
STUDIES IN VARIATION, CONTACTS AND CHANGE IN ENGLISH

VOLUME 22

Data Visualization in Corpus Linguistics: Critical Reflections and Future Directions

Drawing on principles of perception: The line plot

Lukas Sönning
University of Bamberg

Please cite this article as:

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

BibTeX format

@incollection{Sönning2023,
  author = "Lukas Sönning",
  title = "Drawing on principles of perception: The line plot",
  series = "Studies in Variation, Contacts and Change in English",
  year = 2023,
  booktitle = "Data visualization in corpus linguistics: Critical reflections and future directions",
  number = "22",
  editor = "Sönning, Lukas and Schützler, Ole",
  publisher = "VARIENG",
  address = "Helsinki",
  url = "https://urn.fi/URN:NBN:fi:varieng:series-22-2",
  issn = "1797-4453"
}

Abstract

This paper draws attention to an underused display type for corpus data visualization: the line plot. While this graph type is commonly associated with time series data, its true potential arguably unfolds in the application to multifactorial data sets involving discrete (categorical) variables. Data layouts of this kind are typical of corpus-based work and the preferred vehicle for their visualization is currently the bar chart. It is sometimes argued that line plots should only be used when the horizontal axis represents a continuous trait. However, once we allow for the levels of binary and categorical variables to be connected by lines, we recognize that this form offers several distinct advantages over bar charts. This is especially true for visualization tasks involving multiple predictor variables. The paper starts out by providing some theoretical background for the comparative evaluation of graph types, with a focus on quantitative comparisons and perceptual processing. Drawing on empirical insights into visual perception, evidence-based recommendations for the design of line plots are given. These include the choice of line types and plotting symbols, the use of direct labeling, and the arrangement of variables in the display. Following this, key advantages of line plots are illustrated. These include pictorial minimalism, the availability of extended encoding strategies and the scaffolding provided by perceptual grouping laws. The paper closes by emphasizing limitations of this display type. These concern the depiction of non-continuous x-variables and the asymmetric perception of interactions among predictor variables. While ample attention should be paid to these issues, we argue for a (more) routine use of line plots in corpus data visualization.

1. Introduction

Linguistics has developed into a strongly quantitative science, a development that has been partly driven by the emergence of corpus linguistics as a core methodological stream. Corpus-based research often deals with multiple dimensions of variation and complex mixtures of conditions, which are a typical feature of natural language data. Strategies for the effective communication of quantitative insights therefore gain importance. This paper draws attention to a display type that outperforms many competitors in the face of complexity: the line plot. Its primary context of application are time series data, the setting for which

it was originally invented (Playfair 1786). The objective of the present chapter is to argue that the line plot unlocks its true potential in settings with multiple discrete (categorical) predictors. In these contexts, line plots are currently outnumbered by bar charts in the research literature (see Sönning & Schuetzler, this volume). Our aim, then, is to draw attention to this alternative tool, discuss its added value, and argue for a more widespread usage of line plots for the communication of (corpus) linguistic data.

The paper is structured as follows. Section 2 introduces concepts and terminology for the discussion of graph types and graph design. Section 3 illustrates key components of line plots, and Section 4 spells out design recommendations. In Section 5, we highlight key advantages of line plots over bar charts, and Section 6 reflects on important limitations of this format. Section 7 concludes with a summary of the main points.

2. Theoretical preliminaries

The comparative assessment and evaluation of statistical graphs can draw on theoretical and empirical insights gained across various disciplines. This section sets forth a conceptual and terminological foundation for the ensuing discussion.

2.1 Comparisons

As noted by Tufte (1990: 67, emphasis in original), "at the heart of quantitative reasoning is a single question: *Compared to what*?" In other words: Numbers are rarely interpretable in isolation, but gain meaning through context and contrast. For statistical graphics, Gelman (2009, emphasis in original) goes as far as stating that "*all* graphs are comparisons". To create efficient representations, then, it is essential to be clear about which comparisons a display should show. This communicative goal guides the choice of graph type and the arrangement of variables in a display. The success of a visual representation depends on whether the viewer can draw the intended comparisons with ease.

In multivariate displays of data, where graphical elements can be grouped in different ways, two types of comparisons are critical. [1] We will use the term 'group' to refer to a grouping of data points, or visual elements, according to a particular attribute. Groups may reflect category membership (e.g. animate vs. inanimate referents). Further, we will use the term 'subgroup' to denote cross-classified groupings, i.e. according to two (or more) attributes (e.g. animate referents whose discourse status is given vs. new).

To make matters more concrete, consider the English dative alternation – that is, the variation between a double object structure (e.g. *I gave* ***her*** *the book*) and a prepositional dative structure (*I gave the book* ***to her***). The choice between these forms varies, among other things, with the animacy and discourse accessibility of the referent. For illustration, we use data from a study by Bresnan et al. (2007). In Figure 1a, the outcome quantity, the share of prepositional datives, is expressed as a proportion and shown on the vertical axis. We can see that prepositional variants are more typical for inanimate referents and for new discourse entities.

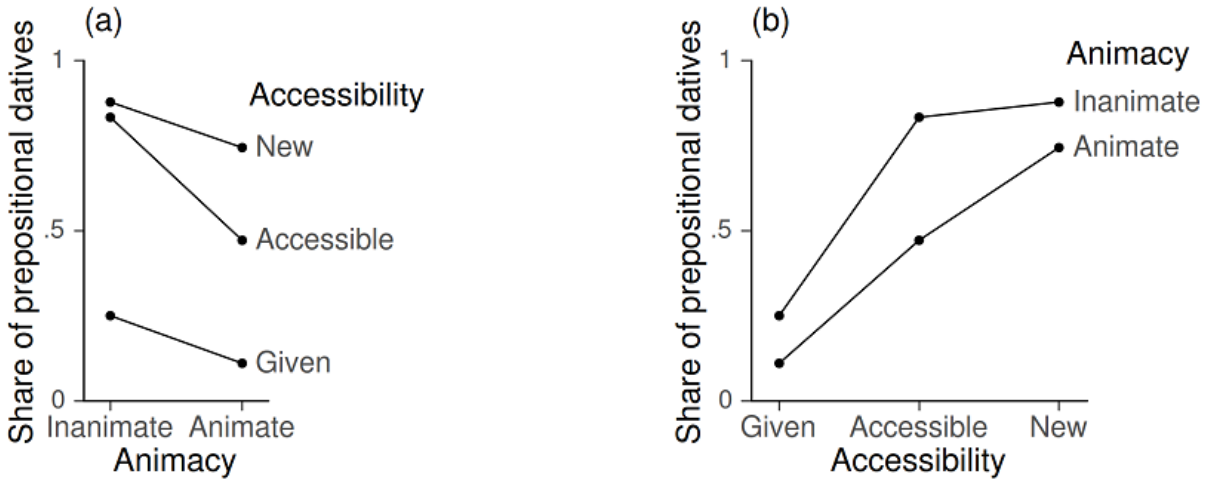

Figure 1. Line plots showing the proportion of prepositional datives by animacy and accessibility (data from Bresnan et al. 2007). Panels (a) and (b) are informationally equivalent, but arrange the data in different ways.

Throughout this paper, we will use the shorthand symbols X, Y, and Z to refer to the variables represented in a line plot. As usual, Y will refer to the outcome quantity on the y-axis. In Figure 1, this is the proportion of prepositional datives. X will denote the variable that is shown along the horizontal scale (x-axis). In Figure 1a, this is the animacy of the recipient. Finally, Z refers to additional variables in the display – in Figure 1a, for instance, the accessibility of the recipient. A graph may include multiple Z-variables.

Of course, formalisms of this kind simplify things for the writer, but not for the reader. We will therefore use concrete labels (e.g. "accessibility" instead of "Z") whenever possible. A key point we wish to stress, however, is that the distinction between X- and Z-variables is important once we discuss the interpretation of line plots. Thus, Figures 1a and b are informationally equivalent, but the arrangement of variables in the display has consequences for the types of comparisons made by the viewer. We will return to this point in section 4.7.

In general, two types of comparisons can be made when reading line plots of the type shown in Figure 1:

- *Average comparisons*: We might be interested in how the share of prepositional datives varies across accessibility levels, ignoring animacy, and vice versa. To answer this query, we need to disregard, i.e. mentally average over, the levels of animacy in the display, which produces the following cline: new > accessible > given. Averaging over accessibility, we find that the prepositional share is larger among inanimate referents. [2]
- *Interaction comparisons*: It might also be of interest whether, say, the difference between inanimate and animate referents is constant across accessibility levels. More generally, we are asking whether a particular predictor-outcome relationship is stable (i.e. roughly the same) across the levels of another predictor. In Figure 1a, the slopes indicate that the directionality of the animacy cline is consistent across accessibility subgroups; 'accessible' referents, however, appear to show the strongest link between animacy and the outcome, at least as far as absolute differences on the proportion scale are concerned. [3]

Importantly, interaction comparisons can be interpreted from different angles. Thus, we could also ask whether the association between accessibility and the outcome is the same for animate and inanimate referents. From this perspective, Figure 1b indicates a stable rank order. For animate referents, however, the increments along the accessibility scale are roughly equidistant (about .3 to .4 in absolute terms). For inanimate referents, on the other hand, we discern a hockey-stick pattern, as accessible and new entities are nearly on a par. The perspective that is chosen for interaction comparisons (here: Figure 1a or b) biases the viewer towards certain comparisons but not others. We will return to this point in section 4.7.

For statistical graphics, then, the notion of comparisons plays a critical role and guides the purposeful selection of visual forms and arrangements. Next, we turn our attention to how people read information from graphs.

2.2 Perceptual processing

To further our understanding of the perception and processing of visual information, we will adopt a heuristic model of perceptual processing, which builds on Rensink (2000) and Ware (2013). Graphical perception is assumed to operate at three levels: (i) a low-level, pre-attentive stage that is driven by rapid and automatic feature extraction in a bottom-up fashion, (ii) a high-level stage at which our visual working memory directs attention and initiates top-down visual search strategies, and (iii) an intermediate stage, which is characterized by the interplay of bottom-up processing and top-down attentional control. Each level

offers insights into the way in which visual information is processed.

The pre-attentive attributes of visual stimuli determine their discriminability in a scene. Knowledge about the factors underlying the perceived distinctness of elements is critical when multiple categories need to be distinguished in a graph. At the pre-attentive stage, the input undergoes rapid parallel processing via neurons in the primary visual cortex. Importantly, these neurons form four groups, each group being tuned to selectively detect certain features. Accordingly, there are four different channels between eye and mind. Corresponding to these channels, the four elementary feature types are form, color, motion and stereoscopic depth (see Ware 2013: 143–152). Since these attributes stimulate different neurons, they undergo separate processing. As a result, they can be parsed independently of each other, which suggests that different grouping variables are ideally mapped onto different channels. Static diagrams rely on form and color, which provide two orthogonal means of signaling category membership. The form channel is particularly versatile, and experimental work has yielded practicable insights into the discriminability of variations in texture, orientation, and size. We will return to these in section 4.1, where we give advice on the use of plotting symbols.

At stage two, processing is constrained by the capacities of our visual working memory, which can retain only few objects from one fixation to the next. Attentional control directs the perception of the visual scene by detecting patterns and grouping entities into visual objects. The term ‘visual object’ will be used to refer to a group of elements in a display that is processed as a chunk. According to Rensink’s (2000) coherence theory, pre-attentive features form so-called proto-objects, or potential chunks. Proto-objects are latent groupings that may be brought into visual working memory through focused attention. In order for visual chunks to be formed, attention must be allocated to certain features, or feature sets, in the input.

The conscious grouping of elements according to certain attributes gains importance in multivariate displays, where the viewer must be able to isolate elements into distinct visual objects. If comparisons between subgroups are critical, selective perception of two or more groupings is required. The ease with which a visual object can be formed and held in working memory depends on the attraction and cohesion of the chunk. Apart from pre-attentive cues, Gestalt laws of perception (Wertheimer 1938; Ware 2013: 181–99) offer insights into surface features that produce cohesive proto-objects. This gives us some understanding of grouping effects in the presence of diverse visual cues.

Of the Gestalt laws that have emerged over the past century, four are relevant for the present discussion (see Wagemans et al. 2012 for a comprehensive overview). These are illustrated in Figure 2:

- The law of *proximity* states that, other things being equal, elements that are close together will be perceived as a group. In Figure 2a, for instance, the six points in the bottom left corner will be perceived as a group.
- If elements constitute regular forms, symmetric shapes, or objects with the same orientation, the law of *good form* states that they will be perceived as single units. In Figure 2a, this applies to the points that form a path from top to bottom.
- Pre-attentive attributes give rise to the law of *similarity*, which holds that elements that are similar in, say, shape or color will be grouped. Figure 2b shows two distinct groups. In order for two (or more) groups to be separable, the elements must be sufficiently discriminable.
- Finally, the law of *connectedness* holds that linked elements will be perceived as groups (Palmer & Rock 1994). In Figure 2c, four points have been conjoined to form a group. Note how panel (c) allows us to perform different groupings in the same visual scene.

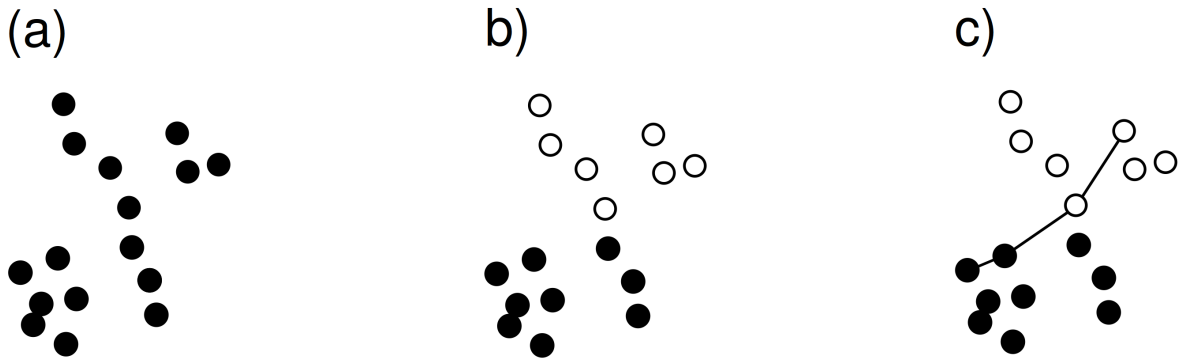

Figure 2. Perceptual grouping laws: (a) proximity, continuity, (b) similarity, (c) connectedness.

Such theoretical insights into graph design and perception provide a foundation for the informed application of statistical graphs in quantitative research. They can guide the choice between different graph types and design options. Before we derive recommendations from these insights, let us have a closer look at the components of line plots to establish a terminological footing for the remainder of this paper.

3. Elements of the line plot

To illustrate the elementary features of line plots, consider Figure 3. The graph shows data from an experimental study on metaphorical language use for the description of static scenes in three different languages (Blomberg 2015). The objective was to examine cross-linguistic regularities in factors triggering dynamic descriptions of static situations (e.g. *The road goes through a river*). Two factors were experimentally varied: (i) the perspective on the scene, i.e. whether the viewer is, say, standing on the road (1st person perspective), or views it from a distance (3rd person perspective); and (ii) the affordance for motion, i.e. whether the object was linked to human motion (afford; e.g. a road) or not (non-afford; e.g. a fence). The outcome quantity is the proportion of non-actual motion (NAM) descriptions, i.e. the share of dynamic verbalizations, where a static situation was described with motion vocabulary.

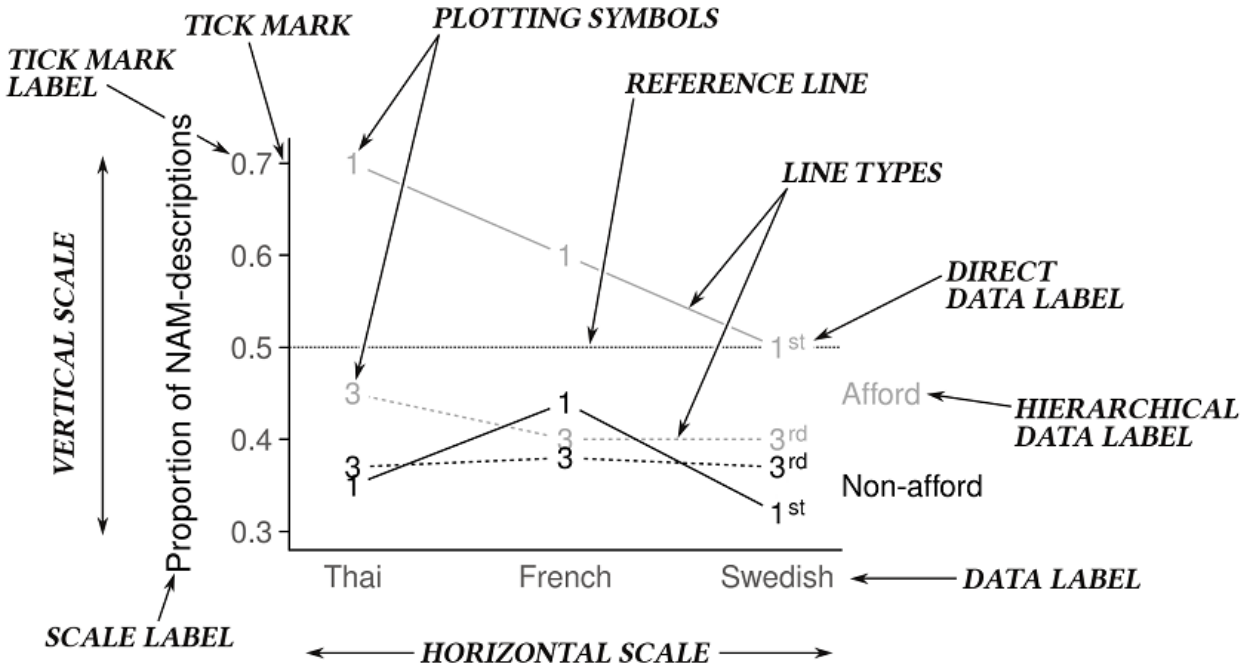

Figure 3. Elements of the line plot. Terminology and style of presentation borrowed heavily from Cleveland (1994: 21–22).

The annotations in the graph highlight several key elements of line plots and serve to establish terminology (largely borrowed from Cleveland 1994: 21–22). The outcome quantity is shown on the *vertical scale*. The *horizontal scale* hosts the focal predictor (language), including *data labels* (Thai, French, Swedish). The experimental conditions are distinguished by color (afford vs. non-afford) and a combination of different *line types* and *plotting symbols* (1st vs. 3rd person perspective). A reference line marks the majority threshold of .5, an informative benchmark. The graph is drawn using an L-shaped framework (rather than a box),

which makes it easier to add annotations and labels to the display. Thus, *direct data labels* and *hierarchical data labels* are attached to the lines.

For practitioners familiar with 'ggplot2' (Wickham 2016), an R package implementing Wilkinson's (2005) *Grammar of Graphics*, the snippet of commented R code below outlines the mapping between data and visual attributes in Figure 3. Language is mapped to x-position, the proportion of NAM-descriptions to y-position. Observations are rendered with two layers of geoms: points and lines. The aesthetic attributes of these are defined as follows: Affordance is mapped to color (gray vs. black). Perspective is mapped to point shape ("1" vs. "3") and to line type (solid vs. dotted). The group aesthetic, which indicates which subgroups should be connected with lines, is mapped to the cross-classification of affordance and perspective. Two further pieces of code specify the shape of plotting symbols and add direct labels:

```
ggplot(                                          # create new plot for
    blomberg,                                    #   the data set 'blomberg'
    aes(                                         # map variables to aesthetics:
        x = language,                            #   language     to x locations
        y = proportion,                          #   proportion   to y locations
        color = affordance,                      #   affordance   to color
        shape = perspective,                     #   perspective  to shape
        linetype = perspective,                  #   perspective  to linetype
        group = affordance:perspective)) +       #   afford:persp (crossed) to group [4]
    geom_line() +                                # add line geom
    geom_point() +                               # add point geom
    scale_shape_manual(values = c("1", "3")) +   # select custom plotting symbols
    geom_dl(aes(label=affordance),               # add direct labels at right margin
      method="last.points")                      #   using R package 'directlabels'
```

Before we go further, let us briefly reiterate our shorthand labels for variables in a line plot: the Y-variable here is the description of the situation (dynamic vs. static) and the X-variable is language (Thai vs. French vs. Swedish). Figure 3 includes two Z-variables: perspective (1st vs. 3rd person) and affordance for motion (afford vs. non-afford).

4. Design

This section draws attention to design options that aim to optimize the resulting display. While this paper focuses on the use of line plots with discrete variables along the x-axis, most of our advice applies to the use of line plots more generally.

4.1 Plotting symbols and line types

As implemented in Figure 1, filled circles and solid lines are the default choice for simple line plots; they are salient, combine well with error bars, and easily handle photocopy and print reproduction. Subgroups in a display can be distinguished with different symbols and line types. In multivariate settings, care must be taken to ensure discriminability between groupings. To this end, the literature on visual perception offers some insights into the contrastive effect of pre-attentive features.

Let us first discuss line types. In general, dashed lines and dotted lines contrast well with solid segments. For dash patterns to be distinguishable, the dashes should differ by at least 2 to 1 in interval length (Kosslyn 2006: 143). This is illustrated in Figure 4a: In the discriminable set, the black segments of adjacent lines (from top to bottom) differ by a ratio of 2 to 1. [5] As Figure 4b demonstrates, dash patterns are more clearly defined if the segments have sharp boundaries, i.e. square rather than round edges. [6]

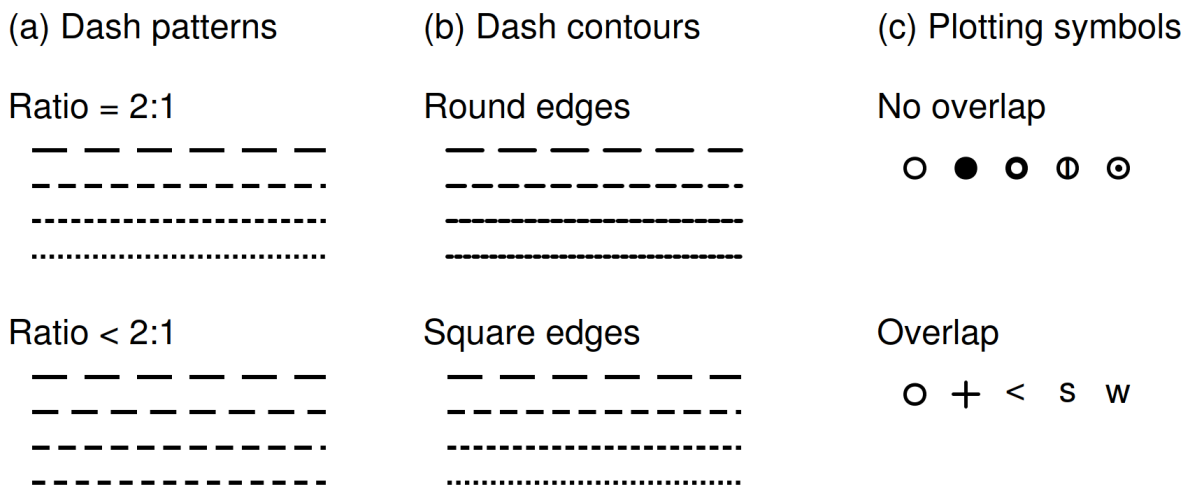

Figure 4. Discriminability: (a) dash patterns, (b) dash contours, and (c) Cleveland's (1994) recommendations for plotting symbols when data points do or do not overlap.

As for plotting symbols, open circles are an ideal companion to filled circles (● ○). Experimental work has shown that human beings discriminate filled and open circles much more accurately than filled circles, triangles and squares (● ▲ ■)

(Chen 1982). A further pair that ensures excellent texture discrimination is open circles and crosses (○ +) (Malik & Perona 1990). Cleveland (1994: 164) recommends the sets shown in Figure 4c. His advice depends on whether overplotting occurs, that is, whether symbols in the graph overlap. [7]

We may also choose meaningful plotting symbols to aid the link-up between visual features and conceptual meaning. Open circles and crosses (○ +), for instance, may serve as indexical symbols, signaling presence/absence of a trait. Effective use can also be made of orthographic symbols, which make it easier to retain group labels in working memory. The viewer then does not have to refer back to the key repeatedly and will parse the graph more quickly. Orthographic symbols must be clearly distinguishable, however. [8]

Recommendations can also be given for the combination of lines and symbols. First, in order for points to remain salient, their diameter should be at least twice the width of the line (Kosslyn 2006: 143). [9] Further, to ensure discriminability between plotting symbols, lines should not obscure the symbols. Two strategies for edge enhancement are (i) haloing (i.e. drawing a white edge around the symbol) and (ii) the use of a layer of white filled circles to interrupt the sequence of line segments (cf. Figure 3). [10]

4.2 Color

The use of color to signal group membership is a key strategy in data visualization, especially since it is processed independently of shape. As illustrated in Figure 3 above, this can be exploited in settings with two or more Z-variables. Thus, we mapped perspective (1st vs. 3rd person) onto the form channel by using different line types and symbols. To distinguish levels of affordance for motion (afford vs. non-afford), we relied on the color channel. This way, the viewer can selectively group elements based on color or form. In general, then, a sensible default is to use variations in form and variations in color to signal different Z-variables.

In many settings we can get by with grey scales. While this may seem outdated, it does yield advantages: (i) grey shades remain distinct under black-and-white reproduction, (ii) they save printing costs, and (iii) color can still be used as a highlighting device (e.g. in presentation slides). [11] The number of categories we can distinguish using gray shades is limited, however, with about 4 values being the maximum (Ware 2013: 75, 122). As Figure 5 illustrates, this value varies with the choice of plotting symbols. If grey serves as a fill color (filled circles on the right hand side), the black contours allow us to exploit the full scale of brightness contrasts, i.e. from black to white. Here, four values appear to maintain acceptable discriminability. If there is no contour, however, the range is limited. The sets

shown on the left-hand side in Figure 5 suggest that three (if not two) shades may be the maximum in this setting. In contrast to filled circles, symbols with thin strokes (w +) and line segments on their own are weaker vehicles for signaling brightness contrasts.

Figure 5. Gray shades for signaling group membership: The number of categories that can be contrasted is limited to 4 (with black contour), or 3 (no contour).

4.3 Direct labeling

Line plots allow for direct encoding of attributes by placing labels inside the display rather than in an external key. Direct labeling accelerates decoding (Milroy & Poulton 1978; Parkin 1983, cited in Pinker 1990: 114) and should therefore be applied whenever feasible. If data labels are placed at the beginning or end of a line, this strengthens the link between label and form due to a visual continuation effect (Gestalt law of good form). If possible, labels should be located in the same part of the display to avoid interference with the data (Kosslyn 2006: 147). If we are free to reorder groups along the horizontal scale, we can choose an arrangement that permits unambiguous labeling. We may also be able to use hierarchical labels (see Figure 3), which is similar to the use of decked column headers in tables. [12]

To illustrate, consider Figure 6, which shows data from a study on noun phrase modification (Biber et al. 2009: 189). Panel (a) resembles the original diagram and depicts change over time in the text frequency of three postmodifying structures. There are two Z-variables: structural type (*of*-phrase, other prepositional phrase, restrictive relative clause) and variety (British vs. American English). The redesign in (b) uses direct (hierarchical) labels, which allows us to proceed to a linguistic interpretation more quickly. [13]

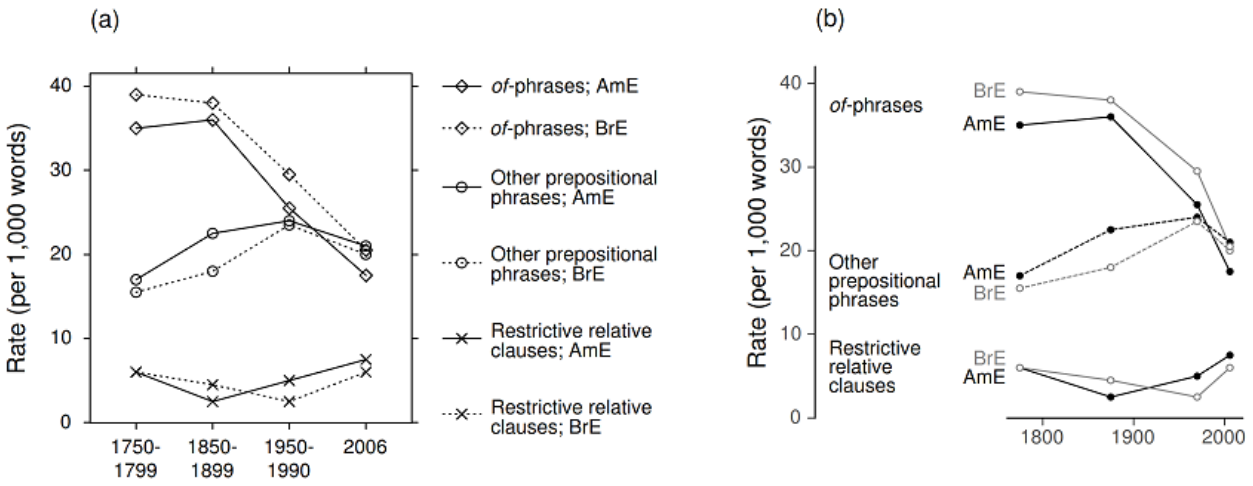

Figure 6. Illustrating of direct and hierarchical labeling: (a) mimics the graph shown in the original publication (Biber et al. 2013: 189), (b) adds direct and hierarchical data labels.

4.4 Error bars

Statistical graphics often include error bars to provide information about statistical variation. These may represent different types of descriptive and inferential estimates (e.g. standard deviations, standard errors, confidence or posterior intervals). Simple line plots with a single predictor on the x-axis can incorporate interval estimates more effectively than bar charts. Minimalism and distinct visual cues ensure that point and interval estimates are visually distinct, which allows the viewer to selectively attend to either type of statistical information. In bar charts, this is more difficult due to the non-distinctiveness of visual elements (right-angled line segments with the same orientation).

These advantages of line plots over bar charts also apply to settings with one or more Z-variables if intervals in the same column (i.e. for the same level of the X-variable) do not overlap. Overlapping error bars impede the ability of the graph to give a clear representation of statistical variation. Various solutions have been put forward. Kosslyn (2006: 146) recommends showing only one arm of each interval. Such one-armed intervals are problematic, however: (i) they fail to indicate the full range of statistical variation, (ii) they do not allow us to assess the extent to which intervals overlap, and (iii) uncertainty intervals for bounded outcome quantities (e.g. proportions, rates, or correlation coefficients) are asymmetric.

A better strategy is to displace the points and intervals horizontally by a small amount. [14] This is illustrated in Figure 7, which shows data from a study on onomasiological variation (Mehl 2019: 70–71). Interest lies in the distribution of three synonymous forms (*inform*, *give information*, *provide information*) across (i)

three varieties of English (Great Britain, Singapore, Hong Kong) and (ii) speech vs. writing. The outcome variable is therefore nominal, and the quantities of interest are the proportions of the variants. [15] Symbols and error bars have been shifted horizontally to avoid overlap.

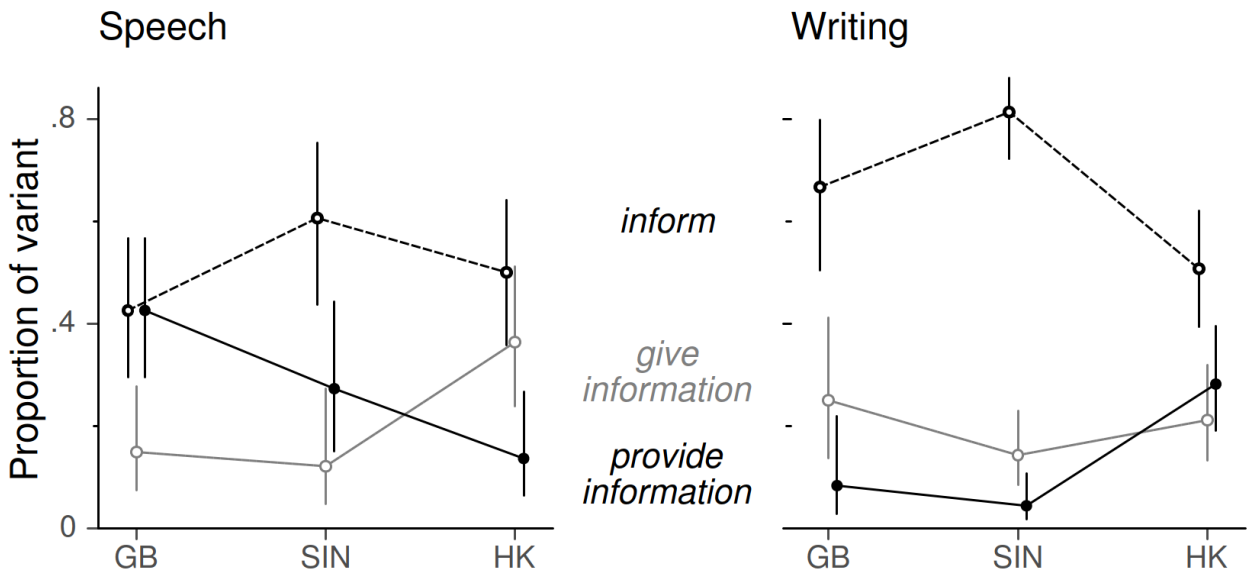

Figure 7. Horizontal displacement to avoid error bar overlap (data from Mehl 2019: 68).

Two words of caution are in order. First, the horizontal displacement should be kept to a minimum, as it distorts the vertical distance between line segments and therefore produces inaccuracies in the perceptual processing of a display. Further, error bars quickly increase the visual clutter in multivariate displays. It may be more appropriate the show (sub)groups in different panels, as was done for speech and writing in the figure above.

4.5 Aspect ratio

Another feature that can be manipulated to facilitate graphical perception is the aspect ratio, the height-to-width relationship of the display. In line plots, changes in the outcome are represented by line orientation (i.e. slopes), with steeper gradients signaling larger differences. The primary perceptual task is therefore the comparison of slopes. Cleveland & McGill (1987) report that comparative judgments of gradients are most accurate when the absolute orientation of lines is 45°, on average. They propose 'banking to 45°' as a design guideline for line graphs: The height-to-width ratio should be adjusted so that the absolute orientation of line segments in the display is centered at about 45°.

Figure 8 compares different aspect ratios. The data are from Crawford's (2009: 265) study on the mandative subjunctive in BrE and AmE. There are certain

lexemes that trigger the mandative subjunctive in a subordinate clause. Figure 8 compares, for a set of 15 verbs, the "triggering strength" in AmE and BrE. The outcome is the proportion of dependent clauses featuring a mandative subjunctive. Grey denotes verbs that are more likely to trigger a mandative subjunctive in BrE. The direct labels for this set are placed in the left part of the display, to quickly reveal inter-varietal differences. The width of the left panel has been chosen to facilitate slope comparisons. In the right panel, it is more difficult to spot verbs with distinct behavior in the standard varieties (e.g. *order*, *determine*).

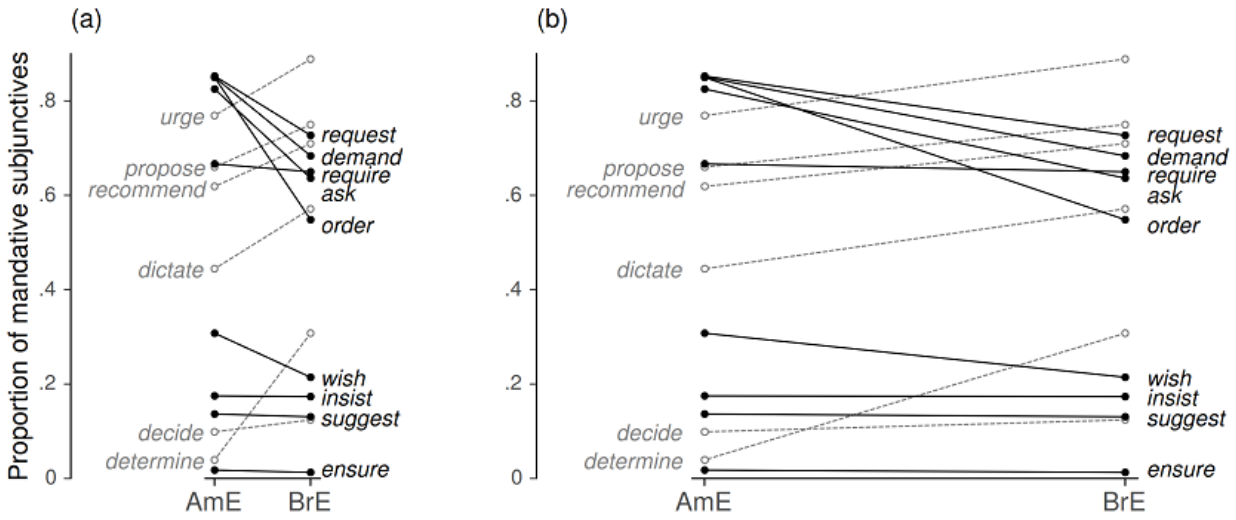

Figure 8. Choice of aspect ratio: Variation among slopes emerges more readily from (a), where the aspect ratio has been set to approximate the recommendations by Cleveland & McGill (1987).

4.6 Display size

Advice about the size of graphical displays applies to data visualization more generally. In visual perception, the concept of the useful field of vision describes the area around the fixation point of our eyes in which stimuli can be parsed without requiring eye movements; in other words: the region (on a sheet of paper or computer screen) we can process (literally) at a glance. Measured in visual angles, the extent of this field has been put at 1 to 4 degrees (Wickens et al. 2013: 51). On a sheet of paper at an average reading distance of 50 cm (20 inches), this is the width of about 1 or 2 thumbs (2–5 cm or 1–2 inches). This diameter minimizes the number of saccadic movements required to scan an image. In compact displays, then, the visual search is more efficient. While this suggests the ideal width and height of a graphical display, it goes without saying that this advice must be applied within reason.

4.7 Arrangement

The effectiveness of a graph depends on how quantitative information is organized in the display. In line plots showing two or more predictor variables, arrangement decisions can be made at different levels. First, we must decide which variable to assign to the x-axis (X-variable). The remaining predictors (Z-variables) are then encoded either in a single display using color, symbols, and/or line types; or, alternatively, by juxtaposing subgroups in side-by-side panels. Finally, the order of categories along the x-axis may be purposefully chosen. We will discuss these options in turn.

X-variables vs. Z-variables

The question of which variable to assign to the horizontal scale may be informed by three considerations (see Keppel & Wickens 2004: 248; Kosslyn 2006: 81). First, and perhaps most importantly, factors denoting (quasi-)temporal information should be assigned to the x-axis. By convention, time series and apparent-time differences (e.g. age groups, developmental stages or pre- and posttest scores) are shown from left to right. Second, the horizontal scale is also preferred for quantitative (perhaps including ordinal) variables. Finally, the horizontal scale should host the most important predictor. These recommendations find support in experimental studies, which suggest that the X-variable plays a primary role in the interpretation of statistical graphs (Shah & Freedman 2011: 570).

To make matters more concrete, let us turn to Figure 9, which shows data from Deshors (2015), a study on the usage of *can* vs. *may* by L1 French and Chinese learners of English. We focus on spoken language for now, and consider two of the factors investigated: voice (active vs. passive) and L1 (Chinese vs. French). The outcome of interest is the proportion of contexts in which *can* was chosen over *may*. In panel (a), we are inclined to infer that the difference between active and passive verb phrases is much greater for French learners. In other words, we conclude that the "effect" of voice (X-variable) varies by L1 (Z-variable).

Now consider panel (b). The immediate message emerging from this arrangement is that the difference between the L1s is more pronounced in passive contexts. In other words, we tend to discern that the "effect" of L1 (X-variable) varies conditional on voice (Z-variable).

While panels (a) and (b) show the same data, they open up different perspectives and direct our attention to different comparisons. This has consequences for how we perceive and interpret the information in the display. The arrangement determines which predictor assumes a primary role and which predictor is understood, at least implicitly, as a moderator (or "effect modifier"). It is the

X-variable that leads the interpretation, as we are invited to interpret that the "effect" of X varies over levels of Z (Shah & Freedman 2011).

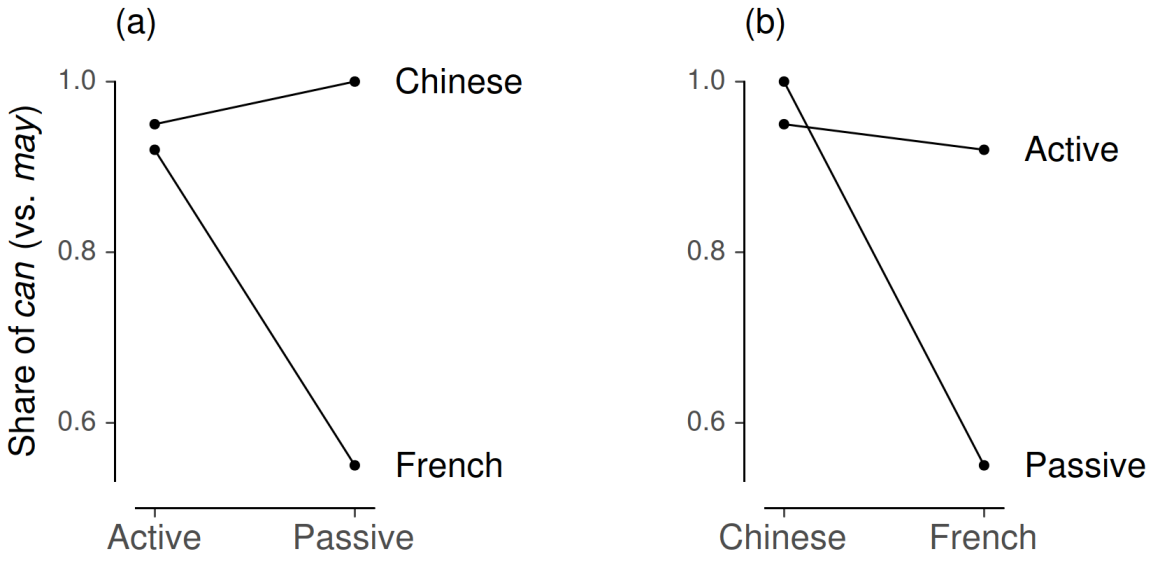

Figure 9. Biased interpretation of line plots: In interaction comparisons, the X-variable assumes the role of the primary predictor, the Z-variable that of the moderator (see text for discussion) (data from Deshors 2015).

In a given setting, it may be the case that only one of these perspectives makes sense or directly addresses a question of linguistic concern. In general, then, care must be taken to ensure that the arrangement matches the intended interpretation. This is because line plots bias the viewer towards one perspective. [16]

Order along the horizontal scale

Once we have decided which variable to assign to the x-axis, we need to arrange categories along the axis. If the categories have no logical order, the following objectives may provide guidance:

- Groups may be ordered according to value or size. This reveals additional information (Wainer 1997: 35) and may yield a regular shape, which facilitates pattern perception due to the Gestalt law of good form.
- Groups may be rearranged to produce the simplest overall pattern, or a pattern that aids the interpretation of the display (Kosslyn 2006: 83).
- The order along the x-axis may be chosen to allow for unambiguous direct labeling.

Superposition vs. juxtaposition

Finally, let us turn to a general strategy for the arrangement of visual information:

the navigation between superposition and juxtaposition. Superposition means that everything is shown in a single display. By contrast, juxtaposition assigns subsets of the data to separate panels (or facets, in the *Grammar of Graphics* terminology). As both strategies have their advantages and drawbacks, they are complementary approaches to the display of multivariate data.

Superposed displays facilitate comparisons between (sub)groups but quickly become cluttered. Juxtaposition (or facetting), on the other hand, strives for clear vision and allows for better comparison within (sub)groups. The use of juxtaposition is illustrated in Figure 10, which re-arranges the information shown in Figure 8 above. The separation of the two groups of verbs into different panels yields a less busy graph and gives a clearer picture of the variation among verbs in each group.

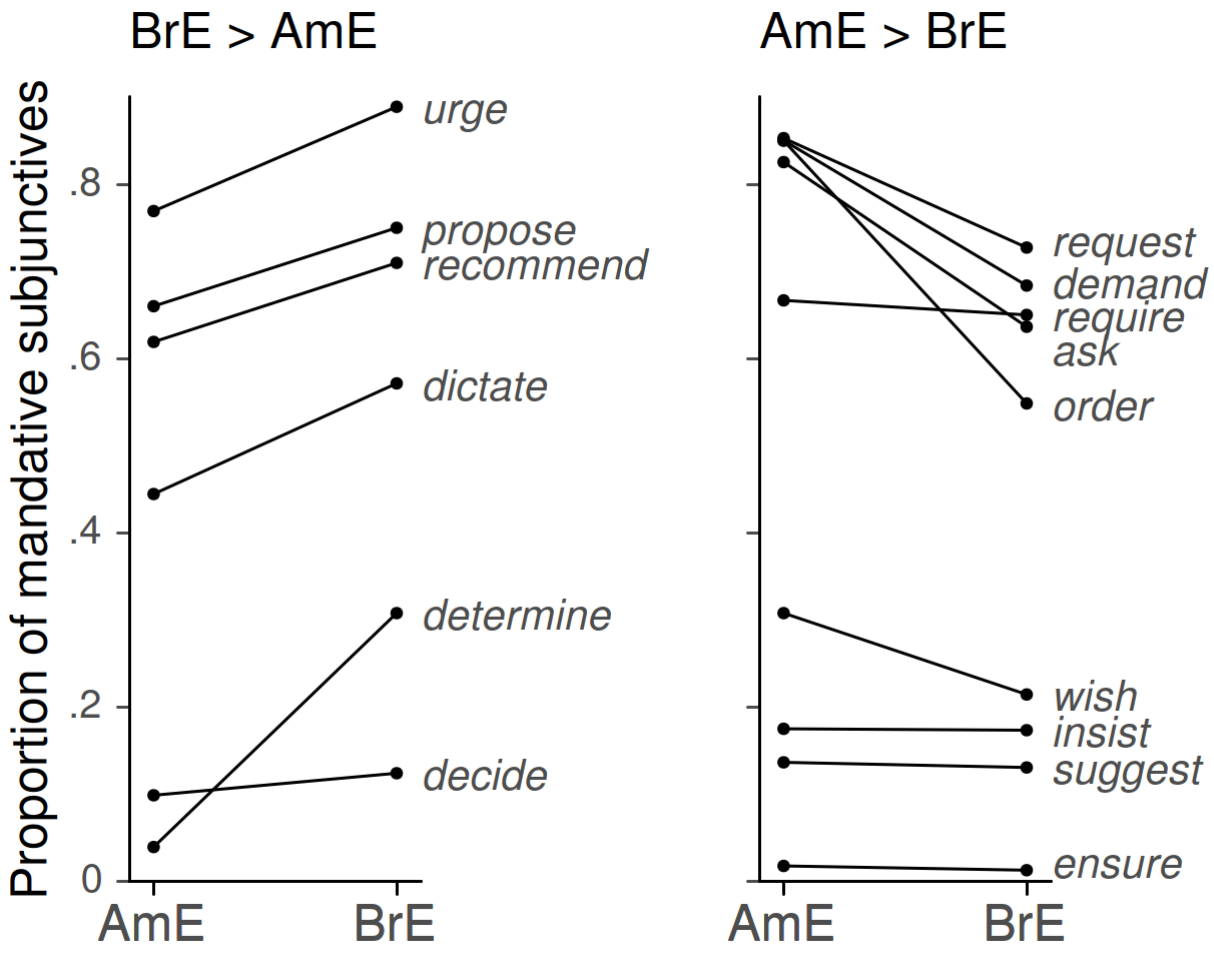

Figure 10. Juxtaposition: In contrast to Figure 8a, the subgroups of verbs are shown in side-by-side panels, yielding a less cluttered graph.

4.8 Scale of measurement of the outcome variable

Before we summarize our advice on the design of line plots, let us pause to

consider the implications of different scales of measurement for visualization tasks. Our focus will be on the outcome variable, which is usually shown on the y-axis. Following Stevens (1951), we will distinguish between binary, nominal, ordinal, and continuous variables. Further attention to continuous traits is given in section 5.3; here, we focus on categorical outcomes.

A key property of binary outcomes, which is of relevance for their visualization, is the fact that they can be represented using a single symbol. This is because the proportion of one category (p) implies the share of the complementary event ($1 - p$). This greatly simplifies the visualization task, since, for a given condition, a single mark gives an exhaustive description of the outcome trait. Judging from the visual treatment of binary outcomes in the literature, however, where binary outcomes are usually graphed using other chart types, it is our impression that some researchers may not be aware of this (cf. Figures 11–14).

A second point that deserves emphasis is the fundamental difference between binary and continuous traits on the one hand, and nominal and ordinal ones on the other. For nominal and ordinal variables, three or more values of the outcome must be graphed. This adds a layer of complexity to our visual representation, since graphical means must be "spent" to represent category membership of the outcome. This effectively constrains the number of predictor variables that can be represented in a display. To the best of our knowledge, the question of whether outcome categories should be encoded on the horizontal axis (X), or encoded via other means (Z) has not been discussed in the literature.

Our recommendations for the design of line plots are summarized in Table 1. We now move on to a discussion of the advantages and opportunities offered by line plots. The suggestions spelled out above will reverberate in our visual illustrations.

Feature	Recommendation
Lines	Dash patterns should differ by at least 2:1; square edges preferred
Plotting symbols	●, ○, + as a set of maximally distinctive forms; meaningful signs such as letters should be employed where possible
Lines and symbols	Symbols should be set off against lines in terms of size and contour
Gray shades	Should be limited to four (when used as fill color) or two (for lines and plotting symbols)
Direct labeling	Should be used to avoid a key; if possible, put labels at the end or beginning of lines and use hierarchical labeling
Error bars	Use horizontal displacement to avoid error bar overlap; show subgroups in different panels to avoid visual clutter
Aspect ratio	Adjust the height/width of the display so that the absolute orientation of line segments averages at about 45°
Display size	Ideally, data points occupy a region of 2 to 5 cm in diameter
Arrangement	Heed conventions when selecting a variable for the *x*-axis; beware of perceptual inclinations for interactions between X- and Z-variables; consider rearranging categories along the *x*-axis; alleviate visual clutter by using juxtaposed panels

Table 1. Summary of design recommendations for line plots.

5. Advantages

Currently, bar charts are the most popular vehicle for corpus data visualization (see Sönning & Schuetzler, this volume), and we will therefore contrast line plots with this graph type. [17] From the viewpoint of scientific data communication, line plots arguably have a number of distinct advantages over bar charts, and we will illustrate these in the following.

5.1 Minimalism

One of the principles of graph design outlined by Tufte (2001) is the avoidance of redundancy. Line plots make effective use of visual elements and avoid superfluous cues. While empirical work has produced no evidence for the superiority of pictorial minimalism (Spence 1990; Gillan & Richman 1994), the elimination of redundant elements yields a less cluttered graph. This can help with the graphical representation of multivariate data patterns.

To illustrate, we turn a bar chart presented in Levin (2009: 67) into a line plot. The study is concerned with British-American differences in the usage of irregular past tense forms (e.g. *learned* vs. *learnt*). Figure 11a shows how the "regularness", i.e. the share of regular *-ed* forms, varies for five verbs in different semantic contexts (durative vs. punctual meaning). In the left hand bar chart, which mimics the original graph, rectangles dominate the display, and plenty of ink is used to show 10 data points. In the line plot, we replace stacked bars with dots (see section 4.8). The *x*-axis hosts the key predictor, semantic context, and verb labels assume the status of a Z-variable. The arrangement guides the interaction comparison: The immediate message is how the "effect" of semantic context varies between verbs. The horizontal order of the categories was chosen to allow for direct labeling in the right-hand part of the display.

The more minimalistic line plot allows us to draw key comparisons with ease: We quickly discern that the iconically motivated cline punctual < durative holds for all verbs except *spill*. Further, the rank-order of verbs emerges readily (*spill* > *burn* > *lean* > *learn* > *leap*). These insights are more difficult to extract from the bar chart.

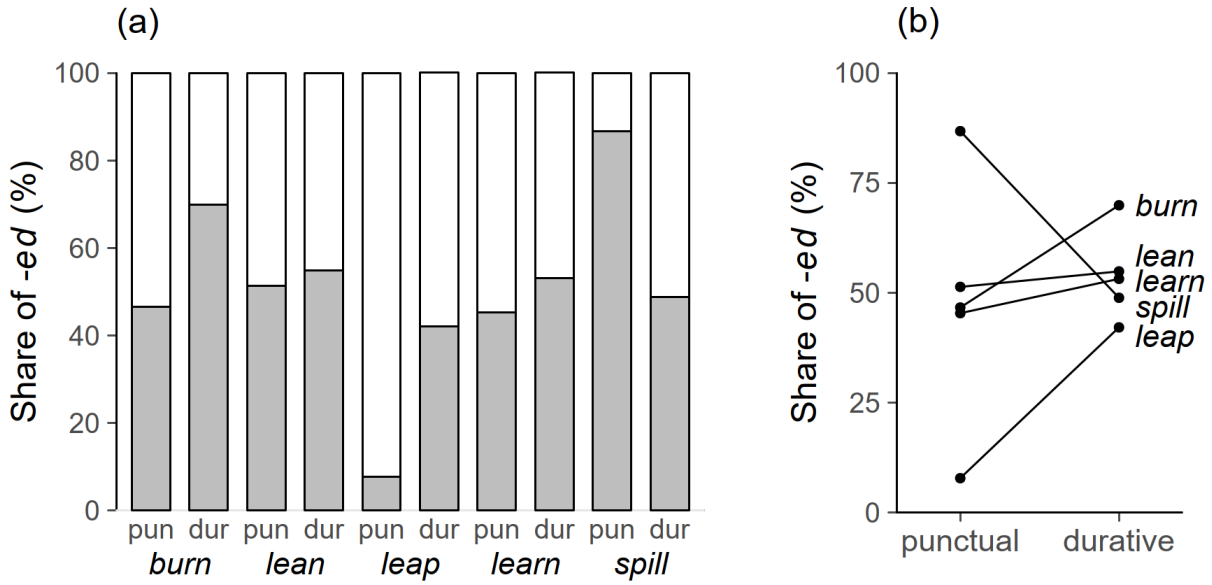

Figure 11. Minimalism: The bar chart shown in Levin (2009: 67) is reworked into a line plot.

5.2 Resolution

In statistical graphics, the term 'resolution' may refer to the ability of a display to reveal variation in a set of data points. Holding constant the size of a graph, its resolution can be enhanced via axis scaling, which includes the choice of limits and the use of non-linear scales (e.g. a logarithmic scale, see Schuetzler, this volume). In contrast to bar charts, line plots permit such scale manipulations. We

will discuss each option in turn.

As for the choice of limits, the question of whether scaling to zero is necessary has been vividly debated in the literature. Excluding zero from the scale is in many (but not all) cases desirable: Zooming in by rescaling enhances the resolution of a display and foregrounds the variability among data points (Tukey 1977: 51; Cleveland 1994: 92; Wainer 1997: 18). It has also been argued that such rescaling is inherently misleading (Huff 1954: 63). While this essentially depends on the type of display chosen (Robbins 2005: 239), this concern cannot be overcome completely. As usual, then, design recommendations are sensitive to the function and context of a display.

If zero is excluded from the outcome scale, however, bar charts should be avoided, as they use position (end of bar) as well as size (length and area) to encode numeric values. Without a baseline of measurement, the size of a bar is meaningless and misleading. Line plots only use position, which allows us to zoom into the interval within which observations vary. However, we may want to add a break between the x- and y-axis to signal that the vertical scale doesn't start at zero (Cobb 1998: 158).

This is illustrated in Figure 12, with data from a study on *g*-dropping (Forrest 2017: 149), i.e. the pronunciation of forms such as *saying* as /ˈseɪ.ɪn/ instead of /ˈseɪ.ɪŋ/. The focus is on how, in four high-frequency verbs, the likelihood of *g*-dropping varies with the place of articulation of the following segment. The quantity of interest ranges between about 45 and 70 percent. The bar chart representation in Figure 12a, which approximates the original version, begins at zero; this impedes the resolution of the graph. Figure 12b zooms into this interval. The categories along the x-axis are ordered by dropping rate (bilabial < vowel < coronal), which aids average comparisons and creates the simplest visual pattern. Conveniently, this also allows us to directly attach verb labels at the right margin.

The line plot arguably gives a clearer picture of similarities and differences among the verbs. It permits quick estimation of the average rank order for verbs (*coming* > *having* > *being*, *getting*) and contexts (bilabial < vowel < coronal). The interaction comparison readily singles out *being* as the deviant type, as it cuts across the consistent cline exhibited by the other verbs. Its exceptional status is foregrounded by making it visually distinct.

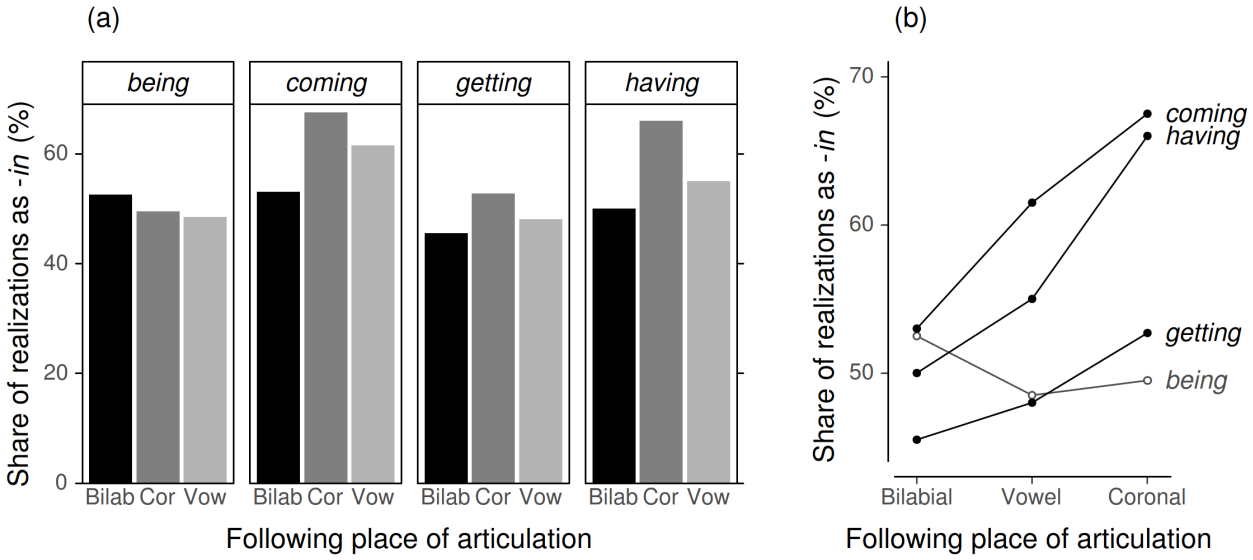

Figure 12. Rescaling by omitting zero from the y-axis: The bar chart presented in Forrest (2017: 149) is recast into a line plot, which zooms into the interval containing the data points.

A further strategy for increasing resolution is the choice of a non-linear scale. Logarithms are a particularly useful tool when the data are skewed towards large values or when relative (i.e. multiplicative) differences are of interest (see Cleveland 1994: 120–126; Schuetzler, this volume). When graphing on a logarithmic scale, we should use dots to encode numerical values, as the length of bars provides meaningless information: a log scale has no logical baseline or origin.

Figure 13 illustrates the added value of non-linear re-scaling. The data are from a study on particle usage by native and non-native speakers of English (Gilquin 2015). The graphs show the occurrence rate (per 10,000 words) of 24 forms in two corpora. The left-hand graph mimics that presented in the paper. The high-frequency forms *out* and *up* reduce the resolution in the lower frequency range. By contrast, the line plot shows log-scaled rates and gives original values as tick mark labels. Note that minimalism works to our advantage here, as Figure 13b is less cluttered. The particles are ordered based on their occurrence rate in native English, which yields a clear reference profile for comparison with the non-native rates. Particles with rates of zero (*aboard*, *under*) were separated using a full scale break (Cleveland 1984; see Schützler, this volume). We discern that, in relative terms, the underrepresentation in non-native English is largest for the particles *round*, *ahead*, *across*, and *in*. Further, the forms *aside*, *by*, *apart*, and *together* stand out as they run counter to the general underrepresentation of particles in non-native speech. These insights are difficult to gain in the grouped bar chart.

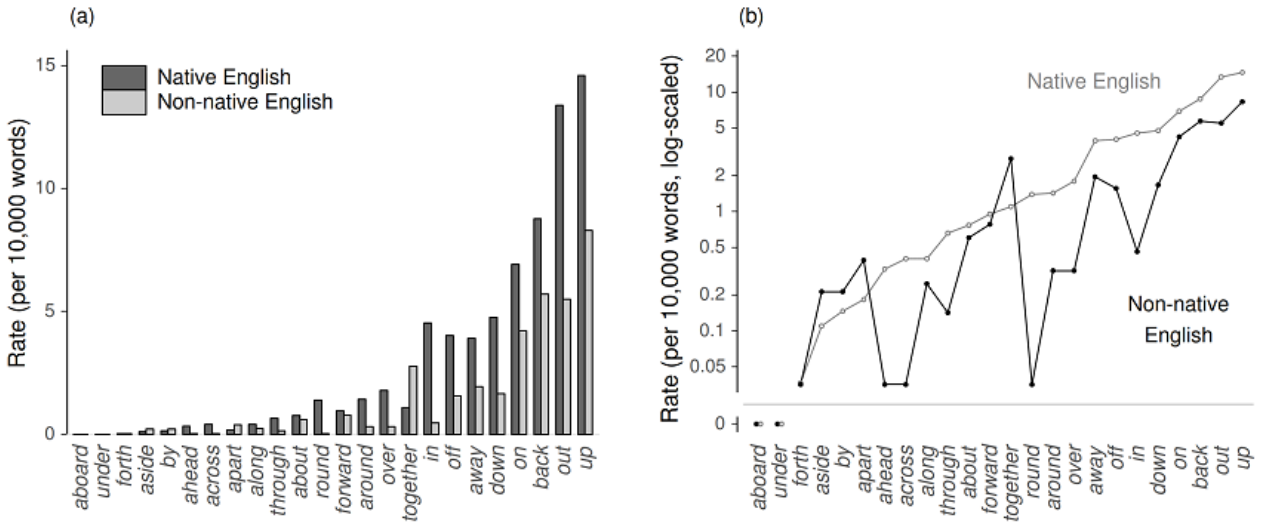

Figure 13. Rescaling using a non-linear outcome scale: The bar chart (Gilquin 2015: 74) is recast into a line plot, which uses log-scaled rates to facilitate comparisons among low-frequency forms.

5.3 Interval scales

Quantitative, or continuous, variables can be divided into interval- and ratio-scaled measures (Stevens 1951). Interval scales allow positive and negative values – examples are difference scores and correlation coefficients. Bar charts are a poor choice for interval-scaled outcomes, especially if a display shows values to both sides of zero. Further, bars with error intervals extending beyond zero arguably look odd. More importantly, however, the lengths of bars encourage the viewer to make ratio comparisons ("A is about twice as large as B"), which may not be warranted on interval scales (e.g. for correlation coefficients). None of these shortcomings apply to line plots.

Figure 14 shows a series of percentage point differences, which express how the usage rate of the modals *shall*, *will*, their short form *'ll*, and BE *going to* developed from the 1960s to the 1990s (Aarts et al. 2013: 38). Positive differences indicate that a form increased over time. We see that *shall* decreased in frequency, by about 60 percentage points, while the contracted form *'ll* showed an increase of about 10 points. The authors use a figure similar to 14a to compare two ways of expressing change. The line plot in Figure 14b merges these into the same panel and adds annotations to clarify the interpretation of points above and below the zero line.

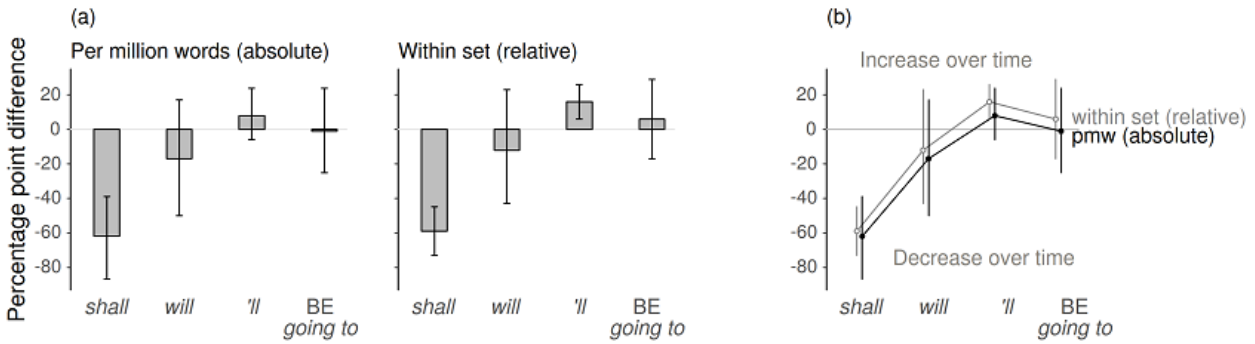

Figure 14. Interval-scaled outcomes: Replacement of bar chart with line plot to show difference estimates (data from Aarts et al. 2013: 38).

5.4 Extended encoding strategies

Graphical displays are limited to the two-dimensional space. Visualizations of three or more variables must therefore resort to visual means other than location in the 2-D plane. In bar charts, attributes are mapped to the area of the rectangles by varying color and/or fill patterns. Line plots offer more diverse means of signaling category membership.

An effective indicator is direct labeling, which was discussed in section 4.3. Two further visual cues are plotting symbols and line types. They can be mapped to different variables, thus opening up further dimensions for data display. Used in tandem, they enhance discriminability via "redundant coding" (Nothelfer et al. 2017; Ware 2013: 159). Figure 15 returns to data on *can* vs. *may* in learner English (Deshors 2015). Above, we concentrated on the factors voice (active vs. passive) and L1 (Chinese vs. French) (see Figure 9). We now add a further dimension of variation: mode of production (speech vs. writing). This leaves us with three predictor variables and a challenging visualization task.

We forego a comparison with the original graph and turn directly to a line plot that exploits four coding strategies: direct (hierarchical) labeling and grey shading, both for French vs. Chinese, and different line types and meaningful plotting symbols, both for speech vs. writing. Let us first examine Figure 15a. There are three average comparison we might be interested in: the average "effect" of voice, L1, and mode. To estimate these, we need to form different sets of visual objects:

- For L1 (French vs. Chinese), we attend to brightness contrasts and group grey and black elements into two distinct objects. The direct labels facilitate this task. We note that the share of *can* for Chinese learners is higher, on average.
- For mode of production (speech vs. writing), the relevant features are line types and plotting symbols. Here, redundant coding aids visual separation and approximation of the average share of *can*. On balance, the share appears

to be slightly higher in speech.

- For the average effect of voice, we approximate two averages, one for the left and one for the right column of points. Overall, the share of *can* appears to be about the same.

These average comparisons illustrate how the choice of surface features generates different proto-objects to allow flexible groupings in a multivariate display. The patterns in Figure 15, however, suggest that averaging over conditions glosses over striking interactions, to which we now turn. Recall that the arrangement of predictors into X- and Z-variables channels our perception and interpretation of data patterns (see section 4.7). We therefore consider two setups: Figure 15a makes voice the primary factor, and Figure 15b assigns this status to mode of production.

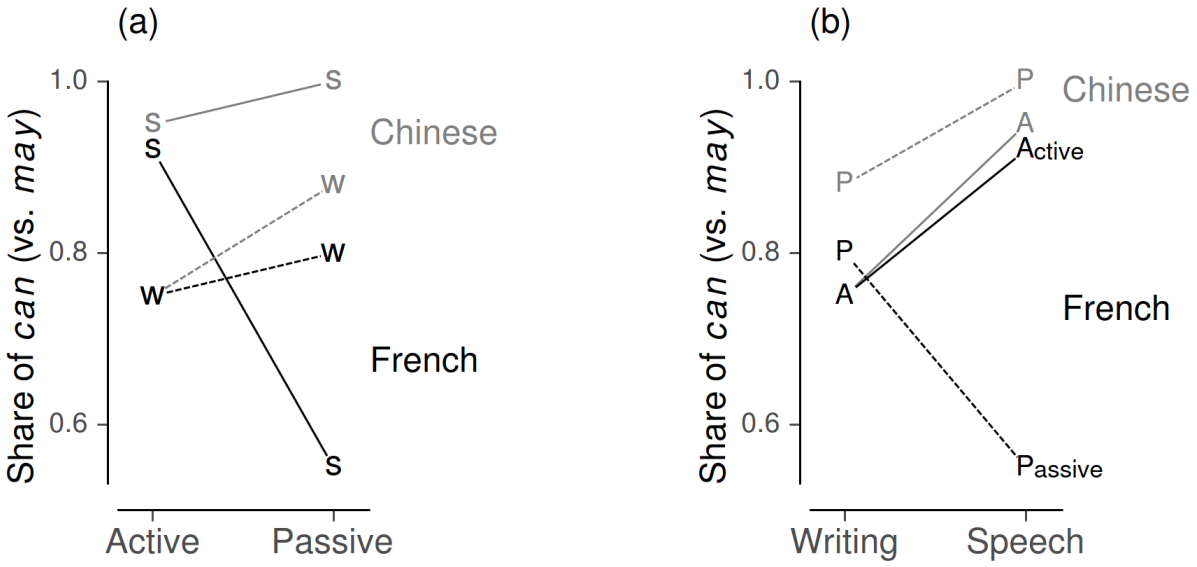

Figure 15. Illustration of extended coding strategies using data from Deshors (2015).

Figure 15a invites us to interpret how the "effect" of voice varies over cross-classifications of L1 and mode of production. In technical terms, this is a three-way interaction. At this level of detail, the main conclusion appears to be that, in writing, we find a consistent pattern in the two L1s, with passive verb phrases showing a slightly higher share of *can*. In speech, Chinese learners lean into the same direction. French learners, however, deviate substantially from this general tendency.

Our reading of Figure 15b proceeds along different lines. Here, slopes express the "effect" of mode of production in four subgroups (cross-classifications of voice and L1). The observation that emerges most readily from this arrangement is that it is passive contexts in French that deviate from the overall trend (i.e. higher rates of *can* in speech vs. writing). The same data therefore generates different insights,

which are informationally equivalent, but cognitively distinct.

5.5 Perceptual grouping laws

In multivariate displays, our ability to draw comparisons depends on how easily we can organize (sub)groups into visual objects. Early-stage visual processing and Gestalt laws of perception determine the attraction and cohesion of chunks. We saw that line plots can draw on a wide range of encoding strategies to effect similarity-based groupings (see section 5.4). They also exploit the law of connectedness by linking data points with lines. Connectedness yields stronger cohesion within subgroups than the law of similarity. As a result, these groups form strong chunks, which allows the viewer to draw comparisons between them with relative ease.

Figure 16 illustrates the added value of line plots in complex settings. The data are from Schuetzler (2018) and show how the usage rate of three concessive subordinators (*although*, *though*, *even though*) varies in speech vs. writing and across 12 varieties of English. Figure 16a sketches the visualization used in the study, a dot plot with mode of production mapped onto plotting symbols, and varieties distributed over panels. Though a sophisticated technique, the dot plot makes it difficult to draw interaction comparisons. Visual queries reflecting such comparisons are, for instance: [18]

- How does the usage rate of the subordinators vary between speech and writing?
- In writing, how does usage rate of the connectives vary between varieties?

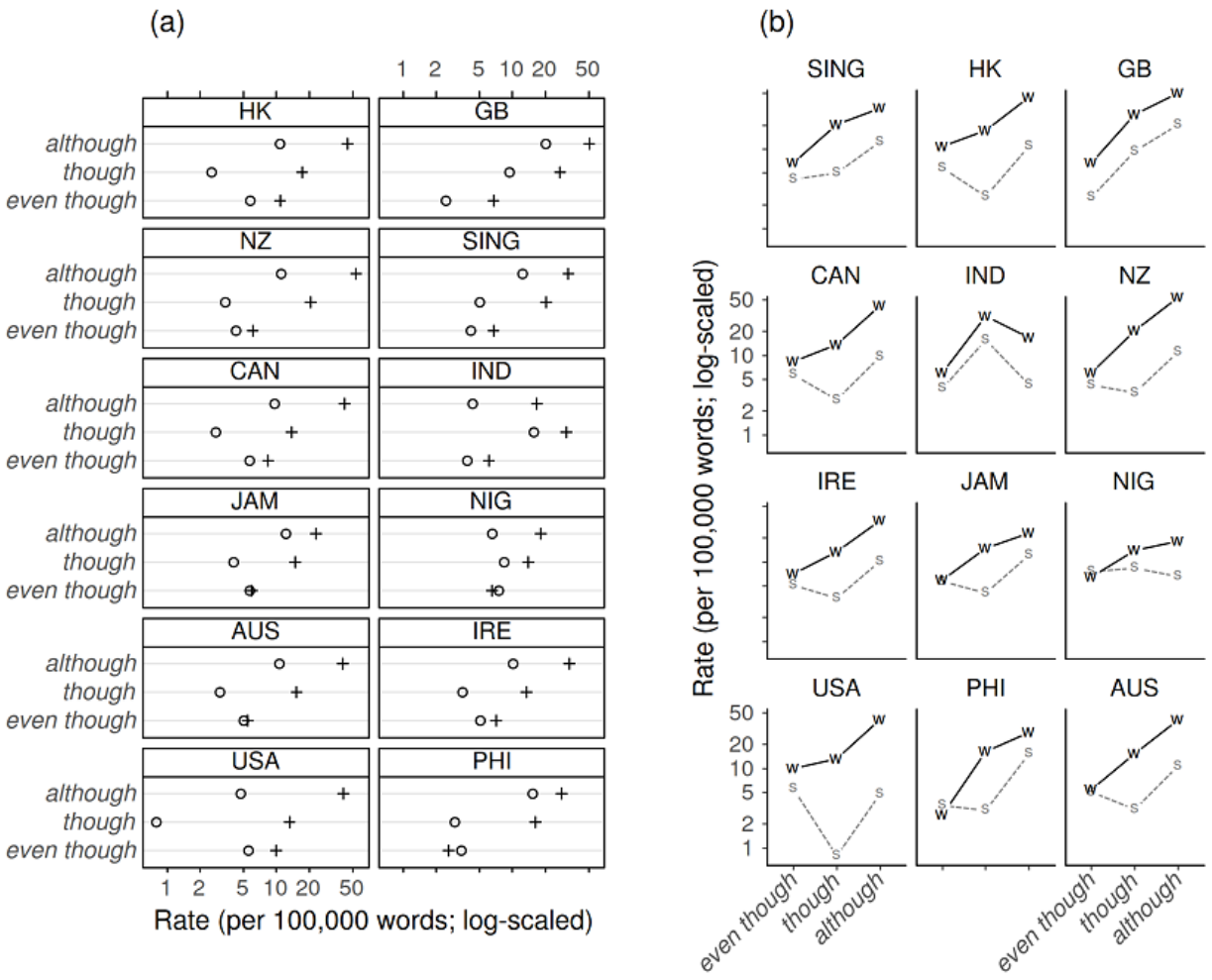

Figure 16. Perceptual grouping laws: While interaction comparisons are difficult to draw in panel (a), connectedness effects in (b) facilitate subgroup comparisons (data from Schuetzler 2018).

Figure 16b casts these data into a line plot. Subgroups based on mode-by-variety cross-classifications form strong chunks, and the resulting profiles, or gestalts, aid interaction comparisons. Thus, in writing, a monotonic increase from left to right seems typical, and we quickly discern that India goes against this trend. The profiles for spoken language show greater variation between varieties, with the majority showing a hockey-stick pattern that reflects a drop in the rate of *though*. A few varieties deviate from this contour, and British English (GB) shows remarkably similar profiles in both modes of production. These insights are difficult to extract from the dot plot in Figure 16a.

We have outlined a number of advantages that line plots bring to data visualization tasks. It is our impression that bar charts can in many cases be constructively replaced by line plots. We should nevertheless be alert to drawbacks of our preferred graph type, which are the topic of the next section.

6. Limitations

Our treatment of line plots for data visualization would be incomplete without a discussion of their limitations. The issues we deal with concern their use with categorical X-variables, the application we have advocated in this chapter.

Experimental studies on graph interpretation have reported that bar charts prompt viewers to describe discrete differences, while line plots encourage trend interpretations (Simcox 1984; Carswell et al. 1993; Zacks & Tversky 1999). To illustrate, we glance back at Figure 9a, where lines represent differences between active and passive verb phrases. Consider a discrete reading: The share of *can* is about .40 (or 40 percentage points) higher in active contexts. Contrast this with a trend interpretation: The more active a context, the higher the share of *can*. For categories that cannot be considered as points along a continuum, this interpretation is nonsensical, of course. Along similar lines, a criticism that is often voiced is that line segments suggest the existence of intermediate values between two categories on the x-axis, which need not exist.

While this criticism is justified, we would argue that, in complex visualization tasks at least, the connectedness of points yields advantages that outweigh these concerns. We should also note that there are categories for which a continuum-interpretation makes sense, e.g. animate vs. inanimate (see Comrie 1989: 185) and ordered categories in general (e.g. discourse accessibility). Finally, it seems that trend interpretations primarily concern two-category settings – based on our intuition, they are less likely for three or more categories. At any rate, these issues may be overcome if this graph type becomes conventionalized for the settings illustrated in this chapter.

A more serious drawback appears to be the unfamiliarity of many viewers with the use of line plots for categorical traits along the x-axis. Multivariate displays in particular may be challenging since viewers may lack experience with the cognitive tasks required to successfully perform visual searches and draw subgroup comparisons. Thus, the purposeful composition of visual objects requires practice. Conceptualizing line-dot profiles metaphorically can help putting complex patterns into words and communicating them to an audience. To describe the reference pattern in Figure 16b, for instance, we used the image of a hockey-stick.

A critical issue, which we have highlighted repeatedly throughout this chapter, is the impact of arrangement decisions, i.e. the assignment of X- or Z-variable status (Shah & Freedman 2011). When processing a display, viewers tend to describe the

relationship as the “effect” of X being moderated by Z. The reverse interpretation (i.e. the “effect” of Z varying over levels of X), appears to be disfavored, if not blocked. In fact, even in simple settings with a binary X- and a binary Z-variable, viewers do not recognize the informational equivalence of these two arrangements (Shah & Carpenter 1995: 45). By the same token, it is difficult to appreciate that in Figures 1, 9 and 15, panels (a) and (b) show the same data.

By contrast, bar charts leave greater interpretational flexibility. If a graphical representation needs to be open to different visual queries and different comparisons, bar charts should be preferred. In settings, however, where we are clear about the status of predictors in interaction comparisons and therefore have a preferred directionality of interpretation, line plots firmly instantiate this perspective. With clear objectives, then, this limitation may also be seen as a strength.

7. Conclusion and outlook

In summary, we have argued for a (more) routine usage of line plots in language data visualization. Using a range of examples from the literature, we demonstrated that this display type may be fruitfully extended to settings where an outcome quantity varies over combinations of categorical traits – a scenario that is typical of (corpus) linguistic research. We discussed a rich set of design options, many of which allow us to exploit principles of human cognition to facilitate graphical communication. In particular, advantage can be taken of Gestalt laws of perceptual organization, to reduce processing and working memory constraints. As a consequence, line plots may outperform bar charts, especially in complex visualization tasks.

These advantages carry a certain cost, however. Thus, viewers may not be familiar with line plots as representations of categorical values along the x-axis, and therefore tempted to interpret discrete attributes as (end)points on a continuous scale. Perhaps most importantly, though, we saw that line plots bias the interpretation of interaction patterns. In multivariate settings with two or more predictor variables, it is therefore advisable to be clear about, and check, which arrangement provides the most informative perspective on the data. In order to make an informed choice, we usually need to try different set-ups.

Given the existing trade-offs between bar charts and line plots, as well as the line plot’s reluctance to flexible interpretation, the goal of producing a static diagram that delivers the intended message and at the same time reaches a broad audience

may be misguided. Thus, our iterative visualization efforts – involving (perhaps many) intermediate versions – may culminate in a display that *we* consider suitable in light of *our* communicative goals and graphical preferences. The viewer, however, may bring to the task different inclinations and queries. More modestly, then, our visual proposal could form the initial set-up of an *interactive* interface. [19] The viewer could then manipulate various features of the display, including (i) the choice of graph type (bar chart or line plot), (ii) the arrangement of variables in the diagram (X- vs. Z-variables), and (iii) the arrangement of variables along the horizontal axis. This would not only allow the user to adjust the display to their preferred configuration, but also provide the opportunity to explore comparisons that were not foregrounded in the initial set-up.

In conclusion, we can state that line plots offer a varied range of opportunities for (corpus) data visualization. Their benefits, however, may not materialize in all settings and for all audiences. We nevertheless believe they deserve to be(come) a core tool in corpus data visualization.

Notes

[1] The terms “multivariate” or “multifactorial” are also used to refer to much more complex data layouts. Here, we use these labels to denote settings with three or more predictor variables, where it is the aim to directly show the data using traditional graphical means that do not require special tools or intermediate dimensionality reduction steps. [Go back up]

[2] Average comparisons are often referred to as “main effects”. [Go back up]

[3] The patterns revealed by the lines in Figure 1 are often referred to as “simple effects”, i.e. the effect of one factor at a particular level of another factor. [Go back up]

[4] In ggplot2, the group aesthetic determines which sets of data points are connected by lines. Here, we want four profiles, one for each combination of perspective and affordance. We therefore cross the two factors using a colon operator, which creates four subgroups. [Go back up]

[5] In R, dash patterns can be defined manually (Murrell 2011: 327) using two numbers: The first sets the length of the black segments, the second that of the spaces. In Figure 4a, from top to bottom: “84”, “42”, “21”, and “11”. In ggplot2, the default line types are well-chosen; custom line types can be specified using an extra line of code, e.g.: `+ scale_linetype_manual(values=c("84", "42"))` [Go

back up]

[6] Square edges are the default in ggplot2. In base R and lattice, however, round edges are the standard. To change this, additional arguments need to be supplied: lend="butt" in base R (see Murrell 2011: 329) and lineend="butt" in lattice. [Go back up]

[7] In lattice, the default set of plotting symbols is geared towards Cleveland's (1994) recommendations. The standard choices in base R and ggplot2, however, are not optimal with regard to discriminability. Custom shapes (e.g. open and filled circles) can be specified: in base R (and lattice), with the argument pch=c(1,19); in ggplot2, with an extra line of code.: `+ scale_shape_manual(values=c(1,19))`. [Go back up]

[8] The arguments pch and values (see previous footnote) also accept orthographic symbols. [Go back up]

[9] When supplying orthographic symbols to scale_shape_manual in ggplot2, we usually need to increase the size of the plotting symbols: `geom_point(size=3)`. [Go back up]

[10] In ggplot2, haloing for plotted text is implemented in the R package shadowtext (Yu 2019). A layer of white filled circles can be added by inserting the following line of code after geom_line() and before geom_point(): `+ geom_point(shape=19,color="white",size=2)`; note that the color should match that of the background. [Go back up]

[11] Some linguistic journals print in black and white and refer to the online PDF for a color version of figures. We have seen examples where grey scales in print are indistinguishable, which renders the graph useless. [Go back up]

[12] In R, direct labeling can be implemented in ggplot2- and lattice-based displays with the package directlabels (Hocking 2021). [Go back up]

[13] Note that time is shown on a continuous scale: Periods have been replaced by their midpoint (i.e. 1775 for "1750-1799") and the estimate for 2006 moves to its proper location relative to the other time points. [Go back up]

[14] For instructions on how to produce line plots with error bars and horizontal displacement in ggplot2, see Chang (2019: Section 7.7): https://r-graphics.org /recipe-annotate-error-bar [Go back up]

[15] In Mehl (2019), there is a mismatch between the information given in Tables 3

and 4 and the corresponding Figures 3 and 4. Here, we have used the counts listed in the tables. [Go back up]

[16] In ggplot2, the arrangement of variables in the display is determined by the mapping: The X-variable is mapped to x-location, and Z-variables are mapped to the aesthetics of points and lines, i.e. color, line type and point shape. [Go back up]

[17] Bar charts, in turn, are usually superior to most other types (e.g. pie charts and mosaic charts). [Go back up]

[18] See Sönning (2016), a plea for dot plots similar in structure to the present chapter. While the dot plot is a useful tool, it often fails in the face of complexity, i.e. data structures with two or more predictor variables. [Go back up]

[19] We thank an anonymous reviewer for pointing us to the added value of interactive applications. [Go back up]

References

Aarts, Bas, Joanne Close & Sean Wallis. 2013. “Choices over time: Methodological issues in investigating current change”. *The Verb Phrase in English: Investigating Recent Language Change with Corpora*, ed. by Bas Aarts, Joanne Close, Geoffrey Leech & Sean Wallis, 14–45. Cambridge: Cambridge University Press.

Biber, Douglas, Jack Grieve & Gina Iberri-Shea. 2009. “Noun phrase modification”. *One Language, Two Grammars? Differences Between British and American English*, ed. by Günter Rohdenburg & Julia Schlüter, 182–193. Cambridge: Cambridge University Press.

Blomberg, Johan. 2015. “The expression of non-actual motion in Swedish, French and Thai”. *Cognitive Linguistics* 24(6): 657–696. doi:10.1515/cog-2015-0025

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. “Predicting the dative alternation”. *Cognitive Foundations of Interpretation*, ed. by Gerlof Bouma, Irene Kraemer & Joost Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Carswell, C. Melody, Cathy Emery & Andrea M. Lonon. 1993. “Stimulus complexity and information integration in the spontaneous interpretation of line graphs”. *Applied Cognitive Psychology* 7: 341–357.

Chang, Winston. 2019. *R Graphics Cookbook: Practical Recipes for Visualizing Data*.

Sebastopol, CA: O'Reilly.

Chen, Lin. 1982. "Topological structure in visual perception". *Science* 218: 699–700.

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

Cleveland, William S. 1994. *The Elements of Graphing Data*. Summit: Hobart Press.

Cleveland, William S. & Robert McGill. 1987. "Graphical perception: The visual decoding of quantitative information on graphical displays of data". *Journal of the Royal Statistical Society: Series* A 150(3): 192–229.

Cobb, George W. 1998. *Introduction to Design and Analysis of Experiments*. New York: Springer.

Comrie, Bernard. 1989. *Language Universals and Linguistic Typology*. Chicago: University of Chicago Press.

Crawford, William J. 2009. "The mandative subjunctive". *One Language, Two Grammars? Differences Between British and American English*, ed. by Günter Rohdenburg & Julia Schlüter, 60–85. Cambridge: Cambridge University Press.

Deshors, Sandra C. 2015. "A multifactorial approach to linguistic structure in L2 spoken and written registers". *Corpus Linguistics and Linguistic Theory* 11(1): 19–50.

Forrest, Jon. 2017. "The dynamic interaction between lexical and contextual frequency: A case study of (ING)". *Language Variation and Change* 29: 129–156. doi:10.1017/S0954394517000072

Gelman, Andrew. 2009. "A diagram of graphs". Blog post, Statistical Modeling, Causal Inference, and Social Science. https://statmodeling.stat.columbia.edu/2009/01/26/a_diagram_of_gr/

Gillan, Douglas J. & Edward H. Richman. 1994. "Minimalism and the syntax of graphs". *Human Factors* 36(4): 619–644.

Gilquin, Gaëtanelle. 2015. "The use of phrasal verbs by French-speaking EFL learners: A constructional and collostructional corpus-based approach". *Corpus Linguistics and Linguistic Theory* 11(1): 51–88.

Hocking, Toby D. 2021. R package 'directlabels': Direct Labels for Multicolor Plots. R package version 2021.1.13. https://cran.r-project.org/web/packages

/directlabels/index.html

Huff, Darrell. 1954. *How to Lie with Statistics*. New York: W. W. Norton.

Keppel, Geoffrey & Thomas D. Wickens. 2004. *Design and Analysis: A Researcher's Handbook*. Upper Saddle River, NJ: Pearson.

Kosslyn, Stephen M. 2006. *Graph Design for the Eye and Mind*. Oxford: Oxford University Press.

Levin, Magnus. 2009. The formation of the preterite and the past participle. *One Language, Two Grammars? Differences Between British and American English*, ed. by Günter Rohdenburg & Julia Schlüter, 257–276. Cambridge: Cambridge University Press.

Malik, Jitendra & Pietro Perona. 1990. "Preattentive texture discrimination with early vision mechanisms". *Journal of the Optical Society of America* A 7: 923–932.

Mehl, Seth. 2019. "Light verb semantics in the International Corpus of English: Onomasiological variation, identity evidence and degrees of lightness". *English Language and Linguistics* 23(1): 55–80.

Milroy, Robert & E. Christopher Poulton. 1978. "Labelling graphs for improved reading speed". *Ergonomics* 21: 55–61.

Murrell, Paul. 2011. *R Graphics*. Boca Raton: CRC Press.

Nothelfer, Christie, Michael L. Gleicher & Steven Franconeri. 2017. "Redundant encoding strengthens segmentation and grouping in visual displays of data". *Journal of Experimental Psychology: Human Perception & Performance* 43(9): 1667–1676. doi:10.1037/xhp0000314

Palmer, Stephen E. & Irvin Rock. 1994. "Rethinking perceptual organization: The role of uniform connectedness". *Psychonomic Bulletin and Review* 1: 29–55.

Pinker, Steven. 1990. "A theory of graph comprehension". *Artificial Intelligence and the Future of Testing*, ed. by Roy Freedle, 73–126. Hillsdale: Erlbaum.

Playfair, William. 1786. *The Commercial and Political Atlas*. London: Corry.

Rensink, Ronald A. 2000. "The dynamic representation of scenes". *Visual Cognition* 7(1–3): 17–42.

Robbins, Naomi B. 2005. *Creating More Effective Graphs*. Hoboken: Wiley.

Schuetzler, Ole. 2018. *Concessive Constructions in Varieties of English*. Post-doctoral thesis, University of Bamberg.

Shah, Priti & Patricia A. Carpenter. 1995. “Conceptual limitations in comprehending line graphs”. *Journal of Experimental Psychology: General* 124(1): 43–61.

Shah, Priti & Eric G. Freedman. 2011. “Bar and line graph comprehension: An interaction of top-down and bottom-up processes”. *Topics in Cognitive Science* 3(3): 560–578. doi:10.1111/j.1756-8765.2009.01066.x

Simcox, William A. 1984. “A method for pragmatic communication in graphic displays”. *Human Factors* 26: 483–487.

Spence, Ian. 1990. “Visual psychophysics of simple graphical elements”. *Journal of Experimental Psychology: Human Perception and Performance* 16: 683–692.

Stevens, Stanley S. 1951. “Mathematics, measurement and psychophysics”. *Handbook of Experimental Psychology*, ed. by Stanley S. Stevens, 1–49. New York: Wiley.

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://fis.uni-bamberg.de/bitstream/uniba/41360/1/BABEL15WernerMaLTopusse_A3a.pdf

Tufte, Edward R. 1990. *Envisioning Information*. Cheshire: Graphics Press.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Cheshire: Graphics Press.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Wagemans, Johan, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh & Rüdiger von der Heydt. 2012. “A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization”. *Psychological Bulletin* 138(6): 1172–1217. doi:10.1037/a0029333

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

Ware, Colin. 2013. *Information Visualization: Perception for Design*. Amsterdam: Elsevier.

Wertheimer, Max. 1938. “Laws of organization in perceptual forms”. A *Source Book of Gestalt Psychology*, ed. by Willis D. Ellis, 71–88. London: Routledge & Kegan Paul.

Wickens, Christopher D., Justin G. Hollands, Simon Banbury & Raja Parasuraman. 2013. *Engineering Psychology and Human Performance*. New York: Routledge.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wilkinson, Leland. 2005. *The Grammar of Graphics*. New York: Springer.

Yu, Guangchuang. 2019. R package ‘shadowtext’. R package version 0.0.7. https://cran.r-project.org/web/packages/shadowtext/index.html

Zacks, Jeff & Barbara Tversky. 1999. ”Bars and lines: A study of graphic communication“. *Memory & Cognition* 27(6): 1073–1079.

Studies in Variation, Contacts and Change in English 22: Data Visualization in Corpus Linguistics: Critical Reflections and Future Directions

Last updated 2023-12-07 by Joe McVeigh