Estimating literary readability through lexical & syntactic complexity.

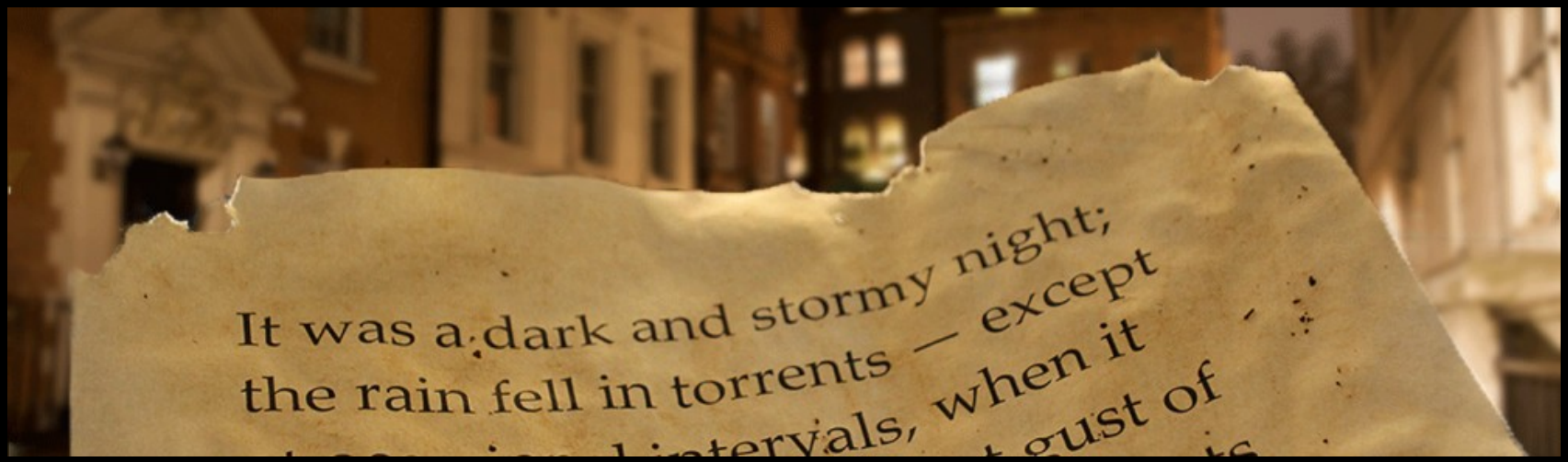
Kim Jautze, Corina Koolen,
Andreas van Cranenburgh

Huygens ING
Royal Netherlands Academy of Arts and Sciences

Institute for Logic, Language and Computation
University of Amsterdam

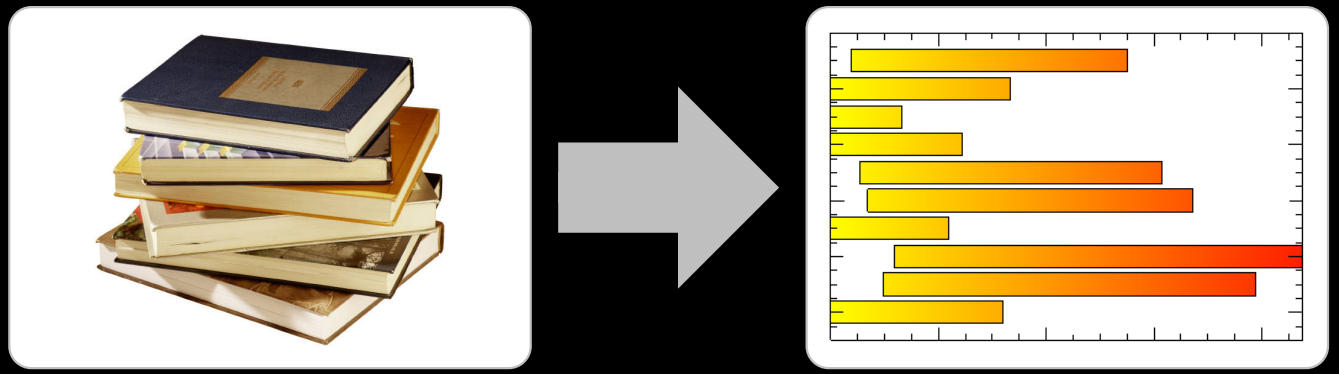
Nov 7, 2013

Meertens, Amsterdam, 2013



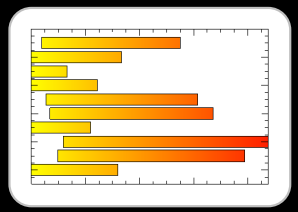
The Riddle of Literary Quality

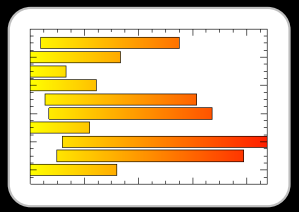
- The Riddle of Literary Quality aims to find properties which make certain texts literary.



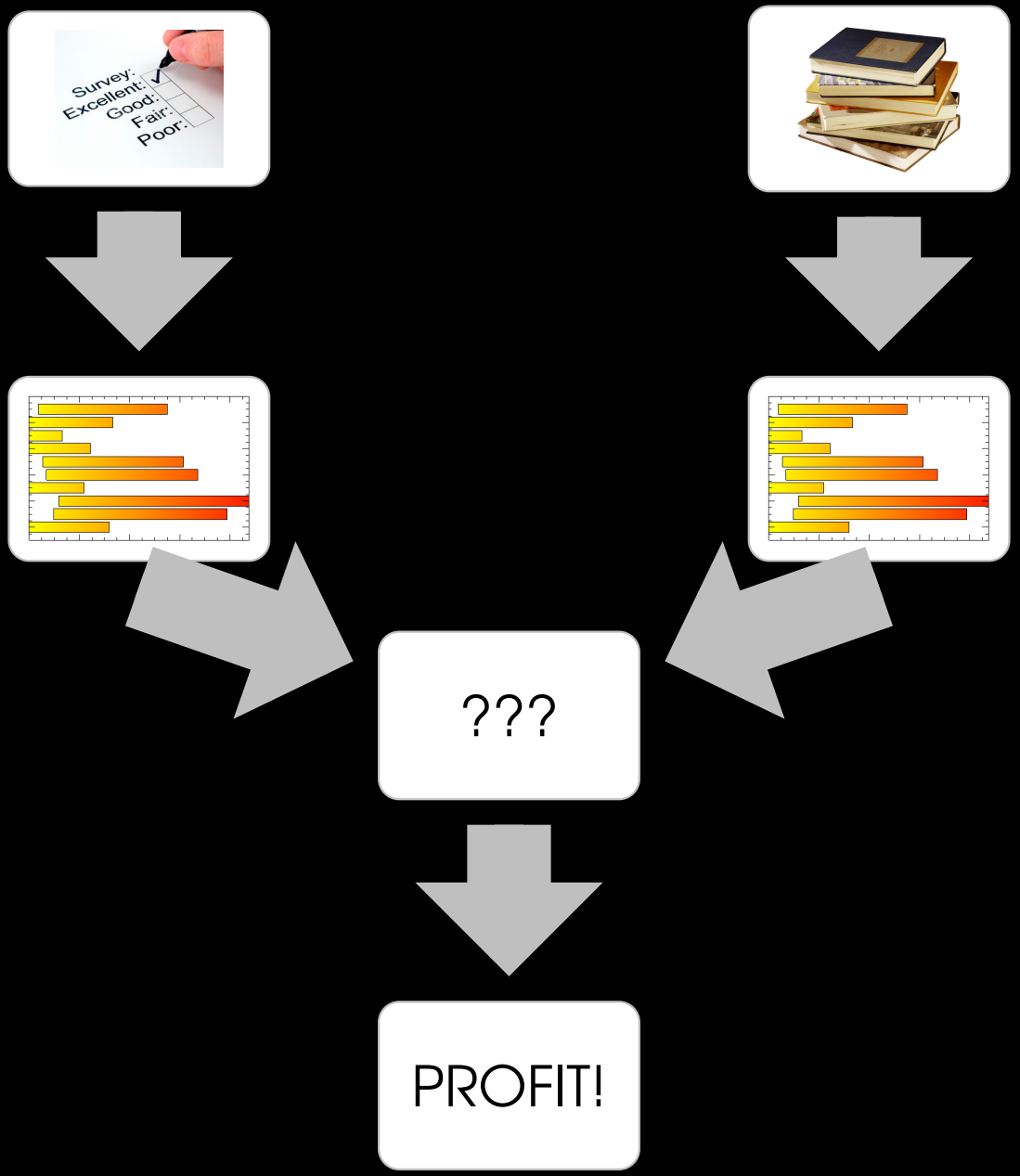
- Literary complexity, the idea that literary texts are more complex than others, is one such property.

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)

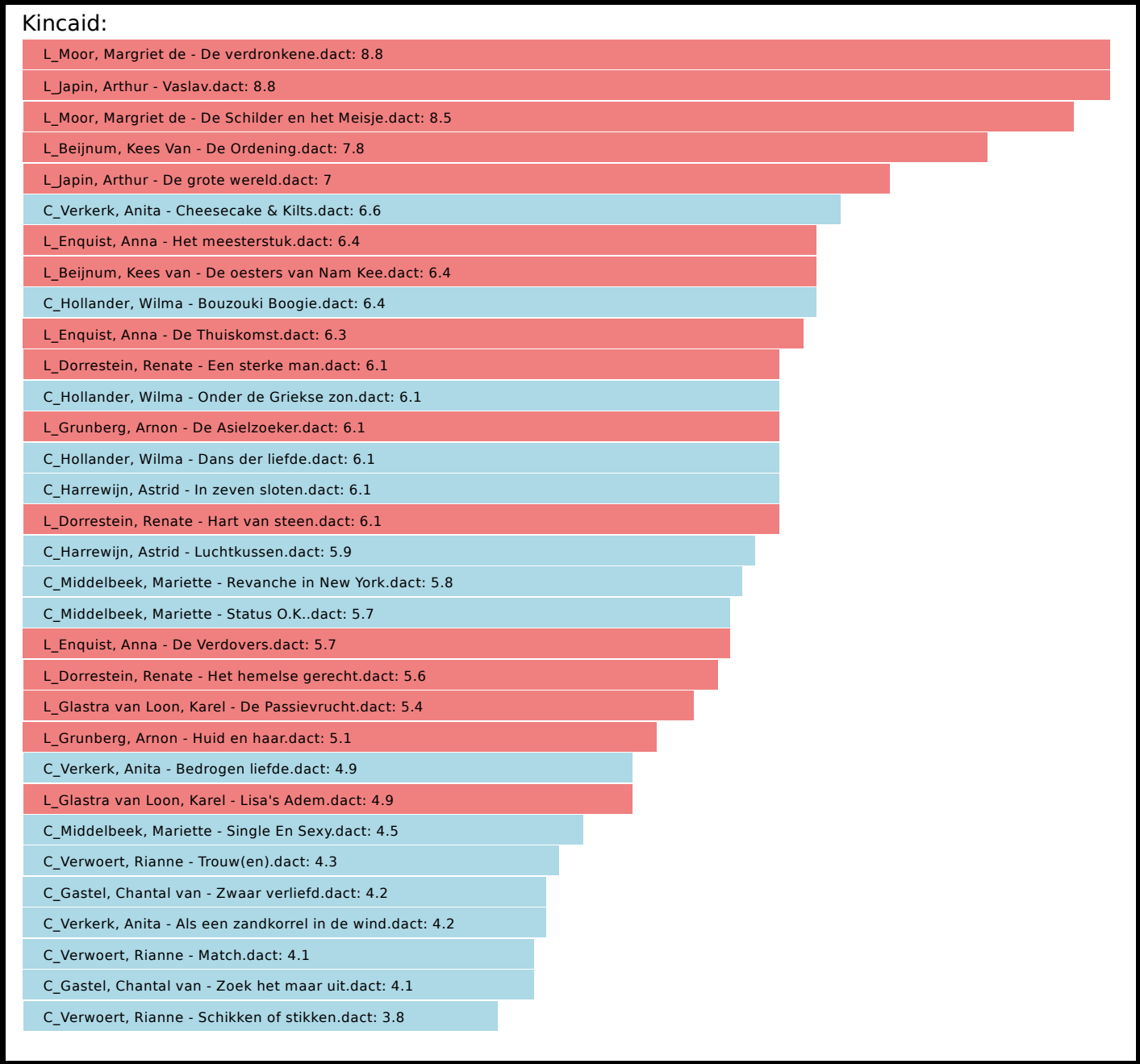
The Lexile framework for reading. http://www.lexile.com Collins-Thompson & Callan (2004). A language modeling approach to predicting reading difficulty. In Proc. of HLT/NAACL, pp. 193–200.

Measuring success

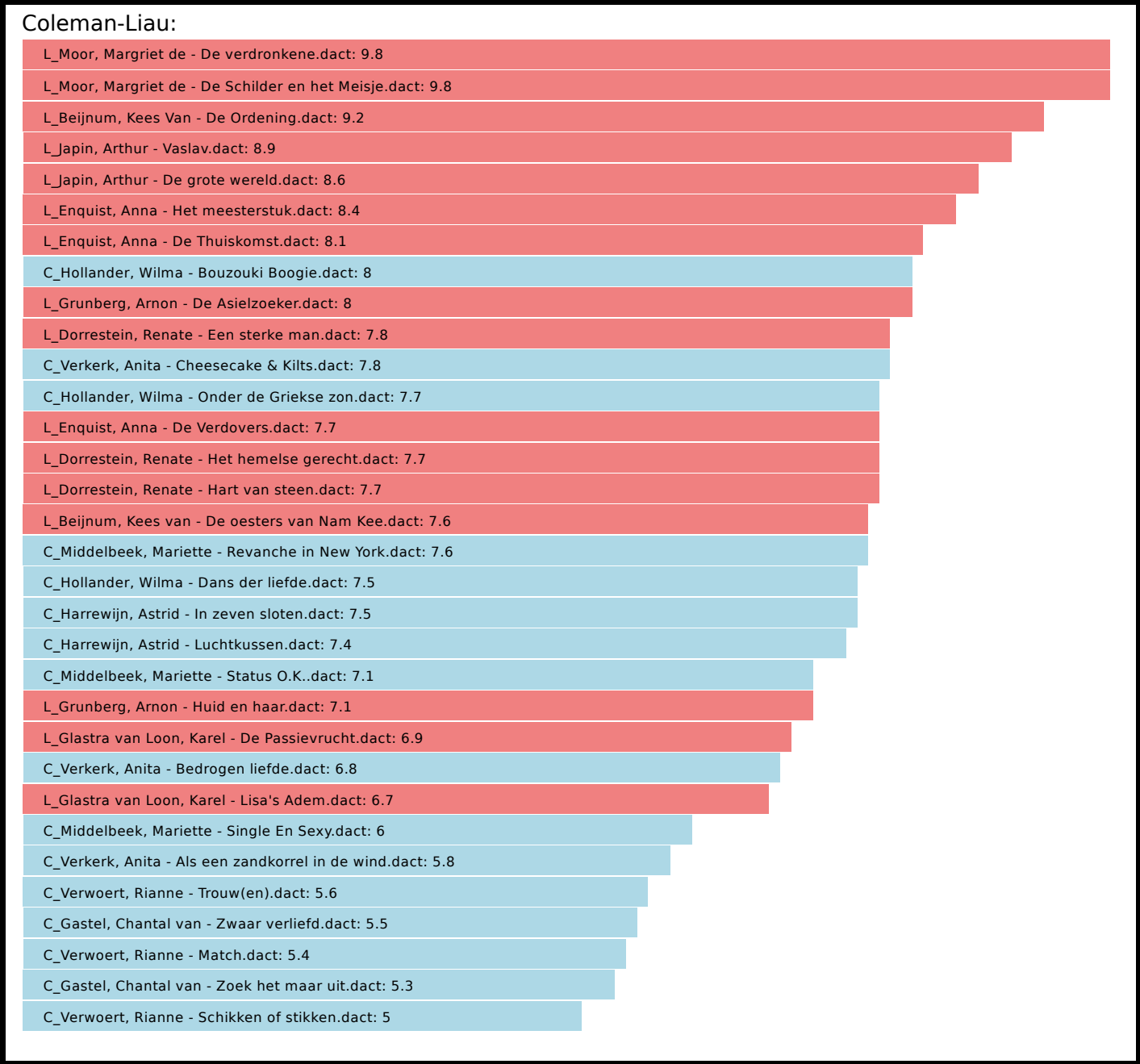
- Ashok et al. (2013) report an inverse correlation between simple readability measures and literary success.
- Basically, more VPs => more readability, but less success.

Ashok, Feng and Choi (2013). Success with Style—Using Writing Style to Predict the Success of Novels. EMNLP

Kincaid measure

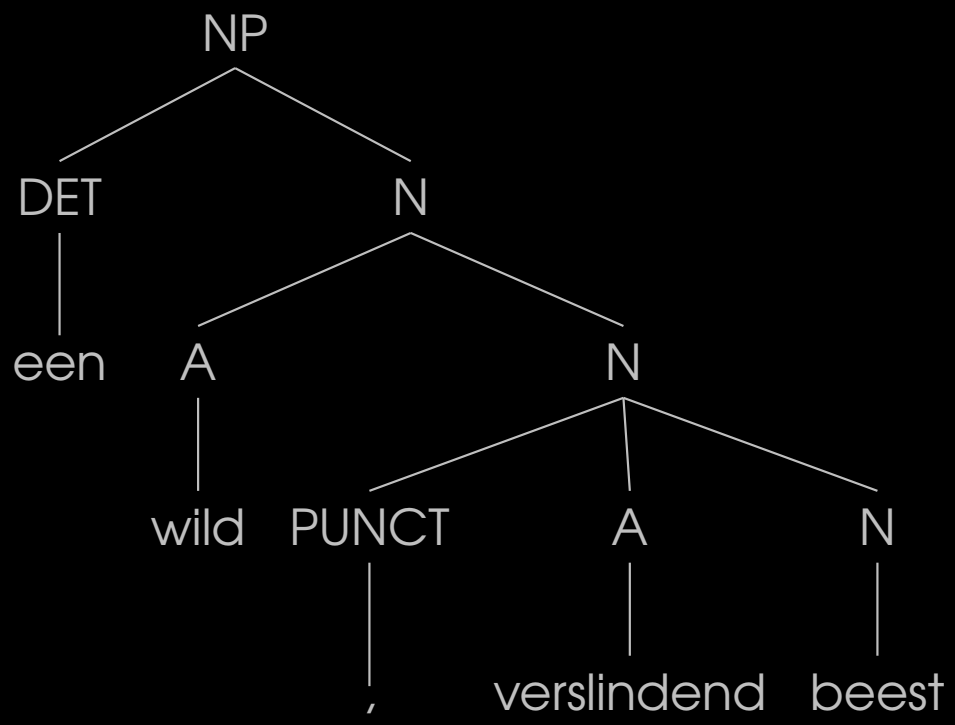


Coleman-Lau measure



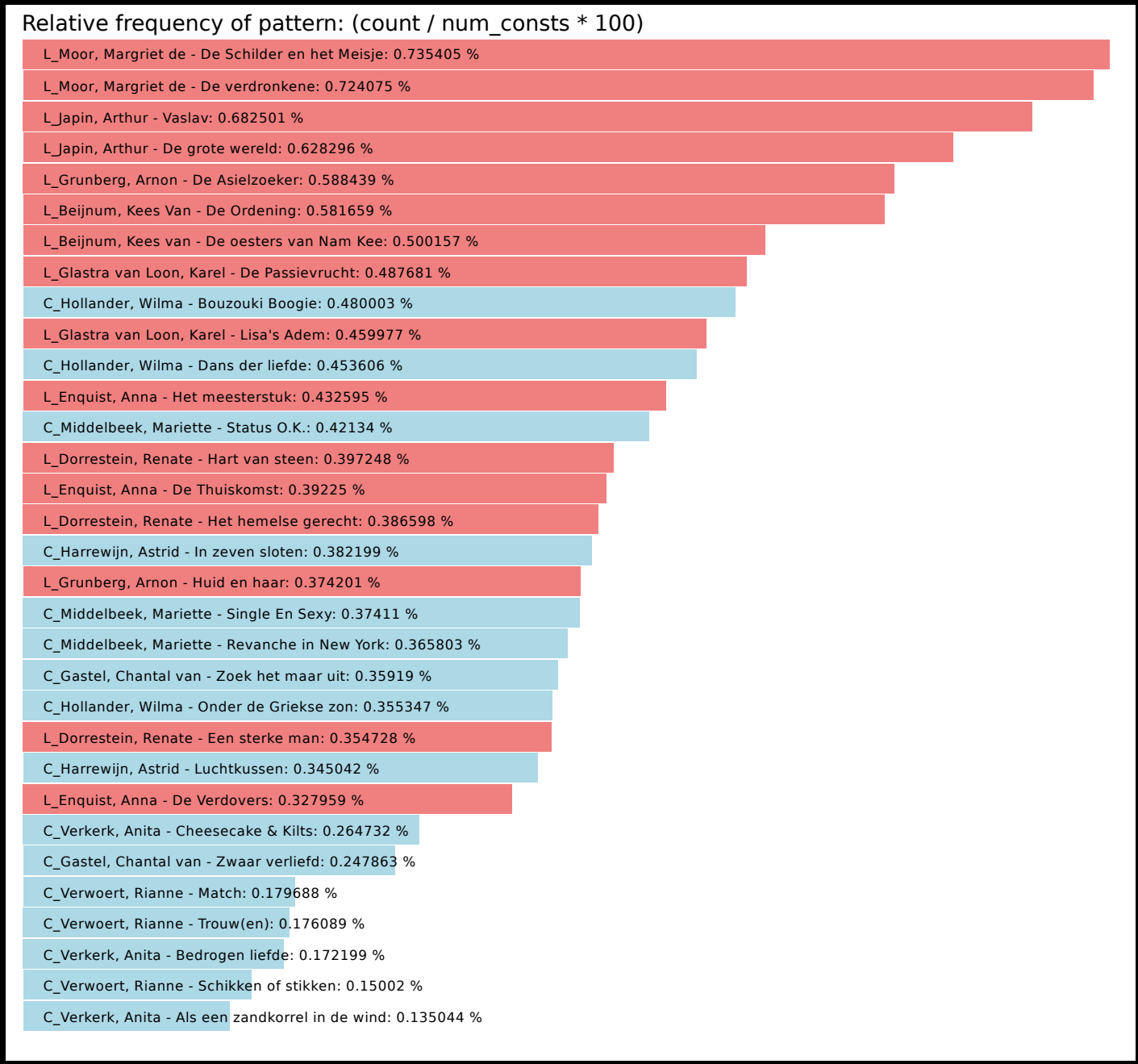
Syntactic patterns

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

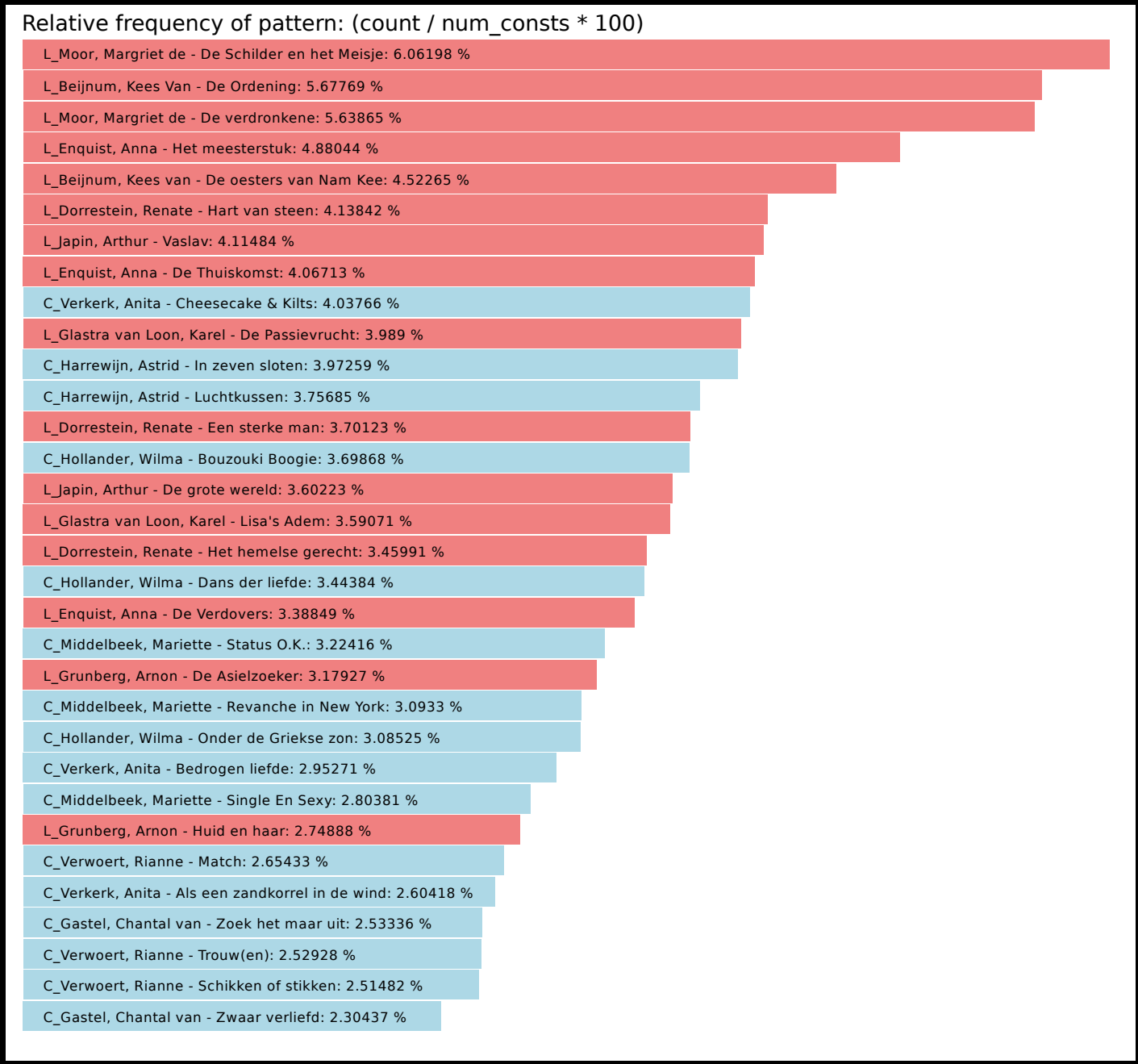


Jautze, Koolen, van Cranenburgh, de Jong (2013). From high heels to (...): a syntactic investigation of chick lit and literature. Proc. of CLFL.

Relative clauses



NP modifiers



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	-	-	+/-	-

+ = `complex'
- = `simple'

Other measures

Syntactic & higher-level complexity

- Perplexity: *n*-gram model (Schwarm & Ostendorf, 2005)
- Coherence (Graesser et al., 2004)
  - Co-reference chains make text more cohesive
  - Connectives between sentences

Schwarm & Ostendorf (2005). Reading level assessment using SVM and stat. lang. models. Proc. of ACL, pp. 523–530.

Graesser et al. (2004). Coh-Metrix (...) BRMIC (36), pp. 193–202.

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!