Text as Data

An introduction to quantitative text analysis and reproducible research with R

Jerid Francom

December 13, 2021 (latest version)

Contents

Welcome

INCOMPLETE DRAFT

This is the coursebook to accompany Linguistics 380 “Language Use and Technology” at Wake Forest University. The working title for this coursebook is *Text as Data: An Introduction to Quantitative Text Analysis and Reproducible Research with R*.

The content is currently under development. Feedback is welcome and can be provided through the hypothes.is[1] service. A toolbar interface to this service is located on the right sidebar. To register for a free account and join the “text_as_data” annotation group follow this link[2]. Suggestions and changes that are incorporated will be acknowledged.

Author

Dr. Jerid Francom is Associate Professor of Spanish and Linguistics at Wake Forest University. His research interests are focused around quantitative approaches to language variation.

License

Credits

Icons made from Icon Fonts are licensed by CC BY 3.0

Acknowledgements

TAD has been reviewed by and suggestions and changes incorporated based on the feedback through the TAD Hypothes.is group[4] by the following people: …

Preface

DRAFT

[1]https://web.hypothes.is/
[2]https://hypothes.is/groups/WkoaXnBX/text-as-data
[3]https://francojc.github.io/
[4]https://hypothes.is/groups/Q3o92MJg/tad

The journey of a thousand miles begins with one step.

— Lao Tzu[5]

The essential questions for this chapter are:

- What is the rationale for creating a coursebook on quantitative text analysis?
- What is the approach taken in this coursebook?
- What are the learning goals and how does the coursebook aim to support attaining these goals?

This chapter aims to provide a brief summary of current research trends that form the context for the rationale for this coursebook It also provides instructors and students an overview of the purpose and approach of the coursebook It will also include a description of the main components of each section and chapter and provide a guide to conventions used in the book and resources available.

Rationale

In recent years there has been a growing buzz around the term 'Data Science' and related terms; data analytics, data mining, *etc.* In a nutshell data science is the process by which an investigator leverages statistical methods and computational power to uncover insight from large datasets. Driven in large part by the increase in computing power available to the average individual and the increasing amount of electronic data that is now available through the internet, interest in data science has expanded to virtually all fields in academia and areas in the public sector. Data scientists are in high demand and this trend is expected to continue into the foreseeable future.

This coursebook is an introduction to the fundamental concepts and practical programming skills from Data Science that are increasingly employed in a variety of language-centered fields and sub-fields applied to the task of text analysis. It is geared towards advanced undergraduates and graduate students of linguistics and related fields. As quantitative research skills are quickly becoming a core aspect of many language programs, this coursebook aims to provide a fundamental understanding of theoretical concepts, programming skills, and statistical methods for doing quantitative text analysis.

Approach

Many textbooks on doing 'Data Science', even those that have a domain-centric approach, such as text analysis, tend to focus on the 'tidy' approach, seen in Figure **??** from **?**.

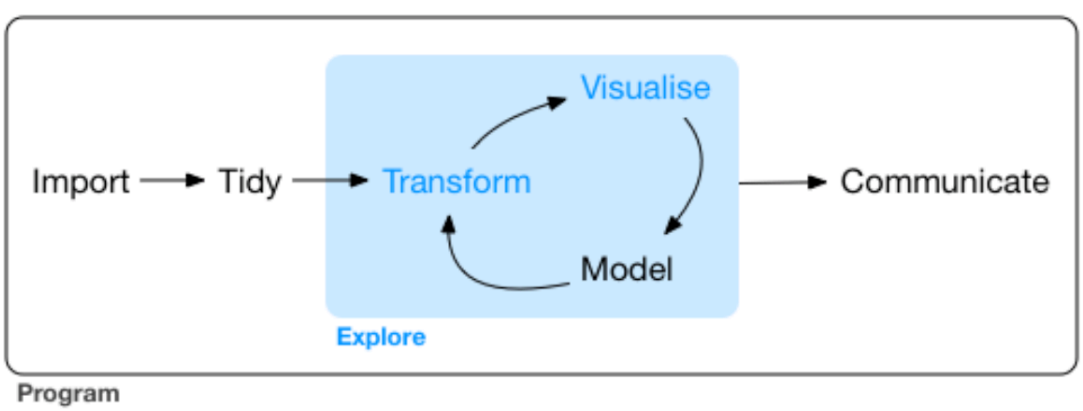

Figure 1: Workflow diagram from R for Data Science.

[5] https://en.wikipedia.org/wiki/Laozi

However these resources tend to underrepresent the importance of establishing a research question and implementation plan. A big part, or perhaps the biggest part of doing quantitative research, and research in general is what is the question to be addressed. I think a central advantage to this coursebook for language researchers is to thread the project goals from a conceptual point of view without technical implementation in mind first. Then, after establishing a viable vision about what the data should look like, how it should be analyzed, and how the analysis will contribute to knowledge in the field, we can move towards implementing these preliminary formulations in R code in a reproducible fashion. In essence this approach reflects the classic separation between content and format[6] –the content of our research should precede the format it should or will take.

Learning goals

This course you will:

Data Literacy (DL): learn to interpret, assess, and contextualize findings based on data.

1. ability to understand and apply data analysis to derive insight from data
2. ability to understand and apply data knowledge and skills across linguistic and language-related disciplines

Research Skills (RS): learn to conduct original research (design, implementation, interpretation, and communication).

1. identify an applicable area of investigation in a linguistic or language-related field
2. develop a viable research question or hypothesis
3. assess, acquire, and document data
4. curate and transform data for analysis
5. select and apply relevant analysis method
6. interpret and communicate findings

Programming Skills (PS): learn to produce your own research and work collaboratively with others.

1. demonstrate proficiency to implement research with R (RD points 3-5)
2. demonstrate ability to produce collaborative and reproducible research using R, RStudio, and GitHub

Prerequisites

This coursebook is aimed at students that have some background in language-related studies or linguistics with a desire to expand their methodological toolbox. It does not assume a strong background in these areas, however. Furthermore, I will make no assumptions about students' experience with programming in general, or programming with R, in particular.

You will need reliable access to the internet and a computer to work with the code in this coursebook and the code found in the accompanying resource site (tadr[7]).

Programming

Programming is the backbone for modern quantitative research. Here are some of the reasons to program:

- *Flexibility* Graphical User Interface (GUI) based software is inherently limited. What you see is what you get. If you have another need, you need to find a tool. If another tool does not implement what you think you need, you are out of luck.
- *Transparency* By taking a programming approach to research analysis you make your decisions explicit and leave a breadcrumb trail to everything you do.
- *Reproducibility* What you do will be clearer to you but also allow you to share the process with others (including your future self!). Insight grows much faster when exposed to light. Sharing your

[6] https://en.wikipedia.org/wiki/Separation_of_content_and_presentation
[7] https://lin380.github.io/tadr/

research with collaborators or on sites such as GitHub or BitBucket brings makes your work visible and accessible to the world. Reproducibility is gaining momentum and is fueled by programmatic approaches to research.

R is a popular programming language with statiticians and was adopted by many other fields in natural and social sciences. There are various reasons why using R for this coursebook is a good choice:

- *One stop shopping* Once known specifically as a statistical programming language, R can now be a round trip tool to acquire, curate, transform, visualize, *and* statistically analyze data. It also allows for robust communication in reports and data and analysis sharing (reproducibility).
- *You are not alone* There is a sizable R programming community, especially in academics. This has two tangible benefits; first, you will likely be able to find user contributed R packages that will satisfy many of the more sophisticated programming goals you will have and second, you will be able to get answers to any of your programming questions on popular sites like StackOverflow.
- *RStudio* RStudio is the envy of many other programmers. It is a very capable interface to R and provides convenient access powerful tools to allow you to be a more efficient and productive R programmer.

Coursebook structure

This coursebook is divided into four parts:

1. In "Foundations", an environmental survey of quantitative research across disciplines and orient language-based research is provided.
2. "Orientation" aims to build your knowledge about what data is, how text is organized into datasets, what role statistics play in quantitative research and the types of statistical approaches that are commonly found in text analysis research, and finally how to develop a research question and a research blueprint for conducting a quantitative text analysis research project.
3. "Preparation" covers a variety of implementation approaches for each stage for deriving a dataset ready for statistical analysis which includes acquiring, curating, and transforming data.
4. "Analysis" elaborates various statistical approaches for data analysis and contextualizes their application in for different types of research questions and aims.

Conventions

This coursebook is about the concepts for understanding and the techniques for doing quantitative text analysis with R. Therefore there will be an intermingling of prose and code presented. As such, an attempt to establish consistent conventions throughout the text has been made to signal reader's attention as appropriate. As we explore concepts, R code itself will be incorporated into the text. This may be a unique coursebook compared to others you have seen. It has been created using R itself –specifically using an R language package called `bookdown` (**?**). This R package makes it possible to write, execute ('run'), and display code and results within the text.

For example, the following text block shows actual R code and the results that are generated when running this code. Note that the hashtag `#` signals a **code comment**. The code follows within the same text block and a subsequent text block displays the output of the code.

```
# Add 1 plus 1
1 + 1
#> [1] 2
```

Inline code will be used when code blocks are short and the results are not needed for display. For example, the same code as above will sometimes appear as `1 + 1`.

When necessary meta-description of code will appear. This is particularly relevant for R Markdown documents.

```
```{r test-code}
1 + 1
```
```

```

In terms of prose, key concepts will be signaled using ***bold italics***. Terms that appear in this typeface will also appear in the [glossary] at the end of the text (TODO). Furthermore, there are four pose text blocks that will be used to signal the reader's attention: *key points*, *notes*, *tips*, *questions*, and *warnings*.

Key points summarize the main points to be covered in a chapter or a subsection of the text.

In this chapter you will learn:

- the goals of this coursebook
- the reasoning for using the R programming language
- important text conventions employed in this coursebook

Notes provide a bit more information on the topic or where to find more information.

R is more than a powerful statistical programming language, it also can be used to perform all the necessary steps in a data science project; including reporting. A relatively new addition to the reporting capabilities of R is the `bookdown` package (this coursebook was created using this very package). You can find out more here[8].

Tips are used to signal helpful hints that might otherwise be overlooked.

During a the course of an exploratory work session, many R objects are often created to test ideas. At some point inspecting the workspace becomes difficult due to the number of objects displayed using `ls()`.

To remove all objects from the workspace, use `rm(list = ls())`.

From time to time there will be points for you to consider and questions to explore.

Consider the objectives in this course: what ways can the knowledge and skills you will learn benefit you in your academic studies and/ or professional and personal life?

Errors will be an inevitable part of learning, but some errors can be avoided. The text will used the warning text block to highlight typical pitfalls and errors.

Hello world!
This is a warning.

Although this is not intended to be a in-depth introduction to statistical techniques, mathematical formulas will be included in the text. These formulas will appear either inline $1 + 1 = 2$ or as block equations.

$$\hat{c} = \underset{c \in C}{\text{argmax}}\ \hat{P}(c) \prod_i \hat{P}(w_i|c) \tag{1}$$

Data analysis leans heavily on graphical representations. Figures will appear numbered, as in Figure **??**.

```
library(ggplot2) # load graphics package
ggplot(mtcars, aes(x = hp, y = mpg)) + # map 'hp' and 'mpg' to coordinate space
  geom_point() + # add points
  geom_smooth(method = "lm") + # draw linear trend line
```
```

```
  labs(x = "Horsepower", # label x axis
       y = "Miles per gallon", # label y axis
       title = "Test plot", # add title
       subtitle = "From mtcars dataset") # add subtitle
```

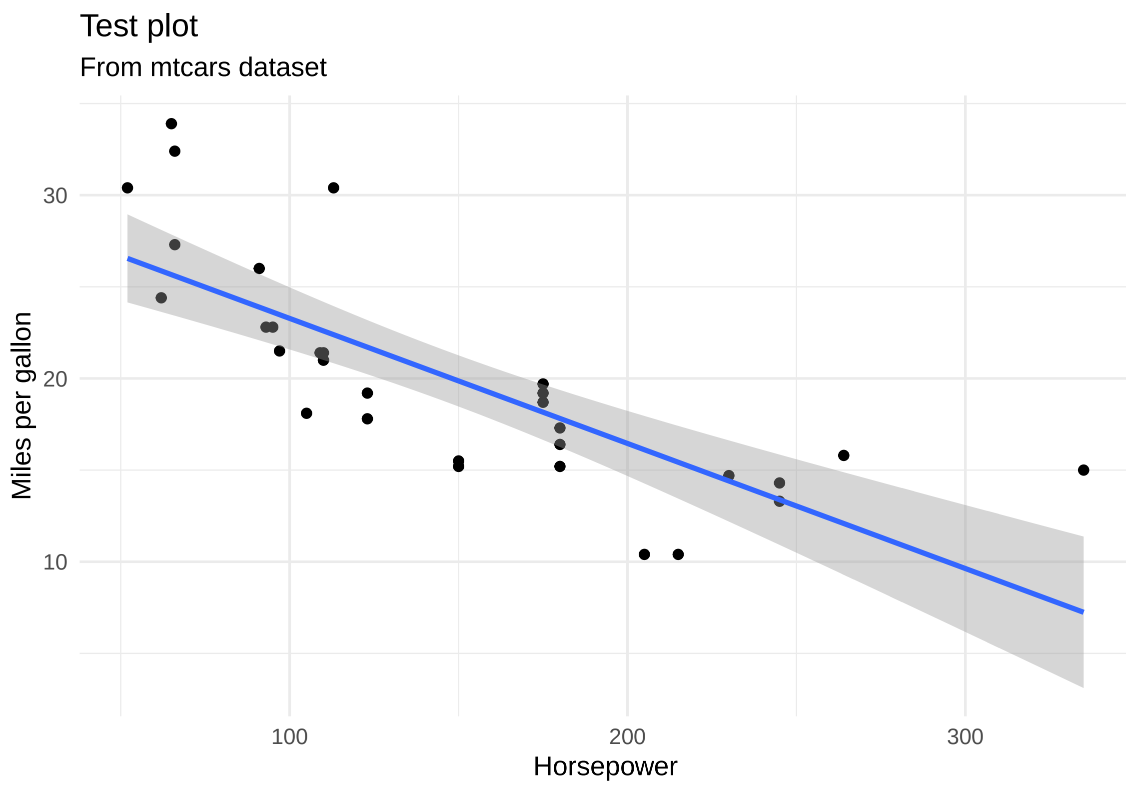

Figure 2: Test plot from mtcars dataset

Tables, such as Table **??** will be numbered separately from figures.

```
knitr::kable(head(iris, 20), caption = "Here is a nice table!", booktabs = TRUE)
```

Resources

This coursebook includes the Text as Data Resources[9] accompany website. This site itself includes resources to learn and extend R skills relevant for conducting reproducible text analysis research. Tutorials[10] are provided which provide video, questions, and interactive coding practice. Recipes[11] includes worked demonstrations of targeted aspects of R programming. Each of these resources are coordinated to provide the programming skills necessary for the stages of text analysis covered in the coursebook.

In addition to the Tutorials and Recipes, students are encouraged to engage with the interactive coding `swirl` activities. In contrast to Tutorials, swirl activities will be performed in an RStudio session in the R console. This provides a more authentic experience for learning to use R.

[9] https://lin380.github.io/tadr/
[10] https://lin380.github.io/tadr/tutorials/index.html
[11] https://lin380.github.io/tadr/articles/index.html

Table 1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Build information

This coursebook was written in bookdown[12] inside RStudio[13]. The website is hosted with GitHub Pages[14] and the complete source is available from GitHub[15].

This version of the coursebook was built with R version 4.1.2 (2021-11-01).

Part I

Foundations

Overview

FOUNDATIONS

In this section the aims are to: (1) provide an overview of quantitative research and their applications, by both highlighting visible applications and notable research in various fields. (2) We will under the hood a bit and consider how quantitative research contributes to language research. (3) I will layout the main types of research and situate quantitative text analysis inside these. (4) I will provide an overview of the rest of the coursebook highlighting the learning goals, the structure of the coursebook, and how this structure will support a robust knowledge of what text analysis is, why it is used, and how to conduct your own research.

[12] http://bookdown.org/
[13] http://www.rstudio.com/ide/
[14] https://pages.github.com/
[15] https://github.com/lin380

1 Data, language, and text analysis

DRAFT

> Science walks forward on two feet, namely theory and experiment...Sometimes it is one foot which is put forward first, sometimes the other, but continuous progress is only made by the use of both.
>
> — Robert A. Millikan[16] (**?**)

The essential questions for this chapter are:

- What is the role and goals of data analysis in and outside of academia?
- In what ways is quantitative language research approached?
- What are some of the applications of text analysis?
- How is this coursebook structured and what are the target learning goals?

In this chapter I will aim to introduce the topic of text analysis and text analytics and frame the approach of this coursebook. The goals of this section are to work from the general field of data science/ data analysis to the particular sub-field of text analysis (where text is defined broadly as corpus). The aim is to introduce the context needed to understand how text analysis fits in a larger universe of data analysis and see the commonalities in the ever-ubiquitous field of data analysis, with attention to how language and linguistics studies employ data analysis down to the particular area of text analysis. To round out this chapter, I will provide a general overview of the rest of the coursebook motivating the general structure and sequencing as well as setting the foundation for programmatic approaches to data analysis.

1.1 Making sense of a complex world

The world around us is full of actions and interactions so numerous that it is difficult to really comprehend. Through the lens each individual sees and experiences this world. We gain knowledge about this world and build up heuristic knowledge about how it works and how we do and can interact with it. This happens regardless of your educational background. As humans we are built for this. Our minds process countless sensory inputs many of which never make it to our conscious mind. They underlie skills and abilities that we take for granted like being able to predict what will happen if you see someone about to knock a wine glass off a table and onto a concrete floor. You've never seen this object before and this is the first time you've been to this winery, but somehow and from somewhere you 'instinctively' make an effort to warn the would-be-glass-breaker before it is too late. You most likely have not stopped to consider where this predictive knowledge has come from, or if you have, you may have just chalked it up to 'common sense'. As common as it may be, it is an incredible display of the brain's capacity to monitor your environment, relate the events and observations that take place, and store that information all the time not making a big fuss to tell you conscious mind what it's up to.

So wait, this is a coursebook on text analytics and language, right? So what does all this have to do with that? Well, there are two points to make that are relevant for framing our journey: (1) the world is full of countless information which unfold in real-time at a scale that is daunting and (2) for all the power of the brain that works so efficiently behind the scene making sense of the world, we are one individual living one life that has a limited view of the world at large. Let me expand on these two points a little more.

First let's be clear. There is no way for any one to experience all things at all times, i.e. omnipotence. But even extremely reduced slices of reality are still vastly outside of our experiental capacity, at least in real-time. One can make the point that since the inception of the internet an individual's ability to experience larger slices of the world has increased. But could you imagine reading, watching, and listening to every file that is currently accessible on the web? Or has been? (See the Wayback Machine[17].) Scale this down even further; let's take Wikipedia, the world's largest encyclopedia. Can you imagine reading every wiki entry? As large

[16] https://www.nobelprize.org/uploads/2018/06/millikan-lecture.pdf
[17] https://web.archive.org/

as a resource such as Wikipedia is [18], it is still a small fragment of the written language that is produced on the web, just the web [20]. Consider that for a moment.

To my second framing point, which is actually two points in one. I made underscored the efficiency of our brain's capacity to make sense of the world. That efficiency comes from some clever evolutionary twists that lead our brain to take in the world but it makes some shortcuts that compress the raw experience into heuristic understanding. What that means is that the brain is not a supercomputer. It does not store every experience in raw form, we do not have access to the records of our experience like we would imagine a computer would have access to the records logged in a database. Where our brains do excel is in making associations and predictions that help us (most of the time) navigate the complex world we inhabit. This point is key –our brains are doing some amazing work, but that work can give us the impression that we understand the world in more detail that we actually do. Let's do a little thought experiment. Close your eyes and think about the last time you saw your best friend. What were they wearing? Can you remember the colors? If your like me, or any other human, you probably will have a pretty confident feeling that you know the answers to these questions and there is a chance you a right. But it has been demonstrated in numerous experiments on human memory that our confidence does not correlate with accuracy (**??**). You've experienced an event, but there is no real reason that we should be our lives on what we experienced. It's a little bit scary, for sure, but the magic is that it works 'good enough' for practical purposes.

So here's the deal: as humans we are (1) clearly unable to experience large swaths of experience by the simple fact that we are individuals living individual lives and (2) the experiences we do live are not recorded with precision and therefore we cannot 'trust' our intuitions, at least in an absolute sense.

What does that mean for our human curiosity about the world around us and our ability to reliably make sense of it? In short it means that we need to approach understanding our world with the tools of science. Science is so powerful because it makes strides to overcome our inherit limitations as humans (breadth of our experience and recall and relational abilities) and bring a complex world into a more digestible perspective. Science starts with question, identifies and collects data, careful selected slices of the complex world, submits this data to analysis through clearly defined and reproducible procedures, and reports the results for others to evaluate. This process is repeated, modifying, and manipulating the procedures, asking new questions and positing new explanations, all in an effort to make inroads to bring the complex into tangible view.

In essence what science does is attempt to subvert our inherent limitations in understanding by drawing on carefully and purposefully collected slices of experience and letting the analysis of this experience speak, even if it goes against our intuitions (those powerful but sometime spurious heuristics that our brains use to make sense of the world).

1.2 Data analysis

At this point I've sketched an outline strengths and limitations of humans' ability to make sense of the world and why science to address these limitations. This science I've described is the one you are familiar with and it has been an indespensible tool to make sense of the world. If you are like me, this description of science may be associated with visions of white coats, labs, and petri dishes. While science's foundation still stands strong in the 21st century, a series of intellectual and technological events mid-20th century set in motion changes that have changed aspects about how science is done, not why it is done. We could call this Science 2.0, but let's use the more popularized term "Data Science". The recognized beginnings of Data Science are attributed to work in the "Statistics and Data Analysis Research" department at Bell Labs during the 1960s. Although primarily conceptual and theoretic at the time, a framework for quantitative data analysis took shape that would anticipate what would come: sizable datasets which would "…require advanced statistical and computational techniques … and the software to implement them." (**?**) This framework emphasized both the inference-based research of traditional science, but also embraced exploratory research and recognized the need to address practical considerations that would arise when working with and deriving insight from an abundance of data.

[18]As of 22 July 2021, there are 6,341,359 articles in the English Wikipedia[19] containing over 3.9 billion words occupying around 19 gigabytes of information.

[20]For reference, Common Crawl[21] has millions of gigabytes collected since 2008

Fast-forward to the 21st century a world in which machine readable data is truly in abundance. With increased computing power and innovative uses of this technology the world wide web took flight. To put this in perspective, in 2019 it was estimated that every minute 511 thousand tweets were posted, 18.1 million text messages were sent, and 188 million emails were sent (**?**). The data flood has not been limited to language, there are more sensors and recording devices than ever before which capture evermore swaths of the world we live in (**?**). Where increased computing power gave rise to the influx of data, it is also on of the primary methods for gathering, preparing, transforming, analyzing, and communicating insight derived from this data (**?**). The vision laid out in the 1960s at Bell Labs had come to fruition.

The interest in deriving insight from the available data is now almost ubiquitous. The science of data has now reached deep into all aspects of life where making sense of the world is sought. Predicting whether a loan applicant will get a loan (**?**), whether a lump is cancerous (**?**), what films to recommend based on your previous viewing history (**?**), what players a sports team should sign (**?**) all now incorporate a common set of data analysis tools.

These advances, however, are not predicated on data alone. As envisioned by researchers at Bell Labs, turning data into insight it takes computing skills (i.e. programming), knowledge of statistics, and, importantly, substantive/ domain expertise. This triad has been popularly represented in a Venn diagram **??**.

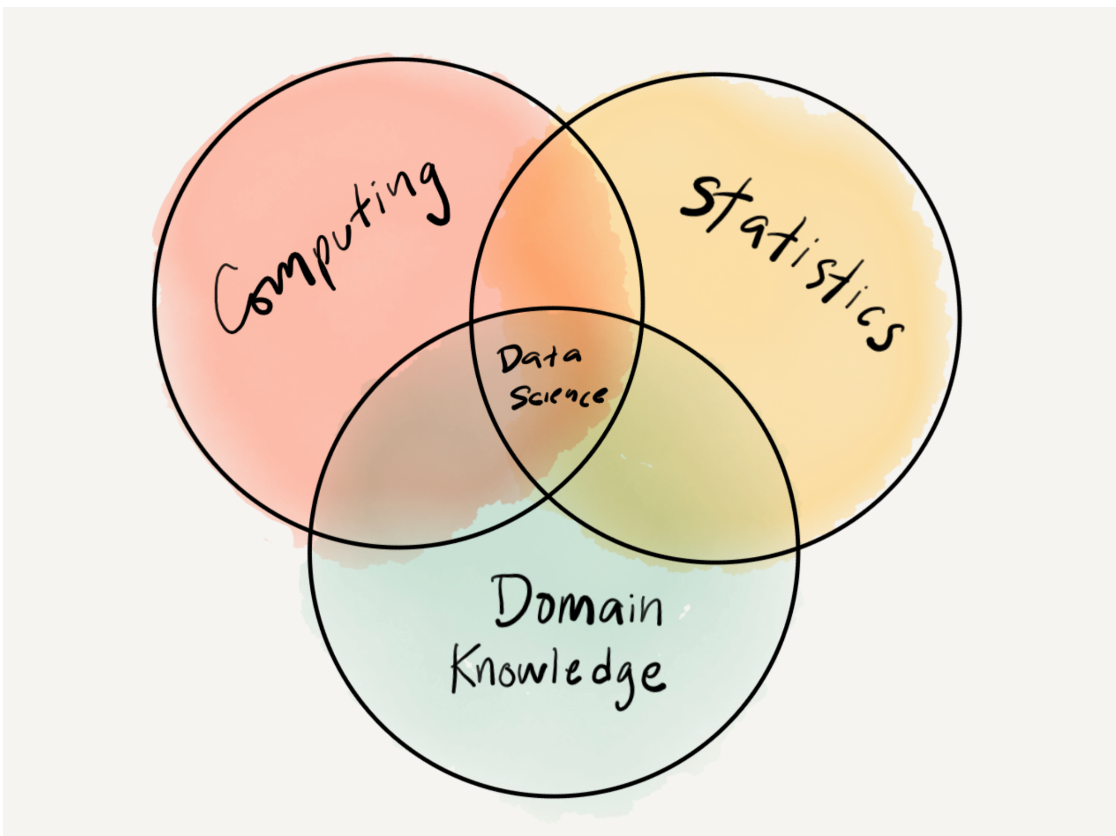

Figure 3: Data Science Venn Diagram adapted from [Drew Conway](http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram).

This same toolbelt underlies well-known public-facing language applications. From the language-capable personal assistant applications, plagiarism detection software, machine translation and search, tangible results of quantitative approaches to language are becoming standard fixtures in our lives.

The spread of quantitative data analysis too has taken root in academia. Even in areas that on first blush don't appear to be approached in a quantitative manner such as fields in the social sciences and humanities,

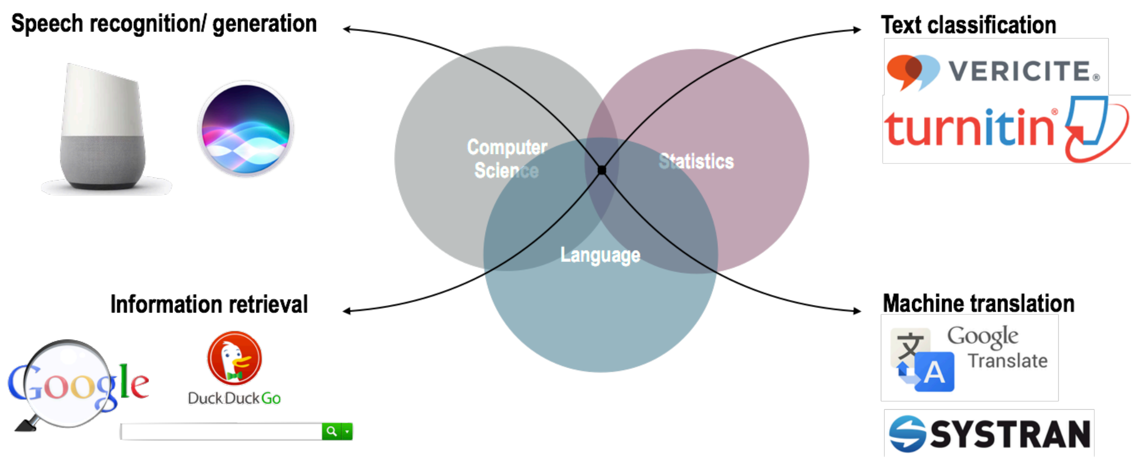

Figure 4: Well-known language applications

data science is making important and sometimes disisplinary changes to the way that academic research is conducted. This coursebook focuses in on a domain that cuts across many of these fields; namely language. At this point let's turn to quantitative approaches to language.

1.3 Language analysis

Language is a defining characteristic of our species. As such, the study of language is of key concern to a wide variety of fields, not just linguists. The goals of various fields, however, and as such approaches to language research, vary. On the one hand some language research traditions, namely those closely associated with Noam Chomsky, eschewed quantitative approaches to language research during the later half of the 20th century and instead turned to qualitative assessment of language structure through introspective methods. On the other hand many language research programs turned to and/or developed quantitative research methods either by necessity or through theoretical principles. These quantitative research trajectories share much of the common data analysis toolbox described in the previous section. This means to a large extent language analysis projects share a common research language with other language research but also with research beyond outside of language. However, there is never a one-size-fits all approach to anything – much less data analysis. And in quantitative analysis there is a key distinction in data collection that has downstream effects in terms of procedure but also in terms of interpretation.

The key distinction, that we need to make at this point, which will provide context for our exploration of text analysis, comes down to the approach to collecting language data and the nature of that data. This distinction is between experimental and observational data collection. Experimental approaches start with a intentionally designed hypothesis and lay out a research methodology with appropriate instruments and a plan to collect data that shows promise for shedding light on the validity of the hypothesis. Experimental approaches are conducted under controlled contexts, usually a lab environment, in which participants are recruited to perform a language related task with stimuli that have been carefully curated by researchers to elicit some aspect of language behavior of interest. Experimental approaches to language research are heavily influenced by procedures adapted from psychology. This link is logical as language is a central area of study in cognitive psychology. This approach looks a much like the white-coat science that we made reference to earlier but, as in most quantitative research, has now taken advantage of the data analysis tool belt to collect and organize much larger quantities of data and conduct statistically more robust analysis procedures and communicate findings more efficiently.

Observational approaches are a bit more of a mixed bag in terms of the rationale for the study; they may either start with a testable hypothesis or in other cases may start with a more open-ended research question to explore. But a more fundamental distinction between the two is drawn in the amount of control the researcher has on contexts and conditions in which the language behavior data to be collected is produced.

Observational approaches seek out records of language behavior that is produced by language speakers for communicative purposes in natural(istic) contexts. This may take place in labs (language development, language disorders, etc.), but more often than not, language is collected from sources where speakers are performing language as part of their daily lives –whether that be posting on social media, speaking on the telephone, making political speeches, writing class essays, reporting the latest news for a newspaper, or crafting the next novel destined to be a New York Times best-seller. What is more, data collected from the 'wild' is varies in more in structure relative to data collected in experimental approaches and requires a number of steps to prepare the data to synch up with the data analysis tool belt.

I liken this distinction between experimental and observational data collection to the difference between farming and foraging. Experimental approaches are like farming; the groundwork for a research plan is designed, much as a field is prepared for seeding, then the researcher performs as series of tasks to produce data, just as a farmer waters and cares for the crops, the results of the process bear fruit, data in our case, and this data is harvested. Observational approaches are like foraging; the researcher scans the available environmental landscape for viable sources of data from all the naturally existing sources, these sources are assessed as to their usefulness and value to address the research question, the most viable is selected, and then the data is collected.

The data acquired from both of these approaches have their trade-offs, just as farming and foraging. Experimental approaches directly elicit language behavior in highly controlled conditions. This directness and level of control has the benefit of allowing researchers to precisely track how particular experimental conditions effect language behavior. As these conditions are an explicit part of the design and therefore the resulting language behavior can be more precisely attributed to the experimental manipulation. The primary shortcoming of experimental approaches is that there is a level of artificialness to this directness and control. Whether it is the language materials used in the task, the task itself, or the fact that the procedure takes place under supervision the language behavior elicited can diverge quite significantly from language behavior performed in natural communicative settings. Observational approaches show complementary strengths and shortcomings. Whereas experimental approaches may diverge from natural language use, observational approaches strive to identify and collected language behavior data in natural, uncontrolled, and unmonitored contexts. In this way observational approaches do not have to question to what extent the language behavior data is or is not performed as a natural communicative act. On the flipside, the contexts in which natural language communication take place are complex relative to experimental contexts. Language collected from natural contexts are nested within the complex workings of a complex world and as such inevitably include a host of factors and conditions which can prove challenging to disentangle from the language phenomenon of interest but must be addressed in order to draw reliable associations and conclusions.

The upshot, then, is twofold: (1) data collection methods matter for research design and interpretation and (2) there is no single best approach to data collection, each have their strengths and shortcomings. In the ideal, a robust science of language will include insight from both experimental and observational approaches (**?**). And evermore there is greater appreciation for the complementary nature of experimental and observational approaches and a growing body of research which highlights this recognition. Given their particular trade-offs observational data is often used as an exploratory starting point to help build insight and form predictions that can then be submitted to experimental conditions. In this way studies based on observational data serve as an exploratory tool to gather a better and more externally valid view of language use which can then serve to make prediction that can be explore with more precision in an experimental paradigm. However, this is not always the case. Observational data is also often used in hypothesis-testing contexts as well. And furthermore, some in some language-related fields, a hypothesis-testing is not the ultimate goal for deriving knowledge and insight.

1.4 Text analysis

Text analysis is the application of data analysis procedures from data science to derive insight from textual data collected through observational methods. I have deliberately chosen the term 'text analysis' to avoid what I see are the pitfalls of using some other common terms in the literature such as Corpus Linguistics, Computational Linguistics, or Digital Humanities. There are plenty of learning resources that focus specifically on one of these three fields when discussing the quantitative analysis of text. But from my perspective what is missing is a resource which underscores the fact that text analysis research and the methods employed span across a wide variety of academic fields and applications in industry. This coursebook aims to introduce you to these areas through the lens of the data and analysis procedures and not through a particular field. This approach, I hope, provides a wider view of the potential applications of using text as data and inspires you to either employ quantitative text analysis in your research and/ or to raise your awareness of the advantages of text analysis for making sense of language-related and linguistic-based phenomenon.

So what are some applications of text analysis? For most the public facing applications that stem from Computational Linguistic research, often known as Natural Language Processing by practitioners, are the most well-known applications of text analysis. Whether it be using search engines, online translators, submitting your paper to plagiarism detection software, etc. the text analysis methods we will cover are at play. These uses of text analysis are production-level applications and there is big money behind developing evermore robust text analysis methods.

In academia the use of quantitative text analysis is even more widespread, despite the lack of public fanfare. Let's run through some select studies to give you an idea of the areas that are employing text analysis, of what researchers are doing with text analysis, and to whet your interest for conducting your own text analysis project.

? track the geographic spread of neologisms from city to city in the United States using Twitter data collected between 6/2009 and 5/2011. They only used tweets with geolocation data and then associated each tweet with a zipcode using the US Census. The most populous metropolitan areas were used. They also used the demographics from these areas to make associations between lexical innovations and demographic attributes. From this analysis they are able to reconstruct a network of linguistic influence. One of the main findings is that demographically-similar cities are more likely to share linguistic influence. At the individual level, there is a strong, potentially stronger role of demographics than geographical location.

? explore potential racial disparities in officer respect in police body camera footage. The dataset is based on body camera footage from the Oakland Police Department during April 2014. At total of 981 stops by 245 different officers were included (black 682, white 299) and resulted in 36,738 officer utterances. The authors found evidence for racial disparities in respect but not formality of utterances, with less respectful language used with the black community members.

? investigate whether the established drop in language complexity of rhetoric in election seasons is associated with election outcomes. The authors used US Democratic Primary Debates from 2004. The results suggest that although there was no overall difference in complexity between winners and losers, their pattern differed over time. Winners tended to drop the complexity of their language closer to the upcoming election.

? explore the extent to which languages are positively, neutrally, or negatively biased. Using Twitter, Google Books (1520-2008), NY Times newspaper (1987-2007), and music lyrics (1960-2007) the authors extract the top 5,000 most frequent words from each source and have participants rate each word for happiness (9-point scale). The results show that positive words strongly outnumber negative words overall suggesting English is positive-, and pro-social- biased.

? investigates possible differences between L1-English and L1-Chinese undergraduate students' use of lexical bundles, multiword sequences which are extended collocations (i.e. as the result of), in argumentative essays. The authors used the Michigan Corpus of Upper-Level Student Papers

(MICUSP) corpus using the argumentative essay section for L1-English and the Corpus of Ohio Learner and Teacher English (COLTE) for the L1-Chinese English essays. They found that L1-Chinese writers used more than 2 times as many bundle types than L1-English peers which they attribute to L1-Chinese writers attempt to avoid uncommon expressions and/or due to their lack of register awareness (conversation has more bundles than writing generally).

? use a corpus study to investigate the phenomenon of syntactic persistence, the increased tendency for speakers to use a particular syntactic form over an alternate when the syntactic form is recently processed. The authors attempt to distinguish between two competing explanations for the phenomenon: (1) transient activation, where the increased tendency is short-lived and time-bound and (2) implicit learning, where the increased tendency is a reflect of learning mechanisms. The use of a speech corpora (Switchboard and spoken BNC) were used to avoid the artificialness that typically occurs in experimental settings. The authors investigated the ditransitive alternation (NP PP/ NP NP), voice alternation (active/ passive), and complementizer/ relativizer omission. In these alternations structural bias was established by measuring the probability for a verb form to appear in one of the two syntactic forms. Then the probability that that form (target) would change given previous exposure to the alternative form (prime) was calculated; what the authors called surprisal. Distance between the prime structure and the target verb were considered in the analysis. In these alternations, the less common structure was used in the target more often when the when it corresponded to the prime form (higher surprisal) suggesting that implicit learning underlies syntactic persistence effects.

? explore differences between British and American English at the lexico-syntactic level in the *into*-causative construction (ex. 'He tricked me into employing him.'). The analysis uses newspaper text (The Guardian and LA Times) and the findings suggest that American English uses this construction in verbal persuasion verbs whereas British English uses physical force verbs.

? provide a method for solving the authorship debate surrounding The Federalist papers [22]. They employ a probabilistic approach using the word frequency profiles of the articles with known authors to predict authorship of the disputed 12 papers. The results suggest that the disputed papers were most likely authored by Madison.

? investigate the extent to which translated texts differ from native texts. In particular the author explores the notion of explicitation in translated texts (the tendency to make information in the source text explicit in the target translation). The study makes use of the Translational English Corpus (TEC) for translation samples and comparable sections of the British National Corpus (BNC) for the native samples. The results suggest that there is a tendency for syntactic explicitation in the translational corpus (TEC) which is assumed to be a subconscious process employed unwittingly by translators.

This sample of studies include research from areas such as translation, stylistics, language variation, dialectology, psychology, psycholinguistics, political science, and sociolinguistics which highlights the diversity of fields and subareas which employ quantitative text analysis. Text analysis is at the center of these studies as they share a set of common goals:

1. To detect and retrieve patterns from text too subtle or too numerous to be done by hand
2. To challenge assumptions and/or provide other views from textual sources
3. To explore new questions and/or provide novel insight

Let's now turn to the last section of this chapter which will provide an overview of the rationale for doing learning to do text analysis, the structure of the content covered, and a justification for the approach we will take to perform text analysis.

1.5 Coursebook overview

In this section I will provide a general overview of the rest of the coursebook motivating the general structure and sequencing as well as setting the foundation for programmatic approaches to data analysis. Let me highlight why I think this is a valuable area of study, what I hope you gain from this coursebook, and how the structure of this coursebook is configured to help scaffold your conceptual and practical knowledge of text analysis.

The target learning outcomes in this coursebook are the following:

1. Data Literacy
2. Research Skills
3. Programming Skills

Data Literacy refers to the ability to interpret, assess, and contextualize findings based on data. Throughout this coursebook we will explore topics which will help you understand how data analysis methods derive insight from data. In this process you will be encouraged to critically evaluate connections across linguistic and language-related disciplines using data analysis knowledge and skills. Data Literacy is an invaluable skillset for academics and professionals (cite) but also is an indispensable aptitude for in the 21st century citizens to navigate and actively participate in the 'Information Age' in which we live (**?**).

Research skills covers the ability to conduct original research, communicate findings, and make meaningful connections with findings in the literature of the field. This target area does not differ significantly, in spirit, from common learning outcomes in a research methods course: identify an area of investigation, develop a viable research question or hypothesis, collect relevant data, analyze data with relevant statistical methods, and interpret and communicate findings. However, working with text will incur a series of key steps in the selection, collection, and preparation of the data that are unique to text analysis projects. In addition, I will stress the importance of research documentation and creating reproducible research as an integral part of modern scientific inquiry (**?**).

Programming skills aims to develop your ability to implement research skills programmatically and produce research that is replicable and collaborative. Modern data analysis, and by extension, text analysis is conducted using programming. There are various key reasons for this: (1) programming affords researchers unlimited research freedom –if you can envision it, you can program it. The same cannot be said for off-the-shelf software which is either proprietary or unmaintained –or both. (2) programming underlies well-documented and reproducible research –documenting button clicks and menu option selections leads to research which is not readily reproduced, either by some other researcher or by your future self! (3) programming forces researchers to engage more intimately with the data and the methods for analysis. The more familiar you are with the data and the methods the more likely you are to produce higher quality work.

Now let me turn to how these learning goals integrate and shape the structure and sequencing of the following chapters.

In Part II "Orientation" we will build our Data Literacy skills working from data to insight. This progression is visualized in Figure **??** [23].

The DIKI Hierarchy highlights the stages and intermediate steps required to derive insight from data. Chapter 2 "Understanding data" will cover both Data and Information covering the conceptual topics of populations versus samples and how language data samples are converted to information and the forms that they can take. In Chapter 3 "Approaching analysis" I will discuss the distinction between descriptive and analytic statistics. In brief they are important for data analysis, but descriptive statistics serve as a sanity check on the dataset before submitting it to interrogation –which is the goal of analytic statistics. We will also cover some of the main distinctions between analytics approaches including inference-, exploration-, and

[23] Adapted from **?**

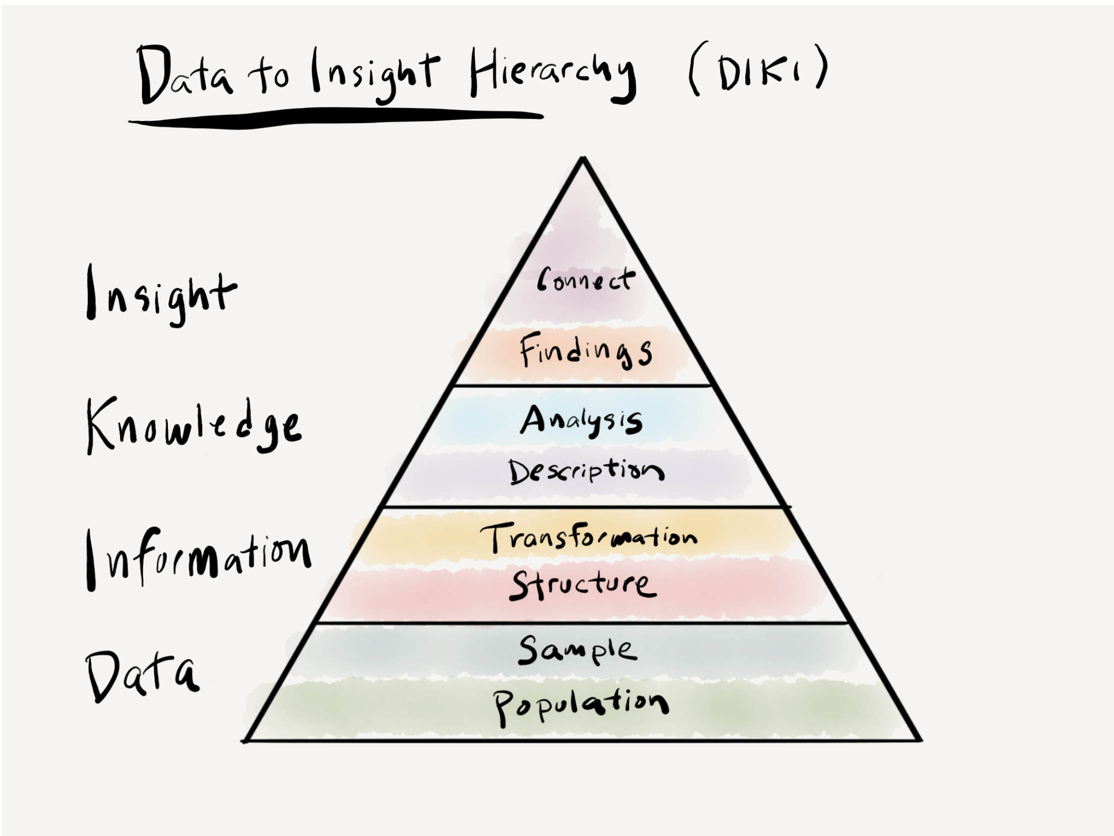

Figure 5: Data to Insight Hierarchy (DIKI)

prediction-based methods. With a fundamental understanding of data, information, and knowledge we will then move to Chapter 4 "Framing research" where we will discuss how to develop a research plan, or what I will call a 'research blueprint'. At this point we will directly address Research Skills and elaborate on how research really comes together; how to bring yourself up to speed with the literature on a topic, how to develop a research goal or hypothesis, how to select data which is viable to address the research goal or hypothesis, how to determine the necessary information and appropriate measures to prepare for analysis, how to perform diagnostic statistics on the data and make adjustments before analysis, how to select and perform the relevant analytic statistics given the research goals, how to report your findings, and finally, how to structure your project so that it is well-documented and reproducible.

Part III "Preparation" and Part IV "Analysis" serve as practical and more detailed guides to the R programming strategies to conduct text analysis research and as such develop your Programming Skills. In Chapter 5 "Acquire data"[24] I will discuss three main strategies for accessing data: direct downloads, Automatic Programming Interfaces (APIs), and web scraping. In Chapter 6 "Curate data"[25] I will outline the process for converting or augmenting the acquired data into a structured format, therefore creating information. This will include organizing linguistic and non-linguistic metadata into one dataset. In Chapter 7 "Transform data"[26] I describe how to work with a curated dataset to derive more detailed information and appropriate dataset structures that are appropriate for the upcoming analysis.

Chapters 8 "Inference"[27], 9 "Prediction"[28], and 10 "Exploration"[29] focus on different categories of statistical analysis each associated with distinct research goals. Inference deals with analysis methods associated with standard hypothesis-testing. This will include some common statistical models employed in text analysis: chi-squared, logistic regression, and linear regression. Prediction covers methods for modeling associations in data with the aim to accurately predict outcomes on new textual data. I will cover some standard methods for text classification including Näive Bayes, *k*-nearest neighbors (*k*-NN), and decisions tree and random forest models. Exploration covers a variety of analysis methods such as association measures, clustering, topic modeling, and vector-space models. These methods are aligned with research goals that aim to interpret patterns that arise in from the data itself.

Summary

In this chapter I started with some general observations about the difficulty of making sense of a complex world. The standard approach to overcoming inherent human limitations in sense making is science. In the 21st century the toolbelt for doing scientific research and exploration has grown in terms of the amount of data available, the statistical methods for analyzing the data, and the computational power to manage, store, and share the data, methods, and results from quantitative research. The methods and tools for deriving insight from data have made significant inroads in and outside academia, and increasingly figure in the quantitative investigation of language. Text analysis is a particular branch of this enterprise based on observational data from real-world language and is used in a wide variety of fields. This coursebook aims to develop your knowledge and skills in three fundamental areas: Data Literacy, Research Skills, and Programming Skills.

In the end I hope that you enjoy this exploration into text analysis. Although learning curve at times may seem steep –the experience you will gain will not only improve your data literacy, research skills, and programmings skills but also enhance your appreciation for the richness of human language and its important role in our everyday lives.

24

25

26

27

28

29

Part II

Orientation

Overview

ORIENTATION

Before we begin working on the specifics of our data project, it is important to establish a fundamental understanding of the characteristics of each of the levels in the DIKI Hierarchy (Figure **??**) and the roles each of these levels have in deriving insight from data. In Chapter 2 we will explore the Data and Information levels drawing a distinction between two main types of data (populations and samples) and then cover how data is structured and transformed to generate information (datasets) that is fit for statistical analysis. In Chapter 3 I will outline the importance and distinct types of statistical procedures (descriptive and analytic) that are commonly used in text analysis. Chapter 4 aims to tie these concepts together and cover the required steps for preparing a research blueprint to conduct an original text analysis project.

2 Understanding data

DRAFT

> The plural of anecdote is not data.
>
> — Marc Bekoff

The essential questions for this chapter are:

- What are the distinct types of data and how do they differ?
- What is information and what form does it take?
- What is the importance of documentation in quantitative research?

In this chapter I cover the starting concepts in our journey to understand how to derive insight from data, illustrated in the DIKI Hierarchy (Figure **??**), focusing specifically on the first two levels: Data and Information. We will see that what is commonly referred to as 'data' everyday uses is broken into three distinct categories, two of which are referred to as data and the third is known as information. We will also cover the importance of documentation of data and datasets in quantitative research.

2.1 Data

Data is data, right? The term 'data' is so common in popular vernacular it is easy to assume we know what we mean when we say 'data'. But as in most things, where there are common assumptions there are important details the require more careful consideration. Let's turn to the first key distinction that we need to make to start to break down the term 'data': the difference between populations and samples.

2.1.1 Populations

The first thing that comes to many people's mind when the term population is used is human populations. Say for example –What's the population of Milwuakee? When we speak of a population in these terms we are talking about the total sum of people living within the geographical boundaries of Milwaukee. In concrete terms, a **population** is the objective make up of an idealized set of objects and events in reality. Key terms here are objective and idealized. Although we can look up the US Census report for Milwaukee and retrieve a figure for the population, this cannot truly be the population. Why is that? Well, whatever method that was used to derive this numerical figure was surely incomplete. If not incomplete, by the time someone recorded the figure some number of residents of Milwaukee moved out, moved in, were born, or passed away –the figure is no longer the true population.

Likewise when we talk about populations in terms of language we dealing with an objective and idealized aspect of reality. Let's take the words of the English language as an analog to our previous example population. In this case the words are the people and English is the bounding characteristic. Just as people, words move out, move in, are born, and pass away. Any compendium of the words of English at any moment is almost instananeously incomplete. This is true for all populations, save those in which the bounding characteristics select a narrow slice of reality which is objectively measurable and whose membership is fixed (the complete works of Shakespeare, for example).

In sum, (most) populations are amorphous moving targets. We objectively hold them to exist, but in practical terms we often cannot nail down the specifics of populations. So how do researchers go about studying populations if they are theoretically impossible to access directly? The strategy employed is called sampling.

2.1.2 Sampling

A **sample** is the product of a subjective process of selecting a finite set of observations from an objective population with the goal of capturing the relevant characteristics of the target population. Although there are strategies to minimize the mismatch between the characteristics of the subjective sample and objective population, it is important to note that it is almost certainly true that any given sample diverges from the population it aims to represent to some degree. The aim, however, is to employ a series of sampling decisions, which are collectively known as a sampling frame, that maximize the chance of representing the population.

What are the most common sampling strategies? First **sample size**. A larger sample will always be more representative than a smaller sample. Sample size, however, is not enough. It is not hard to imagine a large sample which by chance captures only a subset of the features of the population. A next step to enhance sample representativeness is apply **random sampling**. Together a large random sample has an even better chance of reflecting the main characteristics of the population better than a large or random sample. But, random as random is, we still run the risk of acquiring a skewed sample (i.e a sample which does not mirror the target population).

To help mitigate these issues, there are two more strategies that can be applied to improve sample representativeness. Note, however, that while size and random samples can be applied to any sample with little information about internal characteristics of the population, these next two strategies require decisions depend on the presumed internal characteristics of the population. The first of these more informed sampling strategies is called **stratified sampling**. Stratified samples make (educated) assumptions about sub-components within the population of interest. With these sub-populations in mind, large random samples are acquired for each sub-population, or strata. At a minimum, stratified samples can be no less representative than random sampling alone, but the chances that the sample is better increases. Can there be problems in the approach? Yes, and on two fronts. First knowledge of the internal components of a population are often based on a limited or incomplete knowledge of the population. In other words, strata are selected subjectively by researchers using various heuristics some of which are based on some sense of 'common knowledge'. The second front that stratified sampling can err concerns the relative sizes of the sub-components relative to the whole population. Even if the relevant sub-components are identified, their relative size adds another challenge in which researchers must face in order to maximize the representativeness of a sample. To attempt to align, or **balance**, the relative sizes of the samples for the strata is the second population-informed sampling strategy.

A key feature of a sample is that it is purposely selected. Samples are not simply a collection or set of data from the population. Samples are rigorously selected with an explicit target population in mind. In text analysis a purposely sampled collection of texts, of the type defined here, is known as a **corpus.** For this same reason a set of texts or documents which have not been selected along a purposely selected sampling frame is not a corpus. The sampling frame, and therefore the populations modeled, in any given corpus most likely will vary and for this reason it is not a safe assumption that any given corpus is equally applicable for any and every research question. Corpus development (i.e. sampling) is purposeful, and the characteristics of the corpus development process should be made explicit through documentation. Therefore vetting a corpus sample for its applicability to a research goal is a key step in that a research must take to ensure the integrity of the research findings.

The Brown Corpus is widely recognized as one of the first large, machine-readable corpora. It was compiled by **?**. Consult the documentation for this corpus[30]. Can you determine what language population this corpus aims to represent? Given the sampling frame for this corpus (in the documentation and summarized in Figure **??**), what types of research might this corpus support or not support?

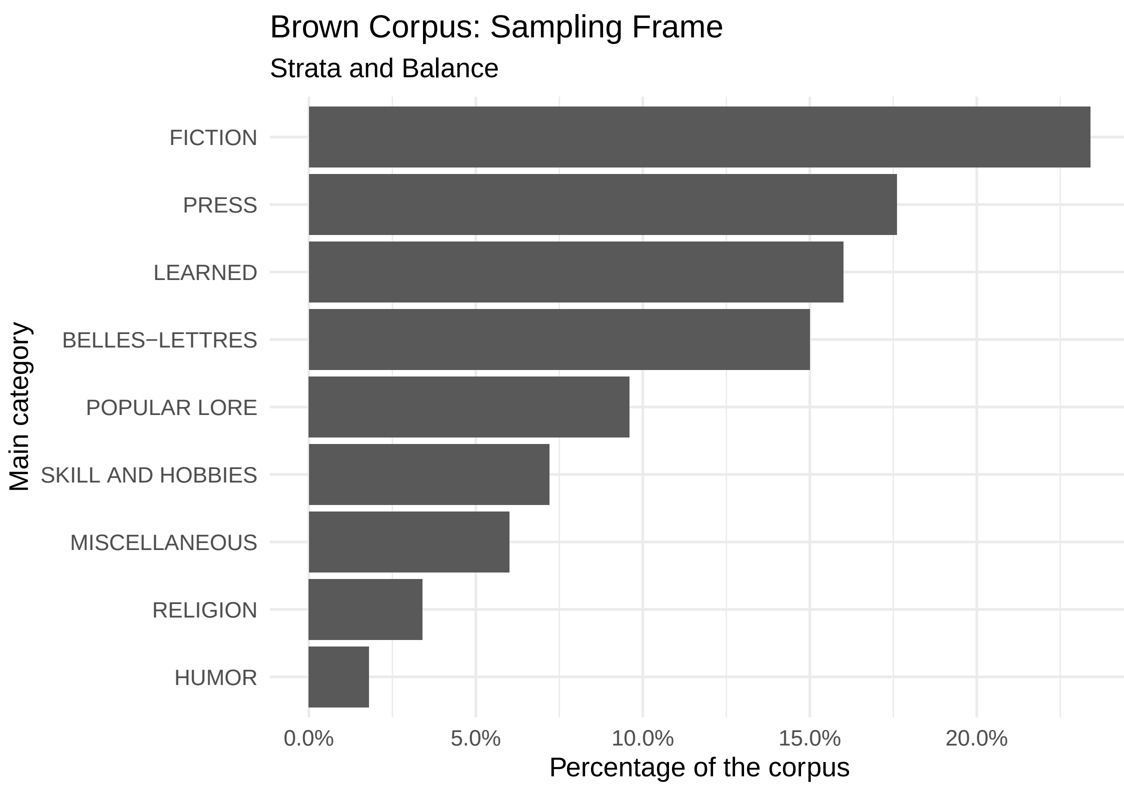

Figure 6: Brown Corpus of Written American English

2.1.3 Corpora

2.1.3.1 Types With the notion of sampling frames in mind, some corpora are compiled with the aim to be of general purpose (general or **reference corpora**), and some with much more specialized sampling frames (**specialized corpora**). For example, the American National Corpus (ANC)[31] or the British National Corpus (BNC)[32] are corpora which aim to model (represent/ reflect) the general characteristics of the English language, the former of American English and the later British English. These are ambitious projects, and require significant investments of time in corpus design and then in implementation (and continued development) that are usually undertaken by research teams (**?**).

Specialized corpora aim to represent more specific populations. The Santa Barbara Corpus of Spoken American English (SBCSAE)[33], as you can imagine from the name of the resource, aims to model spoken American

[31] https://www.anc.org/
[32] http://www.natcorp.ox.ac.uk/
[33] https://www.linguistics.ucsb.edu/research/santa-barbara-corpus

English. No claim to written English is included. There are even more specific types of corpora which attempt to model other types of sub-populations such as scientific writing, computer-mediated communication (CMC)[34], language use in specific regions of the world[35], a country[36], or a region[37], etc.

Another set of specialized corpora are resources which aim to compile texts from different languages or different language varieties for direct or indirect comparison. Corpora that are directly comparable, that is they include source and translated texts, are called **parallel corpora**. Parallel corpora include different languages or language varieties that are indexed and aligned at some linguistic level (i.e. word, phrase, sentence, paragraph, or document), see OPUS[38]. Corpora that are compiled with different languages or language varieties but are not directly aligned are called **comparable corpora**. The comparable language or language varieties are sampled with the same or similar sampling frame, for example Brown[39] and LOB[40] corpora.

The aim of the quantitative text researcher is to select the corpus or corpora (plural of corpus) which best aligns with the purpose of the research. Therefore a general corpus such as the ANC may be better suited to address a question dealing with the way American English works, but this general resource may lack detail in certain areas, such as medical language[41], that may be vital for a research project aimed at understanding changes in medical terminology.

2.1.3.2 Sources The most common source of data used in contemporary quantitative research is the internet. On the web an investigator can access corpora published for research purposes and language used in natural settings that can be coerced by the investigator into a corpus. Many organizations exist around the globe that provide access to corpora in browsable catalogs, or **repositories**. There are repositories dedicated to language research, in general, such as the Language Data Consortium[42] or for specific language domains, such as the language acquisition repository TalkBank[43]. It is always advisable to start looking for the available language data in a repository. The advantage of beginning your data search in repositories is that a repository, especially those geared towards the linguistic community, will make identifying language corpora faster than through a general web search. Furthermore, repositories often require certain standards for corpus format and documentation for publication. A standardized resource many times will be easier to interpret and evaluate for its appropriateness for a particular research project.

In the table below I've compiled a list of some corpus repositories to help you get started.

Repositories are by no means the only source of corpora on the web. Researchers from around the world provide access to corpora and other data sources on their own sites or through data sharing platforms. Corpora of various sizes and scopes will often be accessible on a dedicated homepage or appear on the homepage of a sponsoring institution. Finding these resources is a matter of doing a web search with the word 'corpus' and a list of desired attributes, including language, modality, register, etc. As part of a general movement towards reproducibility more corpora are available on the web than ever before. Therefore data sharing platforms supporting reproducible research, such as GitHub[44], Zenodo[45], Re3data[46], OSF[47], etc., are a good place to look as well, if searching repositories and targeted web searches do not yield results.

In the table below you will find a list of corpus resources and datasets.

[34] https://www.clarin.eu/resource-families/cmc-corpora
[35] http://ice-corpora.net/ice/index.html
[36] https://cesa.arizona.edu
[37] https://cesa.arizona.edu
[38] https://opus.nlpl.eu/
[39] https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0402
[40] https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0167
[41] http://www.hd.uib.no/icame/ij22/vihla.pdf
[42] https://www.ldc.upenn.edu/
[43] http://talkbank.org/
[44] https://github.com/
[45] https://zenodo.org/
[46] http://www.re3data.org/
[47] https://osf.io/

Table 2: A list of some corpus repositories

Resource	Description
BYU corpora	A repository of corpora tha
COW (COrpora from the Web)	A collection of linguisticall
Leipzig Corpora Collection	Corpora in different langua
Linguistic Data Consortium	Repository of language cor
LRE Map	Repository of language reso
NLTK language data	Repository of corpora and
OPUS - an open source parallel corpus	Repository of translated te
TalkBank	Repository of language col
The Language Archive	Various corpora and langua
The Oxford Text Archive (OTA)	A collection of thousands o

Table 3: Corpora and language datasets.

Resource
<a href=”http://www.socsci.uci.edu/~lpearl/CoLaLab/CHILDESTreebank/childestreebank.html”>CHILDES Treebank<
<a href=”http://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html”>Cornell Movie-Dialogs Corpus</a
Corpus Argentino
Corpus of Spanish in Southern Arizona
Europarl Parallel Corpus
Google Ngram Viewer
International Corpus of English (ICE)
OpenSubtitles2011
Russian National Corpus
The Big Bad NLP Database - Quantum Stat
The Switchboard Dialog Act Corpus
Welcome to LANGSNAP - LANGSNAP
<a href=”http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html”>Westbury Lab Web Site

Table 4: R Package interfaces to language corpora and datasets.

Resource	Description
aRxiv	R package interface to qu
crminer	R package interface focus
dvn	R package interface to ac
fulltext	R package interface to qu
gutenbergr	R package interface to do
internetarchive	R package interface to qu
newsflash	R package interface to qu
oai	R package interface to qu
rfigshare	R package interface to qu
rtweet	R client for interacting wi

Table 5: Language data from previous research and meta-studies.

Resource	Description
English Lexicon Project	Access to a large se
lingtypology	R package interface
The Corpus of Linguistic Acceptability (CoLA)	A corpus that consi
The Moby lexicon project	Language wordlists

Language corpora prepared by researchers and research groups listed on repositories or hosted by the researchers themselves is often the first place to look for data. The web, however, contains a wealth of language and language-related data that can be accessed by researcher to compile their own corpus. There are two primary ways to attain language data from the web. The first is through the process of web scraping. Web scraping is the process of harvesting data from the web either manually or (semi-)automatically from the actual public-facing web. The second way to acquire data from the web is through an Application Programming Interface (API). APIs are, as the title suggests, programming interfaces which allow access, under certain conditions, to information that a website or database accessible via the web contains.

The table below lists some R packages that serve to interface language data directly through R.

Data for language research is not limited to (primary) text sources. Other sources may include processed data from previous research; word lists, linguistic features, etc.. Alone or in combination with text sources this data can be a rich and viable source of data for a research project.

Below I’ve included some processed language resources.

The list of data available for language research is constantly growing. I’ve document very few of the wide variety of resources. Below I’ve included attempts by others to provide a summary of the corpus data and language resources available.

Table 6: Lists of corpus resources.

Resource
<a href=”https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html”>Learner corpora a
Machine Learning Datasets | Papers With Code
Stanford NLP corpora
<a href=”https://makingnoiseandhearingthings.com/2017/09/20/where-can-you-find-language-data-on-the-web/”>Where

Explore some of the resources listed above and consider their sampling frames. Can you think of a research question or questions that this resource may be well-suited to support research into? What types of questions would be less-than-adequate for a given resource?

2.1.3.3 Formats A corpus will often include various types of non-linguistic attributes, or **meta-data**, as well. Ideally this will include information regarding the source(s) of the data, dates when it was acquired or published, and other author or speaker information. It may also include any number of other attributes that were identified as potentially important in order to appropriately document the target population. Again, it is key to match the available meta-data with the goals of your research. In some cases a corpus may be ideal in some aspects but not contain all the key information to address your research question. This may mean you will need to compile your own corpus if there are fundamental attributes missing. Before you consider compiling your own corpus, however, it is worth investigating the possibility of augmenting an available corpus to bring it inline with your particular goals. This may include adding new language sources, harnessing software for linguistic annotation (part-of-speech, syntactic structure, named entities, etc.), or linking available corpus meta-data to other resources, linguistic or non-linguistic.

Corpora come in various formats, the main three being: running text, structured documents, and databases. The format of a corpus is often influenced by characteristics of the data but may also reflect an author's individual preferences as well. It is typical for corpora with few meta-data characteristics to take the form of running text.

Running text sample from the Europarle Parallel Corpus[48].

```
> Resumption of the session
> I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I w
> Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people
> You have requested a debate on this subject in the course of the next few days, during this part-sess
> In the meantime, I should like to observe a minute' s silence, as a number of Members have requested,
> Please rise, then, for this minute' s silence.
> (The House rose and observed a minute' s silence)
> Madam President, on a point of order.
> You will be aware from the press and television that there have been a number of bomb explosions and 
> One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam, who had visited th
```

In corpora with more meta-data, a header may be appended to the top of each running text document or the meta-data may be contained in a separate file with appropriate coding to coordinate meta-data attributes with each text in the corpus.

Meta-data header sample from the Switchboard Dialog Act Corpus[49].

```
> FILENAME: 4325_1632_1519
> TOPIC#:       323
> DATE:     920323
> TRANSCRIBER:  glp
> UTT_CODER:    tc
> DIFFICULTY:   1
> TOPICALITY:   3
> NATURALNESS:  2
> ECHO_FROM_B:  1
> ECHO_FROM_A:  4
> STATIC_ON_A:  1
> STATIC_ON_B:  1
```

[48] https://www.statmt.org/europarl/
[49]

```
> BACKGROUND_A: 1
> BACKGROUND_B: 2
> REMARKS:        None.
>
> =========================================================================
>
>
> o          A.1 utt1: Okay.  /
> qw          A.1 utt2: {D So, }
>
> qy^d          B.2 utt1: [ [ I guess, +
>
> +          A.3 utt1: What kind of experience [ do you, + do you ] have, then with child care? /
>
> +          B.4 utt1: I think, ] + {F uh, } I wonder ] if that worked. /
>
> qy          A.5 utt1: Does it say something? /
>
> sd          B.6 utt1: I think it usually does.  /
> ad          B.6 utt2: You might try, {F uh, }  /
> h          B.6 utt3: I don't know,  /
> ad          B.6 utt4: hold it down a little longer,  /
> ad          B.6 utt5: {C and } see if it, {F uh, } -/
```

When meta-data and/ or linguistic annotation increases in complexity it is common to structure each corpus document more explicitly with a markup language such as XML (Extensible Markup Language) or organize relationships between language and meta-data attributes in a database.

XML format for meta-data (and linguistic annotation) from the Brown Corpus[50].

```
> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader><fileDesc><titleStmt><title>Sample A01 from  The A
>  "Hartsfield Files"
>  August 17, 1961, "Urged strongly ..."
>  "Sam Caldwell Joins"
>  March 6,1961, p.1 "Legislators Are Moving" by Reg Murphy
>  "Legislator to fight" by Richard Ashworth
>  "House Due Bid..."
>  p.18 "Harry Miller Wins..."
> </title></titleStmt><editionStmt><edition>A part  of the XML version of the Brown Corpus</edition></ed
> <text xml:id="A01" decls="A">
> <body><p><s n="1"><w type="AT">The</w> <w type="NP" subtype="TL">Fulton</w> <w type="NN" subtype="TL">
> </p>
```

Although there has been a push towards standardization of corpus formats, most available resources display some degree of idiosyncrasy. Being able to parse the structure of a corpus is a skill that will develop with time. With more experience working with corpora you will become more adept at identifying how the data is stored and whether its content and format will serve the needs of your analysis.

2.2 Information

Identifying an adequate corpus resource for the target research question is the first step in moving a quantitative text research project forward. The next step is to select the components or characteristics of this resource that are relevant for the research and then move to organize the attributes of this data into a more useful and informative format. This is the process of converting a corpus into a **dataset** –a tabular representation of the information to be leveraged in the analysis.

[50] http://www.nltk.org/nltk_data/

Table 7: First 10 source and target sentences in the Europarle Corpus.

type	sentence_id	sentence
Source	1	Resumption of the session
Target	1	Reanudación del período de sesiones
Source	2	I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999,
Target	2	Declaro reanudado el período de sesiones del Parlamento Europeo, interrumpido el viernes 17 de d
Source	3	Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the peopl
Target	3	Como todos han podido comprobar, el gran "efecto del año 2000" no se ha producido. En cambio,
Source	4	You have requested a debate on this subject in the course of the next few days, during this part-se
Target	4	Sus Señorías han solicitado un debate sobre el tema para los próximos días, en el curso de este per
Source	5	In the meantime, I should like to observe a minute' s silence, as a number of Members have reques
Target	5	A la espera de que se produzca, de acuerdo con muchos colegas que me lo han pedido, pido que ha
Source	6	Please rise, then, for this minute' s silence.
Target	6	Invito a todos a que nos pongamos de pie para guardar un minuto de silencio.
Source	7	(The House rose and observed a minute' s silence)
Target	7	(El Parlamento, de pie, guarda un minuto de silencio)
Source	8	Madam President, on a point of order.
Target	8	Señora Presidenta, una cuestión de procedimiento.
Source	9	You will be aware from the press and television that there have been a number of bomb explosions
Target	9	Sabrá usted por la prensa y la televisión que se han producido una serie de explosiones y asesinato
Source	10	One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam, who had
Target	10	Una de las personas que recientemente han asesinado en Sri Lanka ha sido al Sr. Kumar Ponnamb

2.2.1 Structure

Data alone is not informative. Only through explicit organization of the data in a way that makes relationships accessible does the data become information. This is a particularly salient hurdle in text analysis research. Some textual data is *unstructured* –that is, the relationships that will be used in the analysis have yet to be explicitly drawn and organized from the text to make the relationships meaningful and useful for analysis.

For the running text in the Europarle Corpus, we know that there are files which are the source text (original) and files that correspond to the target text (translation). In Table **??** we see that this text has been organized so that there are columns corresponding to the `type` and `sentence` with an additional `sentence_id` column to keep an index of how the sentences are aligned.

It is conventional to work with column names for datasets in R using the same conventions that are used for naming objects. It is a matter of taste which convention is used, but I have adopted snake case[51] as my personal preference. There are also alternatives[52]. Regardless of the convention you choose, it is good practice to be consistent.

It is also of note that the column names should be balanced for meaningfulness and brevity. This brevity is of practical concern but can be somewhat opaque. For questions into the meaning of the column and is values consult the resource's documentation.

Other corpus resources are *semi-structured* –that is, there are some characteristics which are structured, but other which are not.

The Switchboard Dialog Act Corpus is an example of a semi-structured resource. It has meta-data associated with each of the 1,155 conversations in the corpus. In Table **??** a language-relevant sub-set of the meta-data is associated with each utterance.

Table 8: First 5 utterances from the Switchboard Dialog Act Corpus.

doc_id	speaker_id	topic_num	topicality	naturalness	damsl_tag	speaker	utterance_num	utterance_text
4325	1632	323	3	2	o	A	1	Okay. /
4325	1632	323	3	2	qw	A	2	{D So, }
4325	1519	323	3	2	qy^d	B	1	[[I guess, +
4325	1632	323	3	2	+	A	1	What kind of exp
4325	1519	323	3	2	+	B	1	I think,] + {F u

Table 9: First 10 words from the Brown Corpus.

document_id	category	words	pos
01	A	The	AT
01	A	Fulton	NP
01	A	County	NN
01	A	Grand	JJ
01	A	Jury	NN
01	A	said	VBD
01	A	Friday	NR
01	A	an	AT
01	A	investigation	NN
01	A	of	IN

Relatively fewer resources are *structured*. In these cases a high amount of meta-data and/ or linguistic annotation is included in the corpus. The format convention, however, varies from resource to resource. Some of the formats are programming general (.csv, .xml, .json, etc.) and others are resource specific (.cha, .utt, .prd, etc.). In Table **??** the XML version of the Brown Corpus is represented in tabular format. Note that along with other meta-data variables, it also contains a variable with linguistic annotation for grammatical category (`pos` part-of-speech) of each word.

In this coursebook, the selection of the attributes from a corpus and the juxtaposition of these attributes in a relational format, or dataset, that converts data into information will be referred to as **data curation**. The process of data curation minimally involves creating a base dataset, or *derived dataset*, which establishes the main informational associations according to philosophical approach outlined by **?**. In this work, a 'tidy' dataset refers both to the structural (physical) and informational (semantic) organization of the dataset. Physically, a tidy dataset is a tabular data structure where each *row* is an observation and each *column* is a variable that contains measures of a feature or attribute of each observation. Each cell where a given row-column intersect contains a *value* which is a particular attribute of a particular observation for the particular observation-feature pair also known as a *data point*.

Semantic value in a tidy dataset is derived from the association of this physical structure along the two dimensions of this rectangular format. First, each column is a **variable** which reflects measures for a particular attribute. In the Europarle Corpus dataset, in Table **??**, for example, the `type` column measures the type of text, either `Source` or `Target`. Columns can contain measures which are qualitative or quantitative, that is character-based or numeric. Second, each row is an **observation** that contains all of the variables associated with the primary unit of observation. The primary unit of observation the variable that is the essential focus of the informational structure. In this same dataset the first observation contains the `type`, `sentence_id`, and the `sentence`. As this dataset is currently structured the primary unit of investigation is the `sentence` as each of the other variables have measures that characterize each value of `sentence`.

The decision as to what the primary unit of observation is is fundamentally guided by the research question, and therefore highly specific to the particular research project. Say instead we wanted to focus on words

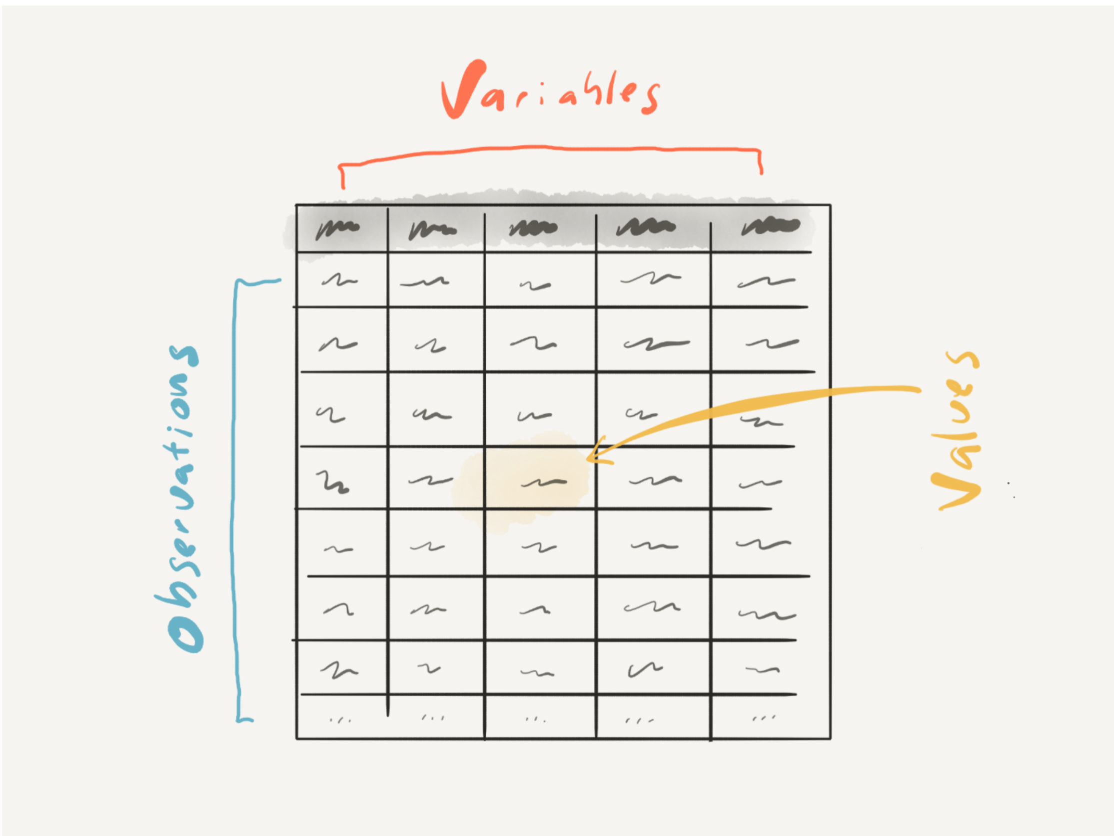

Figure 7: Visual summary of the tidy format.

Table 10: Europarle Paralle Corpus with 'words' as primary unit of investigation.

type	sentence_id	words
Source	1	Resumption
Source	1	of
Source	1	the
Source	1	session
Target	1	Reanudación
Target	1	del
Target	1	período
Target	1	de
Target	1	sesiones

instead of sentences. The dataset would need to be transformed such that a new variable (`words`) would be created to contain each word in the corpus.

The values for the variables `type` and `sentence_id` maintain the necessary description for each `word` to ensure the required semantic relationships to identify the particular attributes for each word observation. This dataset may seem redundant in that the values for `type` and `sentence_id` are repeated numerous times but this 'redundancy' makes the relationship between each variable associated with the primary unit of investigation explicit. This format makes a tidy dataset a versatile format for researchers to conduct analyses in a powerful and flexible way, as we will see throughout this coursebook.

It is important to make clear that data in tabular format in itself does not constitute a dataset, in the tidy sense we will be using. Data can be organized in many ways which do not make relationships between variables and observations explicit.

> All tabular data does not have the 'tidy' format that I have described here. Can you think of examples of tabular information that would not be in a tidy format?

2.2.2 Transformation

At this point have introduced the first step in data curation in which the original data is converted into a relational dataset (derived dataset) and highlighted the importance of this informational structure for setting the stage for data analysis. However, the primary derived dataset is often not the final organizational step before proceeding to statistical analysis. Many times, if not always, the derived dataset requires some manipulation or transformation to prepare the dataset for the specific analysis approach to be taken. This is another level of human intervention and informational organization, and therefore another step forward in our journey from data to insight and as such a step up in the DIKI hierarchy. Common types of transformations include cleaning variables (normalization), separating or eliminating variables (recoding), creating new variables (generation), or incorporating others datasets which integrate with the existing variables (merging). The results of these transformations build on and manipulate the derived dataset and produce an *analysis dataset*. Let's now turn to provide a select set of examples of each of these transformations using the datasets we have introduced in this chapter.

2.2.2.1 Normalization The process of normalization aims to *sanitize* the values within a variable or set of variables. This may include removing whitespace, punctuation, numerals, or special characters or substituting uppercase for lowercase characters, numerals for word versions, acronyms for their full forms, irregular or incorrect spelling for accepted forms, or removing common words (stopwords), etc.

On inspecting the Europarle dataset (Table **??**) we will see that there are sentence lines which do not represent actual parliment speeches. In Table **??** we see these lines.

Table 11: Non-speech lines in the Europarle dataset.

type	sentence_id	sentence
Source	1	Resumption of the session
Target	1	Reanudación del período de sesiones
Source	7	(The House rose and observed a minute' s silence)
Target	7	(El Parlamento, de pie, guarda un minuto de silencio)

Table 12: The Europarle dataset with non-speech lines removed.

type	sentence_id	sentence
Source	2	I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999,
Target	2	Declaro reanudado el período de sesiones del Parlamento Europeo, interrumpido el viernes 17 de d
Source	3	Although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the peopl
Target	3	Como todos han podido comprobar, el gran "efecto del año 2000" no se ha producido. En cambio,
Source	4	You have requested a debate on this subject in the course of the next few days, during this part-se
Target	4	Sus Señorías han solicitado un debate sobre el tema para los próximos días, en el curso de este per
Source	5	In the meantime, I should like to observe a minute' s silence, as a number of Members have reques
Target	5	A la espera de que se produzca, de acuerdo con muchos colegas que me lo han pedido, pido que ha
Source	6	Please rise, then, for this minute' s silence.
Target	6	Invito a todos a que nos pongamos de pie para guardar un minuto de silencio.
Source	8	Madam President, on a point of order.
Target	8	Señora Presidenta, una cuestión de procedimiento.
Source	9	You will be aware from the press and television that there have been a number of bomb explosions
Target	9	Sabrá usted por la prensa y la televisión que se han producido una serie de explosiones y asesinato
Source	10	One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam, who had
Target	10	Una de las personas que recientemente han asesinado en Sri Lanka ha sido al Sr. Kumar Ponnamb

A research project aiming to analyze speech would want to normalize this dataset removing these lines, as seen in Table **??**.

Another feature of this dataset which may require attention is the fact that the English lines include whitespace between possessive nouns.

This may affect another transformation process or subsequent analysis, so it may be a good idea to normalize these forms by removing the extra whitespace.

A final normalization case scenario involves changing converting all the text to lowercase. If the goal for the research is to count words at some point the fact that a word starts a sentence and by convention the first letter is capitalized will result distinct counts for words that are in essence the same (i.e. "In" vs. "in").

Note that lowercasing text, and normalization steps in general, can come at a cost. For example, lowercasing the Europarle dataset sentences means we lose potentially valuable information; namely the ability to identify

Table 13: Lines with possessives with extra whitespace in the Europarle dataset.

type	sentence_id	sentence
Source	5	In the meantime, I should like to observe a minute' s silence, as a number of Members have reques
Source	6	Please rise, then, for this minute' s silence.

Table 14: The Europarle dataset with whitespace from possessives removed.

type	sentence_id	sentence
Source	5	In the meantime, I should like to observe a minute's silence, as a number of Members have request
Source	6	Please rise, then, for this minute's silence.

Table 15: The Europarle dataset with lowercasing applied.

type	sentence_id	sentence
Source	2	i declare resumed the session of the european parliament adjourned on friday 17 december 1999, a
Target	2	declaro reanudado el período de sesiones del parlamento europeo, interrumpido el viernes 17 de dic
Source	3	although, as you will have seen, the dreaded 'millennium bug' failed to materialise, still the people
Target	3	como todos han podido comprobar, el gran "efecto del año 2000" no se ha producido. en cambio, l
Source	4	you have requested a debate on this subject in the course of the next few days, during this part-se
Target	4	sus señorías han solicitado un debate sobre el tema para los próximos días, en el curso de este perí
Source	5	in the meantime, i should like to observe a minute's silence, as a number of members have request
Target	5	a la espera de que se produzca, de acuerdo con muchos colegas que me lo han pedido, pido que ha
Source	6	please rise, then, for this minute's silence.
Target	6	invito a todos a que nos pongamos de pie para guardar un minuto de silencio.
Source	8	madam president, on a point of order.
Target	8	señora presidenta, una cuestión de procedimiento.
Source	9	you will be aware from the press and television that there have been a number of bomb explosions
Target	9	sabrá usted por la prensa y la televisión que se han producido una serie de explosiones y asesinato
Source	10	one of the people assassinated very recently in sri lanka was mr kumar ponnambalam, who had vis
Target	10	una de las personas que recientemente han asesinado en sri lanka ha sido al sr. kumar ponnambala

Table 16: Results for stemming the first words in the Brown Corpus.

document_id	category	words	pos	word_stems
01	A	The	AT	The
01	A	Fulton	NP	Fulton
01	A	County	NN	Counti
01	A	Grand	JJ	Grand
01	A	Jury	NN	Juri
01	A	said	VBD	said
01	A	Friday	NR	Fridai
01	A	an	AT	an
01	A	investigation	NN	investig
01	A	of	IN	of

proper names (i.e. "Mr Kumar Ponnambalam") and titles (i.e. "European Parliament") directly from the orthographic forms. There are, however, transformation steps that can be applied which aim to recover 'lost' information in situations such as this and others.

2.2.2.2 Recoding The process of recoding aims to *recast* the values of a variable or set of variables to a new variable or set of variables to enable more direct access. This may include extracting values from a variable, stemming or lemmatization of words, tokenization of linguistic forms (words, ngrams, sentences, etc.), calculating the lengths of linguistic units, removing variables that will not be used in the analysis, etc.

Words that we intuitively associate with a 'base' word can take many forms in language use. For example the word forms 'investigation', 'investigation', 'investigate', 'investigated', etc. are intuitively linked. There are two common methods that can be applied to create a new variable to facilitate the identification of these associations. The first is stemming. Stemming is a rule-based heuristic to reduce word forms to their stem or root form.

A few things to note here. First there are a number of stemming algorithms both for individual languages and distinct languages [53]. Second not all words can be stemmed as they do not have alternate morphological forms (i.e. "The", "of", etc.). This generally related to the distinction between closed-class (articles, prepositions, conjunctions, etc.) and open-class (nouns, verbs, adjectives, etc.) grammatical categories. Third the stem generated for those words that can be stemmed result in forms that are not words themselves. Nonetheless, stems can be very useful for more easily extracting a set of related word forms.

As an example, let's identify all the word forms for the stem 'investig'.

We can see from the results in Table **??** that searching for `word_stems` that match 'investig' returns a set of stem-related forms. But it is worth noting that these forms cut across a number of grammatical categories. If instead you want to draw a distinction between grammatical categories, we can apply lemmatization. This process is distinct from stemming in two important ways: (1) inflectional forms are grouped by grammatical category and (2) the resulting forms are lemmas or 'base' forms of words.

To appreciate the difference between stemming and lemmatization, let's compare a filter for `word_lemmas` which match 'investigation'.

Only lemma forms of 'investigate' which are nouns appear. Let's run a similar search but for the lemma 'be'.

Again only words of the same grammatical category are returned. In this case the verb 'be' has many more inflectional forms than 'investigate'.

Another form of recoding is to detect a pattern in the values of an existing variable and create a new variable whose values are the extracted pattern or register that the pattern occurs and/ or how many times

[53] https://snowballstem.org/algorithms/

Table 17: Results for filter word stems for ”investig” in the Brown Corpus.

document_id	category	words	pos	word_stems
01	A	investigation	NN	investig
01	A	investigate	VB	investig
03	A	investigation	NN	investig
03	A	investigation	NN	investig
07	A	investigations	NNS	investig
07	A	investigate	VB	investig
08	A	investigation	NN	investig
09	A	investigation	NN	investig
09	A	investigating	VBG	investig
09	A	investigation	NN	investig

Table 18: Results for lemmatization of the first words in the Brown Corpus.

document_id	category	words	pos	word_lemmas
01	A	The	AT	The
01	A	Fulton	NP	Fulton
01	A	County	NN	County
01	A	Grand	JJ	Grand
01	A	Jury	NN	Jury
01	A	said	VBD	say
01	A	Friday	NR	Friday
01	A	an	AT	a
01	A	investigation	NN	investigation
01	A	of	IN	of

Table 19: Results for filter word stems for ”investigation” in the Brown Corpus.

document_id	category	words	pos	word_lemmas
01	A	investigation	NN	investigation
03	A	investigation	NN	investigation
03	A	investigation	NN	investigation
07	A	investigations	NNS	investigation
08	A	investigation	NN	investigation
09	A	investigation	NN	investigation
09	A	investigation	NN	investigation
23	A	investigation	NN	investigation
25	A	investigation	NN	investigation
41	A	investigation	NN	investigation

Table 20: Results for filter word stems for ”be” in the Brown Corpus.

document_id	category	words	pos	word_lemmas
01	A	was	BEDZ	be
01	A	been	BEN	be
01	A	was	BEDZ	be
01	A	was	BEDZ	be
01	A	are	BER	be
01	A	are	BER	be
01	A	be	BE	be
01	A	is	BEZ	be
01	A	was	BEDZ	be
01	A	be	BE	be

Table 21: Disfluency counts in the first 10 utterance text values from the Switchboard Corpus.

utterance_text	disfluency_count
Okay. /	0
{D So, }	0
[[I guess, +	0
What kind of experience [do you, + do you] have, then with child care? /	0
I think,] + {F uh, } I wonder] if that worked. /	1
Does it say something? /	0
I think it usually does. /	0
You might try, {F uh, } /	1
I don’t know, /	0
hold it down a little longer, /	0

it occurs. As an example, let’s count the number of disfluencies (‘uh’ or ‘um’) that occur in each utterance in `utterance_text` from the Switchboard Dialog Act Corpus. *Note I’ve simplified the dataset dropping the non-relevant variables for this example.*

One of the most common forms of recoding in text analysis is tokenization. Tokenization is the process of recasting the text into smaller linguistic units. When working with text that has not been linguistically annotated, the most feasible linguistic tokens are words, ngrams, and sentences. While word and sentence tokens are easily understandable, ngram tokens need some explanation. An ngram is a sequence of either characters or words where n is the length of this sequence. The ngram sequences are drawn incrementally, so the bigrams (two-word sequences) for the sentence “This is an input sentence.” are:

this is, is an, an input, input sentence

We’ve already seen word tokenization exemplified with the Europarle Corpus in subsection Structure in Table **??**, so let’s create (word) bigram tokens for this corpus.

As I just mentioned, ngrams sequences can be formed of characters as well. Here are character trigram (three-character) sequences.

2.2.2.3 Generation The process of generation aims to *augment* a variable or set of variables. In essence this aims to make implicit attributes explicit to that they are directly accessible. This often targeted at the automatic generation of linguistic annotations such as grammatical category (part-of-speech) or syntactic structure.

Table 22: The first 10 word bigrams of the Europarle Corpus.

type	sentence_id	word_bigrams
Source	2	i declare
Source	2	declare resumed
Source	2	resumed the
Source	2	the session
Source	2	session of
Source	2	of the
Source	2	the european
Source	2	european parliament
Source	2	parliament adjourned
Source	2	adjourned on

Table 23: The first 10 character trigrams of the Europarle Corpus.

type	sentence_id	char_trigrams
Source	2	ide
Source	2	dec
Source	2	ecl
Source	2	cla
Source	2	lar
Source	2	are
Source	2	rer
Source	2	ere
Source	2	res
Source	2	esu

Table 24: Automatic linguistic annotation for grammatical category and syntactic structure for an example English sentence from the Europarle Corpus

type	sentence_id	token_id	token	upos	feats	token_id_source	syntactic_r
Target	6	1	Invito	ADP	NA	3	case
Target	6	2	a	DET	Definite=Ind|PronType=Art	3	det
Target	6	3	todos	NOUN	Number=Plur	6	nmod
Target	6	4	a	DET	Definite=Ind|PronType=Art	6	det
Target	6	5	que	ADJ	Degree=Pos	6	amod
Target	6	6	nos	NOUN	Number=Plur	0	root
Target	6	7	pongamos	X	NA	13	goeswith
Target	6	8	de	X	Foreign=Yes	13	goeswith
Target	6	9	pie	X	NA	13	goeswith
Target	6	10	para	X	NA	13	goeswith
Target	6	11	guardar	X	NA	13	goeswith
Target	6	12	un	X	NA	13	goeswith
Target	6	13	minuto	NOUN	Number=Sing	6	appos
Target	6	14	de	PROPN	Number=Sing	15	compound
Target	6	15	silencio	PROPN	Number=Sing	13	flat
Target	6	16	.	PUNCT	NA	6	punct

In the examples below I've added linguistic annotation to a target (English) and source (Spanish) example sentence from the Europarle Parallel Corpus. First, note the variables that are added to our dataset that correspond to grammatical category. In addition to the `type` and `sentence_id` we have an assortment of variables which replace the `sentence` variable. As part of the process of annotation the input text to be annotated `sentence` is tokenized `token` and indexed `token_id`. Then `upos` contains the Universal Part of Speech tags[54], and a detailed list of features is included in `feats`. The syntactic annotation is reflected in the `token_id_source` and `syntactic_relation` variables. These variables correspond to the type of syntactic parsing that has been done, in this case Dependency Parsing (using the Universal Dependencies[56] framework). Another common syntactic parsing framework is phrase constituency parsing (**?**).

Now compare the English example sentence dataset in Table **??** with the parallel sentence in Spanish. Note that the grammatical features are language specific. For example, Spanish has gender which is apparent when scanning the `feats` variable.

There is much more to explore with linguistic annotation, and syntactic parsing in particular, but at this point it will suffice to note that it is possible to augment a dataset with grammatical information automatically.

There are strengths and shortcomings with automatic linguistic annotation that a research should be aware of. First, automatic linguistic annotation provides quick access to rich and highly reliable linguistic information for a large number of languages. However, part of speech taggers and syntactic parsers are not magic. They are resources that are built by training a computational algorithm to recognize patterns in manually annotated datasets producing a language model. This model is then used to predict the linguistic annotations for new language (as we just did in the previous examples). The shortcomings of automatic linguistic annotation is first, not all languages have trained language models and second, the data used to train the model inevitably reflect a particular variety, register, modality, etc. The accuracy of the linguistic annotation is highly dependent on alignment between the language sampling frame of the trained data and the language data to be automatically annotated. Many (most) of the language models available for automatic linguistic annotation are based on language that is most readily available and for most languages this has traditionally been newswire text. It is important to be aware of these characteristics when using linguistic annotation

[54]Descriptions of the UPOS tagset[55]
[56]https://universaldependencies.org/

Table 25: Automatic linguistic annotation for grammatical category and syntactic structure for an example Spanish sentence from the Europarle Corpus

type	sentence_id	token_id	token	upos	feats
Source	6	1	Please	PROPN	Gender=Fem|Number=Sing
Source	6	2	rise	PROPN	Number=Sing
Source	6	3	,	PUNCT	NA
Source	6	4	then	VERB	Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin
Source	6	5	,	PUNCT	NA
Source	6	6	for	ADP	NA
Source	6	7	this	X	NA
Source	6	8	minute's	X	Gender=Masc|Number=Sing
Source	6	9	silence	X	Gender=Masc|Number=Sing
Source	6	10	.	PUNCT	NA

tools.

2.2.2.4 Merging The process of merging aims to *join* a variable or set of variables with another variable or set of variables from another dataset. The option to merge two (or more) datasets requires that there is a shared variable that indexes and aligns the datasets.

To provide an example let's look at the Switchboard Diaglog Act Corpus. Our existing, disfluency recoded, version includes the following variables.

```
#> Rows: 5
#> Columns: 11
#> $ doc_id          <chr> "4325", "4325", "4325", "4325", "4325"
#> $ speaker_id      <dbl> 1632, 1632, 1519, 1632, 1519
#> $ topic_num       <dbl> 323, 323, 323, 323, 323
#> $ topicality      <chr> "3", "3", "3", "3", "3"
#> $ naturalness     <chr> "2", "2", "2", "2", "2"
#> $ damsl_tag       <chr> "o", "qw", "qy^d", "+", "+"
#> $ speaker         <chr> "A", "A", "B", "A", "B"
#> $ turn_num        <chr> "1", "1", "2", "3", "4"
#> $ utterance_num   <chr> "1", "2", "1", "1", "1"
#> $ utterance_text  <chr> "Okay.  /", "{D So, }", "[ [ I guess, +", "What kind ~
#> $ disfluency_count <int> 0, 0, 0, 0, 1
```

It turns out that on the corpus website[57] a number of meta-data files are available, including files pertaining to speakers and the topics of the conversations.

The speaker meta-data for this corpus is in a the `caller_tab.csv` file and contains a `speaker_id` variable which corresponds to each speaker in the corpus and other potentially relevant variables for a language research project including `sex`, `birth_year`, `dialect_area`, and `education`.

Since both datasets contain a shared index, `speaker_id` we can merge these two datasets. The result is found in Table **??**.

In this example case the dataset that was merged was already in a structured format (.csv). Many corpus resources contain meta-data in stand-off files that are structured.

In some cases a researcher would like to merge information that does not already accompany the corpus resource. This is possible as long as a dataset can be created that contains a variable that is shared.

[57] https://catalog.ldc.upenn.edu/docs/LDC97S62/

Table 26: Speaker meta-data for the Switchboard Dialog Act Corpus.

speaker_id	sex	birth_year	dialect_area	education
1632	FEMALE	1962	WESTERN	2
1632	FEMALE	1962	WESTERN	2
1519	FEMALE	1971	SOUTH MIDLAND	1
1632	FEMALE	1962	WESTERN	2
1519	FEMALE	1971	SOUTH MIDLAND	1

Table 27: Merged conversations and speaker meta-data for the Switchboard Dialog Act Corpus.

doc_id	speaker_id	sex	birth_year	dialect_area	education	topic_num	topicality	naturalness
4325	1632	FEMALE	1962	WESTERN	2	323	3	2
4325	1632	FEMALE	1962	WESTERN	2	323	3	2
4325	1519	FEMALE	1971	SOUTH MIDLAND	1	323	3	2
4325	1632	FEMALE	1962	WESTERN	2	323	3	2
4325	1519	FEMALE	1971	SOUTH MIDLAND	1	323	3	2

Without a shared variable to index the datasets the merge cannot take place.

In sum, the transformation steps described here collectively aim to produce higher quality datasets that are relevant in content and structure to submit to analysis. The process may include one or more of the previous transformations but is rarely linear and is most often iterative. It is typical to do some normalization then generation, then recoding, and then return to normalizing, and so forth. This process is highly idiosyncratic given the characteristics of the derived dataset and the ultimate goals for the analysis dataset.

> Note in some cases we may convert our tidy tabular dataset to other data formats that may be required for some particular statistic approaches but at all times the relationship between the variables should be maintained in line with our research purpose. We will touch on examples of other types of data formats (e.g. Corpus and Document-Term Matrix (DTM) objects in R) when we dive into particular statistical approaches that require them later in the coursebook.

2.3 Documentation

As we have seen in this chapter that acquiring data and converting that data into information involves a number of conscious decisions and implementation steps. As a favor to ourselves as researchers and to the research community, it is crucial to document these decisions and steps. This makes it both possible to retrace our own steps and also provides a guide for future researchers that want to reproduce and/ or build on your research. A programmatic approach to quantitative research helps ensure that the implementation steps are documented and reproducible but it is also vital that the decisions that are made are documented as well. This includes the creation/ selection of the corpus data, the description of the variables chosen from the corpus for the derived dataset, and the description of the variables created from the derived dataset for the analysis dataset.

For an existing corpus sample acquired from a repository (e.g. Switchboard Dialog Act Corpus[58], Language Data Consortium), a research group (e.g. CEDEL2[59]), or an individual researcher (e.g. SMS Spam Collection[60]), there is often documentation provided describing key attributes of the resource. This documentation

[58] https://catalog.ldc.upenn.edu/LDC97S62
[59] http://cedel2.learnercorpora.com/
[60] https://www.dt.fee.unicamp.br/~tiago/smsspamcollection/

should be included with the acquisition of the corpus and added to the research project. For a corpus that a researcher compiles themselves, they will need to generate this documentation.

The curation and transformation steps conducted on the original corpus data to produce the datasets should also be documented. The steps themselves can be included in the programming scripts as code comments (or in prose if using a literate programming strategy (e.g. RMarkdown)). The structure of each resulting dataset should include what is called a **data dictionary**. This is a table which includes the variable names, the values they contain, and a short prose description of each variable (e.g. ACTIV-ES Corpus[61]).

Summary

In this chapter we have focused on data and information –the first two components of DIKI Hierarchy. This process is visualized in Figure **??**.

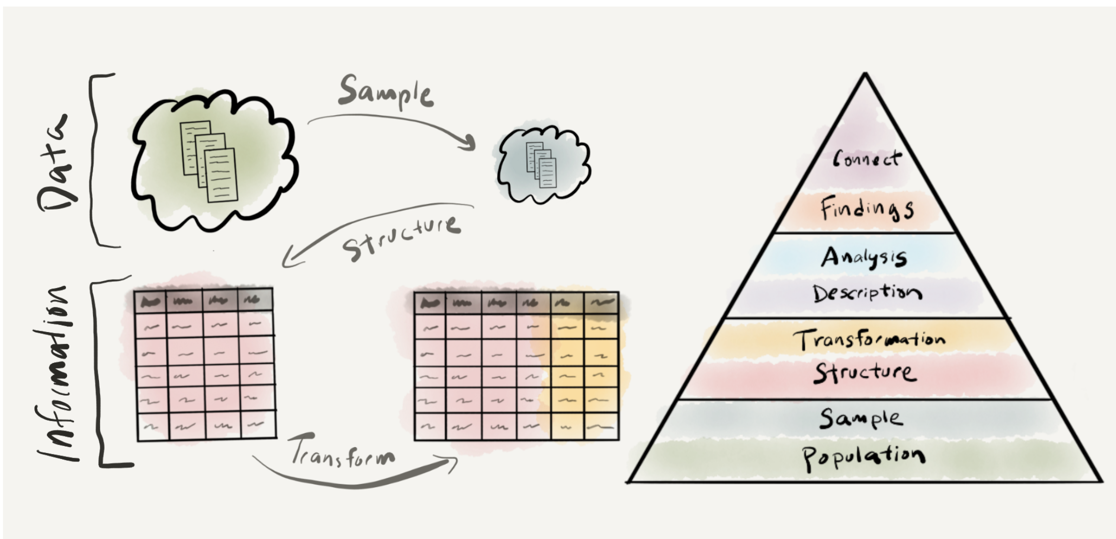

Figure 8: Understanding data: visual summary

First a distinction is made between populations and samples, the latter being a intentional and subjective selection of observations from the world which attempt to represent the population of interest. The result of this process is known as a corpus. Whether developing a corpus or selecting an existing a corpus it is important to vet the sampling frame for its applicability and viability as a resource for a given research project.

Once a viable corpus is identified, then that corpus is converted into a derived dataset which adopts the 'tidy' dataset format where each column is a variable, each row is an observation, and the intersection of columns and rows contain values. This derived dataset serves to establish the base informational relationships from which your research will stem.

The derived dataset will most likely require transformations including normalization, recoding, generation, and/ or merging to enhance the usefulness of the information to analysis. An analysis dataset is the result of this process.

Although covered at the end of this chapter, documentation should be implemented at each stage of the process. Employing a programmatic approach establishes documentation of the implementation steps but the motivation behind the decisions taken and the content of the corpus data and datasets generated also need documentation to ensure transparent and reproducible research.

[61] https://osf.io/9jafz/

3 Approaching analysis

DRAFT

> Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.
>
> — H.G. Wells

The essential questions for this chapter are:

- What is the role of statistics in data analysis?
- What is the importance of descriptive assessment in data analysis?
- In what ways are main approaches to data analysis similar and different?

In this chapter I will build on the notions of data and information from the previous chapter. The aim of statistics in quantitative analysis is to uncover patterns in datasets. Thus statistics is aimed at deriving knowledge from information, the next step in the DIKI Hierarchy (Figure **??**). Where the creation of information from data involves human intervention and conscious decisions, as we have seen, deriving knowledge from information involves even more conscious subjective decisions on what information to assess, and what method to select to interrogate the information, and ultimately how to interpret the findings. The first step is to conduct a descriptive assessment of the information, both at the individual variable level and also between variables, the second is to interrogate the dataset either through inferential, predictive, or exploratory analysis methods, and the third is to interpret and report the findings.

3.1 Description

A descriptive assessment of the dataset includes a set of diagnostic measures and tabular and visual summaries which provide researchers a better understanding of the structure of a dataset, prepare the researcher to make decisions about which statistical methods and/ or tests are most appropriate, and to safeguard against false assumptions (missing data, data distributions, etc.). In this section we will first cover the importance of understanding the informational value that variables can represent and then move to use this understanding to approach summarizing individual variables and relationships between variables.

To ground this discussion I will introduce a new dataset. This dataset is drawn from the Barcelona English Language Corpus (BELC)[62], which is found in the TalkBank repository[63]. I've selected the "Written composition" task from this corpus which contains writing samples from second language learners of English at different ages. Participants were given the task of writing for 15 minutes on the topic of "Me: my past, present and future". Data was collected for many (but not all) participants up to four times over the course of seven years. In Table **??** I've included the first 10 observations from the dataset which reflects structural and transformational steps I've done so we start with a tidy dataset.

The entire dataset includes 79 observations from 36 participants. Each observation in the BELC dataset corresponds to an individual learner's composition. It includes which participant wrote the composition (`participant_id`), the age group they were part of at the time (`age_group`), their sex (`sex`), and the number of English words they produced (`num_tokens`), the number of unique English words they produced (`num_types`). The final variable (`ttr`) is the calculated ratio of number of unique words (`num_types`) to total words (`num_tokens`) for each composition. This is known as the Type-Token Ratio and it is a standard metric for measuring lexical diversity.

3.1.1 Information values

Understanding the informational value, or **level of measurement**, of a variable or set of variables in key to preparing for analysis as it has implications for what visualization techniques and statistical measures we

[62] https://slabank.talkbank.org/access/English/BELC.html
[63] http://talkbank.org/

Table 28: First 10 observations of the BELC dataset for demonstration.

participant_id	age_group	sex	num_tokens	num_types	ttr
L02	10-year-olds	female	12	12	1.000
L05	10-year-olds	female	18	15	0.833
L10	10-year-olds	female	36	26	0.722
L11	10-year-olds	female	10	8	0.800
L12	10-year-olds	female	41	23	0.561
L16	10-year-olds	female	13	12	0.923
L22	10-year-olds	female	47	30	0.638
L27	10-year-olds	female	8	8	1.000
L28	10-year-olds	female	84	34	0.405
L29	10-year-olds	female	53	34	0.642

can use to interrogate the dataset. There are two main levels of measurement a variable can take: categorical and continuous. **Categorical variables** reflect class or group values. **Continuous variables** reflect values that are measured along a continuum.

The BELC dataset contains three categorical variables (`participant_id`, `age_group`, and `sex`) and three continuous variables (`num_tokens`, `num_types`, and `ttr`). The categorical variables identify class or group membership; which participant wrote the composition, what age group they were in, and their biological sex. The continuous variables measure attributes that can take a range of values without a fixed limit and the differences between each value are regular. The number of words and number of unique words for each composition can range from 1 to n and the Type-Token Ratio being derived from these two variables is also continuous for the same reason. Furthermore, the differences between the each of values of these measures is on a defined interval, so for example a composition which has a word count (`num_tokens`) of 40 is exactly two times as large as a composition with a word count of 20.

The distinction between categorical an continuous levels of measurement, as mentioned above, are the main two and for some statistical approaches the only distinction that needs to be made to conduct an analysis. However, categorical and continuous can each be broken down into subcategories and for some descriptive and analytic purposes these distinctions are important. For categorical variables a distinction can be made between variables in which there is a structured relationship between the values and those in which there is not. *Nominal variables* contain values which are labels denoting the membership in a class in which there is no relationship between the labels. *Ordinal variables* also contain labels of classes, but in contrast to nominal variables, there is a relationship between the classes, namely one in which there is a precedence relationship or order. With this in mind, our categorical variables be sub-classified. There is no order between the values of `participant_id` and `sex` and they are therefore nominal whereas the values of `age_group` are ordered, each value refers to a sequential age group, and therefore it is ordinal.

Turning to continuous variables, another subdivision can be made which hinges on the existence of a non-arbitrary zero or not. *Interval variables* contain values in which the difference between the values is regular and defined, but the measure has an arbitrary zero value. A typically cited example of an interval variable is temperature measurements on the Fahrenheit scale. A value of 0 on this scale does not mean there is 0 temperature. *Ratio variables* have all the properties of interval variables but also include a non-arbitrary definition of zero. All of the continuous variables in the BELC dataset (`num_tokens`, `num_types`, and `ttr`) are ratio variables as a value of 0 would indicate the lack of this attribute.

An hierarchical overview of the relationship between the two main and four sub-types of levels of measurement appear in Figure **??**.

A few notes of practical importance; First, the distinction between interval and ratio variables is often not applicable in text analysis and therefore often treated together as continuous variables. Second, the distinction between ordinal and interval/continuous variables is not as clear cut as it may seem. Both

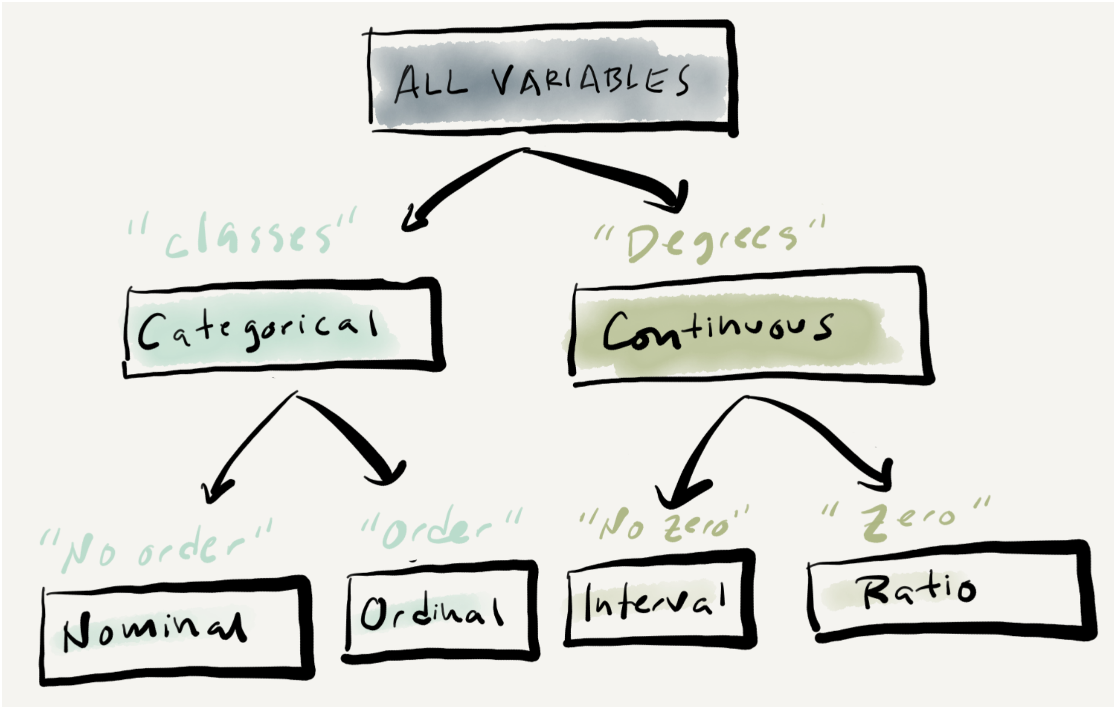

Figure 9: Levels of measurement graphic representation.

variables contain values which have an ordered relationship. By definition the values of an ordinal variable do not reflect regular intervals between the units of measurement. But in practice interval/continuous variables with a defined number of values (say from a Likert scale used on a survey) may be treated as an ordinal variable as they may be better understood as reflecting class membership. Third, all continuous variables can be converted to categorical variables, but the reverse is not true. We could, for example, define a criterion for binning the word counts in `num_tokens` for each composition into ordered classes such as "low", "mid", "high". On the other hand, `sex` (as it has been measured here) cannot take intermediate values on a unfixed range. The upshot is that variables can be down-typed but not up-typed. In most cases it is preferred to treat continuous variables as such, if the nature of the variable permits it, as the down-typing of continuous data to categorical data results in a loss of information –which will result in a loss of information and hence statistical power which may lead to results that obscure meaningful patterns in the data (**?**).

3.1.2 Summaries

It is always key to gain insight into shape of the information through numeric, tabular and/ or visual summaries before jumping in to analytic statistical approaches. The most appropriate form of summarizing information will depend on the number and informational value(s) of our target variables. To get a sense of how this looks, let's continue to work with the BELC dataset and pose different questions to the data with an eye towards seeing how various combinations of variables are descriptively explored.

3.1.2.1 Single variables The way to statistically summarize a variable into a single measure is to derive a **measure of central tendency**. For a continuous variable the most common measure is the (arithmetic) *mean*, or average, which is simply the sum of all the values divided by the number of values. As a measure of central tendency, however, the mean can be less-than-reliable as it is sensitive to outliers which is to say that data points in the variable that are extreme relative to the overall distribution of the other values in the variable affect the value of the mean depending on how extreme the deviate. One way to assess the effects of outliers is to calculate a **measure of dispersion**. The most common of these is the *standard deviation* which estimates the average amount of variability between the values in a continuous variable. Another way to assess, or rather side-step, outliers is to calculate another measure of central tendency, the *median*. A median is calculated by sorting all the values in the variable and then selecting the value which falls in the middle of all the other values. A median is less sensitive to outliers as extreme values (if there are few) only indirectly affect the selection of the middle value. Another measure of dispersion is to calculate quantiles. A *quantile* slices the data in four percentile ranges providing a five value numeric summary of the spread of the values in a continuous variable. The spread between the first and third quantile is known as the Interquartile Range (IQR) and is also used as a single statistic to summarize variability between values in a continuous variable.

Below is a list of central tendency and dispersion scores for the continuous variables in the BELC dataset.

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	iqr
num_tokens	0	1	66.23	43.90	1.00	29.00	55.00	90.00	185	61.00
num_types	0	1	40.25	22.80	1.00	22.00	38.00	54.00	97	32.00
ttr	0	1	0.67	0.13	0.41	0.57	0.64	0.73	1	0.16

The descriptive statistics returned above were generated by the `skimr` package.

In the above summary, we see the mean, standard deviation (sd), and the quantiles (the five-number summary, p0, p25, p50, p75, and p100). The middle quantile (p50) is the median and the IQR is listed last.

These are important measures for assessing the central tendency and dispersion and will be useful for reporting purposes, but to get a better feel of how a variable is distributed, nothing beats a visual summary. A boxplot graphically summarizes many of these metrics. In Figure **??** we see the same three continuous variables, but now in graphical form.

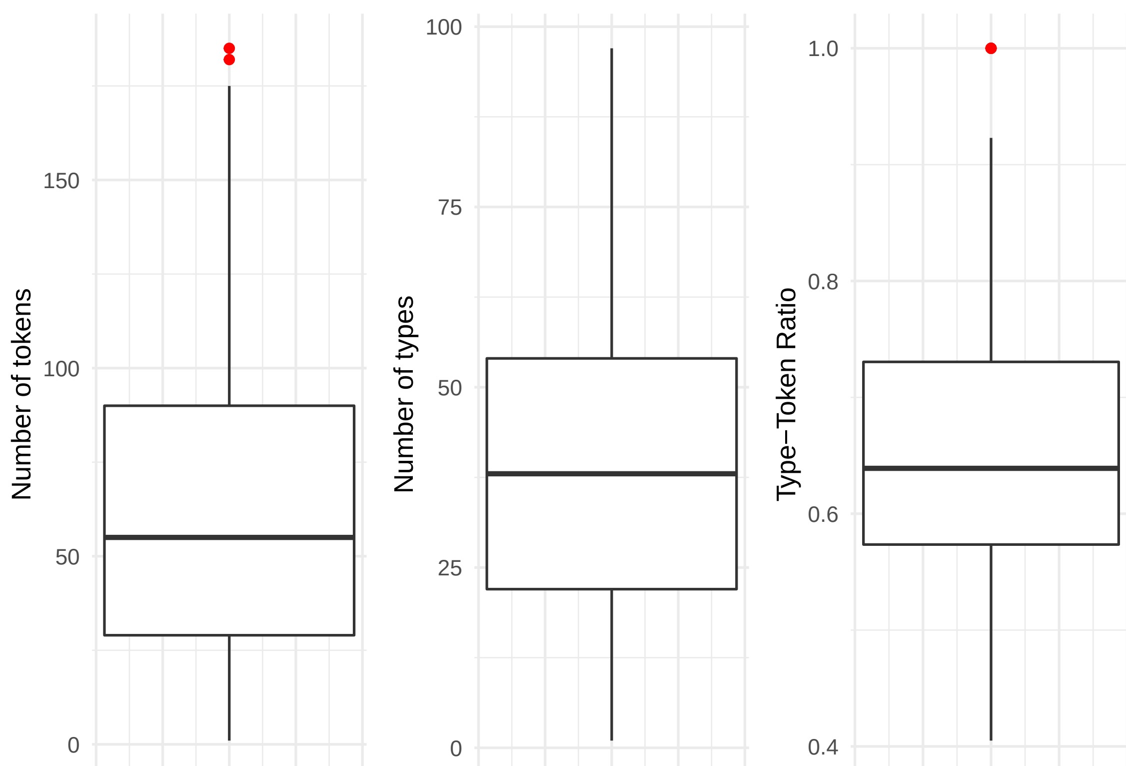

Figure 10: Boxplots for each of the continuous variables in the BELC dataset.

In a boxplot, the bold line is the median. The surrounding box around the median is the interquantile range.

The extending lines above and below the IQR mark the largest and lowest value that is within 1.5 times either the 3rd (top of the box) or 1st (bottom of the box). Any values that fall outside, above or below, the extending lines are considered statistical outliers and are marked as dots (in this case red dots). [64]

Boxplots provide a robust and visually intuitive way of assessing central tendency and variability in a continuous variable but this type of plot can be complemented by looking at the overall distribution of the values in terms of their frequencies. A histogram provides a visualization of the frequency (and density in this case with the blue overlay) of the values across a continuous variable binned at regular intervals.

In Figure **??** I've plotted histograms in the top row and density plots in the bottom row for the same three continuous variables from the BELC dataset.

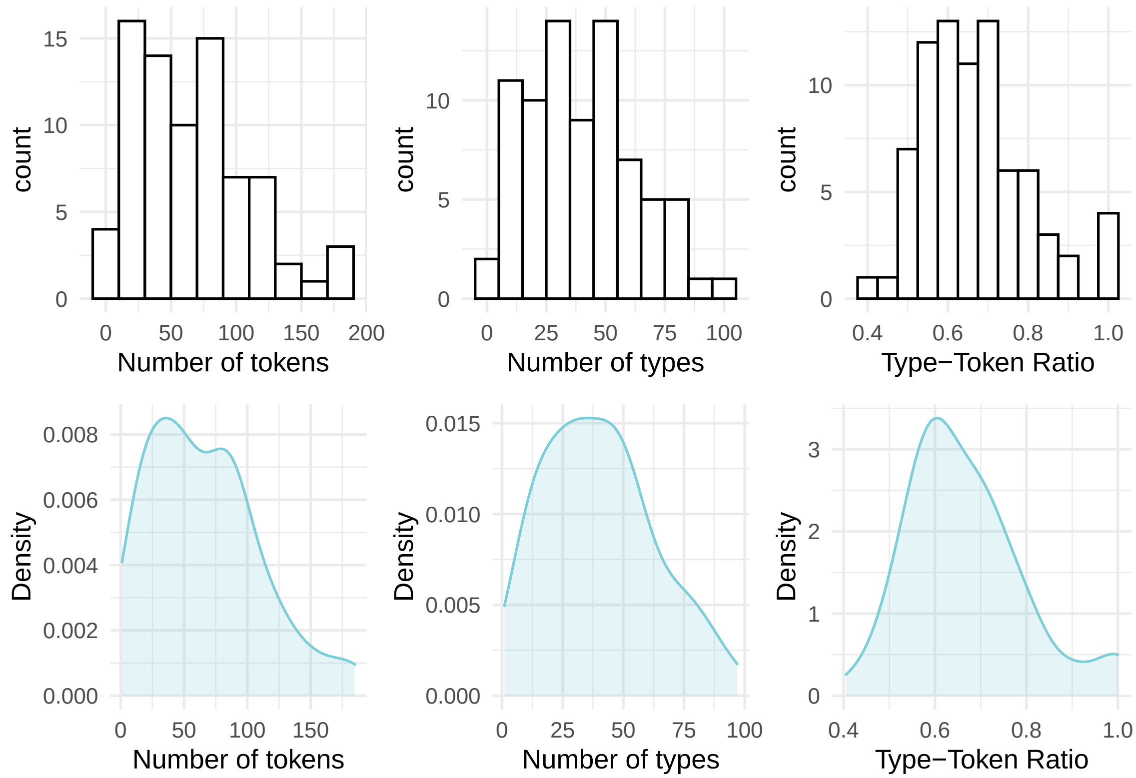

Figure 11: Histograms and density plots for the continuous variables in the BELC dataset.

Histograms provide insight into the distribution of the data. For our three continuous variables, the distributions happen not to be too strikingly distinct. They are, however, not the same either. When we explore continuous variables with histograms we are often trying to assess whether there is skew or not. There are three general types of skew, visualized in Figure **??**.

In histograms/ density plots in which the distribution is either left or right, the median and the mean are not aligned. The *mode*, which indicates the most frequent value in the variable is also not aligned with the other two measures. In a left-skewed distribution the mean will be to the left of the median which is left of the mode whereas in a right-skewed distribution the opposite occurs. In a distribution with absolutely no skew these three measures are the same. In practice these measures rarely align perfectly but it is very typical for these three measures to approximate alignment. It is common enough that this distribution is called the Normal Distribution [65] as it is very common in real-world data.

[64] Note that each of these three variables are to be considered separately here (vertically). Later we will see the use of boxplots to compare a continuous variable across levels of a categorical variable (horizontally).

[65] formally known as a Gaussian Distribution

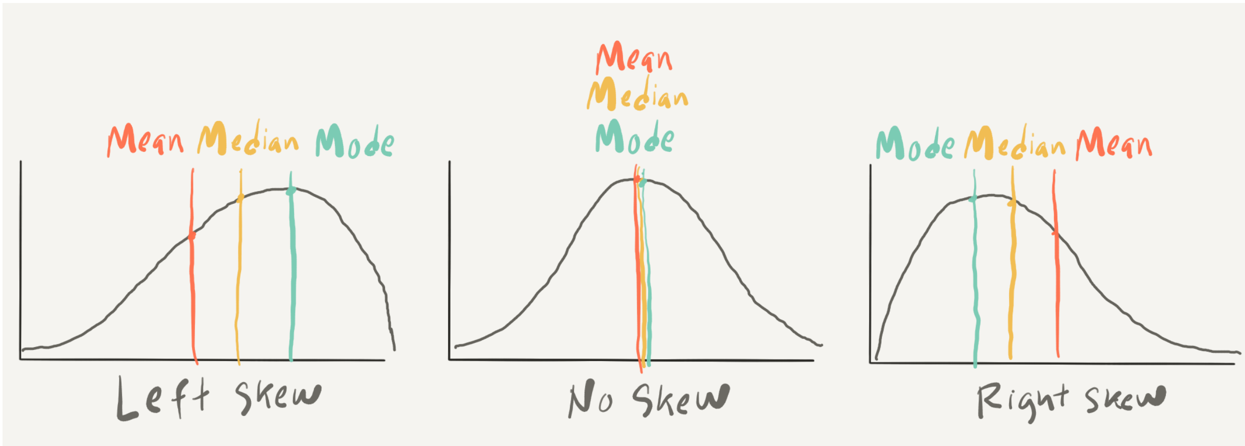

Figure 12: Examples of skew types in density plots.

Table 29: Results from Shapiro-Wilk test of normality for continuous variables in the BELC dataset.

variable	statistic	p_value
Number of tokens	0.942	0.001
Number of types	0.970	0.058
Type-Token Ratio	0.947	0.003

Another and potentially more informative way to inspect the normality of a distribution is to create Quantile-Quantile plots (QQ Plot). In Figure **??** I've created QQ plots for our three continuous variables. The line in each plot is the normal distribution and the more points that fall off of this line, the less likely that the distribution is normal.

A visual inspection can often be enough to detect non-normality, but in cases which visually approximate the normal distribution (such as these) we can perform the Shapiro-Wilk test of normality. This is an inferential test that compares a variable's distribution to the normal distribution. The likelihood that the distribution differs from the normal distribution is reflected in a p-value. A p-value below the .05 threshold suggests the distribution is non-normal. In Table **??** we see that given this criterion only the distribution of `num_types` is normally distributed.

Downstream in the analytic analysis, the distribution of continuous variables will need to be taken into account for certain statistical tests. Tests that assume 'normality' are parametric tests, those that do not are non-parametric. Distributions which approximate the normal distribution can sometimes be transformed to conform to the normal distribution either by outlier trimming or through statistical procedures (e.g. square root, log, or inverse transformation), if necessary. At this stage, however, the most important thing is to recognize whether the distributions approximate or wildly diverge from the normal distribution.

Before we leave continuous variables, let's consider another approach for visually summarizing a single continuous variable. The Empirical Cumulative Distribution Frequency, or *ECDF*, is a summary of the cumulative proportion of each of the values of a continuous variable. An ECDF plot can be useful in determining what proportion of the values fall above or below a certain percentage of the data.

In Figure **??** we see ECDF plots for our three continuous variables.

Take, for example, the number of tokens (`num_tokens`) per composition. The ECDF plot tells us that 50% of the values in this variable are 56 words or less. In the three variables plotted, the cumulative growth is quite steady. In some cases it is not. When it is not, an ECDF goes a long way to provide us a glimpse into key bends in the proportions of values in a variable.

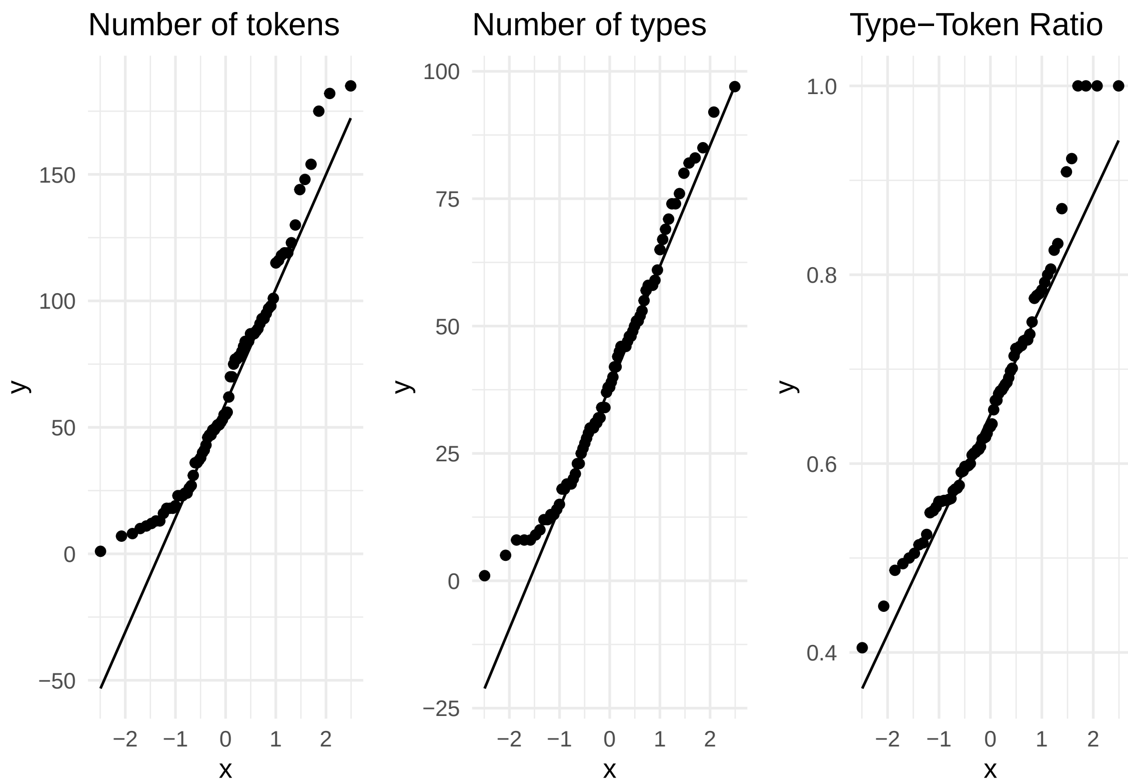

Figure 13: QQ Plots for the continuous variables in the BELC dataset.

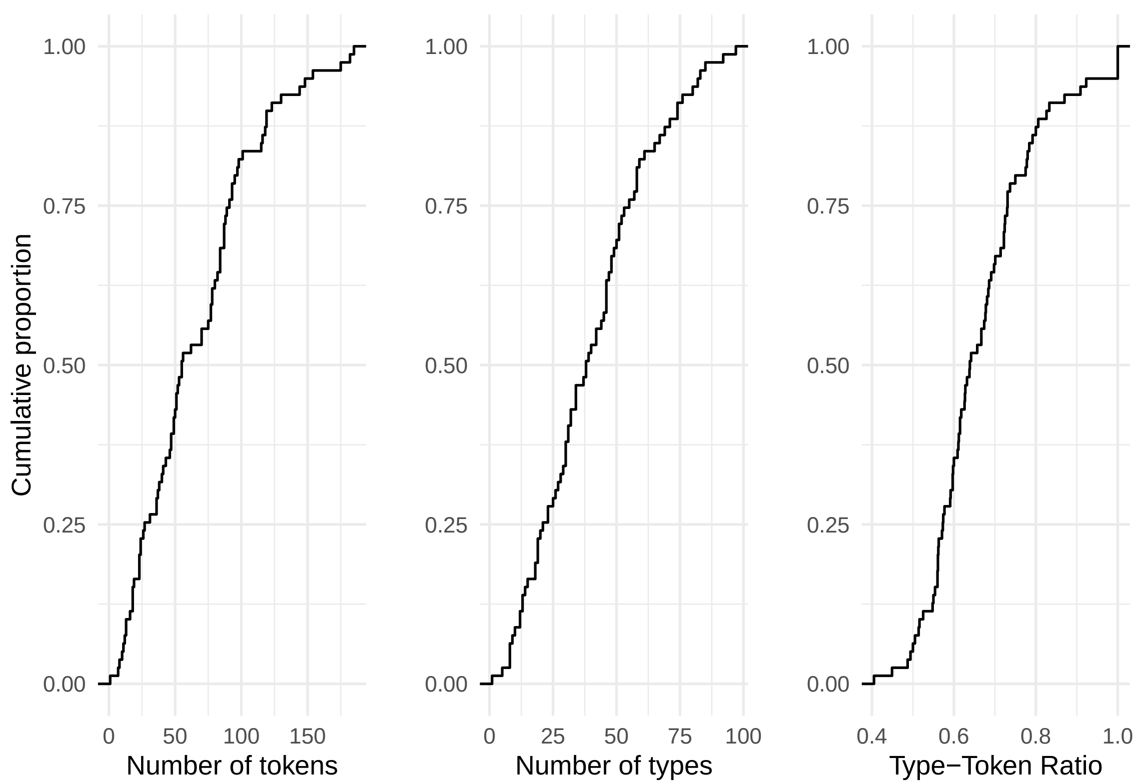

Figure 14: ECDF plots for the continuous variables in the BELC dataset.

Now let's turn to the descriptive assessment of categorical variables. For categorical variables, central tendency can be calculated as well but only a subset of measures given the reduced informational value of categorical variables. For nominal variables where there is no relationship between the levels the central tendency is simply the mode. The levels of ordinal variables, however, are relational and therefore the median, in addition to the mode, can also be used as a measure of central tendency. Note that a variable with one mode is unimodal, two modes, bimmodal, and in variables that have two or more modes multimodal.

> To get numeric value of the median for an ordinal variable the levels of the variable will need to be numeric as well. Non-numeric levels can be recoded to numeric for this purpose if necessary.

Below is a list of the central tendency metrics for the categorical variables in the BELC dataset.

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
participant_id	0	1	FALSE	36	L05: 3, L10: 3, L11: 3, L12: 3
age_group	0	1	TRUE	4	10-: 24, 16-: 24, 12-: 16, 17-: 15
sex	0	1	FALSE	2	fem: 48, mal: 31

In practice when a categorical variable has few levels it is common to simply summarize the counts of each level in a table to get an overview of the variable. With ordinal variables with more numerous levels, the five-score summary (quantiles) can be useful to summarize the distribution. In contrast to continuous variables where a graphical representation is very helpful to get perspective on the shape of the distribution of the values, the exploration of single categorical variables is rarely enhanced by plots.

3.1.2.2 Multiple variables In addition to the single variable summaries (univariate), it is very useful to understand how two (bivariate) or more variables (multivariate) are related to add to our understanding of the shape of the relationships in the dataset. Just as with univariate summaries, the informational values of the variables frame our approach.

To explore the relationship between two continuous variables we can statistically summarize a relationship with a **coefficient of correlation** which is a measure of **effect size** between continuous variables. If the continuous variables approximate the normal distribution *Pearson's r* is used, if not *Kendall's tau* is the appropriate measure. A correlation coefficient ranges from -1 to 1 where 0 is no correlation and -1 or 1 is perfect correlation (either negative or positive). Let's assess the correlation coefficient for the variables `num_tokens` and `ttr`. Since these variables are not normally distributed, we use Kendall's tau. Using this measure the correlation coefficient is -0.563 suggesting there is a correlation, but not a particularly strong one.

Correlation measures are important for reporting but to really appreciate a relationship it is best to graphically represent the variables in a *scatterplot*. In Figure **??** we see the relationship between `num_tokens` and `ttr`.

In both plots `ttr` is on the y-axis and `num_tokens` on the x-axis. The points correspond to the intersection between these variables for each single observation. In the left pane only the points are represented. Visually (and given the correlation coefficient) we can see that there is a negative relationship between the number of tokens and the Type-Token ratio: in other words, the more tokens a composition has the lower the Type-Token Ratio. In this case this trend is quite apparent, but in other cases is may not be. To provide an additional visual cue a trend line is often added to a scatterplot. In the right pane I've added a linear trend line. This line demarcates the optimal central tendency across the relationship, assuming a linear relationship. The steeper the line, or slope, the more likely the correlation is strong. The band, or ribbon, around this trend line indicates the **confidence interval** which means that real central tendency could fall anywhere within this space. The wider the ribbon, the larger the variation between the observations. In this case we see that the ribbon widens when the number of tokens is either low or high. This means that the trend line could be potentially be drawn either steeper (more strongly correlated) or flatter (less strongly correlated).

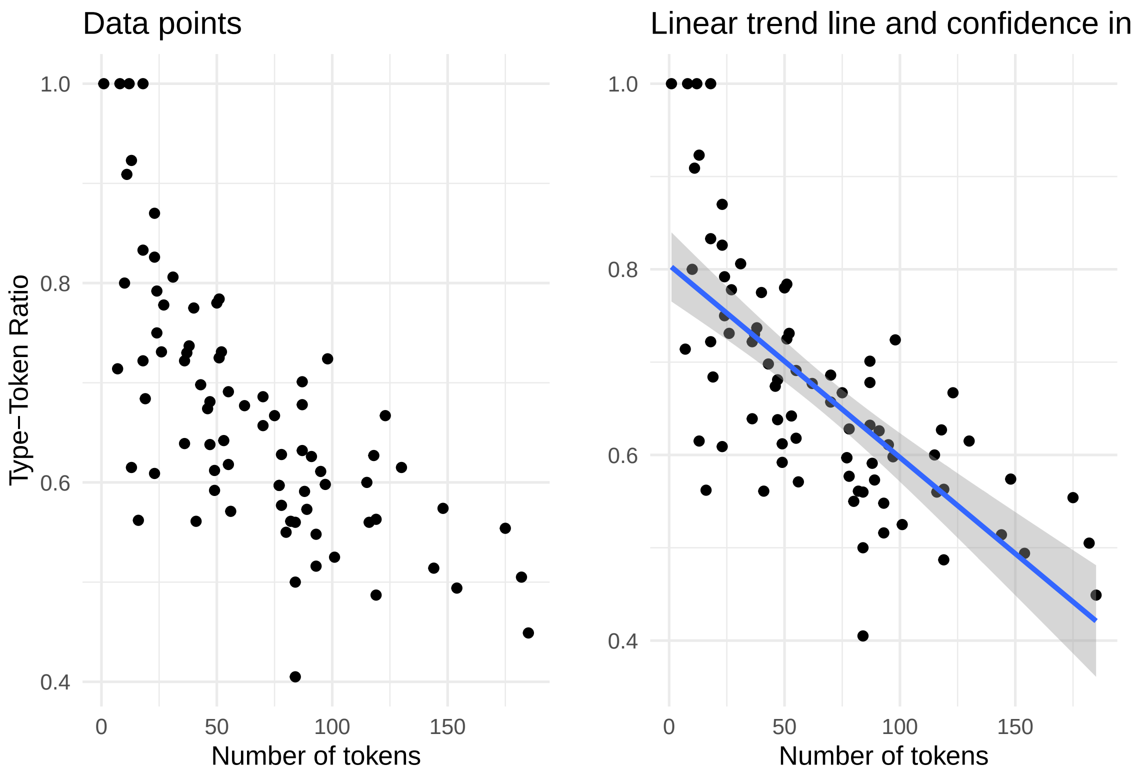

Figure 15: Scatterplot...

In plots comparing two or more variables, the choice of which variable to plot on the x- and y-axis is contingent on the research question and/ or the statistical approach. The language varies between statistical approaches: in inferential methods the x-axis is used to plot what is known as the dependent variable and the y-axis an independent variable. In predictive methods the dependent variable is known as the outcome and the independent variable a predictor. Exploratory methods do not draw distinctions between variables along these lines so the choice between which variable to plot along the x- and y-axis is often arbitrary.

Let's add another variable to the mix, in this case the categorical variable `sex`, taking our bivariate exploration to a multivariate exploration. Again each point corresponds to an observation where the values for `num_tokens` and `ttr` intersect. But now each of these points is given a color that reflects which level of `sex` it is associated with.

In this multivariate case, the scatterplot without the trend line is more difficult to interpret. The trend lines for the levels of `sex` help visually understand the variation of the relationship of `num_tokens`and `ttr` much better. But it is important to note that when there are multiple trend lines there is more than one slope to evaluate. The correlation coefficient can be calculated for each level of `sex` (i.e. 'male' and 'female') independently but the relationship between the each slope can be visually inspected and provide important information regarding each level's relative distribution. If the trend lines are parallel (ignoring the ribbons for the moment), as it appears in this case, this suggests that the relationship between the continuous variables is stable across the levels of the categorical variable, with males showing more lexical diversity than females declining at a similar rate. If the lines were to cross, or suggest that they would cross at some point, then there would be a potentially important difference between the levels of the categorical variable (known as an interaction). Now let's consider the meaning of the ribbons. Since the ribbons reflect the range in which the real trend line could fall, and these ribbons overlap, the differences between the levels of our categorical