Impact and Fiction

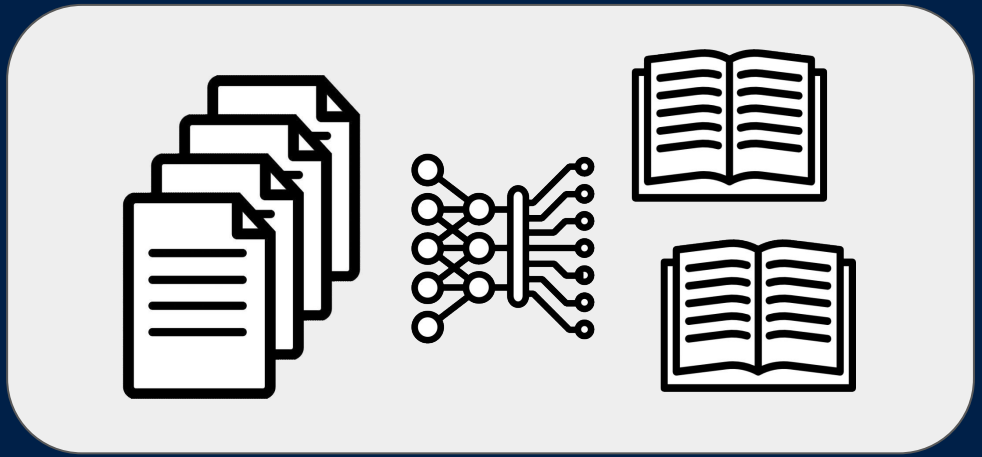
Katja Tereshko, Willem van Hage, Eva Viviani
Marijn Koolen, Peter Boot, Joris J. van Zundert

Huygens Institute, Amsterdam
Netherlands eScience Center, Amsterdam

CLARIAH Annual Event - 30-11-2023 - Utrecht

Impact and Fiction: Goal

Relate **features of impact** as reported in online reviews
to **textual features** of the novels



IMPACT & FICTION
Measuring the impact of fiction on readers

https://impactandfiction.huygens.knaw.nl/

Research Questions

- Three main questions
  - How does reading fiction affect or impact readers?
  - Which textual properties of books contribute to this impact?
  - How does this impact depend on the reader's reading preferences?

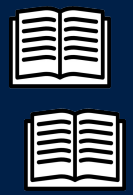
Research Questions

- Three main questions
  - How does reading fiction affect or impact readers?
  - Which textual properties of books contribute to this impact?
  - How does this impact depend on the reader's reading preferences?
- Approach: investigate textual aspects
  - Overall features: mood, topic
  - Stylistic: e.g. concreteness, past/present tense, complexity, unexpectedness
  - Narrative: pace, attractiveness of characters

Data

670,924 online reviews, ~ 150 million words

- 7 platforms, incl. hebban.nl, bol.com etc.
- Heuristic impact model (Boot & Koolen 2020)
- 2.4 million expressions of impact



18,885 full texts of novels, ~1.3 billion words

- 11 genres
- Dutch (original and translated)

Data Issues

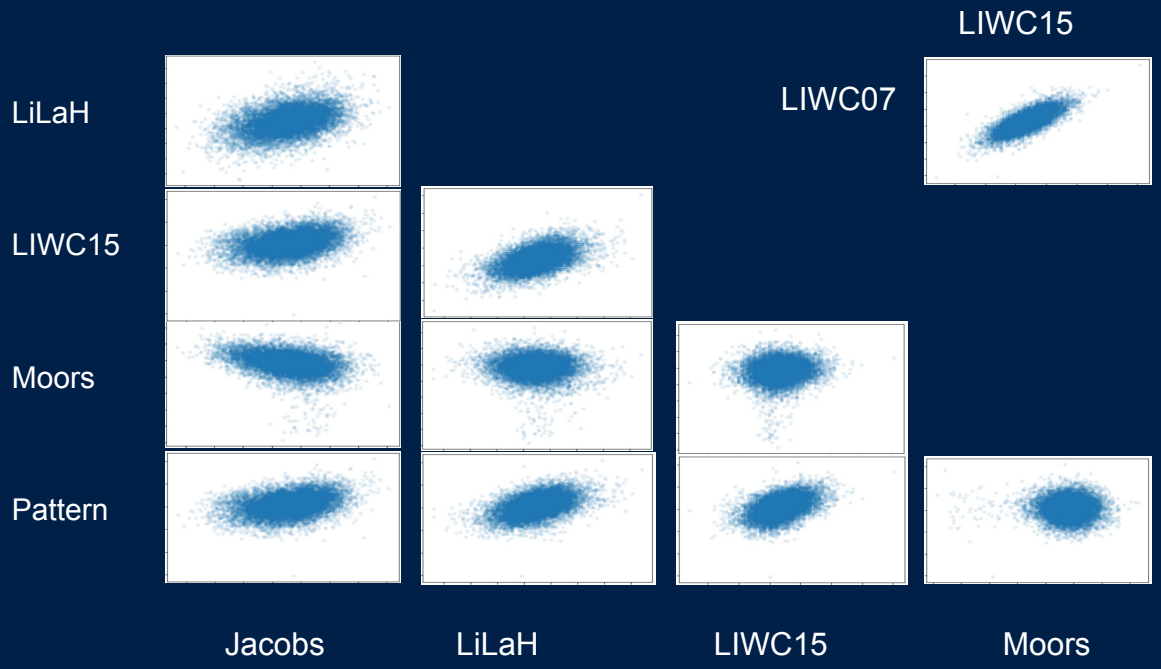
- Copyright issues
  - Books can only be processed on premises of Koninklijke Bibliotheek
  - In future: Secure ANanlysis Environment (SANE, SURF+CLARIAH+ODISSEI)
- Privacy issues
  - Reviews contain user names and other identifying data
  - Anonymisation, sharing with constraints

Methods & Techniques

- Syntactic parsing books and reviews
  - Trankit (and recently Alpino, Frog, Spacy, Stanza)
- Topic Modelling
  - Top2Vec
- Sentiment, Emotion and Mood Analysis (Valence-Arousal-Dominance)
  - Embeddings-based: SentiArt
  - Lexicon-based: LiLaH, LIWC, Moors lexicon, NRC lexicon, Pattern
- Linguistic analysis
  - Pronouns, tense, sentence complexity
- Statistical analysis
  - Keyness, multiple regression, Bayesian modelling

Findings

Association Between Valence Dictionaries



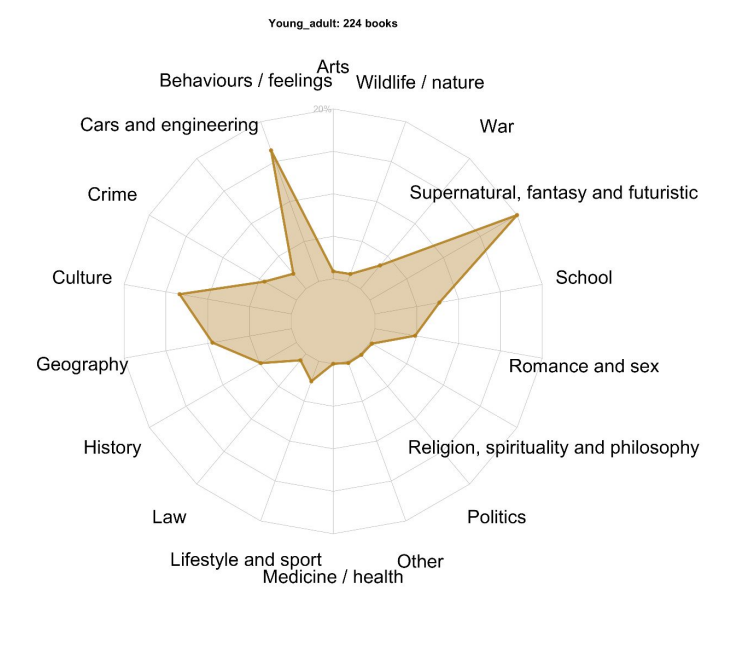
Mood

- Comparison of Valence lexicons (Boot et al. - DH 2023)
  - There is no single well-defined concept of valence
  - Definitions vary between perspective of author, world and reader
  - Most assume that context is irrelevant
  - No methods allow for ambivalence
  - Demographic groups differ in their perception of (word) valence
  - On narrative texts, tools / lexicons (for Dutch) give wildly different results
- Next step: investigate relationship of mood indicators with
  - Other book aspects, e.g. topic, pace, ...
  - Impact in reviews

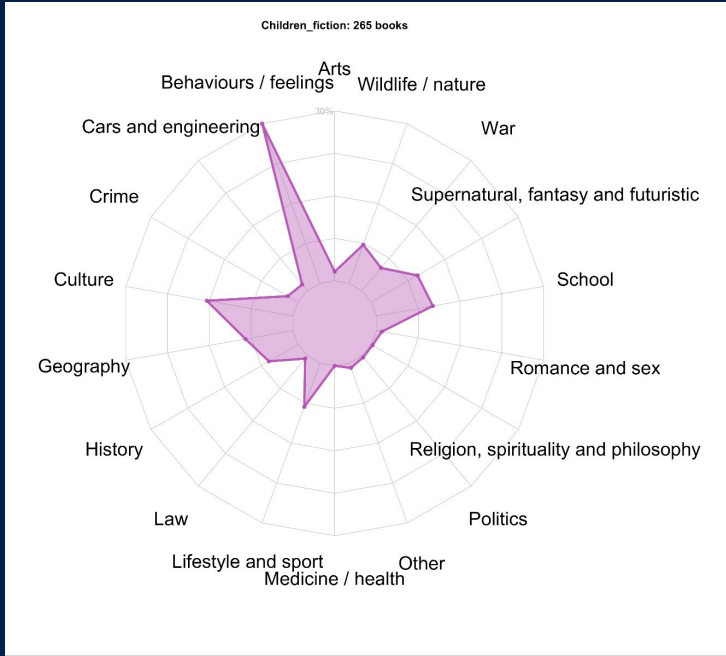
Topic

- Top2Vec on whole books
  - 94 topics, labelled by three people (and a bit of ChatGPT)
  - Grouped into 18 themes/categories

Thematics by Genre

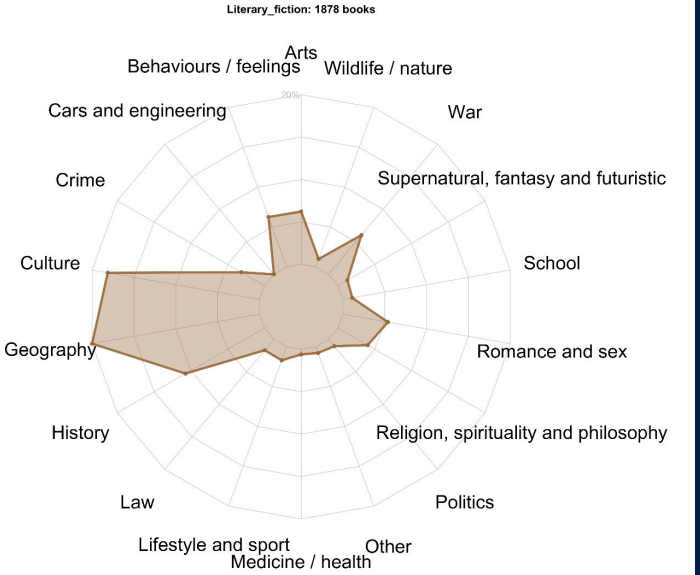


Young Adult

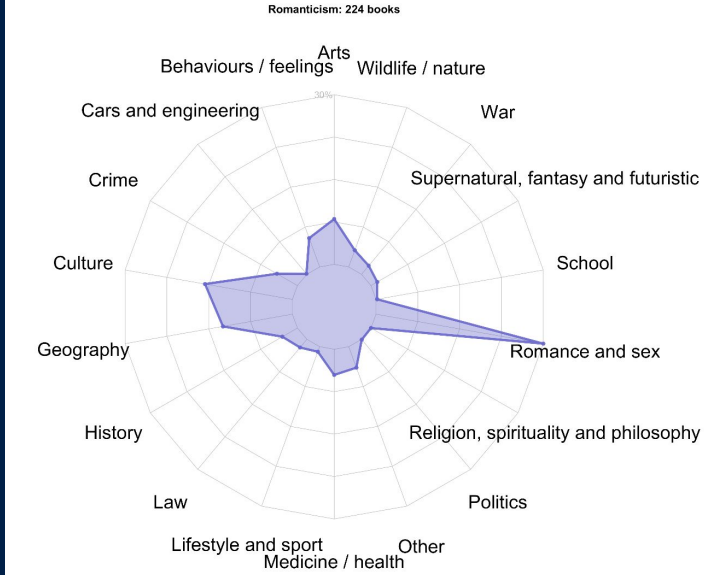


Children's Fiction

Thematics by Genre



Literary Fiction



Romanticism

Results: Prevalence of Impact Terms by Genre

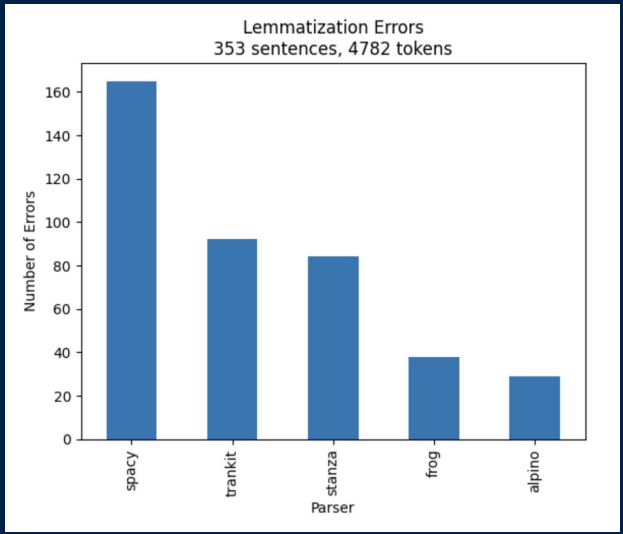
Reviews and Types of Reading

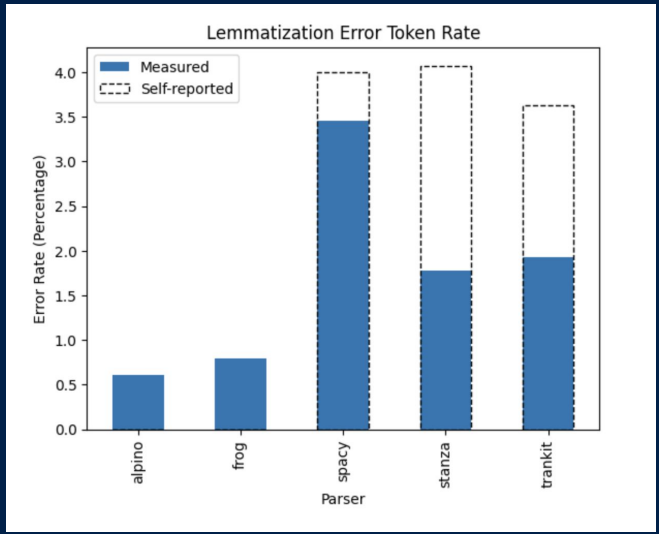
- Types of reading in online book reviews (Koolen et al. - CLIN 2022)
- Reviews on all platforms show:
  - Most reviews mention story, characters
  - Many reviews mention personal reading experience
  - Some reviews mention being transformed by reading the book
- Aligns with theories of reading
  - Interpretive vs. Experiencing reading
  - Blending of reader and story: neither first- nor third-person, but second-person (Fialho 2012, 2019)
  - Revealed by use of verbs of vivid imagination associated with 2nd person pronouns

Parsers for NL Fiction

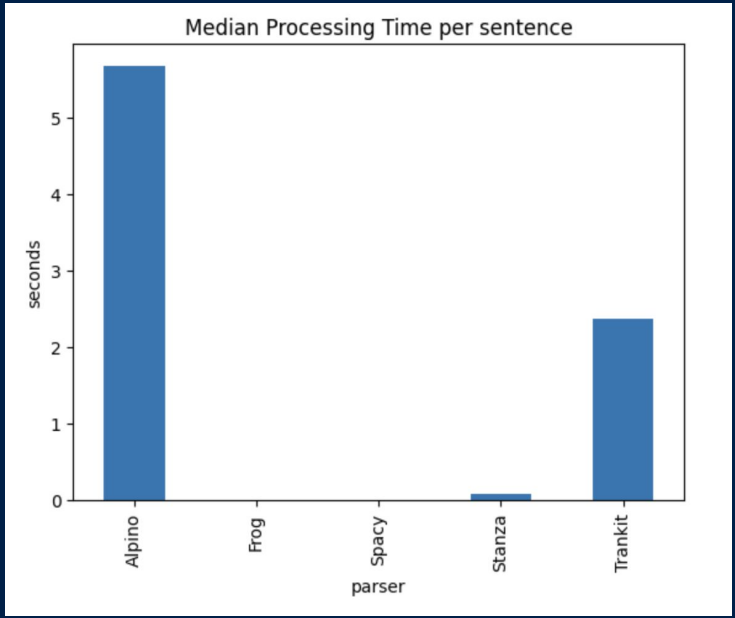
- Comparison of syntactic parsers (van Zundert et al. - CLIN 2023)
  - Alpino, Frog, SpaCy, Stanza, Trankit
- Criteria
  - Accuracy
  - Speed

Lemmatization errors





Speed


Median Processing Time per sentence
seconds
0
1
2
3
4
5
Alpino
Frog
Spacy
Stanza
Trankit
parser

Parsers for NL Fiction

- Comparison of syntactic parsers (van Zundert et al. - CLIN 2023)
  - Alpino, Frog, SpaCy, Stanza, Trankit
- Accuracy
  - Alpino is most accurate, SpaCy least
  - SpaCy and Trankit are better on our ground truth then on self-reported
- Speed
  - SpaCy is BY FAR the fastest,
  - During the project, Frog has become a lot faster
- Conclusion
  - Frog is good balance between speed and accuracy

Impact & Fiction

Katja Tereshko, Willem van Hage,
Eva Viviani, Marijn Koolen,
Peter Boot, Joris J. van Zundert