WORK IN PROGRESS

Multilingual Wikipedia:
Editors of Primary Language Contribute to More Complex Articles

Sungjoon Park, Suin Kim*, Scott A. Hale, Sooyoung Kim, Jeongmin Byun, Alice Oh
{sungjoon.park, suin.kim, sooyoungkim, jmbyun}@kaist.ac.kr, scott.hale@oii.ox.ac.uk, alice.oh@kaist.edu

Research Questions

- Is it possible to *quantify the complexity of language* in Wikipedia articles?
- Do primary users tend to edit parts of the Wikipedia articles with *higher language complexity* than non-primary users?
- Do we observe *more natural language* in the articles after primary users' edits compared to the articles after non-primary users' edits?

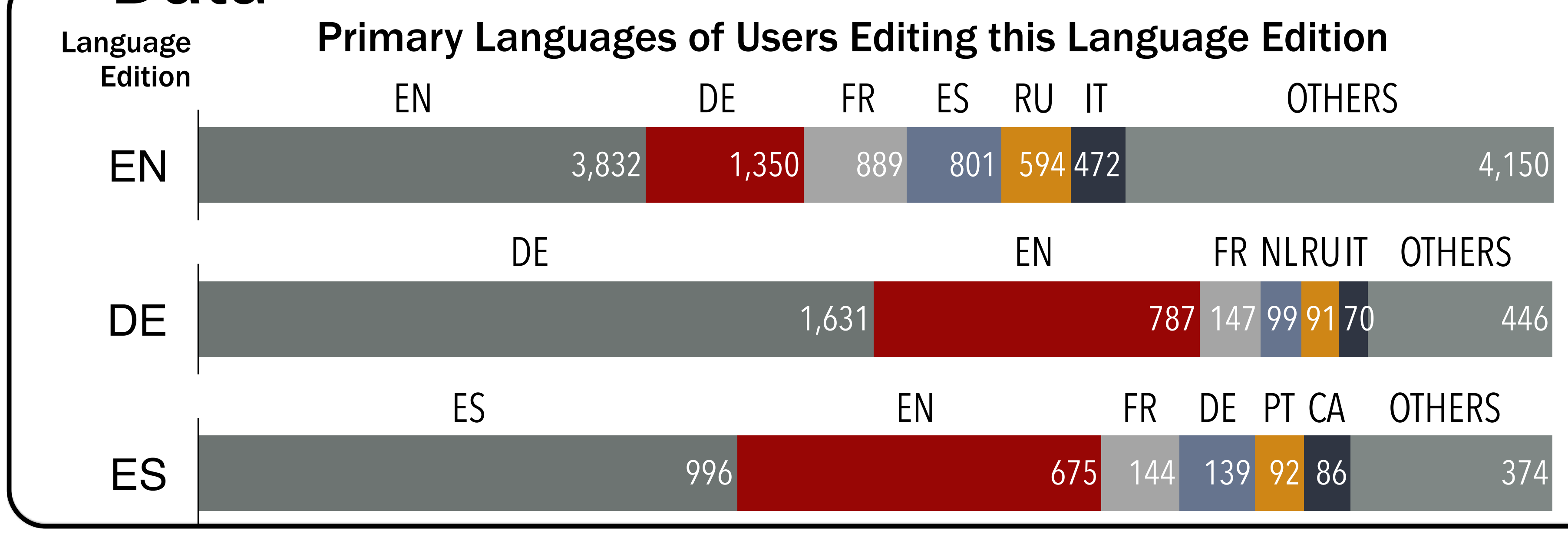
Primary/Non-Primary Users

- Multilingual editors have edited different language editions
- We define a user's primary language as the language that the user edited the most (not necessarily user's native language)
- For each language X edition, we define primary users as those whose primary language is X

Contributions

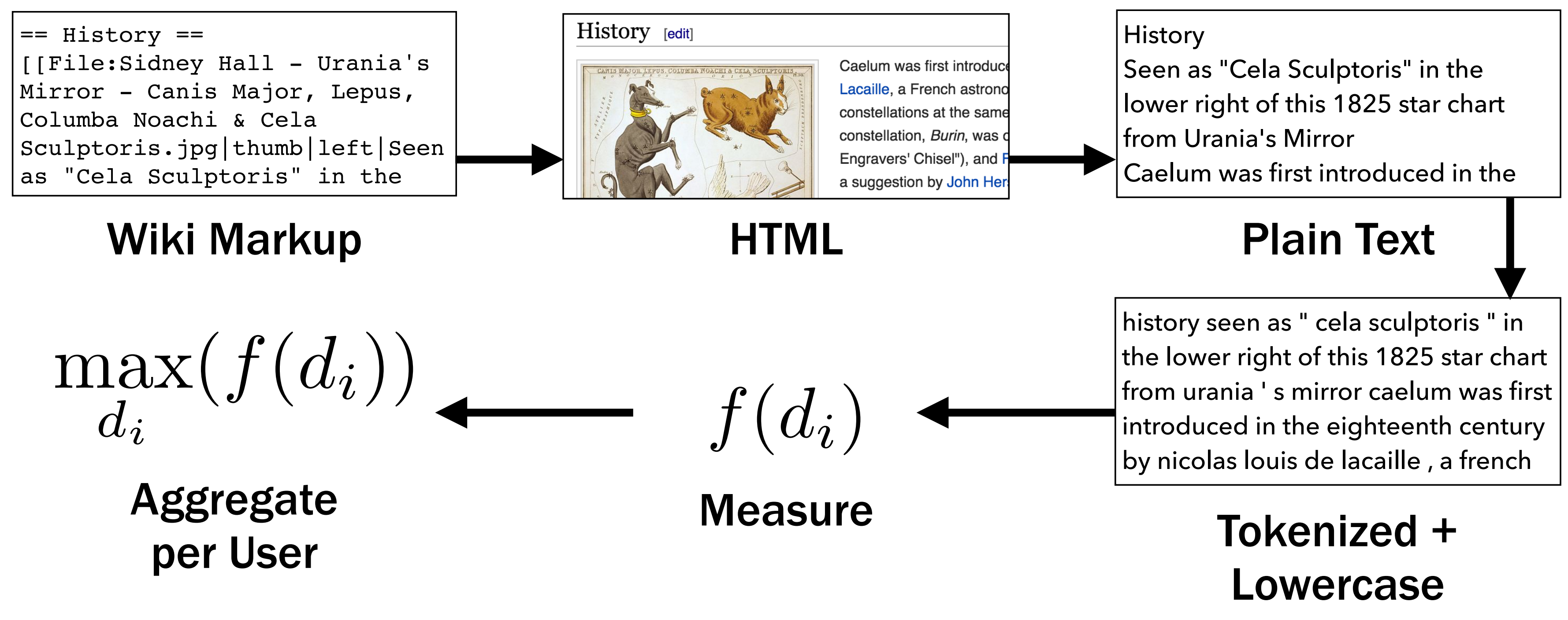
- We tested 20 measures for language complexity and showed they are highly consistent.
- Multilinguals in Wikipedia show relatively high levels of proficiency in their primary languages.

Data



	English	German	Spanish
#Editors	11,616	3,271	2,506
#Article Edit Sessions	237,849	120,123	69,557
#Edits	350,541	160,126	112,099

Quantifying Language Complexity



```
== History ==
[[File:Sidney Hall - Urania's
Mirror - Canis Major, Lepus,
Columba Noachi & Cela
Sculptoris.jpg|thumb|left|Seen
as "Cela Sculptoris" in the
```

Wiki Markup

History [edit]

Caelum was first introduce
Lacaille, a French astrono
constellations at the same
constellation, *Burin*, was c
Engravers' Chisel"), and F
a suggestion by John Her

HTML

History
Seen as "Cela Sculptoris" in the lower right of this 1825 star chart from Urania's Mirror
Caelum was first introduced in the

Plain Text

history seen as " cela sculptoris " in the lower right of this 1825 star chart from urania ' s mirror caelum was first introduced in the eighteenth century by nicolas louis de lacaille , a french

Tokenized + Lowercase

$f(d_i)$

Measure

$\max_{d_i}(f(d_i))$

Aggregate per User

Basic Features

- Number of characters
- Number of words
- Number of unique words
- Number of sentences
- Average word length in characters
- Average sentence length in words

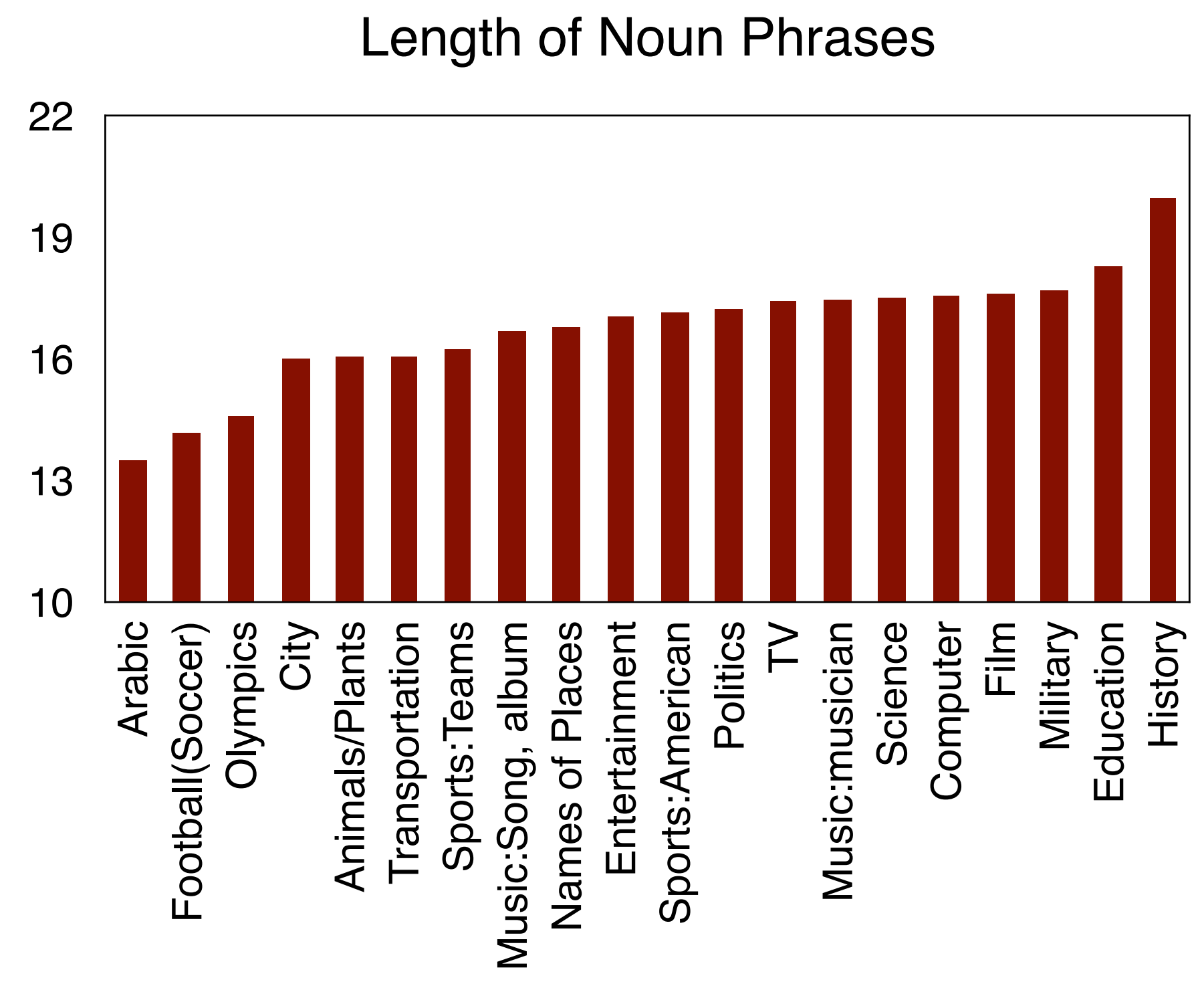
Lexical Diversity

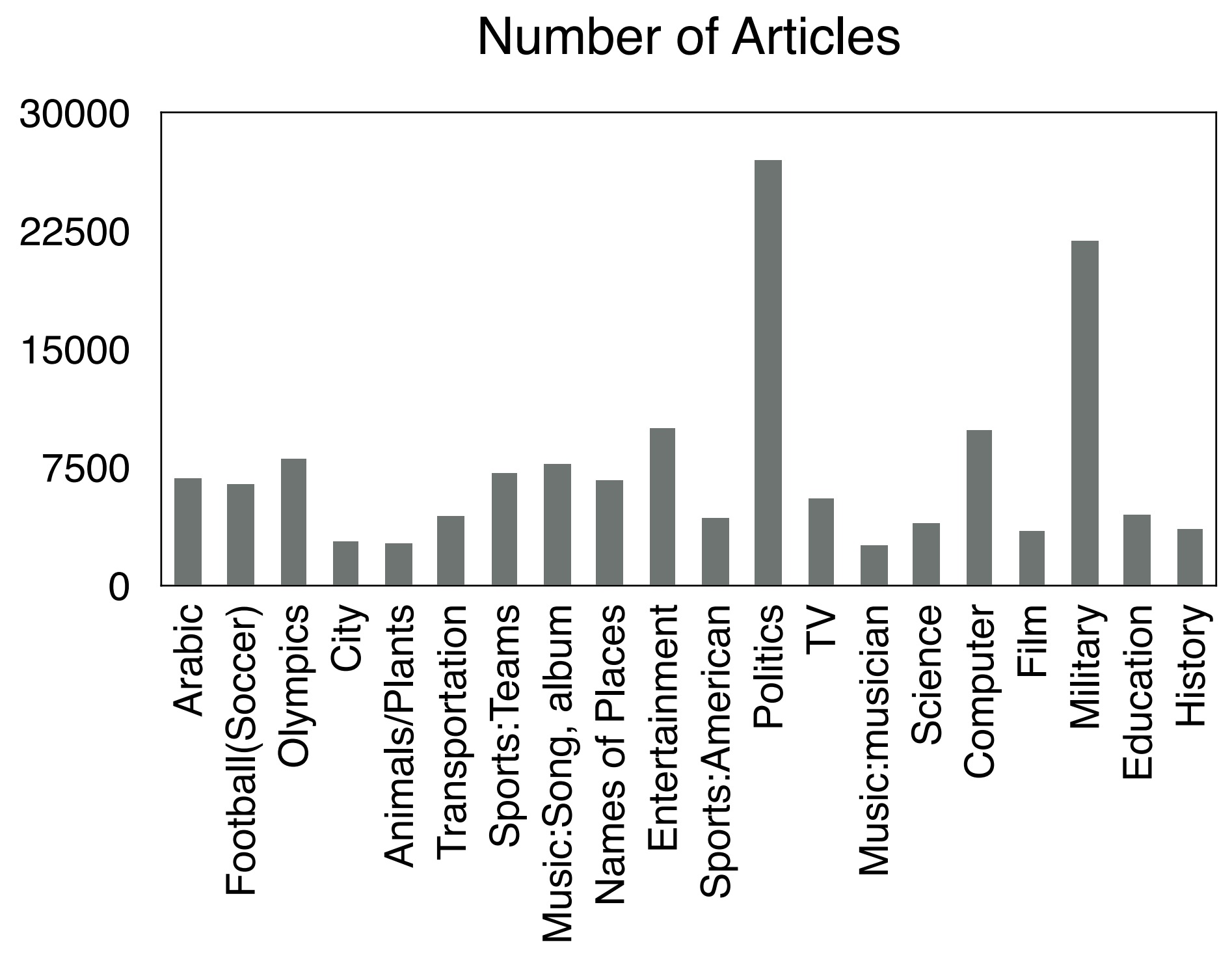
- Entropy of word frequency
- Average word rank
- Average word occurrence frequency
- Error rate

Syntactic Structure

- Entropy of POS frequency
- Mean phrase length (NP, VP)
- Mean parse tree depth

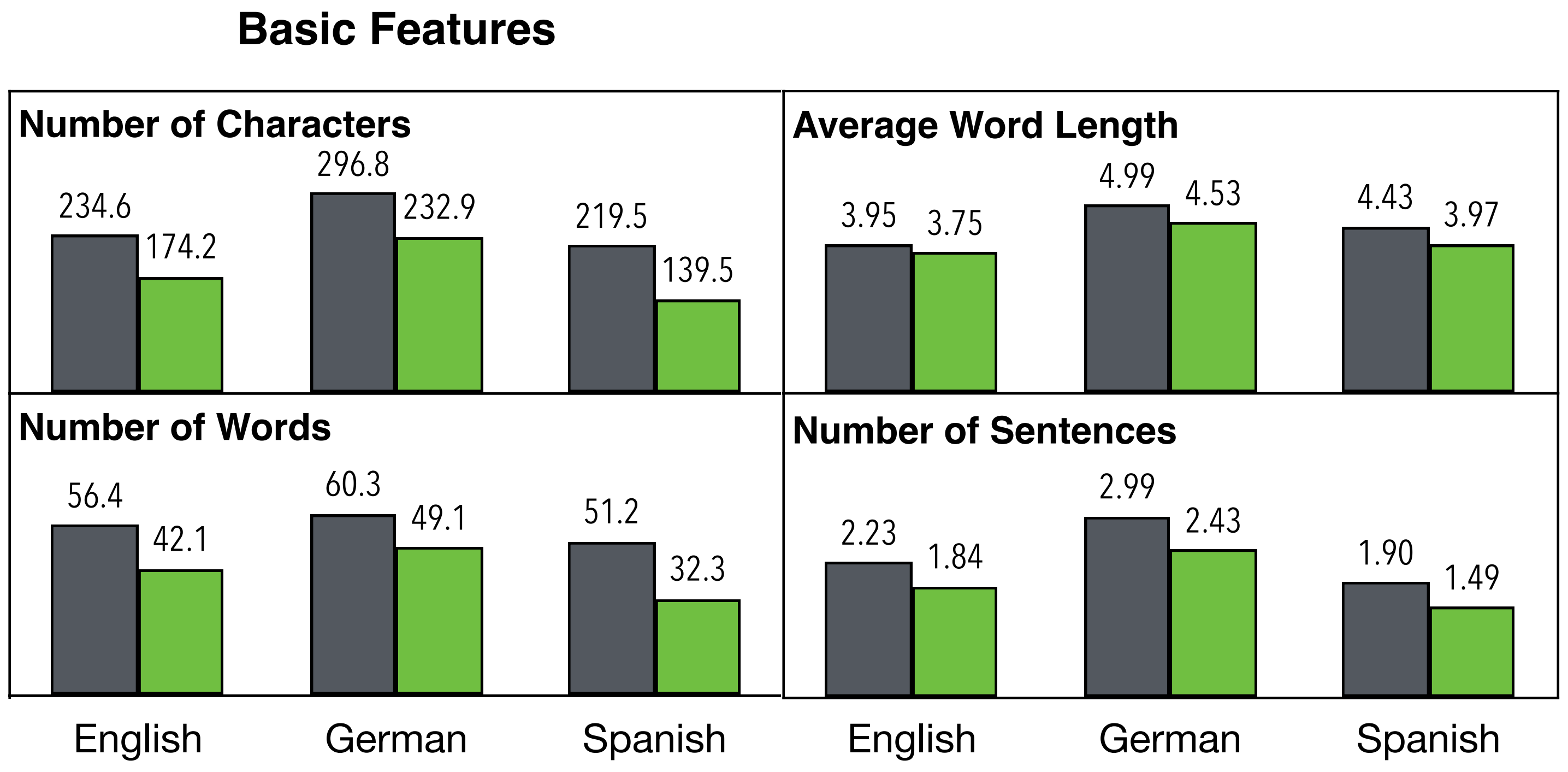
Topic Clusters



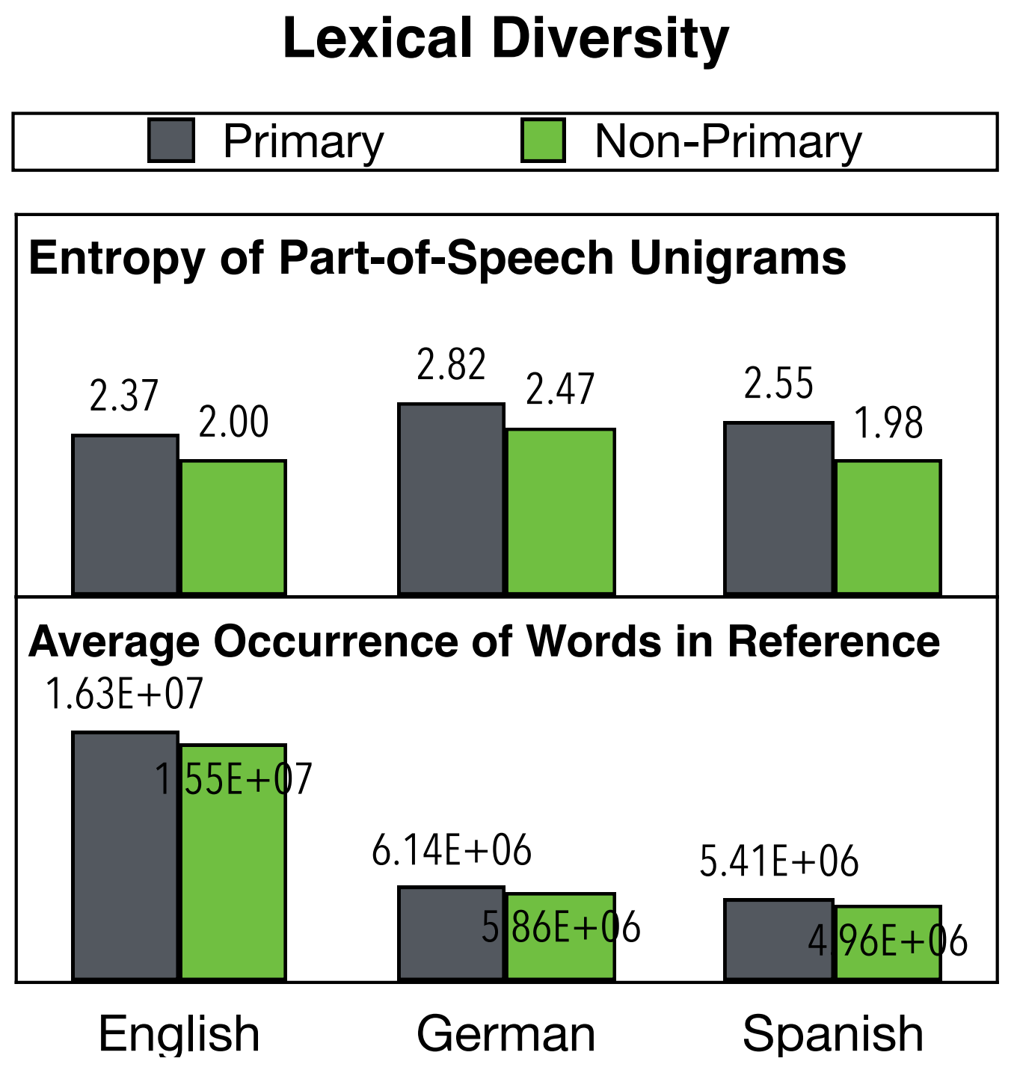


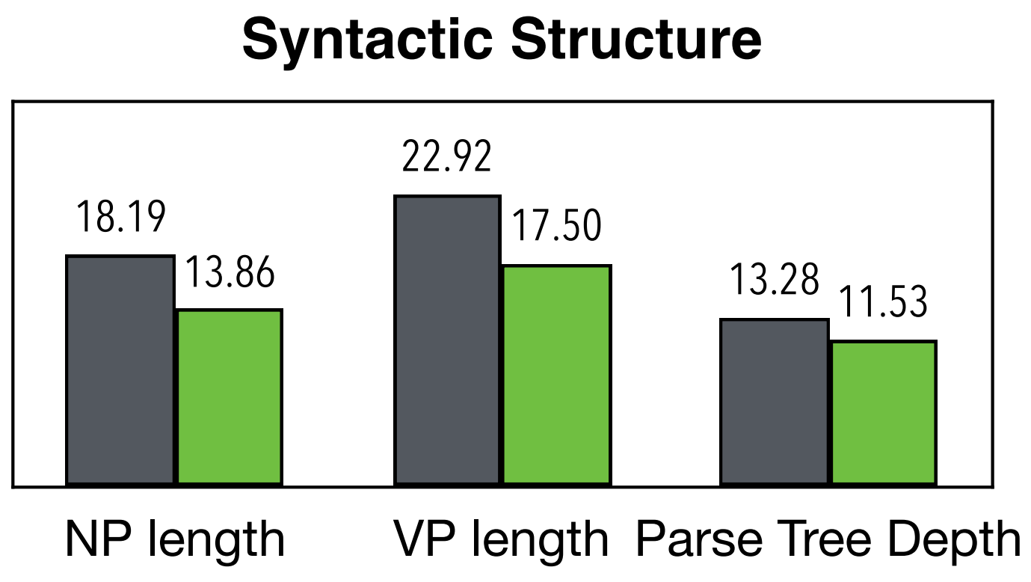
- **LDA+DBSCAN Algorithm**
  - Fit LDA over entire data with 100 topics
    - Each document is reduced to 100 dimensions
  - Cluster documents into 20 clusters with DBSCAN
- **Cluster-level rank of language complexity is consistent over measures**
  - Articles in *history* cluster are more complex Articles in *football* cluster are less complex
  - The complexity measures used are consistent
  - Different topics show different complexity

Results (per Paragraph)



- **Primary editors edit longer articles.**
  - Number of characters, words, average word length, sentences are always higher for the articles edited by primary users





- **Primary editors edit articles with higher lexical diversity.**
  - Entropy of n-gram is always higher in primary edits
  - Primary users edit articles of frequent yet diverse set of words
- **Primary editors edit articles with more complex syntactic structure.**
  - Parse-tree based measures
    - Length of longest noun/verb phrase
    - Depth of parse tree
  - Articles edited by primary users have complex syntactic structure