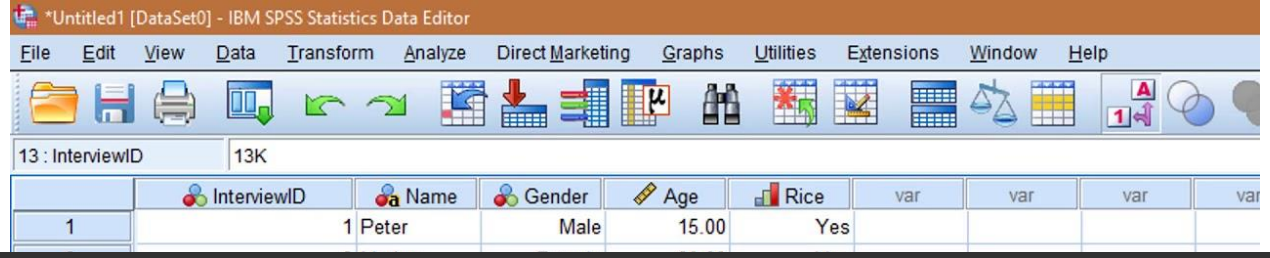
An Introduction to Statistical Analysis for Linguistics using R

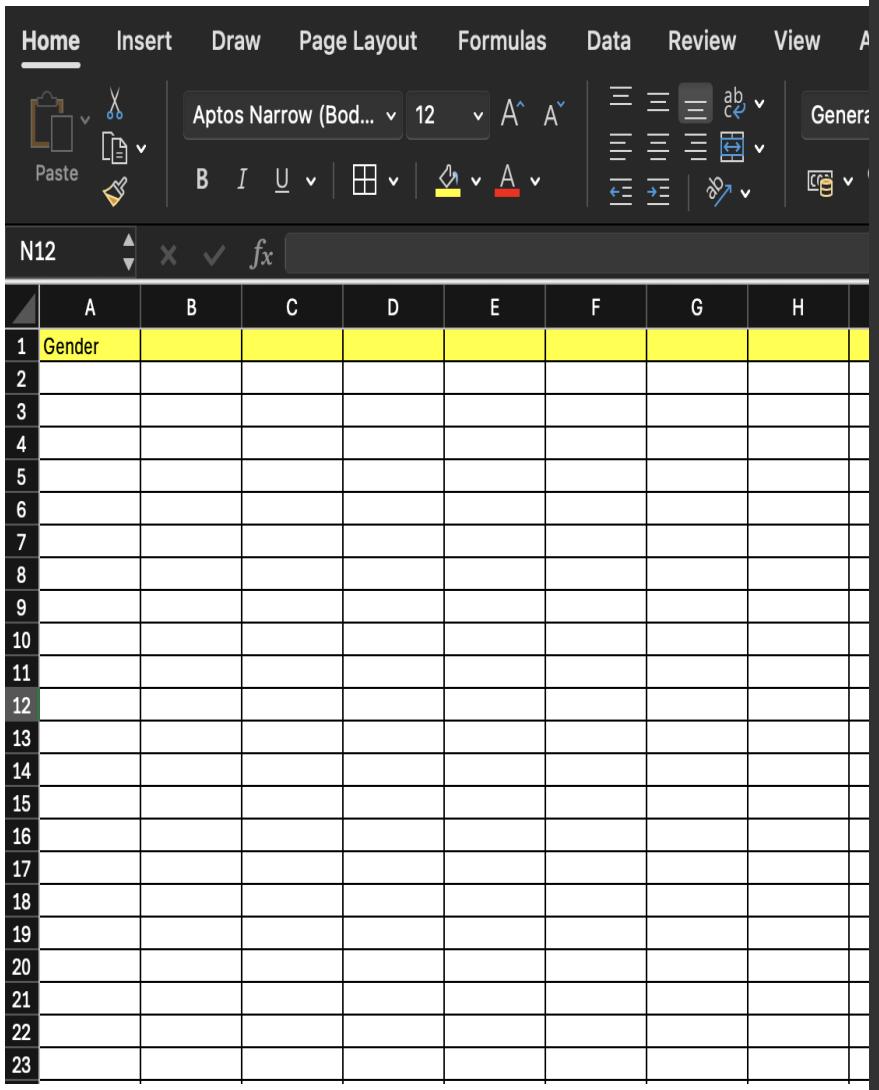
A practical workshop for Redeemers University Linguistics Students on the 9th of July 2025

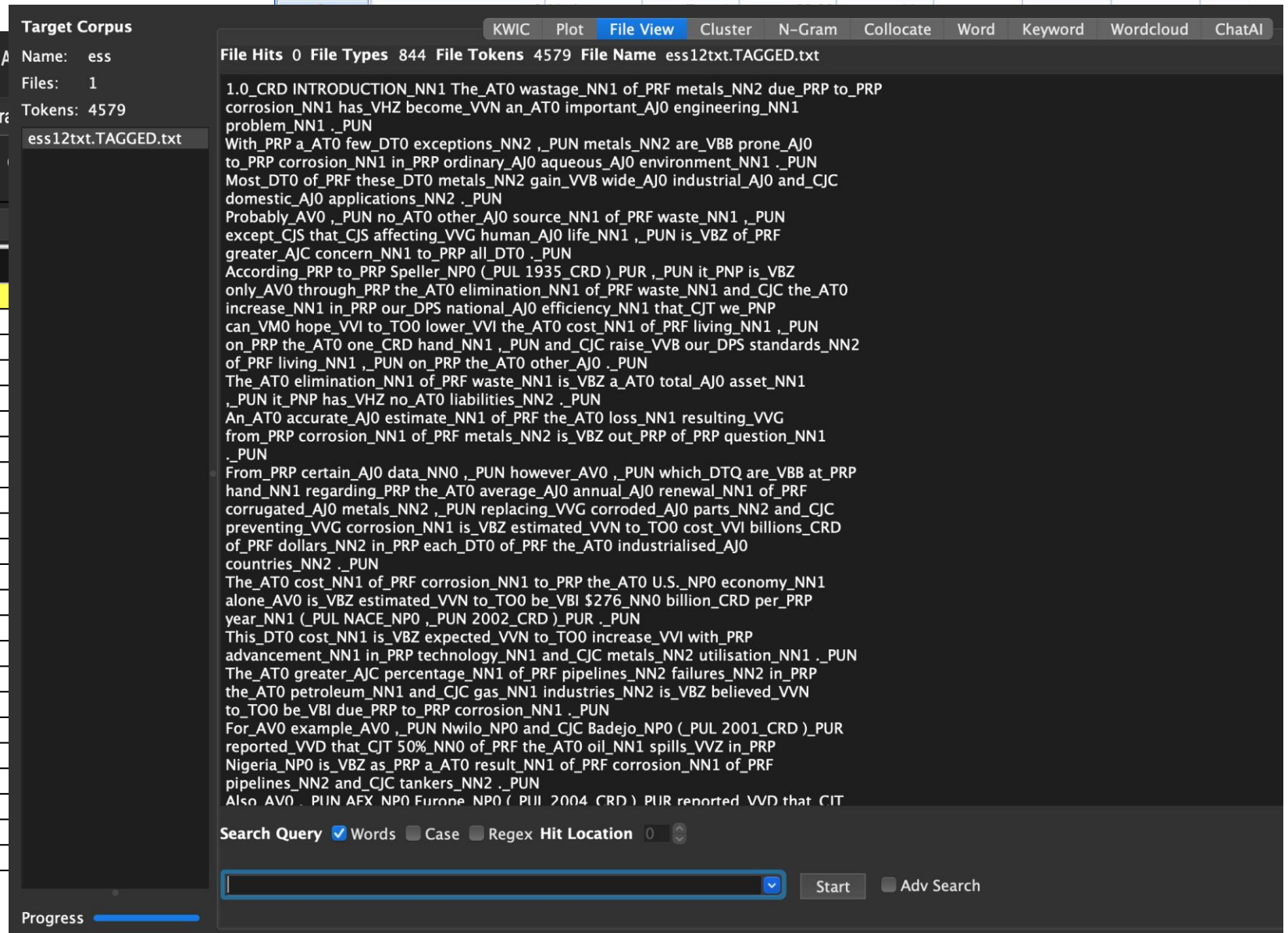
Temitayo Olatoye
University of Eastern Finland, Finland
temitayo.olatoye@uef.fi

Time	Topic	Description
10:00 – 10:10	**Welcome & Overview**	- Brief introduction - Why R and statistics for linguists
10:10 – 10:35	**Basic Statistical Concepts for Linguists**	- Types of data (nominal, ordinal, interval, ratio) - Descriptive vs. inferential statistics - Variables (e.g., reaction times, word frequency, syntactic choice)
10:35 – 11:05	**Getting Started with R & RStudio**	- R/RStudio basics - R syntax essentials (objects, data frames, indexing) - Loading data (read.csv) and basic inspection: *head*(), *str*()
11:05 – 11:40	**Exploratory Data Analysis**	- Summary stats: *summary*(), *mean*(), *sd*(), *table*() - Visualization with *ggplot2* (histograms, boxplots, scatterplots) - Detecting trends and outliers
11:40 – 11:50	**Break**	
11:50 – 12:15	**Statistical Tests**	- t-test *t.test()* – e.g., comparing mean word durations between groups - Chi-square *chisq.test()* – e.g., POS vs. speaker group - Correlation *cor.test()* – e.g., lexical frequency vs. response time
12:15 – 12:50	**Intro to Regression Analysis**	- What is regression analysis? - Linear regression *lm()*: predicting duration from frequency or age group - Interpreting coefficients and p-values - Logistic regression
12:50 – 13:00	**Wrap-Up & Q&A**	- Resources for further learning - Questions and feedback

Which have you used?







Introduction – *why R?*

- Open-source
- A growing adoption in linguistics
- Sharing is easy + transparent
- Produce better research
- Do cool stuff… 😎

RStudio is an Integrative Development Environment (IDE).

Types of data

Data Type	Ordered?	Equal Intervals?	True Zero?	Example
Nominal (categorical)	❌	❌	❌	Lexical variants, morphosyntactic choices
Ordinal	✅	❌	❌	Evaluation scales, proficiency levels
Interval	✅	✅	❌	Temperature (°C)
Ratio	✅	✅	✅	Reaction time, Frequency

What can you do with your data?

- Description
- Explanation
- Prediction

Descriptive Statistics - Describe a data sample

- Location: Mean value, median, mode, sum
- Dispersion: Standard deviation, variance, range
- Tables: Absolute, relative and cumulative Frequencies
- Charts: Histograms, bar charts, box plots, scatter plots

Inferential Statistics

Make a statement about the general population

Simple test procedures

- t-Test
- Binominal Test
- Chi-square test
- Mann-Whitney U Test
- Wilcoxon-Test
- ...

Regression Analysis

- Simple linear regression
- Multiple regression
- Logistic regression
- ...

Correlation analysis

- Pearson Correlation analysis
- Spearman Rank Correlation
- ...

ANOVA

- Single factorial ANOVA
- Two factorial ANOVA
- ANOVA with measurement repetitions
- ...

Operationalizing your variables

	Independent Variable (IV)	Dependent Variable (DV)
What is it?	The variable you **manipulate or group by**	The variable you **measure or observe**
Role in study	Acts as the **cause**, condition, or factor	Acts as the **effect** or outcome
Control over it?	✅ Yes (you choose or categorize it)	❌ No (you observe how it changes)
Examples in Linguistics	- Word Frequency - Sentence Type - Speaker Age	- Reaction Time - Word Duration - Accuracy
Research question	Does **word frequency** affect **reaction time**?	DV: reaction time (measured in ms)

Analysis

	Monofactorial Analysis	Multifactorial Analysis
Definition	Examines the effect of **one independent variable**	Examines the effects of **multiple independent variables**
Complexity	Simpler, easier to interpret	More complex, captures interactions between variables
Example	Studying how **age** affects pronunciation accuracy	Studying how **age, gender, and L1 background** affect pronunciation accuracy
Statistical Methods	t-tests, chi-square tests, simple regression	ANOVA, multiple regression, mixed-effects models
Limitations	May overlook confounding factors	Provides a more nuanced and realistic analysis
Usage	Preliminary studies or when data is limited	Advanced research with rich datasets

Resources

- https://www.stgries.info/research/overview-research.html

- http://www.martinschweinberger.de/blog/resources/
- http://www.martinschweinberger.de/blog/presentations-talks/

- https://dlf.uzh.ch/openbooks/statisticsforlinguists/
- https://www.zora.uzh.ch/id/eprint/183632/1/Statistics-for-Linguists-1580118836.pdf
- https://appliedstatisticsforlinguists.org/bwinter_stats_proofs.pdf

Texts

- Baayen, R Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using r*. Cambridge: Cambridge University press.
- Crawley, Michael J. 2005. *Statistics: An Introduction Using r. 2005*. Chichester, West Sussex: John Wiley & Sons.
- Field, Andy, Jeremy Miles, and Zoe Field. 2012. *Discovering Statistics Using r*. Vol. 50. Sage. https://doi.org/https://doi.org/10.5860/choice.50-2114.
- Gries, Stefan. 2009, 2021. *Statistics for Linguistics Using r: A Practical Introduction*. Berlin & New York: Mouton de Gruyter.
- Levshina, Natalia. 2015. *How to Do Linguistics with r: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins Publishing Company.
- Wilcox, R. 2009. *Basic Statistics: Understanding Conventional Methods and Modern Insights*. Vol. 47. Oxford University Press. https://doi.org/https://doi.org/10.5860/choice.47-5075.
- Winter, Bodo. 2019. *Statistics for Linguists: An Introduction Using r*. Routledge. https://doi.org/https://doi.org/10.4324/9781315165547