

SWELLEX, PRODUCTIVE L2 LEARNER VOCABULARY AND MORE

David Alfter

University of Gothenburg

Språk-
BANKEN



UNIVERSITY OF GOTHENBURG

It all started with...

a demo

<https://spraakbanken.gu.se/larka/texteval>

Texteval

- Text analysis platform
- Assessment of learner written and expert written texts

Texteval

- Machine learning and readability measures
 - ML: predict overall level of text
 - Readability measures
 - Number of sentences
 - Number of tokens
 - Average sentence/token length
 - LIX score

Texteval

- Word-level CEFR highlighting
 - Depends on graded word lists
 - Out-of-vocabulary words

Word lists

- SVALex (François et al., 2016)
 - *COCTAILL corpus (Volodina et al., 2014)*
 - *Receptive knowledge*
- SweLLex (Volodina et al., 2016)
 - *SweLL corpus (Volodina et al., 2016)*
 - *Productive knowledge*
- Kelly list (Volodina and Kokkinakis, 2012)
 - *L1 web corpus*

Word list format (SVALex & SweLlex)

Lemma	POS-tag	A1	A2	B1	B2	C1	Total
bil	NN_UTR	430.218	1234.207	728.9847	422.283	363.5446	618.8567
överge	VB	0	0	7.3203	24.5182	39.6516	17.2695
rättvisa	NN_UTR	0	0	3.6601	25.6189	26.4344	13.6602
kilo	NN_NEU	0	302.0833	145.1229	65.0611	13.2172	89.8907
resa	VB	166.300	375.2582	450.3526	298.4905	330.4297	356.362
låg	JJ	0	49.315	125.922	217.3103	252.1311	156.126
så klart	ABM_MWE	0	16.2635	81.6019	45.5033	13.2172	38.1738

Word list format (Kelly list)

ID	88
Raw Freq	2624.032
Word per Million	23017.26
CEFR level	A1
Source	SweWaC
Grammar marker	att
Item	vara (vardagl. va)
POS	verb
Example	e.g. var så god!

Word list format (Kelly list)

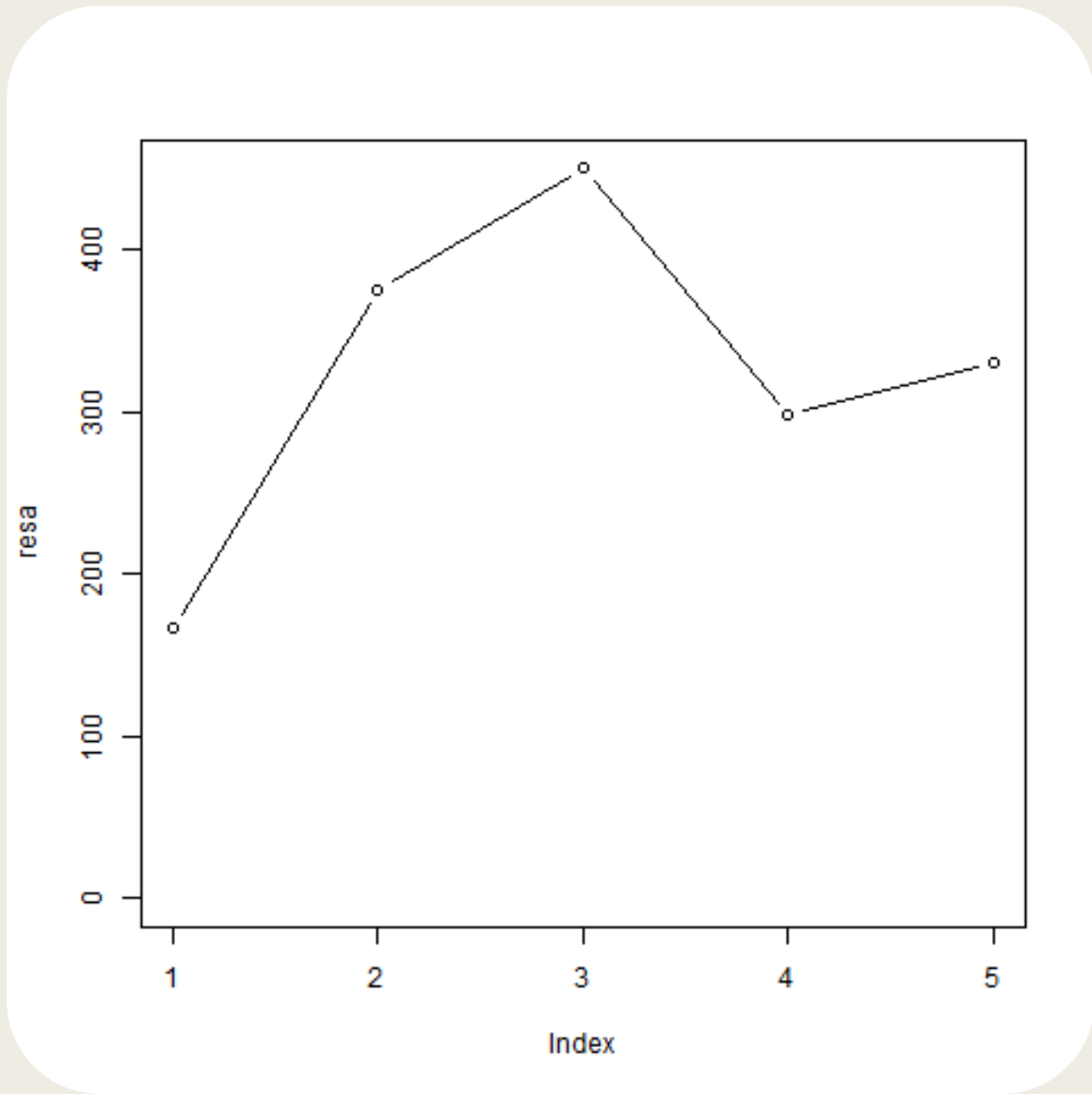
ID	88
Raw Freq	2624.032
Word per Million	23017.26
CEFR level	A1
Source	SweWaC
Grammar marker	att
Item	vara (vardagl. va)
POS	verb
Example	e.g. var så god!

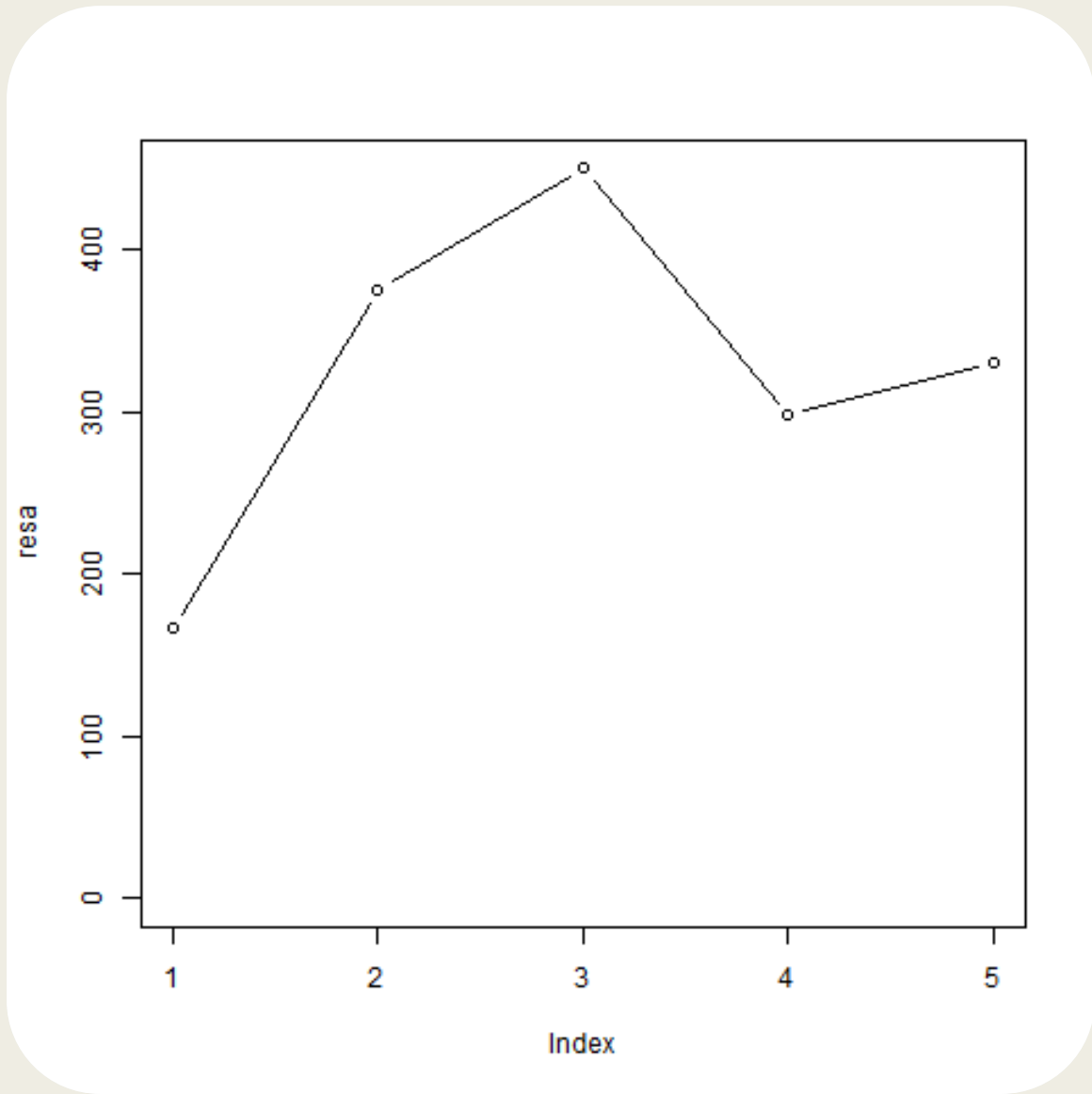
Mapping from distributions to labels

- Different possible approaches:
 - *First occurrence (Gala et al., 2013; Gala et al., 2014)*
 - *Maximum*
 - *Thresholding (Alfter et al., 2016)*
 - ...

Mapping distributions to levels

Lemma	POS-tag	A1	A2	B1	B2	C1	Total
bil	NN_UTR	430.218	1234.207	728.9847	422.283	363.5446	618.8567
överge	VB	0	0	7.3203	24.5182	39.6516	17.2695
rättvisa	NN_UTR	0	0	3.6601	25.6189	26.4344	13.6602
kilo	NN_NEU	0	302.0833	145.1229	65.0611	13.2172	89.8907
resa	VB	166.300	375.2582	450.3526	298.4905	330.4297	356.362
låg	JJ	0	49.315	125.922	217.3103	252.1311	156.126
så klart	ABM_MWE	0	16.2635	81.6019	45.5033	13.2172	38.1738

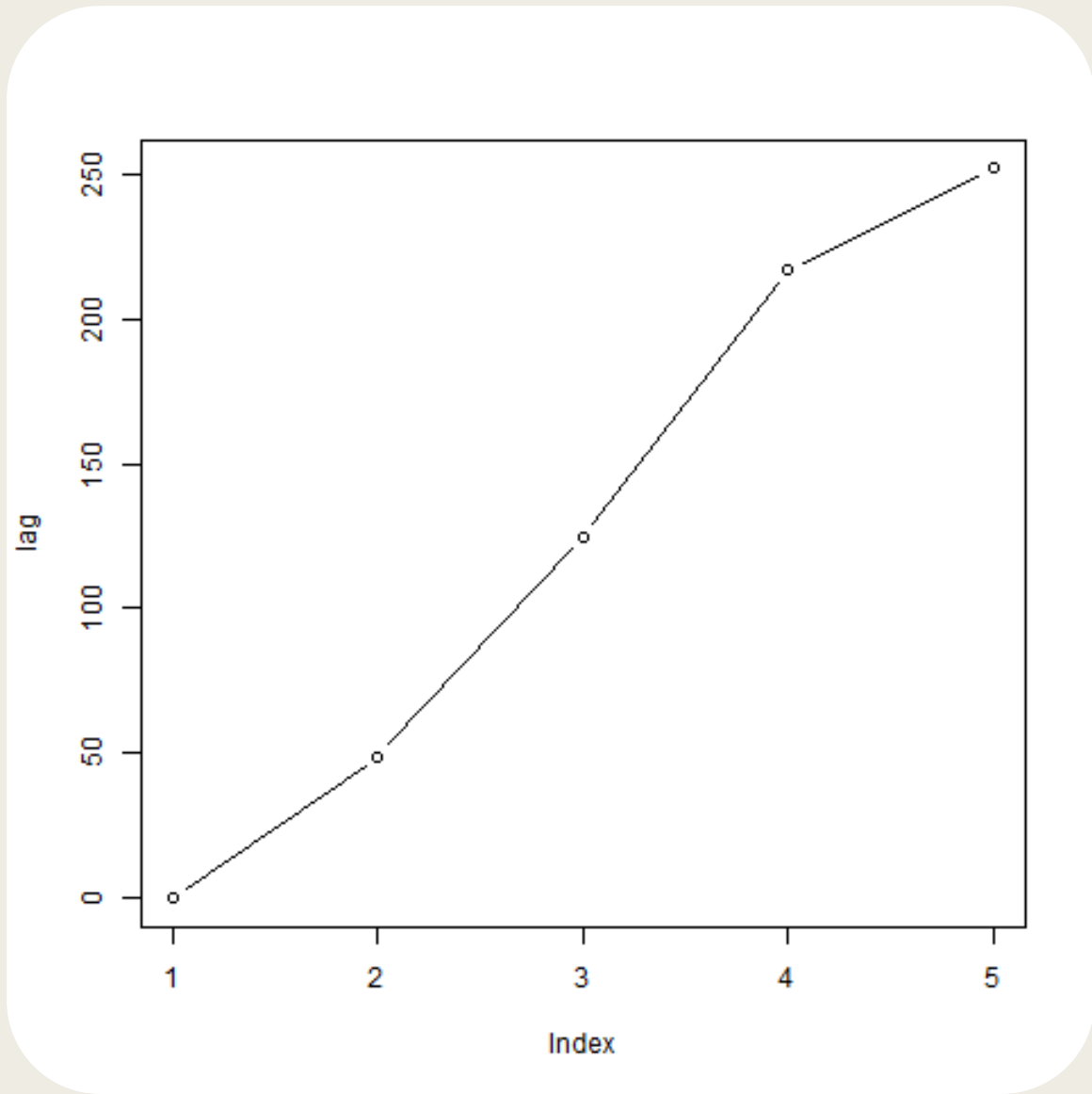


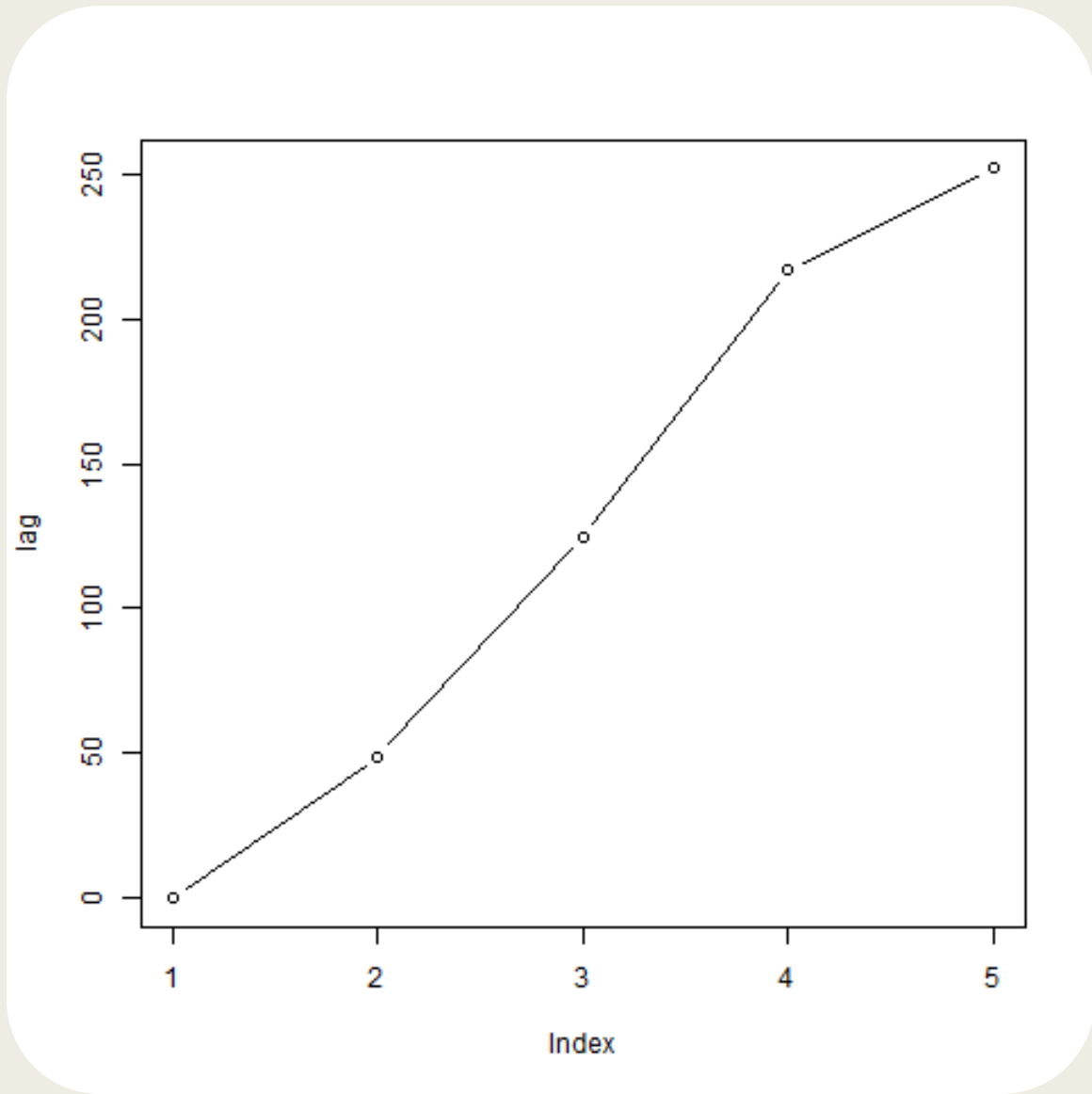


First occurrence: A1
Maximum: B1
Threshold: A2

Mapping distributions to levels

Lemma	POS-tag	A1	A2	B1	B2	C1	Total
bil	NN_UTR	430.218	1234.207	728.9847	422.283	363.5446	618.8567
överge	VB	0	0	7.3203	24.5182	39.6516	17.2695
rättvisa	NN_UTR	0	0	3.6601	25.6189	26.4344	13.6602
kilo	NN_NEU	0	302.0833	145.1229	65.0611	13.2172	89.8907
resa	VB	166.300	375.2582	450.3526	298.4905	330.4297	356.362
låg	JJ	0	49.315	125.922	217.3103	252.1311	156.126
så klart	ABM_MWE	0	16.2635	81.6019	45.5033	13.2172	38.1738



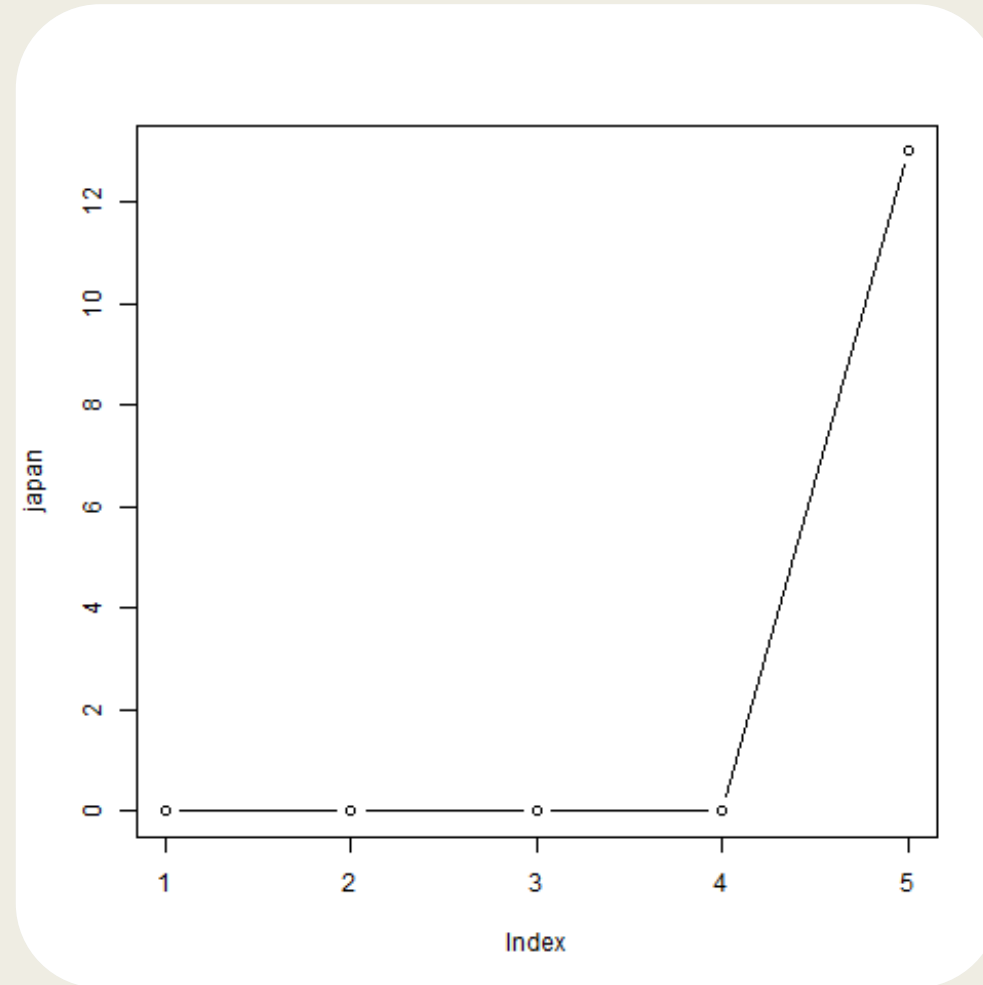


First occurrence: A2
Maximum: C1
Threshold: B2

Sparse data

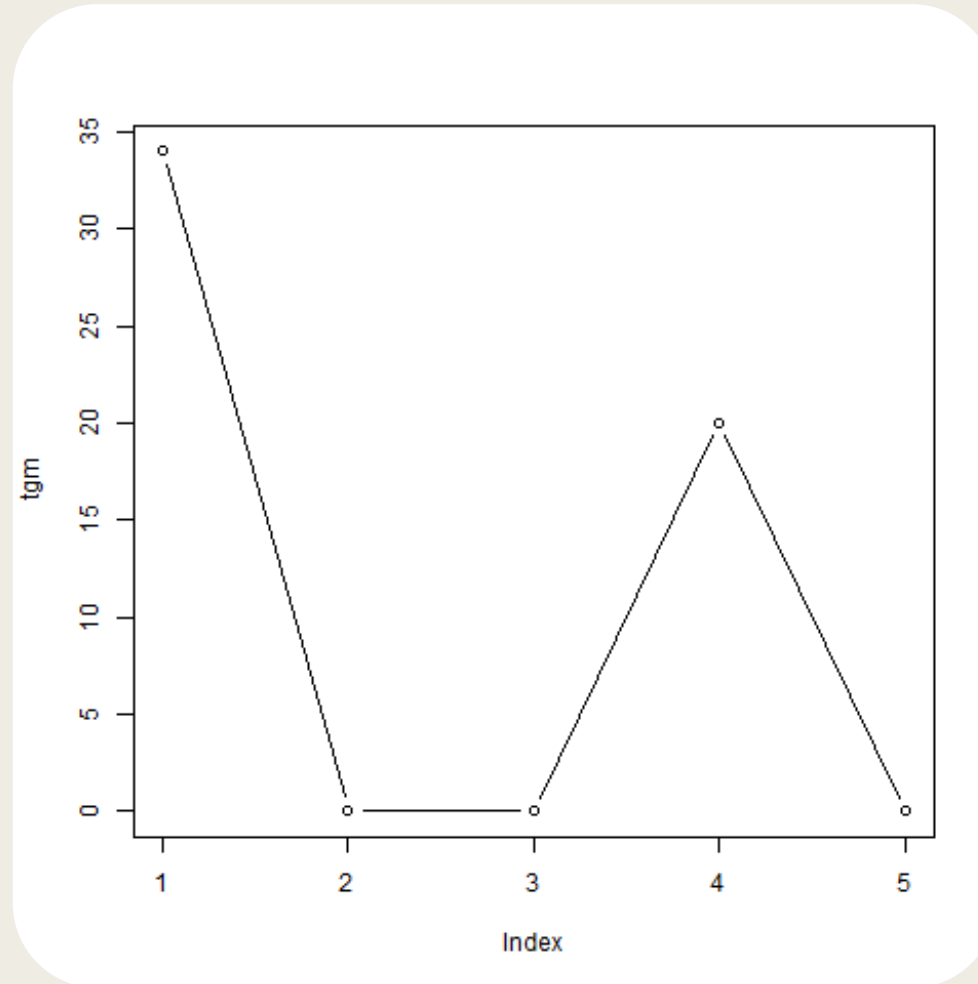
- Word lists created from sparse data

Japan (PM)



C1?

Trädgårdsmästare (NN)



A1?

Out-of-vocabulary words

- katt (cat)

Single-word lexical complexity

- Word length
- Syllables
- Suffix length
- Gender
- Homonymy
- Polysemy
- Compounds
- N-grams

Feature extraction

- Input to machine learning algorithm
- Different algorithms tested
 - *SVM*
 - *Logistic regression*
 - *MLP*
- Different combinations of features tested

Topic distributions

- CEFR proficiency levels correspond roughly to topics
 - *A1: introductions, greetings*
 - *A2: personal life, family*
 - *B1: school, leisure, personal interests*

(Council of Europe, 2001, p. 26)

Topic distributions

- Extract topic lists from
 - *COCTAILL corpus: 33 topics*
 - *Swedish FrameNet (SweFN++): 1010 topics*
- Retain only most predictive words per list (TF-IDF)
- Add topic distribution to feature vector

Results SweLLex

	A1	A2	B1	B2	C1
A1	102	37	21	9	3
A2	41	147	58	33	28
B1	19	56	157	80	37
B2	9	46	71	120	32
C1	8	26	44	58	207

Results SweLLex

- Accuracy: 50%
- F1 score: 0.51

Results SVALex

	A1	A2	B1	B2	C1
A1	244	64	26	15	10
A2	34	342	174	121	42
B1	9	134	474	277	73
B2	7	71	197	1774	172
C1	2	25	67	231	1037

Results SVALex

- Accuracy: 69%
- F1 score: 0.65

Siwoco

- Automatic prediction of single word lexical complexity
- MLP classifier

- Demo

<https://spraakbanken.gu.se/larkalabb/siwoco>

Validation through crowdsourcing

- Idea: use crowd sourcing to validate correctness of target level assignment
- Learner proficiency model predicts current learner level
- Check response time and accuracy for items of different levels
 - *Items of learner level: Faster response time and higher accuracy*
 - *Items of (predicted) learner level with lower response time/lower accuracy possibly higher level*

Resource creation

- Self-updating and self-validating graded vocabulary list
 - *New unseen words graded by algorithm*
 - *Untested words given to learners*
 - *Validation/estimation of level through crowdsourcing*
 - *Improvement of the resource*

Questions? Comments?

david.alfter@gu.se