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

Jautze, K.J.; Koolen, C.W.; van Cranenburgh, Andreas; de Jong, H.A.

***published in***
Proceedings of the Workshop on Computational Linguistics for Literature
2013

***document version***
Publisher's PDF, also known as Version of record

# 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 Dutch Academy of Science
P.O. box 90754, 2509 LT, The Hague, The Netherlands
{Kim.Jautze,Hayco.de.Jong}@huygens.knaw.nl

[†]Institute for Logic, Language and Computation
University of Amsterdam
Science Park 904, 1098 XH, The Netherlands
{C.W.Koolen,A.W.vanCranenburgh}@uva.nl

## Abstract

Stylometric analysis of prose is typically limited to classification tasks such as authorship attribution. Since the models used are typically black boxes, they give little insight into the stylistic differences they detect. In this paper, we characterize two prose genres syntactically: chick lit (humorous novels on the challenges of being a modern-day urban female) and high literature. First, we develop a top-down computational method based on existing literary-linguistic theory. Using an off-the-shelf parser we obtain syntactic structures for a Dutch corpus of novels and measure the distribution of sentence types in chick-lit and literary novels. The results show that literature contains more complex (subordinating) sentences than chick lit. Secondly, a bottom-up analysis is made of specific morphological and syntactic features in both genres, based on the parser's output. This shows that the two genres can be distinguished along certain features. Our results indicate that detailed insight into stylistic differences can be obtained by combining computational linguistic analysis with literary theory.

## 1 Introduction

The gap between literary theory and computational practice is still great. Despite pleas for a more integrated approach (e.g., Ramsay, 2003), and suggestions from literary theorists (e.g., Roque, 2012), literary theory is more often used for illustrative or explicative purposes, rather than as a basis for computational analysis. The hermeneutic nature of most literary theory is a valid cause for caution, as it is not easy to 'translate' discursive arguments into the strict rules a computer needs. Too many intermediary steps are required, if a translation is possible at all.

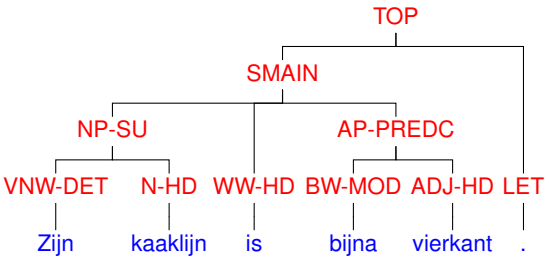


Figure 1: A sentence from 'Zoek Het Maar Uit' by Chantal van Gastel, as parsed by Alpino. Translation: *His jawline is almost square.*

We therefore take a different approach in this paper. Instead of building on hermeneutic theory, we use a literary-linguistic theory about syntactic structures as a basis for developing a computational method for prose genre analysis; in this paper we will focus on chick-lit and literary novels. Because of this tight connection between theory and method, these usually separate sections are combined. In addition to this top-down approach, we report bottom-up findings based on syntactic features encountered in the data. These complementary results will be used to further analyze and interpret genre differences, as opposed to author style. Our aim is not text categorization, but to describe the genres from a syntactic point of view.

We have chosen the genres of chick lit and literature, because many readers have intuitive notions on differences between them. In this paper we want to find out whether it is possible to retrace these notions

in syntactic properties of the texts, by addressing the following questions: (i) are there differences in the distribution of sentence types between chick lit and literary novels, (ii) is the intuitive notion that chick lit is easier to read reflected in a tendency towards simple sentence structures rather than complex ones? In answering these questions, two methodological goals are achieved simultaneously: we discover how a specific literary-linguistic theory can be transformed to a computational method and we explore how well the output of a statistical parser facilitates such an investigation.

This study is a first exploration in a project called The Riddle of Literary Quality,[1] which aims to find patterns in the texts of Dutch current-day novels, that relate to the distinction between high-brow and low-brow literature. Deep syntactic structures as researched in the present paper are an important aspect of this investigation.

## 2 Theory and method

According to linguists Leech and Short (1981) syntactic structure is one of the grammatical features that can be taken into account when analyzing the style of prose texts. To this end, they make a division between six basic sentence types, from simple to parenthetic.

Toolan (2010) applies their theory by close-reading a paragraph from a short story by Alice Munro. He suggests that the six sentence types are part of a hierarchy of increasing complexity, a notion we will explore further by taking a distant reading approach, namely syntactically analyzing a prose corpus. In recent computational work on syntactic stylistics by Feng et al. (2012) and van Cranenburgh (2012) computational explorations of deep syntactic structures in academic and literary writing styles are undertaken on a similar scale. They make use of a machine learning methodology in which the results are evaluated on objective criteria, in this case authorship.

In line with this previous research we want to examine whether the use of certain types of sentence structures can inform our understanding of the difference between two prose genres, chick lit and literature. As opposed to Feng et al. (2012) however, we do not rely on black box machine learning approaches. And instead of extracting arbitrary syntactic patterns as in van Cranenburgh (2012), we target specific syntactic features, based partially on literary-linguistic theory as well as manual exploration of the data. To be more specific, the computational tools we employ deliver syntactic structures by querying the structures for certain syntactic features. During the development of our method, we continually verify our intuitions against the actual data.

To categorize the sentences into types, we devise two classifications, based on a combination of the theory developed by Leech and Short (1981) and Toolan (2010) and computational tests in Feng et al. (2012).

**Class I**

1. Simple: one main clause, no subordination on any level in the parse tree
2. Compound: coordination of sentence-level clauses, no subordination on any level
3. Complex: subordination anywhere in the sentence, no top-level coordination
4. Complex-compound: coordination on top-level and subordination

Leech and Short's definition does not specify whether non-finite or relative clauses that modify noun phrases count towards being a complex sentence. According to the ANS (2013), the Dutch standard reference work on grammar, all sentences with more than one connection between a subject and predicate are 'composed,' thus not 'singular' or simple. We therefore choose to count all subordinating clauses as making a sentence complex.

See (1)–(4) for examples of each sentence type.[2] An (L) indicates a sentence from the literature corpus, and a (C) a sentence from the chick lit corpus.

**Simple sentence:**

(1) a. Sjaak schraapte zijn keel. (L)
*Sjaak cleared his throat.*
b. Mijn knieën voelen als pudding. (C)
*My knees feel like jelly.*

**Compound sentence:**

(2) Ik had dood kunnen zijn en niemand deed iets. (C)
*I could have died and no one did anything.*

[1] Cf. `http://literaryquality.huygens.knaw.nl`

[2] These are examples from the novels in our corpus; cf. table 1.

**Complex sentence:**

(3) Ik weet ook niet waarom ik op van die hoge hakken ga shoppen. (C)
*I really don't know why I go shopping on such high heels.*

**Complex-compound sentence:**

(4) Suzan had een vaag gezoem gehoord terwijl ze bezig was in de keuken en had voor de zekerheid de deur opengedaan. (L)
*Suzan had heard a vague buzzing while she was busy in the kitchen and had opened the door to be safe.*

The second classification describes the distribution of several types of complex sentences, based on Toolan's hierarchical ordering of complex sentence types. This concerns sentences consisting of a dependent and main clause at the top level:

## Class II

1. Trailing: main clause followed by subordinating clause
2. Anticipatory: subordinating clause followed by main clause
3. Parenthetic: subordinating clause interrupting a main clause

Toolan argues that the complex sentences, especially the anticipatory and parenthetic ones, are more demanding to process than the simple and compound sentences, because of a disruption in the linear clause-by-clause processing (Toolan, 2010, p. 321).

This can be explained by two principles: (1) the principle that theme precedes rheme (originally called 'Behaghel's second law') and (2) the 'complexity principle' (originally 'Law of increasing terms') (Behaghel, 1909). The first principle concerns the content: the less informative or important elements are placed before what is important or new. Usually, the new information is introduced by the subordinate clause and is therefore placed after the main clause. The second principle argues that generally the more complex and longer elements–'heavier' constituents containing more words and elaborate syntax–tend to be placed at the end of the sentence (Behaghel, 1909; Bever, 1970). These principles also apply to Dutch; cf. Haeseryn (1997, p. 308) and ANS (2013). With respect to the content and syntactic dependency, subordinate clauses are more demanding and complex, thus at best in this final position.

**Trailing sentence**

(5) Bo is te dik, omdat Floor hem macaroni voert.
*Bo is too fat, because Floor feeds him macaroni.*

**Anticipatory sentence**

(6) Omdat Floor Bo macaroni voert, is hij te dik.
*Because Floor feeds Bo macaroni, he is too fat.*

**Parenthetic sentence**

(7) Bo is, omdat Floor hem macaroni voert, te dik.
*Bo, because Floor feeds him macaroni, is too fat.*

We parse the corpus with the Alpino parser (Bouma et al., 2001; van Noord, 2006) to obtain syntactic parse trees (e.g., figure 1). The output of Alpino is in the form of dependency trees, containing both syntactic categories and grammatical functions. In order to work with tools based on constituency trees, we convert any non-local dependencies to discontinuous constituents, and apply the transformation described by Boyd (2007) to resolve discontinuities. For example, the Dutch equivalent of a phrasal verb such as "Wake [NP] up" might be parsed as a discontinuous `VP` constituent, but will be split up into two separate constituents `VP*0` and `VP*1`, bearing an implicit relation encoded in the label.

In order to categorize the parsed sentences in Class I and II, we build two different sets of queries: one for the trees wherein the main clause is a direct child of the TOP-node, and another for the parsed trees that introduce an extra node (DU) between the TOP and the main clause. The former are the 'regular' sentences that comprise approximately 67 % of the corpus, the latter are the so-called 'discourse units' (DUs) that comprise 33 %. DUs incorporate extensions to the sentence nucleus; cf. (8a) and (8b), constructions which depend on discourse relations (8c), and implicit conjunctions (8d).

(8) a. [$_{\text{DU}}$ [$_{\text{SMAIN-NUCL}}$ dat verbaast me ] , [$_{\text{SAT}}$ dat je dat nog weet ] ]
*that surprises me, that you still remember that*

b. [$_{\text{DU}}$ [$_{\text{SMAIN-TAG}}$ Hij verklaarde ] : [$_{\text{SMAIN-NUCL}}$ "Ik kom niet" ] ]
*He declared: "I won't come"*

c. [$_{\text{DU}}$ dus [$_{\text{SMAIN-NUCL}}$ Jan gaat naar huis. ] ]
*So Jan is going home.*

d. (welke kranten lees jij?) [$_{\text{DU}}$ [$_{\text{DU-DP}}$ bij de lunch de Volkskrant ] ; [$_{\text{DU-DP}}$ s avonds de NRC ]
*(which newspapers do you read?) at lunch the Volkskrant; at night the NRC*

(van Noord et al., 2011, p.182–192)

| CHICK LIT | LITERATURE |
|---|---|
| Gastel, Chantal van - Zoek het maar uit (2011) | Beijnum, Kees van - De oesters van Nam Kee (2000) |
| Gastel, Chantal van - Zwaar verliefd (2009) | Beijnum, Kees Van - De Ordening (1998) |
| Harrewijn, Astrid - In zeven sloten (2007) | Dorrestein, Renate - Een sterke man (1994) |
| Harrewijn, Astrid - Luchtkussen (2009) | Dorrestein, Renate - Hart van steen (1998) |
| Hollander, Wilma - Bouzouki Boogie (2011) | Dorrestein, Renate - Het hemelse gerecht (1991) |
| Hollander, Wilma - Dans der liefde (2010) | Enquist, Anna - De Thuiskomst (2005) |
| Hollander, Wilma - Onder de Griekse zon (2008) | Enquist, Anna - De Verdovers (2011) |
| Middelbeek, Mariette - Revanche in New York (2006) | Enquist, Anna - Het meesterstuk (1994) |
| Middelbeek, Mariette - Single En Sexy (2009) | Glastra van Loon, Karel - De Passievrucht (1999) |
| Middelbeek, Mariette - Status O.K. (2010) | Glastra van Loon, Karel - Lisa's Adem (2001) |
| Verkerk, Anita - Als een zandkorrel in de wind (1994) | Grunberg, Arnon - De Asielzoeker (2003) |
| Verkerk, Anita - Bedrogen liefde (2006) | Grunberg, Arnon - Huid en haar (2010) |
| Verkerk, Anita - Cheesecake & Kilts (2010) | Japin, Arthur - De grote wereld (2006) |
| Verwoert, Rianne - Match (2009) | Japin, Arthur - Vaslav (2010) |
| Verwoert, Rianne - Schikken of stikken (2010) | Moor, Margriet de - De Schilder en het Meisje (2010) |
| Verwoert, Rianne - Trouw(en) (2009) | Moor, Margriet de - De verdronkene (2005) |

Table 1: The corpus

The translation of Alpino-tags into queries is as follows (van Noord et al., 2011):

1. Categories for main clauses: SMAIN (declaratives), SV1 (verb initial: imperatives, polar questions) and WHQ (WH-questions).
2. Categories for finite subordinate clauses: SSUB (V-final), WHSUB (constituent questions), and (WH)REL (relative clauses).
3. Categories for non-finite subordinate clauses: PPART (perfect tense), INF (bare infinitives), TI (to-infinitives), and OTI ('om te' + INF) when accompanied by the BODY-function. Without BODY, PPART and INF can also be part of a simple sentence.
4. Functions used with DU: DP (discourse part), NUCL (sentence nucleus) SAT ("satellite" of the sentence, comparable with subordinate clauses)[3] and TAG (tag questions: 'isn't it?', 'you know?', dialogue markers: 'he said', etc.)

The query language used is TGrep2 (Rohde, 2005). For example, we identify simple sentences using the following query:

```
TOP !< DU < ( /SMAIN|SV1|WHQ/ !< /CONJ/ )
    !<< /WHSUB|SSUB|(PPART|TI|INF)-BODY/
```

This query matches a TOP node which does not have a DU child, but does have a SMAIN, SV1, or WHQ child. This child, in turn, must not have one of the categories signifying a conjunction or subordinate clause, at any level.

[3]The Alpino treebank annotation uses the terminology of nucleus and satellite, originally from Rhetorical Structure Theory (Mann and Thompson, 1988).

| | chick lit | literature |
|---|---|---|
| no. of sentences | 7064.31 | 7237.94 |
| sent. length | 11.90 | **14.12** |
| token length | 4.77 | **4.98** |
| type-token ratio | 0.085 | **0.104** |
| time to parse (hrs) | 2.05 | **5.14** |

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

We test for statistical significance of the syntactic features with a two-tailed, unpaired t-test. We consider $p$-values under 0.05 to be significant. We present graphs produced by Matplotlib (Hunter, 2007), including standard deviations among texts of each genre.

## 3 Data

Our corpus is composed of 32 Dutch novels, equally divided between the genres chick lit and literature, and published between 1991 and 2011, cf. table 1. These novels were selected from a collection of ebooks; the number of each set was restricted by the number of chick-lit novels available. Female and male writers should ideally be equally represented, to avoid gender being a possible confounding factor. Since the chick-lit novels at our disposal were all written by women, this was not possible for that genre. The genre distinctions are based on classifications

| | CHICK LIT % | LIT. % |
|---|---|---|
| simple | 32.36 | 29.87 |
| compound | 8.54 | 6.23 |
| complex | 16.10 | 17.93 |
| complex-compound | 4.94 | 3.86 |
| DU simple | 5.98 | 4.56 |
| DU compound | 8.36 | 11.02 |
| DU complex (compound or not) | 7.64 | **11.52** |

Table 3: Sentence Class I identification, regular and DU-sentences. Bold indicates a significant difference.

by the publisher and reviews on `www.chicklit.nl`. For selecting literature we employed an additional criterion: the writer of the literary novel(s) has had to be accredited by winning at least one Dutch national literary prize.

Table 2 lists basic surface characteristics of the genres. A salient detail is that the literary novels took significantly longer to parse than the chick-lit novels ($p = 0.0001$), which cannot be attributed solely to longer sentence length, because the difference remains when correcting for the cubic time complexity of parsing—viz. $O(nm^3)$, with $n$ the number of sentences, and $m$ average sentence length.

## 4 Results on sentence types

Table 3 shows the results for Class I. The queries could classify approximately 60 % out of the 67 % regular sentences and 24.5 % out of the of 33 % discourse units into one of these four basic sentence types. Since DU-sentences often contain multiple main clauses without an explicit form of conjunction, it is difficult to define when a sentence is a compound rather than a complex sentence. Therefore we do not distinguish between compound and non-compound for complex DU-sentences, cf. 'DU complex' in table 3.

The remaining 15.5 % of the sentences in our corpus cannot be classified by our queries and would therefore fall into a residual category. This is (probably) due to single-word and verbless sentence fragments that do not fit into any of the categories and are therefore not captured by any of the formulated queries.

| | CHICK LIT % | LIT. % |
|---|---|---|
| trailing | 6.50 | 6.32 |
| anticipatory | 1.03 | 1.20 |
| parenthetic | 0.01 | **0.03** |

Table 4: Sentence Class II identification. Bold indicates a significant difference.

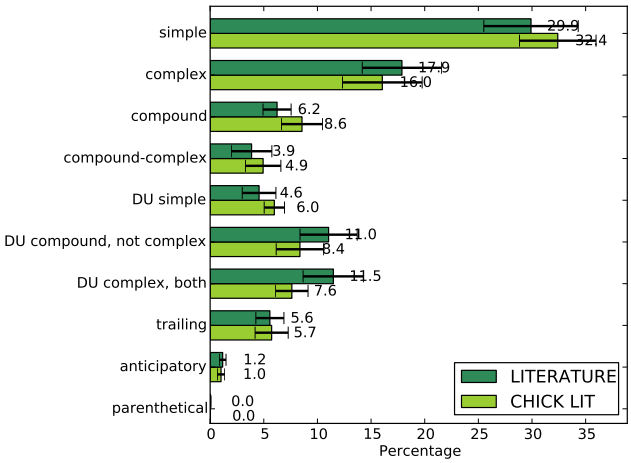


Figure 2: Overview of sentence tests.

The Class I identification shows that chick-lit authors tend to use more simple sentence structures and literary writers prefer complex ones, in both regular and DU-type sentences.[4] Although this difference is not significant for regular sentences, this may have been caused by the relatively small size of the corpus. In the discourse type-sentences DU complex (both with and without coordination) does show a significant difference. DU complex predicts genre adequately ($p = 0.003$; cf. figure 4), indicating that dialogue sentences might be a better predictor for genre differences than narrative sentences.

The results for Class II identification can be found in table 4. Although the difference is not significant, in chick lit we do find a tendency towards the use of more trailing sentences, as opposed to more anticipatory sentences in literary novels. The difference in use of parenthetic structure is significant

[4]When taking a closer look at the constituents, the TI, OTI and BODY-INF clauses are the exception, because they are more often used in chick-lit novels. TI and OTI introduce to-infinitives, e.g., I want *to sleep*, and the BODY-INFs are bare infinitive clauses. These three are the least complex of the subordinating clauses.

| | chick lit % | LIT. % |
|---|---|---|
| noun phrases | 6.4 | **8.0** |
| prepositional phrases | 5.5 | **6.5** |
| prep. phrases (modifiers) | 2.2 | **2.9** |
| relative clauses | 0.32 | **0.50** |
| diminutives (% of words) | 0.79 | 0.49 |

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

($p = 0.014$), but because of the negligible number of occurrences, this is not a reliable predictor. Relating these results to Toolan's theory that sentence types of Leech and Short are ordered according to increasing complexity—i.e., that anticipatory and parenthetic sentences are more demanding to process and therefore more complex—this tendency could be an indicator of a measurably higher syntactic complexity in literary novels.

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 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,

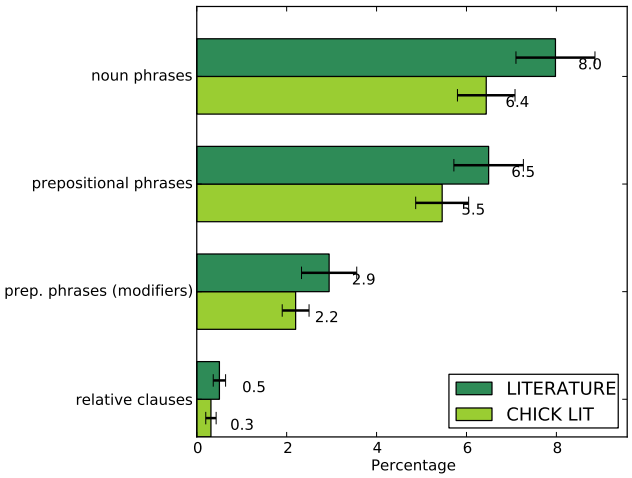


Figure 3: Overview of morphosyntactic tests.

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 [$_{\text{REL}}$ 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

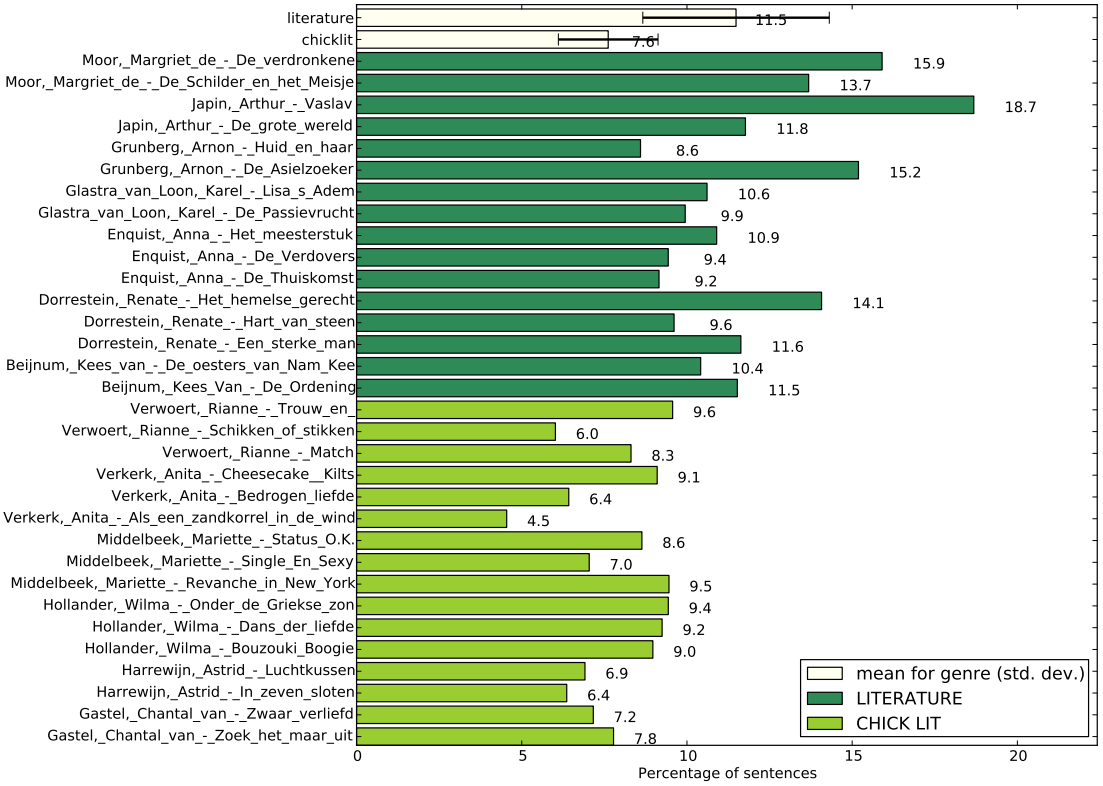


Figure 4: Distribution of complex DU-sentences.

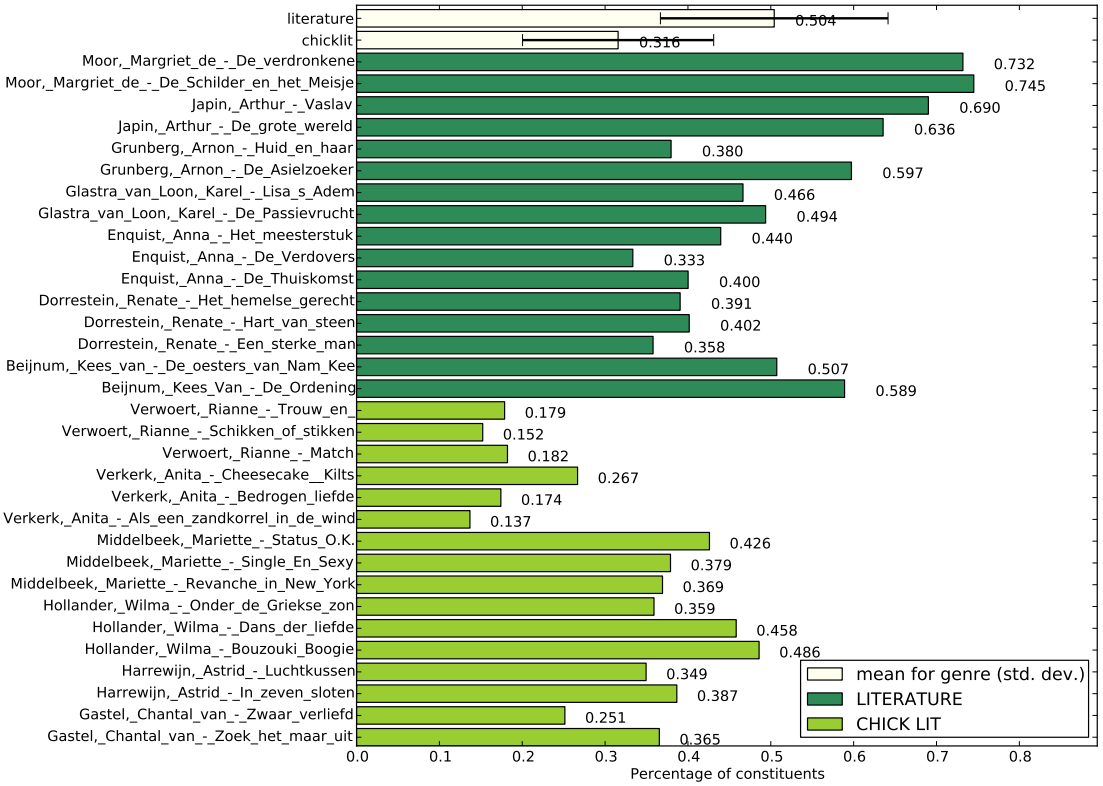


Figure 5: Relative clauses in each text.

$p = 0.0015$, respectively). The aforementioned argument by Wells that chick lit is less descriptive than literature is reflected in the results of the PPs and NPs as well. PPs, especially PP-adjuncts—grammatically optional constituents that function as modifiers of other constituents—are also indicative of descriptive language. Example (10) shows liberal use of prepositional phrases, including the first two PP-MODs that modify the same constituent—although the latter was not attached correctly by the parser.

(10) Ineens had ik zin om te schreeuwen en de gerookte zalm en quiches van tafel te slaan, [$_{\text{PP-MOD}}$ maar [$_{\text{MWU-HD}}$ in plaats daarvan]] troostte ik me [$_{\text{PP-PC}}$ met de wietzolder [$_{\text{PP-MOD}}$ van [$_{\text{N-OBJ1}}$ Emiel]], [$_{\text{PP-MOD}}$ met [$_{\text{NP-OBJ1}}$ de gedachte dat ik nog meer geheimen had en dat het behaaglijk kon zijn]] [$_{\text{NP-OBJ1}}$ het slappe geklets [$_{\text{PP-MOD}}$ van [$_{\text{N-OBJ1}}$ anderen]] te verachten] (L)
*Suddenly I felt an urge to scream and throw the smoaked 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.*

In sum, both the relative clauses and the PPs differentiate between literature and chick lit and point towards more descriptive language in literature.

### 5.3 Diminutives

Another marker for the distinction between chick lit and literature is the use of diminutives (almost significant, p=0.055). In Dutch, the diminutive is a productive part of the language and is typically formed by the suffix '-je'. Alpino marks such words with the morphological feature 'DIM.' The frequent use of the diminutive is a common element in colloquial speech, and aside from the literal meaning of smallness diminutives are also used to express endearment, intimacy, and familiarity:

(11) Ik draai me om en pak mijn telefoontje. (C)
*I turn around and take my telephone*-DIM.

This may indicate that language in chick lit is closer to real-life speech than that of literature and could be explored further when the speech-narrative distinction is operationalized.

## 6 Discussion

A starting point for further exploration is offered by our finding that the complex DU-sentences clearly differentiate between chick lit and literature. Something similar is suggested by Egbert (2012), who uses Multi-Dimensional analysis to explore literary styles. He identifies stylistic variation in the dimensions of Thought Presentation versus Description, and Dialogue versus Narrative. This finding supports our conclusion that it would be fruitful to pursue an intratextual distinction of regular versus dialogue sentences. In future research the method could for instance be expanded by using a discourse analyzer to identify all dialogue sentences. This will require some notion of a text grammar (Nunberg, 1990; Power et al., 2003), to recognize the different ways in which dialogue can be represented in text.

In order to assess the fitness of statistical parsers for literary investigations, a more comprehensive study of the quality of the parse trees is in order. The trees we have inspected were overall of good quality, especially concerning the elements we analyze. These consist mostly of overtly marked syntactic constituents, and do not hinge on correct attachments, which are often difficult to get right for statistical parsers.

Furthermore, we would like to investigate Toolan's claims about the complexity of sentence types, and on more specific morphosyntactic features. Unfortunately, little theory exists on syntactic aspects of literature, let alone its complexity. This could be improved by using results from psycholinguistics on what kinds of syntactic constructions are perceived as complex. Related to this is the work concerning readability measures, such as the Flesch and Kincaid scales, which can be obtained with the `style` program (Cherry and Vesterman, 1981).

Finally, in future work we would like to combine our computational results with literary interpretation. This requires attending to the context of the syntactic features in question.

## 7 Conclusion

We have operationalized a literary-linguistic theory by employing several computational tools and found specific syntactic features that characterize the two prose genres. Especially the Discourse Units showed

a marked difference between the genres: chick lit uses more compound sentences, whereas literature contains more complex sentences. The bottom-up tests showed that chick-lit writers use significantly more diminutives, whereas literary writers employ more prepositional phrases and relative clauses which results in more descriptive language.

Although these findings agree with intuitive notions that literature employs more complex syntactic constructions than chick lit, computational analysis has proven its added value. The distant reading method of sifting through large amounts of text can reveal patterns too subtle or diffused to spot without computational tools; the distribution of the specific sentence structures we have investigated here would have been cumbersome to extract manually.

Our approach of analyzing syntactic features yields promising results on characterizing prose genre and explaining the syntactic differences. The positive results mean that the method that we have applied can be developed further in the context of the project The Riddle of Literary Quality to find out whether syntactic complexity correlates with the perceived aesthetic quality of the texts as well.

## Acknowledgments

We are grateful to Isaac Sijaranamual for supplying us with a collection of ebooks and timely advice on pre-processing, and to Emy Koopman for suggestions on statistical matters. We thank Karina van Dalen-Oskam, Rens Bod, and Sally Wyatt for reading drafts, and the reviewers for helpful comments.

This paper is part of the project The Riddle of Literary Quality, supported by the Royal Netherlands Academy of Arts and Sciences as part of the Computational Humanities program.

## References


ANS. 2013. Algemene Nederlandse Spraakkunst (ANS). URL http://ans.ruhosting.nl/.

Otto Behaghel. 1909. Beziehungen zwischen umfang und reihenfolge von satzgliedern. *Indogermanische Forschungen*, 25:110–142.

Thomas G. Bever. 1970. The cognitive basis for linguistic structures. In J.R. Hayes, editor, *Cognition and the Development of Language*, pages 279–362. Wiley, New York.

Gosse Bouma, Gertjan van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers*, 37(1):45–59.

Adriane Boyd. 2007. Discontinuity revisited: An improved conversion to context-free representations. In *Proceedings of the Linguistic Annotation Workshop*, pages 41–44. URL http://aclweb.org/anthology/W/W07/W07-1506.

Lorinda L. Cherry and William Vesterman. 1981. Writing tools—the STYLE and DICTION programs. Computer Science Technical Report 91, Bell Laboratories, Murray Hill, N.J. Republished as part of the 4.4BSD User's Supplementary Documents by O'Reilly.

Jesse Egbert. 2012. Style in nineteenth century fiction: A multi-dimensional analysis. *Scientific Study of Literature*, 2(2):167–198.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Characterizing stylistic elements in syntactic structure. In *Proceedings of EMNLP*, pages 1522–1533. URL http://www.aclweb.org/anthology/D12-1139.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Walter Haeseryn. 1997. Achteropplaatsing van elementen in de zin. *Colloquium Neerlandicum*, 13:303–326.

John D. Hunter. 2007. Matplotlib: a 2D graphics environment. *Computing In Science & Engineering*, 9(3):90–95.

Geoffrey N. Leech and Michael H. Short. 1981. *Style in Fiction. A linguistic introduction to English fictional prose*. English Language Series 13. London / New York: Longman.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Geoff Nunberg. 1990. *The Linguistics of Punctuation*. volume 18 in CSLI Lecture Notes. CSLI, Stanford, California.

Richard Power, Donia Scott, and Nadjet Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 29(2):211–260.

Stephen Ramsay. 2003. Toward an algorithmic criticism. *Literary and Linguistic Computing*, 18(2):167–174.

Douglas LT Rohde. 2005. *TGrep2 User Manual version 1.15*. Massachusetts Institute of Technology. URL http://tedlab.mit.edu/dr/Tgrep2.

Antonio Roque. 2012. Towards a computational approach to literary text analysis. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 97–104. URL http://www.aclweb.org/anthology/W12-2514.

Michael Toolan. 2010. The intrinsic importance of sentence type and clause type to narrative effect: or, how Alice Munro's "Circle of Prayer" gets started. In *Language and style. In honour of Mick Short*, pages 311–327. Palgrave Macmillan, New York.

Andreas van Cranenburgh. 2012. Literary authorship attribution with phrase-structure fragments. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 59–63. URL http://www.aclweb.org/anthology/W12-2508.

Gertjan van Noord. 2006. At last parsing is now operational. In *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42.

Gertjan van Noord, Ineke Schuurman, and Gosse Bouma. 2011. *Lassy Syntactic Annotation Manual*. URL http://www.let.rug.nl/vannoord/Lassy/sa-man_lassy.pdf.

Juliette Wells. 2005. Mothers of chick lit? Women writers, readers, and literary history. In Suzanne Ferriss and Mallory Young, editors, *Chick lit: the new woman's fiction*, pages 45–70. Routledge, New York.