# Estimating literary readability through lexical & syntactic complexity.

Kim Jautze, Corina Koolen, Andreas van Cranenburgh

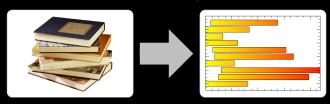
Huygens ING Royal Netherlands Academy of Arts and Sciences Institute for Logic, Language and Computation University of Amsterdam

Nov 7, 2013



# The Riddle of Literary Quality

► The Riddle of Literary Quality aims to find properties which make certain texts literary.



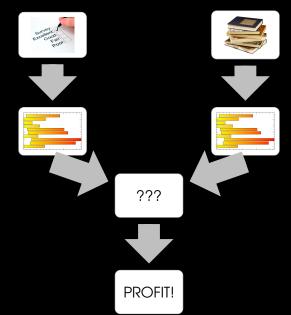
Literary complexity, the idea that literary texts are more complex than others, is one such property.

# The project





# The project



# Readability

Simple readability measures use superficial features:

- words per sentence
- syllables per word
- $\Rightarrow$  two-parameter multiple regression equations

# Readability

#### Vocabulary-based tests

- fixed word lists & frequencies (Lexile)
- smoothed unigram model (C & C, 2004)

# Measuring success

- Ashok et al. (2013) report an inverse correlation between simple readability measures and literary success.
- Basically, more VPs => more readability, but less success.

## Kincaid measure

L Enquist, Anna - De Verdovers, dact: 5.7
L Dorrettein, Renate - Het hemelse gerecht, dact: 5.6
L Glastra van Loon, Karel - De Rassiervrucht, dact: 5.4
L Grunberg, Amon - Huid en haar, dact: 5.1
C, Verkerk, Anna - Bedrogen liefde, dact: 4.9
L, Glastra van Loon, Karel - Lisas & Ammadact: 4.9
C, Middelbeek, Mariette - Single En Sexy, dact: 4.5
C, Verwoert, Rianne - Trouwien), dact: 4.2
C, Verkerk, Anna - Als een zandstorre lin de wind, dact: 4.2
C, Verwoert, Rianne - Match, dact: 4.1
C, Gastel, Chantal van - Zoek het maar uit, dact: 4.1
C, Gastel, Chantal van - Zoek het maar uit, dact: 4.1
C (Verwoert, Rianne - Schikken of stilken dact: 3.8

#### Kincaid: L. Moor, Margriet de - De verdronkene, dact: 8.8 L Japin, Arthur - Vaslav, dact: 8.8 L. Moor, Margriet de - De Schilder en het Meisie,dact: 8.5 L Beijnum, Kees Van - De Ordening.dact: 7.8 L. Japin, Arthur - De grote wereld,dact: 7 C Verkerk, Anita - Cheesecake & Kilts, dact: 6.6 L Enquist, Anna - Het meesterstuk.dact: 6.4 L Beijnum, Kees van - De oesters van Nam Kee.dact: 6.4 C Hollander, Wilma - Bouzouki Boogie, dact: 6.4 L Enquist, Anna - De Thuiskomst.dact: 6.3 L Dorrestein, Renate - Een sterke man.dact: 6.1 C Hollander, Wilma - Onder de Griekse zon.dact: 6.1 L. Grunberg, Arnon - De Asielzoeker, dact: 6.1 C Hollander, Wilma - Dans der liefde.dact: 6.1 C Harrewiin, Astrid - In zeven sloten, dact: 6.1 L. Dorrestein, Renate - Hart van steen,dact: 6.1 C Harrewijn, Astrid - Luchtkussen.dact: 5.9 C Middelbeek, Mariette - Revanche in New York.dact: 5.8 C Middelbeek, Mariette - Status O.K.,dact: 5.7

### Coleman-Lau measure

#### Coleman-Liau:

L. Moor, Margriet de - De verdronkene dact: 9.8

L\_Moor, Margriet de - De Schilder en het Meisje.dact: 9.8

L\_Beijnum, Kees Van - De Ordening.dact: 9.2

L Japin, Arthur - Vaslav.dact: 8.9

L\_Japin, Arthur - De grote wereld.dact: 8.6

L\_Enquist, Anna - Het meesterstuk.dact: 8.4 L Enquist, Anna - De Thuiskomst.dact: 8.1

C Hollander, Wilma - Bouzouki Boogie.dact: 8

L Grunberg, Arnon - De Asielzoeker.dact: 8

L\_Dorrestein, Renate - Een sterke man.dact: 7.8

C\_Verkerk, Anita - Cheesecake & Kilts.dact: 7.8

C\_Hollander, Wilma - Onder de Griekse zon.dact: 7.7

L\_Enquist, Anna - De Verdovers.dact: 7.7
L Dorrestein, Renate - Het hemelse gerecht.dact: 7.7

L Dorrestein, Renate - Hart van steen dact: 7.7

L Beijnum, Kees van - De oesters van Nam Kee.dact: 7.6

C\_Middelbeek, Mariette - Revanche in New York.dact: 7.6

C\_Hollander, Wilma - Dans der liefde.dact: 7.5
C. Harrewiin. Astrid - In zeven sloten.dact: 7.5

C\_marrewijn, Astrid - in zeven sloten.dact: 7.:
C Harrewijn, Astrid - Luchtkussen.dact: 7.4

C\_Middelbeek, Mariette - Status O.K..dact: 7.1

L\_Grunberg, Arnon - Huid en haar.dact: 7.1

L\_Glastra van Loon, Karel - De Passievrucht.dact: 6.9

C\_Verkerk, Anita - Bedrogen liefde.dact: 6.8

L\_Glastra van Loon, Karel - Lisa's Adem.dact: 6.7

C\_Middelbeek, Mariette - Single En Sexy.dact: 6

C\_Verkerk, Anita - Als een zandkorrel in de wind.dact: 5.8

C Verwoert, Rianne - Trouw(en).dact: 5.6

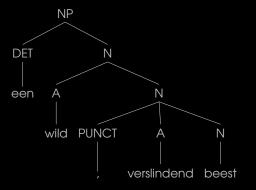
C\_Gastel, Chantal van - Zwaar verliefd.dact: 5.5

C\_Verwoert, Rianne - Match.dact: 5.4
C Gastel, Chantal van - Zoek het maar uit.dact: 5.3

C Verwoert, Rianne - Schikken of stikken.dact: 5

# Syntactic patterns

- Parse sentences in books
- Count occurrences of 'complex' patterns with query engine



Jautze, Koolen, van Cranenburgh, de Jong (2013). From high heels to (...): a syntactic investigation of chick lit and literature. Proc. of CLFL.

### Relative clauses

Relative frequency of pattern: (count / num consts \* 100) L. Moor, Margriet de - De Schilder en het Meisie: 0.735405 % L. Moor, Margriet de - De verdronkene: 0.724075 % L Japin, Arthur - Vaslav: 0.682501 % L Japin, Arthur - De grote wereld: 0.628296 % L. Grunberg, Arnon - De Asielzoeker: 0.588439 % L\_Beijnum, Kees Van - De Ordening: 0.581659 % L Beijnum, Kees van - De oesters van Nam Kee: 0.500157 % L Glastra van Loon, Karel - De Passievrucht: 0.487681 % C Hollander, Wilma - Bouzouki Boogie: 0.480003 % L Glastra van Loon, Karel - Lisa's Adem: 0.459977 % C Hollander, Wilma - Dans der liefde: 0.453606 % L Enquist, Anna - Het meesterstuk: 0.432595 % C Middelbeek, Mariette - Status O.K.: 0.42134 % L Dorrestein, Renate - Hart van steen: 0.397248 % L Enquist, Anna - De Thuiskomst: 0.39225 % L. Dorrestein, Renate - Het hemelse gerecht: 0.386598 % C Harrewijn, Astrid - In zeven sloten: 0.382199 % L Grunberg, Arnon - Huid en haar: 0.374201 % C\_Middelbeek, Mariette - Single En Sexy: 0.37411 % C Middelbeek, Mariette - Revanche in New York: 0.365803 % C Gastel, Chantal van - Zoek het maar uit: 0.35919 % C Hollander, Wilma - Onder de Griekse zon: 0.355347 % L. Dorrestein, Renate - Een sterke man: 0.354728 % C Harrewijn, Astrid - Luchtkussen: 0.345042 % L Enquist, Anna - De Verdovers: 0.327959 % C Verkerk, Anita - Cheesecake & Kilts: 0.264732 % C Gastel, Chantal van - Zwaar verliefd: 0.247863 % C Verwoert, Rianne - Match: 0.179688 % C Verwoert, Rianne - Trouw(en): 0.176089 % C Verkerk, Anita - Bedrogen liefde: 0.172199 % C Verwoert, Rianne - Schikken of stikken: 0.15002 %

C Verkerk, Anita - Als een zandkorrel in de wind: 0.135044 %

#### NP modifiers

Relative frequency of pattern: (count / num consts \* 100) L. Moor, Margriet de - De Schilder en het Meisie: 6.06198 % L Beijnum, Kees Van - De Ordening: 5,67769 % L\_Moor, Margriet de - De verdronkene: 5.63865 % L Enquist, Anna - Het meesterstuk: 4.88044 % L Beijnum, Kees van - De oesters van Nam Kee: 4,52265 % L. Dorrestein, Renate - Hart van steen: 4.13842 % L Japin, Arthur - Vaslav: 4.11484 % L Enquist, Anna - De Thuiskomst: 4.06713 % C Verkerk, Anita - Cheesecake & Kilts: 4.03766 % L Glastra van Loon, Karel - De Passievrucht: 3.989 % C Harrewijn, Astrid - In zeven sloten: 3.97259 % C Harrewiin, Astrid - Luchtkussen: 3,75685 % L Dorrestein, Renate - Een sterke man: 3,70123 % C Hollander, Wilma - Bouzouki Boogie: 3.69868 % L. Japin, Arthur - De grote wereld: 3,60223 % L Glastra van Loon, Karel - Lisa's Adem: 3.59071 % L Dorrestein, Renate - Het hemelse gerecht: 3.45991 % C Hollander, Wilma - Dans der liefde: 3.44384 % L Enquist, Anna - De Verdovers: 3,38849 % C Middelbeek, Mariette - Status O.K.: 3.22416 % L Grunberg, Arnon - De Asielzoeker: 3.17927 % C Middelbeek, Mariette - Revanche in New York: 3.0933 % C Hollander, Wilma - Onder de Griekse zon: 3.08525 % C Verkerk, Anita - Bedrogen liefde: 2.95271 % C Middelbeek, Mariette - Single En Sexv: 2.80381 % L. Grunberg, Arnon - Huid en haar: 2,74888 % C Verwoert, Rianne - Match: 2.65433 % C Verkerk, Anita - Als een zandkorrel in de wind: 2.60418 % C Gastel, Chantal van - Zoek het maar uit: 2,53336 % C Verwoert, Rianne - Trouw(en): 2.52928 % C Verwoert, Rianne - Schikken of stikken: 2.51482 % C Gastel, Chantal van - Zwaar verliefd: 2.30437 %

# Outliers

	Kincaid	Coleman	Rel.	NP mod
Verkerk - Cheesecake	+	+		+
Hollander - Bouzouki	+	+	+	
Dorrestein - Hemelse Gerecht	-	+/-	+/-	+/-
Dorrestein - Sterke Man	+/-	+		+/-
Enquist - Verdovers		+/-		+/-
Glastra van Loon			+	+
Grunberg - Huid en Haar			+/-	

#### Other measures

## Syntactic & higher-level complexity

- Perplexity: n-gram model (Schwarm & Ostendorf, 2005)
- ► Coherence (Graesser et al., 2004)
  - Adjacent sentences which share a noun
  - ► Logical connectives between sentences

# Beyond syntax: narrative complexity

- ► Slow vs. fast pace
- Linear vs. non-linear timeline
- Multiple narrators

#### Conclusion

#### Either ...

- Some chick lit is more complex, some literature is simpler?
- ... or not all complexity is in the lexical & syntactic realm

# That's all, folks!