



Methodology for Quantitative Data Analysis

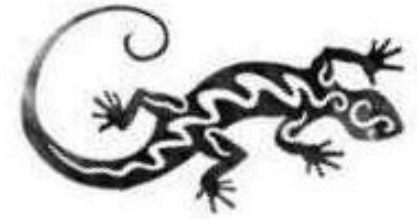
Quantitative Analysis



Hypothesis:

Reference is a distinguishing feature for register variation.

Quantitative Analysis



Methodology:

- Annotation,
- Descriptive Data Analysis,
- Clustering,
- Classification

Quantitative Analysis



Annotation Quality

- Precision and Recall:

	Precision	Recall
EO	0.98	0.73
GO	0.73	0.60

- Inter-annotator agreement human annotators:

	HumanAnnotators	HumanSystem
EO	0.74	0.85
GO	0.78	0.66

Quantitative Analysis



Dataset

- Starting Dataset: EO
- 8 Registers: Essay, Fiction, Instr, Popsci, Share, Web, Speech, Tou
- 11 reference features: types and functions (cf. Halliday and Hasan1976)

COMP: particular, general

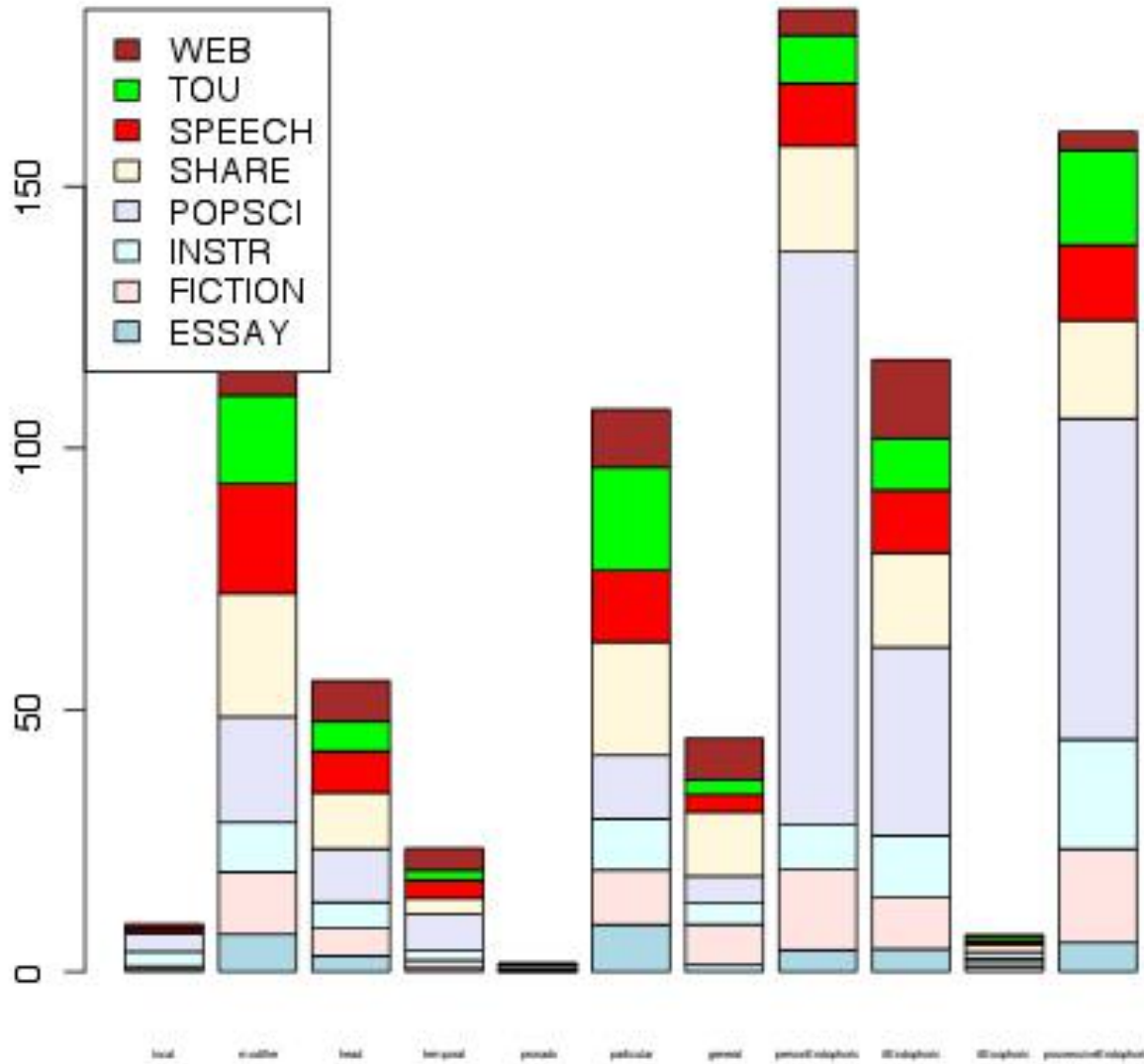
PERS: personEndo,
itEndo, possEndo, itExo

DEM: local, modifier,
head, temporal,
pronadv



Quantitative Analysis

Reference Function by Registers

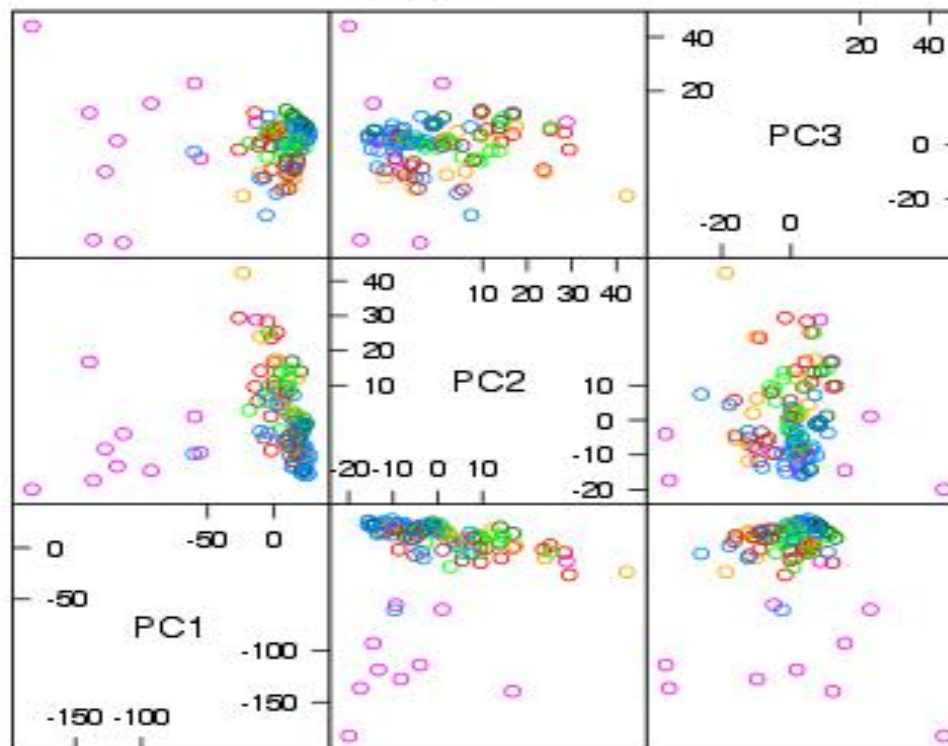


Unsupervised Clustering: PCA



Registers and Reference

ESSAY ○
 FICTION ○
 INSTR ○
 POPSCI ○
 SHARE ○
 SPEECH ○
 TOU ○
 WEB ○

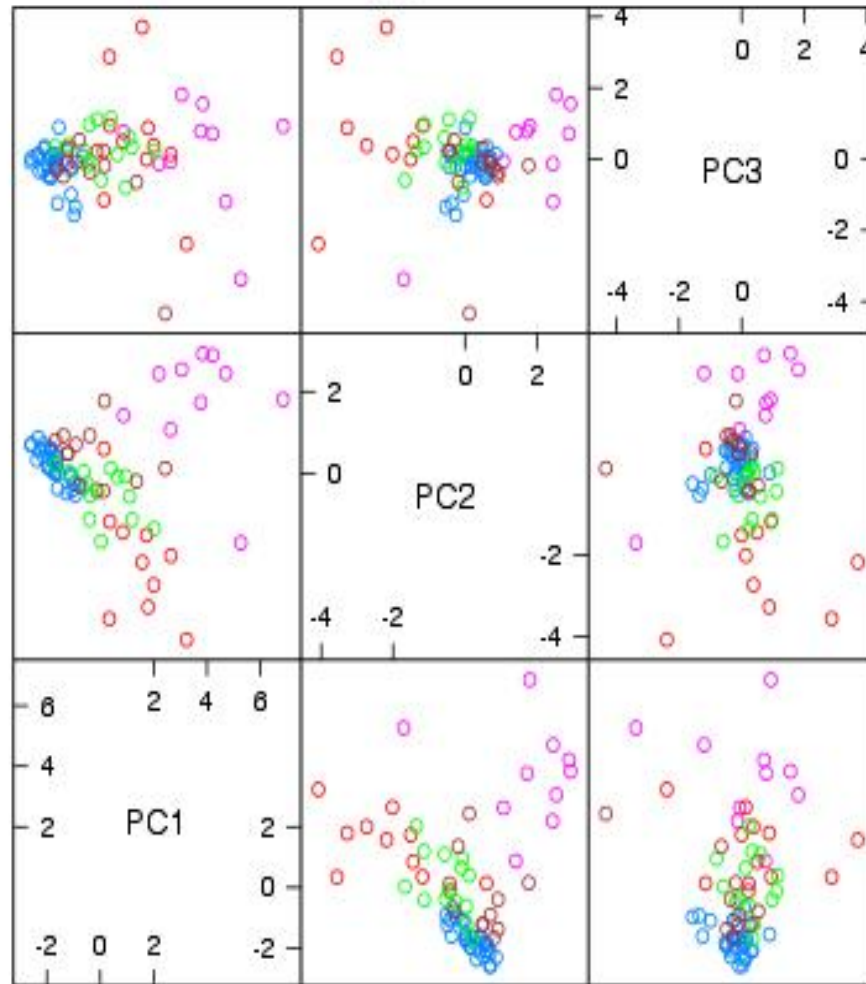


Scatter Plot Matrix



Registers and Reference

ESSAY ○
 FICTION ○
 INSTR ○
 POPSCI ○
 TOU ○

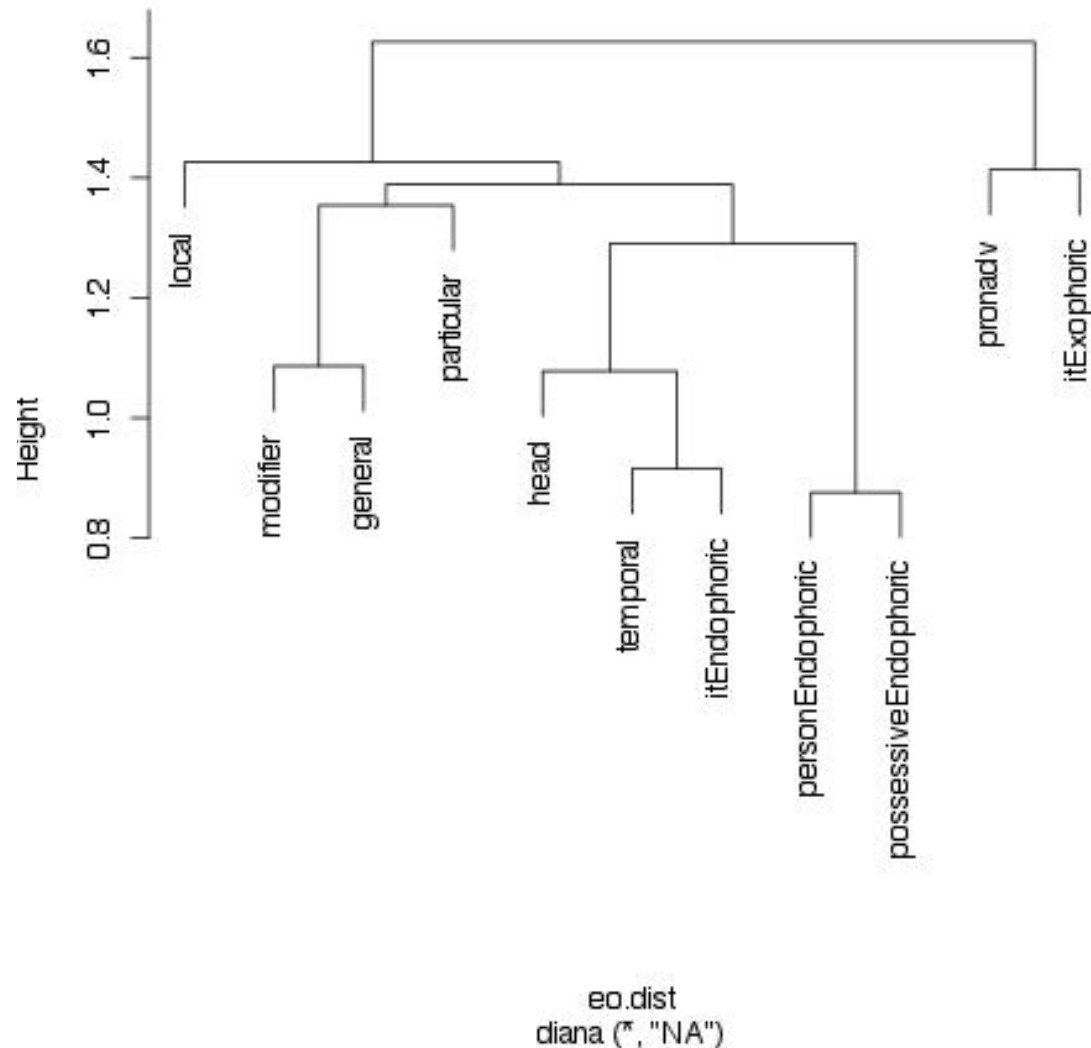


Scatter Plot Matrix



Unsupervised Clustering: PCA

Dendrogram of `diana(x = eo.dist)`



COMP: particular, general

PERS: personEndo,
itEndo, pos-
sEndo, itExo

DEM: local, modifier,
head, temporal,
pronadv

Analysis of Variance



Analysis of variance shows that the major predictors of register variance are the following:

- personEndophoric
- possessiveEndophoric
- general
- local,
- itEndophoric

Classification



svm-Classifiers

Parameters	Accuracy _{EO}	Accuracy _{EO_{s1}}	Accuracy _{EO_{s2}}	Accuracy _{EO_{s3}}
all	50.91%	67.57%	51.28%	66.20%
pred1	49.09%	72.97%	41.03%	76.06%
pred2	48.18%	75.68%	46.15%	78.87%

pred₁: personEndophoric, possessiveEndophoric, general, local, itEndophoric

pred₂: personEndophoric, possessiveEndophoric, general, local

EO_{s1}: ESSAY, FICTION, REGISTER, INSTR, POPSCI, SPEECH

EO_{s2}: WEB, SPEECH, SHARE

EO_{s3}: ESSAY, FICTION, INSTR, POPSCI, TOU