ICAME Journal • 50(1) • 2026 • 69–90
DOI: 10.2478/icame-2026-0005

ICAME Journal

An analysis of bar chart usage in corpus data visualization

Research Article

Lukas Sönning

University of Bamberg

Received 31 July 2025; accepted 15 December 2025

Abstract: A recent survey of graph usage in corpus-based research has shown that the bar chart is the most widely used graph type for corpus data presentation. Motivated by this finding, the present paper offers a systematic review of bar chart usage in corpus-based research articles. It covers all papers (n = 1,183) published in five corpus-linguistic journals up to and including the year 2024 (*International Journal of Corpus Linguistics*, *Corpus Linguistics and Linguistic Theory*, *Corpora*, *Research in Corpus Linguistics*, and *International Journal of Learner Corpus Research*). The aim of this survey is to arrive at a better understanding of the kinds of visualization tasks imposed on bar charts. We observe that they most commonly show percentages or absolute/normalized frequencies, and that they are often used for relatively complex visualization tasks involving multiple variables and subgroups. The survey is carried out against the backdrop of known limitations of this graph type and design recommendations found in data visualization guidebooks. Our critical examination of diagrams pays attention to issues that compromise the ability of the viewer to accurately perceive patterns in the data, and minor issues that affect the efficiency of a display. These observations are then distilled into a set of concrete recommendations, which are grounded in current usage and the advice given in the data visualization literature.

Keywords: *data visualisation • corpus linguistics • bar charts • graph design • graphical criticism • best practices*

© Author

1 Introduction

As a genuinely quantitative branch of the language sciences, corpus linguistics has always shown a particular reliance on statistical methodology. This includes the use of data visualization tools, which have become essential aids for data analysis and presentation. As corpus linguistics is growing into a mature discipline, it seems appropriate to cultivate a methodological discourse on corpus data visualization and reflect on the kinds of data layouts and visualization tasks that are typical of our field. An edited volume by Sönning & Schützler (2023) has taken a step in this direction. Their analysis of corpus-based research articles showed that classical visualization forms continue to account for the vast majority of cases. In particular, the bar chart featured as the most widely used type of display.

This finding stimulated the present paper, which examines in more detail the way in which bar charts are used for corpus data presentation. To this end, I conducted a systematic review of their use in corpus-linguistic journals. This analysis is carried out against the backdrop of known limitations of this graph type, which provide important guardrails for their use. The goals of this literature review are two-fold. The first is to arrive at a better understanding of the kinds of visualization tasks that are tackled using bar charts. This includes an overview of the types of corpus-based quantities that are typically visualized and the complexity of displays, which depends on the number of variables and the number of (sub)groups that appear in the chart. The second goal is to examine the construction and layout of figures in light of relevant constraints and guidelines found in the literature on data visualization. This

* *Corresponding author: Lukas Sönning, E-mail: lukas.soenning@uni-bamberg.de*

Open Access. © 2026 Author.
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License.

allows us to identify aspects where there is room for improvement and to point to contexts where an alternative graph type may be given preference. A distinctive feature of the graphical criticism presented in the current paper is that it is grounded in usage. This means that advantages and disadvantages of bar charts relative to alternative graph types are only considered in so far as they are relevant for the practicing corpus linguist. Critical treatments from a more general perspective may be found elsewhere in the literature (see, e.g., Cleveland 1994; Tufte 2001; Robbins 2005; Sönning 2016; Wilke 2019; Sönning 2023).

The present paper is structured as follows. As a point of departure, Section 2 first establishes a terminological basis and then summarizes limitations of bar charts as well as design recommendations given in the literature. In Section 3, I then describe how I conducted the review of bar chart usage for corpus data presentation. Section 4 summarizes general features of corpus-based bar charts, focusing on the types of quantities that are visualized, the choice of bar chart type, and the complexity of graphs. Section 5 then presents the results of my critical examination of bar charts in the report pool, dealing in turn with major and minor design issues. Section 6 distils the observations into a set of recommendations that are specifically geared toward corpus linguists. Finally, Section 7 closes with a summary and discussion.

2 Bar charts for data visualization: Background

The current section sets the scene for our review, starting with an outline of relevant terminology for the discussion of structural elements of bar charts. Section 2.2 then recapitulates key limitations of this graph type and Section 2.3 summarizes some general design recommendations given in the literature.

2.1 Bar charts: Terminology

Figure 1 illustrates three basic forms that are commonly distinguished: simple, grouped, and stacked bar charts. Simple bar charts show quantitative values for a single categorical variable. The bars are aligned at zero and their length encodes the quantity of interest. A grouped bar chart has quantitative values for two or more cross-classified factors. Bars are arranged in clusters and the bars within a cluster are contrasted using fill colours. Again, all bars are aligned at zero. In stacked bar charts (sometimes called divided bar charts), bars are stacked into columns. There are two versions: one where columns have the same height (as in Figure 1), with bars representing percentages or proportions, and one where bars show absolute numbers, with the height of columns being proportional to their total.

While the bars in Figure 1 are all in an upright position, they can also be arranged horizontally; in fact, historically this seems to be the preferred orientation (see, e.g., Playfair 1787: Plate 24; Karsten 1925: 99–123). As we will see further below, a horizontal arrangement in fact offers advantages in certain situations.

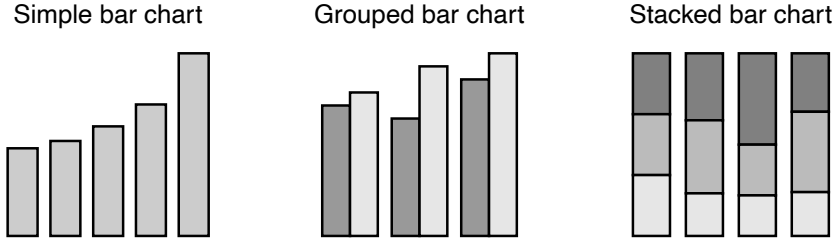


Figure 1. Bar charts: three basic variants.

All images in this paper have been published under the Creative Commons Attribution 4.0 license (CC BY 4.0, http://creativecommons.org/licenses/by/4.0) in the accompanying OSF project (https://osf.io/a5p4s/). All figures were drawn using the R packages ‘lattice’ (Sarkar 2008) and ‘ggplot2’ (Wickham 2016).

Figure 2 sets out the terminology used in the present paper to refer to various structural elements of bar charts, with relevant labels appearing in italics. In general, we can distinguish between four types of variables that may surface in such displays. It is useful to organize terminological notes along these lines:

– Y-*variable*. This is the quantity of interest, which in corpus-based work is often a normalized frequency, a frequency count, or a percentage (see Section 4.1). The *y*-variable usually occurs on the vertical scale, but it may also appear on the horizontal scale if the axes are swapped. The *axis title* states what type of quantity is shown, and the *tick mark labels* denote the scale units.
– X-*variable*. This variable refers to the categories that are compared, and it is encoded using position along the *x*-axis. In the illustrative charts that appear in Figure 2, Corpus is always the *x*-variable. The levels of the categorical *x*-variable, or x-*levels*, are marked using *data labels*. In a simple bar chart, each *x*-level appears as a bar, in grouped bar charts as a *cluster* of bars, and in stacked bar charts as a *column* of bars.
– *Fill variable*. Grouped and stacked bar charts include a *fill variable* to distinguish between a further set of categories or groups. The levels of the fill variable, which are shown using different fill colours, will be called *fill levels*. Information about the mapping between fill colours and category labels can either be shown with a *key* (illustrative grouped bar chart) or using *direct data labels* (illustrative stacked bar chart).
– *Faceting variable*. Finally, bar charts may generally include what is often referred to as a *faceting variable*. It is used to form subsets of the data, which then appear in different panels or *facets*. The different subsets correspond to the *facet levels*, and they are typically designated directly using a *facet label*.

As the distinction between these four types of variables is critical for the ensuing discussion, Table 1 recapitulates their key features.

Table 1. Summary of the different variables that (may) appear in bar charts.

Type of variable	Description	Levels
Y-variable	The type of quantity represented by the length of the individual bars	---
X-variable	The variable representing the categories that are compared using different bars (or clusters/columns of bars)	*X*-levels: the different bars (or clusters/columns of bars) that appear in the chart
Fill variable	In grouped and stacked bar charts, the variable representing the categories that are distinguished using fill colours	Fill levels: the different fill colours in each cluster or column of bars
Faceting variable	The variable that divides the data into subsets, which then appear in different panels or facets	Facet levels: the different facets in the figure

2.2 Limitations of bar charts

As with any graph type, a number of limitations must be borne in mind when considering the use of bar charts for data visualization. In the present section, potential weaknesses will be discussed and illustrated using figures that mimic charts encountered in our survey. Of course, the weight of these limitations depends on the type of data shown and the intended visual message.

Start-at-zero constraint. In bar charts, the primary visual cue to the values shown in the display is the length of the bars. Since viewers will compare these lengths, a faithful representation of the data requires the bars to start at zero – otherwise comparisons are misleading (Huff 1954; Schmid 1983: 58; Tufte 2001; Schwabish 2021: 69). This constraint can undermine the resolution of a chart if it has long bars of similar length (Wilke 2019: 54). This is illustrated in Figure 3, where the *y*-variable is a measure of lexical density, and two corpora are compared across 15 genres. The range of variation in bar heights amounts to 15 points on the scale, which is less than 20% of the vertical space taken up by the bars.

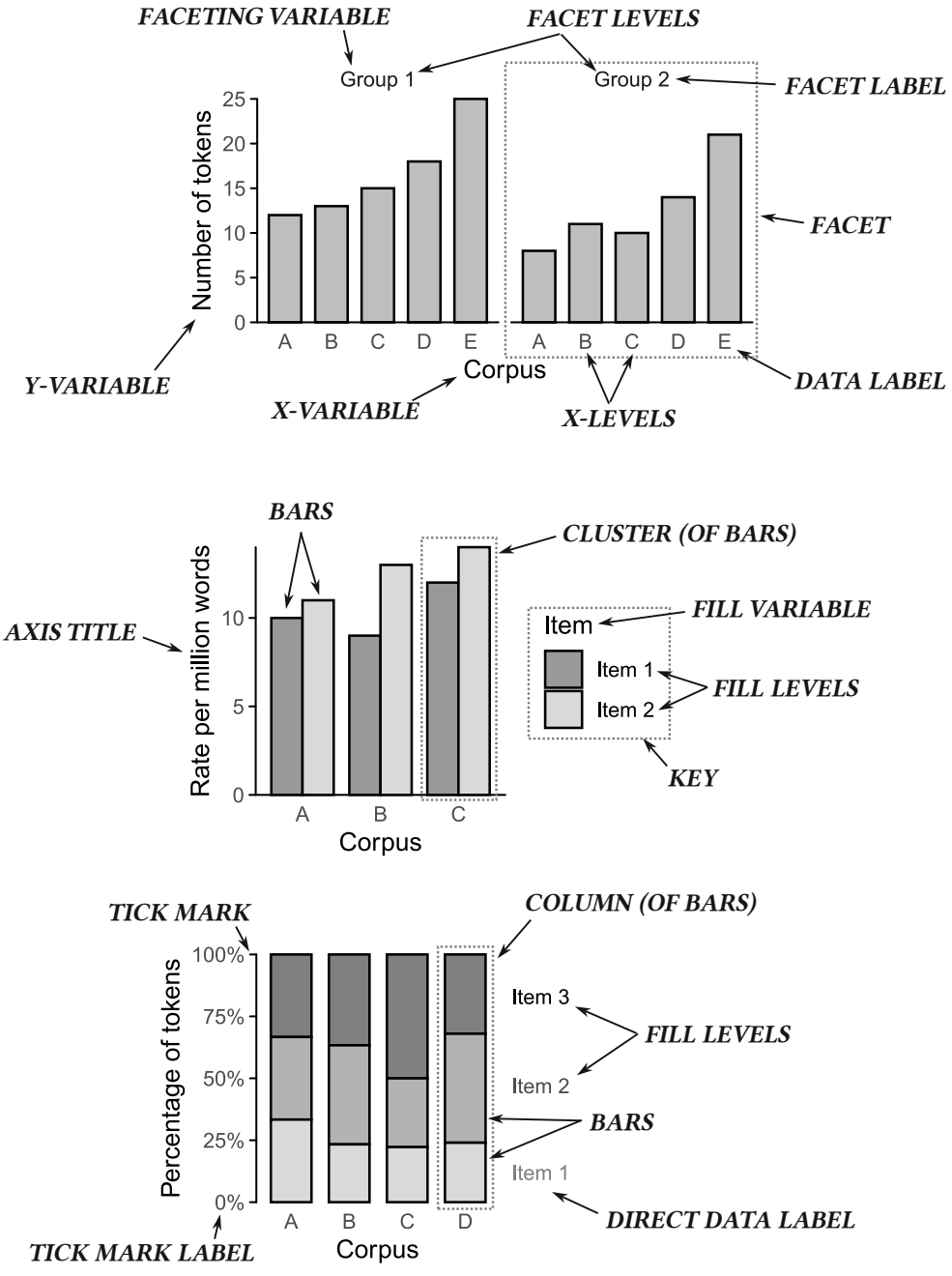


Figure 2. Structural elements of bar charts. Terminology and style of presentation borrow heavily from Cleveland (1994: 21–22).

If the graph is meant to show that the two corpora differ systematically in lexical density, the display arguably fails to communicate this clearly.

Constraints on y*-variables*. The start-at-zero constraint also limits the type of quantity (or *y*-variable) that may be sensibly visualized using a bar chart. This is because zero must be an existing and meaningful value. It is therefore inappropriate to use bar charts to show log-transformed scores, and care is needed when graphing ratios or fractions, for which the relevant point of reference is 1 rather than 0. This is illustrated in Figure 4, which visually compares two ratios across ten features. The bars start at zero,

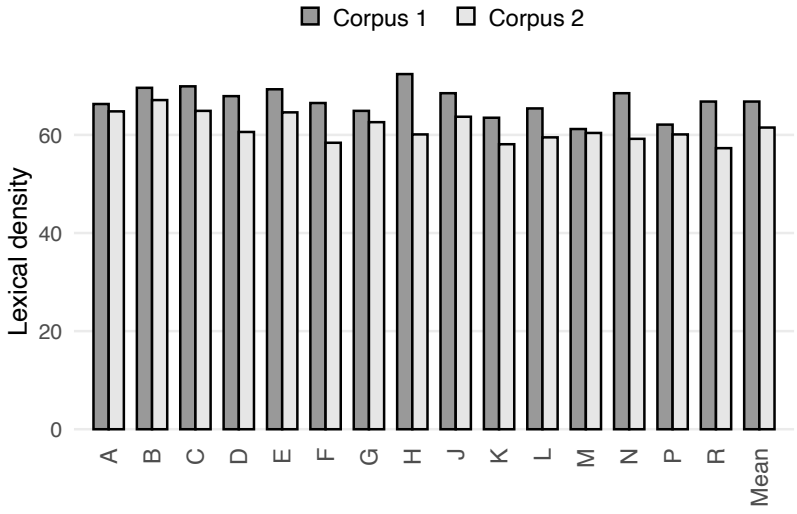


Figure 3. This grouped bar chart offers poor resolution: since all bars necessarily start at zero, there is little room for the variation in the data to unfold visually (IJCL 15-1_1).

which is inappropriate for ratios and fractions. Even a redesign using (diverging) bars anchored at 1 would be problematic, however, since fractions (i.e. ratios below 1) are crammed into the [0, 1] interval, while ratios above 1 are unbounded. As a result, even though a ratio of 2 to 1 (2.0) represents the same relative difference as a ratio of 1 to 2 (0.5), its associated bar would be twice as long in such a chart.

Visual weight. Another potential limitation of bar charts is their use of multiple visual cues to mark a specific point value (see Tufte 2001: 96). While this is rarely a concern in simple bar charts, grouped and stacked variants can become busy when the number of *x*-levels and/or fill levels in the display increases (Wallgren et al. 1996: 26–27; Schmid 1983: 61–62). Visual clutter can then make it difficult to form relevant groupings and draw certain comparisons. In the crowded bar chart in Figure 5, for instance, it takes a slow sequential search to identify (i) the category with the smallest percentage of modifier nouns (light grey), (ii) the one with the largest percentage of head nouns (dark grey), and (iii) those categories that have a larger percentage of head nouns (dark bar taller than light bar).

Non-aligned comparisons. The final limitation only applies to stacked bar charts. In general, length comparisons are much easier to make when bars are aligned. In stacked bar charts, this only applies to the fill level at the lower end of the column (and if columns have the same height, also to the fill level at the upper end of the column). The intermediate bars, however, are shifted relative to each other and comparisons among them are much harder to make (Wallgren et al. 1996: 27; Camões 2016: 217). In Figure 6, the white bars and the dark grey bars share a baseline. We can therefore quickly identify the adverb with the largest and smallest proportion of position “I” (initial) and “M3” (medial, after the verb phrase). This is much harder to determine for categories M1 and M2.

2.3 Bar chart design: General recommendations in the literature

The data visualization literature provides a number of guidelines for the design of bar charts. Many of these derive from the limitations summarized in the previous section.

Zero baseline. First and foremost is the imperative that bars should start at zero (or rather at the relevant baseline; see Figure 4). If the axis showing the *y*-variable does not include the relevant baseline, bars provide an invalid representation of the quantities they show (Huff 1954; Schmid 1983: 58; Tufte 2001; Schwabish 2021: 69).

Natural origin. Since bar charts encourage length comparisons, they should only be used for quantities with a natural baseline, or origin (Cleveland 1984: 272; Robbins 2005: 250–251; Few 2012: 196). If the scale on which data are shown does not have such a starting point, the length of segments

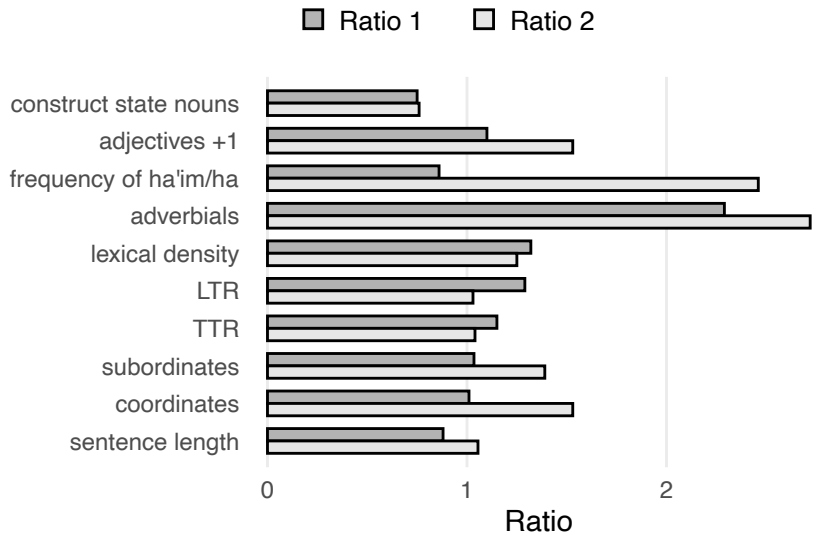


Figure 4. This bar chart provides an inadequate representation of ratios, where the relevant baseline is 1 (rather than zero) (CLLT 6-2_3)

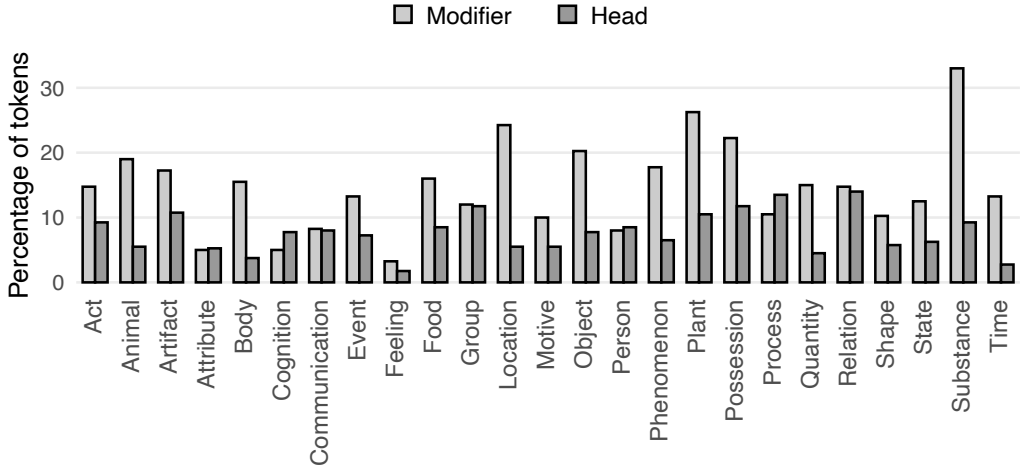


Figure 5. This chart is quite cluttered, which makes it hard to draw comparisons (CLLT 6-1_3).

is an irrelevant visual feature. A case in point are transformed scales such as the log (or square-root) scale.

Order. A fundamental design consideration with categorical *x*-variables concerns the arrangement of categories along the horizontal (or vertical) scale. If the categories have no logical or preferred order, it usually makes sense to sort them according to value or size (Karsten 1925: 107; Schmid 1983: 42; Wainer 1997: 35–38; Camões 2016: 347–351; Wilke 2019: 48).

Horizontal format. The prototypical bar chart is in an upright position, with the bars resting on the horizontal axis. This arrangement leaves limited space for data labels. Longer labels are then often abbreviated or rotated, which makes them difficult to read. The literature on bar chart design encourages researchers to consider switching to a horizontal format in the case of long data labels or when many

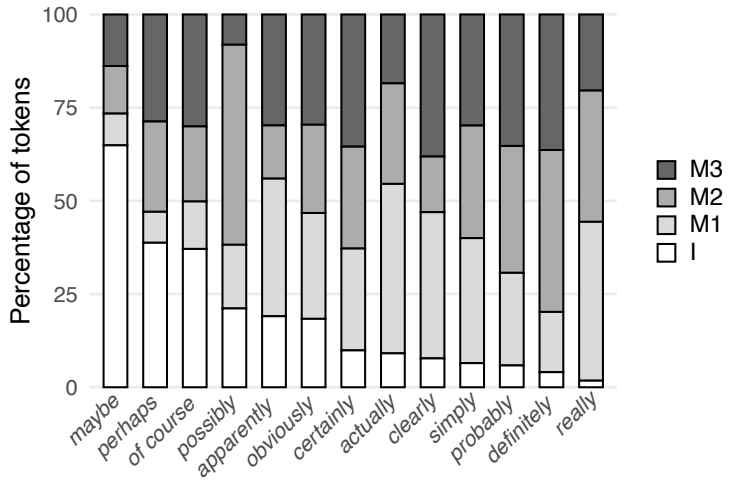


Figure 6. In this stacked bar chart, the intermediate categories M1 and M2 are difficult to compare (IJCL 25-2_2).

categories are compared (Schmid 1983: 50–52; Wallgren et al. 1996: 25; Kosslyn 2006: 54; Wilke 2019: 47; Schwabish 2021: 79).

Choice between grouped and stacked bar charts. The choice between grouped and stacked variants depends on the intended visual message as well as features of the data. Since all bars in a grouped arrangement are aligned at zero, this variant is typically preferable if bars are meant to be compared to one another (Wallgren et al. 1996: 27; Kosslyn 2006: 56; Wilke 2019: 97). Further, Camões (2016: 298–299) argues that the clusters in a grouped bar chart are suitable to profile the different *x*-levels with respect to the distribution of the fill variable. If the total over all fill levels is the quantity of primary concern, stacked variants (with columns of different heights) are preferable (Wallgren et al. 1996: 27; Wilke 2019: 52). Further, stacked bar charts, especially their equal-height version, clearly show part-whole relations in the data (Kosslyn 2006: 44). They are also generally a good choice for ordinal fill variables, since the cumulative summation of category counts then constitutes a meaningful quantity (Wilke 2019: 97). Finally, Wilke (2019: 97) argues that stacked forms work better if the *x*-variable has many levels, and grouped bar charts if the fill variable has many levels.

Complexity of grouped and stacked bar charts. With an increasing number of fill levels, grouped bar charts can become difficult to read, and more mental effort is required to parse the display. Some authors therefore suggest a maximum of two or three fill levels (Wallgren et al. 1996: 26). Similar concerns apply to stacked bar charts, which become difficult to digest when there are too many fill levels (Schmid 1983: 61–62; Wallgren et al. 1996: 27). As the number of fill levels in grouped bar charts increases, it is easier to make visual comparisons between the fill levels, i.e. within a cluster of bars; comparison across *x*-levels, on the other hand, will be obscured (Wilke 2019: 50).

3 Graph types in corpus-based work: A survey

To gain a better understanding of how bar charts are used for corpus data visualization, I conducted a survey of published research articles.

3.1 Report pool and general graph usage

To this end, I selected five corpus-linguistic journals and downloaded all articles that had appeared in these publication channels up to and including the year 2024. The five journals will hereafter be referred to using the following abbreviations:

– *International Journal of Corpus Linguistics* (IJCL)
– *Corpus Linguistics and Linguistic Theory* (CLLT)
– *Corpora*
– *Research in Corpus Linguistics*, (RiCL)
– *International Journal of Learner Corpus Research* (IJLCR)

Before we take a closer look at the report pool, it should be noted that this survey covers one particular usage context of statistical graphics, namely presentation graphs in a printed medium. Clearly, this is just one function of statistical graphics (see Fienberg 1979; Schmid 1983). However, presentation graphs play an important role in scientific communication, as they are used to communicate the key insights of a study. Further, they allow us to map out the nature and range of visualization tasks corpus linguists often confront. Finally, presentation graphs should be held to the highest standards, which warrants a critical engagement with their usage in our discipline.

In total, the report pool includes 1,183 research articles. Since the journals were launched in different years, the distribution of papers across the five outlets is imbalanced. This is shown graphically in Figure 7: IJCL, which started in 1996, contributes by far the largest number of articles (n = 476, 40%). The equally-apportioned launch dates of ensuing journals yield a near-linear increase in article counts across the 28-year period covered by the survey.

For data coding, I used the qualitative data analysis software MAXQDA (VERBI Software 2021) and started by identifying all occurrences in the report pool of statistical tables (4,552 in total) and statistical graphs (n = 3,060). Not all papers present or interpret data. If we take the occurrence of at least one table with quantitative information (e.g., frequencies, percentages, or other statistics) *or* one statistical graph as an indicator, 86% of the articles (n = 1,018) classify as empirical.

The prevalence of graphs and tables permits a first glance at the usage rate of these different means of data presentation: 81% of all articles used at least one data table, and 54% of all articles showed at least one graph. There is a diachronic increase in the proportion of articles presenting data visualizations, from roughly 20% in the late 1990s to around 65% in the early 2020s.

The next step was to categorize figures in terms of the graph type they represent, which turned out to be an iterative process requiring several revisions of the labels assigned to images. Before we take a look at the usage rate of graph types, two comments are in order. First, the number of figures varies quite markedly across papers, ranging from 0 to 37. For this reason, I decided to pay attention to the dispersion of graph types, i.e. their relative range. This means that I will not refer to the number of figures (out of the 3,060) that represented a particular graph type, but rather the number of articles that used a graph type at least once. Further, the relevant baseline for judging the pervasiveness of graph types seems to be the number of empirical articles in the report pool, whose number we estimate at 1,018.

Figure 8 shows usage rates for the major graph types in our report pool. Bar charts clearly emerge as the most popular vehicle for visualizing corpus data. About 33% of the empirical papers in our survey use this display type. This rate is relatively stable across the time period covered by the survey, with the late 1990s averaging at around 40% and the early 2020s at about 30%. Together with line plots and scatterplots, bar charts account for the vast majority of visuals found in the database.

This initial overview underscores the prevalence of bar charts in corpus data visualization. We now take a closer look at how this graph type is used by corpus linguists.

3.2 Focus on bar charts

In total, the report pool includes n = 1,061 unique instances of bar charts, which are distributed across 335 articles. I created an Excel spreadsheet and coded each figure for a number of features.

– Metadata: article id, journal/issue/number, year of publication, figure number in paper
– Bar chart variant: simple, grouped, stacked
– *Y*-variable: percentage, normalized frequency, absolute frequency, other
– Dimensions: physical size (height and width) of the figure excluding figure caption

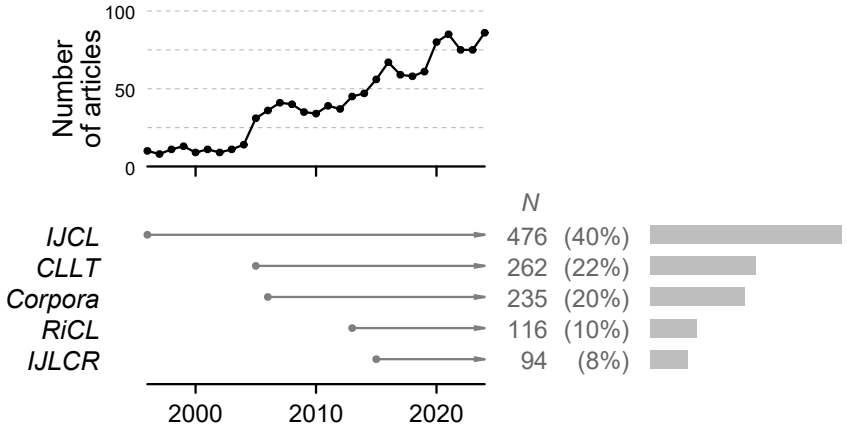


Figure 7. Report pool: number of articles (*n* = 1,183) across time and journals.

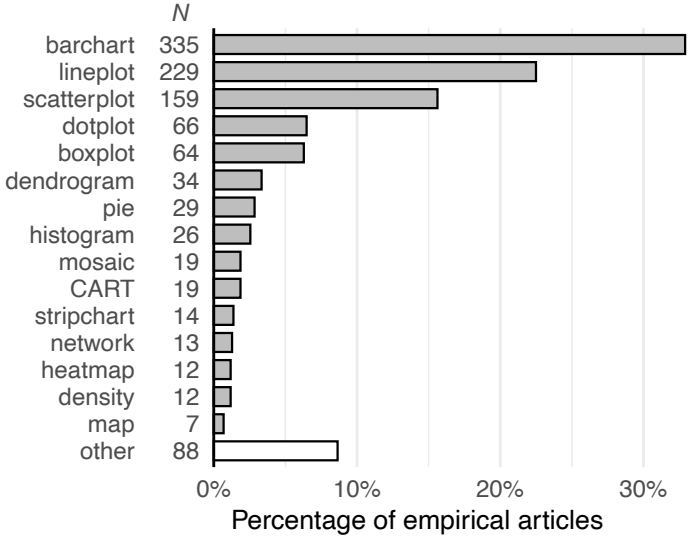


Figure 8. Distribution of graph types in the sample.

- Content: number of values shown, number of variables (fill, faceting, and *x*-variables), number of levels for each variable
- Form: use of fill colours/greyscales, whether the chart includes an (informative) axis title, format (horizontal vs. vertical), whether data labels are rotated, number of tick mark labels, use of faceting, visual cues to facets, whether the chart is drawn in 3D
- Scaling: whether the axis showing the quantity of interest is scaled appropriately, whether the *y*-variable is log-scaled
- *X*-variable: whether *x*-levels are ordered, whether the *x*-variable is time, or the speaker ID

For more details on the coding process and the categories used, please refer to the TROLLing post (Sönning 2025).

4 Bar charts in corpus-based work: General features

Let us start by considering some general features of bar charts in the published corpus-linguistic literature. This allows us to appreciate the variety of visualization tasks imposed on bar charts, which is an essential starting point for a critical discussion of their use.

4.1 Quantity of interest and bar chart type

For each bar chart in the survey, I determined the type of *y*-variable that is visualized, which led to three major classes – percentages/proportions, absolute frequencies, and normalized frequencies – and a residual category with less common quantities of interest (e.g., ratios, measures of collocational strength, or variable importance scores in random-forest models). Percentages turned out to be the most common quantity (43% of the charts in our survey), followed by absolute frequencies (24%) and normalized frequencies (20%). The category "other" includes 13% of the charts. While this distribution is perhaps unsurprising, it provides important guideposts for the ensuing discussions.

The three major types of bar charts appear at similar rates, with grouped (38%) and simple forms (36%) being somewhat more prevalent than stacked variants (25%). Figure 9 shows how the choice of chart type varies with the type of quantity shown:[1]

– Percentages are most commonly visualized using stacked bar charts (46%), followed by grouped (30%) and simple forms (23%).
– Absolute frequencies most often appear as simple bar charts (49%), followed by grouped bar charts (39%); stacked versions are rather untypical (13%).
– Normalized frequencies are typically presented with grouped bar charts (57%), less commonly with simple forms (34%) and rarely with stacked bar charts (8%).

4.2 Complexity

It is also of interest to examine the complexity of bar charts, which is reflected in the number of variables they show, and – more critically – their number of levels. We start with an overview of the number of variables that appear in the display (necessarily disregarding the *y*-variable, which is always present).

The rightmost column in Table 2 gives the overall distribution in the sample. We note that most bar charts involve two factors (58%), followed by monofactorial displays (32%). Around 10% of the charts in our sample are quite complex in the sense that they depict three or more variables. If we break down this distribution by bar chart type, we observe the following patterns:

– Simple bar charts are typically monofactorial (83%), but faceting is sometimes used to incorporate an additional variable into the display.
– Grouped bar charts predominantly show two factors (89%), an *x*-variable and a fill variable. A relatively small share of 8% includes an additional faceting variable.
– While stacked bar charts also usually have two factors (74%), an *x*-variable and a fill variable, a considerable proportion (25%) involve additional variables using faceting.

The shaded cells in Table 2 identify the most commonly used arrangement(s) for each graph type. These will now be examined in more detail.

We will start with simple bar charts that include a single *x*-variable. The visual complexity of this subset mainly depends on the number of bars in the display. Figure 10 shows the distribution of this count in the 317 charts occupying this cell. Note that since the tallies are skewed, the *x*-axis is log-scaled. Bar counts range from 2 to 177, with a median number of 8 categories. What seems remarkable about the distribution in Figure 10 is the relatively high share of simple bar charts with a large number of categories: about a quarter of them depict 20 or more categories.

1 This mosaic chart was drawn with the R package 'ggmosaic' (Jeppson & Hofmann 2023).

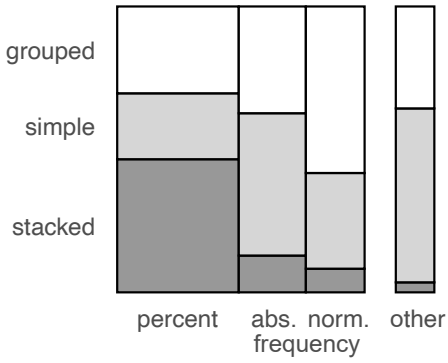


Figure 9. Mosaic chart showing the distribution of bar chart variants by quantity of interest.

Table 2. Number of variables in bar charts, by type of bar chart.

	Simple		Grouped		Stacked		Total	
Number of variables	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
4	--	--	--	--	2	1%	2	0%
3	7	2%	31	8%	63	24%	101	10%
2	60	16%	360	89%	196	74%	616	58%
1	317	83%	14	3%	3	1%	334	32%
Total	384	100%	405	100%	264	100%	1,083	100%

To map the visual complexity of grouped and stacked bar charts, two features are of primary interest: the number of *x*-levels and the number of fill levels. We will examine the distribution of these tallies for all grouped and stacked charts in our survey.

Table 3 starts by considering the marginal tallies. The overall distribution, which sums across both chart types, appears in the rightmost column. The number of fill levels is usually two (52%), three (25%), or four (10%), but even a higher number is not infrequent (13%). If we consider this distribution separately for grouped and stacked displays, we note that stacked bar charts show a stronger tendency toward a large number of fill levels, with 20% of the charts contrasting 5 or more categories.

Table 3 also shows the distribution of the number of *x*-levels in our sample of grouped and stacked bar charts – higher counts are represented by two bins (“5–9” and “10+”). The distribution of charts across these five classes is remarkably balanced, with the marginal shares ranging between 15% and 25%. Grouped bar charts show a somewhat stronger tendency toward a greater number of *x*-levels: 49% of the charts feature five or more clusters of bars, while only 36% of stacked bar charts show 5 or more columns.

Since the visual complexity of a multifactorial bar chart depends both on the number of fill levels and the number of *x*-levels, we also need to examine their joint distribution. To do so, I cross-tabulated these features for each chart type. This allows us to see how many charts of a given type include, say, a fill variable with two levels and an *x*-variable with 10 levels (as is the case in Figure 4).

Figure 11 represents this two-way distribution graphically. The horizontal axis marks the number of categories of the *x*-variable, and the vertical axis denotes the number of fill levels. The area of the circles is proportional to the number of displays with a specific combination. The frequencies in the attested cells range from 1 (dot in the upper panel right below the facet label “Grouped bar charts”) to 46 (grouped bar chart with two fill levels and three *x*-levels). Three observed combinations are outside of the range covered by the graph, and these are flagged at the top or right margin of the panels. For instance, the mark “(200)” to the right of the lower panel denotes a bar chart with two fill levels and two hundred *x*-levels.

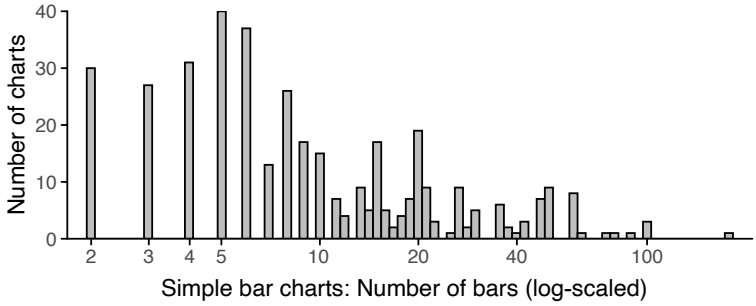


Figure 10. Histogram showing the distribution of the number of categories that appear in the 317 monofactorial simple bar charts in our sample.

Table 3. Number of fill levels and *x*-levels in bar charts, by type of bar chart.

	Grouped		Stacked		Total	
Number of levels	**N**	**%**	**N**	**%**	**N**	**%**
Fill variable						
5+	32	8%	52	20%	84	13%
4	35	9%	33	12%	68	10%
3	110	27%	59	22%	169	25%
2	228	56%	120	45%	348	52%
(Total)	(405)		(264)		(669)	
X-variable						
10+	84	21%	40	15%	124	19%
5–9	112	28%	56	21%	168	26%
4	51	13%	47	18%	98	15%
3	73	18%	54	20%	127	19%
2	71	18%	64	24%	135	21%
(Total)	(391)		(261)		(652)	

In Figure 11, it can be seen that most bar charts are located in the lower left corner of the grid, with two to five *x*-levels and two or three fill levels. A key insight that emerges from Figure 11, however, is that an appreciable share of stacked and grouped bar charts is visually quite complex.[2] The overall spread of charts across the graph shows that stacked and grouped bar charts more readily incorporate a larger number of *x*-levels than fill-levels – the panels are therefore wider than high. Apparently, this tendency is more pronounced for grouped bar charts.[3]

The two-way distribution in Figure 11 also reveals differences between the two types of bar chart in the combination of these level counts. In grouped bar charts, the number of fill levels clearly decreases as the number of *x*-levels increases. With the exception of a single outlier (13 fill levels, 18 *x*-levels), most arrangements with 10 or more clusters feature at most four fill levels (i.e. four bars per cluster). This relationship does not hold for stacked bar charts, where the number of fill levels does not vary systematically with the number of *x*-levels.

2 It may be worthwhile, at this point, to recall that some guidebooks suggest that at most three fill levels should be compared. We note that stacked bar charts commonly exceed this threshold.

3 We note, in passing, that this runs counter to Wilke's (2019) suggestion to give preference to grouped bar charts when there are many fill levels, and to prioritize the stacked form when there are many *x*-levels.

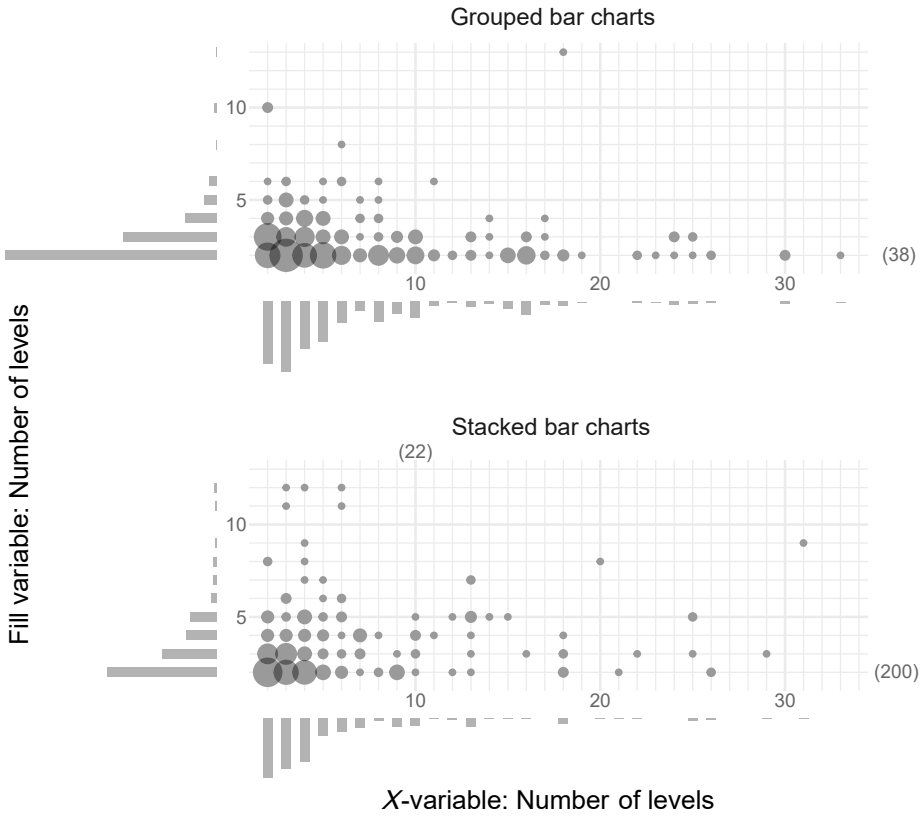


Figure 11. Grouped and stacked bar charts: cross-tabulation of the number of categories of the fill variable (vertical scale) and the *x*-variable (horizontal scale). The area of the circles is proportional to the number of bar charts showing a given constellation, and the histograms show the marginal distribution of the two tallies.

We can summarize by stating that bar charts for corpus data presentation usually show percentages and absolute or normalized frequencies, and that the choice between simple, stacked, and grouped forms varies with the type of quantity. Our analysis further revealed a considerable share of complex visualization tasks, where the chart includes a relatively large number of *x*-levels and fill levels.

5 Bar charts in corpus-based work: Design issues

We will now take a critical look at the use and design of bar charts in our report pool, dealing in turn with what may be considered major and minor aspects. We start with issues that affect the ability of the viewer to extract valid and meaningful information from the display.

5.1 Major issues

Axis title. There is considerable variation in our report pool with regard to whether and how information is provided about the *y*-variable and its units. The major strategies[4] were the following:

– 48% use an axis title (n = 512)
– 25% relegate this information to the figure caption (n = 265)

4 Minor formats were the addition of numeric percentage values directly to the bars (n = 15, 1%), the use of the axis title (n = 14, 1%), and the legend of the chart (n = 1).

– 11% add percentage signs to the tick mark labels (*n* = 116)
– 13% fail to provide information about the *y*-variable (*n* = 135)

I also decided to take a look at whether the information given was instructive:
– For percentages, the details provided were usually sufficient, with only 4% of charts remaining vague (e.g., "Frequency", "Relative frequency") and 1% giving opaque cues.
– For normalized frequencies, 18% of the charts lacked specificity about the *y*-variable (e.g., "Frequency", "Relative frequency").
– For absolute frequencies, 31% of the figures failed to be clear about the kind of frequency shown (e.g., "Frequency").

Proper scaling. The vast majority of bar charts in our survey are scaled appropriately (94%, *n* = 1,002). The two most common issues occurred at similar rates: 28 charts (2.6%) do not include a zero baseline and in 25 figures (2.4%), the facets showing subsets of the data do not have the same scale, which distorts visual comparisons across the panels (Wallgren et al. 1996: 20; see Unwin 2015: 216–218 for a helpful summary of "graphics principles for comparisons").

Time as an x-variable. Trends over time are shown by 10% of the bar charts (*n* = 111). In most cases, the spacing between the columns is proportional to the time intervals; in 30 figures, however, the position of bars along the *x*-axis gives a distorted representation of diachronic trends.

5.2 Minor issues

We now turn to a number of issues that may be considered minor in the sense that they mainly revolve around aesthetic preferences, figure layout, and the efficiency of graphs. Nevertheless, all of these points of critique are grounded in the data visualization literature.

Ordering. When it comes to the arrangement of categories in the chart, I only examined whether the *x*-levels are ordered or not. Note, however, that ordering likewise matters for fill variables and faceting variables. We disregard charts with only two *x*-levels (17%, *n* = 182) or an ordinal *x*-variable (30%, *n* = 314) and concentrate on *x*-variables with three or more unordered levels. My assessment of the remaining 565 charts shows more or less a 50-50 split: in 47% of these (*n* = 263), the *x*-levels are ordered by size, and in 53% (*n* = 302) there is no discernible systematic arrangement.

Figure size. The physical dimensions of a figure on the printed page are relevant for two reasons. For one, a graph should not be wasteful of space – its size should be in proportion to the amount of information it contains. Further, the larger the figure, the longer it takes the eye to scan and interpret its contents. In this regard, a relevant concept from vision science is the useful field of view. It refers to the area from which we can extract information quickly, i.e. without the need for eye movements (see Ware 2013: 151, 173). This region is about 2 to 5 degrees of visual angle, which, at a normal reading distance of 40 cm, corresponds to a diameter of 1.4 to 3.5 cm. Figure 12 shows the size of the bar charts in our survey compared to the approximate extension of the useful field of view, and relative to the text area of our journals, which is the region in which material is printed. Each chart is represented by a partly transparent frame. Two points are noteworthy:
– There is a tendency for figures to use up the horizontal space available, and most figures are therefore wider than high.
– The size of all figures exceeds the useful field of view, most of them by a considerable margin.

Information density. Whether a figure is actually wasteful of space depends on the amount of information it contains – other things being equal, graphs that show more quantitative values will require more space. We must therefore consider the area of bar charts in relation to the number of values they depict (which range from 2 to 1,980 in our sample). Figure 13 shows this relationship using a scatterplot, where each circle represents a bar chart in our survey. The dot diagram at the right margin reflects the overall distribution: most bar charts occupy somewhere around (to just over) one quarter of the text area. While we observe the expected positive association, there is considerable variation at each

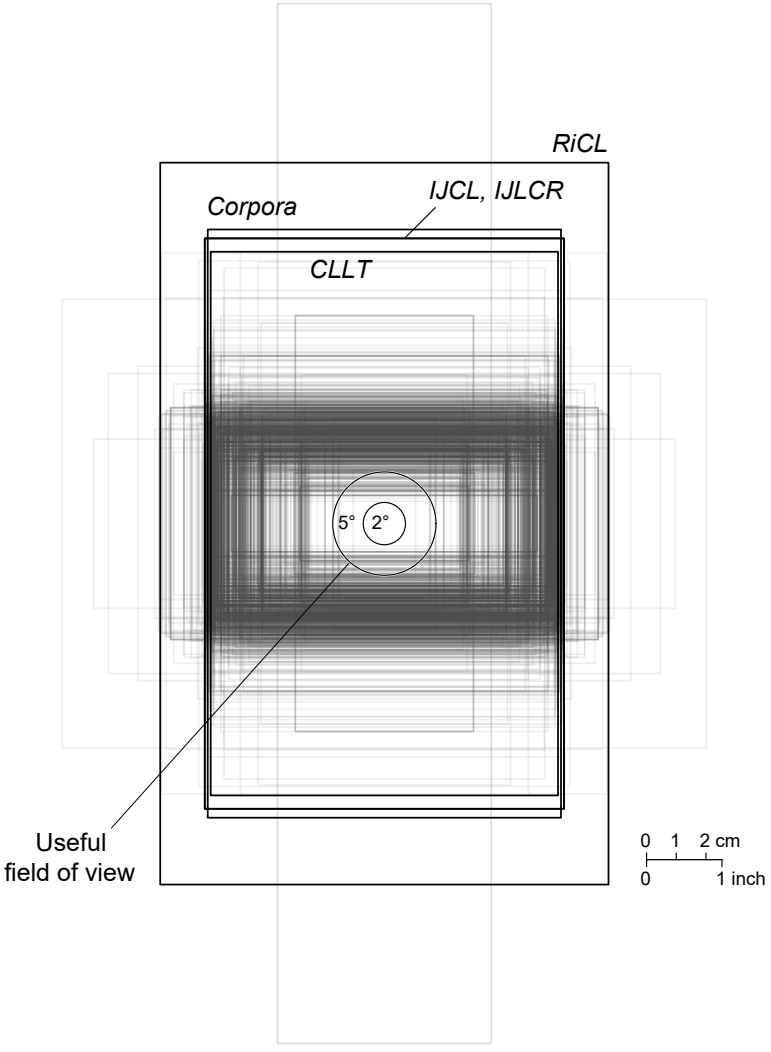


Figure 12. The size of bar charts in comparison with the text area of journals and the useful field of view.

location of the *x*-axis. At the far left, for instance, the charts showing only two values differ considerably in size – the smallest one falls comfortably within the useful field of view, while the largest ones take up more than a third of the typical text area in our selection of journals. The floating axis in the lower right corner of the plot can be used as a measuring stick: it shows that for a given number of numeric values that are visualized, the size of bar charts can vary by up to a factor of 8 to 1, indicating that for many authors, space-saving visualization is not a priority.

Fill colour. Overall, 45% (n = 299) of the grouped and stacked bar charts in our survey used colour to distinguish between fill levels, meaning that 55% (n = 370) relied on other means such as greyscales or fill patterns. From the viewpoint of chart design, a typical recommendation is to start with greyscales (Schwabish 2021: 44–45; Nussbaumer Knaflic 2015: 117–119). Colour is usually necessary when the number of fill levels increases. In grouped and stacked bar charts with black contours around the bars, it is usually possible to clearly discriminate between more than four shades of grey, a value that has been suggested for point symbols in scatterplots (see Ware 2013: 75, 122). Reference back to Table 3 suggests that greyscales would probably suffice for the vast majority of grouped and stacked bar charts: only 13% depict five or more fill levels. As Figure 14 shows, there is, perhaps surprisingly, no clear relationship in our sample between the use of fill colour and the number of fill levels: about 36%

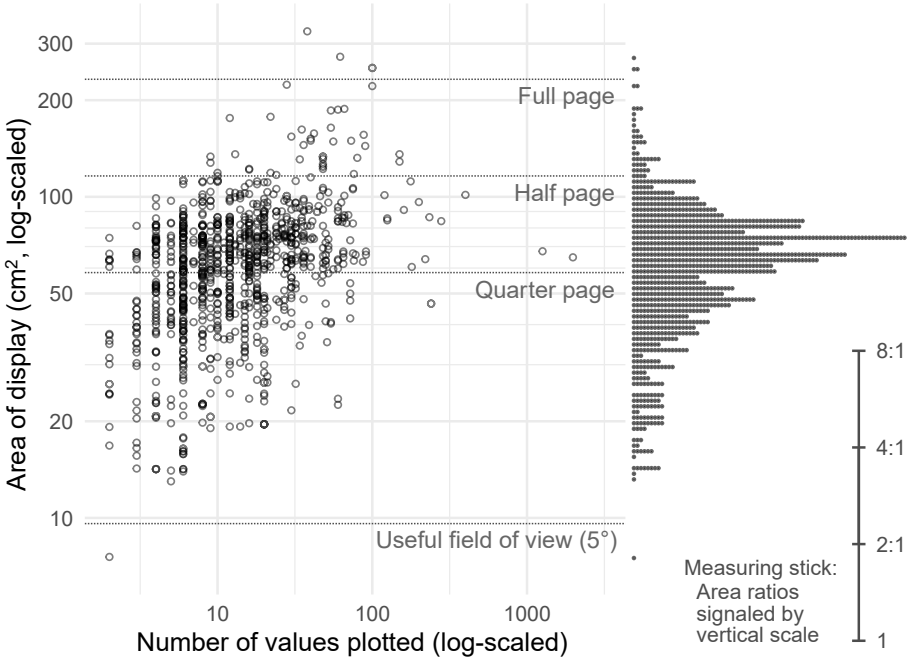


Figure 13. The relationship between the area of the chart and the number of values that are graphed.

of charts with two fill levels use colour, and the proportion climbs to around 55% when the number of fill levels exceeds 2. From the viewpoint of chart design, then, the decision to use color to distinguish fill levels seems to be based on other considerations.[5]

Direct labelling. Data visualization guides recommend the use direct data labels instead of a legend whenever possible (e.g., Schmid 1983: 19; Kosslyn 2006: 106–107; Nussbaumer Knaflic 2015: 96; Camões 2016: 346; Schwabish 2021: 34–36). This accelerates the decoding of a graph, since it obviates repeated eye movements back and forth between data and key (Milroy & Poulton 1978). Direct labelling is an option for stacked bar charts, where they can be placed to the left of the first (or to the right of the last) column (as in Figures 9 and 14; see Kosslyn 2006: 190–191). Out of the 264 stacked bar charts in our survey, only 6% (n = 17) use direct labels to denote the fill levels.

Rotated data labels. The vast majority of the bar charts in our survey are vertical, with the bars in an upright position (87%, n = 922). In 200 of these vertical bar charts, the data labels for the x-levels were too long and therefore rotated. This means that the readability of about 1 in 5 bar charts could perhaps be improved if the chart were arranged horizontally. In the 135 horizontal bar charts that appear in our survey, data labels always have a horizontal orientation.

Number of tick mark labels. When marking the units of the y-variable, a balance must be struck between too few and too many labels. The literature cautions against too many tick marks, to avoid a cluttered display. Cleveland (1994: 39) recommends between 3 and 10, some style guides[6] suggest starting with about 5 to 7, and Wilkinson (2005: 96) proposes a search algorithm that starts at 5, which is also implemented in the 'ggplot2' package (Wickham 2016). Figure 15 shows the distribution of the number of tick mark labels in our bar chart survey. The most popular choice is 11, and the vast majority of these charts demarcate percentages with 10-point intervals running from 0 to 100.

5 It should also be noted that, since around 2015, the journals published by Benjamins (IJCL and IJLCR) redraw graphics to enforce a house style that uses custom color palettes to distinguish fill levels.

6 See, e.g., https://www.quanthub.com/designing-charts-axis-tick-marks/



Figure 14. Mosaic chart showing the proportional distribution of greyscales vs. fill colours by number of fill levels.

Cues to facets. In 157 bar charts (15%), an additional variable is incorporated into the graph using faceting. Different design features may be used to clearly distinguish between the subgroups in the different panels. The clearest separation is achieved with a combination of facet labels and interspaces between panels (see Kosslyn 2006: 130–131); 96 charts were drawn in this way. 32 charts only used a facet label, and 20 figures provided no grouping cues, with the key to the facets hidden in the data labels for the *x*-levels.

3D effects. The data visualization literature routinely discourages the use of an additional third dimension in a two-dimensional chart (e.g., Nussbaumer Knaflic 2015: 65–66; Few 2012: 197–202). In our survey, only few bar charts (n = 23, 2%) were drawn with pseudo-3D effects.

6 Implications and recommendations

Our survey of the use of bar charts in corpus data visualization allows us to consider their limitations and design issues from a discipline-specific perspective. The current section distils our observations into a number of recommendations targeted at corpus linguists.

6.1 Countering the limitations of bar charts

The types of quantities that are most commonly visualized in corpus linguistics have a natural zero point. If the start-at-zero constraint leads to poor resolution, the researcher may consider excluding the baseline from the display (see Franconeri et al. 2021: 116–117 for a careful discussion of this design choice). This decision must be made carefully, however, since the zero referent line is a relevant anchor when showing relative or absolute frequencies. At any rate, if zero ends up being removed from the chart, one must switch to a graph type that does not use length but only position to represent *y*-values. Two options are dot plots (Cleveland 1984) and line plots. The advantages of both graph types over bar charts are discussed in detail in Sönning (2016) and Sönning (2023).

Another strategy for addressing poor resolution is to graph the *y*-variable on a transformed scale. In corpus-based work, this makes sense for normalized and absolute frequencies. Two options are the square-root transformation, which allows us to include a zero baseline into the graph, and a logarithmic transformation, which has the advantage of showing multiplicative differences (see Cleveland 1994: 120–126; Robbins 2005: 73–75). In both cases, the constraints on *y*-variables force us to use another graph type such as a dot plot or a line plot (see Sönning 2016, 2023 for several illustrations). Seeing that 18 bar charts in our survey showed ratios, it is perhaps also worth stressing that the log scale is the natural scale on which to compare ratios (see Wilke 2019: 21).

Our survey has revealed that bar charts are quite often used for relatively complex visualization tasks. If the number of *x*-levels and the number of fill levels is large, visual clutter can obscure the patterns in the

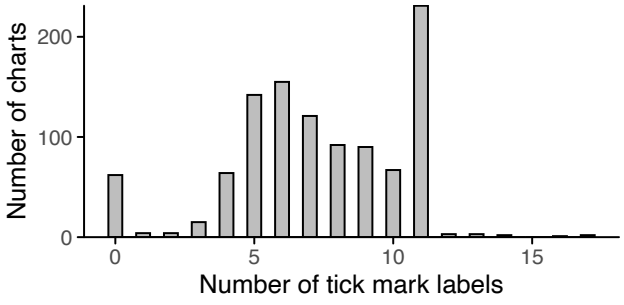


Figure 15. Histogram showing the distribution of the number of tick mark labels used to demarcate the units of the *y*-variable.

data. In such scenarios, it can help to switch to a more minimalistic graph type. To illustrate this, Figure 16 re-expresses Figure 5 above using a dot plot.[7] Two additional strategies are used: the categories are sorted according to the percentage of the Modifier category, which creates a (more) regular pattern, against which the percentage for the other category can be compared. Further, the graph includes a second panel, which shows the differences between the two percentages directly. This technique is sometimes referred to as "directly depicting deltas" (Franconeri et al. 2021: 124; see also Schwabish 2021: 85) – it allows us to quickly identify categories with positive and negative differences, as well as categories with particularly large differences. The fact that 56% of grouped bar charts in our survey have two fill levels (see Table 3) suggests that this technique may be of wider applicability. Note that, apart from obtaining a minimalist and clean aesthetic, the issue of rotated data labels has also been resolved.

It is perhaps also worth mentioning, at this point, that stacked bar charts with a two-level fill variable can be simplified to a simple bar chart (or a dot plot/line plot). This is because a single proportion (or percentage) fully describes the distribution of a binary variable. With 45% of the stacked bar charts in our survey having two fill levels (see Table 3), it seems that corpus data quite frequently afford the opportunity to do so. However, if it is important to depict the part-whole relation in the data, this simplification is not advised (see Schwabish 2021: 91).

In stacked bar charts, the problem of non-aligned intermediate fill levels seems to be particularly relevant in corpus data visualization: reference back to Table 3 shows that 22% of the stacked forms have three, 12% four, and 20% five or more fill levels. Strategies for addressing this issue depend on the comparisons that (are meant to) take centre stage. If the fill levels need to be compared across columns, Kosslyn (2006: 118, 134) recommends putting the category that varies least at the bottom of the columns. Another option is illustrated in Figure 17, which reorganizes Figure 6 by separating the columns; bars then rest on the same baseline. This is sometimes referred to as a panel chart[8] (see Camões 2016: 217–218; Schwabish 2021: 88–89). Comparison of bars across *x*-levels (i.e., adverbs) can now be made with ease. As in the original chart, the adverbs are ordered systematically; further, the position categories now form an iconic sequence starting with "I" (initial) at the far left. Within the 'ggplot2' framework, switching between Figures 17 and 6 is relatively straightforward, as Figure 17 merely changes the status of Position from a fill variable to a faceting (as well as fill) variable. In general, faceting is a useful strategy to (try to) alleviate overloaded displays (see Schwabish 2021: 87).

6.2 Addressing major issues in bar chart design

One major issue stands out in terms of prevalence: axis labelling. The literature on chart design states that statistical graphics should generally have clear and informative axis titles (e.g., Kosslyn 2006: 26;

7 For an illustration of how to draw such charts with the R package 'ggplot2' (Wickham 2016), see https://lsoenning.github.io/posts/2025-11-28_drawing_grouped_dotplots/.

8 For an illustration of how to draw such charts with the R package 'ggplot2' (Wickham 2016), see https://lsoenning.github.io/posts/2025-11-28_drawing_panel_charts/.

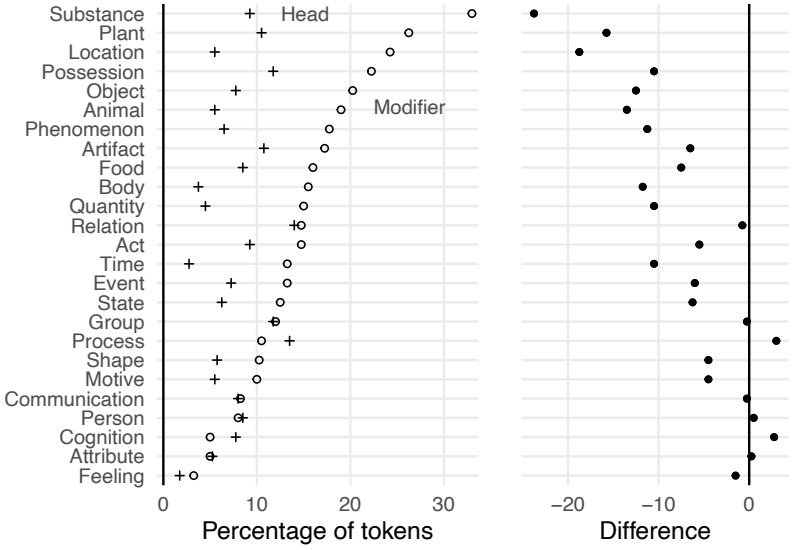


Figure 16. Redesign of Figure 5: a dot plot alleviates the visual clutter, the *x*-levels are ordered systematically, and an appended panel shows the differences directly.

Schwabish 2021: 354). It would seem unusual to give this information in the figure caption, which instead serves to add notes and explanations that make the figure as self-contained as possible (Cleveland 1994: 54–59). In general, then, any numeric scale in a statistical graph should have an axis title. Deficiencies in the informativeness of the description were also apparent. Importantly, in the context of corpus data visualization, the label "Frequency" is too vague, as it fails to distinguish between normalized and absolute (or raw) figures. The same is arguably true for the designation "Relative frequency", which may refer to a percentage/proportion or to a normalized occurrence rate. Another piece of information that is often missing from axis titles is the kind of units that were counted, which could be (word) tokens, types, or texts. And for normalized frequencies, the basis for normalization should be stated (e.g., "per million words"). For convenience, we list a number of exemplary axis titles we found in our survey:

– Percentages: "Deontic modal usage (% of tokens)", "Observed percentage of *this*", "Percentage of files using feature", "Percentage of tokens"
– Absolute frequencies: "Number of occurrences", "Number of bundles (tokens)"
– Normalized frequencies: "Frequency per 1,000 words", "Occurrences per million words", "Rate per million words", "Normalized frequency (pmw)"

While only 1 in 20 bar charts in our survey was scaled inappropriately, it is worth repeating two principles: bars should start at the relevant baseline (usually zero), and charts that appear side by side and show the same *y*-variable for different subsets of the data should be scaled identically.

One out of 10 bar charts in our report pool was used to show trends over time. Researchers considering the use of bar charts for temporal *x*-variables should keep in mind two things: first, the canonical graph type for time trends is the line plot, which makes more effective use of ink and is less cluttered (Schwabish 2021: 87). Further, if time intervals vary in length, the spacing of bars should be adjusted (Schmid 1983: 54, 56).

6.3 Addressing minor issues in chart design

Finally, let us recapitulate the minor design issues:

– Our survey showed that the systematic arrangement of *x*-levels is an underutilized strategy in corpus data visualization. We should routinely check whether a display can be made more informative if the *x*-levels are rearranged, typically according to size (see Figures 6 and 16).

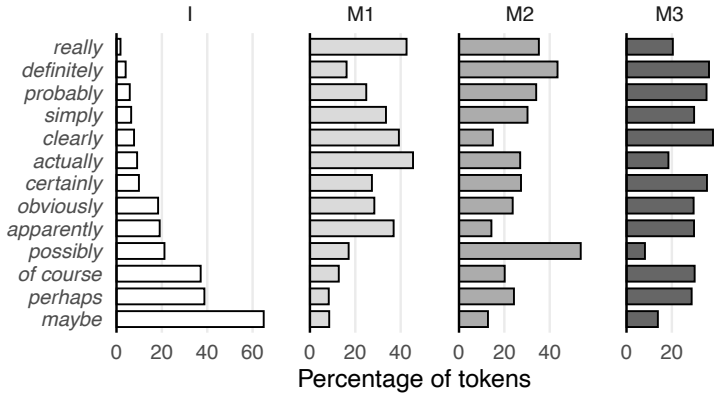


Figure 17. Redesign of Figure 6: the intermediate categories (M1, M2) are now easier to compare.

– A considerable share of bar charts in our survey was larger than they need to be, which suggests that figure size should be reduced (within reason).
– In grouped and stacked bar charts, it makes sense for fill colour to be used judiciously and strategically; often greyscales suffice.
– Researchers may routinely consider the option of switching to a horizontal format, especially if there are many categories (see Figure 10) and/or the data labels are long.
– Stacked bar charts can profit from the use of direct data labels (see Figure 9 and 14).
– The number of tick mark labels should be reasonable, perhaps somewhere around five to seven.[9]
– When faceting variables are used, the reader profits from unambiguous cues to the subgroups, ideally a combination of direct facet labels and interspaces between facets.
– 3D effects should routinely be avoided.

7 Summary and conclusion

The present paper was motivated by the observation that the bar chart is the most commonly used graph type for corpus data presentation. Since it is (and will likely continue to be) such an important vehicle for the communication of corpus findings, it is worthwhile to examine in more detail the way in which we use them. Our systematic review showed that bar charts in corpus-based research articles are most commonly used to show percentages as well as absolute and normalized frequencies. Perhaps this makes the bar chart a natural choice: it is anchored at zero, which is the baseline and critical reference point for these quantities. Our survey has also demonstrated, however, that bar charts in corpus-based publications are often visually quite complex, including several variables with multiple levels each. These visualization tasks push bar charts to (or perhaps beyond) their limits. We therefore looked at how alternative graph types can alleviate the display and elucidate the visual message. Our stocktaking exercise also suggested that certain visualization strategies, such as "directly depicting deltas" or the use of panel charts (see Section 6.1) could potentially be used at a much larger scale.

Since presentation graphs in published research articles should arguably be held to the highest standards, we also scrutinized what we considered as major and minor design issues, and condensed these into a set of concrete recommendations. This will hopefully help corpus linguists to continue to make good use of bar charts for communicating their data. To remind ourselves of the tremendous importance of this graph type for (corpus) data visualization, it seems appropriate to end by quoting Cole

9 Further, they should represent "natural" scale values, which means that intervals between them are 1, 2, or 5 (or 10, 20, or 50, etc.); in the case of percentages, intervals of 25 are natural (see Wallgren et al. 1996: 20; Schmid 1983: 42).

Nussbaumer Knaflic (2015: 40), who put it this way: “Sometimes bar charts are avoided because they are common. This is a mistake. Rather, bar charts should be leveraged *because they are common*, as this means less of a learning curve for your audience.”

Acknowledgements

I would like to thank two anonymous reviewers and the editors, Merja Kytö and Patricia Ronan, for their constructive feedback. This research was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; grant number 548274092).

References

Camões, Jorge. 2016. *Data at work: Best practices for creating effective charts and information graphics in Microsoft Excel*. Hoboken, NJ: New Riders.

Cleveland, William S. 1984. Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician* 38 (4). 270–280. https://doi.org/10.2307/2683401

Cleveland, William S. 1994. *The elements of graphing data*. Summit: Hobart Press.

Few, Stephen. 2012. *Show me the numbers: Designing tables and graphs to enlighten*. Oakland, CA: Analytics Press.

Fienberg, Stephen E. 1979. Graphical methods in statistics. *American Statistician* 33 (4). 165–178. https://doi.org/10.1080/00031305.1979.10482688

Franconeri, Steven L., Lace M. Padilla, Priti Shah, Jeffrey M. Zacks & Jessica Hullman. 2021. The science of visual data communication: What works. *Psychological Science in the Public Interest* 22 (3). 110–161. https://doi.org/10.1177/15291006211051956

Huff, Darrell. 1954. *How to lie with statistics*. New York: W. W. Norton.

Jeppson, Haley & Heike Hofmann. 2023. Generalized mosaic plots in the ggplot2 framework. *The R Journal* 14 (4). 50–78. https://doi.org/10.32614/RJ-2023-013

Karsten, Karl G. 1925. *Charts and graphs: An introduction to graphic methods in the control and analysis of statistics*. New York: Prentice Hall.

Kosslyn, Stephen M. 1985. Graphics and human information processing: A review of five books. *Journal of the American Statistical Association* 80 (391). 499–512. https://doi.org/10.2307/2288463

Kosslyn, Stephen M. 2006. *Graph design for the eye and mind*. Oxford: Oxford University Press.

Milroy, Robert & E. Christopher Poulton. 1978. Labelling graphs for improved reading speed. *Ergonomics* 21 (1). 55–61. https://doi.org/10.1080/00140137808931693

Nussbaumer Knaflic, Cole. 2015. *Storytelling with data: A data visualization guide for business professionals*. Hoboken, NJ: Wiley.

Playfair, William. 1787. *The commercial and political atlas*. London: Corry.

Robbins, Naomi B. 2005. *Creating more effective graphs*. Hoboken: Wiley.

Sarkar, Deepayan. 2008. *Lattice: Multivariate data visualization with R*. New York: Springer.

Schmid, Calvin F. 1983. *Statistical graphics: Design principles and practices*. New York: Wiley.

Schwabish, Johnathan. 2021. *Better data visualizations: A guide for scholars, researchers, and wonks*. New York: Columbia University Press.

Sönning, Lukas. 2016. The dot plot: A graphical tool for data analysis and presentation. In Hanna Christ, Daniel Klenovšak, Lukas Sönning & Valentin Werner (eds.), *A blend of MaLT: Selected contributions from the Methods and Linguistic Theories Symposium*, 101–129. Bamberg: University of Bamberg Press. https://doi.org/10.20378/irbo-51101

Sönning, Lukas. 2023. Drawing on principles of perception: The line plot. In Lukas Sönning & Ole Schützler (eds.), *Data visualization in corpus linguistics: Reflections and future directions* (Studies in Variation, Contacts and Change in English 22). Helsinki: VARIENG, University of Helsinki. https://urn.fi/URN:NBN:fi:varieng:series-22-2

Sönning, Lukas. 2025. *Background data for: An analysis of bar chart usage in corpus data visualization*, https://doi.org/10.18710/ILJY4U. DataverseNO, V1.

Sönning, Lukas & Ole Schützler. 2023. Data visualization in corpus linguistics: Critical reflections and future directions. In Lukas Sönning & Ole Schützler (eds.), *Data visualization in corpus linguistics: Critical reflections and future directions*. Helsinki: VARIENG, University of Helsinki. https://urn.fi/URN:NBN:fi:varieng:series-22-0.

Tufte, Edward R. 2001. *The visual display of quantitative information*. Cheshire: Graphics Press.

Unwin, Antony. 2015. *Graphical data analysis with R*. Boca Raton: CRC Press.

VERBI Software. 2021. MAXQDA 2022 [computer software]. Berlin, Germany: VERBI Software. Available from maxqda.com.

Wainer, Howard. 1997. *Visual revelations*. Mahwah: Erlbaum.

Wallgren, Anders, Britt Wallgren, Rolf Persson, Ulf Jorner & Jan-Aage Haaland. 1996. *Graphing statistics and data: Creating better charts*. London: Sage.

Ware, Colin. 2013. *Information visualization: Perception for design*. Amsterdam: Elsevier.

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Wilke, Claus O. 2019. *Fundamentals of data visualization*. Sebastopol, CA: O'Reilly.

Wilkinson, Leland. 2005. *The grammar of graphics*. New York: Springer.