ANNOTATION GUIDELINES FOR THE BAR CHART DATASET OF PLOTS AND THEIR CROWDSOURCES SUMMARIES

Contributors: Rudy Khalil, Iza Škrjanec, Emeka Udeh, Vera Demberg

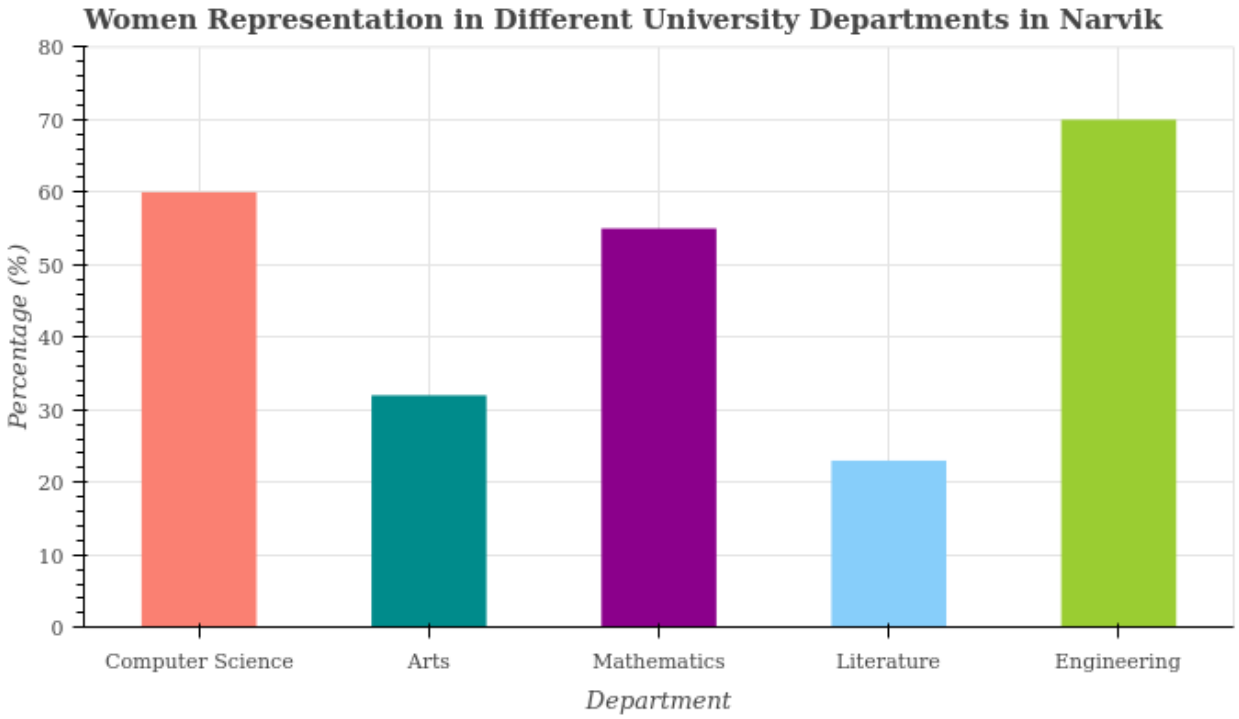
These guidelines present the set of labels used for annotating phenomena in summaries of bar charts. Previous versions of the guidelines can be found here: version 1.0 and version 1.1.

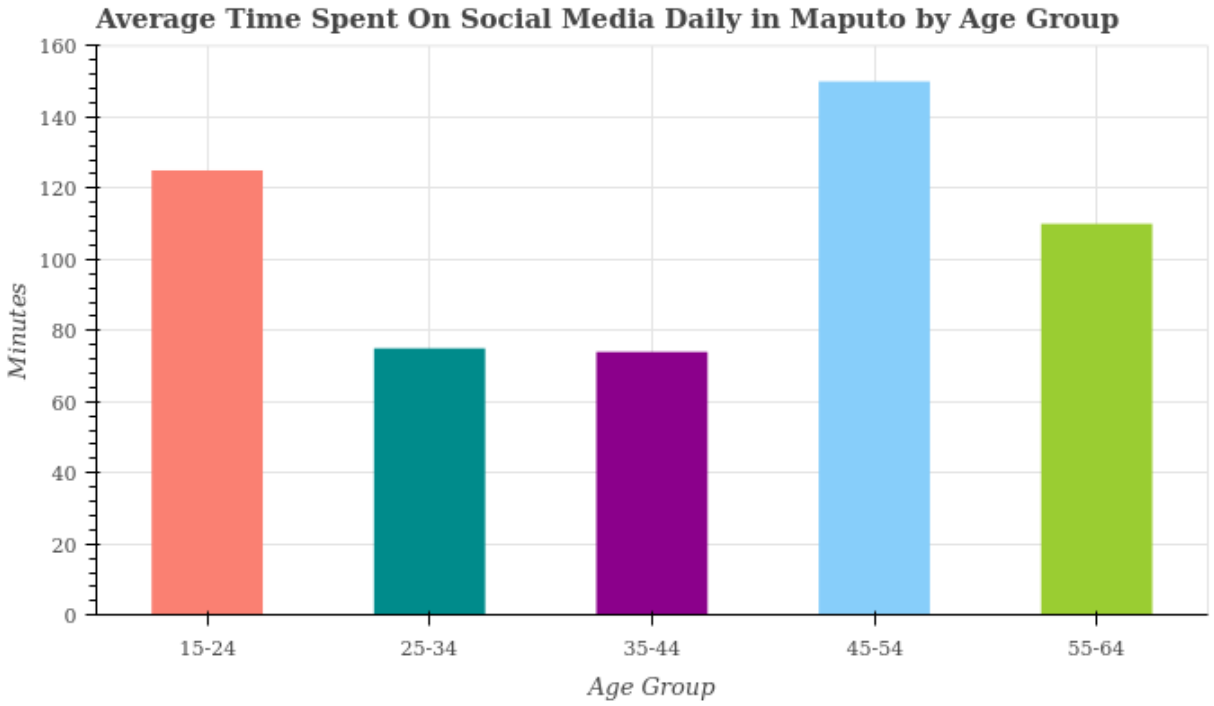
To collect data, we first generated bar chart plots and then collected their summaries via crowdsourcing. The current dataset has 47 plots, and for each of these, around 23 summaries.

When designing a tagging system, our main principle was to have a manageable set of labels which would help us 1) track discourse phenomena of chart summaries relevant to NLG, 2) delexicalize chart-specific tokens so as to make the input data more general, but also to relexicalize the output accordingly.

Each label roughly belongs to one of the following types of labels:

1) Chart-specific: names and heights of bars, other like the chart title and related vocabulary
2) Discourse: words/phrases typical of chart summaries, but not dependent on particular charts and their data
3) Interpretation: any kind of notions that are not grounded in the chart, be it interpretation of the data or non-relevant comment about the chart ("but the chart doesn't tell us which year the numbers were measured")


Women Representation in Different University Departments in Narvik
Percentage (%)
0
10
20
30
40
50
60
70
80
Computer Science
Arts
Mathematics
Literature
Engineering
Department


Average Time Spent On Social Media Daily in Maputo by Age Group
Minutes
0
20
40
60
80
100
120
140
160
15-24
25-34
35-44
45-54
55-64
Age Group

1) x_axis

References to the name of the x axis is labeled as x_axis. The reference may be verbatim as in the chart or taking a different surface form.

Examples for chart 1: department, departments, university departments, courses
Examples for chart 2: age group, age groups, group, groups, age range, user group

2) y_axis

This label accounts for the quantity presented on the y axis. This may or may not overlap with the name of the y axis in the chart.

Chart 1: percentage*, amount, share of women, women representation
Chart 2: time, time spent on social media, minutes*

*In case the name of the y axis and its unit overlap, pay attention to how a word is used in context. For example:

"The percentage [y_axis] of women varies across different departments. In computer science it is 60 percent [y_axis_inferred_label].

3) x_axis_label_*_value

References to bar names are annotated with these labels. Currently, we code each bar name in terms of its rank given the height. For charts with an ordinal X variable, we are considering coding in terms of rank given the x axis (order of appearance in the chart).

The label includes the rank information as follows:

- the highest bar: x_axis_label_highest_value
- second highest: x_axis_label_Scnd_highest_value
- third highest: x_axis_label_3rd_highest_value
- fourth highest: x_axis_label_4th_highest_value
- …
- least/lowest one: x_axis_label_least_value

Note that the label for the lowest bar doesn't include the rank number, but rather "least". If a chart has 4 bars of different heights, the subset of labels to be applied is {x_axis_label_highest_value, x_axis_label_Scnd_highest_value, x_axis_label_3rd_highest_value, x_axis_label_least_value}.

The reference to the bar may or may not match the bar name completely.
Chart 1, x_axis_label_3rd_highest_value: mathematics, maths, math
Chart 2, x_axis_label_Scnd_highest_value: 15-24, between 15 and 24, from 15 to 24

In case two or more bars are of exactly the same height, they should carry the same label. TODO check and give examples

4) y_axis_*_value (_val)

This label annotates bar heights. More specifically, the exact values of bar heights as they appear in the plotting data and in the chart itself.

The label names include the height rank:

- y_axis_highest_value_val
- y_axis_Scnd_highest_value
- y_axis_3rd_highest_value
- …
- y_axis_least_value_val

Note that the label for highest and lowest heights have a suffix _value_val.

Chart 1:

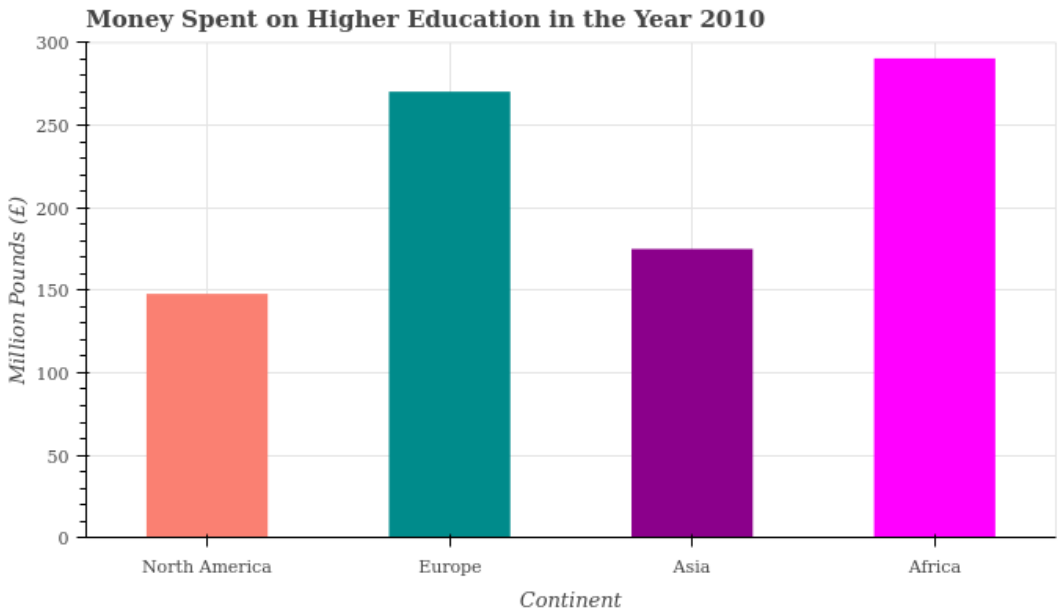
- y_axis_highest_value_val: 23
- y_axis_Scnd_highest_val: 32
- y_axis_3rd_highest_val: 55
- y_axis_4th_highest_val: 60
- y_axis_least_value_val: 70

For some charts, the y axis ticks mark lower integers, while the name of the y axis gives information about their magnitude (e.g. thousand or million). In that case, we still annotate the exact bar height, splitting the magnitude and labeling it separately.

For a description applying to the bottom chart, the labeling for this sentence is as follows:

The amount of money spent on Higher Education in 2010 in Asia was £ 175,000,000.

The
amount of money <y_axis>
spent
on
Higher Education in 2010 <topic_related_property>
in
Asia <x_axis_label_3rd_highest_value>
was
£ <y_axis_inferred_label>
175 <y_axis_3rd_highest_val>
,000,000 <y_magnitude>



TODO: if two bars are of the exact same height, do they share the label too?

5) y_axis_inferred_*_value_approx

Crowdsourcers often describe bars in terms of inexact heights. This could be because the top of the bar is not aligned with a tick mark stating the value. The bar heights are then inferred from the chart and given with an approximated value. Such instances are annotated with y_axis_inferred_*_value_approx.

The label names encode the height rank:

- y_axis_inferred_highest_value_approx
- y_axis_inferred_Scnd_highest_value_approx
- …
- y_axis_inferred_least_value_approx

The
least <y_axis_least_value>
amount of time <y_axis>
was
spent
by
the
35-44 <x_axis_label_least_value>
age group <x_axis>
totalling
68 <y_axis_inferred_least_value_approx> ## actual bar height: 74
minutes <y_axis_inferred_label>
For treating magnitudes, the same rule as for y_axis_*_value (_val) applies.

6) y_axis_inferred_label

The quantity on the y axis is measurable in particular units, which are often given as the whole or a part of the name of the y axis. The label used should be y_axis_inferred_label.

The unit can be verbalized as a word ("pounds") or with a symbol ("£").
For chart 1, the label can be applied to e.g. "percents" or "%".
For chart 2, to "minutes", "min", "mins", "hours".
For chart 3, "pounds", "pound", "£".

Note that sometimes participants use a wrong currency, e.g. dollar instead of pound. We label such instances with y_axis_inferred_label and correct the token, keeping both, the incorrect and corrected one.

7) y_magnitude

To keep the chart clear and tick marks on the y axis visible, the tick marks are often smaller numbers, while the name of the y axis gives information about their magnitude.

See chart 3 as an example. The highest tick mark goes to 300, the magnitude is million.
Magnitudes can be verbalized in different forms:

- thousand: "thousand", "thousands", "k", "000", ",000"
- million: "million", "millions", "m", "M", "000000", ",000,000"

If the bar height appears together with the magnitude, we split them and label accordingly. See below:

It actually decreased to nearly £ 19,500 by the year 2010.

it
actually
decreased <slope_down>
to
nearly <y_axis_approx>
£ <y_axis_inferred_label>
19.5 <y_axis_inferred_3rd_highest_value_approx>
000 <y_magnitude>
by
the
year <x_axis>
2010 <x_axis_label_3rd_highest_value>

last was North America who only spent nearly £ 150 m on higher education.

last <order_last>
was
North America <x_axis_label_least_value>
who
only <y_axis_least_value>

spent
nearly <y_axis_approx>
£ <y_axis_inferred_label>
150 <y_axis_inferred_least_value_approx>
m <y_magnitude>
on
higher education <topic_related_property>

8) y_axis_inferred_value_add_v1=*_v2=*

When the relative relation between bar heights is calculated via addition/subtraction, we label the height value as y_axis_inferred_value_add.

The label for additive relation assumes two bar heights are involved, v1 and v2. The relative bar height d is calculated, such that v1 - v2 = d. The given relative height does not have to be exactly equal to d; it may be an approximation.

The attributes v1 and v2 get as values the names of bars, which are coded in terms of their height rank. For example, y_axis_inferred_value_add_v1=Scnd_v2=4th annotates the relative height that we get from height_Scnd minus height_4th.

Insurance has the highest representation at 65% whereas law firms has the lowest at 35% - a 30% difference.

Insurance <x_axis_label_highest_value>
has
the
highest <y_axis_highest_value>
representation <y_axis>
at
65 <y_axis_highest_value_val>
% <y_axis_inferred_label>
whereas <y_x_comparison>
law firms <x_axis_label_least_value>
has
the
lowest <y_axis_least_value>
at
35 <y_axis_least_value_val>
% <y_axis_inferred_label>
-
a
30 <y_axis_inferred_value_add_v1=highest_v2=least>
% <y_axis_inferred_label>
difference

Those in the 15-24 age bracket spent the most time on social media, clocking over 100 minutes more than those in the 55-64 age bracket.

Those
in
the
15-24 <x_axis_label_highest_value>
age bracket <x_axis>
spent
the
most <y_axis_highest_value>
time on social media <y_axis>
,
clocking
over <y_axis_approx>
100 <y_axis_inferred_value_add_v1=highest_v2=least>
minutes <y_axis_inferred_label>
more <y_axis_trend_up>
than
those
in
the
55-64 <x_axis_label_least_value>
age bracket <x_axis>
.

the following group 25-34 spend about 160 minutes, 20 less that the previous group
the
following <order_Scnd>
group 25-34 <x_axis_label_Scnd_highest_value>
spend
about <y_axis_approx>
160 <y_axis_inferred_Scnd_highest_value_approx>
minutes <y_axis_inferred_label>
,
20 <y_axis_inferred_value_add_v1=Scnd_v2=highest>
less <y_axis_trend_down>
that
the
previous group <x_axis_label_highest_value>

9) y_axis_inferred_value_mul_v1=*_v2=*

When the relative relation between bar heights is calculated via multiplication/division, we label the height value as y_axis_inferred_value_mul.

Similarly as for _add_, we assume two bars are involved in the relation. The relative height k is calculated as follows: v1 / v2 = k. The given relative height does not have to be exactly equal to k; it may be an approximation.

Note that k can be expressed as a number (2), a word (twice) or a combination (2 times).

STEM subjects are half as popular as Philosophy with 15% opting for this.

```
STEM subjects    <x_axis_label_least_value>
are
half    y_axis_inferred_value_mul_v1=least_v2=Scnd>
as
popular
as
Philosophy    <x_axis_label_Scnd_highest_value>
with
15    <y_axis_least_value_val>
%    <y_axis_inferred_label>
opting
for
this
```

In law firms this is 3 times higher than in tech where only 20% are women.

```
In
law firms          <x_axis_label_highest_value>
This
is
3 times <y_axis_inferred_value_mul_v1=highest_v2=least>
higher   <y_axis_trend_up>
than
in
tech      <x_axis_label_least_value>
where
only
20        <y_axis_least_value_val>
%         <y_axis>
are
woman  <topic_related_property>
```

10) slope_x_value

When the variable on the x axis is interval/ratio or even ordinal, the slope of the regression line can be calculated and described in the chart summary.

Typically, the label slope_x_value tells us the step size on the x axis in the slope expression. The slope value usually applies to the entire data, not a subset. The given step size can be exact or an approximation.

Take for example the following data showing the number of fatal injuries in the span of 5 years:

- x: '2012', '2013', '2014', '2015', '2016'
- y: 30, 25, 16, 15, 12

- We can see that the slope is negative; the regression line is falling; the number of injuries drops year by year
- In the example sentence below, the slope is 5 injuries per 1 year. Note that "year" here means "one year", so it will be annotated this way.

The trend decreases from 30 in 2012 to 13 in 2016, with an average decrease of 5 per year.

The
trend <y_axis_trend>
decreases <slope_down>
from
30 <y_axis_highest_value_val>
in
2012 <x_axis_label_least_value>
to
13 <y_axis_inferred_least_value_approx>
in
2016 <x_axis_label_least_value>
,
with
an
average <y_axis_trens>
decrease <slope_down>
of
5 <slope_y_value>
per
year <slope_x_value>

11) slope_y_value

See slope_x_value. Similarly, we annotate the step size on the y axis. The size can be exact or an approximation
In the example above, the step size is -5: 5 death less per year.

The trend decreases from 30 in 2012 to 13 in 2016, with an average decrease of 5 per year.

12) x_axis_range_start and x_axis_range_end

The markers of spans between bars are labeled with the range start and range end label. This applies especially to charts with a temporal variable on the x-axis.
The typical pairs of start and end markers are: between-and, from-to, from--

The
significant
results
show

that
the
number of fatal injuries <y_axis>
has
been
declining <slope_down>
from <x_axis_range_start>
year <x_axis>
2012 <x_axis_label_highest_value>
up
to <x_axis_range_end>
year <x_axis>
2016 <x_axis_label_least_value>

the
number of fatal injuries at the Pula Steel factory <topic>
from <x_axis_range_start>
2012 <x_axis_label_highest_value>
- <x_axis_range_end>
2016 <x_axis_label_least_value>

13) x_interval

The label x_interval can apply to charts with an ordinal/interval/ratio variable on the x-axis. It is currently used in 3 different contexts:

a) Referring to the total interval on the x_axis (between the first and the last bar)
b) Referring to the interval between two bars
c) Referring to the interval within a bar (for bars representative of time interval)

Note that it does not annotate slope values.

Example for a)
This chart represents the representation of minorities in the Parliament of Libya over <u>two decades</u> from 1990 to 2019.

This
Chart
represents
the
representation of minorities in the Parliament of Libya <topic>
over
two decades <x_interval>
from <x_axis_range_start>
1990 <x_axis_label_highest_value>

to <x_axis_range_end>
2019 <x_axis_label_5th_highest_value>

Example for b)
The graph shows that the median salaries of women have risen each half decade.
The
graph
shows
that
the
median salaries of women <topic>
have
risen <slope_up>
each
half decade <x_interval>

Example for c)
This graph shows the average time spent on social media by age group in Maputo. Five 10 year spans from 15 to 64 are given.

This
graph
shows
the
average time spent on social media by age group in Maputo <topic>
.
Five <x_axis_labels_count>
10 year <x_interval>
spans <x_axis>
from <x_axis_range_start>
15 <x_axis_label_highest_value>
to <x_axis_range_end>
64 <x_axis_label_least_value>
are
given

14) x_axis_labels_count

Explicit references to the number of bars in the chart are labeled as x_axis_labels_count. This applies only to the exact number of bars; whereby the numeral can be written as an integer ("5") or spelled out ("five").

The
graph
details
the

gender pay gap <y_axis>
in
three <x_axis_labels_count>
European <t=0_a=x>
countries <x_axis>

15) topic

The topic label largely fits the title of the chart. The title often includes the names of the x- and y-axis, and their units, but this varies across charts. This typically appears at the beginning of a description as an introductory sentence.

It is important to note that the topic label encompasses only what appears in the title and the synonyms.

Chart title: *Women representation in different university department in Narvik*
The
chart
shows
the
percentage <y_axis>
of
women in different university departments in Narvik <topic>

Chart title: *What causes obesity in Kiribati*
This
chart
looked
at
causes of Obesity in Kiribati <topic>

Chart title: *Women representation in different sectors in Benoni*
The
following
graph
shows
the
female representation in different work sectors in Benoni <topic>

Note that the title is often broken down into parts in the description. The parts are then labeled accordingly as topic or with other fitting labels.

Chart title: Average time spent on social media daily in Maputo by age group
Daily social media use by time in Maputo <topic>
is
shown

here
by
age group <x_axis>

16) group_* [in progress]

This label can account for cases when participants join bars and talk about them as a group. The label should include the names of bars which are inside the group. The names have been coded given their height so far.

For example, group_Scnd_3rd_4th is a group consisting of the second, third and the fourth highest bar.

So far, the second batch of data has shown examples of:

a) **conceptual grouping**: the bars share some hypernym-like concept. For example, "lifestyle choices" (referring to the bars of *Fast Food* and *Lack of Exercise*, as opposed to *Genetics* as causes of obesity)
b) **grouping by similar height**: bars have a similar height, so their descriptions share a predicate.
c) **grouping by different heights**: based on our observations, this applies to bar charts with an ordinal X variable. So far, we've seen the following case: there was a similarly large height difference between 4 bars (2 bars at the beginning of the x axis, 2 bars at the end). This is the single case so far, so we might not stick to it.
d) ** **one versus rest**: highest/lowest bar versus rest: this grouping might appear when the isolated bar has a very different height than the rest of the bars.
e) **Bars of exact same height**:

Example for b):

The lowest users by time are in the 25-34 and 35-44 age groups, which use social media for around 75 minutes each day on average.

The
lowest <y_axis_least_value>
users <x_axis>
by
time <y_axis>
are
in
the
25-34 and 35-44 <group_4th_least>
age groups <x_axis>
,
which
use
social media <topic_related_object>
for
around <y_axis_approx>

75 <group_y_4th_least>
minutes <y_axis_inferred_label>
each day ##
on average

Another example for b)

It indicates a downward trend in fatalities between 2013 and 2015.

It
indicates
a
downward <slope_down>
trend
in
fatalities <y_axis>
between 2013 and 2015<group_highest_3rd_least>

Example for d) based on this chart:

We can see that a high percentage choose STEM whilst a much lower percentage choose literature, philosophy and medicine.

We
can
see
that
a
high <y_axis_highest_value>
percentage <y_axis>
choose
STEM <x_axis_label_highest_value>
whilst <y_x_comparison>
a
much <y_axis_trend>
lower <y_axis_least_value>
percentage <y_axis>
choose
literature, philosophy or medicine <group_Scnd_3rd_least>

Example for e)

with literature and medicine both at 20%

with
literature and medicine <group_Scnd_3rd>
both
at
20 <group_y_Scnd_3rd>
% <y_axis>

17) group_y_*[in progress]

When a subset of bars is treated as a group, their heights are usually described in a single predicate as well. This group height can be the mean of heights or even a sum.

The label name is similar to the group_* label with the only difference that it indicated the y: group_y_*

We are not sure yet if the label name should also include information on how the group height is calculated: mean, sum or something else.

In the example for the group_* label, the group height is the mean.

The lowest users by time are in the 25-34 and 35-44 age groups, which use social media for around 75 minutes each day on average.

18) y_mean

The label marks references to the mean height of all the bars in the plot.

Fatal injuries at the Pula Steel Factory have varied from 2012 to 2016, averaging above 25.

Fatal injuries <y_axis>
at
the
Pula Steel Factory <t=1_a=b>
have
varied <y_axis_trend>
from <x_axis_range_start>
2012 <x_axis_label_Scnd_highest_value>
to <x_axis_range_end>
2016 <x_axis_label_4th_highest_value>
,
averaging
above <y_axis_approx>
25 <y_mean>

19) y_axis_highest_value

The vocabulary (words or phrases) describing the highest bar without providing the bar height are labeled as y_axis_highest value. This includes expressions of superlatives and majority.

Some examples include: "highest", "largest", "biggest", "most popular/common/frequent".

Note that the phrases might get between labels. In general, determiners are not included into the label span. For example, "the most common cause of obesity" is a noun phrase, which will be labeled as follows:

- the (no label)
- most common <y_axis_highest_value>
- cause of obesity <x_axis>

Note that this label annotates expressions that pre-modify bar names, the name of the y axis, or some topic-related entity.

More examples

most young people spend their evenings reading a book
most <y_axis_highest_value>
young people <topic_related_property>
spend
their evenings <topic_related_property>
reading a book <x_axis_label_highest_value>

the most popular choice is reading a book
The
most popular <y_axis_highest_value>
choice <x_axis>
is
reading a book <x_axis_label_highest_value>

Africa has the highest amount at £ 290 million
Africa <x_axis_label_highest_value>
has
the
highest <y_axis_highest_value>
amount <y_axis>
at
£ <y_axis_inferred_label>
290 <y_axis_inferred_highest_value_approx>
million <y_magnitude>

20) y_axis_least_value

The lowest bar or its height reference may be modified by an expression that is labeled with y_axis_least_value. Such expressions include "lowest", "smallest", "least common/popular".

See y_axis_highest_value for general principles for the span of the label.
Some examples:

in 2015 it was the lowest at £ 16 k
in

2015 <x_axis_label_least_value>
it
was
the
lowest <y_axis_least_value>
at
£ <y_axis_inferred_label>
16 <y_axis_least_value_val>
k <y_magnitude>

Literature has the <u>fewest</u> women
Literature <x_axis_label_least_value>
has
the
fewest <y_axis_least_value>
women <topic_related_object>

As it has been observed in the data, the bar height of the lowest bar is often pre-modified by "only" or similar expressions. We label them as y_axis_least_value as well.

the group aged 35-44 with <u>only</u> 65 minutes a day
the
group <x_axis>
aged
35-44 <x_axis_label_least_value>
with
only <y_axis_least_value>
65 <y_axis_inferred_least_value_approx>
minutes <y_axis_inferred_label>
a day <topic_related_property>

21) y_axis_approx

Hedge expressions (e.g. "about", "around", "approximately", "more than", "less than", "over", "under", "just over") explicitly signal approximations. When these words play the role of a hedge, we label them as y_axis_approx.

The affiliated y values may or may not be inexact or rounded.

Some examples
<u>Roughly</u> 31% of students chose to study medicine
Roughly <y_axis_approx>
31 <y_axis_highest_value_val>
% <y_axis_inferred_label>
of
students <topic_related_property>

chose
to <x_axis_range_end>
study
medicine <x_axis_label_highest_value>

and fast food the lowest, at <u>just under</u> 25%
and
fast food <x_axis_label_least_value>
the
lowest <y_axis_least_value>
,
at
just under <y_axis_approx>
25 <y_axis_inferred_least_value_approx>
% <y_axis_inferred_label>

Between 2000 and 2004 minorities had <u>more than</u> 7% representation.
Between 2000 and 2004 <x_axis_label_4th_highest_value>
minorities <topic_related_property>
had
more than <y_axis_approx>
7 <y_axis_inferred_4th_highest_value_approx>
% <y_axis_inferred_label>
representation <y_axis>

22) y_axis_trend, y_axis_trend_up, y_axis_trend_down

When two or more bars are compared in a more vague manner, i.e. with adjectives and adverbs instead of providing their heights, one of the three labels is typically used: y_axis_trend, y_axis_trend_up, y_axis_trend_down.

Note that the label name is misleading: the label does not signify trends in terms of a regression line, but rather comparisons.

If a comparison has no direction, but just expresses a big, small, (in)significant difference without stating which bar is higher, the label used should be **y_axis_trend**.

In that year the gender pay gap is <u>similar</u> in Germany and the UK at just over 20%.
In
that
year
the
gender pay gap <y_axis>
is
similar <y_axis_trend>
in

Germany <x_axis_label_highest_value> ## group_highest_Scnd
and
the
UK <x_axis_label_Scnd_highest_value>
at
just over <y_axis_approx>
20 <y_axis_inferred_Scnd_highest_value_approx>
% <y_axis_inferred_label>

Mathematics follows closely with 56%
Mathematics <x_axis_label_3rd_highest_value>
follows <order_3rd>
closely <y_axis_trend>
with
56 <y_axis_3rd_highest_val>
% <y_axis_inferred_label>

This increases largely with those between age 45-54 spending the most time on social media
This
increases <slope_up>
largely <y_axis_trend>
with
those
between age 45-54 <x_axis_label_highest_value>
spending
the
most <y_axis_highest_value>
time <y_axis>
on
social media <topic_related_object>

The label y_axis_trend_(up|down) signifies base or comparative forms of adjectives/adverbs denoting size (in some way). Greatness is labeled with **y_axis_trend_up**, e.g. "higher", "bigger", "larger", "more popular". Smallness (e.g. "smaller", "lower", "less common") with **y_axis_trend_down**. Both of these labels can be modified by y_axis_trend.

Note that in case of phrases like "more than" and "less than", only the adverb is labeled. We leave "than" unlabeled; in the opposite case a new/different label would have to be used for "than" when the phrase is split. Compare "*Asia spends more than North America*" and "*Asia spends more money than North America*".

Asia was the bigger of the two with around 170 millions.
Asia <x_axis_label_3rd_highest_value>
was
the

bigger <y_axis_trend_up>
of
the
two
with
around <y_axis_approx>
170 <y_axis_inferred_3rd_highest_value_approx>
millions <y_magnitude>

The chart shows that there is more money spent on Higher Education in Asia than in North America
The
chart
shows
that
there
is
more <y_axis_trend_up>
money <y_axis>
spent
on
Higher Education <topic_related_object>
in
Asia <x_axis_label_highest_value>
than
North America <x_axis_label_Scnd_highest_value>

The amount in Europe was a little lower at £ 270,000,000
The
amount <y_axis>
in
Europe <x_axis_label_Scnd_highest_value>
was
a
little <y_axis_trend>
lower <y_axis_trend_down>
at
£ <y_axis_inferred_label>
270 <y_axis_inferred_Scnd_highest_value_approx>
,000,000 <y_magnitude>

Note that the same expression in different contexts carries a different role and thus label. For example, "more than":

- y_axis_trend_up: Asia has spent more than Europe.
- y_axis_approx: Africa has spent more than £ 60 millions on higher education.

23) slope_up, slope_down and slope_mix

The labels slope_[up|down] annotate lexical expressions of rising or falling slope in descriptions of ordinal/interval/ratio bars.

2010-2014 saw a 6.2% minority and this <u>increased</u> to 7% between 2015-2019.

2010-2014 <x_axis_label_least_value>
saw
a
6.2 <y_axis_inferred_least_value_approx>
% <y_axis_inferred_label>
minority <y_axis>
and
this
increased <slope_up>
to
7 <y_axis_inferred_5th_highest_value_approx>
% <y_axis_inferred_label>
between 2015-2019 <x_axis_label_5th_highest_value>

On Monday the stock closed at £50, then <u>rose</u> to £63 on Tuesday.
On
Monday <x_axis_label_3rd_highest_value>
the
stock <t=1_a=y>
closed <t=1_a=y>
at
£ <y_axis_inferred_label>
50 <y_axis_inferred_3rd_highest_value_approx>
,
then <order_Scnd>
rose <slope_up>
to
£ <y_axis_inferred_label>
63 <y_axis_inferred_Scnd_highest_value_approx>
on
Tuesday <x_axis_label_Scnd_highest_value>

It shows that there was a steady <u>decline</u> in injuries from 30 people being fatally injured in 2012 to 12 people being injured in 2016.
It
shows

that
there
was
a
steady <y_axis_trend>
decline <slope_down>
in
injuries <y_axis>
from
30 <y_axis_highest_value_val>
people being fatally injured <y_axis_inferred_label>
in
2012 <x_axis_label_highest_value>
to
12 <y_axis_least_value_val>
people being injured <y_axis_inferred_label>
in
2016 <x_axis_label_least_value>

References to a dynamic trend of the bar heights are labeled as slope_mix.

The salary of women in Najaf fluctuated between £ 16,000 to £ 26,000

The
salary of women in Najaf <y_axis>
fluctuated <slope_mix>
between
£ <y_axis_inferred_label>
16 <y_axis_least_value_val>
,000 <y_magnitude>
to
£ <y_axis_inferred_label>
26 <y_axis_inferred_highest_value_approx>
,000 <y_magnitude>

24) x_axis_labels_rest

When bars are split and referred to as one-versus-rest or some-versus-rest, the label x_axis_lables_rest is used for the reference to the rest bars. Note that while this is a case of grouping, it is not necessarily motivated by conceptual similarity or height.

Female representation in the Insurance sector in Benoni is higher than in Law Firms, Tech and Financial Groups.

Female representation <y_axis>
in

the
Insurance sector <x_axis_label_highest_value>
in
Benoni <t=1_a=b>
is
higher <y_axis_trend_up>
than
in
Law firms , Tech and Financial Groups <x_axis_labels_rest>

25) x_axis_trend

In charts with an ordinal/interval/ratio variable on the x-axis, descriptions of rising (or in rare cases, falling; x_axis_trend_down) of this variable are annotated with x_axis_trend. The rising/falling refers to the bars' value on the x-axis, not on the y-axis.

[week days on x-axis] *Overall we can see a gradual decrease in closing stock across the week*
Overall
we
can
see
a
gradual <y_axis_trend>
decrease <slope_down>
in
closing stock <y_axis>
across the week <x_axis_trend>

[age groups, younger to older on x-axis] *The older the group the less time they spend daily on social media*
The older the group <x_axis_trend>
he
less <slope_down>
time <y_axis>
they
spend
daily <t=1_a=y>
on
social media <t=1_a=y>

Note that, these expressions are realized as steps ("year on year") or frequency ("each year").

[years on x-axis] *The fatality rate has fallen each year.*
The

fatality rate <y_axis>
has
fallen <slope_down>
each year <x_axis_trend>

Describing the bars in terms of right-to-left orientation on the x-axis should be labeled as x_axis_trend_down.

[age groups in ascending order on x-axis] *The chart shows that the younger you are the more time you spend on social media.*

The
chart
shows
that
the
younger you are <x_axis_trend_down>
the
more <slope_up>
time <y_axis>
spent
on
social media <t=1_a=y>

26) order_*

The label order_* applies to words/phrases that mark the order of bars, in terms of how the bars appear on the chart or in the description.

The order is explicit in the annotated word/token ("second", "thirdly") or inferred from context ("the highest is Spain, followed by Germany").

The label set includes order_Sncd, order_3rd, order_4th, … order_last.

Insurance is the highest at 64%, tech next at 61%, then financial groups at 50%, then law at 35%

Insurance <x_axis_label_highest_value>
is
the
highest <y_axis_highest_value>
at
64 <y_axis_inferred_highest_value_approx>
% <y_axis_inferred_label>
,
tech <x_axis_label_Scnd_highest_value>

next <order_Scnd>
at
61 <y_axis_inferred_Scnd_highest_value_approx>
% <y_axis_inferred_label>
,
then <order_3rd>
financial groups <x_axis_label_3rd_highest_value>
at
50 <y_axis_3rd_highest_val>
% <y_axis_inferred_label>
,
then <order_last>
law <x_axis_label_least_value>
at
35 <y_axis_least_value_val>
% <y_axis_inferred_label>

The third is Asia which drops significantly compared to Europe

The
third <order_3rd>
is
Asia <x_axis_label_3rd_highest_value>
which
drops <y_axis_trend_down>
significantly <y_axis_trend>
compared <y_x_comparison>
to
Europe <x_axis_label_Scnd_highest_value>

Computer science 60% and finally engineering at 70%

Computer science <x_axis_label_Scnd_highest_value>
60 <y_axis_Scnd_highest_val>
% <y_axis_inferred_label>
and
finally <order_last>
engineering <x_axis_label_highest_value>
at
70 <y_axis_highest_value_val>
% <y_axis_inferred_label>

27) y_x_comparison

Lexical expressions marking comparison are annotated with y_x_comparison. This applies to comparisons where the bar comparison is made by providing their heights, but also in cases where the heights are not mentioned explicitly. It can refer to a comparison of two or more bars.

The label annotates verbs ("compared"), nouns ("comparison"), conjunctions ("whilst") or any other part of speech that carries the same meaning.

Insurance has the highest representation at 65% whereas law firms has the lowest.

Insurance <x_axis_label_highest_value>
has
the
highest <y_axis_highest_value>
representation <y_axis>
at
65 <y_axis_highest_value_val>
% <y_axis_inferred_label>
whereas <y_x_comparison>
law firms <x_axis_label_least_value>
has
the
lowest <y_axis_least_value>

32% spend time with family, while 38% prefer to read a book.

32 <y_axis_Scnd_highest_val>
% <y_axis_inferred_label>
spend time with family <x_axis_label_Scnd_highest_value>
,
while <y_x_comparison>
38 <y_axis_highest_value_val>
% <y_axis_inferred_label>
prefer
to
read a book <x_axis_label_highest_value>

This has reduced by over 50% compared to the year 2012.

This
has
reduced <slope_down>
by

over <y_axis_approx>
50 <y_axis_inferred_value_mul_v1=highest_v2=least>
% <y_axis_inferred_label>
compared <y_x_comparison>
to
the
year <x_axis>
2012 <x_axis_label_highest_value>

28) t=*_a=*

The above listed labels sometimes do not suffice to annotate all chart- or topic-specific vocabulary. This vocabulary and its derivations can appear in the title or not, and modify the variable on y-, x- or both axes. We introduce the label t=[0|1]_a=[x|y|b], which should cover for such cases.

In the label

- "t" stands for "title" and is a binary marker of whether or not a word/phrase appears in the title (derivations included; synonyms not)
- "a" stands for axis; the value is either x, y or b for both axes.

Chart: Women representation in different sectors

- t=1_a=y : "tech has the lowest with 20% of women"
- t=0_a=y: "more women than men are found in law firms"

Chart: Median salary of software engineers per year wrt. their degree

- t=1_a=b: "63 k is median salary for software engineers who have PhD"
- t=1_a=y: "the median salary is at 55 k"

Chart: Gender pay gap

- t=0_a=x: "this chart show gender pay gap across 3 European countries"

Chart: Average time spent on social media daily in Maputo by age group
Sentence: In Maputo, persons of 45-54 spend around 150 minutes daily on social media.

- t=1_a=b: Maputo
- t=1_a=x: persons (fix label to t=0)
- t=1_a=y: daily; social media

29) other_operation

Currently, the label set supports only pairwise comparisons, so all other cases are annotated with other_operation

Spain is exactly halfway between the two countries.

Spain <x_axis_label_Scnd_highest_value>
is
exactly
halfway **<other_operation>**
between
the
two
countries <x_axis>

STEM being the lowest at 15% and medicine being the highest at 31% with the other course of studies being in between them.

STEM <x_axis_label_least_value>
being
the
lowest <y_axis_least_value>
at
15 <y_axis_least_value_val>
% <y_axis_inferred_label>
and
medicine <x_axis_label_highest_value>
being
the
highest <y_axis_highest_value>
at
31 <y_axis_highest_value_val>
% <y_axis_inferred_label>
with
the
other
course of studies <x_axis_labels_rest>
being
in
between **<other_operation>**
them

30) separator

In the preprocessing pipeline we wanted to keep the original structure of the text intact as much as possible. Newline characters often delimit prepositions, In the summaries, they are annotated with the separator label.

This
chart
shows
the
causes of obesity in Kiribati <topic>
.
\\n <separator>
The
biggest <y_axis_highest_value>
cause <x_axis>
,
shown
to
be
responsible
for
40 <y_axis_inferred_highest_value_approx>
% <y_axis_inferred_label>
of
cases <t=1_a=y>
,
is
genes <x_axis_label_highest_value>

31) interpretation

Some descriptions include a reading of the data that is not grounded in the chart or includes information that goes beyond what is accessible in the chart. In such cases, the <interpretation> label is used.
Some examples:

- However it doesn't show how this relates to inflation
- If this trend is to continue there were be a decrease in the following years
- This could be due to the economic situation in each respective continent, and also due to societal norms and values. In Asia , there is a lot of emphasis placed on the value of having a good education, which explains why more is spent on higher education in Asia than any other continent.

Note that inside the span of the interpretation label, we do not label entities as we normally would, e.g. "Asia" in the last example is left unlabeled.