Bar charts are useful for comparing frequencies or counts of one category to the next in a qualitative discrete data set and are useful to visualize categorical data (boek statistiek). For example, the height of each bar represents the frequency of the corresponding category. Line graphs are helpful for illustrating trends in continuous data. For example, a line graph could communicate whether or not a patient’s survival time increases or decreases with a given treatment. It is important that each observation is collected at equal intervals (http://www4.ncsu.edu/~aelarsen/vet/display/index.html).

In this study, I aim to identify and label meaningful text passages from (news website) articles where the information of the text could be used for constructing two different kinds of data visualizations, a bar chart and a line graph.

For this, I make use of a rule-based approach with the use of a codebook derived by derived by research and data analysis during a pilot study. During this approach we read the whole article text as input, on the text we perform text segmentation, (http://www4.ncsu.edu

/~aelarsen/vet/display/index.html) say that “The first step to displaying your data is to identify the independent and dependent variables in your experiment”, so in the next step we want to perform a variable identification where we identify the independent and dependent variable.

After that, we want to identify the data types of the variables summarized in table 2, data-driven codebook, because a bar chart is often visualized by a categorical and a quantitative data type variable where a line chart is often visualized by variables of two quantitative data type variables (http://www4.ncsu.edu/~aelarsen/vet/display/index.html + https://onlinecourses.science.psu.edu/stat500/node/113/ + Bronnen toevoegen).

(http://www4.ncsu.edu/~aelarsen/vet/display/index.html) and (statistiek boek) say that nominal and ordinal variable values have a qualitative/categorical data type, and numerical (interval) and continuous (ratio) variable values have a quantitative data type so at this stage we want to identify the data types of the variables by looking at the variable values.

The next step is the identification of how the text is represented for bar or line classification looking at indicators described in the codebook. During this research, we classify nominal, ordinal and numerical variable values as discrete variable types where the units of measurements cannot be split up and continuous variable values as a continuous variable type where the scale of measurement is meaningful at all points between the measurements (bron http://www4.ncsu.edu/~aelarsen/vet/display/index.html , boek statistiek).

During the text representation analysis for the bar annotation, I search for (discrete variable type) 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) therefore I try to identify (continuous variable type) trend assessment indicators (bar and line paper) that serves to express a trend between data points over time in the article texts. In this research, a meaningful text fragment contains at least an independent and dependent variable and bar/line indicators.

The output will be a text fragment labelled as (fit for) bar and/or line when meaningful or labelled as not meaningful for a bar or line visualization.

A detailed explanation of the rule-based approach annotation method: from raw into meaningful text:

1. **Read the text**

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

1. **Text segmentation by topic**

We do this by analyzing the text and perform topic segmentation, which aims to find the boundaries between topic blocks in a text.

* Automatic or manually?

After the segmentation, we do variable identification on the text fragments.

1. **Variable identification**In this step, we perform variable identification where we want to identify the independent and dependent variables required for visualization (<http://www4.ncsu.edu/~aelarsen/vet/display/index.html>) and we want to identify the data types of the variables depending on the graph for visualization (bar or line) (bron: boek statistiek, nog 1 van categorical vs quantitative…).   
   1. **Independent and dependent variables**

We want to identify at least one independent and a dependent variable in a single text fragment because both a bar graph and line chart make use both variables for data visualization.

We do this by looking at:

* An independent variable that represents a quantity that is being manipulated in an experiment.
* A dependent variable that represents a quantity whose value depends on how the independent variable is manipulated.
* Does it need to be written explicitly?
* Can we find variables in the text by looking at references?

<https://www.khanacademy.org/math/pre-algebra/pre-algebra-equations-expressions/pre-algebra-dependent-independent/a/dependent-and-independent-variables-review>

* 1. **Data types of the variables**We want to identify the kind of variable data types by looking at the variable values as described in the data-driven codebook (table 2) for variable identification by analyzing the variables found 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. **Textual indicators**

Bar and line chart assumptions: (bron zoeken: data visualization, comparison/trend)   
As bar charts are useful for comparisons and line graphs are helpful for illustrating trends we want to identify comparison or trend indicators inside the text fragments and set the assumption that there must be at least one indicator in the text?. We want to find indicators like; morphological, syntactical or correlative conjunctions a between two or more entities or trend assessment indicators as described in the codebook.

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" \o "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, Examples, and variable information

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Category | Code | Description | Example | Variable information | |
| 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, variable value = nominal (discrete movies), data type = categorical)* released from 1970 to 2013, we found that only half had at least one scene *(dependent variable = scenes, variable value = numerical (#), variable data type = quantitative)* in which women talked to each other about something other than a man | 1.  Variable:  Kind:  Data type:  Value:  2.  Variable:  Kind:  Data type:  Value: | Movie  Independent  Categorical  Nominal  Scenes  Dependent  Quantitative  Numerical |
| 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:

