lexical density & readability

TM-BOOTCAMP @ KU

kristoffer l nielbo
kln@cas.au.dk
github.com/kln-courses/tmg_bootcamp
tmbootcamp.slack.com

DAI|IMC|AARHUS UNIVERSITY

AARHUS UNIVERSITET

IMC
INTERACTING MINDS CENTRE

`type-token ratio`

$$TTR = \frac{\sum w_{lex}}{\sum Fr(w_i)} \times 100 \tag{1}$$

`information theory`

$$H = -\sum_{i=1}^{n} p_i \times \log_2(p_i) \tag{2}$$

$$p_i = \frac{Fr(w_i)}{\sum_i^n Fr(w_i)} \tag{3}$$

– average amount of information in a text string in terms of its units (e.g., characters, words, n-grams) is closely related to other measures of lexical variety og density (readability)

AARHUS UNIVERSITET

IMC
INTERACTING MINDS CENTRE

läsbarhetsindex or LIX

$$O = \sum_{i=1}^{n} Fr(w_i) \tag{4}$$

$$P = Fr(\circ) \tag{5}$$

$$L = \sum_{i=1}^{n} Fr(w_i^{>6}) \tag{6}$$

$$LIX = \frac{O}{P} \times \frac{L \times 100}{O} \tag{7}$$

LIX	KEY
$\geq$55	very difficult (academic literature)
45-54	difficult (popular science)
35-44	35-44 middle (newspapers)
25-34	25-34 easy (fiction)
$\leq$24	$<$24 very easy (childrens literature)