SINGLE WORD LEXICAL COMPLEXITY FOR HOLISTIC TEXT ASSESSMENT

David Alfter
University of Gothenburg

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

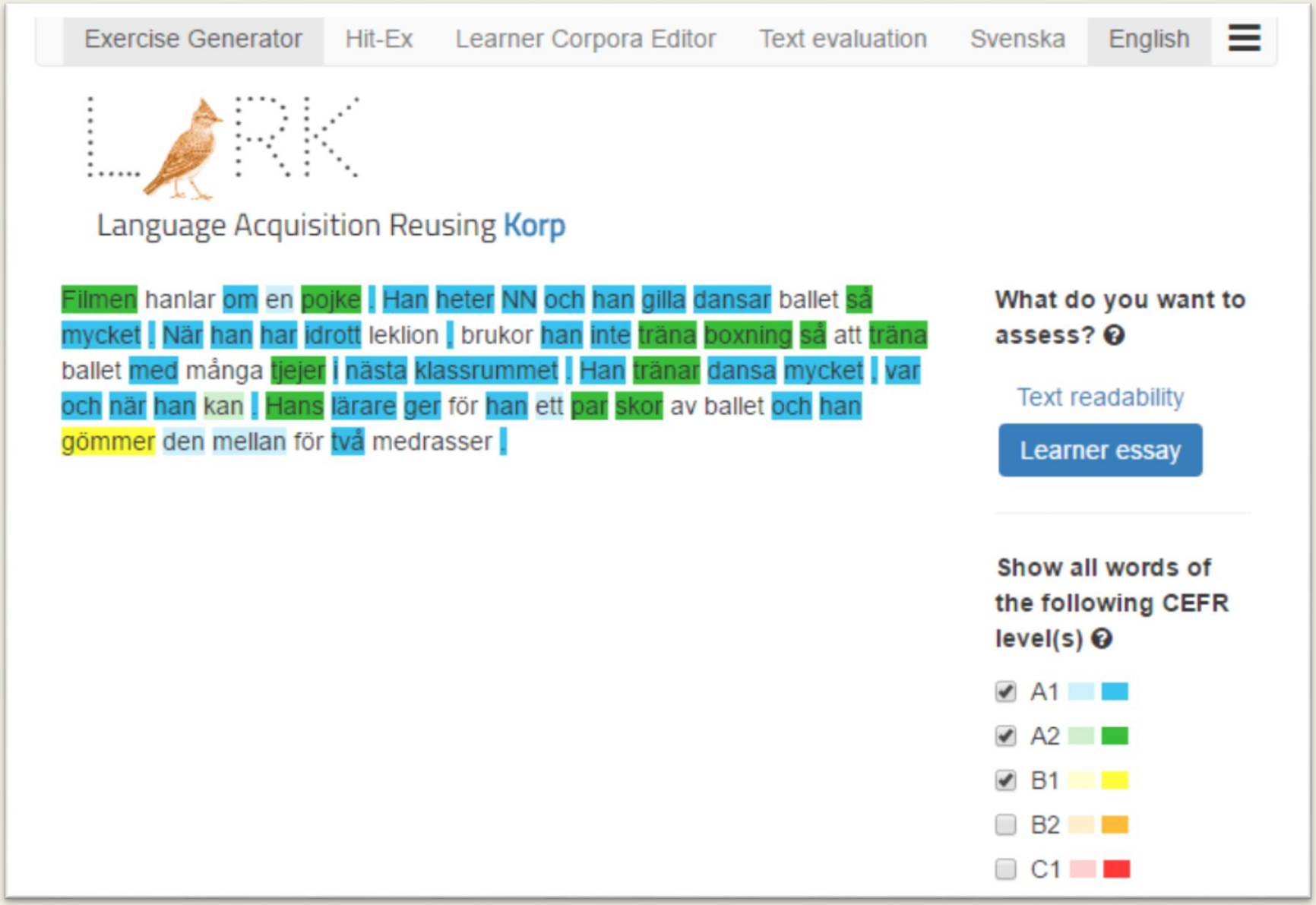
It all started with...

a demo

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

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG



Språk-BANKEN

UNIVERSITY OF GOTHENBURG

Texteval

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

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

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

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Texteval

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

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

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!

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Mapping from distributions to labels

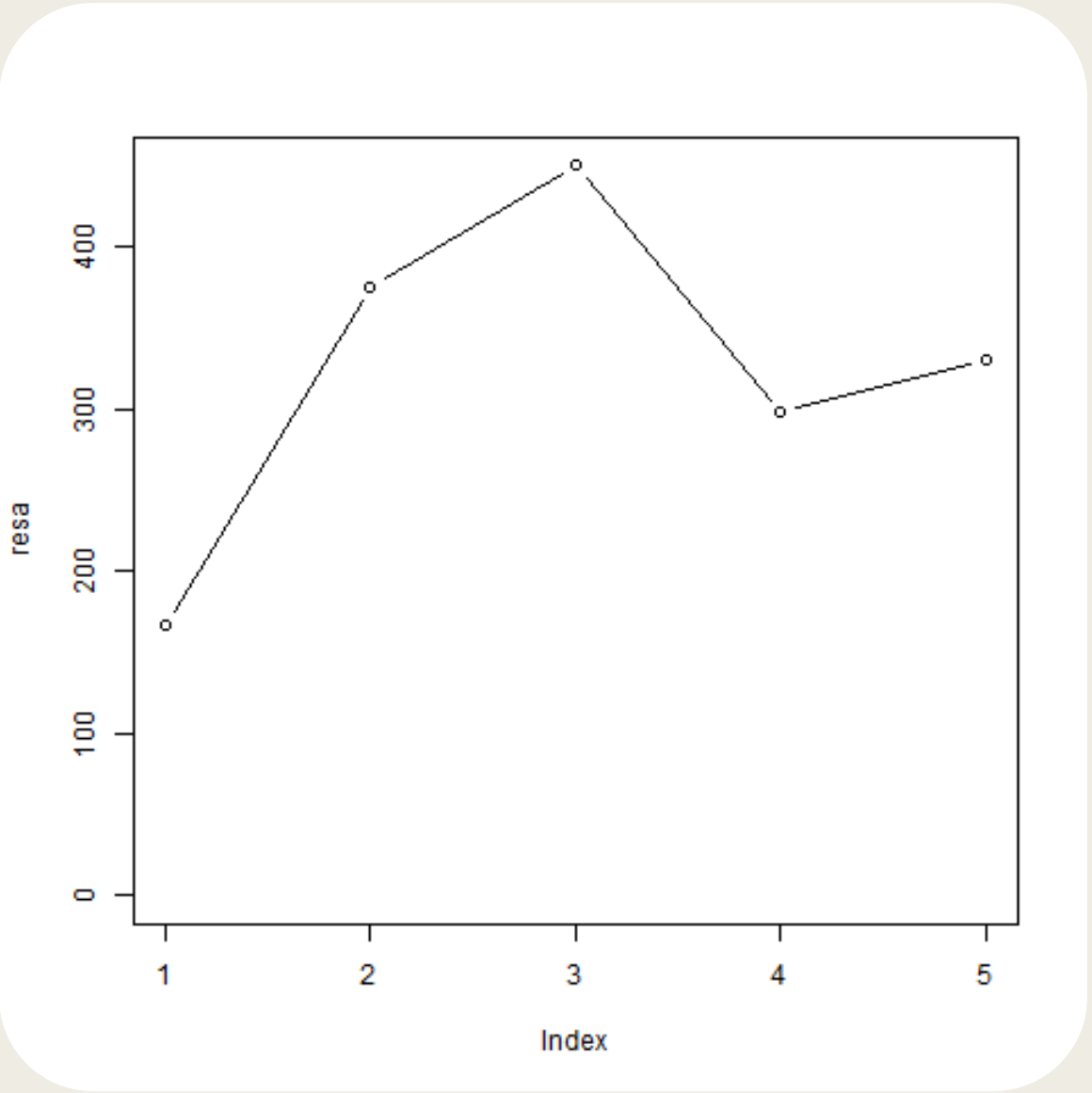
- Different possible approaches:
  - *First occurrence (Gala et al., 2013; Gala et al., 2014)*
  - *Maximum*
  - *Thresholding (Alfter et al., 2016)*
  - *…*

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

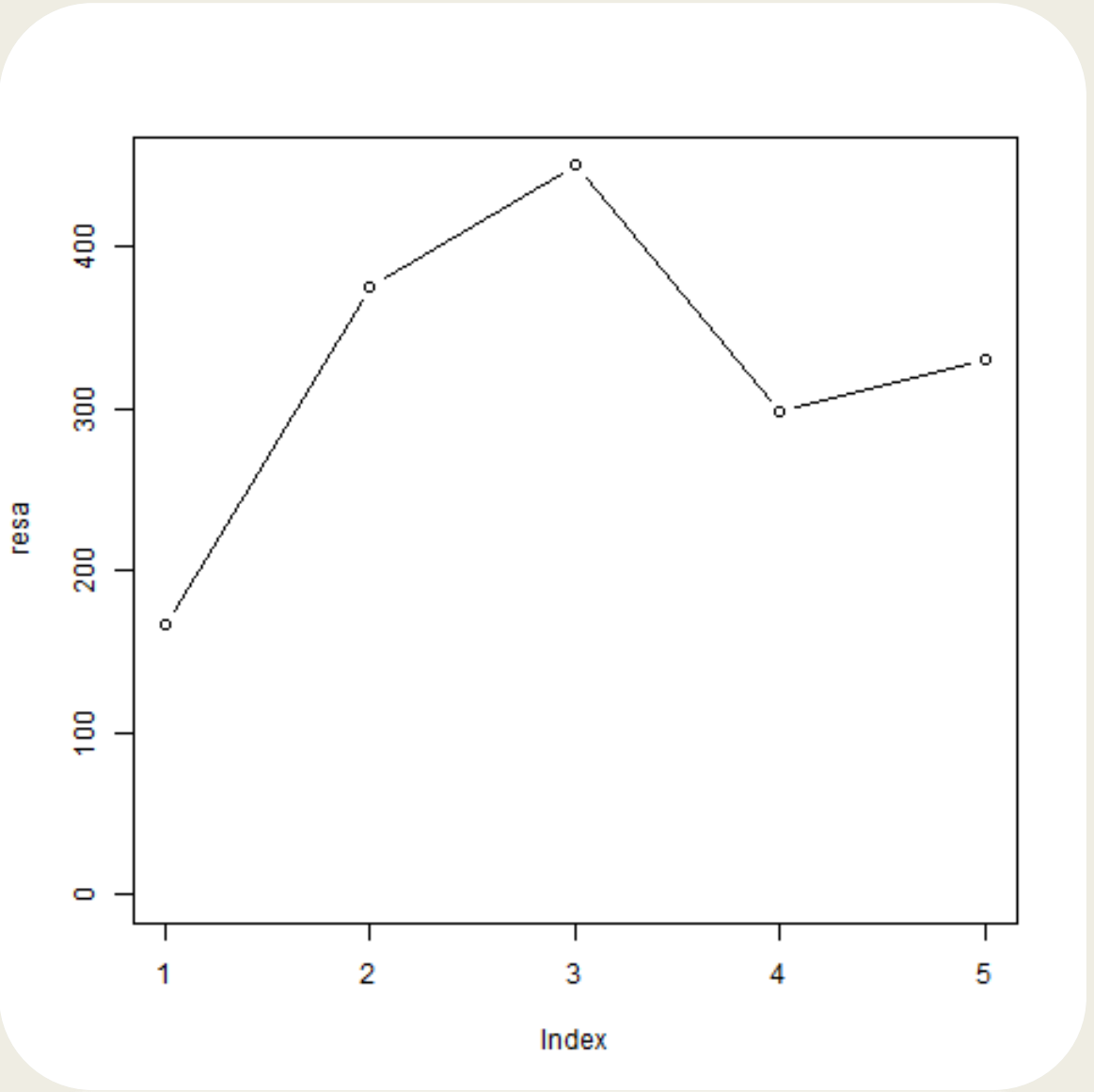
Språk-
BANKEN

UNIVERSITY OF GOTHENBURG



Språk-
BANKEN

UNIVERSITY OF GOTHENBURG



First occurrence: A1
Maximum: B1
Threshold: A2

Språk-
BANKEN

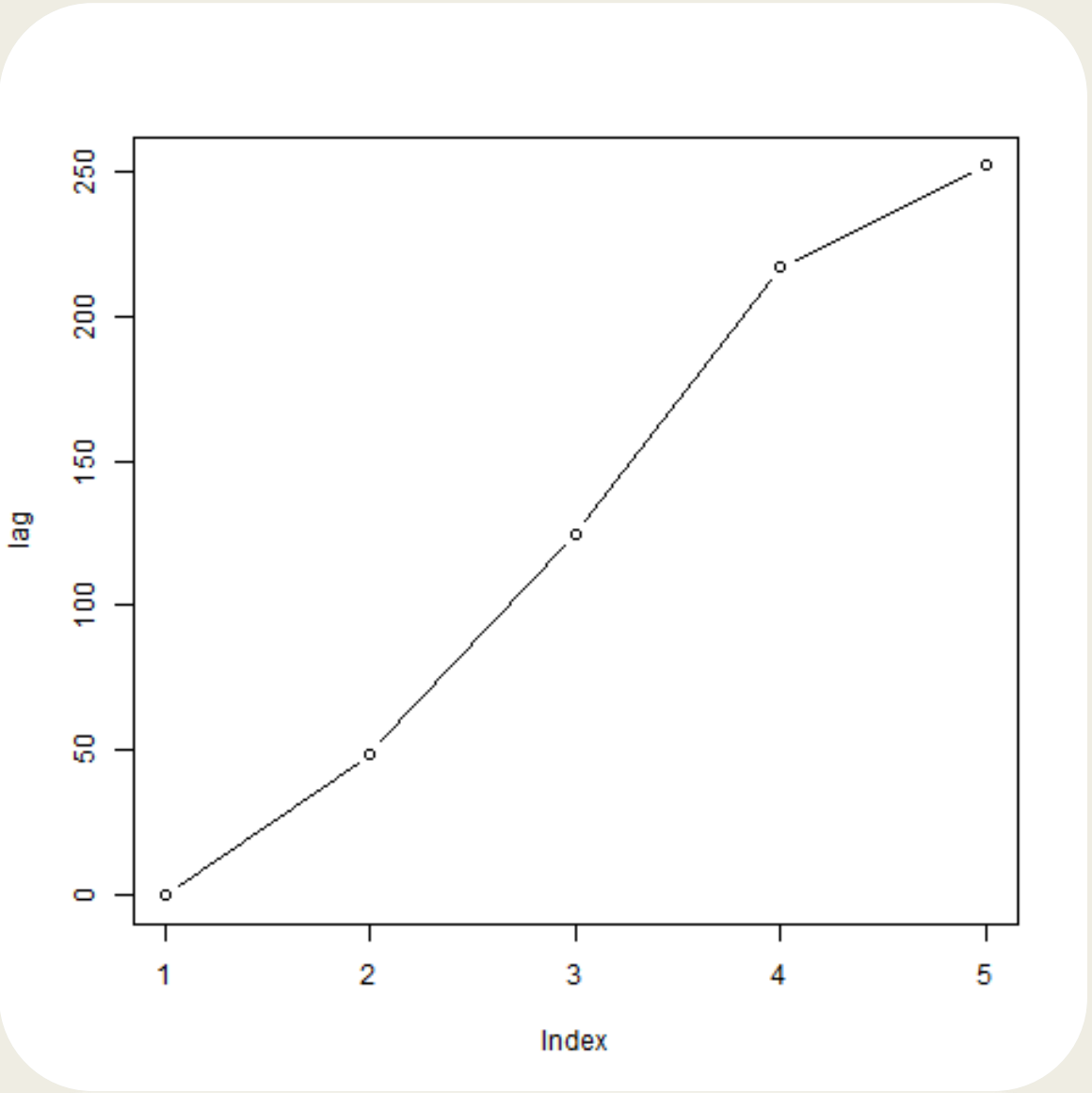
UNIVERSITY OF GOTHENBURG

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

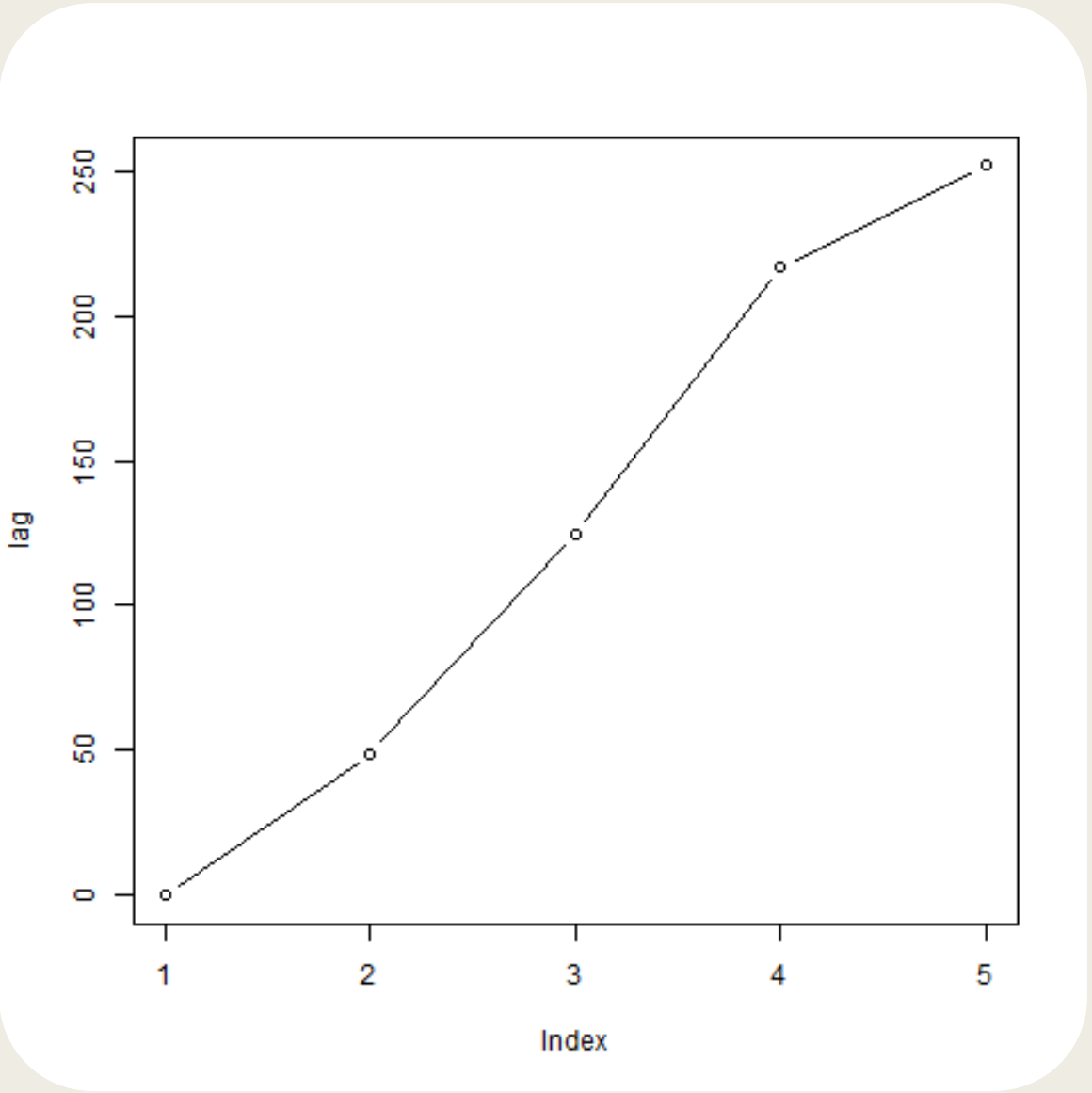
Språk-BANKEN

UNIVERSITY OF GOTHENBURG



Språk-
BANKEN

UNIVERSITY OF GOTHENBURG



First occurrence: A2
Maximum: C1
Threshold: B2

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

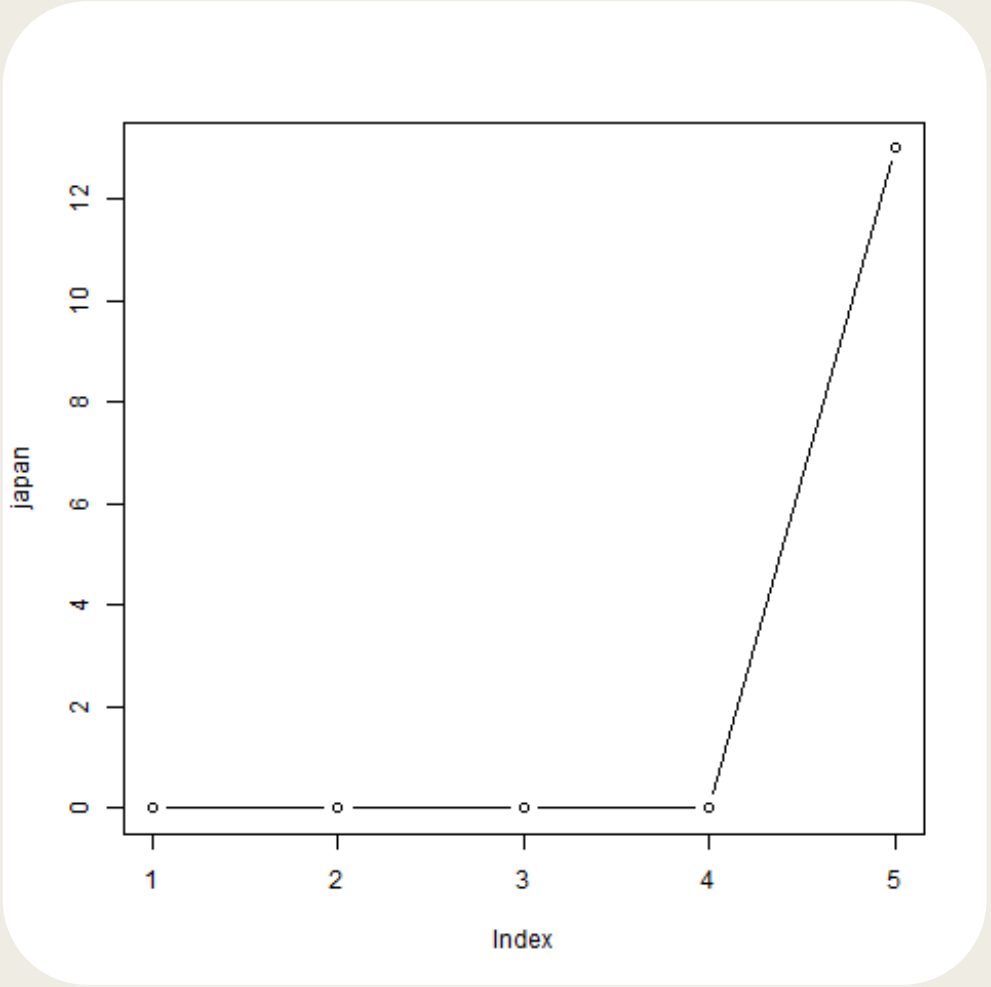
Sparse data

- Word lists created from sparse data

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Japan (PM)

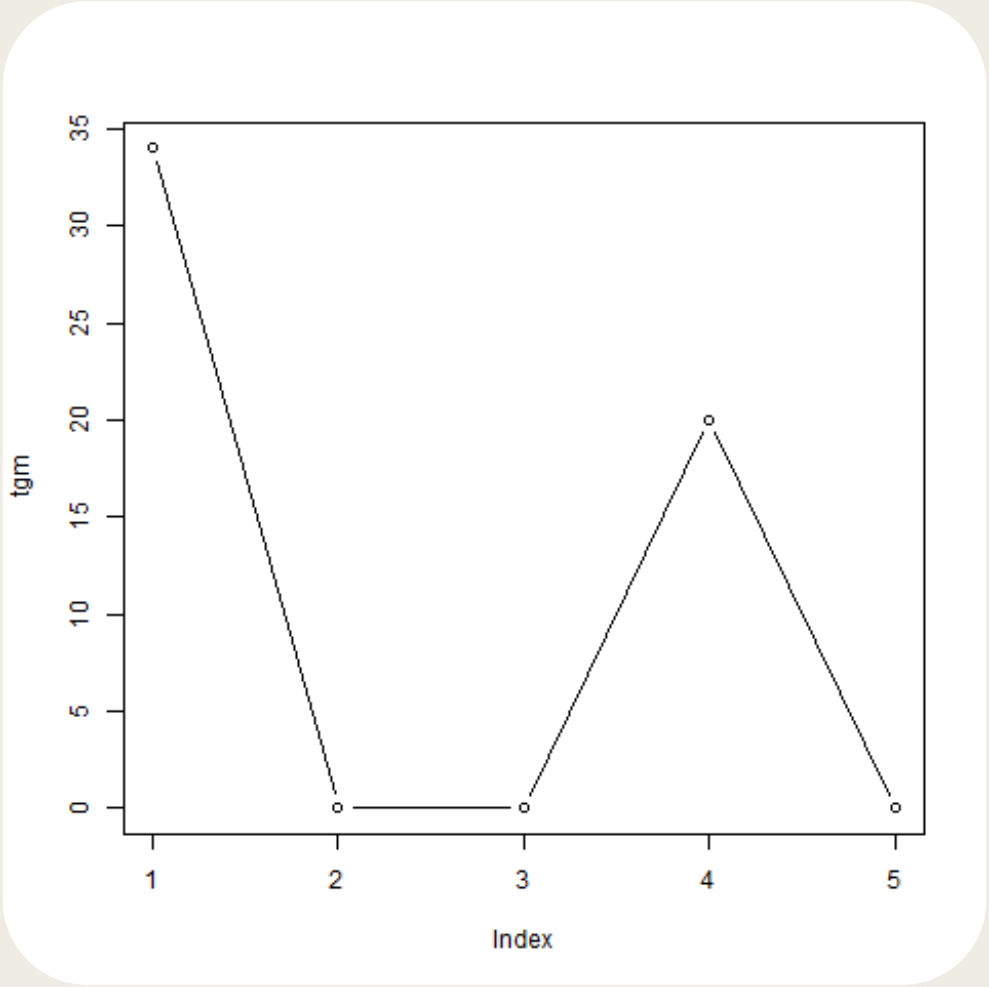


C1?

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Trädgårdsmästare (NN)



A1?

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Out-of-vocabulary words

- katt (cat)

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Single-word lexical complexity

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

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Feature extraction

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

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

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

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

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**

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Results SweLLex

- Accuracy: 50%
- F1 score: 0.51

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

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**

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Results SVALex

- Accuracy: 69%
- F1 score: 0.65

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Siwoco

- Automatic prediction of single word lexical complexity
- MLP classifier

- Demo

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

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Conclusion

- Single word lexical complexity as factor in holistic text assessment

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

Questions? Comments?

david.alfter@gu.se

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG

References

- Alfter, D., Bizzoni, Y., Agebjörn, A., Volodina, E., & Pilán, I. (2016, November). From Distributions to Labels: A Lexical Proficiency Analysis using Learner Corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016* (No. 130, pp. 1-7). Linköping University Electronic Press.
- Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching. *Assessment*.
- François, T., Volodina, E., Pilán, I., & Tack, A. (2016, May). SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. In *LREC*.
- Gala, N., François, T., & Fairon, C. (2013). Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper., Tallin, Estonia*.
- Gala, N., François, T., Bernhard, D., & Fairon, C. (2014, July). Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN'2014* (pp. 91-102).
- Volodina, E., Pilán, I., Eide, S. R., & Heidarsson, H. (2014, November). You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University* (No. 107). Linköping University Electronic Press.
- Volodina, E., Pilán, I., Llozhi, L., Degryse, B., & François, T. (2016, November). SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016* (No. 130, pp. 76-84). Linköping University Electronic Press.
- Volodina, E., & Kokkinakis, S. J. (2012). Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *LREC* (pp. 1040-1046).

Språk-
BANKEN

UNIVERSITY OF GOTHENBURG