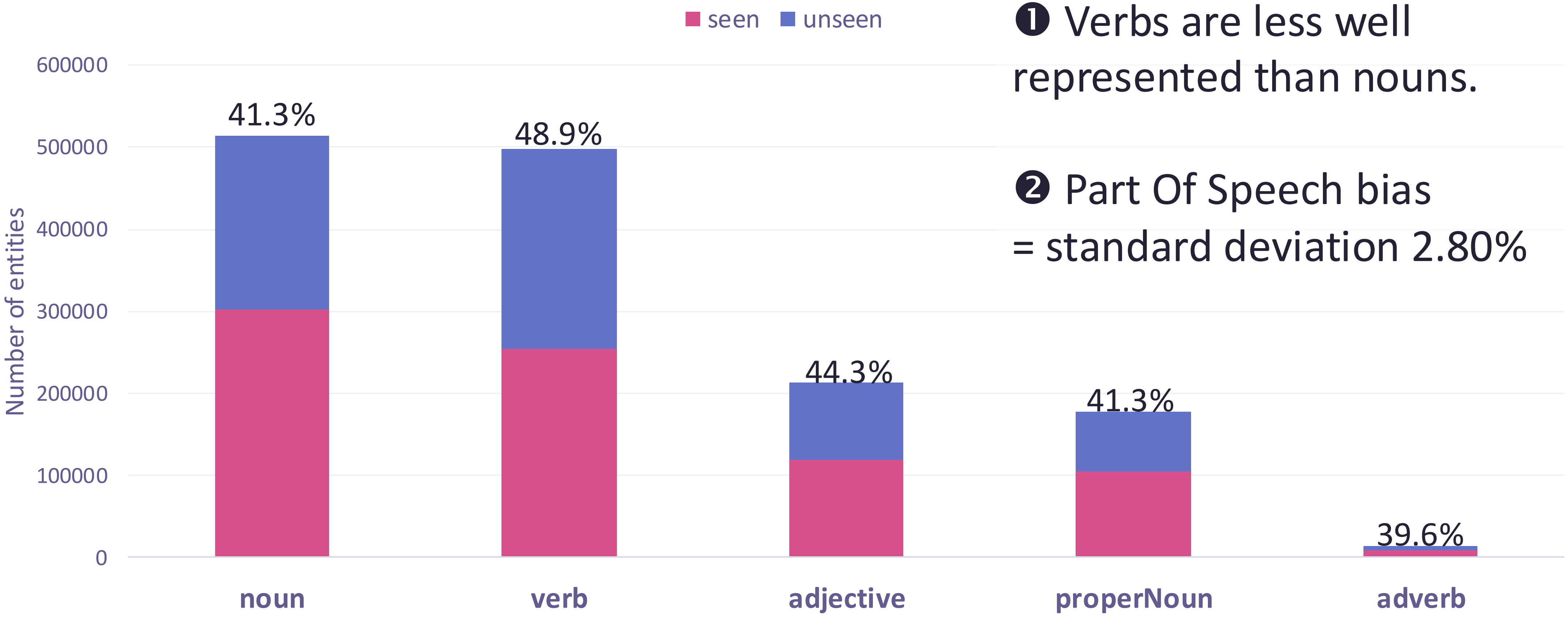
**=**

**#entities seen only once**

That’s it!

*[Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. Biometrika, 40(3-4), 237-264.]*

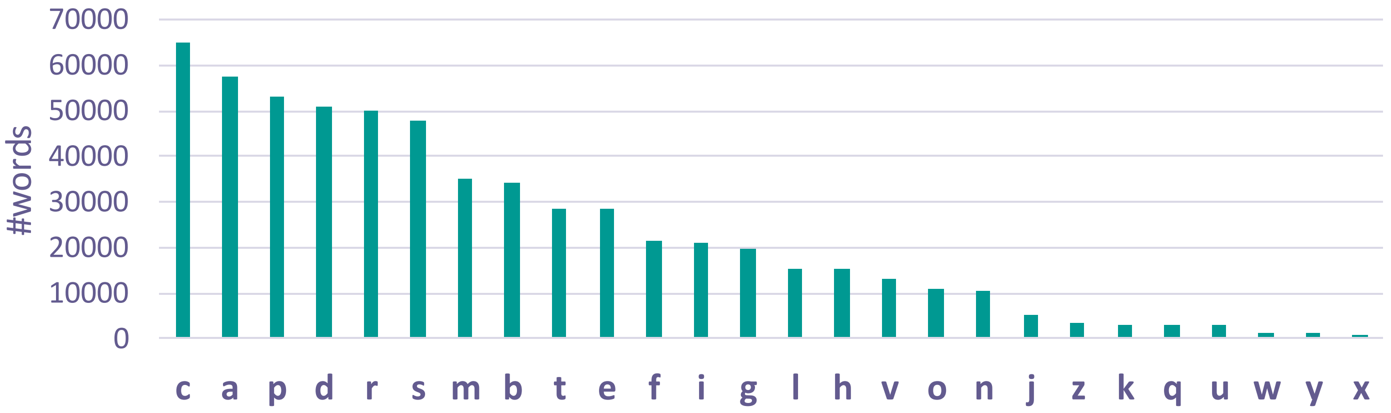
# Example: Part Of Speech for French words



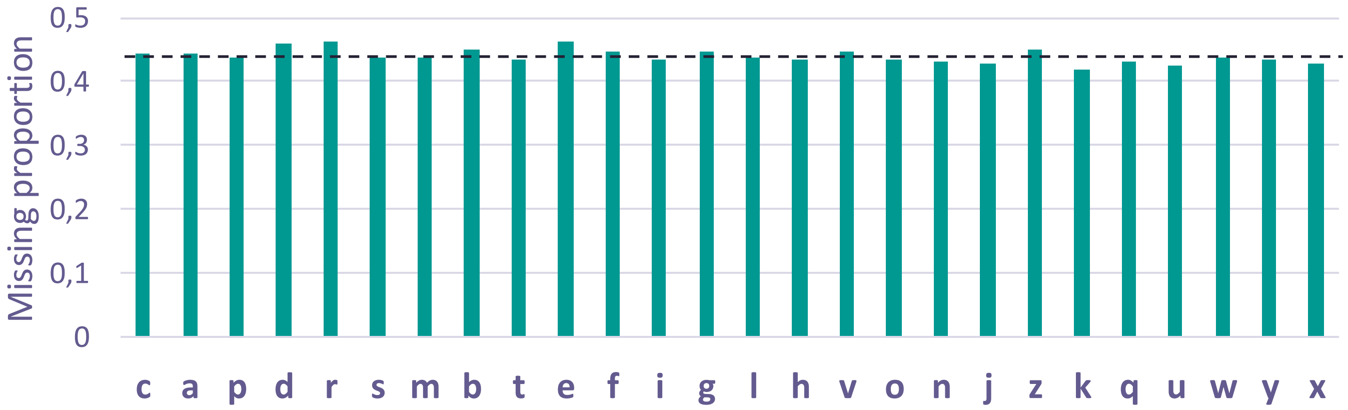
❶ Verbs are less well represented than nouns.

❷ Part Of Speech bias = standard deviation 2.80%

# Experimental evaluation on French Words



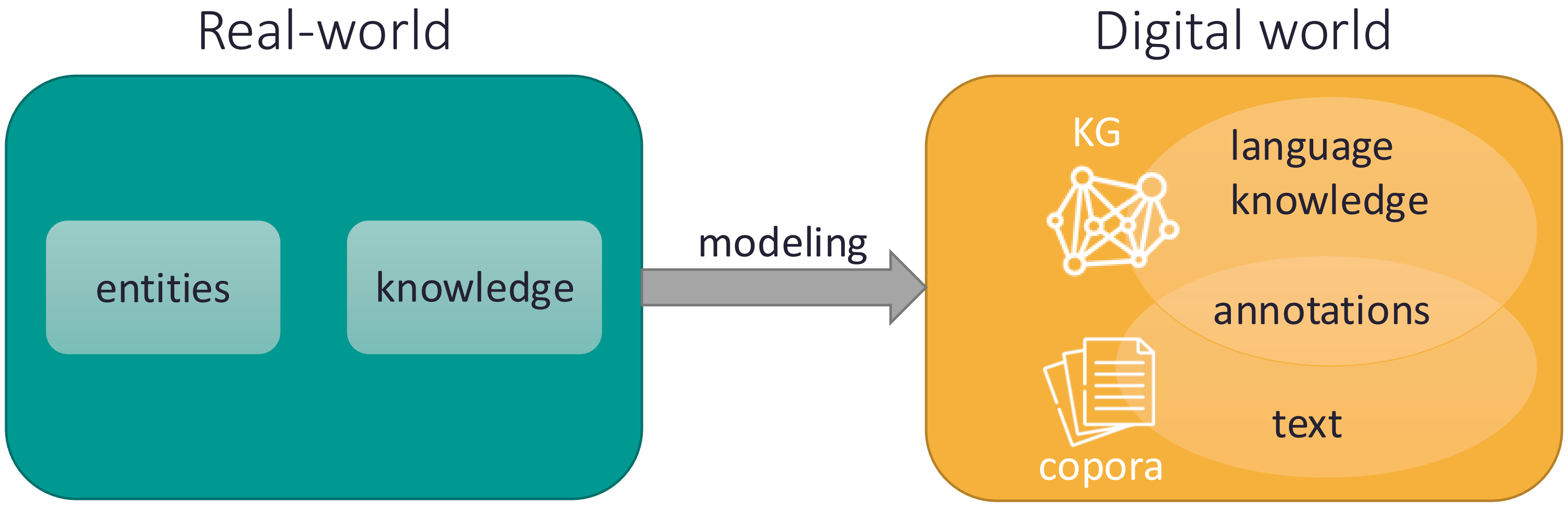
Distribution of words by first letter ➔ a priori, weak bias



First letter bias

= standard deviation 1.12%

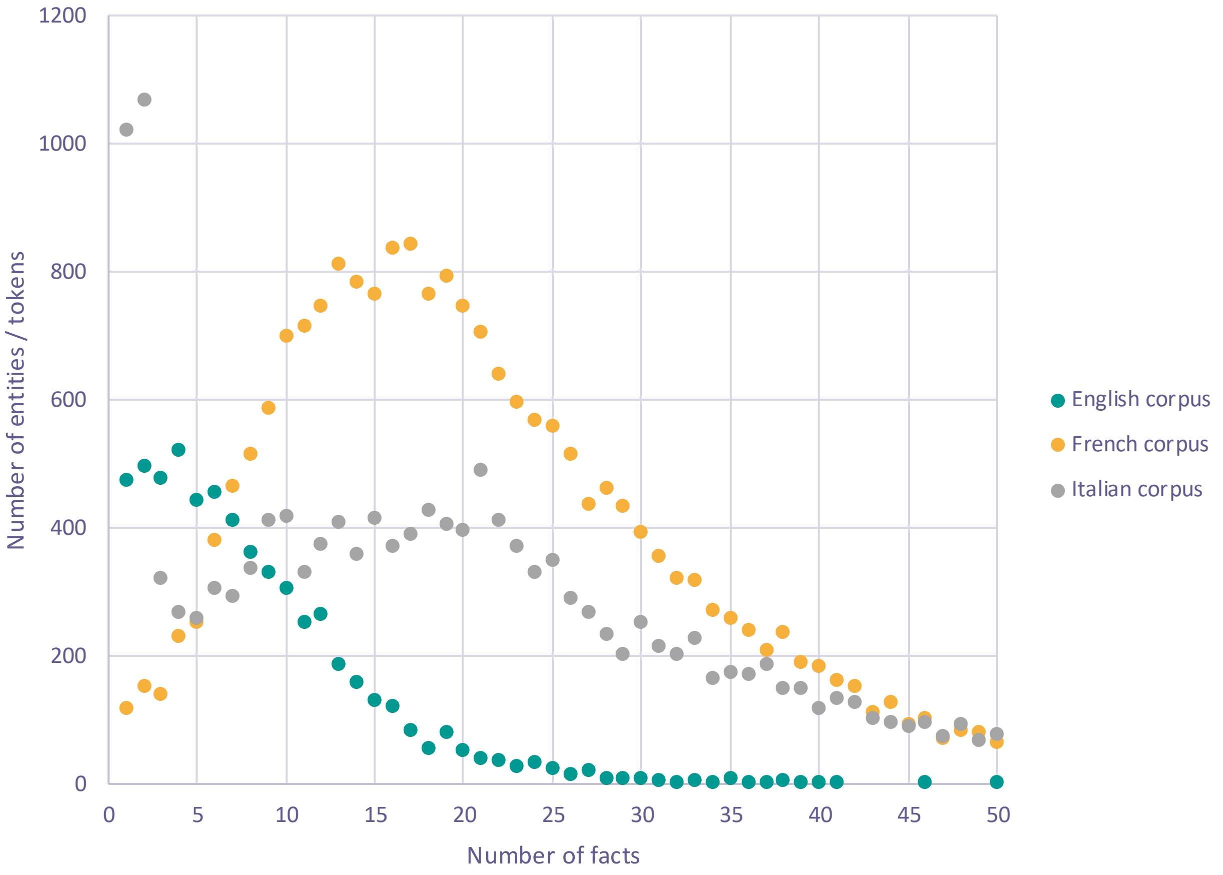
# What's the link between knowledge graphs and corpora?



# How can Good Turing be applied to texts?

- Text is discrete data by nature.
- PARSEME sample:

| Language | EN | FR | IT |
|---|---|---|---|
| Corpus Tokens | 109856 | 457505 | 352985 |
| MWE Tokens | 2386 | 12730 | 9778 |



# How can Good Turing be applied to texts?

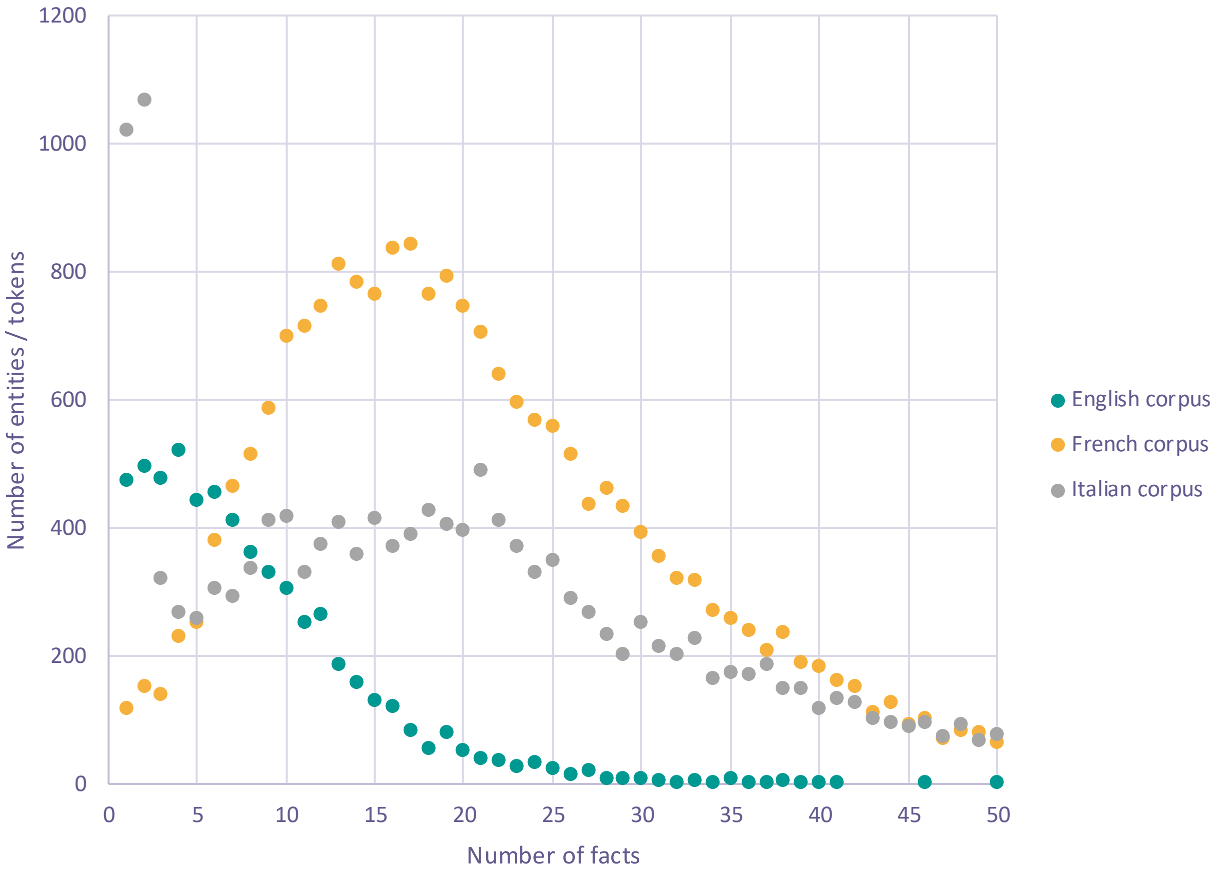
❑ Text is discrete data by nature.

❑ PARSEME sample:

| Language | EN | FR | IT |
|---|---|---|---|
| Corpus Tokens | 109856 | 457505 | 352985 |
| MWE Tokens | 2386 | 12730 | 9778 |

❑ Least complete : Italian, English

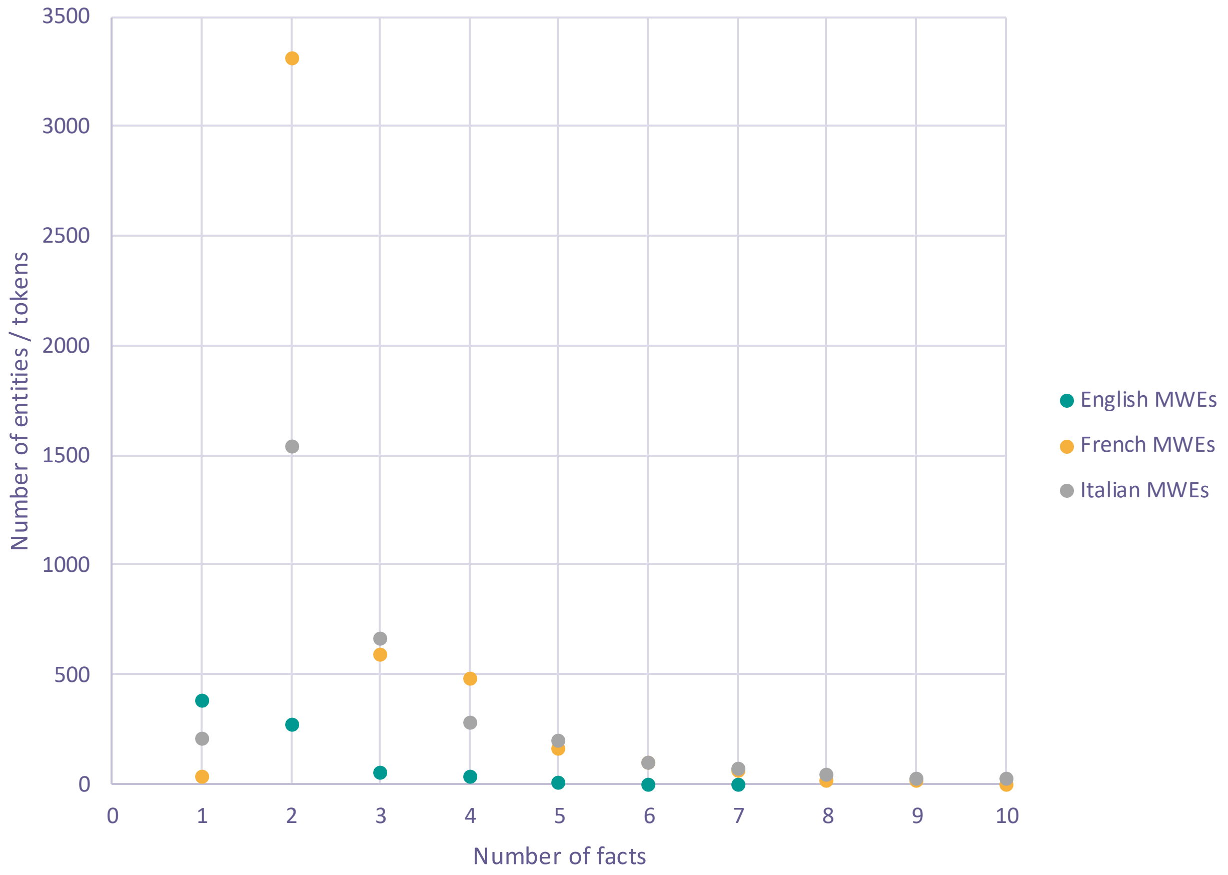
❑ Most complete : French



# How can Good Turing be applied to MWEs?

- Apply on MWEs as a list of tokens.
- PARSEME sample:

| Language | EN | FR | IT |
|---|---|---|---|
| Corpus Tokens | 109856 | 457505 | 352985 |
| MWE Tokens | 2386 | 12730 | 9778 |

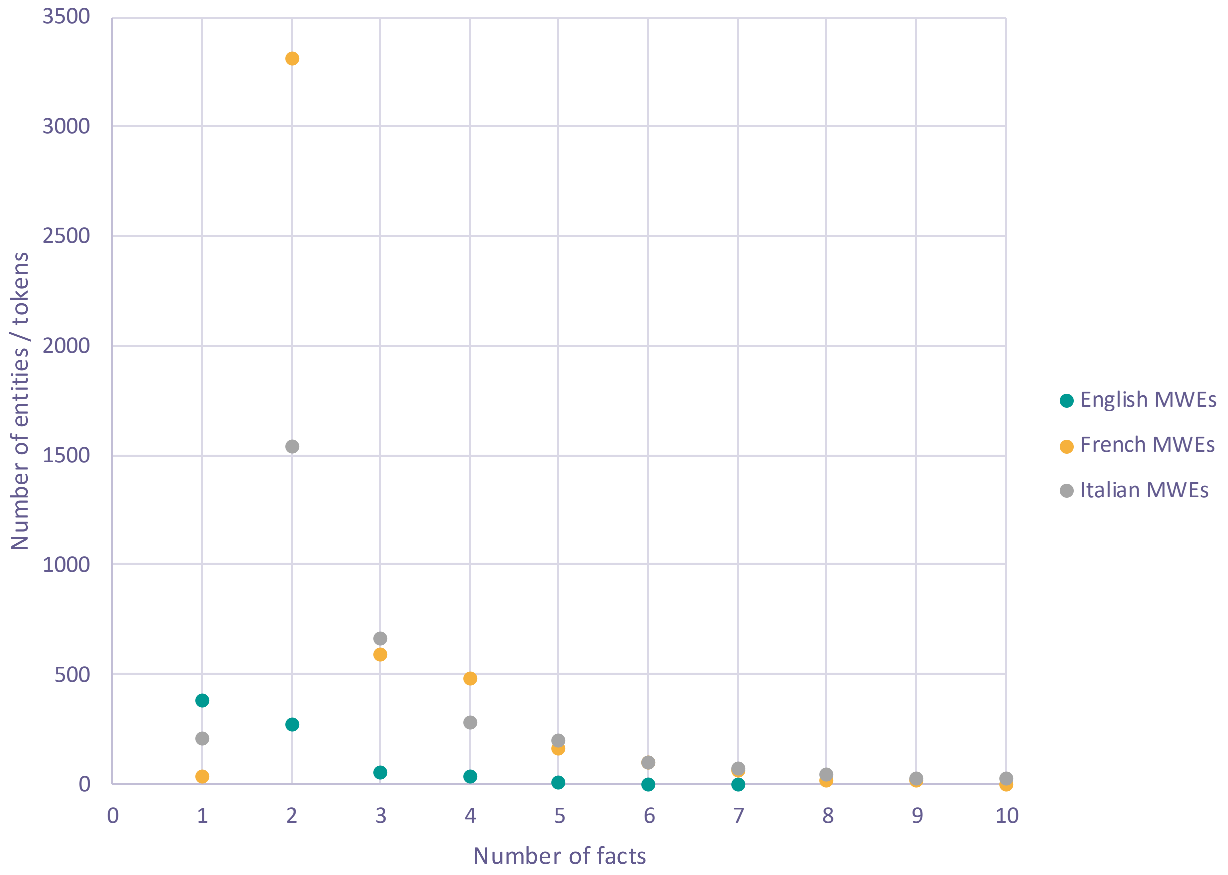


# How can Good Turing be applied to MWEs?

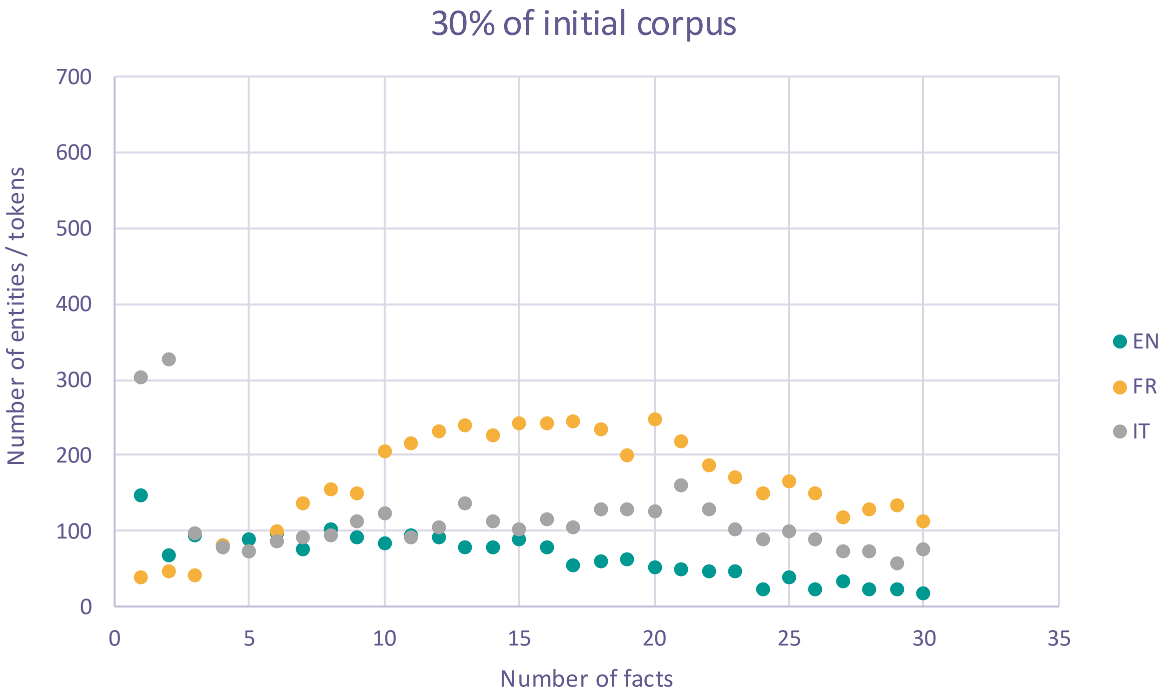
- ❑ Apply on MWEs as a list of tokens.
- ❑ PARSEME sample:

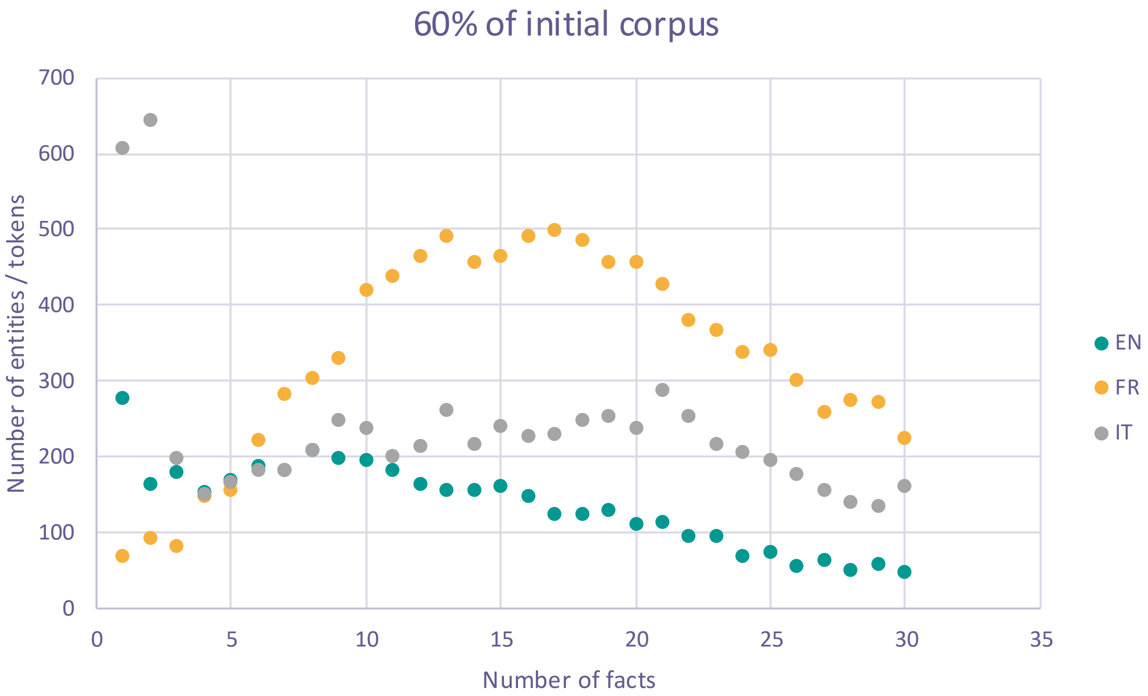
| Language | EN | FR | IT |
|---|---|---|---|
| Corpus Tokens | 109856 | 457505 | 352985 |
| MWE Tokens | 2386 | 12730 | 9778 |

- ❑ Sample might be too small for a definite conclusion
- ❑ Most complete corpus has less missing tokens

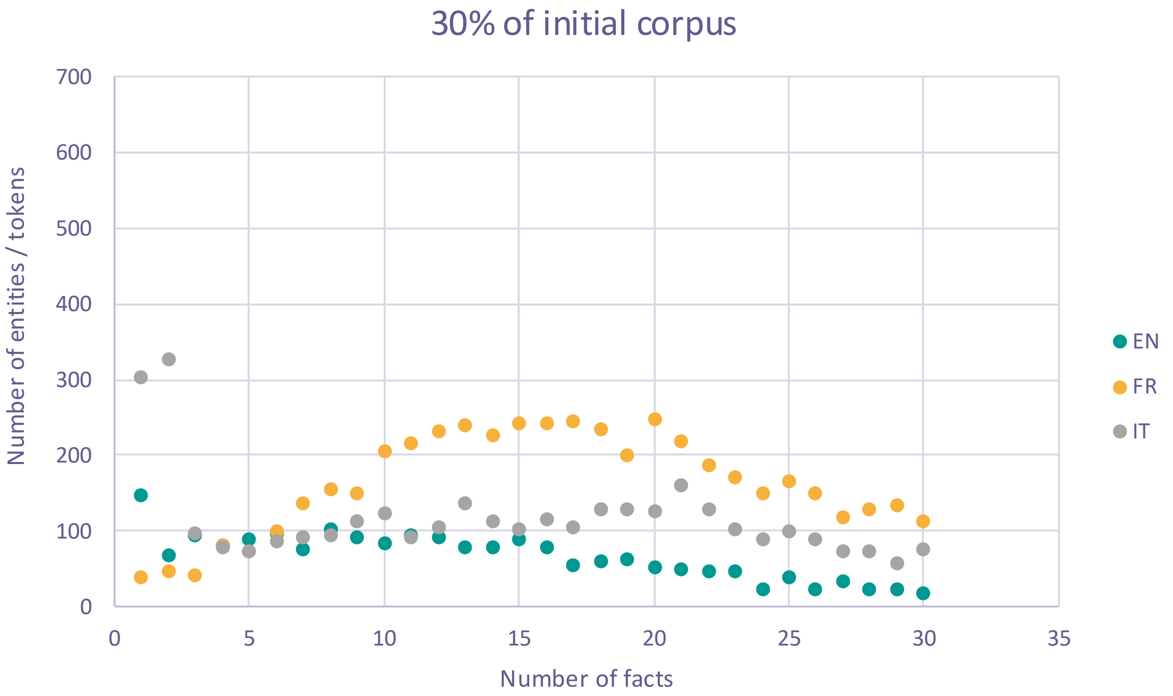


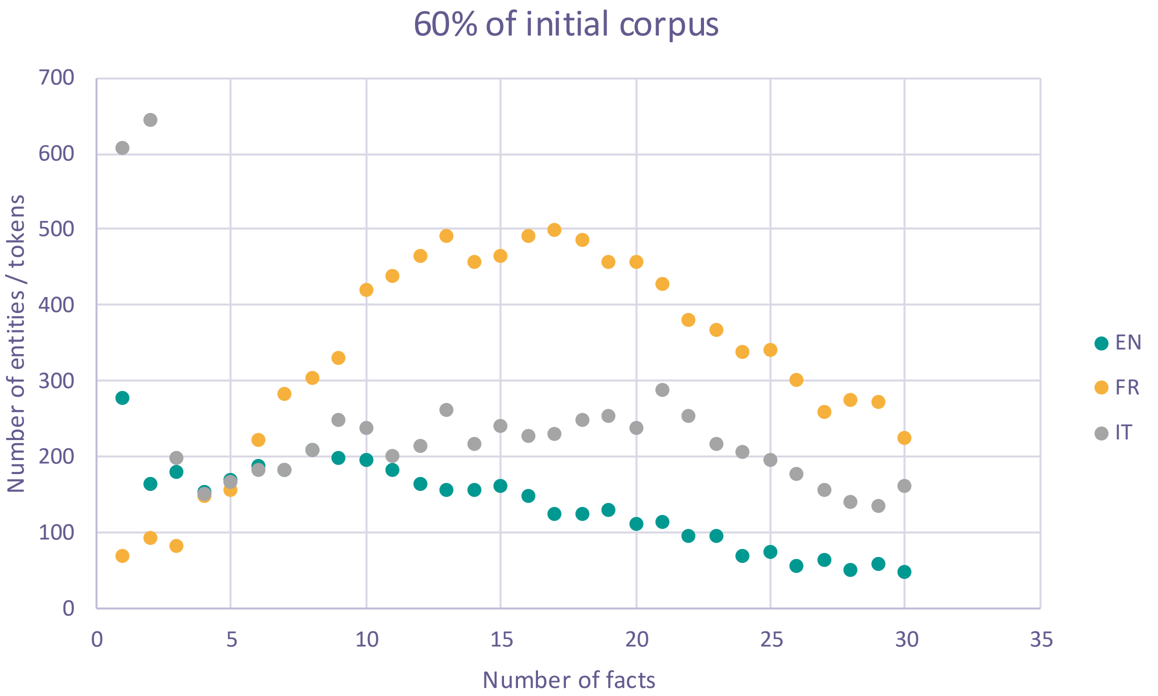
# What is the minimum size for completeness?





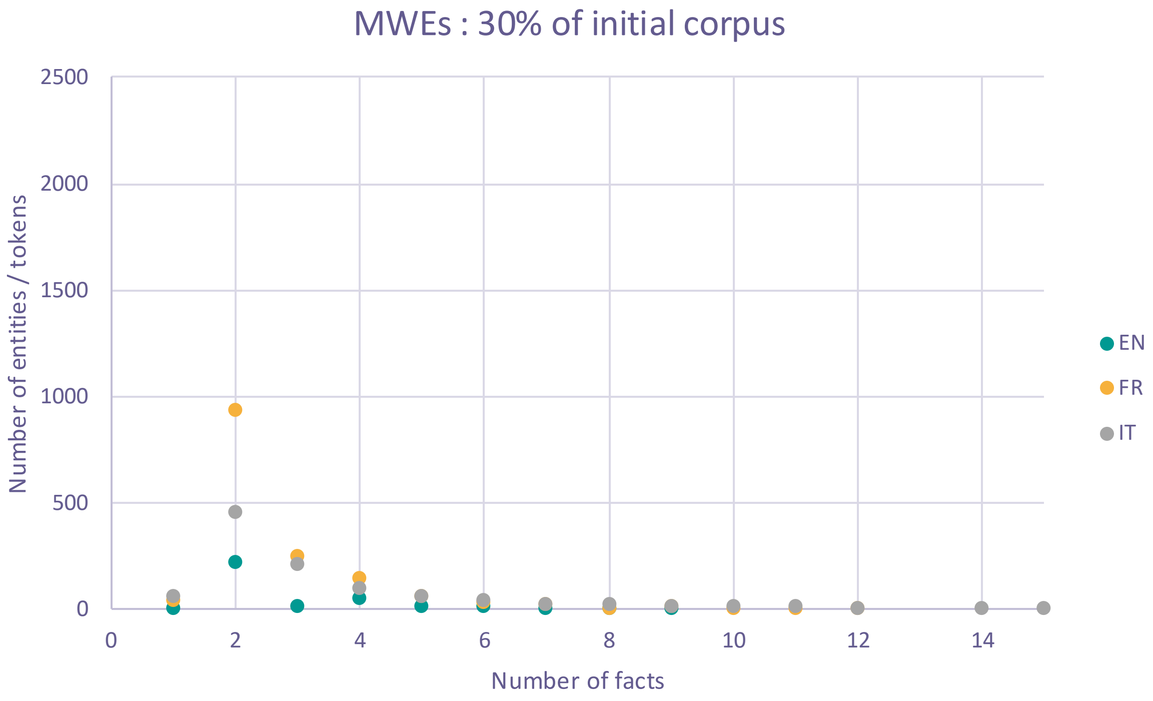
# What is the minimum size for completeness?

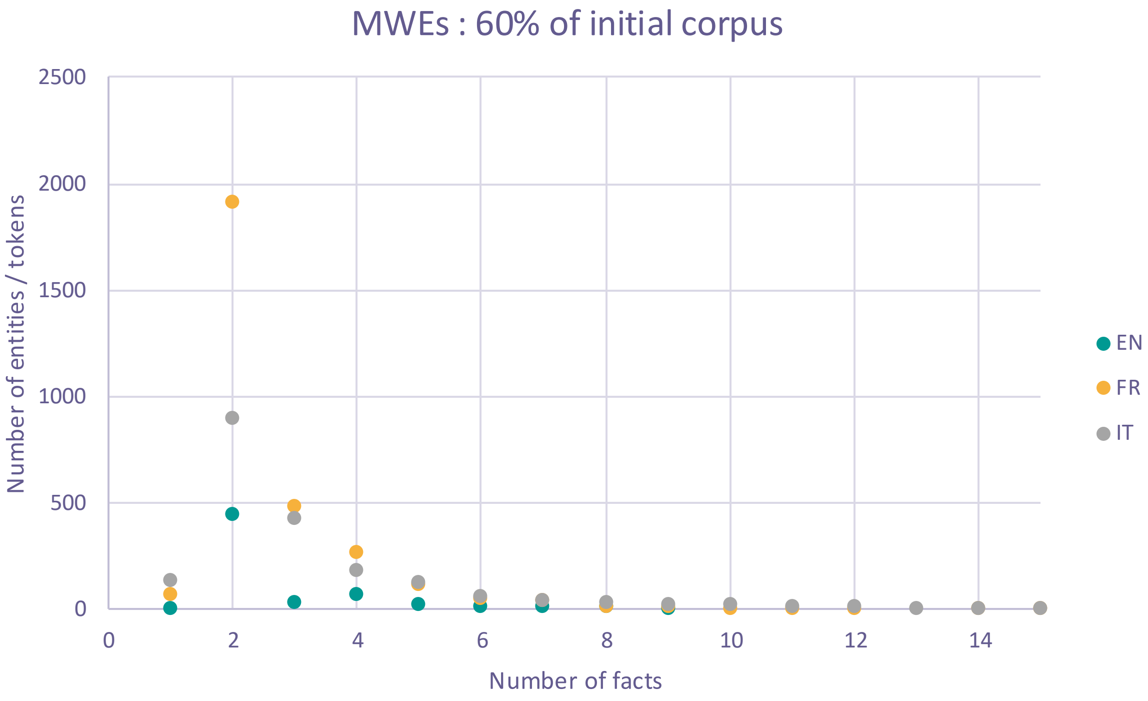




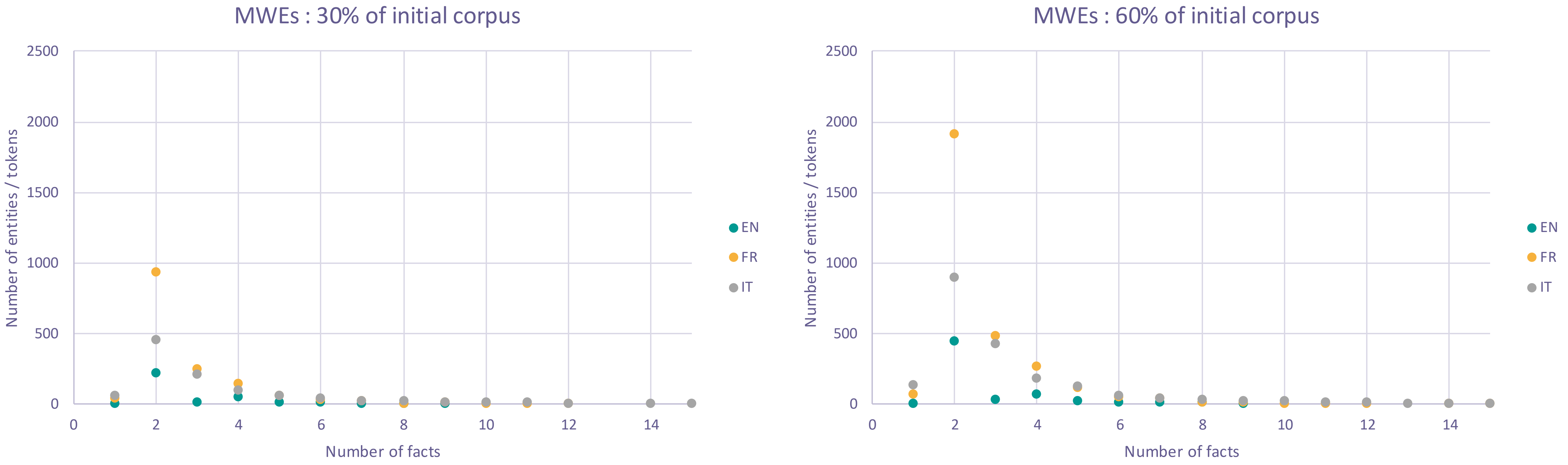
- Linear variation with random sampling

# What is the minimum size for completeness?





# What is the minimum size for completeness?

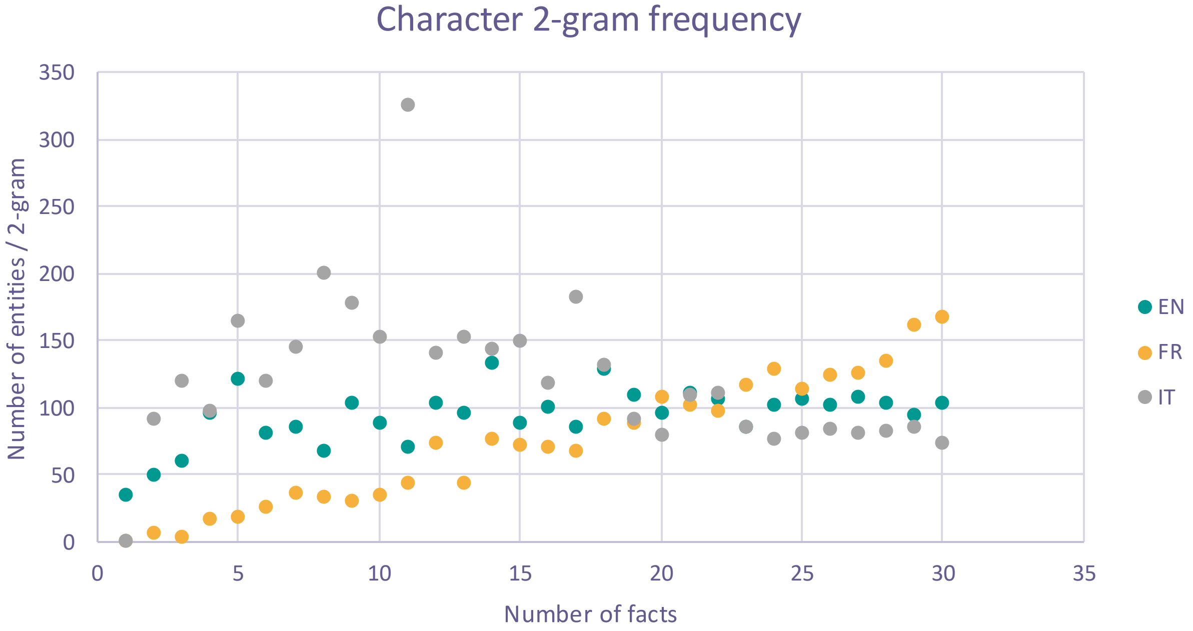


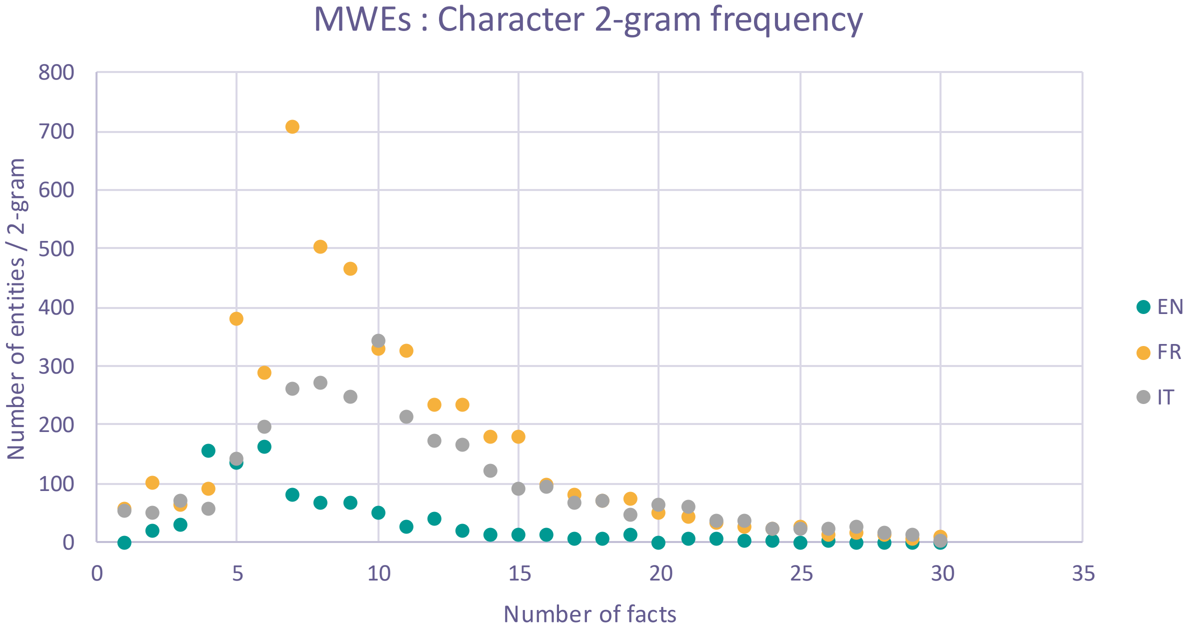
- ❑ Too few tokens in corpus, linear variation

# Is token a small enough feature?

- ❑ **We experiment on different character N-gram for lexical diversity analysis**

| Language | EN | FR | IT |
|---|---|---|---|
| Corpus uniq 2-grams | 7 420 | 20 960 | 15 720 |
| Corpus 2-grams | 364 263 | 1 682 170 | 1 382 702 |
| MWE uniq 2-grams | 958 | 4 805 | 3 192 |
| MWE 2-grams | 7 459 | 49 779 | 38 848 |

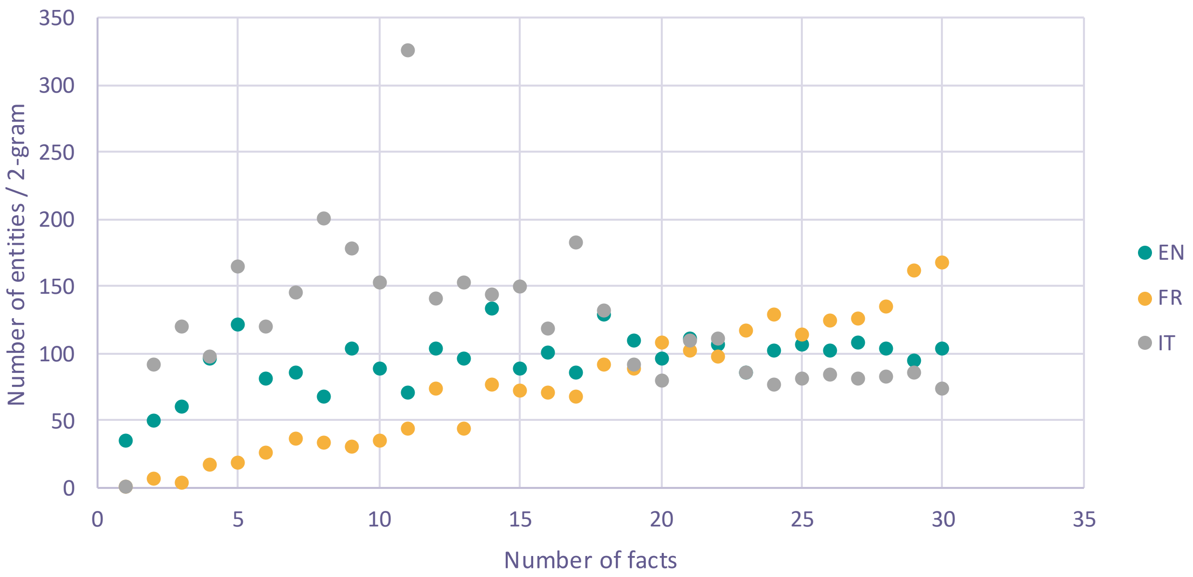




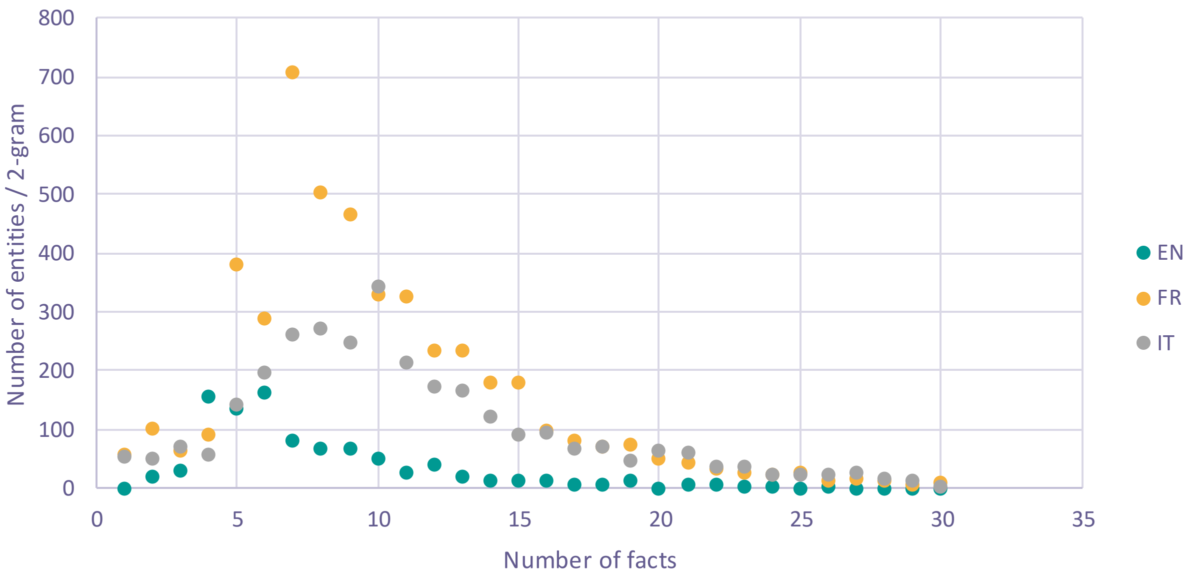
# Is token a small enough feature?

- ❑ **We experiment on different character N-gram for lexical diversity analysis**
- ❑ **Small N □ Completeness but no informativity**
  **□ Too small feature**

| Language | EN | FR | IT |
|---|---|---|---|
| Corpus uniq 2-grams | 7 420 | 20 960 | 15 720 |
| Corpus 2-grams | 364 263 | 1 682 170 | 1 382 702 |
| MWE uniq 2-grams | 958 | 4 805 | 3 192 |
| MWE 2-grams | 7 459 | 49 779 | 38 848 |



Character 2-gram frequency

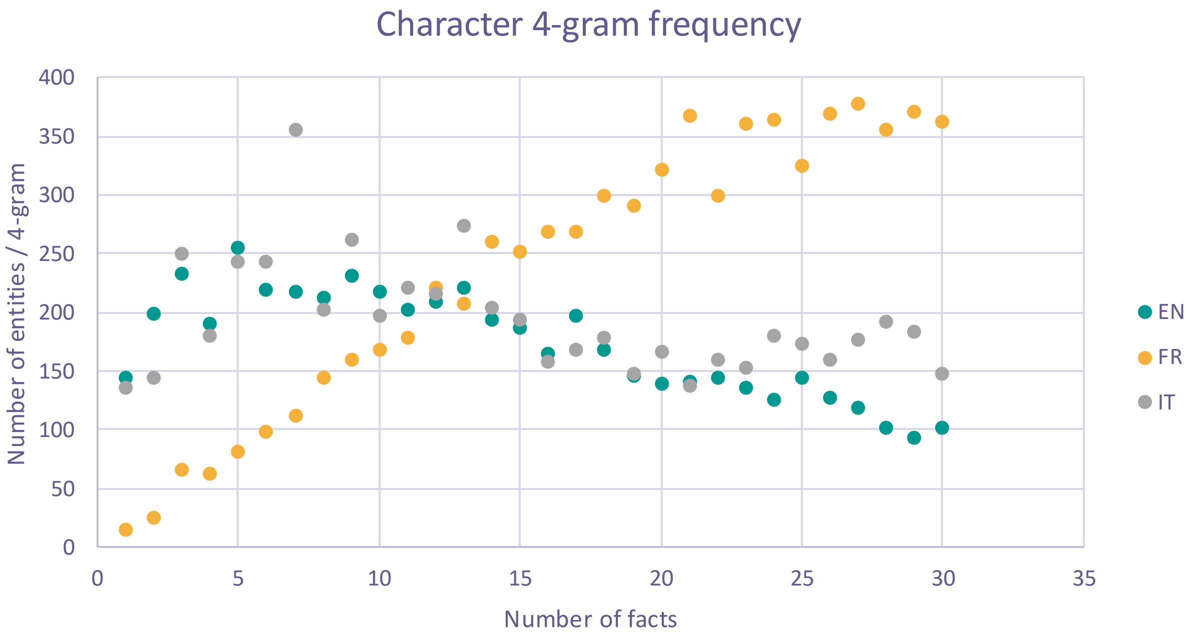


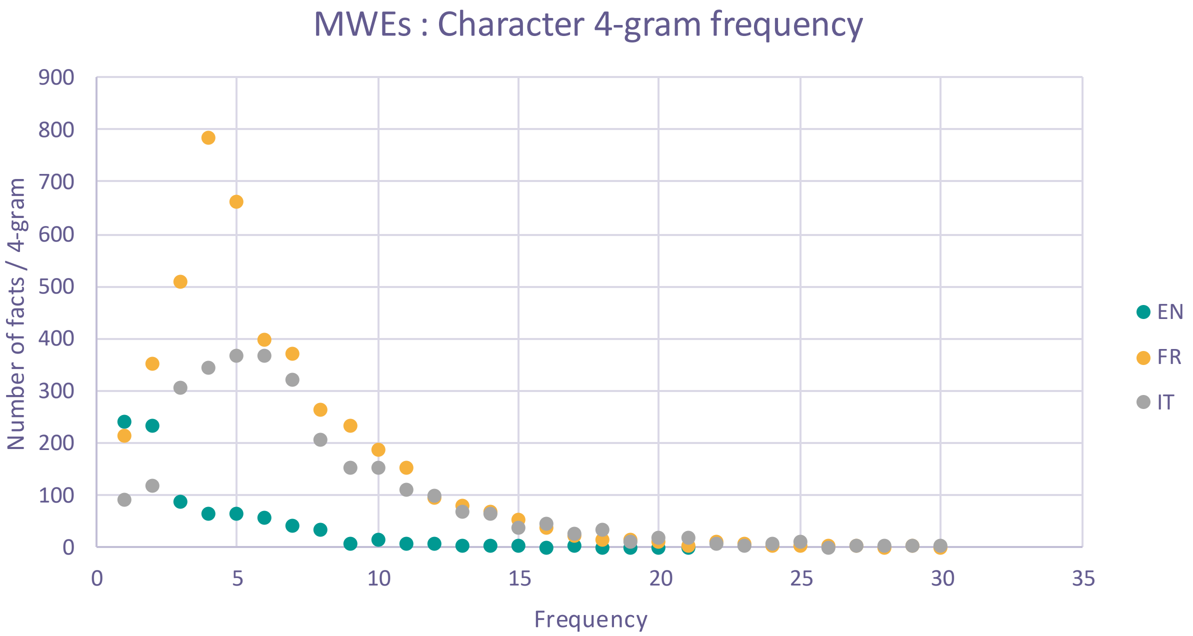
MWEs : Character 2-gram frequency

# Is token a small enough feature?

- ❑ **We experiment on different character N-gram for lexical diversity analysis**

| Language | EN | FR | IT |
|---|---|---|---|
| Corpus uniq 4-grams | 5 196 | 20 948 | 15 616 |
| Corpus 4-grams | 176 861 | 957 905 | 819 255 |
| MWE uniq 4-grams | 958 | 4 584 | 3 029 |
| MWE 4-grams | 7 459 | 28 717 | 22 702 |

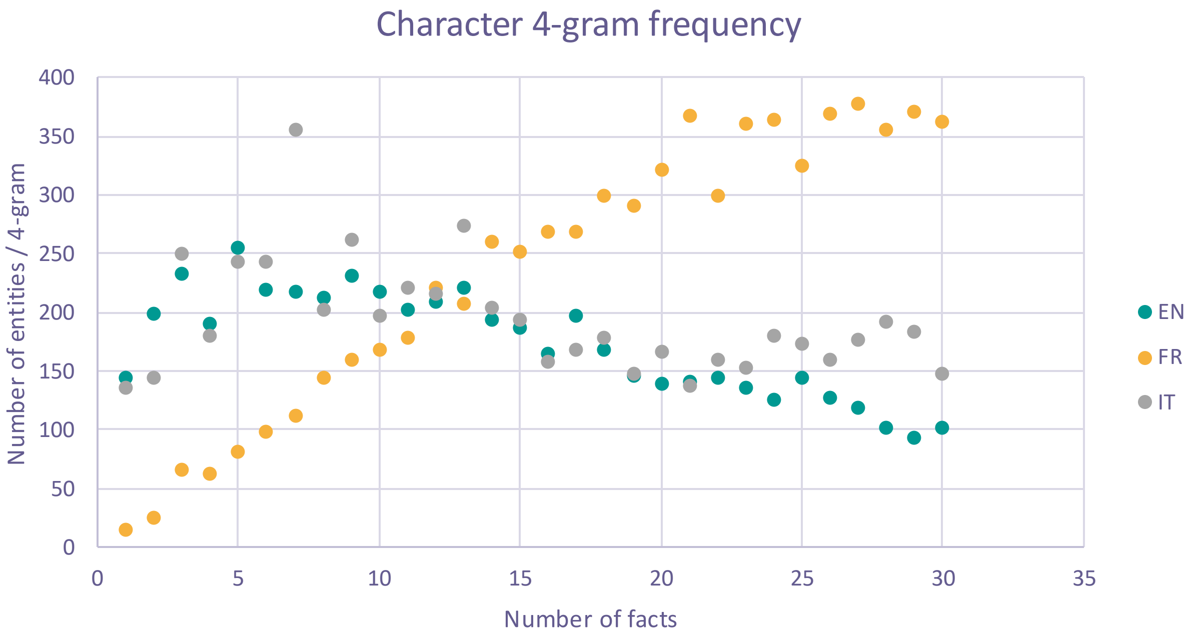


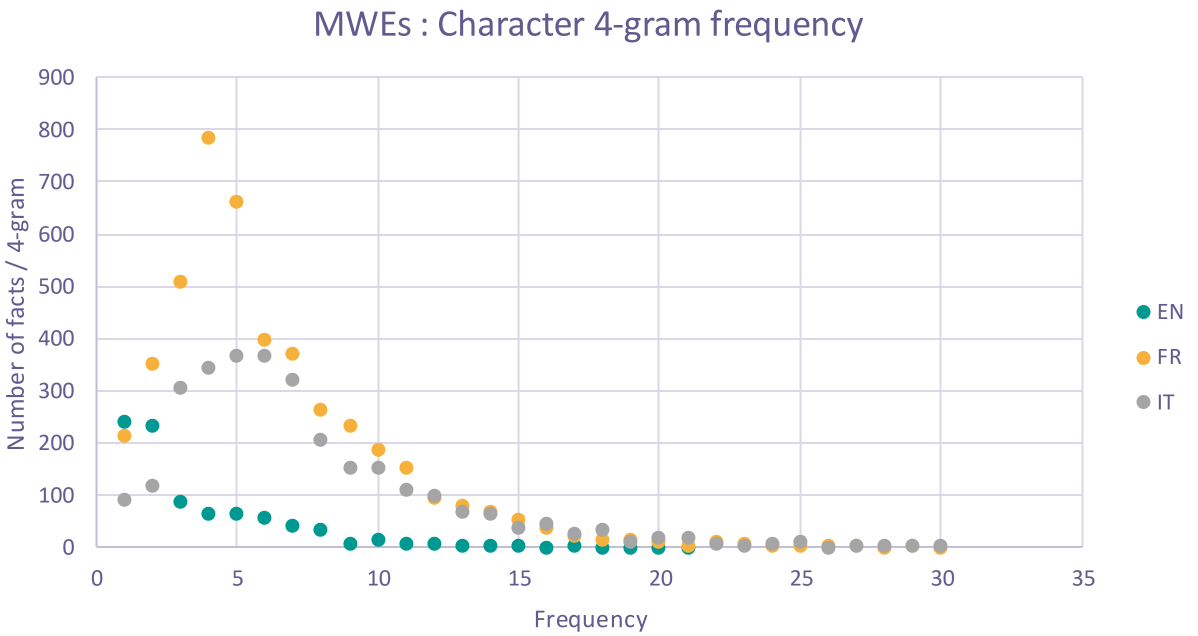


# Is token a small enough feature?

- ❑ **We experiment on different character N-gram for lexical diversity analysis**
- ❑ **Average N ⍰ Completeness for French corpus only**
  **⍰ But not for MWEs inside it**

| Language | EN | FR | IT |
|---|---|---|---|
| Corpus uniq 4-grams | 5 196 | 20 948 | 15 616 |
| Corpus 4-grams | 176 861 | 957 905 | 819 255 |
| MWE uniq 4-grams | 958 | 4 584 | 3 029 |
| MWE 4-grams | 7 459 | 28 717 | 22 702 |


Character 4-gram frequency

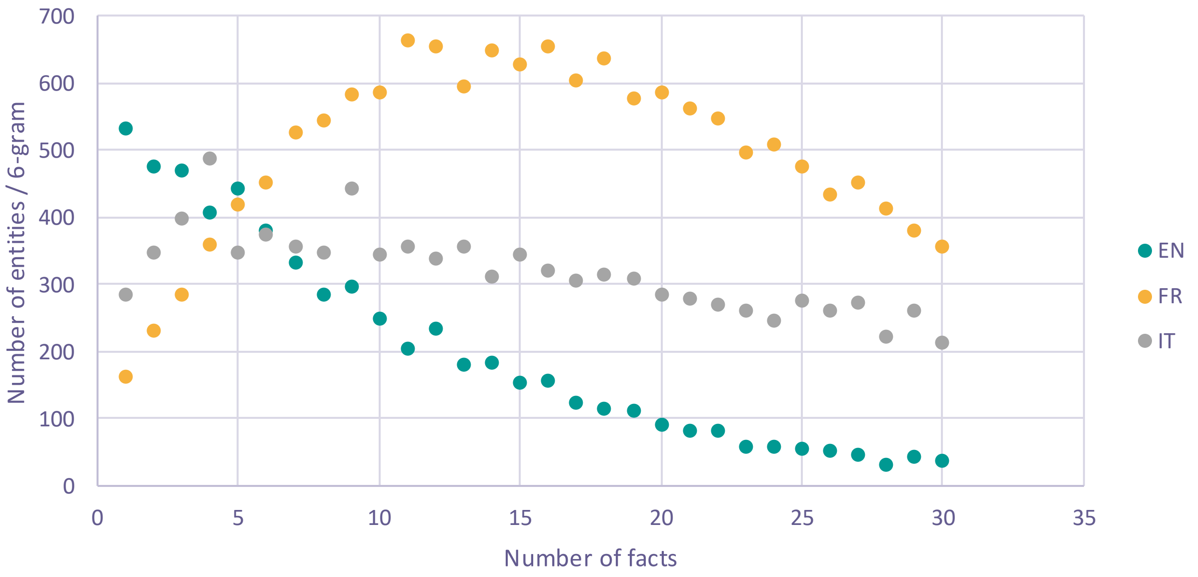

MWEs : Character 4-gram frequency

# Is token a small enough feature?

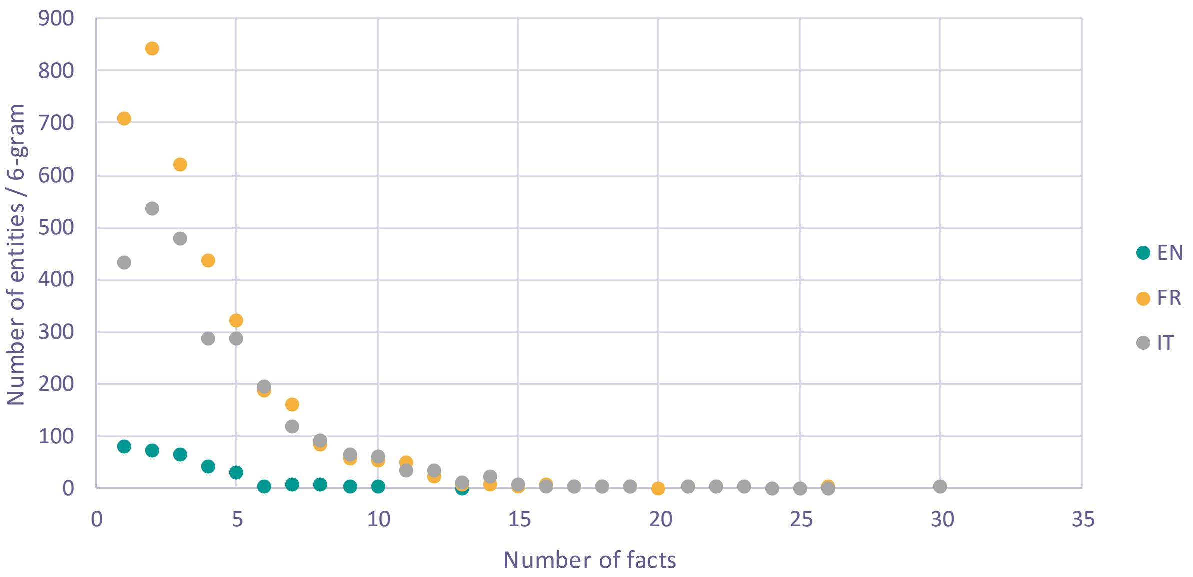
❑ **We experiment on different character N-gram for lexical diversity analysis**

| Language | EN | FR | IT |
|---|---:|---:|---:|
| Corpus uniq 6-grams | 6 502 | 20 813 | 15 187 |
| Corpus 6-grams | 77 544 | 503 010 | 442 811 |
| MWE uniq 6-grams | 315 | 3 588 | 2 696 |
| MWE 6-grams | 943 | 13 114 | 11 364 |

Character 6-gram frequency



MWEs : Character 6-gram frequency

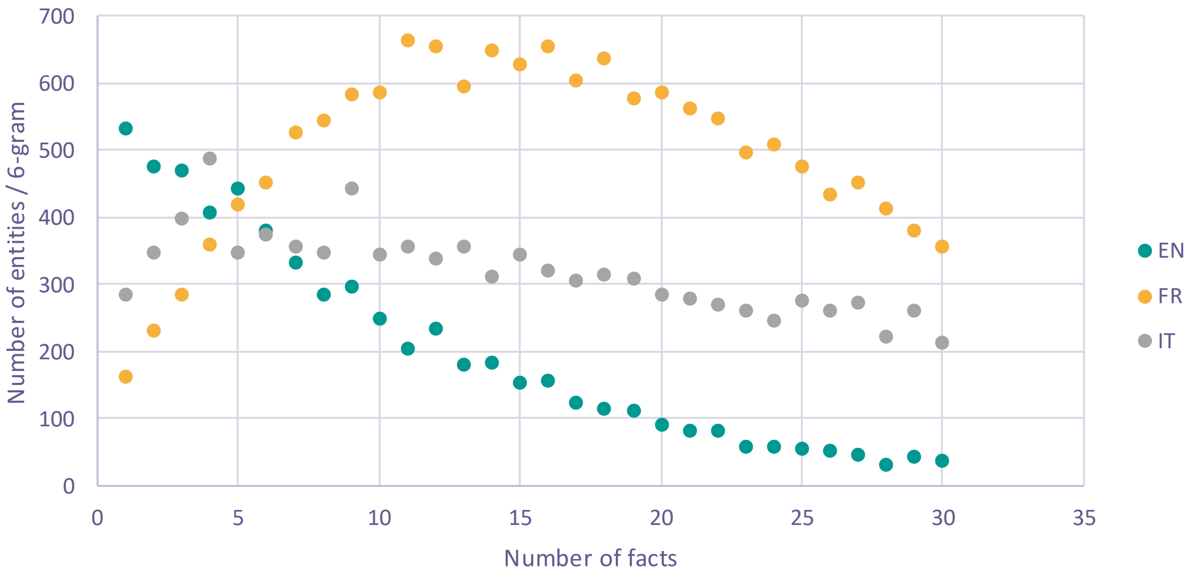


# Is token a small enough feature?

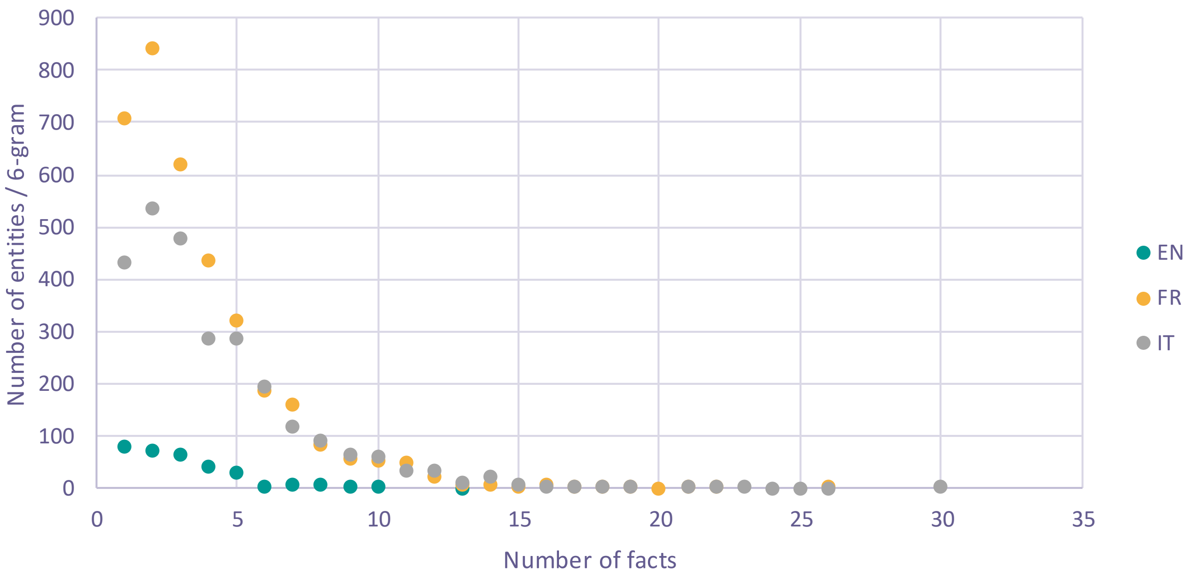
- ❑ **We experiment on different character N-gram for lexical diversity analysis**
- ❑ **Large N ⍰ No completeness for French corpus only**
  **⍰ Sample too small**

| Language | EN | FR | IT |
|---|---|---|---|
| Corpus uniq 6-grams | 6 502 | 20 813 | 15 187 |
| Corpus 6-grams | 77 544 | 503 010 | 442 811 |
| MWE uniq 6-grams | 315 | 3 588 | 2 696 |
| MWE 6-grams | 943 | 13 114 | 11 364 |

Character 6-gram frequency



MWEs : Character 6-gram frequency



# Conclusion

- ❑ **Dbnary bias analysis:**
  - Word/lexical bias detection (language bias, part of speech bias,…)
- ❑ **Corpora**
  - Insight about lexical completeness of a corpora ➔ evaluating the diversity of corpora
  - Next step: Subword study for reducing the vocabulary size
- ❑ **Future work:**
  - Study of synonym, antonym,… ➔ nym diversity
  - Study of syntaxic dependencies ➔ syntax diversity
  - How to detect relational biases?

# Thank you for your attention!

- ❑ Mamadou Balde
- ❑ Béatrice Markhoff
- ❑ Sophie Nung
- ❑ Manon Ovide
- ❑ Ryohta Shiojiri