The State of the Stats: Current Use of Statistical Methods Across Linguistics Subfields

Rachael Tatman (rctatman@uw.edu), Department of Linguistics

Work Supported by NSF Grant DGE-1256082. I would also like to extend my thanks to the University of Washington Linguistics department, Richard Wright, Gretchen McCulloch and Ethan Campbell-Taylor.

Motivation

Problem:

Statistical methods are widely used in linguistics, but vary widely by subfield. This makes it difficult to get a good idea of the state of the field as a whole.

Solution:

Create a database containing a principled sampling of the statistical methods used in recently-published linguistics research.

Applications:

- Guiding course design
- Offering insight into differences between subfields
- As an instructional tool

Methodology

Selection of Journals

Journals were selected by asking faculty and current graduate students at the University of Washington to provide a list of what they considered the top journals in their subfield. The final list was then vetted by the faculty. Other scholar's selections may differ, but the journals listed here offer broad coverage that will help to give a general idea of many of the subfields of linguistics.

List of Journals (by subfield)

Acquisition (L1):	**General Linguistics:**	**Semantics & Pragmatics:**
Journal Of Child Language Acquisition And Development	Language And Linguistics Compass	Journal Of Pragmatics
Language Acquisition	Language	Journal Of Semantics
Acquisition (L2):	Linguistic Inquiry	Linguistics And Philosophy
Second Language Research	Natural Language & Linguistic Theory	Natural Language Semantics
Studies In Second Language Acquisition	**Phonetics:**	Semantics And Pragmatics
Annual Review Of Applied Linguistics	Journal Of Speech, Language, And Hearing Research	**Sign Linguistics:**
Applied Linguistics	Journal Of Phonetics	Sign Language Studies
Reading In A Foreign Language	Journal Of The Acoustical Society Of America (Speech Communication)	Sign Language And Linguistics
Areal Linguistics:	Phonetica	Deaf Studies Digital Journal
International Journal Of American Linguistics	Journal Of The International Phonetic Association	Journal Of Interpretation (Registry Of Interpreters For The Deaf)
Journal Of East Asian Linguistics	Language And Speech	Journal Of Deaf Studies And Deaf Education
Journal Of Comparative Germanic Linguistics	Speech Communication	**Syntax**:**
Journal Of Slavic Linguistics	**Morphology:**	Journal Of Linguistics
Oceanic Linguistics	Mental Lexicon	Lingua
Probus	Morphology	Linguistic Inquiry
Studies In African Linguistics	**Phonology:**	Studia Linguistica
Computational Linguistics*:	Phonology	Syntax
Computational Linguistics	Laboratory Phonology	
IEEE/ACM Transactions On Audio, Speech, And Language Processing	**Psycholinguistics:**	*Computational linguistics includes proceedings, as computational linguistics conferences tend to be more influential than journals.
Language Resources And Evaluation	Journal Of Memory And Language	
Natural Language Engineering	Language Cognition And Neuroscience	**Many journals listed under "General Linguistics" were also listed as top syntax journals.
Transactions Of The Association For Computational Linguistics	Cognition	

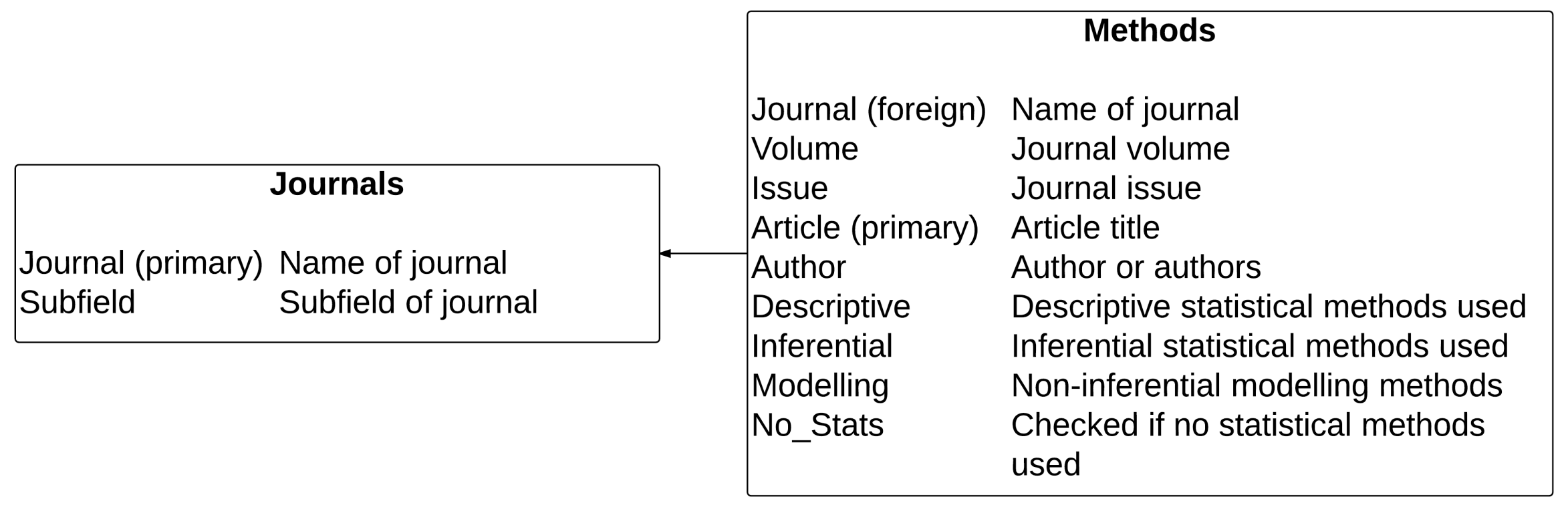
Methodology (cont.)

Selection of Articles

For each journal, the most recent full issue was selected. For purely linguistic journals, statistical techniques (or lack thereof) were recorded for all articles. For multi-field journals (e.g. the Journal of the Acoustical Society of America) only linguistics articles were included.

Database

All information was entered into a locally-hosted mySQL relational database made for this project. The structure of the database is shown in the schema below. It will be made publicly available within the next year.



Journals

Journal (primary)	Name of journal
Subfield	Subfield of journal

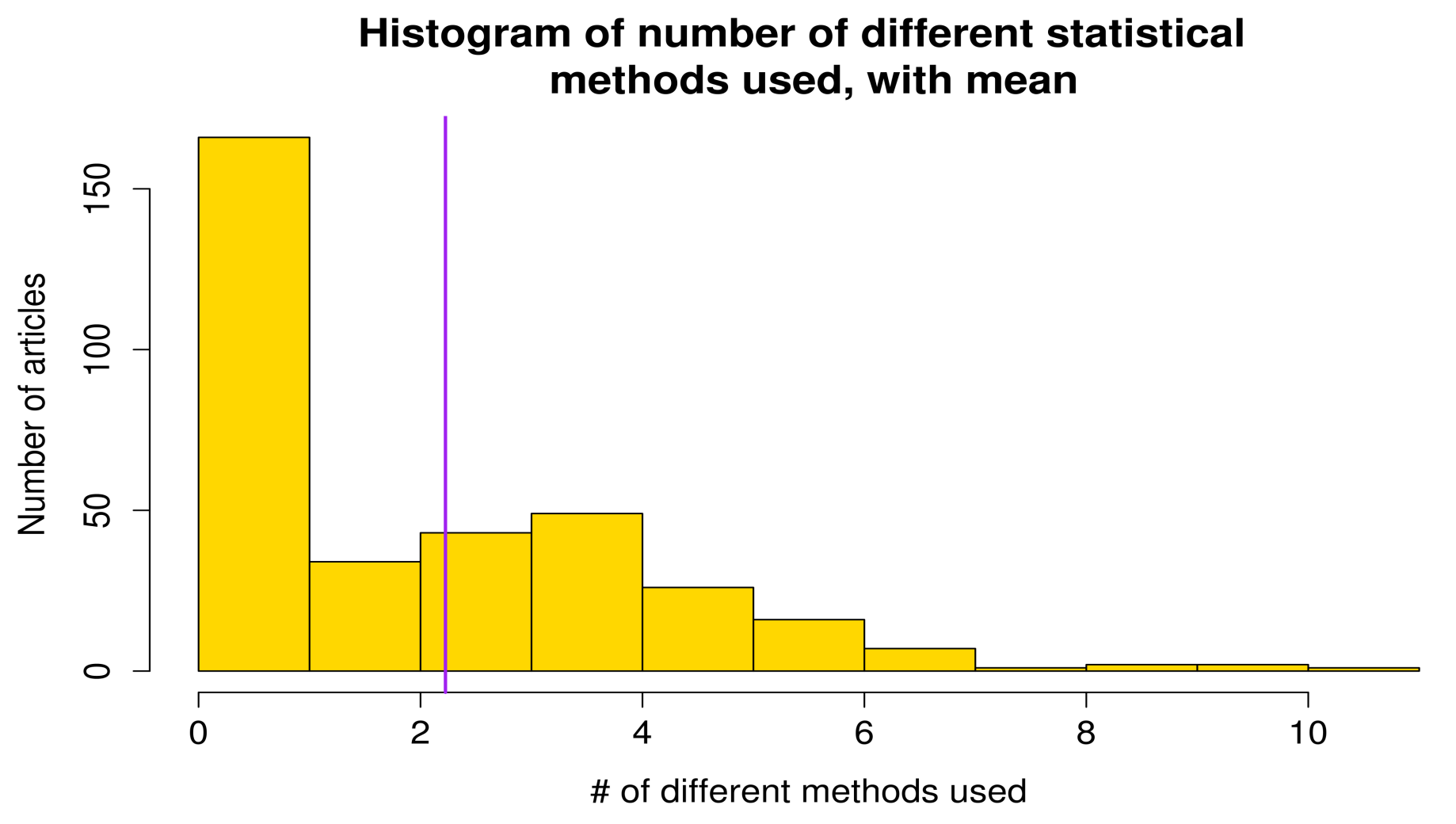
Methods

Journal (foreign)	Name of journal
Volume	Journal volume
Issue	Journal issue
Article (primary)	Article title
Author	Author or authors
Descriptive	Descriptive statistical methods used
Inferential	Inferential statistical methods used
Modelling	Non-inferential modelling methods
No_Stats	Checked if no statistical methods used

Analysis and Results

Overall

Of the 348 journal articles included, 65.8% included at least one method of statistical analysis. Of those that did include at least one method, the average number of different methods used was 2.24. This is summarized in the histogram below.



Common Inferential Methods

Method	Number of Times Used
ANOVA	64
t-test	60
Pearson's r	33
χ2	22
Bonferroni correction	10
Fisher's test	8
Linear regression	8
Tukey's honestly significant difference [HSD]	7
Linear mixed effects model	7
Bootstrapping (inferential only)	6
Wilcoxon rank sum test	5
Wilcoxon signed-rank test	5
R-squared	5
Multiple regression	5

Results by Subfield

In the tables below, the cell that summarizes the percent of articles that use statistical inference is also color-coded:

RED = High levels of use (over 70% of articles)
PURPLE = Medium levels of use (between 70% and 30%)
BLUE = Low levels of use (below 30%)

Acquisition(L1)

# of articles:	8
% of articles that use statistical inference:	100%
Common inferential methods:	T-tests, linear modeling, χ2

Acquisition(L2)

# of articles:	37
% of articles that use statistical inference:	54%
Common inferential methods:	ANOVAs, Tukey's HSD, linear regression

Areal Linguistics

# of articles:	26
% of articles that use statistical inference:	15%
Common inferential methods:	T-test, Wilcoxon signed rank test

Computational Linguistics

# of articles:	57
% of articles that use statistical inference:	98%
Common inferential methods:	T-test, Bootstrapping

General Linguistics

# of articles:	24
% of articles that use statistical inference:	29%
Common inferential methods:	Linear regression, logistic regression

Morphology

# of articles:	9
% of articles that use statistical inference:	55%
Common inferential methods:	Linear mixed-effects

Phonetics

# of articles:	70
% of articles that use statistical inference:	97%
Common inferential methods:	PLDA, Wilcoxon rank-sum tests, Wilcoxon sign-rank tests, mixed models, ANOVA, t-test

Phonology

# of articles:	9
% of articles that use statistical inference:	55%
Common inferential methods:	ANOVA, mixed models

Psycholinguistics

# of articles:	20
% of articles that use statistical inference:	100%
Common inferential methods:	χ2, t-tests, ANOVA, mixed models

Semantics & Pragmatics

# of articles:	25
% of articles that use statistical inference:	24%
Common inferential methods:	χ2, t-tests, mixed models

Sign Linguistics

# of articles:	36
% of articles that use statistical inference:	55%
Common inferential methods:	Pearson's r, t-tests, ANOVA, mixed models

Sociolinguistics

# of articles:	15
% of articles that use statistical inference:	60%
Common inferential methods:	Mixed effects (Goldvarb, Rbrul)

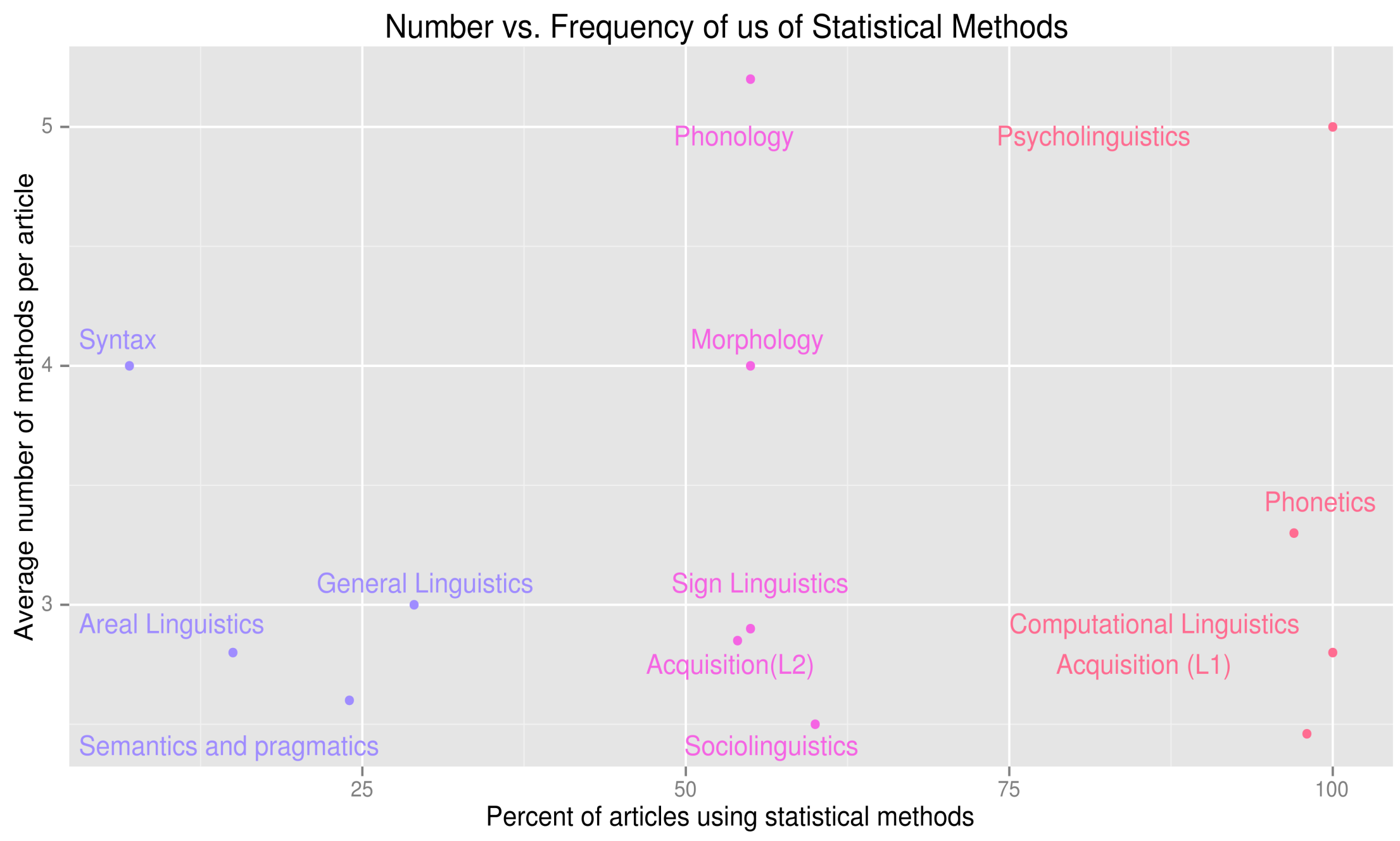
Syntax

# of articles:	13
% of articles that use statistical inference:	7%
Common inferential methods:	------------

Results by Subfield (cont.)

The average number of methods used in each subfield is summarized below. It did not correlated with the percentage of articles in each subfield using statistical methods, r(11) = 0.07, p = 0.79.

Average number of methods in articles that used statistical methods, by discipline	
Phonology	5.2
Psycholinguistics	5
Syntax	4 (one article)
Morphology	4
Phonetics	3.3
General Linguistics	3
Sign Linguistics	2.9
Acquisition(L2)	2.85
Acquisition (L1)	2.8
Areal Linguistics	2.8
Semantics and pragmatics	2.6
Sociolinguistics	2.5
Computational Linguistics	2.46
Overall Average:	3.3



Conclusion

Summary

- Statistical methods were used in every subfield
- Different subfields vary widely in how commonly statistical methods are used
- Different subfields make use of different statistical methods
- Most studies which used statistical methods used between 2 and 4, though this also varied by subfield
- There is no correlation between popularity of statistical methods in a subfield and how many are used together

Applications

- Customizing statistics course design to include methods students are likely to encounter
- Database can be used to quickly find examples of methods used "in the field"
- Students can add to the database to increase coverage and practice identifying types of statistical methods (descriptive, inferential, non-inferential modelling)
- With diachronic data, can be used to track changes in the field over time