Linguistic Bias in Crowdsourced Articles and the Aspect of Time

Vangelis Mathioudis - University of Nicosia - COMP-592DL

Objective

- To compare if the biographies about females contain more subjective words than biographies about males in Wikipedia

What is linguistic bias?

Beukeboom’s definition [1]

A systematic asymmetry in the way that one uses language, as a function of the social group of the person(s) is being described.

What kinds of linguistic bias exist?

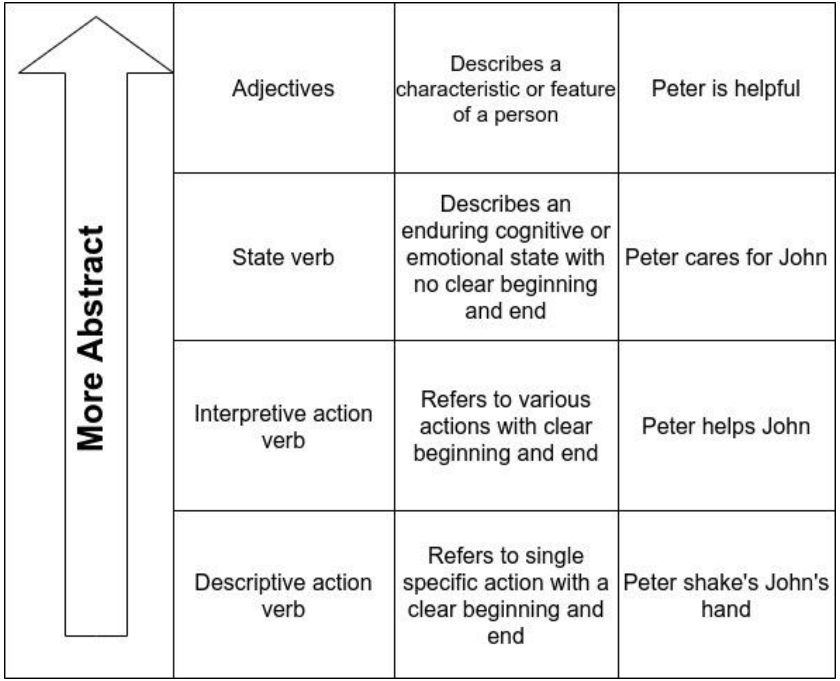
1. Linguistic Expectancy Bias
2. Linguistic Intergroup Bias

Both are built on the **Linguistic Category Model**

[1] C. Beukeboom, Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies., 01 2014, pp. 313–330.

Linguistic Category Model

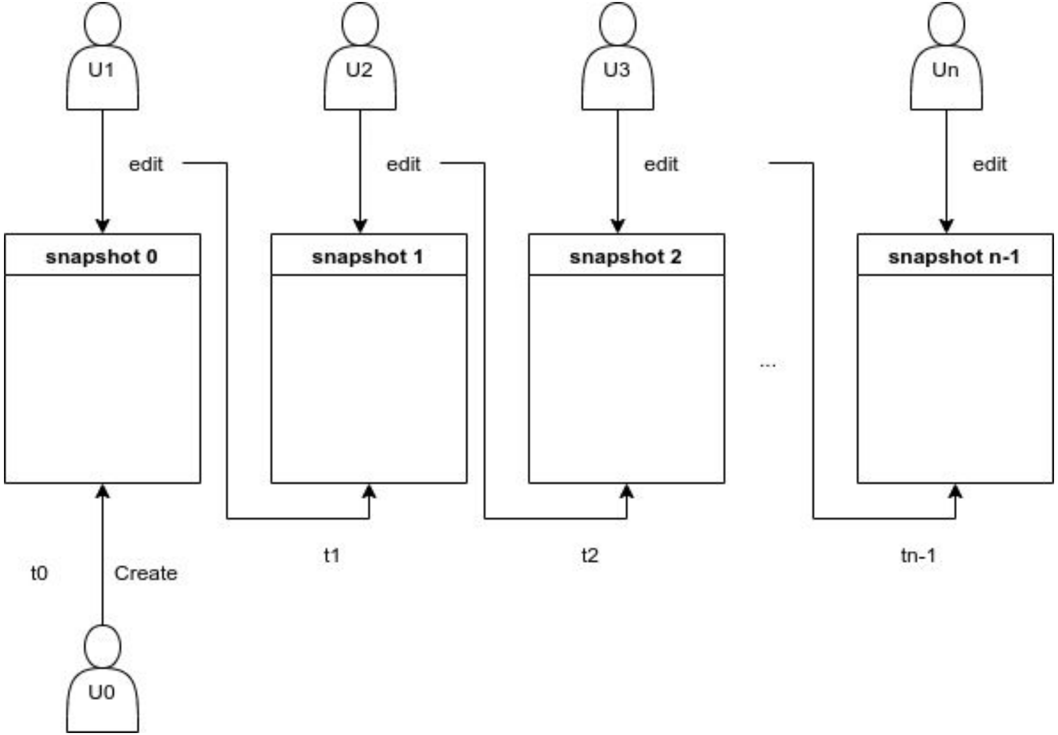
- The four level of abstraction for Linguistic Category Model based on Semin [1]



More Abstract	Adjectives	Describes a characteristic or feature of a person	Peter is helpful
	State verb	Describes an enduring cognitive or emotional state with no clear beginning and end	Peter cares for John
	Interpretive action verb	Refers to various actions with clear beginning and end	Peter helps John
	Descriptive action verb	Refers to single specific action with a clear beginning and end	Peter shake's John's hand

[1] Gün R. Semin. The Linguistic Category Model. SAGE Publications Ltd, 309–326. https://doi.org/10.4135/9781446249215.n16

Revisions in Wikipedia



Building our dataset (1)

Steps:

1. Build our dataset in the form of a python dictionary
   a. We download the history of revisions of biographies of 7 European politicians using the MediaWiki API and the library MWParserFromHell
   b. We keep track of the Revision Id, Users Id, Timestamp, Content and Tags
2. Drop any entry that doesn't have content
3. Plug the dataset in Didaxto[1] to create two domain specific dictionaries (sets) with positive and negative words

[1] https://deixto.com/didaxto/

Building our dataset (2)

- This is the structure of the dataset:

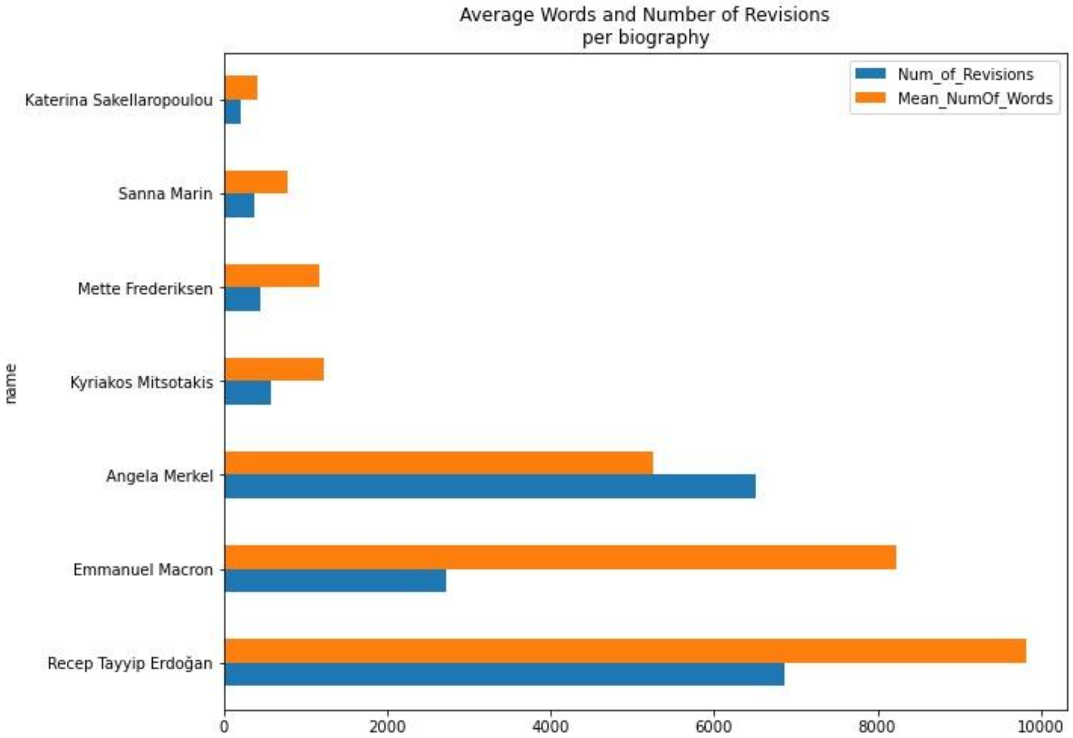
  ```
  {
  "name": {"revid": {"userid": userid,
                       "timestamp": timestamp,
                "content": biography's snapshot,
                "tags": tags that are associated with this revid},
            "revid2":{...}, . . . },
  "name2": {"revid": {...}, "revid2": {...}, . . .}
  }
  ```

Building our dataset (3)

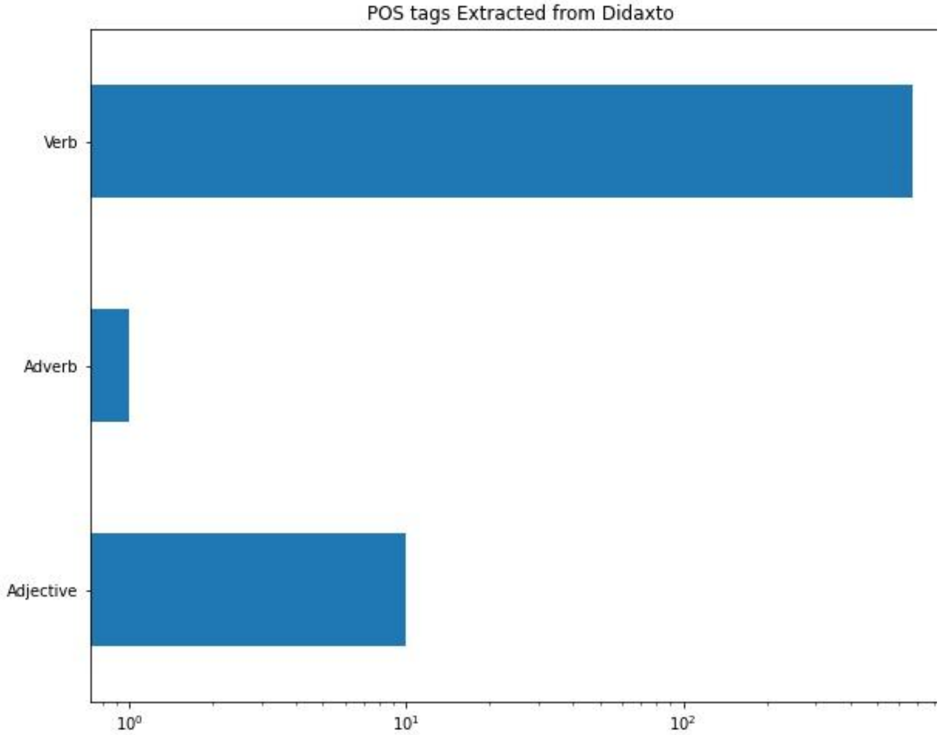
Steps:

- Build our Pandas dataframes by:
    - Extracting the number of verbs, adverbs, adjectives, positive, negative words and total words
    - Extracting ratios for positive, negative words and adjectives
    - Measuring the Mean Abstract Level
- Resample the dataset to daily and weekly periods

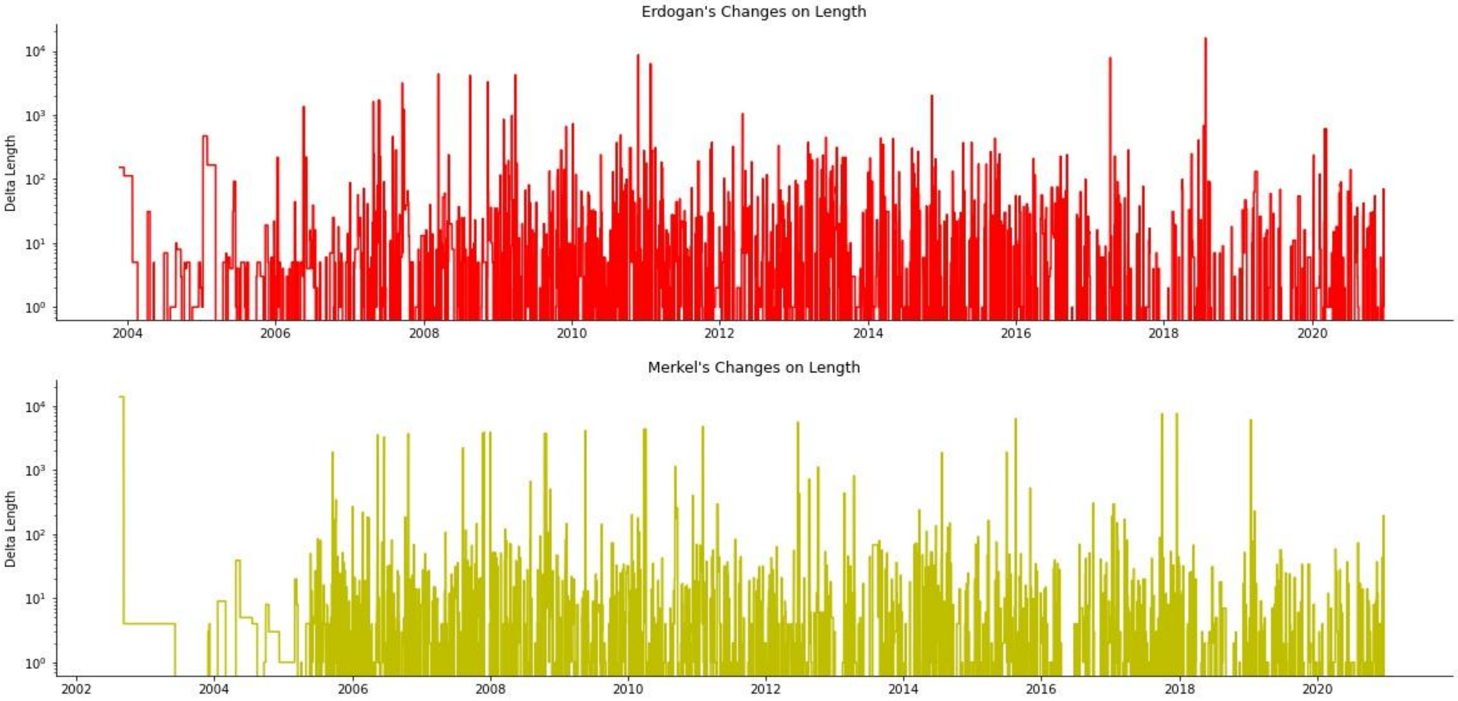
Exploratory Data Analysis (1)



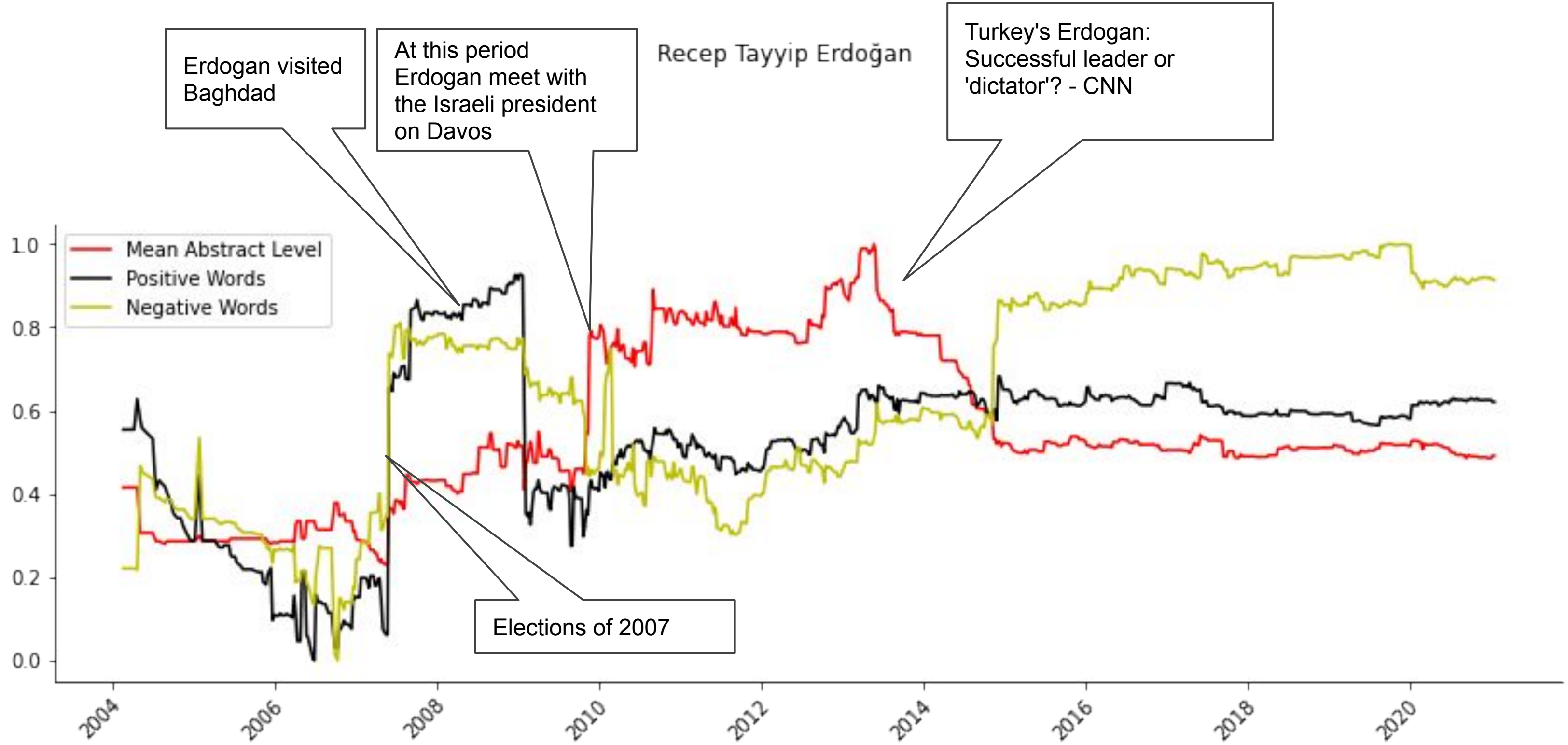
Exploratory Data Analysis (2)



Exploratory Data Analysis (3)



Weekly Changes of MAL, Positive and Negative Ratios



Erdogan visited Baghdad

At this period Erdogan meet with the Israeli president on Davos

Turkey's Erdogan: Successful leader or 'dictator'? - CNN

Elections of 2007

Hypothesis

- Null Hypothesis H_0: The average of female MAL score is equal to average of male MAL
- Alternative Hypothesis H_1: The average of female MAL score is higher than the average of male MAL

Results:

- Confidence level: 95%
- T-statistic = 4.06077
- p-value = $2*10^{-5}$

We conclude that the average Mean Abstract Level for females is different to the average Mean Abstract Level for males.

Future work

- Build a model to categorize verbs as “State”, “Interpretive” and “Descriptive Action Verbs” and use these to calculate MAL scores,
- Work with more biographies about people in diverse domains and across the world,
- Extend to historical events