



LT³

Towards an Improved Methodology for Automated Readability Prediction

Philip van Oosten, Dries Tanghe, Véronique Hoste

LT³ 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

- 1 Introduction: the concept of readability (prediction)

Outline



LT3

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

Outline



LT3

- 1 Introduction: the concept of readability (prediction)
- 2 Experiments on large corpora
- 3 Discussion

Outline: introduction



LT3

- 1 Introduction: the concept of readability (prediction)
- 2 Experiments on large corpora
- 3 Discussion



LT3

What is readability?



LT3

What is readability?

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



LT3

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

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

What is readability prediction?

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

Introduction: readability prediction



LT3

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 → weights;
 - variables → text characteristics.
- e.g. Flesch Reading Ease [Flesch1948]:
$$207 - \textit{avgsentenceLen} - 85 * \textit{avgnumsyl}$$



LT3

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



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)

Outline: experiments



LT3

- 1 Introduction: the concept of readability (prediction)
- 2 Experiments on large corpora
 - Correlation matrices
 - Principal Component Analysis
 - Collinearity tests
- 3 Discussion



Large-scale calculation of readability scores and text characteristics

LT³

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

Calculated correlations between

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



LT³

	brouwer	kincaid	fog	smog	ari	douma	flesch	lix	cil	cilt	cilib	rgs	avgpolsyisent	avgwordlen	avgnumsyI	ppolsyIword	ratiolongword	psw	avgsentencelen	freq3000	tr	
brouwer																						
kincaid	-1,00																					
fog	-0,92	0,92																				
smog	-0,90	0,89	0,99																			
ari	-0,93	0,93	0,82	0,75																		
douma	0,91	-0,89	-0,88	-0,89	-0,77																	
flesch	0,91	-0,89	-0,88	-0,89	-0,76	1,00																
lix	-0,86	0,86	0,81	0,80	0,87	-0,82	-0,82															
cil	-0,66	0,64	0,64	0,66	0,66	-0,82	-0,82	0,75														
cilt	-0,47	0,44	0,45	0,49	0,48	-0,68	-0,68	0,58	0,93													
cilib	-0,35	0,34	0,38	0,38	0,36	-0,46	-0,46	0,47	0,59	0,61												
rgs	-0,31	0,30	0,33	0,33	0,32	-0,42	-0,42	0,44	0,55	0,52	0,55											
avgpolsyisent	-0,90	0,89	0,99	1,00	0,75	-0,89	-0,89	0,80	0,66	0,49	0,38	0,33										
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									
avgnumsyI	-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								
ppolsyIword	-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							
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						
psw	0,71	-0,74	-0,80	-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					
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				
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	-1,00			
tr	-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

	brouwer	kincald	fog	smog	ari	douma	flesch	lix	cli	cilt	clib	rgs	avgpolyisent	avgworden	avgnumsy1	ppolyword	ratilongword	paw	avgsentencelen	freq3000	trr	
brouwer																						
kincald	-1.00																					
fog	0.92	0.92																				
smog	-0.90	0.89	0.99																			
ari	-0.93	0.93	0.82	0.79																		
douma	0.91	-0.89	-0.88	-0.89	-0.77																	
flesch	0.91	-0.89	-0.88	-0.89	-0.76	1.00																
lix	-0.86	0.86	0.81	0.80	0.87	-0.82	-0.82															
cli	-0.68	0.64	0.64	0.66	0.66	-0.82	-0.82	0.79														
cilt	-0.47	0.44	0.45	0.49	0.48	-0.59	-0.58	0.58	0.93													
clib	-0.35	0.34	0.38	0.38	0.36	-0.46	-0.46	0.47	0.59	0.61												
rgs	-0.31	0.30	0.33	0.33	0.32	-0.42	-0.42	0.44	0.55	0.52	0.55											
avgpolyisent	-0.90	0.89	0.99	1.00	0.79	-0.89	-0.89	0.80	0.86	0.49	0.38	0.33										
avgworden	-0.54	0.51	0.54	0.57	0.53	-0.76	-0.76	0.68	0.98	0.96	0.60	0.55	0.57									
avgnumsy1	-0.68	0.65	0.70	0.74	0.50	-0.92	-0.92	0.68	0.85	0.76	0.49	0.45	0.74	0.85								
ppolyword	-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							
ratilongword	-0.52	0.49	0.55	0.58	0.47	-0.72	-0.72	0.71	0.85	0.79	0.57	0.53	0.58	0.86	0.80	0.72						
paw	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					
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				
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	0.18	-0.16			
trr	-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

	brouwer	kincaid	fog	smog	ari	douma	flesch	lix	cli	cilt	clib	rgs	avgpolylsent	avgwordlen	avgnumsyll	ppolyshword	ratilongword	paw	avgsentencelen	freq3000	trr	
brouwer																						
kincaid	-1.00																					
fog	-0.92	0.92																				
smog	-0.90	0.89	0.99																			
ari	-0.93	0.93	0.82	0.79																		
douma	0.91	-0.89	-0.88	-0.89	-0.77																	
flesch	0.91	-0.89	-0.88	-0.89	-0.76	1.00																
lix	-0.86	0.86	0.81	0.80	0.87	-0.82	-0.82															
cli	-0.68	0.64	0.64	0.66	0.66	-0.82	-0.82	0.78														
cilt	-0.47	0.44	0.45	0.49	0.48	-0.58	-0.58	0.58	0.93													
clib	-0.35	0.34	0.38	0.38	0.36	-0.46	-0.46	0.47	0.59	0.61												
rgs	-0.31	0.30	0.33	0.33	0.32	-0.42	-0.42	0.44	0.55	0.52	0.55											
avgpolylsent	-0.90	0.89	0.99	1.00	0.79	-0.89	-0.89	0.80	0.88	0.49	0.38	0.33										
avgwordlen	-0.54	0.51	0.54	0.57	0.53	-0.78	-0.76	0.68	0.98	0.96	0.60	0.55	0.57									
avgnumsyll	-0.68	0.65	0.70	0.74	0.50	-0.92	-0.92	0.68	0.85	0.78	0.49	0.45	0.74	0.85								
ppolyshword	-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							
ratilongword	-0.52	0.49	0.55	0.58	0.47	-0.72	-0.72	0.71	0.85	0.79	0.57	0.53	0.58	0.86	0.80	0.72						
paw	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					
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				
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	0.18	-0.16			
trr	-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



LT³

	brouwer	kincaid	fog	smog	ari	douma	flesch	lix	cli	cilt	clib	rgs	avgpolyisent	avgwordlen	avgnumsyll	ppolyisword	ratilongword	paw	avgsentencelen	freq3000	trr
brouwer	1.00	-0.92	-0.90	-0.93	0.91	0.91	-0.88	-0.88	-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	0.84	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	0.84	0.45	0.38	0.33	0.99	0.54	0.70	0.78	0.55	-0.60	0.60	0.22	0.10	
smog	-0.90	0.89	0.99	1.00	-0.89	-0.89	0.80	0.86	0.49	0.38	0.32	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	1.00	-0.77	-0.76	0.87	0.66	0.48	0.38	0.32	0.77	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	0.86	0.66	0.86	0.85	0.77	-0.57	0.57	0.33	0.13	
cli	-0.68	0.64	0.64	0.66	0.66	-0.82	-0.82	0.78	0.93	0.59	0.56	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	-0.46	-0.46	0.58	0.93	0.61	0.52	0.49	0.98	0.78	0.65	0.78	-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	0.21	
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	
avgpolyisent	-0.90	0.89	0.99	1.00	0.79	-0.89	-0.89	0.80	0.80	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	0.68	0.98	0.96	0.60	0.55	0.57	0.85	0.85	0.74	0.86	0.02	-0.02	0.55	0.27
avgnumsyll	-0.68	0.65	0.70	0.74	0.50	-0.92	-0.92	0.68	0.85	0.76	0.49	0.45	0.74	0.85	0.88	0.80	-0.04	0.04	0.44	0.22	
ppolyisword	-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	
ratilongword	-0.52	0.49	0.55	0.58	0.47	-0.72	-0.72	0.71	0.85	0.79	0.57	0.53	0.58	0.86	0.80	0.72	0.01	-0.01	0.53	0.32	
paw	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	0.18	-0.18	0.42	
trr	-0.02	0.00	0.10	0.09	-0.01	-0.14	-0.14	0.13	0.22	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

The goal of PCA

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



The goal of PCA

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

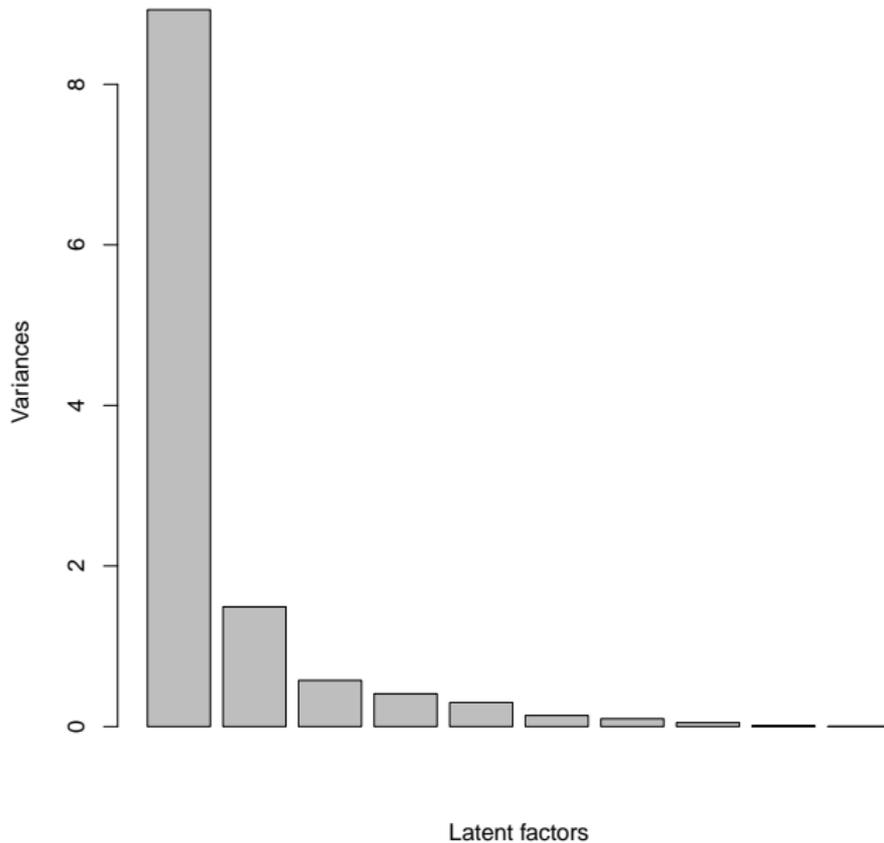
Performed PCA

- on all readability scores
- on all text characteristics



LT3

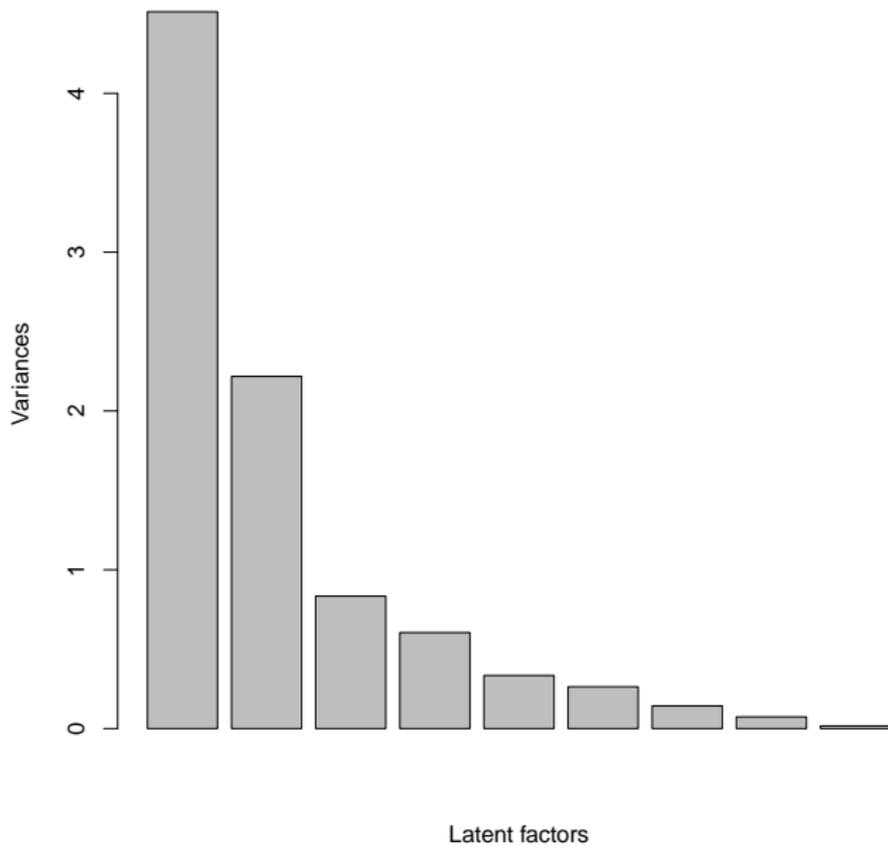
wsj - Readability formulas





LT3

wsj - Text characteristics





LT3

Determining the interdependence of variables in a formula

- Readability formulas $<$ multiple regression
- Collinearity: variables are correlated
 - found in all formulas
 - extrapolating to other data can be problematic

Outline: discussion



LT3

- 1 Introduction: the concept of readability (prediction)
- 2 Experiments on large corpora
- 3 Discussion

Towards an improved feature selection



LT3

Features that are used

- Strongly overlap
- Language-independent
- Strictly superficial

Towards an improved feature selection



LT3

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

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



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



<p>Een domeinnaam verhuizen - stap voor stap</p> <p>Kies een nieuwe agent</p> <p>U benadert een nieuwe agent uit onze lijst van geregistreerde agenten. U zal opnieuw de nodige gegevens aan deze agent moeten bezorgen, zoals bij een nieuwe registratie.</p> <p>Opzetten nameserver</p> <p>De nieuwe agent zal een nameserver op met uw gegevens. Dit kan best gebeuren vóór stap 3 om zoveel mogelijk stringen te voorkomen.</p> <p>Indienen transferaanvraag</p> <p>De agent dient een elektronische transferaanvraag in bij DNS. Controleer zeker of deze gegevens correct zijn! Belangrijk is dat het uw gegevens zijn en niet die van de agent of een tussenpersoon.</p> <p>Uitsturen e-mails</p> <p>Zodra de transferaanvraag bij DNS toekomt, stuurt het automatisch registratiesysteem 2 e-mails uit:</p> <ul style="list-style-type: none"> - één naar de nieuwe agent om ontvangst van zijn aanvraag te bevestigen - één naar de domeinnaamhouder met de vraag of u akkoord bent met de transfer! deze bevat een link naar een beveiligde webpagina <p>Reageren op de e-mail</p>	<p>Link: veel moeilijker Rechts: veel makkelijker</p> <p>Link: info moeilijker Rechts: info makkelijker</p> <p>Deze even moeilijk</p> <p>Link: info moeilijker Rechts: info moeilijker</p> <p>Link: veel makkelijker Rechts: veel moeilijker</p>	<p>Vivart is een Belgische politieke partij en een organisatie (" " bewegingspartij " +), in 1987 opgericht door Robert Duchateau, bewijshouder van het auto-elektronica bedrijf Melosix. Vivart staat zowel in Vlaanderen, Wallonië als de Oostkantons mee aan verkiezingen. Vivart werd voortgezet door de partij BANANA die alleen in 1995 aan verkiezingen deelnam.</p> <p>Het geschiedboek van Vivart, vooral het economische deel, is gebaseerd op het boek « Niv België - verslag aan de aandeelhouders », dat Duchateau schreef in 1994. Vivart staat voor individuele vrijheid en een sterke sociale zekerheid in een context van vrije markteconomie. Bij de parlementsverkiezingen van 1999 haalde Vivart 130.701 (2,1%) stemmen voor de Kamer. Bij de federale verkiezingen van 18 mei 2003 behaalde de partij, na herhaaldelijke interne strubbelingen, nog slechts 1,3% van de stemmen.</p> <p>Voor de verkiezingen van juni 2004 werd, mede vanwege de inwerking van de kieswet van 2001, een kans gevormd met de VLD. In maart 2006 wordt Heke Lijnen senator voor de partij na opletstaple door de VLD.</p>
--	---	--

<p>Minister van Sociale Zaken Frank Vandenbroucke (SP.A) kondigt drastische besparingsmaatregelen aan in de kinderopvoeder. De maatregelen moeten jaarlijks bijna 45 miljard euro opleveren.</p> <p>Kinvasiden krijgen niet alleen een strengere controle op wat ze aanvragen aan de ziekteverzekering, er komt ook meer controle op de kwaliteit en de duur van een behandeling.</p> <p>Zo betaalt de patiënt meer uit eigen zak als twee reizen van negen beurten niet volstaan om een flicke of murige aandoening te genezen. Mensen met zware aandoeningen kunnen onbepaald een beroep blijven doen op de diensten van de kinésist. Vandenbroucke wil een aantal kinésisten er ook toe bewegen met personen te gaan of naar andere beroepen over te stappen, het liefst binnen de verzorgingssector.</p> <p>Verzorgingsverleners van verschillende kinésistenverenigingen reageren enthousiast op de besparingsplannen. Zij vrezen dat het beroep van kinésist op termijn niet meer leefbaar zal zijn. De verenigingen zijn wel te spreken over de plannen voor de inktering van het aantal kinésisten.</p>	<p>Link: veel moeilijker Rechts: veel moeilijker</p> <p>Link: info moeilijker Rechts: info moeilijker</p> <p>Link: veel makkelijker Rechts: veel moeilijker</p>	<p>Minister van Sociale Zaken Frank Vandenbroucke (SP.A) kondigt drastische besparingsmaatregelen aan in de kinderopvoeder. De maatregelen moeten jaarlijks bijna 45 miljard euro opleveren.</p> <p>Kinvasiden krijgen niet alleen een strengere controle op wat ze aanvragen aan de ziekteverzekering, er komt ook meer controle op de kwaliteit en de duur van een behandeling.</p> <p>Zo betaalt de patiënt meer uit eigen zak als twee reizen van negen beurten niet volstaan om een flicke of murige aandoening te genezen. Mensen met zware aandoeningen kunnen onbepaald een beroep blijven doen op de diensten van de kinésist. Vandenbroucke wil een aantal kinésisten er ook toe bewegen met personen te gaan of naar andere beroepen over te stappen, het liefst binnen de verzorgingssector.</p> <p>Verzorgingsverleners van verschillende kinésistenverenigingen reageren enthousiast op de besparingsplannen. Zij vrezen dat het beroep van kinésist op termijn niet meer leefbaar zal zijn. De verenigingen zijn wel te spreken over de plannen voor de inktering van het aantal kinésisten.</p>
---	---	---

Difficult

Text E

Readability score: **82**

Why did you assign score 82 to Text E?

Average

Easy

+ Get new text

Unmark all

✓ Submit scores

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.



LT3