In this study, I aim to identify and label meaningful text passages from (news website) articles where the information of a text passage could be used for constructing two different kinds of data visualization tools, a bar chart and a line graph. Labelling will be done after text analysis where we aim to find indicators described in a codebook derived by research and data analysis during a pilot study. During the data annotation process, a text passage will be labelled as (fit for) bar or line. I will do this in two different ways.   
  
For the bar annotation, I search for comparative syntactic constructions (https://en.wikipedia.org/wiki/Comparative) that serves to express a comparison between two (or more) entities or groups of entities in quality, degree or quantity in the article texts.   
  
A line chart is often used to visualize a trend in data over intervals of time (<https://en.wikipedia.org/wiki/Line_chart>). For line annotation, I try to identify trend assessment indicators (bar and line paper) that serves to express a trend between data points over time in the article texts. I will do this by following the rules below:

1. **Read the text sentence by sentence**

We start the annotation process by reading the article text sentence by sentence.

1. **Find a sentence about a comparison between two or more entities**

In this step we want to identify indicators described in the codebook. We want to find a sentence with comparison indicators; morphological, syntactical or correlative conjunctions a between two or more entities or trend assessment indicators as described in the codebook.

1. **Set the scope**

When we find a sentence which meets the conditions mentioned in step 2 (described in the theory-driven codebook), we have to set a scope on the text fragment we want to annotate.

We do this by analyzing the previous and next sentence(s) of the sentence we found. Does the previous or the next sentence give information about the comparison or a trend, we stretch the scope of the text fragment we want to annotate by including the useful sentences. Repeat this step until we do not find any information about the comparison.

1. **Identify the independent and the dependent variable**

“The first step to displaying your data is to identify the independent and dependent variables in your experiment” (<http://www4.ncsu.edu/~aelarsen/vet/display/index.html>) because both a bar graph and line chart make use of these two variables (bron nog bijvoegen). If we translate this into the annotation process where we determine how to transform raw information into meaningful text passages, we have to be sure that a text passage contains both variables.

1. **Data types of the variables**In this step we want to identify the kind data types (described in the data-driven codebook) of the variables described in the text fragment.
   1. **Categorical vs quantitative**

Information presented as bar graphs should be described categorically in terms of discrete comparisons using comparative terms and are useful for comparing frequencies or counts of one category to the next in a qualitative discrete data set. (bar and line paper, http://www4.ncsu.edu/~aelarsen/vet/display/index.html)

In this step, we have to check the assumption that the data type for the independent variable is categorical (nominal or ordinal) and the data type for the dependent variable is quantitative (numerical or continuous).

* 1. **Quantitative vs quantitative**

Information presented as lines should be described as (continuous) trends between data points (at equal intervals) (bron bar and line paper, boek: statistics for people who)

In this step, we have to check the assumption that the data type for both the independent and dependent variable is quantitative (numerical or continuous).

1. **Annotate**

If the text meets the conditions of the previous steps we annotate the text fragment as bar or line.

## Codebook

The development of the codebook has been done by creating theory-driven codes (Developing and Using a Codebook for the Analysis of Interview Data). These codes are generated from background literature that guide the research. These codes has been set up as baseline assumptions a text passage should met, if not, the text will not be annotated.

Tabel 1 Theory-Driven Codes, Definitions, and Examples

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Code | Description | Example |
| **Bar** | Morphological comparison | Morphological comparison uses the [suffixes](https://en.wikipedia.org/wiki/Suffix) *-er* (the "comparative") and *-est* (the "superlative"). They are typically added to shorter words, words of [Anglo-Saxon](https://en.wikipedia.org/wiki/Old_English_language) origin, and borrowed words which have been fully assimilated into the English vocabulary. This system also contains a number of irregular forms, some of which, like "good", "better", and "best", contain [suppletive](https://en.wikipedia.org/wiki/Suppletion) forms. | Good -> better -> best  Male's height is high**er** than that of females's |
|  | Syntactic comparison | The second system of comparison in English appends the [grammatical particles](https://en.wikipedia.org/wiki/Grammatical_particle) "more" and "most", themselves the irregular comparatives of "many" and "much", to the adjective or adverb being modified. This series can be compared to a system containing the [diminutives](https://en.wikipedia.org/wiki/Diminutive) "less" and "least". | Beautiful ->  more beautiful ->  most beautiful  B is bought **more** often that A |
|  | Correlative conjunction | [Correlative](https://en.wikipedia.org/wiki/Correlative) conjunctions work in pairs to join words and groups of words of equal weight in a sentence.  Pairs:   * either...or * as...as * … | Just as many Americans love basketball, so many Canadians love ice hockey.  X scored 300 points, as many as Y and Z scored 55 points during the match. |
|  |  |  |  |
| **Line** | Trend assessment/analysis | Trend descriptions use terms like function, relationship, correlation, varies, trend; the tend to refer to continuous changes in the variables.  It is the widespread practice of collecting information and attempting to spot a pattern. It could be used to estimate uncertain events in the past (or maybe in the future). | The graph shows a positive correlation between a child's increases in age and height between the ages of 10 and 12.  How many ancient kings probably ruled between two dates, based on data such as the average years which other known kings reigned. |
|  | Time series  (discrete (time order)) | A time series is a series of [data points](https://en.wikipedia.org/wiki/Data_point) indexed (or listed or graphed) in **time order.** Most commonly, a time series is a [sequence](https://en.wikipedia.org/wiki/Sequence) taken at successive equally spaced points in time. Thus it is a sequence of [discrete-time](https://en.wikipedia.org/wiki/Discrete-time) data.  A discrete variable over a particular range of real values is one for which, for any value in the range that the variable is permitted to take on, there is a positive minimum distance to the nearest other permissible value. The number of permitted values is either finite or countably infinite.  ([https://en.wikipedia.org/wiki/ Discrete\_time\_and\_continuous\_time](https://en.wikipedia.org/wiki/Discrete_time_and_continuous_time))  [https://en.wikipedia.org/wiki/ Continuous\_or\_discrete\_variable#Continuous\_variable](https://en.wikipedia.org/wiki/Continuous_or_discrete_variable#Continuous_variable) | When we take a look at the data we see an increasing trend of mobile phones connected to the internet **each year** from 1990 till 2000 |
|  | Time series  (continuous (variable)) | Time as a continuous variable is one which can take on infinitely many, [uncountable](https://en.wikipedia.org/wiki/Uncountable_set) values and depends on the context.  ([https://en.wikipedia.org/wiki/ Discrete\_time\_and\_continuous\_time](https://en.wikipedia.org/wiki/Discrete_time_and_continuous_time))  [https://en.wikipedia.org/wiki/ Continuous\_or\_discrete\_variable#Continuous\_variable](https://en.wikipedia.org/wiki/Continuous_or_discrete_variable#Continuous_variable)  <https://wirelesspi.com/continuous-time-vs-discrete-time-signals/> | We recorded the running speed of a soccer player at **each instant of time** during a 45 minutes match. When we take a look at the data we see that his maximum speed peaks at 34,23min. |

Tabel 2 Data-Driven Codes, Definitions, and Examples

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Code | Description | Example |
| Bar | Categorical vs quantitative reference(s) | Author makes direct/indirect or general/specific references that the independent variable (X-Axis) is categorical (nominal or ordinal) and the dependent variable (Y-axis) is quantitative | In a larger sample of 1,794 movies *(independent variable = movie)* released from 1970 to 2013, we found that only half had at least one scene *(dependent variable = # of scenes)* in which women talked to each other about something other than a man |
| Line | Quantitative vs quantitative reference(s) | Author makes direct/indirect or general/specific references that the Independent variable (X-Axis) is quantitative and the dependent variable (Y-axis) is also quantitative | Year can be a discretization of time. For example, you might have data for a child’s height on January 1 of years from 2010 to 2018. It’s meaningful to ask for height at (say) 2013.5, that would just be on June 30, 2018. So year is a discretized measure of a continuous interval variable, so quantitative. |

Reflection of pilot study:

After reading several articles, I realized that coding line by line and on the paragraph level were often not meaningful. The paragraph level often featured a combination of more sub-code categories from both coding categories, making it impossible to label the text with only one code. Based on this, I decided to code on the sentence level.

Annotation process:

