Digital Humanities: Dealing with text

Jelle Zuidema

ILLC, Universiteit van Amsterdam

10 March 2014

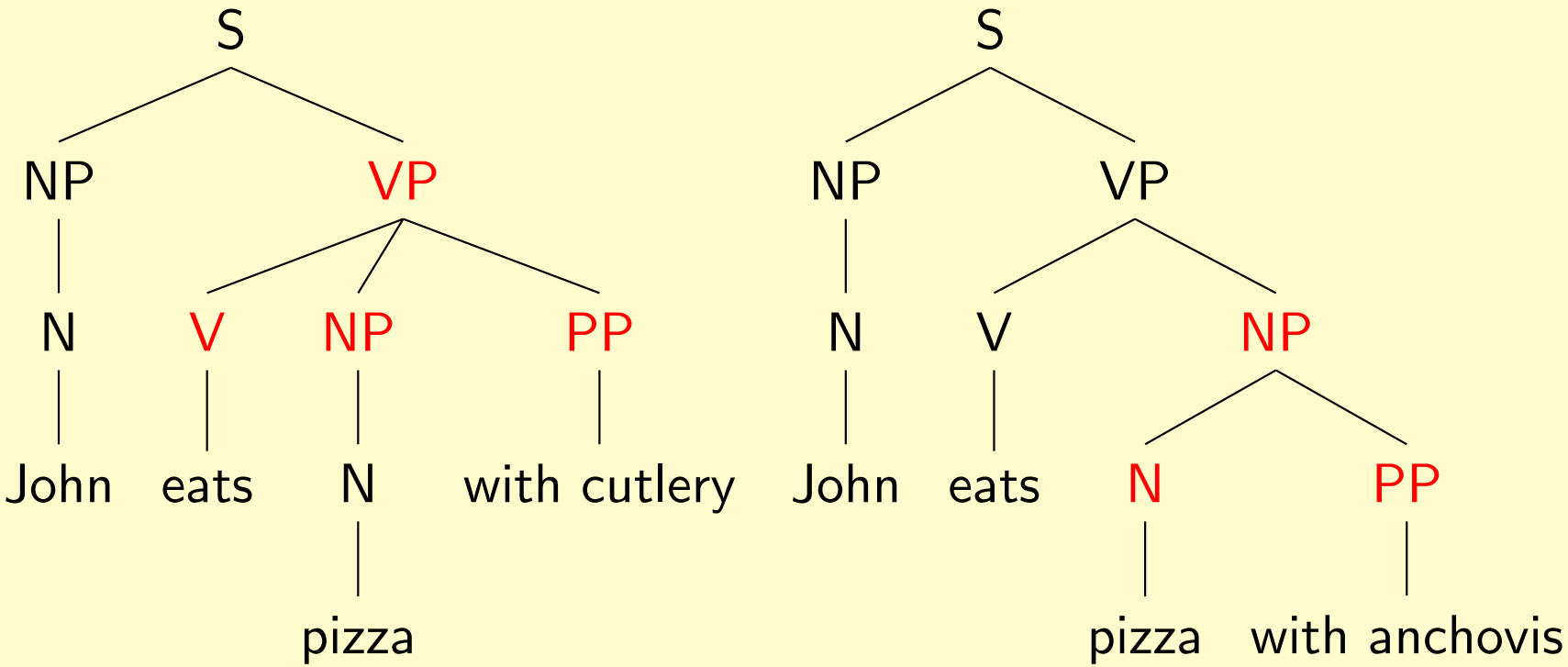
Can digital techniques contribute to the humanities?

- Structural ambiguity
- Frequency-irregularity correlation
- Long tail

Regular expressions for search and gathering statistics

- Description
- Demo: Cygwin.org
- Demo: Comparisons
- Demo: Zipf's law

Structural Ambiguity



Digitally-born Discoveries I

- Structural ambiguity is *pervasive*
- Sentences of 20-40 words often have thousands of possible grammatical analyses.
- Marcus et al., 1993, Building a Large Annotated Corpus of English: The Penn Tree Bank

Digitally-born Discoveries II

- $>$ 2000 past tense verbs commonly used;
- $\sim$ 200 commonly used irregular past tenses
- Strong correlation frequency-(ir)regularity
- Francis & Kucera. Brown Corpus (1967). Frequency Analysis of English Usage: Lexicon and Grammar (1982).
- Assignment I.

Digitally-born Discoveries III

- Zipf's (first) law (1935)
- Few words are extremely frequent, extremely many words are very infrequent (the "long tail")
- If you rank words on frequency, and plot frequency against rank on a log-log scale, you get an approximately straight line:

$$\log(\textit{frequency}) = a - b \times \log(\textit{rank})$$

How do we find patterns and gather statistics?

- Many custom-made tools for particular corpora / annotations;
- One completely general query language: regular expressions

Regular expressions

Allows you define complex patterns. E.g.,

- [a-z] matches any lower case character
- [a-z]+ matches any string of lower case characters
- Q[a-z]+ any lower case string preceded by a capital Q
- [^Q][a-z]+ any lower case string not preceded by a Q

Unix/linux shell commands

Linux computers have tools for operating on large text files installed by default;

- `grep` is a filter tool, finds the lines that match an expression;
- `sed` is a 'stream editor' that replaces matches with something else;
- `sort` is a sorting tool;
- `uniq` is a tool that removes, and counts, duplicate lines in a text file;
- `bash` is a 'shell' that allows you to type in commands, and send the output of one tool directly to the next ('pipe');
- `cygwin` is a free linux-emulator that works on MS Windows and can do all these things.

Case study: Child-directed speech

(Kunert, Fernandez & Zuidema, 2011)

- Using the Brown corpora “Adam”, “Eve” and “Sarah”
- Using regular expressions to get sentence and word length distributions

Data

We use the Brown Corpus from the CHILDES database:

- 3 children: Adam (2;3–5;2), Sarah (2;3–5;1), and Eve (1;6–2;3)
- 214 transcribed longitudinal conversations (one per corpus file)

		number of utterances			
corpus	#files	child	mother	oth.adults	total
Adam	55	46733	20354	6344	73431
Sarah	139	38089	29481	16752	84322
Eve	20	12119	10446	4359	26924

An excerpt from the Adam sub-corpus:

```
CHI : why it got a little tire?
MOT : because it's a little truck.
CHI : can't it be a bigger truck?
MOT : that one can't be a bigger truck but there are bigger trucks.
```

Measures of Speech Complexity

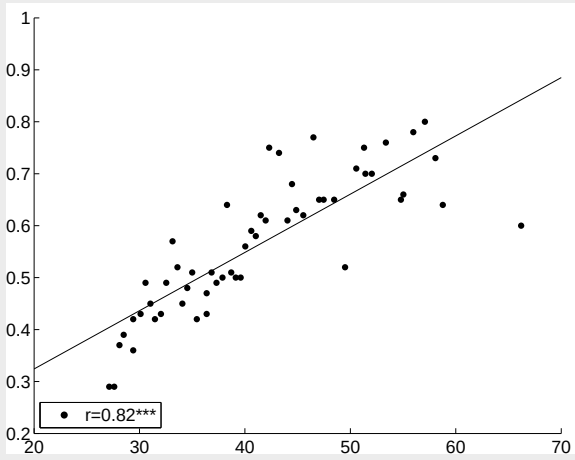
Four simple measures to quantify the complexity of each DP's speech:

- Mean Utterance Length (UL): length of utterance measured in words, averaged over a dialogue ($\sim$ *syntactic complexity*)
- Mean Word Length (WL): length of words measured in characters, averaged over a dialogue ($\sim$ *morphological complexity*)
- Mean Number of Word Types (WT): the number of distinct word types in a dialogue divided by the number of utterances by the relevant speaker in that dialogue ($\sim$ *lexical complexity*)
- Mean Number of Consonant Triples (CT): the number of consonant triples (in the surface orthographic form) per utterance per dialogue ($\sim$ *phonological complexity*)

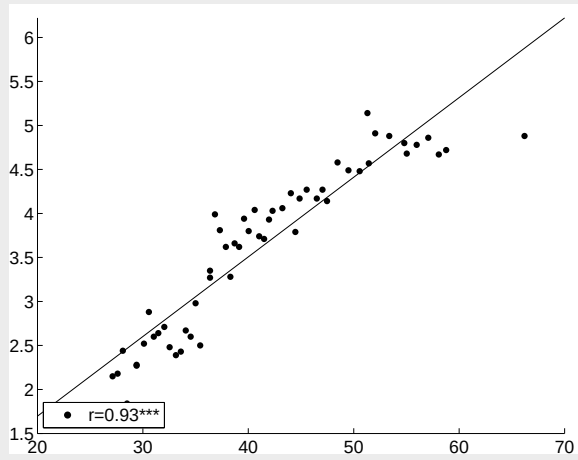
Complexity against Age

Correlation between WT and UL complexity (vertical axis) and the age of the child in months (horizontal axis) in the Adam corpus.

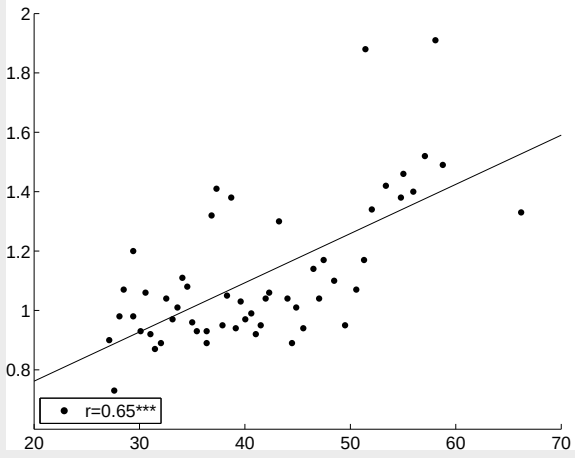
Child-WT vs. age



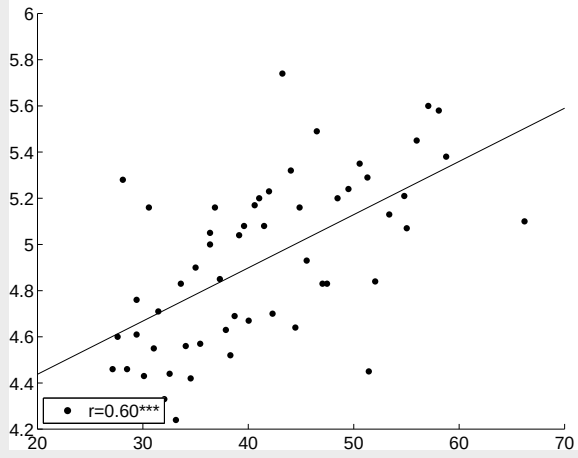
Child-UL vs. age



Mother-WT vs. age

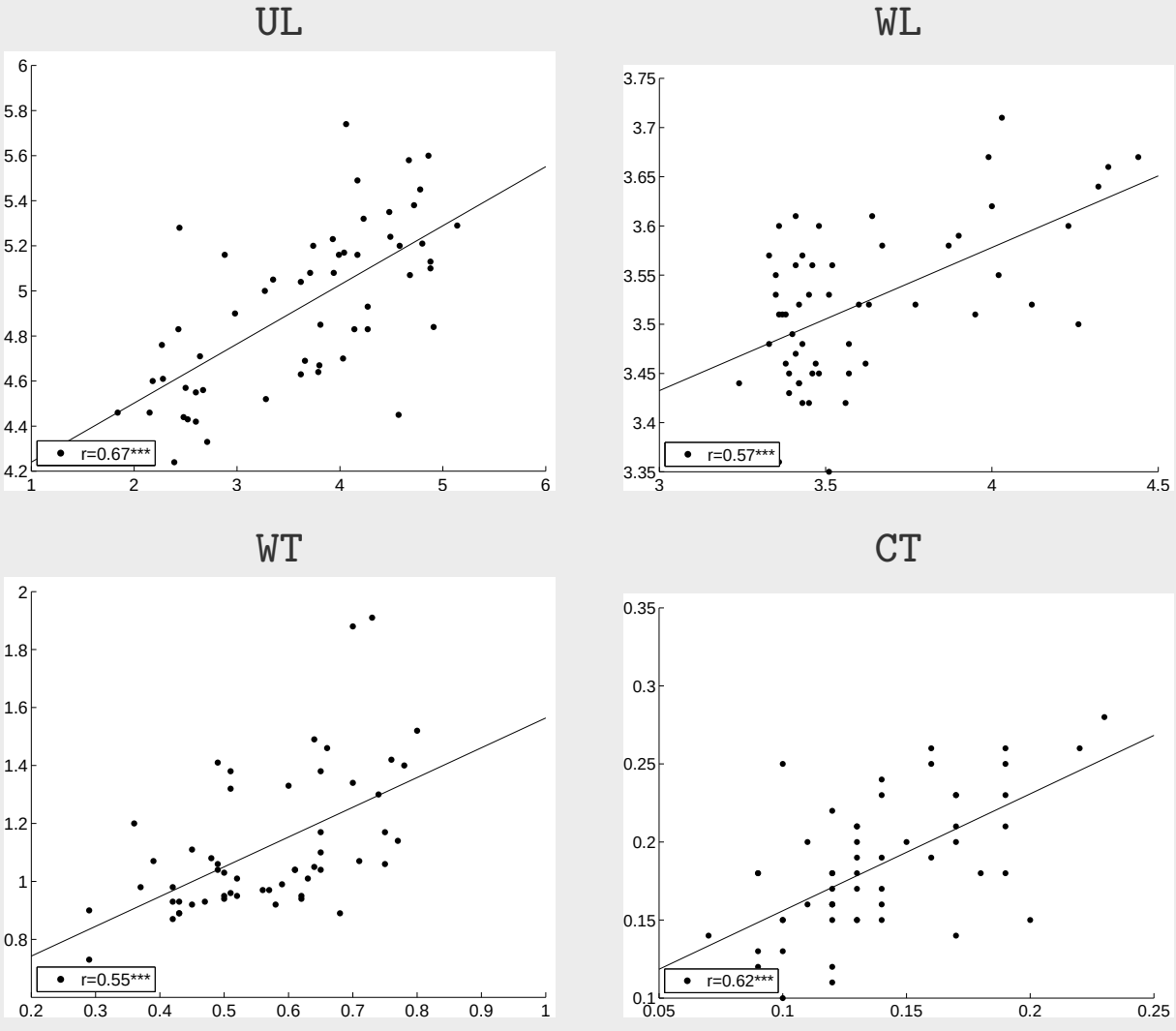


Mother-UL vs. age

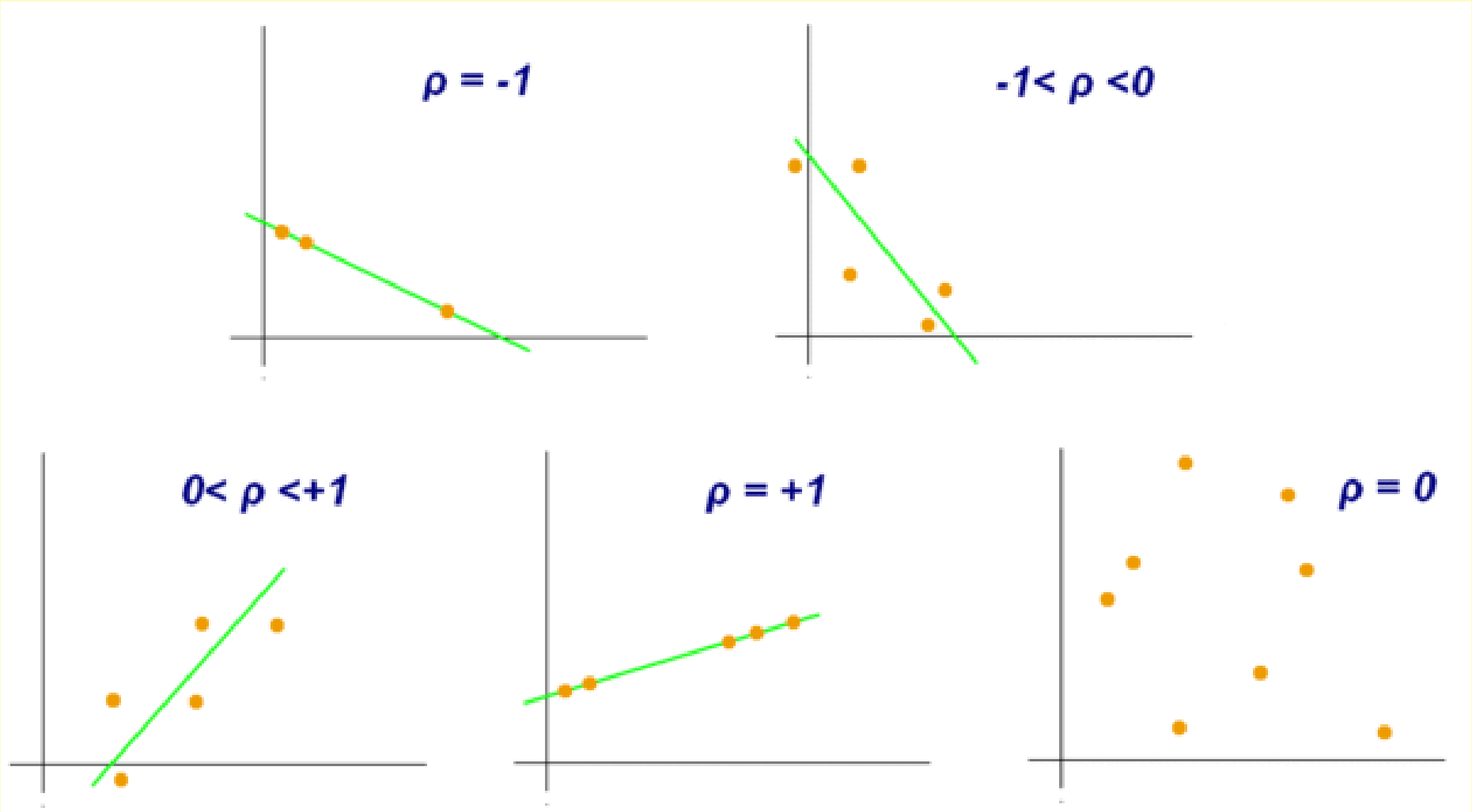


Baseline Results

Corelation between complexity of child utterances (horizontal axis) and the mother's CDS (vertical axis) in the Adam corpus:



Intermezzo: Correlation



Correlation

1 Perfect positive correlation: for each increase on the x-value we can see an increase in the y-value (a straight line with positive slope)

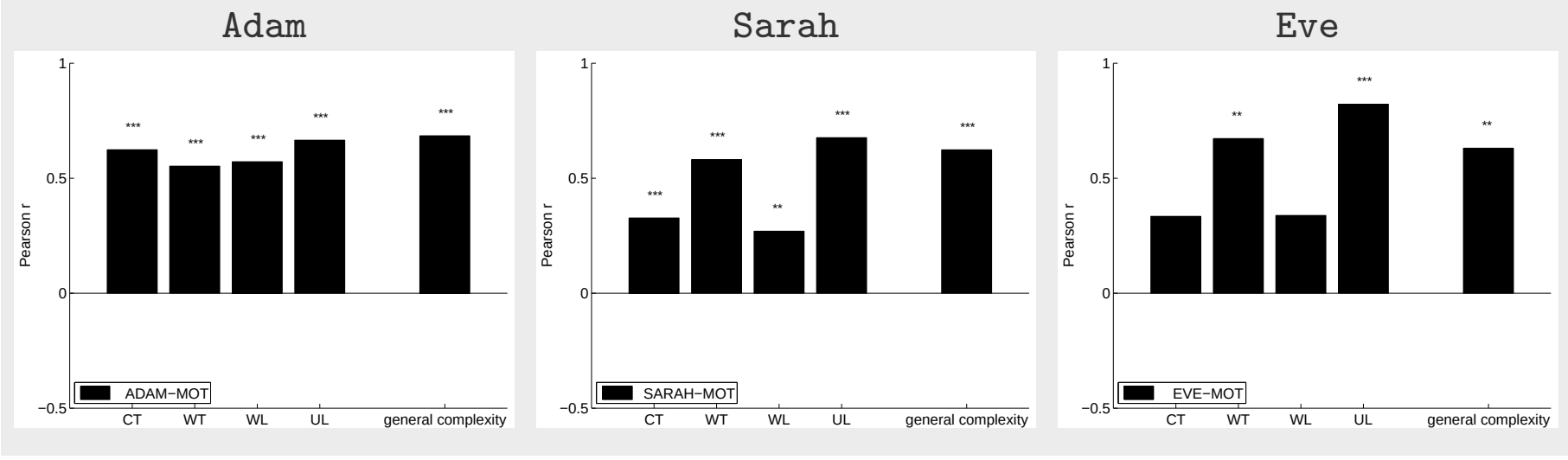
-1 Perfect negative correlation: for each increase on the x-value we can see a decrease in the y-value (a straight line with negative slope)

0 No correlation: knowing the x-value tells you nothing about y-value (a flat straight line, or a cloud of points with no upward or downward direction)

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

$$s_X = \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Baseline Results across Corpora



Correlations are robust across measures and child-mother pairs.

http://www.illc.uva.nl/laco/clas/dighum14