LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Towards an Improved Methodology for Automated Readability Prediction

Philip van Oosten, Dries Tanghe, Véronique Hoste

LT^3 Language and Translation Technology Team
Faculty of Translation Studies
University College Ghent
{philip.vanoosten, dries.tanghe, veronique.hoste}@hogent.be

LREC 2010 - 19 May 2010

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Outline

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Outline

1. Introduction: the concept of readability (prediction)
2. Experiments on large corpora

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Outline

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Outline: introduction

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Introduction: readability

What is readability?

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Introduction: readability

What is readability?

- "The characteristic of text that makes readers willing to read on." [McLaughlin1969]

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Introduction: readability

What is readability?

- "The characteristic of text that makes readers willing to read on." [McLaughlin1969]
- "The reading proficiency that is needed for text comprehension." [Staphorsius1994]

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Introduction: readability

What is readability?

- "The characteristic of text that makes readers willing to read on." [McLaughlin1969]
- "The reading proficiency that is needed for text comprehension." [Staphorsius1994]
- "What makes some texts easier to read than others." [DuBay2004]

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Introduction: readability prediction

What is readability prediction?

- Automated analysis of an unseen text
- Result: readability assessment
  - score
  - grade level
  - ranking
- Sometimes used for assistance in writing process

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Introduction: readability prediction

What is readability prediction?

- Automated analysis of an unseen text
- Result: readability assessment
  - score
  - grade level
  - ranking
- Sometimes used for assistance in writing process

What is a readability formula?

- A readability prediction method
- Mathematical formula consisting of
  - constants $\longrightarrow$ weights;
  - variables $\longrightarrow$ text characteristics.
- e.g. Flesch Reading Ease [Flesch1948]:
  $207 - avgsentencelen - 85 * avgnumsyl$

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Introduction: content of our paper

In-depth analysis of 12 existing readability formulas

- Behaviour when tested on large corpora:
  - correlation matrices
  - Principal Component Analysis (PCA)
- Methodological (in)validity:
  - collinearity tests

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Introduction: content of our paper

In-depth analysis of 12 existing readability formulas

- Behaviour when tested on large corpora:
  - correlation matrices
  - Principal Component Analysis (PCA)
- Methodological (in)validity:
  - collinearity tests

Our findings

- Readability formulas are more or less interchangeable
  - all formulas are based on a limited set of variables
  - regardless of the language for which they were designed (English, Dutch, Swedish)

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Outline: experiments

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Large-scale calculation of readability scores and text characteristics

Data sets

- Dutch Corpora
  - *Eindhoven Corpus*: 740k tokens, 5k fragments
  - *SoNaR*: 81M tokens, 213k texts
- English Corpora
  - *Penn Treebank*: 1M tokens, 2.5k texts
  - *British National Corpus*: 85M tokens, 3.1k texts

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Correlation matrices

Calculated correlations between

- characteristics – characteristics
- characteristics – formulas
- formulas – formulas

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

	brouwer	kincaid	fog	smog	ari	douma	flesch	lix	cli	cilt	clib	rgs	avgpolysylsent	avgwordlen	avgnumsyl	ppolysylword	ratiolongword	psw	avgsentencelen	freq3000	ttr
brouwer		-1,00	-0,92	-0,90	-0,93	0,91	0,91	-0,86	-0,66	-0,47	-0,35	-0,31	-0,90	-0,54	-0,68	-0,59	-0,52	0,71	-0,71	-0,18	-0,02
kincaid	-1,00		0,92	0,89	0,93	-0,89	-0,89	0,86	0,64	0,44	0,34	0,30	0,89	0,51	0,65	0,56	0,49	-0,74	0,74	0,16	0,00
fog	-0,92	0,92		0,99	0,82	-0,88	-0,88	0,81	0,64	0,45	0,38	0,33	0,99	0,54	0,70	0,76	0,55	-0,60	0,60	0,22	0,10
smog	-0,90	0,89	0,99		0,79	-0,89	-0,89	0,80	0,66	0,49	0,38	0,33	1,00	0,57	0,74	0,81	0,58	-0,54	0,54	0,23	0,09
ari	-0,93	0,93	0,82	0,79		-0,77	-0,76	0,87	0,66	0,48	0,36	0,32	0,79	0,53	0,50	0,41	0,47	-0,80	0,80	0,18	-0,01
douma	0,91	-0,89	-0,88	-0,89	-0,77		1,00	-0,82	-0,82	-0,68	-0,46	-0,42	-0,89	-0,76	-0,92	-0,80	-0,72	0,39	-0,39	-0,34	-0,14
flesch	0,91	-0,89	-0,88	-0,89	-0,76	1,00		-0,82	-0,82	-0,68	-0,46	-0,42	-0,89	-0,76	-0,92	-0,80	-0,72	0,39	-0,39	-0,34	-0,14
lix	-0,86	0,86	0,81	0,80	0,87	-0,82	-0,82		0,75	0,58	0,47	0,44	0,80	0,66	0,65	0,57	0,77	-0,57	0,57	0,33	0,13
cli	-0,66	0,64	0,64	0,66	0,66	-0,82	-0,82	0,75		0,93	0,59	0,55	0,66	0,98	0,85	0,72	0,85	-0,14	0,14	0,51	0,22
cilt	-0,47	0,44	0,45	0,49	0,48	-0,68	-0,68	0,58	0,93		0,61	0,52	0,49	0,96	0,78	0,65	0,79	0,05	-0,05	0,53	0,22
clib	-0,35	0,34	0,38	0,38	0,36	-0,46	-0,46	0,47	0,59	0,61		0,55	0,38	0,60	0,49	0,46	0,57	-0,02	0,02	0,53	0,81
rgs	-0,31	0,30	0,33	0,33	0,32	-0,42	-0,42	0,44	0,55	0,52	0,55		0,33	0,55	0,45	0,41	0,53	-0,02	0,02	0,98	0,39
avgpolysylsent	-0,90	0,89	0,99	1,00	0,79	-0,89	-0,89	0,80	0,66	0,49	0,38	0,33		0,57	0,74	0,81	0,58	-0,54	0,54	0,23	0,09
avgwordlen	-0,54	0,51	0,54	0,57	0,53	-0,76	-0,76	0,66	0,98	0,96	0,60	0,55	0,57		0,85	0,74	0,86	0,02	-0,02	0,55	0,27
avgnumsyl	-0,68	0,65	0,70	0,74	0,50	-0,92	-0,92	0,65	0,85	0,78	0,49	0,45	0,74	0,85		0,89	0,80	-0,04	0,04	0,44	0,22
ppolysylword	-0,59	0,56	0,76	0,81	0,41	-0,80	-0,80	0,57	0,72	0,65	0,46	0,41	0,81	0,74	0,89		0,72	0,00	0,00	0,41	0,27
ratiolongword	-0,52	0,49	0,55	0,58	0,47	-0,72	-0,72	0,77	0,85	0,79	0,57	0,53	0,58	0,86	0,80	0,72		0,01	-0,01	0,53	0,32
psw	0,71	-0,74	-0,60	-0,54	-0,80	0,39	0,39	-0,57	-0,14	0,05	-0,02	-0,02	-0,54	0,02	-0,04	0,00	0,01		-1,00	0,16	0,22
avgsentencelen	-0,71	0,74	0,60	0,54	0,80	-0,39	-0,39	0,57	0,14	-0,05	0,02	0,02	0,54	-0,02	0,04	0,00	-0,01	-1,00		-0,16	-0,22
freq3000	-0,18	0,16	0,22	0,23	0,18	-0,34	-0,34	0,33	0,51	0,53	0,53	0,98	0,23	0,55	0,44	0,41	0,53	0,16	-0,16		0,42
ttr	-0,02	0,00	0,10	0,09	-0,01	-0,14	-0,14	0,13	0,22	0,22	0,81	0,39	0,09	0,27	0,22	0,27	0,32	0,22	-0,22	0,42	

Correlation matrix

- Formulas: upper / left
- Characteristics : lower / right
- light green: $\rho > 0.8$
- dark green: $0.8 \geq \rho > 0.6$

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

	brouwer	kincaid	fog	smog	ari	douma	flesch	lix	cli	cilt	clib	rgs	avgpolysylsent	avgwordlen	avgnumsyl	ppolysylword	ratiolongword	psw	avgsentencelen	freq3000	ttr
brouwer		-1,00	-0,92	-0,90	-0,93	0,91	0,91	-0,86	-0,66	-0,47	-0,35	-0,31	-0,90	-0,54	-0,68	-0,59	-0,52	0,71	-0,71	-0,18	-0,02
kincaid	-1,00		0,92	0,89	0,93	-0,89	-0,89	0,86	0,64	0,44	0,34	0,30	0,89	0,51	0,65	0,56	0,49	-0,74	0,74	0,16	0,00
fog	-0,92	0,92		0,99	0,82	-0,88	-0,88	0,81	0,64	0,45	0,38	0,33	0,99	0,54	0,70	0,76	0,55	-0,60	0,60	0,22	0,10
smog	-0,90	0,89	0,99		0,79	-0,89	-0,89	0,80	0,66	0,49	0,38	0,33	1,00	0,57	0,74	0,81	0,58	-0,54	0,54	0,23	0,09
ari	-0,93	0,93	0,82	0,79		-0,77	-0,76	0,87	0,66	0,48	0,36	0,32	0,79	0,53	0,50	0,41	0,47	-0,80	0,80	0,18	-0,01
douma	0,91	-0,89	-0,88	-0,89	-0,77		1,00	-0,82	-0,82	-0,68	-0,46	-0,42	-0,89	-0,76	-0,92	-0,80	-0,72	0,39	-0,39	-0,34	-0,14
flesch	0,91	-0,89	-0,88	-0,89	-0,76	1,00		-0,82	-0,82	-0,68	-0,46	-0,42	-0,89	-0,76	-0,92	-0,80	-0,72	0,39	-0,39	-0,34	-0,14
lix	-0,86	0,86	0,81	0,80	0,87	-0,82	-0,82		0,75	0,58	0,47	0,44	0,80	0,66	0,65	0,57	0,77	-0,57	0,57	0,33	0,13
cli	-0,66	0,64	0,64	0,66	0,66	-0,82	-0,82	0,75		0,93	0,59	0,55	0,66	0,98	0,85	0,72	0,85	-0,14	0,14	0,51	0,22
cilt	-0,47	0,44	0,45	0,49	0,48	-0,68	-0,68	0,58	0,93		0,61	0,52	0,49	0,96	0,78	0,65	0,79	0,05	-0,05	0,53	0,22
clib	-0,35	0,34	0,38	0,38	0,36	-0,46	-0,46	0,47	0,59	0,61		0,55	0,38	0,60	0,49	0,46	0,57	-0,02	0,02	0,53	0,81
rgs	-0,31	0,30	0,33	0,33	0,32	-0,42	-0,42	0,44	0,55	0,52	0,55		0,33	0,55	0,45	0,41	0,53	-0,02	0,02	0,98	0,39
avgpolysylsent	-0,90	0,89	0,99	1,00	0,79	-0,89	-0,89	0,80	0,66	0,49	0,38	0,33		0,57	0,74	0,81	0,58	-0,54	0,54	0,23	0,09
avgwordlen	-0,54	0,51	0,54	0,57	0,53	-0,76	-0,76	0,66	0,98	0,96	0,60	0,55	0,57		0,85	0,74	0,86	0,02	-0,02	0,55	0,27
avgnumsyl	-0,68	0,65	0,70	0,74	0,50	-0,92	-0,92	0,65	0,85	0,78	0,49	0,45	0,74	0,85		0,89	0,80	-0,04	0,04	0,44	0,22
ppolysylword	-0,59	0,56	0,76	0,81	0,41	-0,80	-0,80	0,57	0,72	0,65	0,46	0,41	0,81	0,74	0,89		0,72	0,00	0,00	0,41	0,27
ratiolongword	-0,52	0,49	0,55	0,58	0,47	-0,72	-0,72	0,77	0,85	0,79	0,57	0,53	0,58	0,86	0,80	0,72		0,01	-0,01	0,53	0,32
psw	0,71	-0,74	-0,60	-0,54	-0,80	0,39	0,39	-0,57	-0,14	0,05	-0,02	-0,02	-0,54	0,02	-0,04	0,00	0,01		-1,00	0,16	0,22
avgsentencelen	-0,71	0,74	0,60	0,54	0,80	-0,39	-0,39	0,57	0,14	-0,05	0,02	0,02	0,54	-0,02	0,04	0,00	-0,01	-1,00		-0,16	-0,22
freq3000	-0,18	0,16	0,22	0,23	0,18	-0,34	-0,34	0,33	0,51	0,53	0,53	0,98	0,23	0,55	0,44	0,41	0,53	0,16	-0,16		0,42
ttr	-0,02	0,00	0,10	0,09	-0,01	-0,14	-0,14	0,13	0,22	0,22	0,81	0,39	0,09	0,27	0,22	0,27	0,32	0,22	-0,22	0,42	

Observations

- Formulas correlate strongly with each other

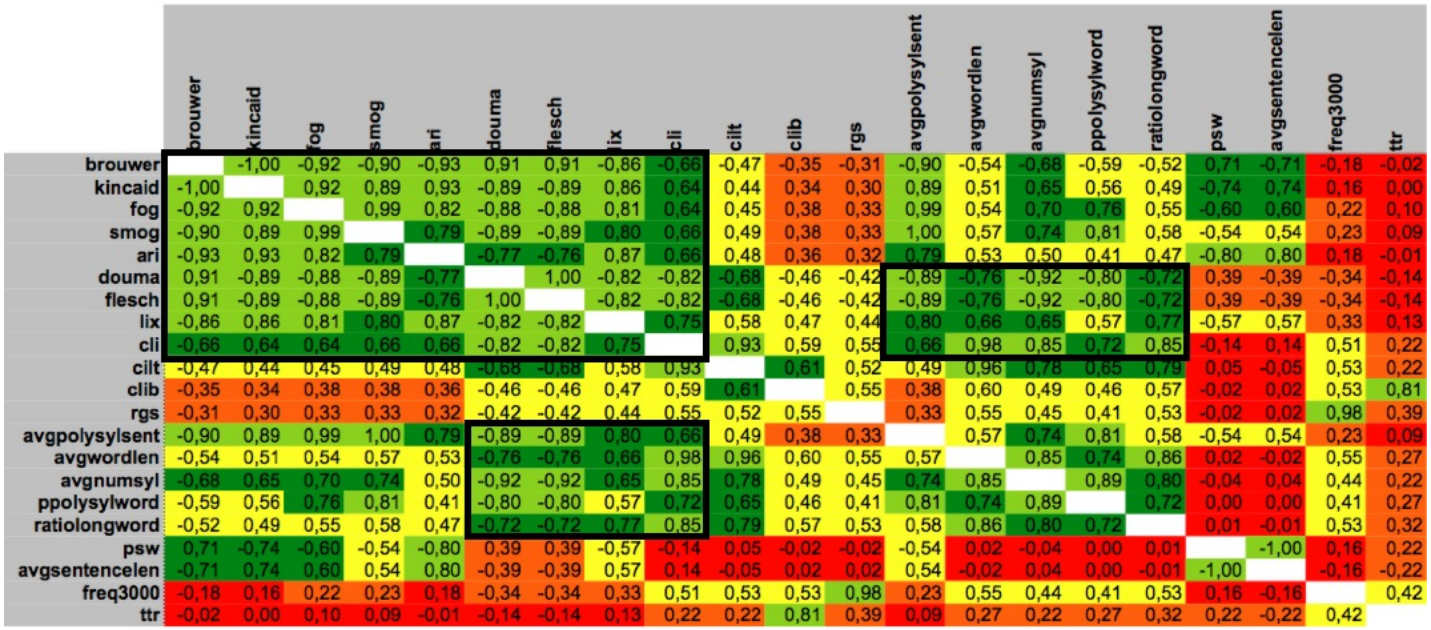
LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

	brouwer	kincaid	fog	smog	ari	douma	flesch	lix	cli	cilt	clib	rgs	avgpolysylsent	avgwordlen	avgnumsyl	ppolysylword	ratiolongword	psw	avgsentencelen	freq3000	ttr
brouwer		-1,00	-0,92	-0,90	-0,93	0,91	0,91	-0,86	-0,66	-0,47	-0,35	-0,31	-0,90	-0,54	-0,68	-0,59	-0,52	0,71	-0,71	-0,18	-0,02
kincaid	-1,00		0,92	0,89	0,93	-0,89	-0,89	0,86	0,64	0,44	0,34	0,30	0,89	0,51	0,65	0,56	0,49	-0,74	0,74	0,16	0,00
fog	-0,92	0,92		0,99	0,82	-0,88	-0,88	0,81	0,64	0,45	0,38	0,33	0,99	0,54	0,70	0,76	0,55	-0,60	0,60	0,22	0,10
smog	-0,90	0,89	0,99		0,79	-0,89	-0,89	0,80	0,66	0,49	0,38	0,33	1,00	0,57	0,74	0,81	0,58	-0,54	0,54	0,23	0,09
ari	-0,93	0,93	0,82	0,79		-0,77	-0,76	0,87	0,66	0,48	0,36	0,32	0,79	0,53	0,50	0,41	0,47	-0,80	0,80	0,18	-0,01
douma	0,91	-0,89	-0,88	-0,89	-0,77		1,00	-0,82	-0,82	-0,68	-0,46	-0,42	-0,89	-0,76	-0,92	-0,80	-0,72	0,39	-0,39	-0,34	-0,14
flesch	0,91	-0,89	-0,88	-0,89	-0,76	1,00		-0,82	-0,82	-0,68	-0,46	-0,42	-0,89	-0,76	-0,92	-0,80	-0,72	0,39	-0,39	-0,34	-0,14
lix	-0,86	0,86	0,81	0,80	0,87	-0,82	-0,82		0,75	0,58	0,47	0,44	0,80	0,66	0,65	0,57	0,77	-0,57	0,57	0,33	0,13
cli	-0,66	0,64	0,64	0,66	0,66	-0,82	-0,82	0,75		0,93	0,59	0,55	0,66	0,98	0,85	0,72	0,85	-0,14	0,14	0,51	0,22
cilt	-0,47	0,44	0,45	0,49	0,48	-0,68	-0,68	0,58	0,93		0,61	0,52	0,49	0,96	0,78	0,65	0,79	0,05	-0,05	0,53	0,22
clib	-0,35	0,34	0,38	0,38	0,36	-0,46	-0,46	0,47	0,59	0,61		0,55	0,38	0,60	0,49	0,46	0,57	-0,02	0,02	0,53	0,81
rgs	-0,31	0,30	0,33	0,33	0,32	-0,42	-0,42	0,44	0,55	0,52	0,55		0,33	0,55	0,45	0,41	0,53	-0,02	0,02	0,98	0,39
avgpolysylsent	-0,90	0,89	0,99	1,00	0,79	-0,89	-0,89	0,80	0,66	0,49	0,38	0,33		0,57	0,74	0,81	0,58	-0,54	0,54	0,23	0,09
avgwordlen	-0,54	0,51	0,54	0,57	0,53	-0,76	-0,76	0,66	0,98	0,96	0,60	0,55	0,57		0,85	0,74	0,86	0,02	-0,02	0,55	0,27
avgnumsyl	-0,68	0,65	0,70	0,74	0,50	-0,92	-0,92	0,65	0,85	0,78	0,49	0,45	0,74	0,85		0,89	0,80	-0,04	0,04	0,44	0,22
ppolysylword	-0,59	0,56	0,76	0,81	0,41	-0,80	-0,80	0,57	0,72	0,65	0,46	0,41	0,81	0,74	0,89		0,72	0,00	0,00	0,41	0,27
ratiolongword	-0,52	0,49	0,55	0,58	0,47	-0,72	-0,72	0,77	0,85	0,79	0,57	0,53	0,58	0,86	0,80	0,72		0,01	-0,01	0,53	0,32
psw	0,71	-0,74	-0,60	-0,54	-0,80	0,39	0,39	-0,57	-0,14	0,05	-0,02	-0,02	-0,54	0,02	-0,04	0,00	0,01		-1,00	0,16	0,22
avgsentencelen	-0,71	0,74	0,60	0,54	0,80	-0,39	-0,39	0,57	0,14	-0,05	0,02	0,02	0,54	-0,02	0,04	0,00	-0,01	-1,00		-0,16	-0,22
freq3000	-0,18	0,16	0,22	0,23	0,18	-0,34	-0,34	0,33	0,51	0,53	0,53	0,98	0,23	0,55	0,44	0,41	0,53	0,16	-0,16		0,42
ttr	-0,02	0,00	0,10	0,09	-0,01	-0,14	-0,14	0,13	0,22	0,22	0,81	0,39	0,09	0,27	0,22	0,27	0,32	0,22	-0,22	0,42	

Observations

- Formulas correlate strongly with each other
- Regardless of language
- No adaptation, only rescaling

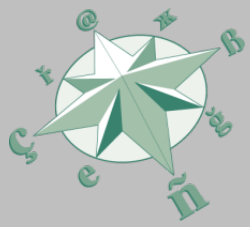


	brouwer	kincaid	fog	smog	ari	douma	flesch	lix	cli	cilt	clib	rgs	avgpolysylsent	avgwordlen	avgnumsyl	ppolysylword	ratiolongword	psw	avgsentencelen	freq3000	ttr
brouwer		-1,00	-0,92	-0,90	-0,93	0,91	0,91	-0,86	-0,66	-0,47	-0,35	-0,31	-0,90	-0,54	-0,68	-0,59	-0,52	0,71	-0,71	-0,18	-0,02
kincaid	-1,00		0,92	0,89	0,93	-0,89	-0,89	0,86	0,64	0,44	0,34	0,30	0,89	0,51	0,65	0,56	0,49	-0,74	0,74	0,16	0,00
fog	-0,92	0,92		0,99	0,82	-0,88	-0,88	0,81	0,64	0,45	0,38	0,33	0,99	0,54	0,70	0,76	0,55	-0,60	0,60	0,22	0,10
smog	-0,90	0,89	0,99		0,79	-0,89	-0,89	0,80	0,66	0,49	0,38	0,33	1,00	0,57	0,74	0,81	0,58	-0,54	0,54	0,23	0,09
ari	-0,93	0,93	0,82	0,79		-0,77	-0,76	0,87	0,66	0,48	0,36	0,32	0,79	0,53	0,50	0,41	0,47	-0,80	0,80	0,18	-0,01
douma	0,91	-0,89	-0,88	-0,89	-0,77		1,00	-0,82	-0,82	-0,68	-0,46	-0,42	-0,89	-0,76	-0,92	-0,80	-0,72	0,39	-0,39	-0,34	-0,14
flesch	0,91	-0,89	-0,88	-0,89	-0,76	1,00		-0,82	-0,82	-0,68	-0,46	-0,42	-0,89	-0,76	-0,92	-0,80	-0,72	0,39	-0,39	-0,34	-0,14
lix	-0,86	0,86	0,81	0,80	0,87	-0,82	-0,82		0,75	0,58	0,47	0,44	0,80	0,66	0,65	0,57	0,77	-0,57	0,57	0,33	0,13
cli	-0,66	0,64	0,64	0,66	0,66	-0,82	-0,82	0,75		0,93	0,59	0,55	0,66	0,98	0,85	0,72	0,85	-0,14	0,14	0,51	0,22
cilt	-0,47	0,44	0,45	0,49	0,48	-0,68	-0,68	0,58	0,93		0,61	0,52	0,49	0,96	0,78	0,65	0,79	0,05	-0,05	0,53	0,22
clib	-0,35	0,34	0,38	0,38	0,36	-0,46	-0,46	0,47	0,59	0,61		0,55	0,38	0,60	0,49	0,46	0,57	-0,02	0,02	0,53	0,81
rgs	-0,31	0,30	0,33	0,33	0,32	-0,42	-0,42	0,44	0,55	0,52	0,55		0,33	0,55	0,45	0,41	0,53	-0,02	0,02	0,98	0,39
avgpolysylsent	-0,90	0,89	0,99	1,00	0,79	-0,89	-0,89	0,80	0,66	0,49	0,38	0,33		0,57	0,74	0,81	0,58	-0,54	0,54	0,23	0,09
avgwordlen	-0,54	0,51	0,54	0,57	0,53	-0,76	-0,76	0,66	0,98	0,96	0,60	0,55	0,57		0,85	0,74	0,86	0,02	-0,02	0,55	0,27
avgnumsyl	-0,68	0,65	0,70	0,74	0,50	-0,92	-0,92	0,65	0,85	0,78	0,49	0,45	0,74	0,85		0,89	0,80	-0,04	0,04	0,44	0,22
ppolysylword	-0,59	0,56	0,76	0,81	0,41	-0,80	-0,80	0,57	0,72	0,65	0,46	0,41	0,81	0,74	0,89		0,72	0,00	0,00	0,41	0,27
ratiolongword	-0,52	0,49	0,55	0,58	0,47	-0,72	-0,72	0,77	0,85	0,79	0,57	0,53	0,58	0,86	0,80	0,72		0,01	-0,01	0,53	0,32
psw	0,71	-0,74	-0,60	-0,54	-0,80	0,39	0,39	-0,57	-0,14	0,05	-0,02	-0,02	-0,54	0,02	-0,04	0,00	0,01		-1,00	0,16	0,22
avgsentencelen	-0,71	0,74	0,60	0,54	0,80	-0,39	-0,39	0,57	0,14	-0,05	0,02	0,02	0,54	-0,02	0,04	0,00	-0,01	-1,00		-0,16	-0,22
freq3000	-0,18	0,16	0,22	0,23	0,18	-0,34	-0,34	0,33	0,51	0,53	0,53	0,98	0,23	0,55	0,44	0,41	0,53	0,16	-0,16		0,42
ttr	-0,02	0,00	0,10	0,09	-0,01	-0,14	-0,14	0,13	0,22	0,22	0,81	0,39	0,09	0,27	0,22	0,27	0,32	0,22	-0,22	0,42	

Observations

- Formulas correlate strongly with each other
- Regardless of language
- No adaptation, only rescaling
- Formulas correlate strongly with word length

LT3
UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION



LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Principal Component Analysis

The goal of PCA

- possibly correlated variables $\longrightarrow$ uncorrelated variables
- latent factors $\approx$ maximal variance

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

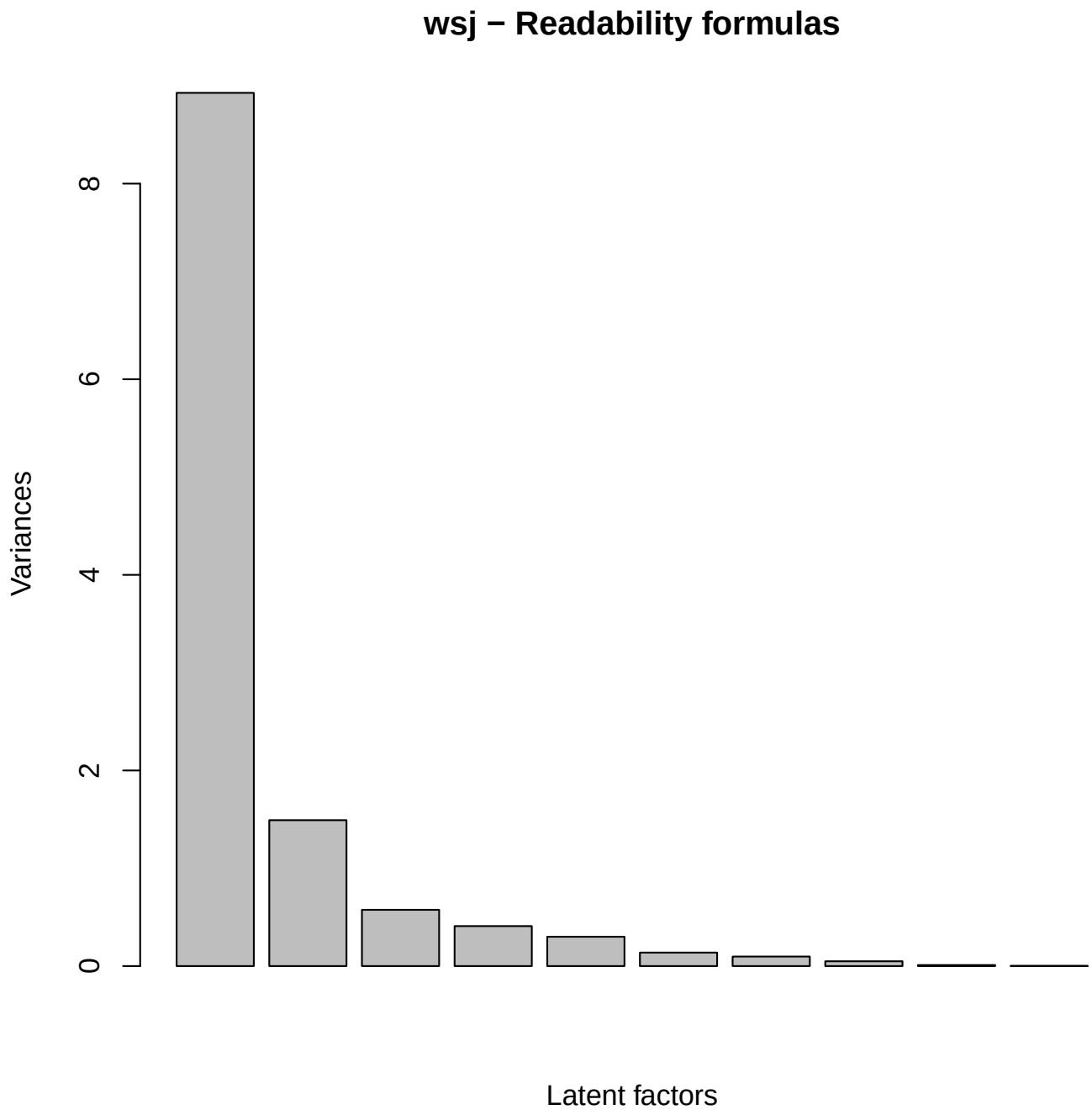
Principal Component Analysis

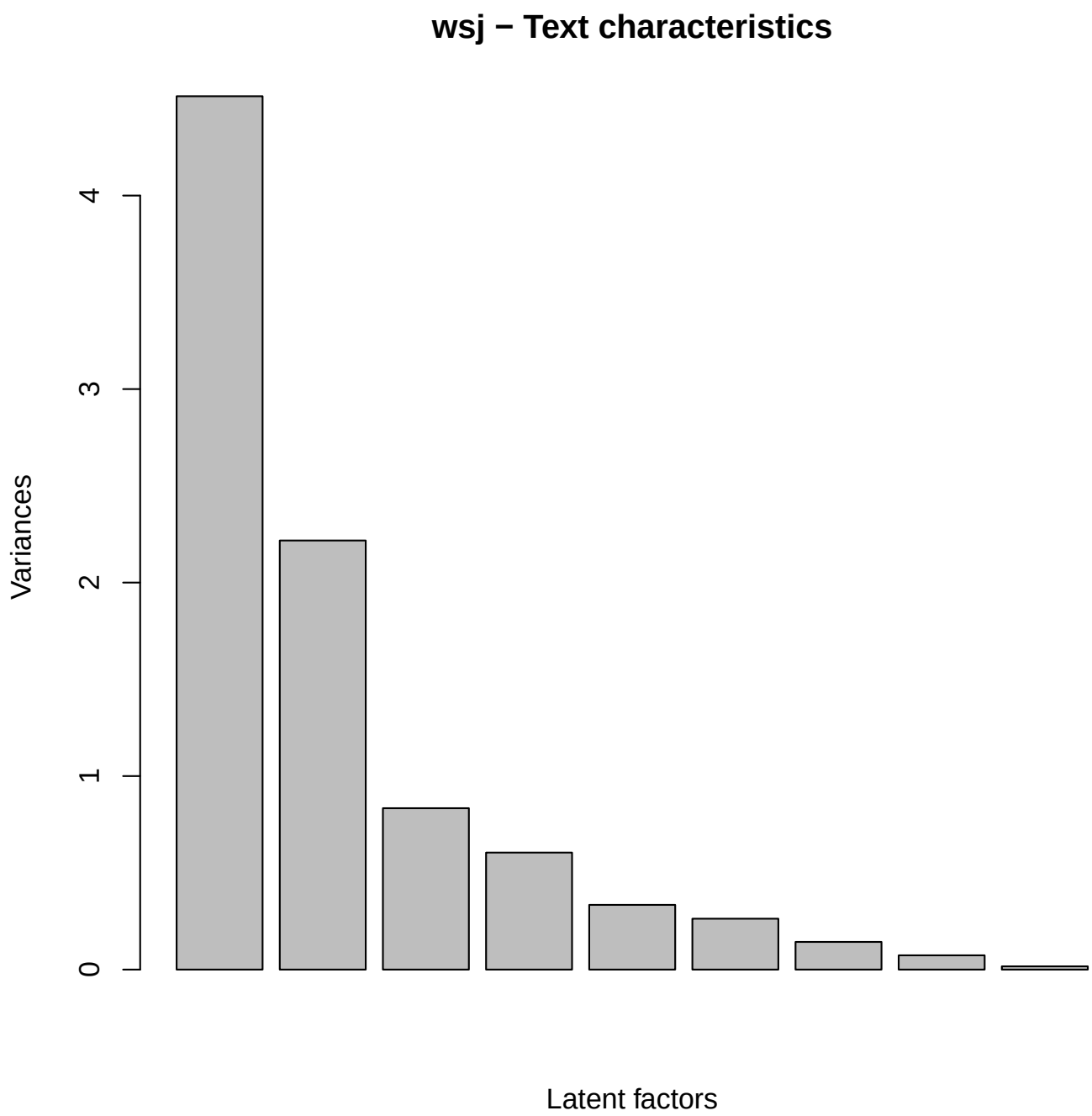
The goal of PCA

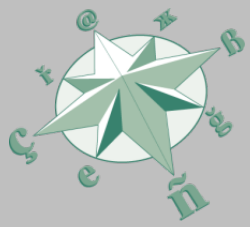
- possibly correlated variables $\rightarrow$ uncorrelated variables
- latent factors $\approx$ maximal variance

Performed PCA

- on all readability scores
- on all text characteristics


wsj − Readability formulas
Variances
0
2
4
6
8
Latent factors


wsj − Text characteristics
Variances
0
1
2
3
4
Latent factors



LT3

Collinearity tests [Belsley et al.1980]

Determining the interdependence of variables in a formula

- Readability formulas $<$ multiple regression
- Collinearity: variables are correlated
  - found in all formulas

  $\longrightarrow$ extrapolating to other data can be problematic

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Outline: discussion

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Towards an improved feature selection

Features that are used

- Strongly overlap
- Language-independent
- Strictly superficial

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Towards an improved feature selection

Features that are used

- Strongly overlap
- Language-independent
- Strictly superficial

Features that should be used

- On several levels
  - lexis, syntax, structural
- Language-dependent
  - e.g. compounding in Dutch
- Underlying causes of readability
  - e.g. cohesion and coherence

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

Towards an improved methodology

Existing readability formulas

- constructed and validated by means of limited corpora
  - typically a few hundred texts
- based on a single method of readability assessment
  - standard reading tests

LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

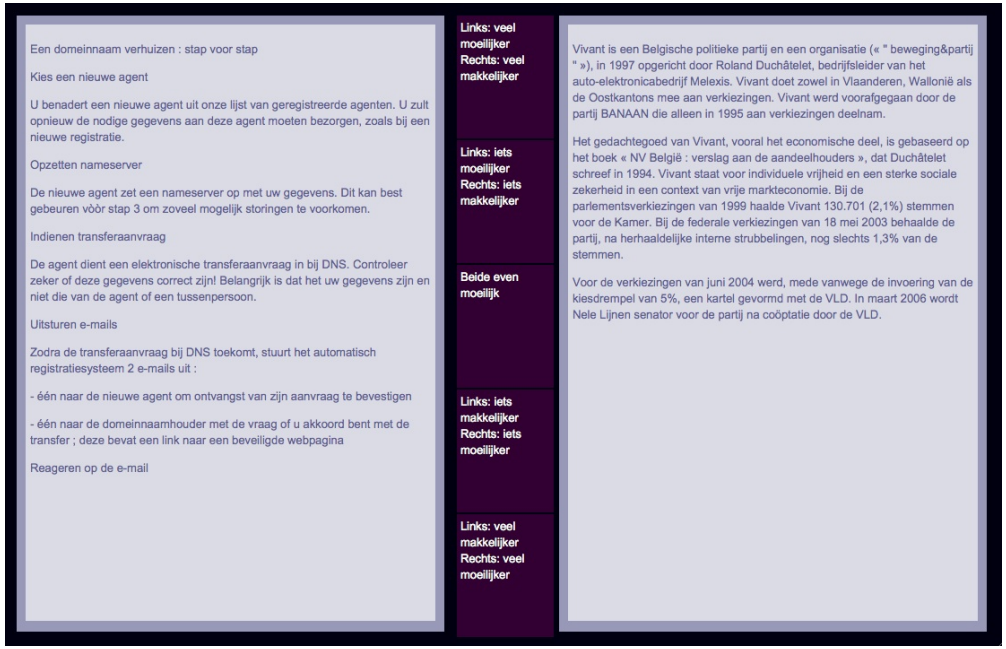
Towards an improved methodology

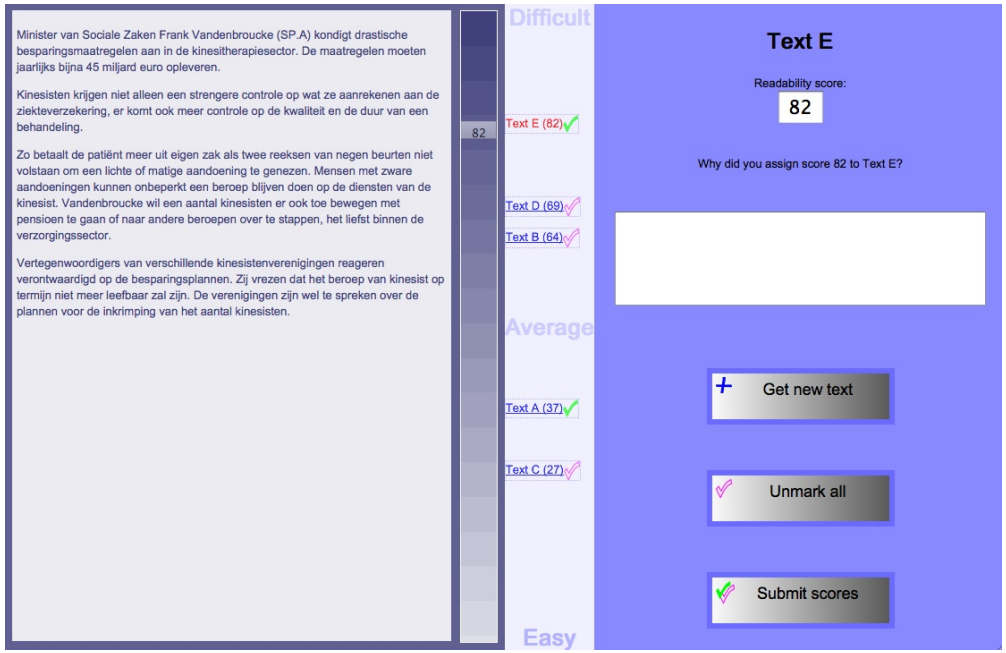
Existing readability formulas

- constructed and validated by means of limited corpora
  - typically a few hundred texts
- based on a single method of readability assessment
  - standard reading tests

Future readability prediction methods

- validation against large corpora
  - embedding in corpus research
- based on different kinds of readability assessment
  - collecting assessments from reading community





LT3

UNIVERSITY COLLEGE GHENT
MEMBER OF GHENT UNIVERSITY ASSOCIATION

References

David A. Belsley, Edwin Kuh, and Roy E. Welsch. 1980.
Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.
Wiley, August.

William H. DuBay. 2004.
The Principles of Readability.
Impact Information.

Rudolph Flesch. 1948.
A new readability yardstick.
Journal of Applied Psychology, 32(3):221–233.

G. Harry McLaughlin. 1969.
SMOG grading – a new readability formula.
Journal of Reading, pages 639–646.

Gerrit Staphorsius. 1994.
Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument.
Cito, Arnhem.