

LIN 380 Coursebook

Text as Data: An introduction to quantative text analysis and reproducible research in R

Jerid Francom

June 25, 2021 (latest version)

Contents

Welcome	2
License	2
Course	3
0.1 Rationale	3
0.2 Learning goals	3
0.3 Approach	4
0.4 Prerequisites	5
0.5 Programming	5
0.6 Conventions	6
0.7 Build information	7
 I Foundations	 8
Overview	8
1 Data, language, and text analysis	9
1.1 Making sense of a complex world	9
1.2 Data analysis	10
1.3 Language analysis	12
1.4 Text analysis	13
1.5 Coursebook overview	16
1.6 Summary	18
 II Orientation	 18
Overview	18
2 Understanding data	18
2.1 Data	19
2.2 Information	25
2.3 Documentation	36
2.4 Summary	37
3 Statistical approaches	38
3.1 Descriptive statistics	38
3.2 Analytic statistics	39

3.3 Reporting	39
4 Framing research	39
4.1 ...chapter subsection	39
4.2 Annotated readings	39
III Preparation	40
Overview	41
5 Acquire data	41
5.1	41
6 Curate data	41
6.1	41
7 Transform data	42
7.1	42
IV Modeling	42
Overview	42
8 Exploration	42
8.1	42
9 Inference	43
9.1	43
10 Prediction	43
10.1	43
A ...	43

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 in R*. The content is currently under development. Feedback is welcome and can be provided through the [hypothes.is](https://web.hypothes.is/)¹ service. A toolbar interface to this service is located on the right sidebar. Note: you will need to register for a free account to make comments and suggestions.

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

This work by Jerid C. Francom² is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States License.

¹<https://web.hypothes.is/>

²<https://francojc.github.io/>

Course

INCOMPLETE DRAFT

The journey of a thousand miles begins with one step. –Lao Tzu³



The essential questions for this chapter are:

– ...

This chapter aims to provide a brief summary of current research trends that form the context for the rationale for this textbook. It also provides instructors and students an overview of the purpose and approach of the textbook. 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.

0.1 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. 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.

0.2 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

³<https://en.wikipedia.org/wiki/Laozi>

In each chapter of this coursebook specific learning objectives will be specified that target these learning outcomes so it is clear what we are doing and why we are doing it.

0.3 Approach

Many textbooks on doing ‘Data Science’, even those that have a domain-centric approach, such as text analysis, tend to focus on the basic ‘tidy’ approach, seen in Figure 1 from Wickham and Grolemund (2017). However these resources tend to underrepresent the importance of leading with a research question. A big part, or perhaps the biggest part of doing quantitative research, and research in general is what is the question to be addressed. Then comes how to orient the research approach to best address this question (or questions). Then we move on to matching data sources, organizing data, modeling data, and finally reporting findings.

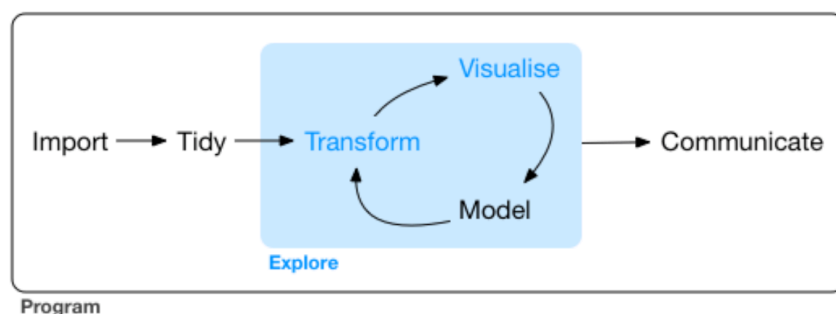


Figure 1: Workflow diagram from R for Data Science.

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 a general idea 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 essence this approach reflects the classic separation between content and format⁴—the content of our research should precede the format it should or will take.

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. (Provide historical and research context for quantitative text analysis)
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. (Develop an understanding of what quantitative research is and how it is approached)
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. (Dive into coding practices produce data ready for statistical analysis)
4. “Modeling” elaborates various statistical approaches for data analysis and contextualizes their application in for types of research questions. (Conducting statistical text analysis)

⁴https://en.wikipedia.org/wiki/Separation_of_content_and_presentation

0.4 Prerequisites



Change this subsection:

- Move the R, RStudio, Packages, Git, GitHub to the **tadr** package vignettes/ articles
- Make reference here to the **tadr** package (Coursebook support package)
-

Before we continue, make sure you have all the software you need for this book:

- **R:** ...
- **RStudio:** RStudio is a free and open source integrated development environment (IDE) for R. ...
- **R packages:** This coursebook uses a bunch of R packages. You can install them all at once by running:

```
install.packages(c("bookdown"))
```

- Coursebook support package **tadr**⁵ is a support R package and resource site for this coursebook. The package includes data, functions, and interactive R programming tutorials which make use of the **swirl** package. The website includes programming demonstrations called ‘Worked’ examples and reference to documentation and other resources for doing quantitative research with R.

```
install.packages("devtools")  
devtools::install_github("lin380/tadr")
```

0.5 Programming

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 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.

Reasons to use R:

- *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.

⁵<https://lin380.github.io/tadr/>

0.6 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 textbook compared to others you have seen. It has been created using R itself –specifically using an R language package called **bookdown** (Xie, 2021). 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. Furthermore, there are four prose text blocks that will be used to signal the reader's attention: *key points*, *notes*, *tips*, 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 textbook
- the reasoning for using the R programming language
- important text conventions employed in this textbook

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 textbook was created using this very package). You can find out more here⁶.

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())`.

Errors will be an inevitable part of learning, but some errors can be avoided. The text will use 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.

$$\hat{c} = \operatorname{argmax}_{c \in C} \hat{P}(c) \prod_i \hat{P}(w_i | c) \quad (1)$$

Data analysis leans heavily on graphical representations. Figures will appear numbered, as in Figure 2.

```
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
```

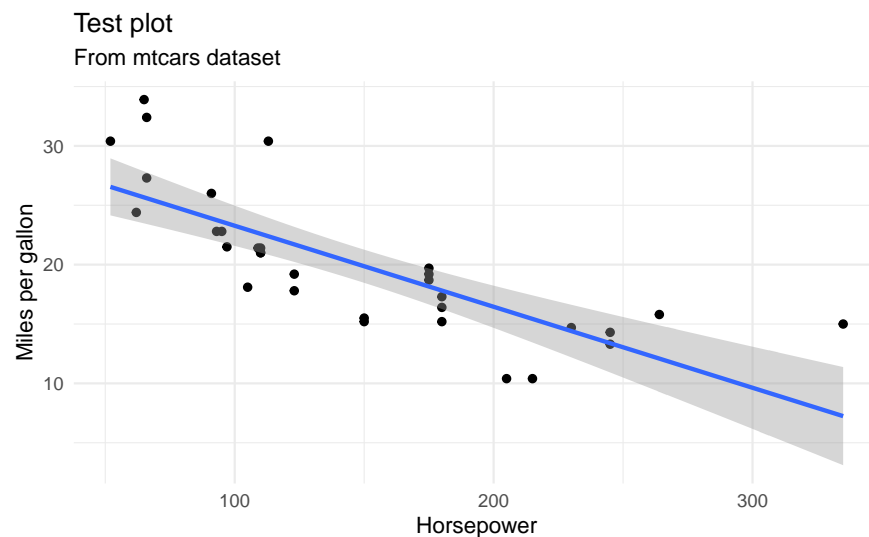


Figure 2: Test plot from mtcars dataset

Tables, such as Table 1 will be numbered separately from figures.

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

0.7 Build information

This coursebook was written in bookdown⁷ inside RStudio⁸. The website is hosted with GitHub Pages⁹ and the complete source is available from GitHub¹⁰.

⁷<http://bookdown.org/>

⁸<http://www.rstudio.com/ide/>

⁹<https://pages.github.com/>

¹⁰<https://github.com/lin380>

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 |

This version of the coursebook was built with R version 4.0.2 (2020-06-22) and the following packages:

| package | version | source |
|----------|---------|----------------|
| bookdown | 0.22 | CRAN (R 4.0.2) |

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. Some attention will be given to the historical background to understand how theory (generative and usage-based grammar) has framed and to some degree continues to frame language research. (4) We will discuss how the programmatic approaches to language, which are fundamental for quantitative text analysis, also provide the opportunity to further science through process documentation and research reproducibility.

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. Millikan¹¹ (1923)



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 experiential 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)? Scale this back even further; let’s take Wikipedia, the world’s largest encyclopedia. Can you imagine reading every wiki entry? As large as a resource such as Wikipedia is ¹², it is still a small fragment of the written language that is

¹¹<https://www.nobelprize.org/uploads/2018/06/millikan-lecture.pdf>

¹²ADD: size of Wikipedia

produced on the web, just the web. 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 than 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 you 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 are right. But it has been demonstrated in numerous experiments on human memory that our confidence does not correlate with accuracy. (where were you when ...? JFK, 9/11, ...other example) 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 inherent 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, carefully 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 indispensable 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." (Chambers, 2020) 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.

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, focusing in on language, the amount of text [here add stats on amount of data added

to the web every day/month/year compared to all the literature from .. to ..?]. 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 (Desjardins, 2019). 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 (Donoho, 2017). 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 [cite], whether a lump is cancerous [cite], what films to recommend based on your previous viewing history [cite], what players a sports team should sign (Lewis, 2004) 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 by Drew Conway in a Venn diagram 3.

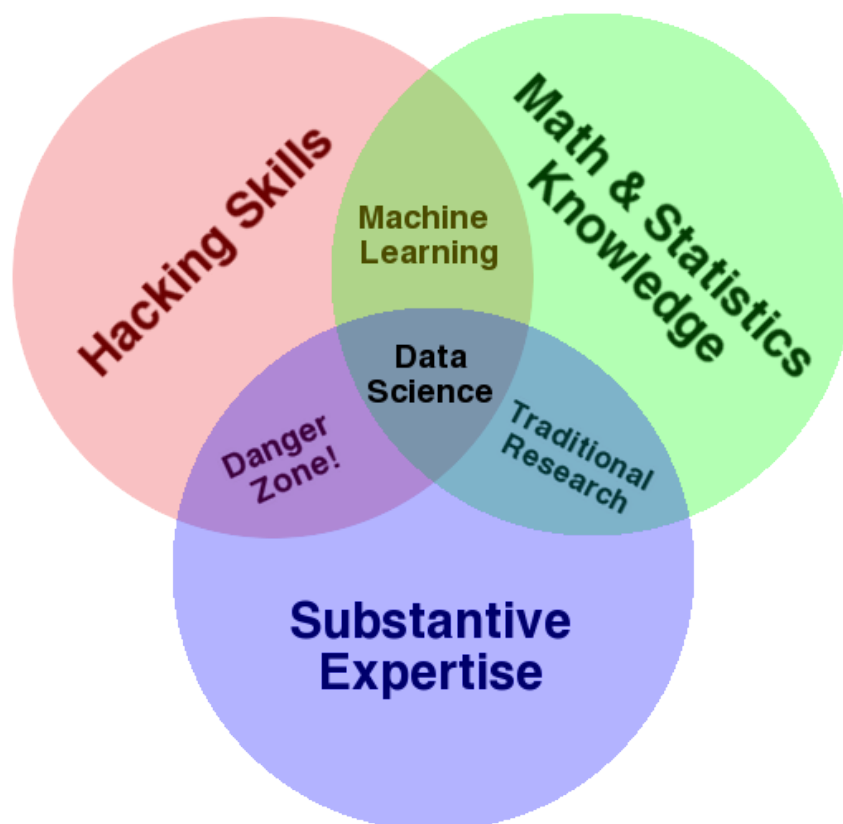


Figure 3: [Venn diagram](<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>) by Drew Conway

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, data science is making important and sometimes disiplinary changes to the way that academic research is

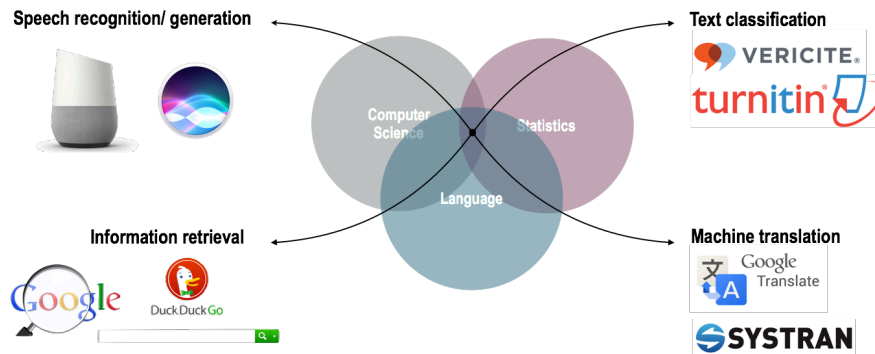


Figure 4: Well-known language applications

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 (Gilquin and Gries, 2009). 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 from the 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 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 practioners, 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 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.



Eisenstein et al. (2012) 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.



Voigt et al. (2017) 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.



Conway et al. (2012) 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.



Kloumann et al. (2012) 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.



Bychkovska and Lee (2017) 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 expres-

sions and/or due to their lack of register awareness (conversation has more bundles than writing generally).



Jaeger and Snider (2007) 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.



Wulff et al. (2007) 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.



Mosteller and Wallace (1963) provide a method for solving the authorship debate surrounding The Federalist papers¹³. 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.



Olohan (2008) 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 (Carmi et al., 2020).

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 5 ¹⁴.

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 “Statistical approaches” 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 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

¹⁴Adapted from Ackoff (1989)

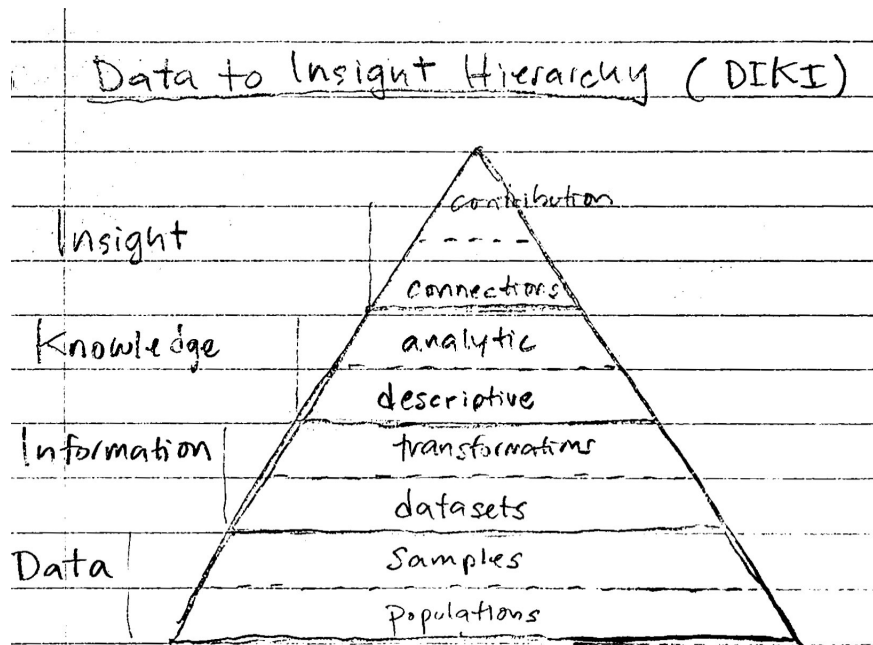


Figure 5: Data to Insight Hierarchy (DIKI)

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 “Modeling” 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” I will discuss three main strategies for accessing data: direct downloads, Automatic Programming Interfaces (APIs), and web scraping. In Chapter 6 “Curate data” 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” 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 “Exploration”, 9 “Inference”, and 10 “Prediction” focus on different categories of statistical analysis each associated with distinct research goals. Exploration covers ‘bottom-up’-style, or ‘unsupervised learning’, 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. 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 explores a set of statistical methods known as ‘supervised learning’. Similar to unsupervised learning, prediction models are employed in a bottom-up fashion. However, the key distinction is that the dataset includes an organizing ‘class’ variable which the statistical methods aim to model in order to formulate a generalization that can correctly classify new textual data. I will cover some standard methods for text classification including Naïve Bayes, k -nearest neighbors (k -NN), and decisions tree and random forest models.

1.6 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.

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 5) 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 5), 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 that 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 Milwaukee? 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 (cite). 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 instantaneously 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 Kucera and Francis (1967). Consult the documentation for this corpus¹⁵. 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 6), what types of research might this corpus support or not support?

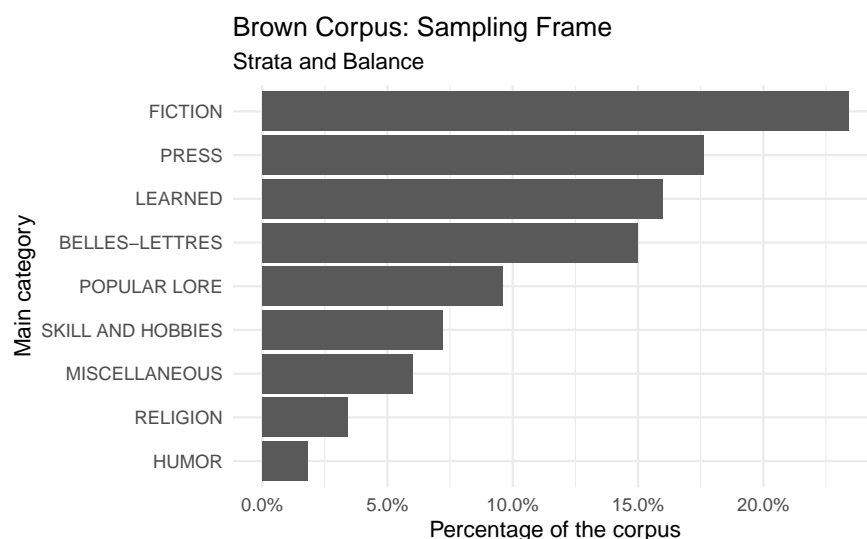


Figure 6: Brown Corpus of Written American English

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)¹⁶ or the British National Corpus (BNC)¹⁷ are corpora which aim to model (represent/ reflect) the general characteristics of the English

¹⁶

¹⁷

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 (Ädel, 2020).

Specialized corpora aim to represent more specific populations. The Santa Barbara Corpus of Spoken American English (SBCSAE)¹⁸, as you can imagine from the name of the resource, aims to model spoken American 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)¹⁹, language use in specific regions of the world²⁰, or a country²¹, 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) OPUS example²². 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 Brown and LOB example²³.

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 language²⁴, 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 Consortium²⁵ or for specific language domains, such as the language acquisition repository TalkBank²⁶. 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 GitHub²⁷, Zenodo²⁸, Re3data²⁹, OSF³⁰, etc.,

¹⁸

¹⁹<https://www.clarin.eu/resource-families/cmc-corpora>

²⁰<http://ice-corpora.net/ice/index.html>

²¹<https://cesa.arizona.edu>

²²

²³

²⁴<http://www.hd.uib.no/icame/ij22/vihla.pdf>

²⁵<https://www ldc.upenn.edu/>

²⁶<http://talkbank.org/>

²⁷<https://github.com/>

²⁸<https://zenodo.org/>

²⁹<http://www.re3data.org/>

³⁰<https://osf.io/>

Table 3: A list of some corpus repositories

| Resource | Description |
|---|--------------------------------|
| https://corpus.byu.edu/ >BYU corpora | A repository of corpora tha |
| http://corporafromtheweb.org/ >COW (COpora from the Web) | A collection of linguistically |
| http://wortschatz.uni-leipzig.de/en/download/ >Leipzig Corpora Collection | Corpora in different langua |
| https://www ldc.upenn.edu/ >Linguistic Data Consortium | Repository of language cor |
| http://www.resourcebook.eu/searchll.php >LRE Map | Repository of language reso |
| http://www.nltk.org/nltk_data/ >NLTK language data | Repository of corpora and |
| http://opus.lingfil.uu.se/ >OPUS - an open source parallel corpus | Repository of translated te |
| http://talkbank.org/ >TalkBank | Repository of language col |
| https://corpus1.mpi.nl/ds/asv/?4 >The Language Archive | Various corpora and langua |
| http://ota.ox.ac.uk/ >The Oxford Text Archive (OTA) | A collection of thousands o |

Table 4: Corpora and language datasets.

| Resource |
|--|
| http://www.socsci.uci.edu/~lpearl/CoLaLab/CHILDESTreebank/childestreebank.html >CHILDES Treebank |
| http://www.cs.cornell.edu/~cristian/Cornell_Movie-Diologs_Corpus.html >Cornell Movie-Diologs Corpus |
| http://www.llf.uam.es/~fmarcos/informes/corpus/coarginl.html >Corpus Argentino |
| https://cesa.arizona.edu/ >Corpus of Spanish in Southern Arizona |
| https://www.statmt.org/europarl/ >Europarl Parallel Corpus |
| http://storage.googleapis.com/books/ngrams/books/datasetv2.html >Google Ngram Viewer |
| http://opus.lingfil.uu.se/OpenSubtitles_v2.php >OpenSubtitles2011 |
| http://www.ruscorpora.ru/en/ >Russian National Corpus |
| https://quantumstat.com//dataset >The Big Bad NLP Database - Quantum Stat |
| https://catalog.ldc.upenn.edu/docs/LDC97S62/ >The Switchboard Dialog Act Corpus |
| http://langsnap.soton.ac.uk/ >Welcome to LANGSNAP - LANGSNAP |
| http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html >Westbury Lab Web Site |

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.

If your corpus search ends in a dead-end, either because a suitable resource does not appear to exist or an existing resource is unattainable given licensing restrictions or fees, it may be time to compile your own corpus. Turning to machine readable texts on the internet is usually the logical first step to access language for a new corpus. Language texts may be found on sites as uploaded files, such as pdf or doc (Word) documents, or found displayed as the primary text of a site. Given the wide variety of documents uploaded and language behavior recorded daily on social media, news sites, blogs and the like, compiling a corpus has never been easier. Having said that, how the data is structured and how much data needs to be retrieved can pose practical obstacles to collecting data from the web, particularly if the approach is to acquire the data by hand instead of automating the task. Our approach here, however, will be to automate the process as much as possible whether that means leveraging R package interfaces to language data, converting hundreds of pdf documents to plain text, or scraping content from web documents.

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

Table 5: 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 |

Table 6: Language data from previous research and meta-studies.

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

variety of resources. Below I’ve included attempts by others to provide a summary of the corpus data and language resources available.



Here I can work with real or simplified research questions and have students consider which of set of corpus resources would most likely be the better 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

Table 7: Lists of corpus resources.

| Resource |
|--|
| Learner corpora a |
| Machine Learning Datasets Papers With Code |
| Stanford NLP corpora |
| Where |

of running text.

Running text sample from the Europarl Parallel Corpus³¹.

```
> Resumption of the session
> I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I v
> 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 l
> One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam, who had visited t
```

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 Corpus³².

```
> 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
> 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,  /
```

³¹<https://www.statmt.org/europarl/>

³²


```
> 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 Corpus³³.

```
> <TEI xmlns="http://www.tei-c.org/ns/1.0"><teiHeader><fileDesc><titleStmt><title>Sample A01 from The
> "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></e
> <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.

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 8 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 case³⁴ as my personal preference. There are also alternatives³⁵. 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

³³http://www.nltk.org/nltk_data/

Table 8: First 10 source and target sentences in the Europarle Corpus.

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

Table 9: 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 |

column and its 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 9 a language-relevant sub-set of the meta-data is associated with each utterance.

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 10 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

Table 10: 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 |

Table 11: Europarle Paralle Corpus with ‘words’ as primary unit of investigation.

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

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 Wickham (2014). 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* for the particular observation-feature pair.

Consider adding a visual to highlight these relationships

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 8, 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 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.

Table 12: Non-speech lines in the Europarle dataset.

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

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.

Consider adding some ‘messy’ data and/ or summary tables which do not reflect the relational structure we are aiming to create to base our research on.

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 8) we will see that there are sentence lines which do not represent actual parliament speeches. In Table 12 we see these lines.

A research project aiming to analyze speech would want to normalize this dataset removing these lines, as seen in Table 13.

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

Table 13: The Europarle dataset with non-speech lines removed.

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

Table 14: Lines with possessives with extra whitespace in the Europarle dataset.

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

Table 15: The Europarle dataset with whitespace from possessives removed.

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

Table 16: The Europarl dataset with lowercasing applied.

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

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 Europarl dataset sentences means we lose potentially valuable information; namely the ability to identify 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 ³⁶. Second not all words can be stemmed as they do not have derivative 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'.

\begin{table}

\caption{Results for a `word_stems` filter for "investig" in the Brown Corpus.}

³⁶<https://snowballstem.org/algorithms/>

Table 17: 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 |

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 |

| 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 |

\end{table}

We can see from the results in Table 2.2.2.2 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) derivative 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’.

`\begin{table}`

`\caption{Results for a word_lemmas filter 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 |

`\end{table}`

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

`\begin{table}`

`\caption{Results for a word_lemmas filter 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 |

`\end{table}`

Again only words of the same grammatical category are returned. In this case the verb ‘be’ has many more derivative 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 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.*

`\begin{table}`

`\caption{Disfluency counts in the first 10 utterance_text values from the Switchboard Corpus.}`

Table 19: The first 10 word bigrams of the Europarle Corpus.

| type | sentence_id | word_bigrams |
|--------|-------------|----------------------|
| Target | 2 | i declare |
| Target | 2 | declare resumed |
| Target | 2 | resumed the |
| Target | 2 | the session |
| Target | 2 | session of |
| Target | 2 | of the |
| Target | 2 | the european |
| Target | 2 | european parliament |
| Target | 2 | parliament adjourned |
| Target | 2 | adjourned on |

| 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 |

\end{table}

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 11, 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.

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

Table 20: The first 10 character trigrams of the Europarle Corpus.

| type | sentence_id | char_trigrams |
|--------|-------------|---------------|
| Target | 2 | ide |
| Target | 2 | dec |
| Target | 2 | ecl |
| Target | 2 | cla |
| Target | 2 | lar |
| Target | 2 | are |
| Target | 2 | rer |
| Target | 2 | ere |
| Target | 2 | res |
| Target | 2 | esu |

Table 21: 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_re |
|--------|-------------|----------|---------|-------|--------------------------|-----------------|--------------|
| Target | 6 | 1 | Please | INTJ | NA | 2 | discourse |
| Target | 6 | 2 | rise | VERB | Mood=Imp VerbForm=Fin | 0 | root |
| Target | 6 | 3 | , | PUNCT | NA | 2 | punct |
| Target | 6 | 4 | then | ADV | PronType=Dem | 10 | advmod |
| Target | 6 | 5 | , | PUNCT | NA | 10 | punct |
| Target | 6 | 6 | for | ADP | NA | 10 | case |
| Target | 6 | 7 | this | DET | Number=Sing PronType=Dem | 8 | det |
| Target | 6 | 8 | minute | NOUN | Number=Sing | 10 | nmod:poss |
| Target | 6 | 9 | 's | PART | NA | 8 | case |
| Target | 6 | 10 | silence | NOUN | Number=Sing | 2 | conj |
| Target | 6 | 11 | . | PUNCT | NA | 2 | punct |

annotated **sentence** is tokenized **token** and indexed **token_id**. Then **upos** contains the Universal Part of Speech tags³⁷, 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 Dependencies³⁹ framework). Another common syntactic parsing framework is phrase constituency parsing (Jurafsky and Martin, 2020).

Now compare the English example sentence dataset in Table 21 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

³⁷Descriptions of the UPOS tagset³⁸

³⁹<https://universaldependencies.org/>

Table 22: Automatic linguistic annotation for grammatical category and syntactic structure for an example Spanish sentence from the Europarl Corpus

| type | sentence_id | token_id | token | upos | feats |
|--------|-------------|----------|----------|-------|---|
| Source | 6 | 1 | Invito | VERB | Gender=Masc Number=Sing VerbForm=Fin |
| Source | 6 | 2 | a | ADP | NA |
| Source | 6 | 3 | todos | PRON | Gender=Masc Number=Plur PronType=Tot |
| Source | 6 | 4 | a | ADP | NA |
| Source | 6 | 5 | que | SCONJ | NA |
| Source | 6 | 6 | nos | PRON | Case=Acc,Dat Number=Plur Person=1 PrepCase=Npr PronType=Obj |
| Source | 6 | 7 | pongamos | VERB | Mood=Ind Number=Plur Person=1 Tense=Pres VerbForm=Fi |
| Source | 6 | 8 | de | ADP | NA |
| Source | 6 | 9 | pie | NOUN | Gender=Masc Number=Sing |
| Source | 6 | 10 | para | ADP | NA |
| Source | 6 | 11 | guardar | VERB | VerbForm=Inf |
| Source | 6 | 12 | un | DET | Definite=Ind Gender=Masc Number=Sing PronType=Art |
| Source | 6 | 13 | minuto | NOUN | Gender=Masc Number=Sing |
| Source | 6 | 14 | de | ADP | NA |
| Source | 6 | 15 | silencio | NOUN | Gender=Masc Number=Sing |
| Source | 6 | 16 | . | PUNCT | NA |

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 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 Dialog 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
```

Table 23: 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 24: 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 |

It turns out that on the corpus website⁴⁰ 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 24.

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.

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.

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

⁴⁰<https://catalog.ldc.upenn.edu/docs/LDC97S62/>

from the corpus for the derived dataset, and the description of the variables created from the derived dataset for the analysis dataset.

Consider adding more specifics on the characteristics and formats for documenting corpus data and datasets.

2.4 Summary

In this chapter we have focused on data and information –the first two components of DIKI Hierarchy. This process is visualized in Figure 7.

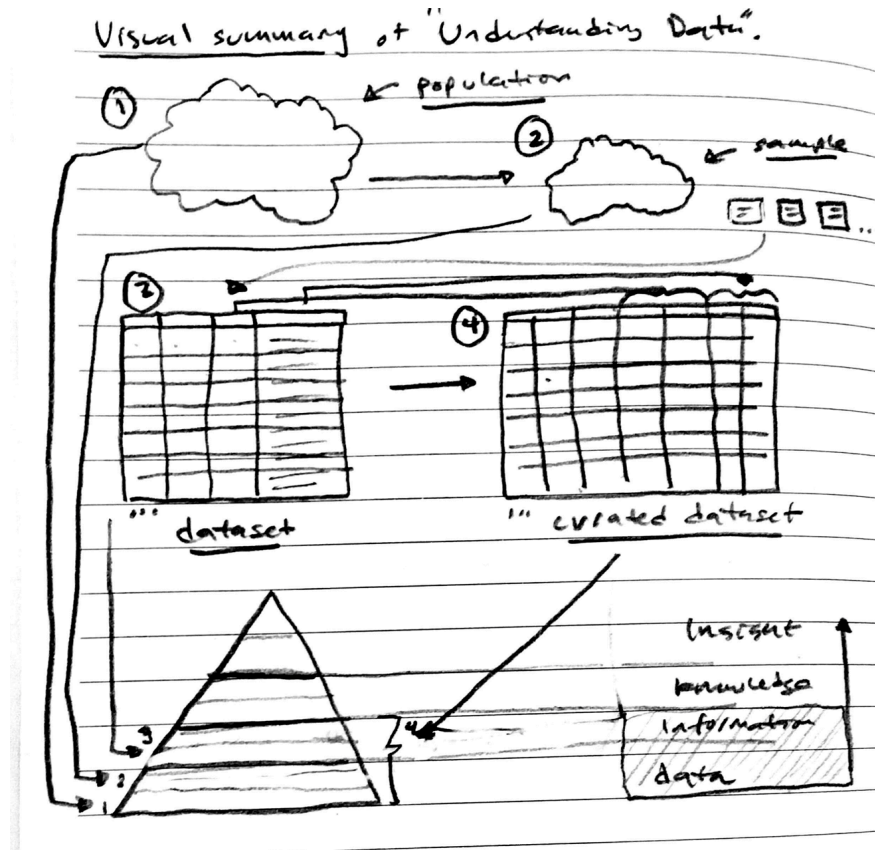


Figure 7: Understanding data: visual summary

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 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.

3 Statistical approaches

INCOMPLETE DRAFT

Lies, damn lies, and statistics
—Benjamin Disraeli, popularized by Mark Twain



In this chapter you will learn:

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

Packages

Overview ... DRAFT

In this chapter I will build on the notions of data and dataset from the previous chapter. I will underscore that the aim of statistical analysis is to uncover patterns in datasets. There are two basic aims that underlie statistical analysis. Either the goal is to probe the data to test a hypothesis about the relationship between the observations or to explore the potential relationships. These distinctions are grouped into Inference or Exploration statistics. Fundamental to either approach is the use of descriptive statistics. These methods aim to provide researchers a better understanding of the structure of the data, prepare the researcher to make decisions about which statistical tests are most appropriate, and to safeguard against false assumptions (missing data, data distributions, etc.)

3.1 Descriptive statistics

3.1.1 Information values

- Categorical variables
- Continuous variables

Maximizing information, minimizing loss

3.1.2 Summaries

- Central tendency
- ...

3.1.3 Distributions

3.1.4 Correction

- Missing data
- Skew
 - Continuous variables
- Outliers
 - Trimming

3.2 Analytic statistics

3.2.1 Exploratory data analysis (EDA)

Bottom-up approach, hypothesis generating

3.2.2 Inferential data analysis (IDA)

Top-down approach, hypothesis testing

3.2.3 Predictive data analysis (PDA)

Mixed approach, can be used for the generation of hypotheses or to test hypotheses

3.3 Reporting

- Descriptive analysis
 - Procedures to diagnose and correct
- Analytic analysis
 - Communicate findings in statistical appropriate forms
 - * Depends on the analytic statistic(s) applied

4 Framing research

INCOMPLETE DRAFT

It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts. —Sir Arthur Conan Doyle, Sherlock Holmes



In this chapter you will learn:

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

Before jumping into the code, every researcher must come to a project with a clear idea about the purpose of the analysis. This means doing your homework in order to understand what it is exactly that you want to achieve; that is, you need to identify a research question. The first step is become versed in the previous literature on the topic. What has been written? What are the main findings? Secondly, it is important to become familiar with the standard methods for approaching the topic of interest. How has the topic been approached methodologically? What are the types, sources, and quality of data employed? What have been the statistical approaches employed? What particular statistical tests have been chosen? Getting an overview not only of the domain-specific findings in the literature but also the methodological choices will help you identify promising plan for carrying out your research.

4.1 ...chapter subsection

text

Packages

4.2 Annotated readings

Ignatow, G., & Mihalcea, R. (2017). **An introduction to text mining: Research design, data collection, and analysis.** Sage Publications. (Ignatow and Mihalcea, 2017)

Chapter 5 “Designing your research project”

Research design is essentially concerned with the basic architecture of research projects, with designing projects as systems that allow theory, data, and research methods to interface in such a way as to maximize a project's ability to achieve its goals (see Figure 5.1). Research design involves a sequence of decisions that have to be taken in a project's early stages, when one oversight or poor decision can lead to results that are ultimately trivial or untrustworthy. Thus, it is critically important to think carefully and systematically about research design before committing time and resources to acquiring texts or mastering software packages or programming languages for your text mining project.

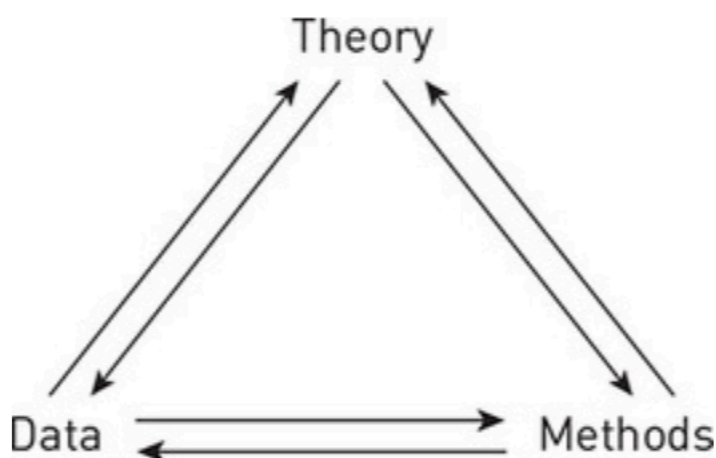


Figure 5.1 ■ The Research Design Triad

Egbert, J., Larsson, T., &
Biber, D. (2020). *Doing Linguistics with a Corpus: Methodological Considerations for the
Everyday User*. Cambridge University Press. (Egbert et al., 2020)

Chapter 3 “Research Designs: Linguistically Meaningful Research Questions, Observational Units,
Variables, and Dispersion”

- Research questions should drive decisions about the choice of observational unit, how variables are defined, and the choice of research design.
- Observational units can be defined at the level of the linguistic feature, the text, or the corpus.
- Variables can be measured qualitatively, according to variants of a linguistic feature, or quantitatively, using rates of occurrence for features.

Chapter 7 “Interpreting Quantitative Results”

- Linguistics is done by linguists, not by computers.
- In order to be useful, quantitative corpus linguistic analysis should be coupled with sound qualitative interpretation.
- Researchers can rely on linguistic context, text-external context, and linguistic theory to guide their interpretation of quantitative corpus findings.

Part III

Preparation

Overview

5 Acquire data

INCOMPLETE DRAFT

...



In this chapter you will learn:

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

Overview...

5.1 ...

text

5.1.1 Packages

```
library(rvest) # full-fledged web scraping
library(datapasta) # copy/paste approach to HTML tables
```

6 Curate data

INCOMPLETE DRAFT

...



In this chapter you will learn:

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

,

```
# Packages
```

Overview...

6.1 ...

text

Data Organization in Spreadsheets (Broman and Woo, 2018). Although based on spreadsheets, many of the best practices discussed apply to good data organization regardless of the technology.

7 Transform data

INCOMPLETE DRAFT

...



In this chapter you will learn:

- ...
- ...

text

7.1 ...

text

Packages

NOTE:

- Cover Corpus and Document-Term Matrices (DTM)s

Part IV

Modeling

Overview

8 Exploration

INCOMPLETE DRAFT

...



In this chapter you will learn:

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

text

8.1 ...

text

8.1.1 Packages

Packages

9 Inference

INCOMPLETE DRAFT

...



In this chapter you will learn:

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

text

9.1 ...

text

9.1.1 Packages

Packages

10 Prediction

INCOMPLETE DRAFT

...



In this chapter you will learn:

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

text

10.1 ...

text

10.1.1 Packages

Packages

A ...

References

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, 16(1):3–9.

Ädel, A. (2020). Corpus compilation. In Paquot, M. and Gries, S. T., editors, *A Practical Handbook of Corpus Linguistics*, pages 3–24. Springer, Switzerland.

- Broman, K. W. and Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician*, 72(1):2–10.
- Bychkovska, T. and Lee, J. J. (2017). At the same time: Lexical bundles in l1 and l2 university student argumentative writing. *Journal of English for Academic Purposes*, 30:38–52.
- Carmi, E., Yates, S. J., Lockley, E., and Pawluczuk, A. (2020). Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation. *Internet Policy Review*, 9(2).
- Chambers, J. M. (2020). S, r, and data science. *Proceedings of the ACM on Programming Languages*, 4(HOPL):1–17.
- Conway, L. G., Gornick, L. J., Burfeind, C., Mandella, P., Kuenzli, A., Houck, S. C., and Fullerton, D. T. (2012). Does complex or simple rhetoric win elections? an integrative complexity analysis of u.s. presidential campaigns. *Political Psychology*, 33(5):599–618.
- Desjardins, J. (2019). How much data is generated each day?
- Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766.
- Egbert, J., Larsson, T., and Biber, D. (2020). *Doing Linguistics with a Corpus: Methodological Considerations for the Everyday User*. Elements in Corpus Linguistics. Cambridge University Press.
- Eisenstein, J., O’Connor, B., Smith, N. A., and Xing, E. P. (2012). Mapping the geographical diffusion of new words. *Computation and Language*, pages 1–13.
- Gilquin, G. and Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1):1–26.
- Ignatow, G. and Mihalcea, R. (2017). *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*. Sage Publications.
- Jaeger, T. F. and Snider, N. (2007). Implicit learning and syntactic persistence: Surprisal and cumulativity. *University of Rochester Working Papers in the Language Sciences*, 3(1).
- Jurafsky, D. and Martin, J. H. (2020). *Speech and Language Processing*.
- Kloumann, I., Danforth, C., Harris, K., and Bliss, C. (2012). Positivity of the english language. *PloS one*.
- Kucera, H. and Francis, W. N. (1967). *Computational Analysis of Present Day American English*. Brown University Press Providence.
- Lewis, M. (2004). *Moneyball: The Art of Winning an Unfair Game*. WW Norton & Company.
- Millikan, R. A. (1923). *The Electron and the Light-Quant from the Experimental Point of View*.
- Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309.
- Olohan, M. (2008). Leave it out! using a comparable corpus to investigate aspects of explicitation in translation. *Cadernos de Tradução*, pages 153–169.
- Voigt, R., Camp, N. P., Prabhakaran, V., Hamilton, W. L., Hetey, R. C., Griffiths, C. M., Jurgens, D., Jurafsky, D., and Eberhardt, J. L. (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526.
- Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10).
- Wickham, H. and Grolemund, G. (2017). *R for Data Science*. O’Reilly Media, first edit edition.
- Wulff, S., Stefanowitsch, A., and Gries, S. T. (2007). Brutal brits and persuasive americans. *Aspects of Meaning*.

Xie, Y. (2021). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.22.