FACULTY OF ARTS AND PHILOSOPHY

LT³, LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

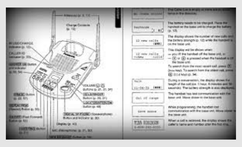
Orphée De Clercq and Véronique Hoste

LT³'S READABILITY PREDICTION SYSTEM

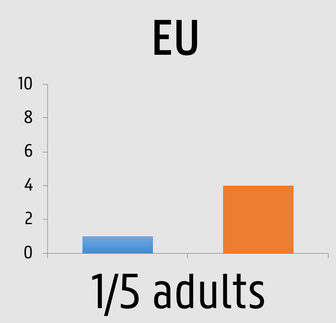
EVERYONE CAN READ

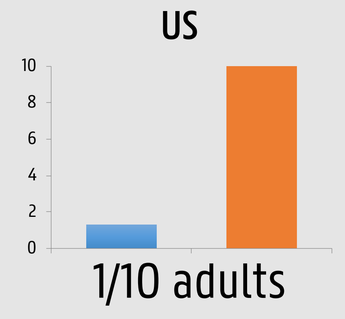
Literacy level of the general adult public

"People understand most texts"



REALITY CHECK





How to write clearly

Plain Language.gov
Improving Communication from the Federal Government to the Public

HEERLIJK HELDER
HAUTEKIET SCHENKT KLARE TAAL

COMPUTERS TO THE RESCUE: FROM FORMULAS TO MACHINE LEARNING

Classical readability formulas

Readability = weight 1 * characteristic a + weight 2 * characteristic b + …

TEXT CHARACTERISTICS	
avgnumsyll	Average word length in number of syllables
avgsentencelen	Average sentence length in number of words
avgwordlen	Average word length in number of characters
freq	Percentage of words also found in a word list (.e.g the Dale Chall (1948) 3000 word list with words familiar to 80% of American fourth-graders)
psw	Percentage of sentences per word
avgpolysyntsent	Average number of words of 3 or more syllables per word
ppolysyntword	Percentage of words of 3 or ore syllables
ratiolongword	Ratio of words of more than 6 characters
ttr	Type token ratio (the number of unique words divided by the total number of words

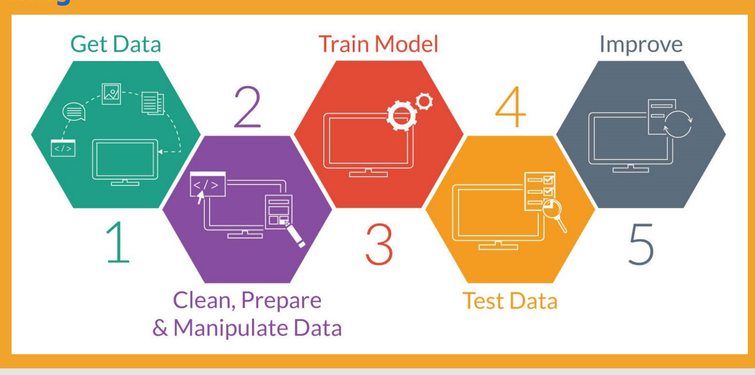
Flesch Reading Ease (EN)

Flesch Douma (DU)

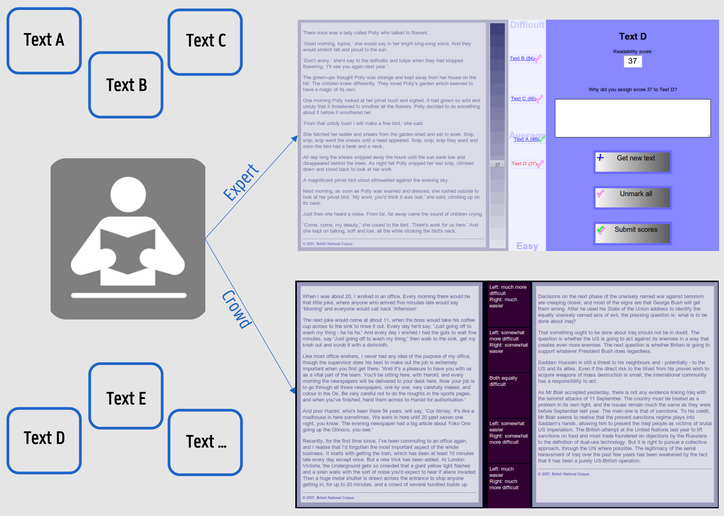
- limited
- superficial
- children
- …

Advances Computer Science & Natural Language Processing

Machine learning

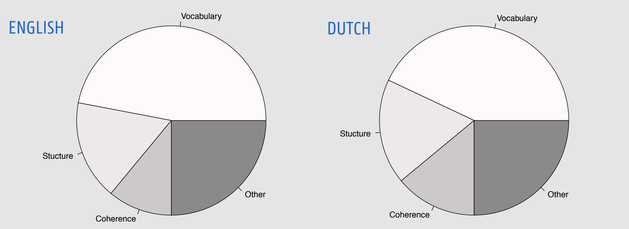


STEP 1: Collect and assess data



STEP 2: Extract meaningful features

Text ID	Score	Comment
dpc-med-000674-nl-sen_3	76	Text demands a lot of concentration; too long sentences
dpc-vla-001161-nl-sen_4	62	Very abstract, both with respect to topic and vocabulary use ("government instruments", "finance products", "investment products")
WR-P-P-F-0000000007_1	66	Pffff: starting with a subordinate sentence, sentences between brackets, passive constructions, constructions such as "in that case, it might be wise to", too much subordinate clauses. Very difficult manual.
dpc-riz-001056-nl-sen_1	67	I cannot really determine whether the text is difficult. It's mainly badly written. The references are unclear. What does the 3rd paragraph refer to? 4th paragraph: "proposal". Which proposal? Last sentence: "the means". Which means?



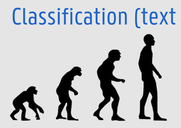
STEP 3-4-5: Train-test and improve

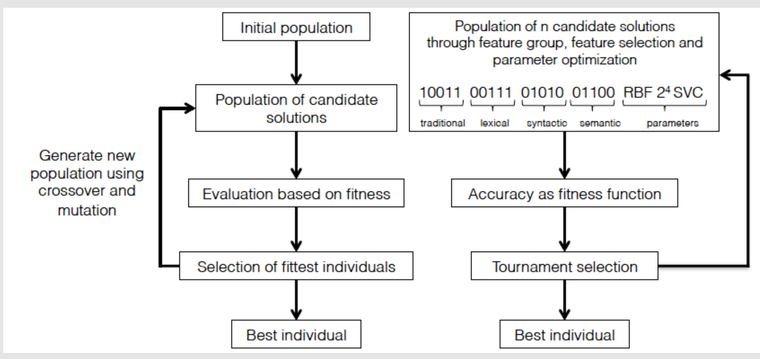
Regression (Text a has readability of X)

Classification (text a is easier than text b)

Optimal machine learning model

➔ Genetic Algorithms approach





STATE-OF-THE-ART READABILITY PREDICTION

		Regression		Classification BINARY		Classification MULTI	
		EN	DU	EN	DU	EN	DU
Baseline	Default, all features	0.1489	0.1813	85.31	92.83	57.35	59.49
Round 1	Feature groups	0.1242	0.1492	85.60	93.16	57.38	60.87
	Individual features	0.0985	0.1470	86.28	93.61	58.14	61.31
Round 2	Joint feature groups	0.0060	**0.0003**	96.27	98.01	70.35	73.35
	Joint individual features	**0.0059**	0.0004	**96.88**	**98.24**	**71.00**	**73.62**

- Generic readability prediction (De Clercq and Hoste, 2016)
- Translation quality and post editing (De Sutter et al. 2017, Daems et al. 2017)
- Retrained on specific genre: sustainability reporting (Smeuninx, 2018)

Assessing Readability demo

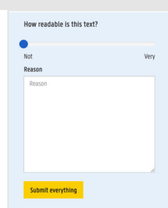
Read all the texts and give each a readability score from 0 to 100.

Tekst 1

Tekst 2

Tekst 3

Het STD-systeem meet allerhande fysico-chemische parameters (het zoutgehalte, de temperatuur, de concentratie aan opgeloste zuurstof, de troebelheid, de hoeveelheid chlorofyl, enz…) in de waterkolom tot 1600 meter diepte. Het STD-syteem wordt in de waterkolom neergelaten met een stalen kabel waarin zich een coaxiale kern bevindt. Terwijl het systeem neerdaalt aan een snelheid van ongeveer 0.8 meter per seconde, worden via de coaxiale kabel voortdurend alle parameters gemeten en naar de verwerkingscomputer aan boord gestuurd waar ze door de wetenschappers geëvalueerd worden.



DEMO 1

www.lt3.ugent.be/tools/assessing-readability/

Machine Learning readability demo

Upload a UTF-8-text file (max. 1500 characters) to assess its readability.

Base features, such as the average sentence length, type token ratio, ratio of content and function words are always calculated. You have the option to derive three other feature groups.

Processing will take about 30 seconds, but depending on the length of your text and/or the amount of traffic on our server this can be longer since the processing tasks are queued.

File

Choose file No file chosen

If you don't have a text-file ready, try one of these: Text 1, Text 2, Text 3, Text 4, Text 5, Text 6, Text 7, Text 8, Text 9 or Text 10.

Lexical features

Calculated based on a large reference corpus

Deep syntax features

Based on dependency parsing

Deep semantics features

Based on named entity recognition and coreference resolution

Submit

DEMO 2

https://www.lt3.ugent.be/tools/machine-learning-readability/

FUTURE

- Translatability prediction
- Automated writing evaluation
- Collaborate: check out our demos or **contact me**

References:

Daems, J., De Clercq, O., & Macken, L. (2017). Translationese and post-editese : how comparable is comparable quality? *LINGUISTICA ANTVERPIENSIA NEW SERIES-THEMES IN TRANSLATION STUDIES*, *16*, 89–103.

De Clercq, Orphée, & Hoste, V. (2016). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. COMPUTATIONAL LINGUISTICS, 42(3), 457–490.

De Sutter, G., Cappelle, B., De Clercq, O., Loock, R., & Plevoets, K. (2017). Towards a corpus-based, statistical approach of translation quality : measuring and visualizing linguistic deviance in student translations. (G. S. Koby & I. Lacruz , Eds.)*LINGUISTICA ANTVERPIENSIA NEW SERIES-THEMES IN TRANSLATION STUDIES*, *16*, 25–39.

Smeuninx, N. (2018). Dear stakeholder : exploring the language of sustainability reporting : a closer look at readability, sentiment and perception. Ghent University. Faculty of Arts and Philosophy, Ghent, Belgium.

Contact

orphee.declercq@ugent.be
www.lt3.ugent.be/people/orphee-de-clercq/

LT³ Research Group

UGentNLP

https://www.linkedin.com/in/orphee/

GHENT UNIVERSITY