Interactive Techniques for Visualising Categorical Data in Linguistics

David Trye, PhD Candidate in Computer Science
Supervised by Mark Apperley, David Bainbridge & Andreea Calude
Presentation Aims

1. To introduce a novel visualisation tool called **Staircase Plots**
2. To encourage you to use this tool in your own analyses
Motivation

• Categorical data are prevalent in linguistics
  • The most common type of data in corpus linguistics (Stefanowitsch, 2020: 177)
  • Phonological, lexical, grammatical features (among others!)

• 192 WALS features (wals.info) with 2-28 categories
  • Rhythm Type (17A) has 5 categories, 323 items (languages)

<table>
<thead>
<tr>
<th>Value</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trochaic: left-hand syllable in the foot is strong</td>
<td>153</td>
</tr>
<tr>
<td>Iambic: right-hand syllable in the foot is strong</td>
<td>31</td>
</tr>
<tr>
<td>Dual: system has both trochaic and iambic feet</td>
<td>4</td>
</tr>
<tr>
<td>Undetermined: no clear foot type</td>
<td>37</td>
</tr>
<tr>
<td>Absent: no rhythmic stress</td>
<td>98</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>323</strong></td>
</tr>
</tbody>
</table>
Motivation

- **Visualisation** can enhance linguistic analysis
  - Sanity checks
  - Anomaly detection
  - Knowledge discovery
  - Hypothesis testing
  - Statistical modelling
  - Presentation of results

- Few visualisation techniques effectively support **3+ categorical variables**
  - Limited scalability and interaction
  - Lack of user-friendly (no-code) tools available

Insights that might otherwise be missed!
Existing Techniques

- Dimensional Stacking
- Correspondence Analysis
- Generalised Pairs Plots
- Categorical Treemaps
- Conditional Inference Trees
- Contingency Wheel++
- Table Lens
- Mosaic Plots
- Parallel Sets
- GPLOM
- Balloon Plots
- Faceted Bar Charts
- Multidimensional Scaling
Disclaimer

• Staircase Plots are currently under development
  • Design aspects are subject to change
  • Not available until next year
Dataset 1: Hybrid Hashtags

- 80 hashtags (rows) x 4 categorical variables (columns)
- Small **sample size**

<table>
<thead>
<tr>
<th>Hashtag</th>
<th>Words (3)</th>
<th>Class (8)</th>
<th>Semantic Domain (6)</th>
<th>Head (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#proudkiwi</td>
<td>2</td>
<td>CNP</td>
<td>Sport</td>
<td>Māori</td>
</tr>
<tr>
<td>#AotearoaNZ</td>
<td>3</td>
<td>PNP</td>
<td>NZ Identity</td>
<td>NA</td>
</tr>
<tr>
<td>#maoripride</td>
<td>2</td>
<td>CNP</td>
<td>Māori Culture</td>
<td>English</td>
</tr>
</tbody>
</table>

Source: Trye et al. (2020)
The Basics
Why this approach?

• “A good **starting point** for any data exploration is a simple summary table” (Brezina, 2018: 108)
  • An even better starting point is a heatmap!
• “It is always useful to do **cross-tabulation** of all categorical predictors and the response before beginning your analysis in order to detect configurations with zero frequencies or a large number of cells with **very low frequencies**” (Levshina, 2015: 273)
Removing Text Labels

- Easier to perceive general patterns
- Interactive tooltips reveal details on demand
Proportions

- Cells show joint probability, $P(X \cap Y)$, where $X$ and $Y$ are the categories on each axis.
Pearson Residuals

- Non-grey cells (>|2|) correspond to sig. residuals when $\alpha \approx 0.05$
- Formula: $r_{ij} = (O_{ij} - E_{ij}) / \sqrt{E_{ij}}$

Visual Properties:
- Variable 1: Pearson Residuals
- Variable 2: None
- Text: Pearson Residuals
- Tooltip: Customise

Chi-Squared Test:
- Significance: 0.01, 0.05, 0.1
- Level: Panels, Cells

Legend:
- <-2
- >2

Table:

<table>
<thead>
<tr>
<th>Class</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADJP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADVP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clause</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMLA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PNP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VP</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>2</th>
<th>3</th>
<th>4+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Māori</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>Flora and Fauna</th>
<th>Generic</th>
<th>Humour</th>
<th>Māori Culture</th>
<th>NZ Identity</th>
<th>Sport</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.5 -0.4 0.7 0.1 -0.7 2.0 -0.7 -0.1</td>
<td>-0.3 -0.2 -0.4 -0.1 2.5 -0.2 -0.4 -0.6</td>
<td>-0.5 -0.4 -0.7 0.4 -0.6 -0.4 -0.7 1.0</td>
<td>0.5 -0.7 -1.1 -0.4 -1.0 -0.7 1.5 1.3</td>
<td>0.0 1.6 0.6 0.2 0.9 -0.8 -0.1 -1.7</td>
<td>0.3 -0.7 0.4 -0.2 -0.2 0.7 -0.4 0.4</td>
</tr>
</tbody>
</table>
Pearson Residuals

- Non-grey cells (>|2|) correspond to sig. residuals when $\alpha \approx 0.05$
- Formula: $r_{ij} = (O_{ij} - E_{ij}) / \sqrt{E_{ij}}$
Bivariate Colour Scheme

- Show both counts (intensity) and residuals (colour)
Bivariate Colour Scheme

- Show both counts (intensity) and residuals (colour)
Chi-Squared Test

• Staircase Plots provide built-in support for the chi-squared test of independence
  • Used to determine whether there is an **association** between two categorical variables

• Ability to calculate & display results for all pairs of variables that satisfy the basic **test conditions**
  • Panels coloured according to strength of association
  • **Effect size** measured using Cramer’s V

• Advantages:
  • Removes burden of manual computation
  • Visually reinforces correct interpretation
  • All results conveniently displayed in one place
Chi-Squared Test Conditions

1. Nominal (preferred) or ordinal variables
   • Quantitative variables can be binned

2. Independent observations
   • Requires manual verification

3. Mutually-exclusive categories
   • Each observation contributes to one cell per panel

4. Expected frequency >1 in all cells and >5 in at least 80% of cells
   • Requires decent sample size
   • Typically at least 5x number of cells

\[
\begin{array}{ccc}
A & X & \text{Row} \\
C & \text{-} & \text{-} \\
D & \text{-} & \text{-} \\
\text{Col} & \text{-} & N \\
\end{array}
\]
Chi-Squared Test

- **Insufficient sample size** for this dataset!
- No pairings meet the expected frequency criterion
Chi-Squared Test

- Example of a larger dataset (N= 2,201)
- Each panel reports the test statistic, (degrees of freedom), p-value & Cramer’s V

<table>
<thead>
<tr>
<th>Class</th>
<th>Adult</th>
<th>Child</th>
<th>Age</th>
<th>Sex</th>
<th>Survived</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult</td>
<td>5.59 (3)</td>
<td>p &lt; 0.001, $\phi_c = 0.139$</td>
<td>27.12 (1)</td>
<td>p &lt; 0.001, $\phi_c = 0.111$</td>
<td>456.87 (1)</td>
</tr>
<tr>
<td>Child</td>
<td>349.91 (3)</td>
<td>p &lt; 0.001, $\phi_c = 0.399$</td>
<td>20.96 (1)</td>
<td>p &lt; 0.001, $\phi_c = 0.098$</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dataset 2: Covid Directives (Burnette & Calude, 2022)

- 754 directives (rows) from tweets containing #covid19nz
- 10 variables (columns)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stance (5)</td>
<td>against, pro, for stronger measures, neutral, unclear</td>
</tr>
<tr>
<td>Force (7)</td>
<td>advice, criticism, indirect, offer, plea, prototypical, well wishers</td>
</tr>
<tr>
<td>Politeness (4)</td>
<td>no redress, on record negative, on record positive, off record</td>
</tr>
<tr>
<td>Verb (4)</td>
<td>let, main verb, modal, no</td>
</tr>
<tr>
<td>Clause (3)</td>
<td>declarative, imperative, interrogative</td>
</tr>
<tr>
<td>Addressees (2)</td>
<td>explicit, implicit</td>
</tr>
<tr>
<td>Hashtags (2)</td>
<td>none, yes</td>
</tr>
<tr>
<td>Loanwords (2)</td>
<td>none, yes</td>
</tr>
<tr>
<td>Subjects (2)</td>
<td>individuated, non-individuated</td>
</tr>
<tr>
<td>Vocative (2)</td>
<td>none, yes</td>
</tr>
</tbody>
</table>
Key Limitations

- Inner variables are **split** across columns and rows
  - Displaying only half the matrix saves space but makes comparison with other variables difficult
- Layout restricts total **number of categories** that can be displayed
  - Don’t want multiple variables with 10+ categories
  - Exact limit varies according to screen resolution
- **Loss of precision** when using bivariate colour maps
  - Fewer distinct shades for each variable
- Not optimised for **ordinal data**
  - Chi-squared test doesn’t consider ordering information
Interactive Features – Coming Soon!

- Display selected items in **scrollable table**
- **Associative highlighting** for categories (rows/columns) & variables (related panels)
  - Related: search feature
- Flexible **re-ordering** of categories & variables
  - Alphabetically, by frequency/cardinality, manually via drag-and-drop
- Basic **data transformations**
  - Collapse/expand existing categories
  - Add/remove variables
  - Filter by selection
Help needed!

• Online participants needed for a **user study** about Staircase Plots
  • ~1 hour in Feb/March 2023
• Please fill out this quick Google Form (name + email) if you might be interested in taking part
• Thank you!

https://forms.gle/evL7j3jed8ZgfgVM6
Contact me
David Trye
dgt12@students.waikato.ac.nz
(Or talk to me on Stream 1 during the breaks!)
References (1)

References (2)


