From high heels to weed attics: a syntactic investigation of chick lit and literature

Kim Jautze, Corina Koolen, Andreas van Cranenburgh & Hayco de Jong

Huygens ING
Royal Netherlands Academy of Arts and Sciences

Institute for Logic, Language and Computation
University of Amsterdam

CLfL, NAACL 2013, June 14, Atlanta
Outline

Background

The Riddle of Literary Quality Project

Paper

Syntactic complexity in chick lit and literature
Background

The Riddle of Literary Quality

It was a dark and stormy night; the rain fell in torrents — except at intervals, when it ceased abating with a gust of wind.
Literary Quality

Power?

- Sociological factors
- Literary institutions

Beauty?

- Intrinsic features
- Formal texts features
Paper

A syntactic investigation of chick lit and literature
I really don’t know why I go shopping on such high heels.

Suddenly I felt an urge to scream and to throw the smoked salmon and quiches off the table, but instead I consoled myself with the weed attic of Emiel, with the idea that I had yet more secrets and that it could be comfortable to despise the petty banter of others.
Main questions

1. What is the distribution of sentence types in chick lit and literature?

2. Does literature have more complex syntax than chick lit?
Research purposes

1. Interpret and analyze genre differences from a syntactic point of view;

2. Transform a literary-linguistic theory about syntactic structures to a computational method;

3. Explore how well the output of a statistical parser facilitates such an investigation.
Literary-linguistic theory

Syntactic structure for analyzing style of prose texts

→ Sentences types: from simple to parenthetic

→ Hierarchy of increasing complexity

(Leech and Short 1981; Toolan 2010)
Sentence types

1. Simple sentence
   [My knees feel like jelly]

2. Compound sentence
   [I could have died] and [no one did anything]

3. Complex sentence
   [I really don’t know [why I go shopping on such high heels]]

4. Complex-compound sentence
   [Suzan had heard a vague buzzing [while she was busy in the kitchen]] and [had opened the door to be safe]
Complex sentence types

3a. Trailing sentence
   Bo is too fat, because Floor feeds him macaroni

3b. Anticipatory sentence
   Because Floor feeds Bo macaroni, he is too fat

3c. Parenthetic sentence
   Bo, because Floor feeds him macaroni, is too fat
Two Principles

(1) Theme precedes rheme  
   (originally called ‘Behaghel’s second law’)

(2) The ‘complexity principle’  
   (originally ‘Law of increasing terms’)

(Behaghel, 1909; Bever 1970; Haeseryn 1997; ANS 2013)
Data

• 32 Dutch novels:
  16 chick lit novels, female authors
  16 literary novels, male & female authors

• Published between 1991 and 2011

• Texts extracted from e-books
Basic statistics

<table>
<thead>
<tr>
<th></th>
<th>chick lit</th>
<th>literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>no. of sentences</td>
<td>7064.31</td>
<td>7237.94</td>
</tr>
<tr>
<td>sent. length</td>
<td>11.90</td>
<td>14.12</td>
</tr>
<tr>
<td>token length</td>
<td>4.77</td>
<td>4.98</td>
</tr>
<tr>
<td>type-token ratio</td>
<td>0.085</td>
<td>0.104</td>
</tr>
<tr>
<td>time to parse (hrs)</td>
<td>2.05</td>
<td>5.14</td>
</tr>
</tbody>
</table>

Table 2: Basic statistics, mean by genre. Bold indicates a significant difference.
Alpino parser

Figure 1: A sentence from ‘Zoek Het Maar Uit’ by Chantal van Gastel, as parsed by Alpino. Translation: *His jawline is almost square.*
Alpino output

- Syntactic categories: NP, VP, &c.
- Grammatical functions: SBJ, OBJ, &c.
- Parts-of-speech: ADJ, VERB, &c.
- Morphological features: plural, past, &c.

Special category: Discourse Unit (DU), signifies:
- Asyndetic construction ("But... why?")
- Extensions to main clause ("Great, isn’t it?")
Queries

Formulation simple sentence:

- contains a main clause
- that does not introduce a conjunction
- does not contain subordination at any level
- is not a discourse unit
Treebank Search

A web-interface to TGrep2 on a corpus of parse trees. Enter a TGrep2 pattern (toggle TGrep overview) to search the corpus. Results can be displayed in various ways and morphological tags of the form TAG[MORPHI,...] and grammatical functions of the form LABEL-FUNC can optionally be filtered out.

Show surface characteristics of texts (using style/11)

Query: 

Input:
16 texts selected: show/hide selection
(Select all; none)
- C_Gastel_Chantal_van__Zoek_het_maar_uit.mrg
- C_Gastel_Chantal_van__Zwaar_verliefd.mrg
- C_Harrewijn_Astrid__In_zeven_sloten.mrg
- C_Harrewijn_Astrid__Luchtkussen.mrg
- C_Middelbeek_Mariette__Ravanche_in_New_York.mrg
- C_Middelbeek_Mariette__Single_En_Sexy.mrg
- C_Middelbeek_Mariette__Status_OK..mrg
- C_Middelbeek_Mariette__Status_OK..mrg
- L_Beiijnum_Kees_Van__De_Ordering.mrg
- L_Beiijnum_Kees_van__De_oesters_van_Nam_Kee.mrg
- L_Glastra_van_Loon_Karel__De_Passievrucht.mrg
- L_Grunberg_Amon__De_Asiazoeker.mrg
- L_Grunberg_Amon__Huid_en_haar.mrg
- L_Japin_Arthur__De_grote_wereld.mrg
- L_Japin_Arthur__Vaslav.mrg
- L_Moor_Margriet_de__De_Schilder_en_het_Melsje.mrg
- L_Moor_Margriet_de__De_verdronkene.mrg

Output:
- Counts; ☐ Matching sentences (words only); ☐ Draw matching trees;
- Matches in bracket notation; ☐ Export all matching trees as text file;

Normalization:
- % of Sentences; ☐ % of Constituents; ☐ % of Words;
Treebank Search

A web-interface to TGrep2 on a corpus of parse trees. Enter a TGrep2 pattern (toggle TGrep overview) to search the corpus. Results can be displayed in various ways and morphological tags of the form \texttt{TAG\_MORPH} and grammatical functions of the form \texttt{LABEL\_FUNC} can optionally be filtered out.

Show surface characteristics of texts (using style(1))

| Query: | CONJ < /PP-/ |

Input: 16 texts selected; show/hide selection

Output:
- Counts
- Matching sentences (words only)
- Draw matching trees
- Matches in bracket notation
- Export all matching trees as text file

Normalization:
- \% of Sentences
- \% of Constituents
- \% of Words

NB: the counts reflect the total number of times a pattern matches for each tree.
Results on sentence types

Figure: Overview of sentence tests.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>LITERATURE</th>
<th>CHICK LIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple</td>
<td>29.9</td>
<td>32.4</td>
</tr>
<tr>
<td>complex</td>
<td>17.9</td>
<td>16.0</td>
</tr>
<tr>
<td>compound</td>
<td>6.2</td>
<td>8.6</td>
</tr>
<tr>
<td>compound-complex</td>
<td>3.9</td>
<td>4.9</td>
</tr>
<tr>
<td>DU simple</td>
<td>4.6</td>
<td>6.0</td>
</tr>
<tr>
<td>DU compound, not complex</td>
<td>11.0</td>
<td>8.4</td>
</tr>
<tr>
<td>DU complex, both</td>
<td>11.5</td>
<td>7.6</td>
</tr>
<tr>
<td>trailing</td>
<td>5.6</td>
<td>5.7</td>
</tr>
<tr>
<td>anticipatory</td>
<td>1.2</td>
<td>1.0</td>
</tr>
<tr>
<td>parenthetical</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Percentage
Morphosyntactic features

<table>
<thead>
<tr>
<th></th>
<th>chick lit %</th>
<th>LIT. %</th>
</tr>
</thead>
<tbody>
<tr>
<td>noun phrases</td>
<td>6.4</td>
<td>8.0</td>
</tr>
<tr>
<td>prepositional phrases</td>
<td>5.5</td>
<td>6.5</td>
</tr>
<tr>
<td>prep. phrases (modifiers)</td>
<td>2.2</td>
<td>2.9</td>
</tr>
<tr>
<td>relative clauses</td>
<td>0.32</td>
<td>0.50</td>
</tr>
<tr>
<td>diminutives (% of words)</td>
<td>0.79</td>
<td>0.49</td>
</tr>
</tbody>
</table>

Table 5: Tests on morphosyntactic features. Bold indicates a significant difference.

In sum, although not significantly different for regular sentences, the Class I and II identification show that the genres tend to differ in the distribution of sentence types and complexity. With more data, the other tests may show significant differences as well.

Especially the complex discourse units are good predictors of the two genres. This is crucial as DUs in general appear to be characteristic of narrative text, which typically contain extensive dialogue and informal speech. However, not all dialogue is identified as a discourse unit, because we did no preprocessing to identify all sentences in quoted speech as being part of dialogue. Therefore, the actual amount of dialogue per novel remains unclear.

5 Results on morphosyntactic features

In addition to the deep syntactic results based on the top-down approach, we take a closer look at the syntactic categories in the generated trees. The results can be found in figure 3 and table 5.

5.1 Relative clauses

Figure 5 shows a substantial difference in the number of relative clauses used in literature and chick lit ($p = 0.0005$). Relative clauses modify noun phrases to describe or identify them. Therefore the relative clause makes the NP 'heavier'. The syntax prefers the relative clause to be placed directly after the NP, although they may be extraposed for pragmatic reasons. When the NP is a subject, this causes the head noun of the NP to be distant from the main verb:

(9) De mensen [REL die even eerder nog zo rustig op de vloer hadden zitten mediteren], sprongen nu dansend en schreeuwend om elkaar heen. (L)

The people who just moments before had been meditating quietly on the floor, were now jumping around each other dancing and screaming.

The relative clause interrupts the relation between the subject and the predicate, but to a lesser extent than in a parenthetic sentence structure. With relative clauses there is also a disruption of the expected information flow, and this contributes to making such sentences more complex to process (Gibson, 1998).

Furthermore, the higher number of relative clauses in the literary novels makes the sentences more elaborate. In *Chick lit: the new woman’s fiction* Wells argues a similar point to make a distinction between the genres:

“[T]he language of chick-lit novels is unremarkable, in a literary sense. Richly descriptive or poetic passages, the very bread and butter of literary novels, both historical and contemporary, are virtually nonexistent in chick lit.” (Wells, 2005, p. 65)

5.2 Prepositional phrases

Given the longer average sentence length of literature, it is to be expected that the prepositional phrases (PPs; as well as noun phrases; NPs) occur more frequently in literary novels than in chick lit ($p = 0.0044$ and...
Morphosyntactic features

Relative Clauses

The people who just moments before had been meditating quietly on the floor, were now jumping around each other dancing and screaming.

Noun phrases and prepositional phrases

The flower in the corner by the room in the window in the sun said it all.
Morphosyntactic features

Ineens had ik zin om te schreeuwen en de gerookte zalm en quiches van tafel te slaan, [PP-MOD maar [MWU-HD in plaats daarvan]] troostte ik me [PP-PC met de wietzolder [PP-MOD van [N-OBJ Emiel]], [PP-MOD met [NP-OBJ de gedachte dat ik nog meer geheimen had en dat het behaaglijk kon zijn]] [NP-OBJ het slappe geklets [PP-MOD van [N-OBJ anderen]] te verachten]

Suddenly I felt an urge to scream and throw the smoked salmon and quiches of the table, but instead I consoled myself with the weed attic of Emiel, with the idea that I had yet more secrets and that it could be comfortable to despise the petty banter of others.
Literary language may be more complex & descriptive than the language of chick lit

“The language of chick-lit novels is unremarkable, in a literary sense. Richly descriptive or poetic passages, the very bread and butter of literary novels, both historical and contemporary, are virtually nonexistent in chick lit.”

(Wells, 2005, p. 65)
Conclusions

1. literature is more complex

2. distant reading is useful

3. correlates with aesthetic quality of the texts?
Thank you!

An ear-shattering applause breaks loose.
Summary

Chick lit  more **simple** and **compound** sentences
tendency more **trailing** sentence structures
more **diminutives**

Literature  more **complex** sentences
tendency more **anticipatory & parenthetical**
more **relative clauses, PPs and NPs**

URL project:  [http://literaryquality.huygens.knaw.nl/](http://literaryquality.huygens.knaw.nl/)
Dutch original sentences

(1) Mijn knieën voelen als pudding. (C)

(2) Ik had dood kunnen zijn en niemand deed iets. (C)

(3) Ik weet ook niet waarom ik op van die hoge hakken ga shoppen. (C)

(4) Suzan had een vaag gezoem gehoord terwijl ze bezig was in de keuken en had voor de zekerheid de deur opengedaan. (L)

(5) De mensen [REL die even eerder nog zo rustig op de vloer hadden zitten mediteren], sprongen nu dansend en schreeuwend om elkaar heen. (L)