Lexical and syntactic complexity of narrative English texts for children versus for adults: computational and corpus-based approaches to a comparison[*]
[draft from 2017/01/14]

Carsten Breul
(University of Wuppertal)

1 Introduction

Studies on the syntactic and lexical characteristics of texts aimed at child readers of English beyond the first grade level are relatively rare.[1] And only a few of them make use of computational tools and corpus data for the purpose of describing and analysing more general tendencies in the linguistic properties of such texts as evidenced by a larger batch of them.[2] None, as far as I can see, compares English texts for children with texts for adults on the basis of corpora.

The growing concern with the learning and teaching to read English in German primary schools – to hint at the educational-societal context in which the investigation reported on here is embedded; see also footnote * – suggests that it may be helpful to enlarge our knowledge of the linguistic properties of texts for learners in the early years of their career as readers of English. The present paper aims to make a contribution in this respect by presenting some computational tools and their application in the analysis and comparison of two sections of a corpus, one containing 125 narrative texts aimed at adults, the other containing 125 narrative texts aimed at children, with each text having a length of 1000 to 1074 word tokens.[3] The computational tools and methods employed here are largely based on publications by Lu (2010, 2011, 2012, 2014), who, working within the context of second language (L2) acquisition research, devised and designed them and employed them in his own studies concerned with syntactic and lexical complexity.

The motivation for investigating syntactic and lexical complexity in texts, including texts for young readers, stems from insights gained and judgements expressed such as those by Mesmer & Cunningham & Hiebert (2012: 236):

> For us, the difficulty of a text or text feature always implies a dependent or criterion variable: the actual or predicted performance of multiple readers on a task based on that text or

[*] Laura Lindau and Miriam Lotz helped me with the compilation of the corpus of texts for children to be described below, for which I am very grateful. Thanks also to Miriam Lotz for proofreading the paper and identifying a number of errors. Thanks to Annette Becker, Bärbel Diehr, Stefanie Frisch and Laura Lindau for discussion. All remaining errors are mine. The work of Annette Becker and Laura Lindau in connection with this paper is funded by the project *Entwicklung von Unterrichtskonzepten zum Lesen lernen im Englischunterricht der Grundschule (EULE)* as part of the project *Kohärenz in der Lehrerbildung (KoLBi)* of the University of Wuppertal. KoLBi is funded by the German Federal Ministry of Education and Research within the common federal and state framework *Qualitätsoffensive Lehrerbildung*.

[1] See, among others, Anderson 1985, Fitzgerald et al. 2015, Gressenich 2011, Knowles & Malmkaer 1996, Lindgren 2003, Munat 2007, Plakans & Bilki 2016, Puurtinen 1998, Stamou 2012, Stephens 2005, Thompson & Sealey 2007, Wild & Kilgariff & Tugwell 2012.

[2] One reason for the scarcity of corpus-based linguistic research into the literature for children is probably the lack of publicly available corpora compiled from this literature. Researchers may gain partial access to the vast Oxford Children's Corpus (OCC) via Sketch Engine (see https://www.sketchengine.co.uk/oxford-childrens-corpus/ (last access 16/07/02); see also Wild & Kilgariff & Tugwell 2012. This, however, does not include access to the digital texts as such that make up the OCC. For some corpus-linguistic purposes, including those targeted in the present paper, direct access to texts that constitute a corpus is necessary.

[3] 'Word token is here to be understood as a character or character sequence that matches the regular expression \w+, i.e. an alphanumeric character or _, or a sequence therof.

> feature. In contrast, the complexity of a text or text feature, as we use the term, always implies independent or predictor variables: textual elements or factors that can be analyzed, studied, or manipulated. Treating the terms *text complexity* and *text difficulty* as synonymous conflates causes with effects.
>
> In reading education, publishers, practitioners, and researchers have long considered text difficulty to be an important issue. Readability research and various text-leveling schemes represent attempts to achieve valid and reliable means of assigning a probable difficulty to a text. […] However, it is text complexity that must be understood, and it is this understanding that will increase knowledge about the specific points of interaction among the characteristics of text, reader, and task. It is also an understanding of complexity that is needed when going beyond the question of difficulty to align specific text characteristics with reading curricula or instruction, a perennial concern in the early grades.[4]

The specific issue the present investigation is concerned with is this: The great majority of beginning readers in English have so far been young people growing up in an English speaking environment, that is, individuals whose syntactic competence is already at or near the English native adult level at the time when they begin to learn to read (L1 learners/readers). However, learners of English as an L2, such as almost all learners of English at German primary schools, do not begin their careers as readers of English texts with an already established (near) adult-native syntactic competence in place (L2 learners/readers). Arguably, this points to different requirements for texts one of whose aims or attributed functions it is to help young L1 readers on the one hand and young L2 readers on the other hand to acquire the reading competence of an educated adult. The primary concern of the former may be argued to lie with the lexical complexity of texts rather than their syntactic complexity. Such a grading of concerns may be argued to make much less sense for the latter; for them the lexical and syntactic complexity of texts may be issues of equal concern.[5] Thus, the assumption that considerations of both lexical and syntactic complexity are, in general, equally relevant for the selection of texts that are appropriate at a given stage of learning to read by L2 learners raises the question of whether the genre of literature for children as such provides for this requirement. Is the genre of literature for children distinguishable in terms of lexical and/or syntactic complexity from the genre of literature for adults – and if so, to what extent? The specific purpose of the present report is to provide empirically challengeable answers on the basis of the small corpus already mentioned above and on the basis of a selection of complexity indices. It ought to be stated clearly at the outset, though, that the corpus consists of text samples of about 1,000 word tokens and that it is one characteristic of the (narrative) literature for children that a large part of it consists of texts that are shorter than 1,000 words. Thus, the corpus section that contains the texts for children can at best be claimed to be representative of the segment of the genre of children's literature that is made up of texts that are at least about 1,000 word tokens long.

It ought to be pointed out that there is an impressive body of research based on an impressive computational tool which occasionally addresses, and arguably has the potential to address to a larger extent, issues related to those in focus in the present paper. This is the Coh-Metrix system and part of the research associated with it (see http://cohmetrix.com/, last access 16/03/03; see

---

[4] The paper by Mesmer & Cunningham & Hiebert (2012) provides, beside a very helpful overview of the field of text and reading research with a focus on beginning readers, a rich mine of directions and questions for research where expertise in certain domains of linguistics is key.

[5] One of the questions Mesmer & Cunningham & Hiebert (2012: 243) raise is "*What is the contribution of specific syntactic features and patterns to sentence difficulty?*". Their comment accompanying this question goes like this: "For both English-first students and ELs [i.e. 'English learners', i.e. learners of English], there is a general lack of research on the difficulty of written sentences with particular syntactic characteristics. The results of such studies could improve text authoring and selection".

McNamara & Graesser & McCarthy & Cai 2014, and the references given in these sources).[6] An assessment of the insights provided by this body of research with respect to its relevance for the learning and teaching to read English in German primary schools is a task that ought to be tackled in the near future; and the same holds for the potential to generate further insights in these respects by Coh-Metrix. However, what is clear is that Coh-Metrix at the present stage of its development, for all its sophistication and potential, does not afford the researcher-user the flexibility to computationally investigate linguistic patterns of their own choice. The user of Coh-Metrix will be informed about the score on a wide range of linguistic and textual indices for a text or corpus of texts to be uploaded by them (see McNamara & Graesser & McCarthy & Cai 2014: 69-72, 85-95 and the "Documentation" at http://cohmetrix.com/). But they are in no position to choose other measures than those provided by the system. In this respect, the methodology to be described in the present paper is different in principle from what can be done with Coh-Metrix. On the other hand, this methodology by no means affords those who apply it the high degree of user-friendliness that Coh-Metrix does.

2 The corpus sections

Almost all of the texts that constitute the two corpus sections underlying the research reported on here are text extracts in the sense that they have been randomly extracted from their larger sources so as to comprise 1,000 orthographic word tokens plus the number of word tokens that complete the sentence where the 1,000 word limit cuts through it. The sources of the texts for the adult section have been taken from the FLOB and FROWN corpora;[7] the sources of the texts for the child section have been compiled from texts in the British National Corpus (BNC),[8] from e-books that do not prevent access to their digital content, from OCR-scanned digital versions of printed books and from freely available sources on the internet. I call the adult section FLOWN (to be pronounced /flaʊn/) and the child section SCETCH, acronymising *Small Corpus of English Texts for Children*. Bibliographical and some basic statistical information on the texts that constitute FLOWN and SCETCH can be accessed either in or from an MS Excel sheet available at http://leute.uni-wuppertal.de/~breul/Homepage/flown-info.xlsx and http://leute.uni-wuppertal.de/~breul/Homepage/scetch-info.xlsx.

The selection of texts for FLOWN and SCETCH was largely opportunistic. But at least in the case of FLOWN, the availability of sources from the FLOB and FROWN corpora happened to afford what I was aiming at, namely a small-size corpus that can nevertheless claim to represent narrative writing in present-day British and American English across a range of narrative genres as reflected by the text categories K ("General Fiction"), L ("Mystery and Detective Fiction"), M ("Science Fiction"), N ("Adventure and Western") and P ("Romance and Love Story") of FLOB and FROWN. In the case of SCETCH, the selection of texts was guided by the criteria that their sources had to be advertised as literature for children, to the exclusion of juvenile literature, and that their sources had to be at least 1,000 word tokens long. In one case a text was allowed into SCETCH after its source had been made longer than 1,000 word tokens by combining two children's books by the same author with a similar content and obviously the same target group (Donaldson, *The Gruffalo* and *The Gruffalo's Child*). In one other case (Baum, *Bad Bat Stories*

---

[6] "Assigning texts to students in school" is mentioned as one of the prominent tasks the developers of Coh-Metrix have in mind for applications of the system (McNamara & Graesser & McCarthy & Cai 2014: 9).

[7] Short introductions to FLOB and FROWN with links to the manuals accompanying them are provided at the following internet addresses (last access 16/03/21): http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/; http://www.helsinki.fi/varieng/CoRD/corpora/FROWN/.

[8] For information about the BNC see http://www.natcorp.ox.ac.uk/ (last access 16/03/21).

for Early Readers), several similar very brief stories that are individually available on the internet were combined in order to provide a source longer than 1,000 word tokens. And in several other cases a collection of short stories for children constitutes the source. As pointed out above, the 1,000+ text extracts for the two corpus sections were randomly extracted from their sources. Chapter numbers and titles contained in these extracts were retained. Textual material such as front and back matter and notes on the author were eliminated from the sources. All of the texts in the two corpus sections had to be modified from the way they exist in their digital or digitalised sources in additional ways. These are the following ones.

The BNC version available to me was the XML edition as described online at http://www.natcorp.ox.ac.uk/docs/URG/ (last access 16/03/03). The BNC-documents had to be cleared of their XML codes and of the textual material added by the corpus compilers in the header of each document. FLOB and FROWN texts were available to me from the second edition (1999) of the ICAME CD (see http://clu.uni.no/icame/newcd.htm, last access 16/03/03). Apart from the stripping away of mark-up codes and of the textual material added by the corpus compilers (editorial comments), several decisions had to be taken concerning the treatment of what are called "unusable characters" (i.e. non-ASCII characters), in the manuals accompanying the corpora, of passages or expressions in languages other than English, including transliterations from Greek and misspellings. If a text was considered to be too heavily burdened by such characteristics, it was excluded from consideration for FLOWN. Passages in texts that did show such characteristics to a lesser extent were not discarded and modified in the following ways: Sentences containing foreign expressions that are not commonly used in English and whole sentences in a language other than English were deleted; misspellings were corrected; the position for a character that was coded as 'unusable' was replaced by an ASCII-character (e.g. *naïve* > *naive*, *señor* > *senor*).

Further modifications of the texts were carried out in order to standardise their format in view of the syntactic parsing procedure to be described below and, possibly, to improve the results of parsing. A period was inserted after chapter headings and chapter numbers in all texts so as to prevent the parser from 'assuming' that they and what follows them belong to the same sentence. All kinds of quotation marks were replaced by a pair of single quotation marks. The ellipsis marker (...) was substituted by a period. Paragraph breaks were deleted. Variations in the punctuation of the end of reported speech or thought between cases exemplified by *'You're ill' said the doctor. / 'You're ill,' said the doctor. / 'You're ill', said the doctor. / 'You're ill.' said the doctor.* were all standardised to the pattern exemplified by *'You're ill.' said the doctor.* This will have lead the parser to a false analysis in some cases (if the absence of a period actually indicates that the reported speech or thought is to be continued and if the continuation would have to be syntactically integrated with the stretch that it continues). But these cases were judged to be much less frequent than those where a closing quotation mark not preceded by a period ends a stretch of discourse that is marked by a period in many other text passages, either of the same text or of other texts in the corpus. In other words, this modification was carried out in order to induce the parser to treat all ends of reported speech or thought marked by a closing quotation mark as the end of a unit to be independently parsed as a 'sentence' (ROOT unit). This invariably causes the parser to consider the syntactic environment of the reported speech or thought to belong to a different sentence (ROOT unit) than the reported speech or thought, which may not generally be accepted as tolerable for syntactic analysis. However, this potential disadvantage was considered to be more than compensated for by the advantage of having the parser treat many cases uniformly, cases in which the employment of a comma or a period varies only according to orthographic usage, with the variation being syntactically unmotivated.[9]

[9] That is, the cases in (a) below are analysed as *one* ROOT unit by the parser while (b) is analysed into *two*. From a syntactic perspective independently of computational parsing, an analysis into one ROOT unit may be preferable.

3 The computational tools and methods employed

As mentioned earlier, the computational tools and methods employed for the research reported on here are based on work by Lu (2010, 2011, 2012, 2014). He discusses corpus and computational linguistic aspects of syntactic complexity and of lexical complexity, also called lexical richness, in English. The syntactic complexity indices that he focuses on are measures of the length of certain syntactic units as well as the frequency of several syntactic constructions. On the side of indices for lexical richness, he considers a range of measures for lexical density, lexical sophistication and lexical variation, conceived of as dimensions of lexical richness. With respect to syntactic complexity, Lu's (2010, 2011) aim is to investigate to what extent syntactic complexity in the writings of L2 learners of English and as measured by these indices correlates with and is predictive of the stage in the L2 acquisition process the learners are in. With respect to lexical richness, he (2012) studies to what extent the scores achieved by L2 learners of English in oral narratives on the three dimensions of lexical richness correlate with expert raters' quality judgments of their oral performance.[10]

The first step in my study of syntactic complexity and lexical richness in SCETCH as compared to FLOWN consisted in extracting the scores for syntactic complexity and lexical richness measures for each of the corpus texts. These scores provide the data that further analyses are based on.

3.1 Lexical richness indices

As far as lexical richness is concerned, the computational tool devised by Lu (2012, 2014), the Lexical Complexity Analyzer (LCA; see Lu 2014: 85; see also http://www.personal.psu.edu/xxl13/downloads/lca.html, last access 16/03/03), allows for the extraction from digital texts of 25 indices that have been proposed in the literature, especially the L2 acquisition literature, as providing measures for three dimensions of lexical complexity or richness of texts, namely lexical density, lexical variation and lexical sophistication. Lexical density "refers to the ratio of the number of lexical (as opposed to grammatical) words to the total number of words in a text" (Lu 2012: 191). While lexical density is represented by only one specific measure in the LCA, lexical sophistication is represented by five measure variants. What they are each supposed to capture in conceptually and partly mathematically different ways is "the degree to which advanced or sophisticated words are used in a text" (Lu 2014: 84). For the LCA, the notion of sophisticated words is defined as those words "that are not among the first 2,000 most frequent lemmas found in either the British National Corpus (for texts with British spelling) or the American National Corpus (for texts with American spelling)" (Lu 2014: 85). There are 19 measure variants for lexical variation incorporated in the LCA, each supposed to provide information about "the range of different lexical items used in a text" (Lu 2014: 82). Four of them are variants of the simple concept of the number of different word tokens in a text, and 15 are variants of type-token ratio (TTR). Some of these variants differ from basic type-token ratio (number of different word types divided by the number of different word tokens) by restricting the counting and computation to a specific word class (verbs, nouns, adjectives, ad-

---

However, the point for the purposes in the present research is that both orthographic usages are common in the sources for FLOWN and SCETCH, but would be treated differently by the parser while they should be treated uniformly.

(a) 'You're ill' said the doctor. / 'You're ill,' said the doctor. / 'You're ill', said the doctor.

(b) 'You're ill.' said the doctor.

[10] Vajjala & Meurers (2012), who point out that Lu's work is "independent of the readability research" (166), conclude that taking his lexical richness and syntactic complexity indices into account is useful for readability classification.

verbs). Other variants are devised so as to cope with the well-known problem that the basic type-token ratio computed for different text samples from a text tends to decrease when sample length increases. One of the results of Lu's (2012) own research into the relationship between these measures for lexical richness and the quality of ESL learners' oral narratives as judged by expert raters is this:

> the three dimensions of lexical richness posited in this and previous research did not correlate strongly with each other, which suggests that they are indeed different constructs. Of these dimensions, lexical variation correlated most strongly with the raters' judgments of the quality of ESL learners' oral narratives. No effect for lexical density emerged, and a very small effect was found for lexical sophistication. (Lu 2012: 203)

For the work described in the present paper, I applied Lu's LCA to the texts in FLOWN and SCETCH. This presupposes previous POS-tagging and subsequent lemmatisation of the texts. The tagging was done by using the Stanford POS Tagger (Toutanova et al. 2003; see also http://nlp.stanford.edu/software/tagger.shtml, last access 16/03/03), the lemmatisation by using Morpha (Minnen & Carroll & Pearce 2001; see also http://users.sussex.ac.uk/~johnca/morph.html, last access 16/03/03; see also Lu 2014: 72f., 85f.).

3.2 Syntactic complexity indices

For computationally measuring the syntactic complexity of texts, Lu (2010, 2011, 2014) developed software called L2 Syntactic Complexity Analyzer (L2SCA; here abbreviated as SCA). SCA embeds the Stanford Parser (Klein & Manning 2003; see also http://nlp.stanford.edu/software/lex-parser.shtml, last access 16/03/03) for syntactically parsing the sentences of a text and Stanford Tregex (Levy & Andrew 2006; see also http://nlp.stanford.edu/software/tregex.shtml, last access 16/03/03) for identifying patterns in the tree structure output produced by the parser. Tregex is used to identify the occurrences of those configurations in the parse trees that are required for the computation of the complexity indices and to count their occurrences in individual texts. Statistics software is then employed to perform statistic computations considered to be appropriate for answering his research questions about the relation between the complexity indices obtained for the texts produced by his L2 subjects and the developmental L2 acquisition stage they are in.[11]

The 14 syntactic complexity indices that Lu considers are among those that have been argued in the L2 acquisition literature to correlate at least weakly with L2 proficiency. They can be divided into two groups. First, there are three syntactic length indices: mean length of clause (MLC), mean length of sentence (MLS) and mean length of T-unit (MLT), with MLC defined as "number of words divided by number of clauses", MLS defined as "number of words divided by number of sentences" and MLT defined as "number of words divided by number of T-units" (Lu 2011: 43).[12] Then, there is the group of 11 constructional indices (definitions quoted from Lu 2011: 43):

(1) a. clauses per sentence (C/S), i.e. "number of clauses divided by number of sentences";
b. clauses per T-unit (C/T), i.e. "number of clauses divided by number of T-units";
c. complex T-units per T-unit (CT/T), i.e. "number of complex T-units divided by

[11] Developmental stage is operationalised by Lu as the school level the respective L2 learner is at. "Following Wolfe-Quintero (1998), I assumed that if a measure progresses linearly in a way that is significantly related to school level, it is potentially a good candidate for a developmental index" (Lu 2011: 45).

[12] A T-unit contains "one main clause with all the subordinate clauses attached to it. The number of subordinate clauses can, of course, be none" (Hunt 1965: 20). That is, two paratactically combined sentences constitute two T-units, three paratactically combined sentences constitute three T-units, and so on.

number of T-units";[13]
d. dependent clauses per clause (DC/C), i.e. "number of dependent clauses divided by number of clauses";
e. dependent clauses per T-unit (DC/T), i.e. "number of dependent clauses divided by number of T-units";
f. coordinate phrases per clause (CP/C), i.e. "number of coordinate phrases divided by number of clauses";
g. coordinate phrases per T-unit (CP/T), i.e. "number of coordinate phrases divided by number of T-units";
h. T-units per sentence (T/S), i.e. "number of T-units divided by number of sentences";
i. complex nominals per clause (CN/C), i.e. "number of complex nominals divided by number of clauses";
j. complex nominals per T-unit (CN/T), i.e. "number of complex nominals divided by number of T-units";
k. verb phrases per T-unit (VP/T), i.e. "number of verb phrases divided by number of T-units".

For my study of syntactic complexity in FLOWN and SCETCH, I did not directly make use of SCA, which is written in the programming language Python, but emulated it by a script written in the programming language Perl. Not being sufficiently familiar with Python, but intending to make use of the Stanford Parser in combination with Stanford Tregex independently of Lu's 14 indices, I decided to write EmuSCA, a Perl script that I can easily use or adapt for the capture of data for additional syntactic indices. Just like SCA, EmuSCA integrates calls of the Stanford Parser for syntactic parsing and of the Tregex software for the identification and counting of syntactic patterns in the output of the parsing procedure. The syntactic patterns read in by EmuSCA and processed by Tregex are taken over from a Python script (analyzeText.py) that is part of Lu's SCA. The Perl instructions for the computation of the indices on the basis of word and pattern counts in the parser output were likewise adopted and adapted from Lu's analyzeText.py script. In sum, running EmuSCA over the FLOWN and SCETCH texts results in a tabular output of each of Lu's length and syntactic complexity indices for each of the texts, rounded to the third decimal place.

3.3 Some more details on parsing and syntactic pattern identification

As just pointed out, the syntactic parsing of the FLOWN and SCETCH texts was carried out by the Stanford Parser, more specifically, version 3.5.2 of the Stanford Parser and the parser model that works on the basis of an unlexicalised probabilistic context-free grammar (PCFG) for English (called englishPCFG.ser.gz in the Stanford Parser software package).[14] The parser was instructed to ignore sentences longer than 100 word tokens. The parsing procedure resulted in separate files containing labeled phrase structure trees for each of the parsed expressions in each of the texts in FLOWN and SCETCH. These trees files were then subjected to Tregex, which identifies and counts patterns of syntactic nodes in labeled phrase structure trees.

In order to provide an idea of the functioning of Tregex, I pick 4 of Lu's 14 indices for illustration: C/S, DC/C, CP/C, CN/C. The computation of these indices requires the identification and

[13] "A complex T-unit is then one containing at least one dependent clause (or subordinate clause in Hunt's term […])" (Lu 2011: 44).
[14] "For English, although the grammars and parsing methods differ, the average quality of englishPCFG.ser.gz and englishFactored.ser.gz is similar, and so many people opt for the faster englishPCFG.ser.gz, though englishFactored.ser.gz sometimes does better because it does include lexicalization" (http://nlp.stanford.edu/software/parser-faq.shtml#y; last access 2016/02/02).

counting of sentences, dependent and independent clauses, coordinated phrases and complex nominals. For this task, the following expressions processable by Tregex are relevant:

(2) a. ROOT
b. S|SINV|SQ < (VP <# MD|VBP|VBZ|VBD)
c. SBAR < (S|SINV|SQ < (VP <# MD|VBP|VBZ|VBD))
d. ADJP|ADVP|NP|VP < CC
e. NP !> NP [<< JJ|POS|PP|S|VBG| << (NP $++ NP !$+ CC)]
f. SBAR [<# WHNP | <# (IN < That|that|For|for) | <, S] & [$+ VP | > VP]
g. S < (VP <# VBG|TO) $+ VP
h. FRAG > ROOT !<< (S|SINV|SQ < (VP <# MD|VBP|VBZ|VBD))

These expressions represent search patterns for identifying labeled syntactic trees or parts of trees as generated by the parser. Before explaining the symbols occurring in these search expressions I will briefly indicate what syntactic structures these expressions are supposed to identify.

→(2)a: maximal expression that is independently parsed; mostly orthographic sentence, including orthographic sentence that contains coordinated (non-orthographic) sentences → the 'sentences' (S) of (1) above.
→(2)b: minimal expression that may stand on its own as a declarative, interrogative, or imperative sentence; i.e. finite clause → some of the 'clauses' (C) of (1).
→(2)c: embedded finite clause → the 'dependent clauses' (DC) of (1).
→(2)d: coordinated adjective phrase, adverb phrase, noun phrase, or verb phrase → the 'coordinate phrases' (CP) of (1).
→(2)e: complex nominal phrase where the top node dominates at least one adjective, possessive marker, prepositional phrase, sentence or *-ing* verb form respectively, or another noun phrase → some of the 'complex nominals' (CN) of (1).
→(2)f: relative clause as part of complex nominal expression; noun clause → some of the 'complex nominals' (CN) of (1).
→(2)g: *-ing* or *to*-infinitive clause as subject or object → some of the 'complex nominals' (CN) of (1).
→(2)h: fragment clause → some of the 'clauses' (C) of (1).

The number of occurrences of (2)a in a parsed text is (the frequency of) S. The number of occurrences of (2)b plus that of (2)h is (the frequency of) C. The number of occurrences of (2)c is (the frequency of) DC. The number of occurrences of (2)d is (the frequency of) CP. The sum of the number of occurrences of (2)e-g is (the frequency of) CN. C/S is C divided by S, and analogously for DC/C, CP/C, CN/C.

In the Tregex pattern expressions in (2), capital letters or sequences of capital letters match labels of nodes of syntactic trees generated by the parser. The expressions *That*, *that*, *For* and *for* in (2)f represent terminal nodes in trees and match the corresponding English lexical items. All other symbols or sequences of symbols belong to the Tregex query language for encoding patterns of relations in syntactic trees. *ROOT* is always provided by the parser as the starting label for each expression that it attempts to assign an independent tree. The following explanations of the other labels mentioned in (2) are quoted from Lu 2014: 44, 98:

Clause level:

S	Simple declarative clause, imperative, infinitive
SBAR	Relative clause, subordinate clause, including indirect question
SINV	Inverted declarative sentence
SQ	Inverted yes/no question, or main clause of a wh-question in SBARQ[15]

[15] SBARQ: "Direct question introduced by a wh-word or wh-phrase" (Lu 2014: 98).

Phrase level:

ADJP	Adjective phrase
ADVP	Adverb phrase
FRAG	Fragment
NP	Noun phrase
PP	Prepositional phrase
VP	Verb phrase
WHNP	Wh-noun phrase

Lexical category level:

CC	Coordinating conjunction
IN	Preposition or subordinating conjunction
JJ	Adjective
MD	Modal verb
POS	Possessive ending (*'s*)
TO	*to*
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBP	Verb, non-third person singular present
VBZ	Verb, third person singular present

The following explanations concerning the symbols for relations in trees are quoted from the file README-tregex.txt that is part of the Tregex package.

A < B	A immediately dominates B
A > B	A is immediately dominated by B
A << B	A dominates B
A <, B	B is the first child of A
A $+ B	A is the immediate left sister of B
A $++ B	A is a left sister of B
A <# B	B is the immediate head of phrase A

The function of a pair of opening and closing round brackets can be most easily explained in the following way: *A < (B < C)* means 'A immediately dominates B, and B immediately dominates C'; by contrast, *A < B < C* means 'A immediately dominates B and C'. Similarly, *A >(B > C)* means 'A is immediately dominated by B, and B is immediately dominated by C'. The exclamation mark '!' negates a relation. That is, *NP !>> NP* captures an NP that is not dominated by an NP, and *NP !$+ CC* captures an NP that is not an immediate left sister of a CC. The symbols '&' and '|' connect relations according to logical conjunction ('and') and disjunction ('or'). That is, *A << B | << C*, for instance, means 'an A that dominates a B or a C; *A < B & > C*, for instance, means 'an A that immediately dominates a B and is immediately dominated by a C'. The '|'-symbol can also be written in between strings designating nodes so that (2)d above means 'an ADJP or an ADVP or an NP or a VP that immediately dominates a CC'. A pair of opening and closing angled brackets serves to group relations. For instance, *A [< B | < C] < D* means 'an A that immediately dominates a D while also immediately dominating a B or a C'. By contrast, *A < B | < C < D* means 'an A that immediately dominates a B, or an A that immediately dominates a C and a D.

As far as I can see, the Tregex patterns designed by Lu do what they are supposed to and have thus been taken over verbatim from Lu's analyzeText.py into EmuSCA. One problematic component of the work described in the present paper is the accuracy of the parsing. Following Lu (2010, 2011, 2014), the Stanford Parser was used, which is popular due to its good balance between accuracy and speed. One additional reason for using it is the fact that it delivers grammatical dependency analyses alongside constituent structure analyses, which I intend to use in re-

search that extends the present one. Nevertheless it must be granted that it is unclear what the effect is of the errors commonly used parsers produce on the results of studies such as the present one.[16]

4 Statistical observations and analyses for FLOWN and SCETCH

The guiding questions in this section on statistic properties and tendencies in SCETCH in comparison to FLOWN as manifested by the data for syntactic complexity and lexical richness indices are these: 1) To what extent do the two corpora statistically differ with respect to their indices data? 2) How well can texts for children be statistically distinguished from texts for adults on the basis of their indices data gathered from the two corpus sections? I will approach these questions in a stepwise manner, first using basic statistical concepts and techniques, and then applying a more elaborate statistical method, Linear Discriminant Analysis. The software employed for carrying out the statistical analyses is the environment known as R (see https://www.r-project.org/about.html; last access 16/07/05) in connection with R Studio (https://www.rstudio.com/; last access 16/07/05). Use is also made of MS Excel for making data available in the form of downloadable spreadsheets.

The syntactic complexity indices for FLOWN and SCETCH gained from applying EmuSCA to the two corpus sections are provided in two MS Excel sheets retrievable from http://leute.uni-wuppertal.de/~breul/Homepage/flown-sca.xlsx and http://leute.uni-wuppertal.de/~breul/Homepage/scetch-sca.xlsx. The lexical richness indices that I will be using are the means of the indices obtained by running Lu's LCA over FLOWN and SCETCH with the help of the lemma list generated from the British National Corpus (British English) on the one hand and the lemma list generated from the American National Corpus (American English) on the other hand. These averaged indices data can be retrieved from http://leute.uni-wuppertal.de/~breul/Homepage/flown-lca-averaged.xlsx and http://leute.uni-wuppertal.de/~breul/Homepage/scetch-lca-averaged.xlsx.

4.1 Statistical observations on lexical richness

I will concentrate on LD as the index for lexical density, two indices for lexical sophistication, LS1 and LS2, and two indices for lexical variation, CTTR and LV. These indices are defined and computed in the following way by LCA:

(3) Lexical density (LD): "the ratio of the number of lexical words […] to the number of words" (Lu 2012: 192); the Python scripts that constitute LCA reveal that the sense in which the term *words* is used here is that of 'word tokens'.
Lexical sophistication 1 (LS1): "the ratio of the number of sophisticated lexical words […] to the number of lexical words" (Lu 2012: 192); again, *words* in the sense of 'word tokens' is meant.
Lexical sophistication 2 (LS2): "the ratio of the number of sophisticated word types […] to the total number of word types" (Lu 2012: 192).

[16] To give just one indication of the level of (in)accuracy of parsers, I may refer to the part of a study by Hempelmann & Rus & Graesser & McNamara (2006: 140) where they report the results of a test based on the following consideration: "A [parsed] sentence was considered usable for further processing [by Coh-Metrix 2.0] if the parse had no error or only one error of the least problematic type one". On this criterion, the Stanford parser produced an average output of about 83% usable parses across seven narrative and expository texts, following in rank the best one of the four parsers tested, the Charniak parser, which produced an output of about 89% usable parses. On the accuracy and performance of parsers including the Stanford parser see also Klein & Manning 2003.

Corrected type-token ratio (CTTR): the ratio of the number of word types to the square root of the number of word tokens multiplied by 2 ($T/\sqrt{2N}$) (see Lu 2012: 193-195).
Lexical word variation (LV): the ratio of the number of lexical word types to the number of lexical word tokens (see Lu 2012: 195).

As pointed out above, since the determination of lemmas and sophisticated words based on the BNC and ANC word lists may result in different values for the indices, the mean of the BNC-based and ANC-based values are taken for the statistical observations in this section.

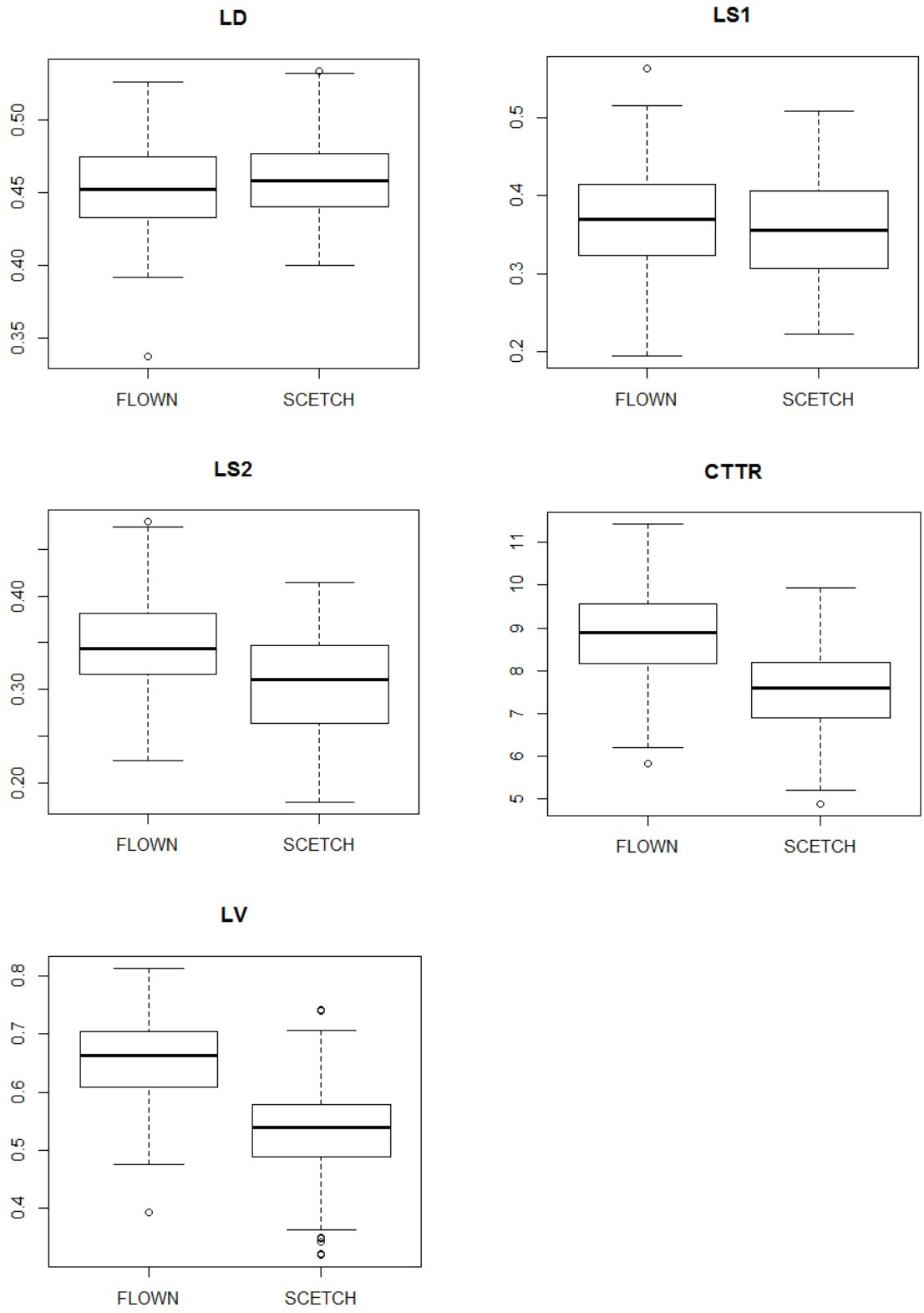
Performing bootstrapped independent samples *t*-tests on the FLOWN and SCETCH data for each of the five lexical richness indices suggests the following:[17] There is no statistical difference between the two groups for LD (lexical density); by contrast, the two groups do differ statistically with respect to the other four indices LS1 and LS2 (lexical sophistication) and CTTR and LV (lexical variation).[18] These results confirm the impressions that we get from boxplots for the comparison between FLOWN and SCETCH for at least four of the five indices; see (4) below. In a boxplot, the values for the indices are located along the *y*-axis of the graph. The boldfaced line in a box represents the value for the median of the observed data. The lower boundary of a box represents the value of the first quartile, the upper boundary represents the value of the third quartile. A box thus comprises the middle 50% of the data. In the version of boxplots used here, where outliers are represented by small circles, the lower end of the so-called whiskers of a box represents the smallest value that is not more than 1.5 interquartile ranges lower than the first quartile;[19] the upper end of the whiskers represents the largest value that is not more than 1.5 interquartile ranges higher than the third quartile.

[17] The tests were performed by using yuenbt() from the R WRS package with a trim level of 0.0 and a number of bootstrap samples of 10,000; see Larson-Hall 2016: 296f. See Plonsky 2015 and Larson-Hall 2016: 73ff., passim for arguments that robust tests such as bootstrapped tests are generally preferable over parametric and classical non-parametric tests.

[18] LD: mean difference: -0.005; 95% confidence interval: [-0.013, 0.002];
LS1: mean difference: 0.019; 95% confidence interval: [0.023, 0.034];
LS2: mean difference: 0.043; 95% confidence interval: [0.029, 0.057];
CTTR: mean difference: 1.326 95% confidence interval: [1.073, 1.575];
LV: mean difference: 0.125; 95% confidence interval: [0.105, 0.144].
The difference between the groups (FLOWN, SCETCH) counts as statistically significant if the confidence interval does not span 0.

[19] The interquartile range is the difference between the first and the third quartiles.

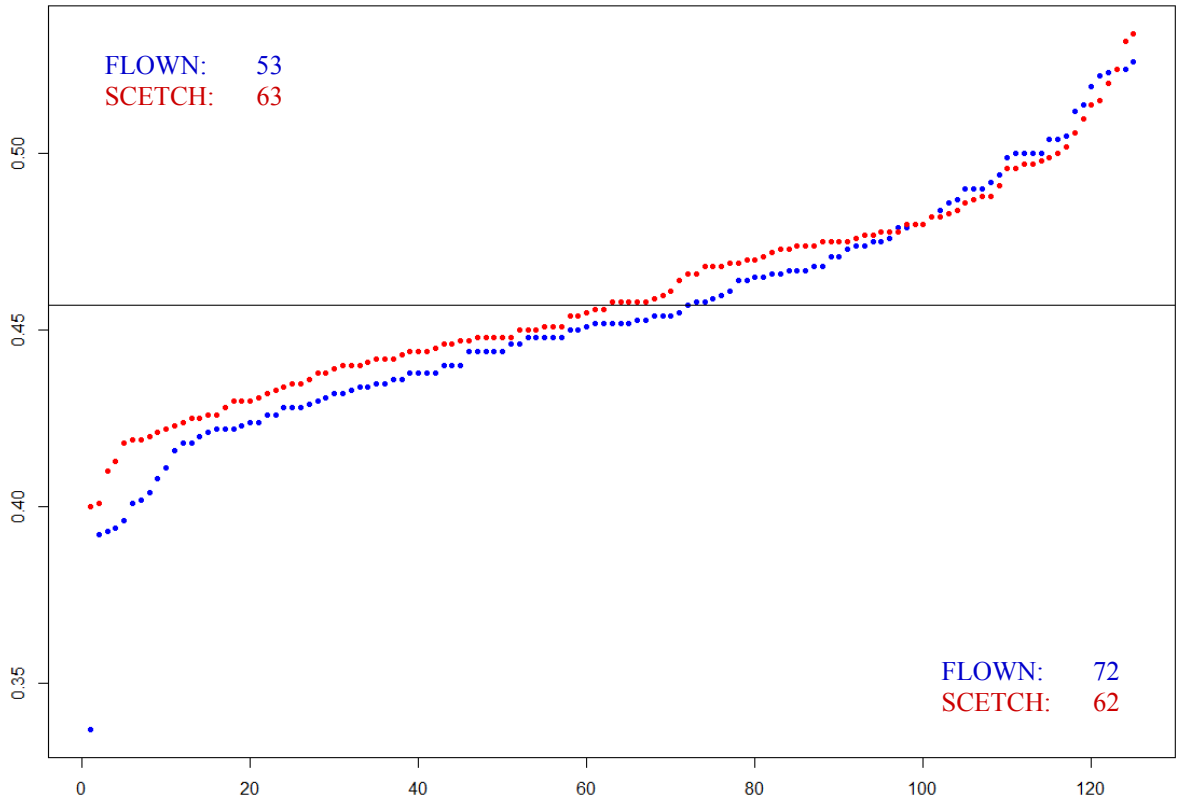
(4) Boxplots for five lexical richness indices for FLOWN vs. SCETCH



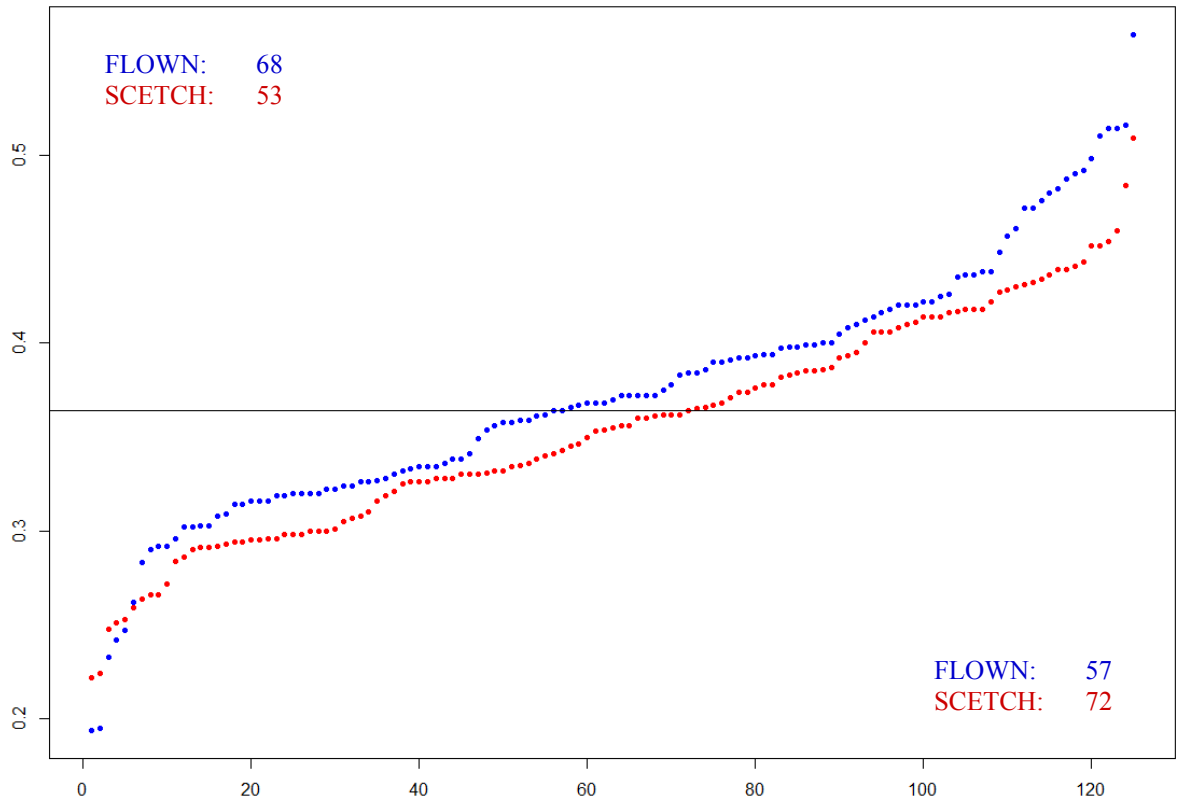
The boxplots for LD are very similar, visualising the lack of a significant statistical difference between FLOWN and SCETCH. By contrast, the boxplots are very different for LS2, CTTR and LV, visualising clearly the presence of a significant statistical difference. A judgement of the degree of similarity or difference between the two boxplots for LS1 is less clear. That is, the presence of a significant statistical difference for LS1 is less clearly visualised in this case.

The five diagrams in (5) below provide scatterplots in which the scores for the five lexical richness indices for each text in FLOWN and SCETCH are represented by a dot (blue for FLOWN, red for SCETCH) with the value of a dot on the *y*-axis corresponding to the index score. The dots have been sorted in increasing order. The black horizontal line marks the mean score for the respective index for all the texts in the corpus. The number of texts above the mean are provided in the upper left hand corner, and the number of texts below the mean are provided in the lower right hand corner.

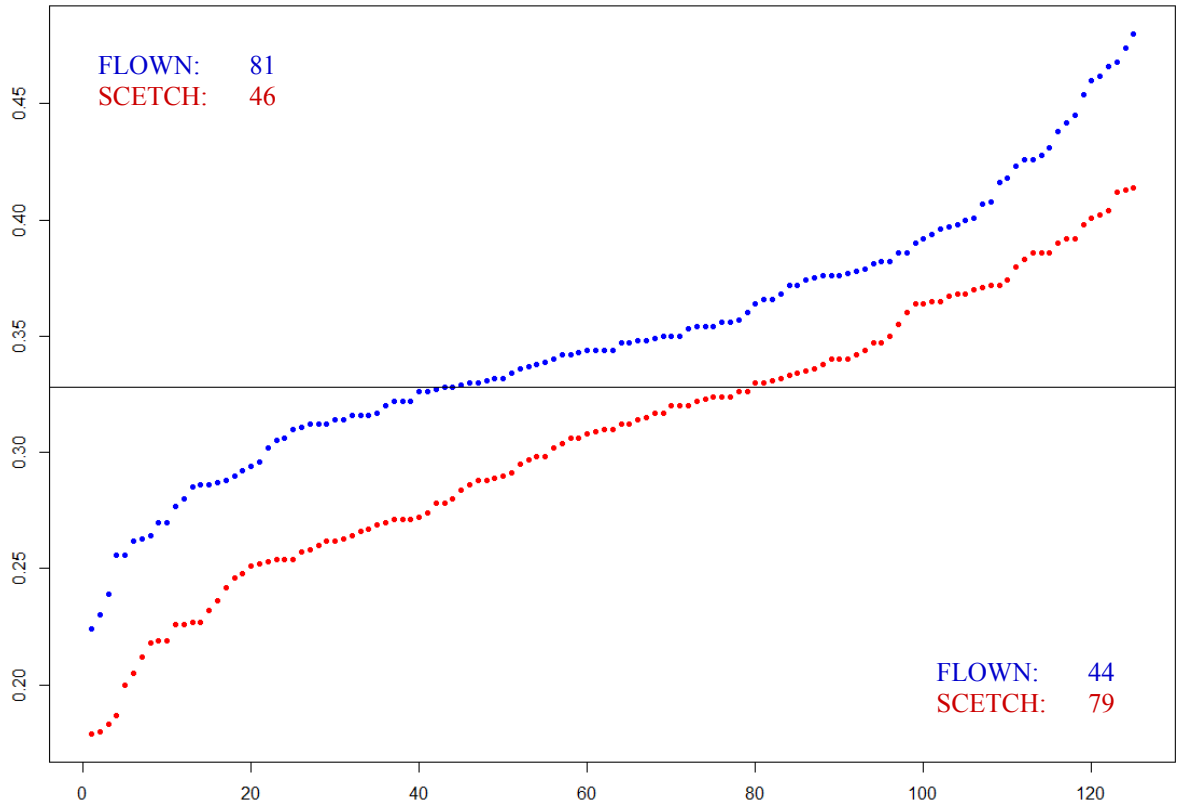
(5) a. LD plot: FLOWN vs. SCETCH



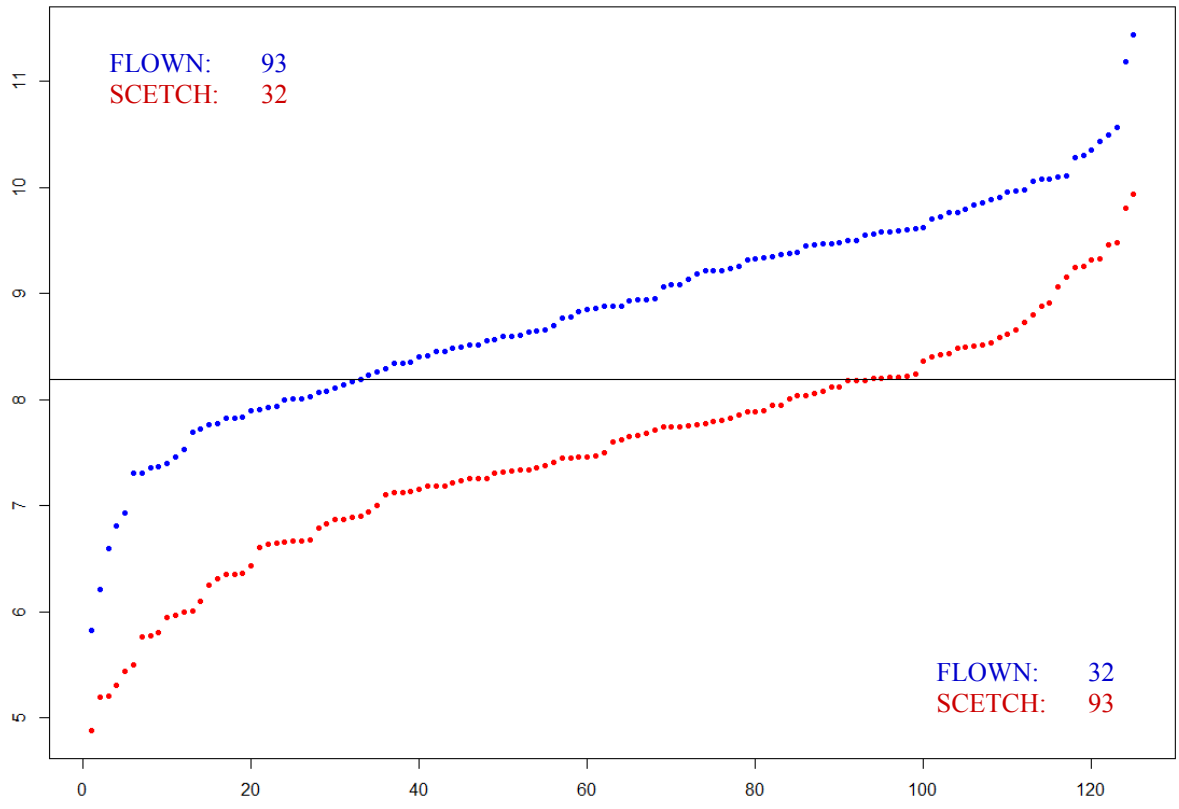
b. LS1 plot: FLOWN vs. SCETCH



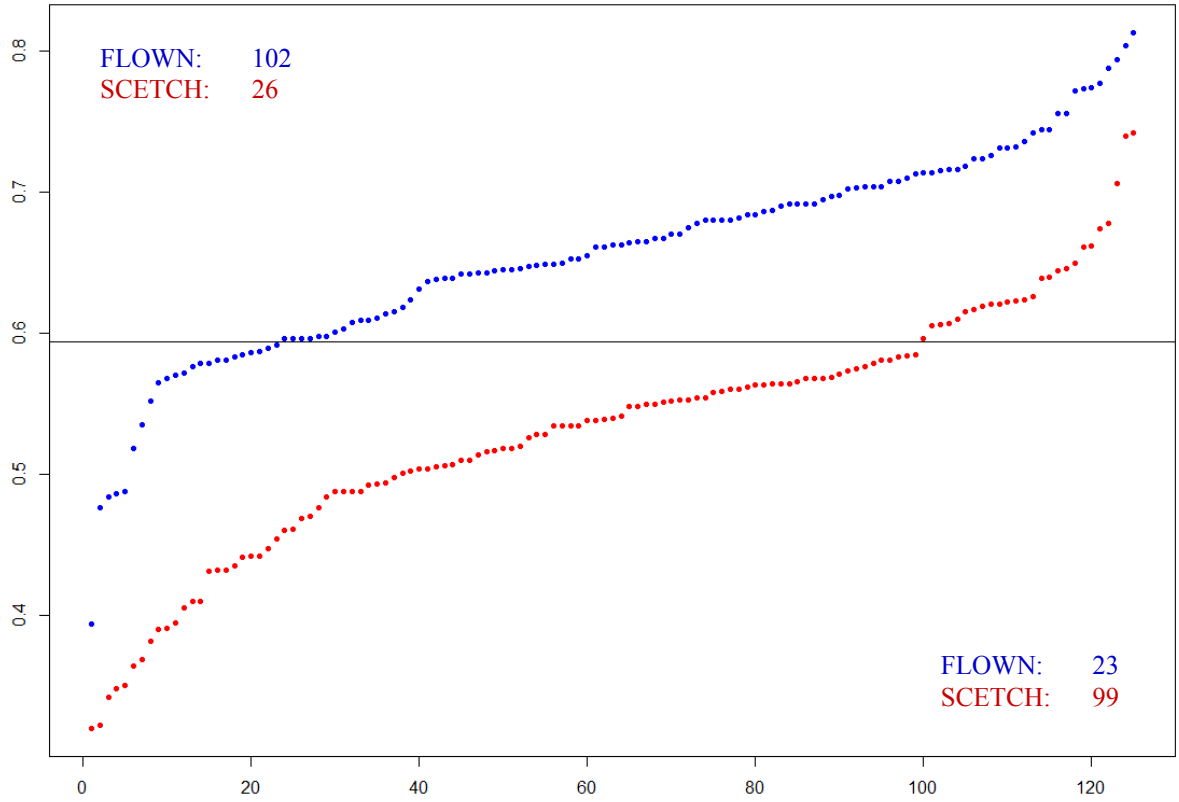
c. LS2 plot: FLOWN vs. SCETCH



d. CTTR plot: FLOWN vs. SCETCH



e. LV plot: FLOWN vs. SCETCH



The plots show that FLOWN and SCETCH differ much more clearly with respect to the second lexical sophistication index (LS2) and the two lexical variation indices CTTR and LV than with respect to the lexical density index (LD) and first lexical sophistication index (LS1). As also discernible from the boxplots in (4), there appears to be a slight tendency for more SCETCH texts than FLOWN texts to achieve a higher score for LD. As already pointed out earlier, this somewhat surprising result is however not statistically significant. The difference between FLOWN and SCETCH with respect to LS1 is, though statistically significant, not that big. Not many more than half of the 125 FLOWN texts, namely 68, score above the mean for LS1, while not many less than half of the 125 SCETCH texts, namely 53, do so. The differences for LS2, CTTR and LV are more pronounced, by contrast. With LS2, 81 texts from FLOWN score above the mean and about 46 texts from SCETCH do so. As for CTTR, 93 texts from FLOWN score above the mean and 32 texts from SCETCH do so. With LV, as many as 102 texts from FLOWN score above the mean while only 26 from SCETCH do so.[20]

The Spearman correlation matrix for the four lexical richness indices, see the table in (6) below, reveals that CTTR and LV are very strongly correlated ($r = 0.945$). This means that the same texts tend very strongly to occupy the same region on the scales represented by these two indi-

[20] Recall the passage from Lu (2012: 203) quoted in section 3.1 above, saying "lexical variation correlated most strongly with the raters' judgments of the quality of ESL learners' oral narratives. No effect for lexical density emerged, and a very small effect was found for lexical sophistication". There appear to be interesting parallels here: between the insignificance of lexical density for the raters' judgement of quality and the insignificance of the average difference in terms of lexical density between FLOWN and SCETCH; between the strong significance of lexical variation for the raters' judgement of quality and the strong significance of the average difference in terms of lexical variation between FLOWN and SCETCH; between the weak significance of lexical sophistication for the raters' judgement of quality and the weaker significance of the average difference in terms of lexical sophistication between FLOWN and SCETCH.

ces. In other words, for these corpora it makes no big difference whether one looks at CTTR or LV. LS1 and LS2 also correlate strongly (r = 0.842).

(6) Spearman correlation matrix for lexical richness indices

```
        LD   LS1   LS2  CTTR    LV
LD   1.000 0.496 0.385 0.368 0.114
LS1  0.496 1.000 0.842 0.531 0.405
LS2  0.385 0.842 1.000 0.728 0.661
CTTR 0.368 0.531 0.728 1.000 0.945
LV   0.114 0.405 0.661 0.945 1.000
```

4.2 Statistical observations on syntactic complexity

One of the questions Mesmer & Cunningham & Hiebert (2012: 243) raise in their survey of research into linguistic aspects of texts for early grade readers of English is "*What is the contribution of specific syntactic features and patterns to sentence difficulty?*". I take this as motivation for concentrating on the constructional indices rather than the length indices (MLS, MLC, MLT) for FLOWN and SCETCH. In addition, it is arguable that the length of an expression in terms of word tokens is dependent on its constructional complexity rather than the other way round. However, it is also conceivable that a cause for a tendency towards shorter sentences, T-units and/or clauses in children's literature than in adults' literature (see scetch-sca.xlsx and flown-sca.xlsx; see above for the links for downloading these files) is that authors set out to write with the intention to construct relatively 'short' rather than 'constructionally less complex' syntactic units. On the level of intentions that are ultimately responsible for the statistical output, then, length considerations would be causal for constructional effects. Nevertheless, it seems to me that constructional issues are more interesting and more important for a comparison of texts for children and for adults if the study is to provide insights that are useful in the context of second language education.[21]

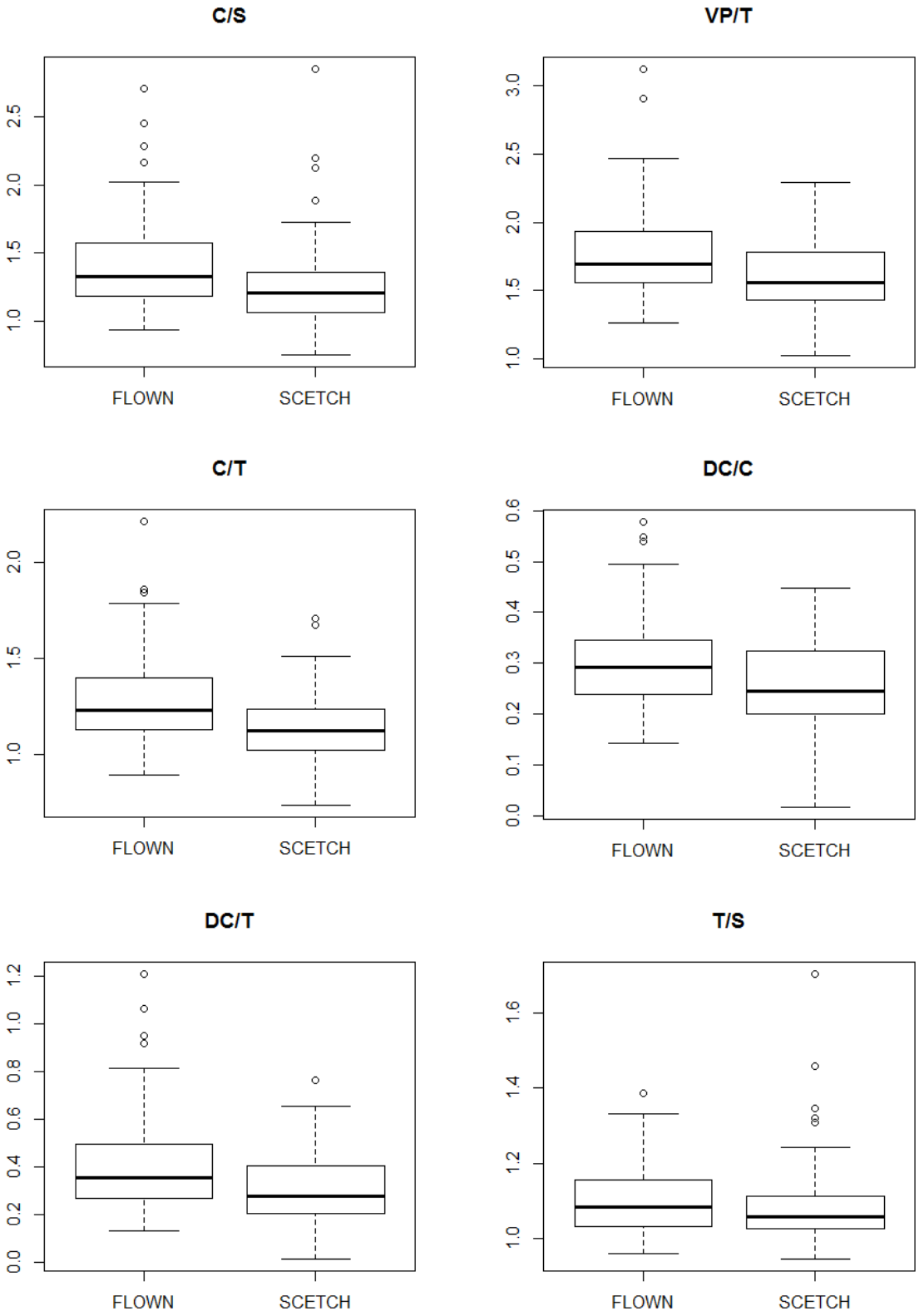
Performing bootstrapped independent samples *t*-tests (see above, footnote 17) at the 95% confidence level on the FLOWN and SCETCH data for each of the constructional indices suggests that there is no statistical difference between the two groups for the indices T/S (T-units per sentence), CP/T (coordinate phrases per T-unit) and CP/C (coordinate phrases per clause).[22] The boxplots for these indices do not correspond as clearly to the results of the *t*-tests as in the case of the lexical richness indices; see (7) below.

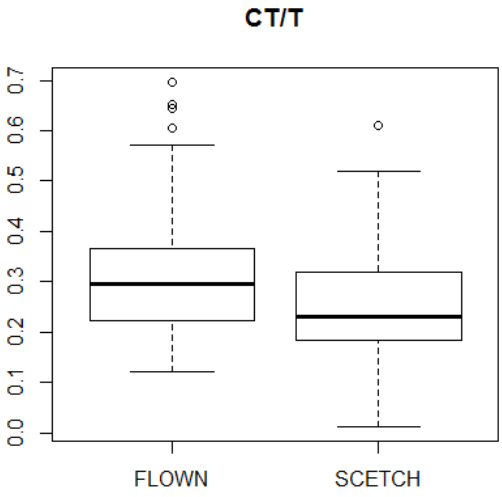
---

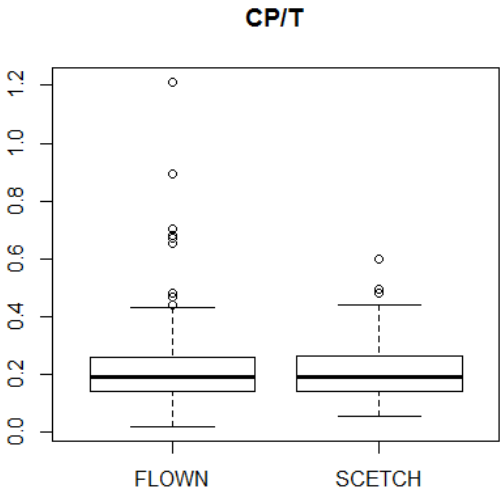
[21] But see also the result of an experiment reported by Koirala & Jee (2015). Investigating the perceived difficulty of sentences judged by five English language instructors and distinguishing between "traditional features (*sentence length* and *number of low-frequency words*) and six nontraditional features (counts of *clauses*, *dependent clauses*, *coordinate phrases*, *t-units*, *complex t-units*, and *Wh nominals* (Lu, 2010)" (Koirala & Jee 2015: 325), the authors come to the conclusion "that traditional features outperform the nontraditional features as single predictors of difficulty" (328). This emphasises the importance of the distinction between linguistic complexity and difficulty mentioned above (section 1) on the basis of a quotation from Mesmer & Cunningham & Hiebert 2012 and points to interesting questions for future research. Note that Koirala & Jee (2015: 329) "plan to conduct this same experiment on language learners and compare those findings with these current findings".

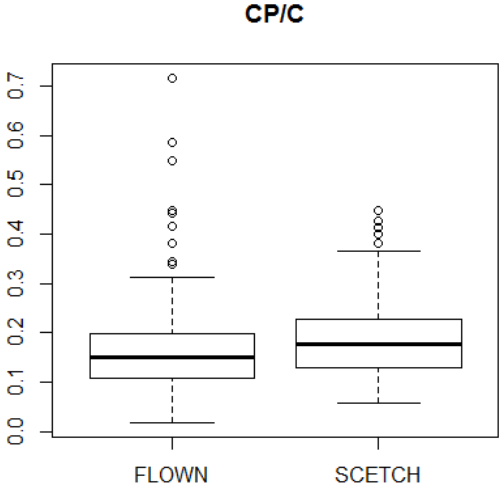
[22] C/S: mean difference: 0.173; 95% confidence interval: [0.099, 0.248];
VP/T: mean difference: 0.163; 95% confidence interval: [0.097, 0.234];
C/T: mean difference: 0.140; 95% confidence interval: [0.093, 0.189];
DC/C: mean difference: 0.047; 95% confidence interval: [0.027, 0.068];
DC/T: mean difference: 0.099; 95% confidence interval: [0.061, 0.141];
T/S: mean difference: 0.016; 95% confidence interval: [-0.008, 0.038];
CT/T: mean difference: 0.059; 95% confidence interval: [0.033, 0.086];
CP/T: mean difference: 0.011; 95% confidence interval: [-0.020, 0.051];
CP/C: mean difference: -0.015; 95% confidence interval: [-0.037, 0.011];
CN/T: mean difference: 0.342; 95% confidence interval: [0.245, 0.457];
CN/C: mean difference: 0.191; 95% confidence interval: [0.131, 0.257].

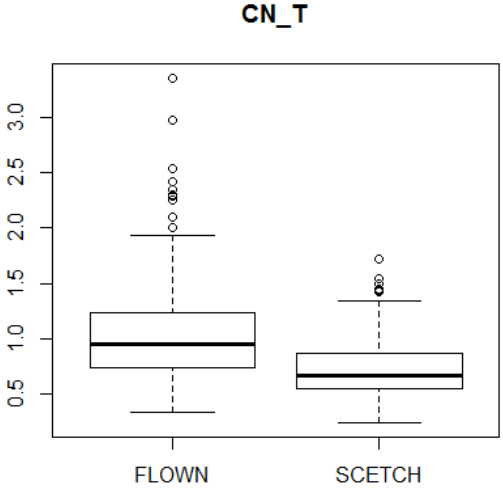
(7) Boxplots for 11 syntactic complexity indices for FLOWN vs. SCETCH

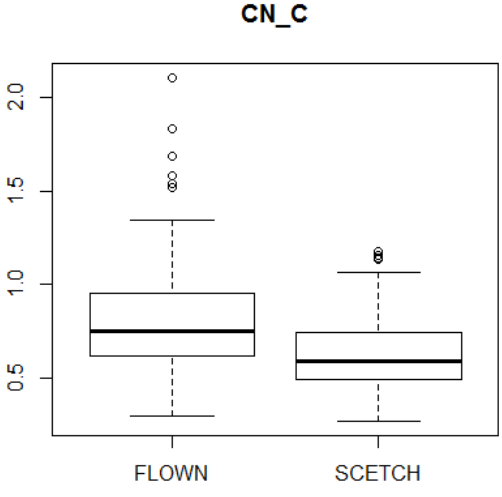



CT/T
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
FLOWN
SCETCH


CP/T
0.0
0.2
0.4
0.6
0.8
1.0
1.2
FLOWN
SCETCH


CP/C
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
FLOWN
SCETCH


CN_T
0.5
1.0
1.5
2.0
2.5
3.0
FLOWN
SCETCH


CN_C
0.5
1.0
1.5
2.0
FLOWN
SCETCH

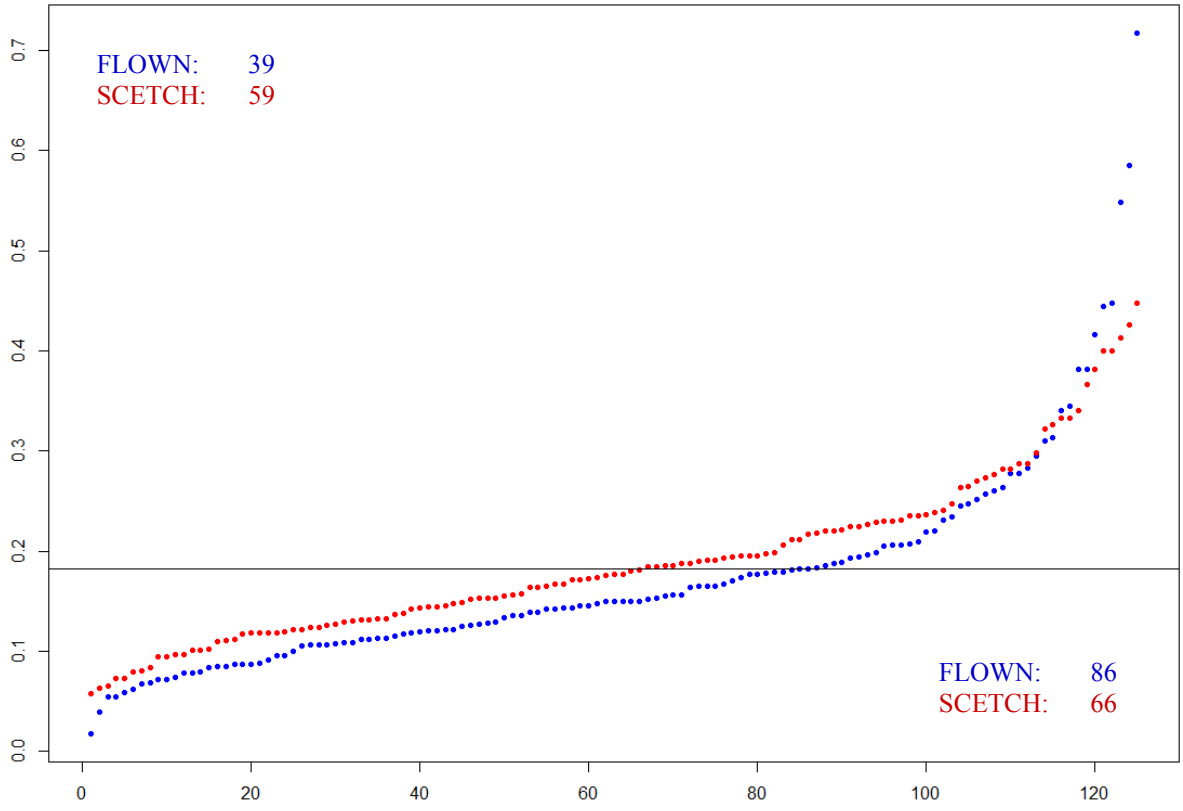
It can be seen that the FLOWN and SCETCH boxplots for CP/T look very similar; and this corresponds to the lack of a significant statistical difference as suggested by the *t*-test for this index. But the boxplots for all other constructional indices do not readily suggest the results of the corresponding *t*-tests; it would have been hard to predict from the boxplots that the *t*-tests for T/S and CP/C, beside that for CP/T, point to a non-significant statistical difference and to significance in the other cases.

As we will see in more detail below, many of the syntactic complexity indices correlate (very) strongly with one another. The scatterplots for the highly correlated indices would not differ much so that there is no need to present them all.[23] I provide the scatterplots for one of the two very strongly correlated coordination indices (CP/C), one of the two very strongly correlated indices concerned with complex noun phrases (CN/C), the scatterplot for one of several (very) strongly correlated indices (DC/C), and for T/S. These indices correlate with one another only weakly to moderately.

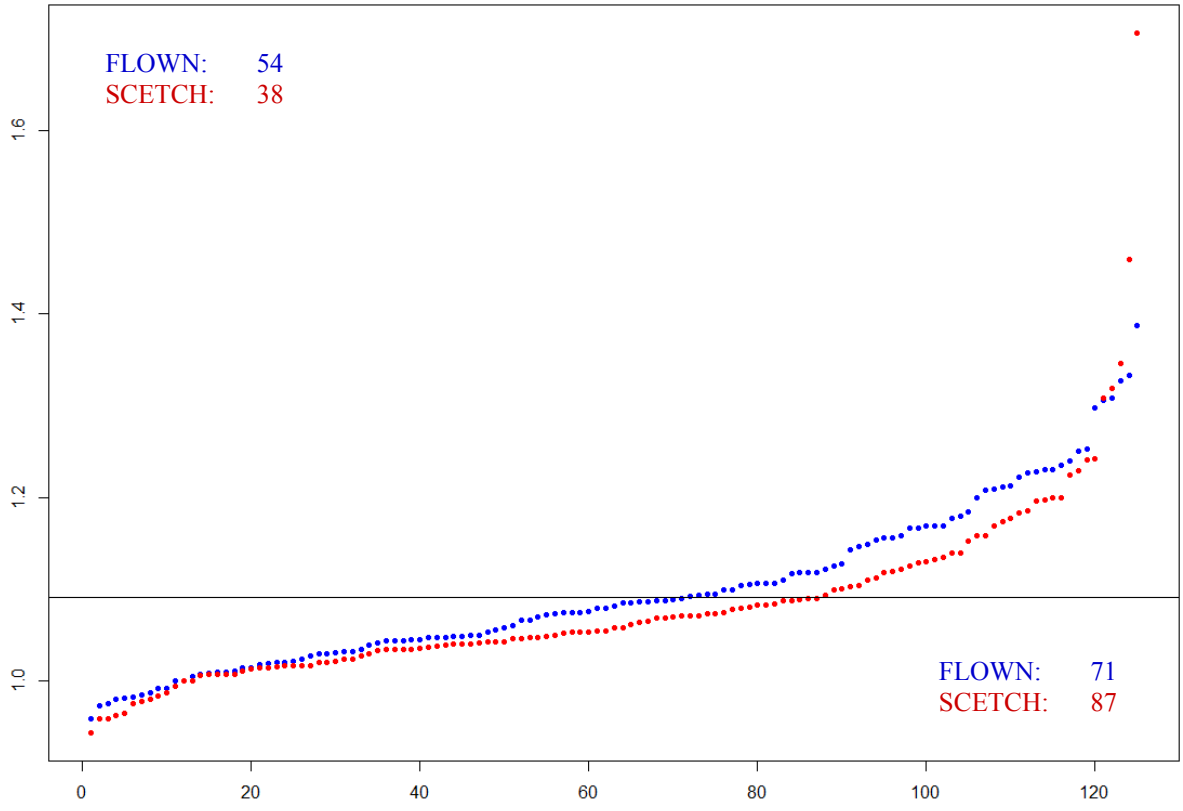
[23] But I give here number of the FLOWN and SCETCH texts above and below the mean index score for the other syntactic complexity indices as well:

		C/S	VP/T	C/T	DC/T	CT/T	CP/T	CN/T
above mean	FLOWN	63	66	70	65	68	41	72
above mean	SCETCH	39	42	39	40	43	51	30
below mean	FLOWN	62	59	55	60	57	84	53
below mean	SCETCH	86	83	86	85	82	74	95

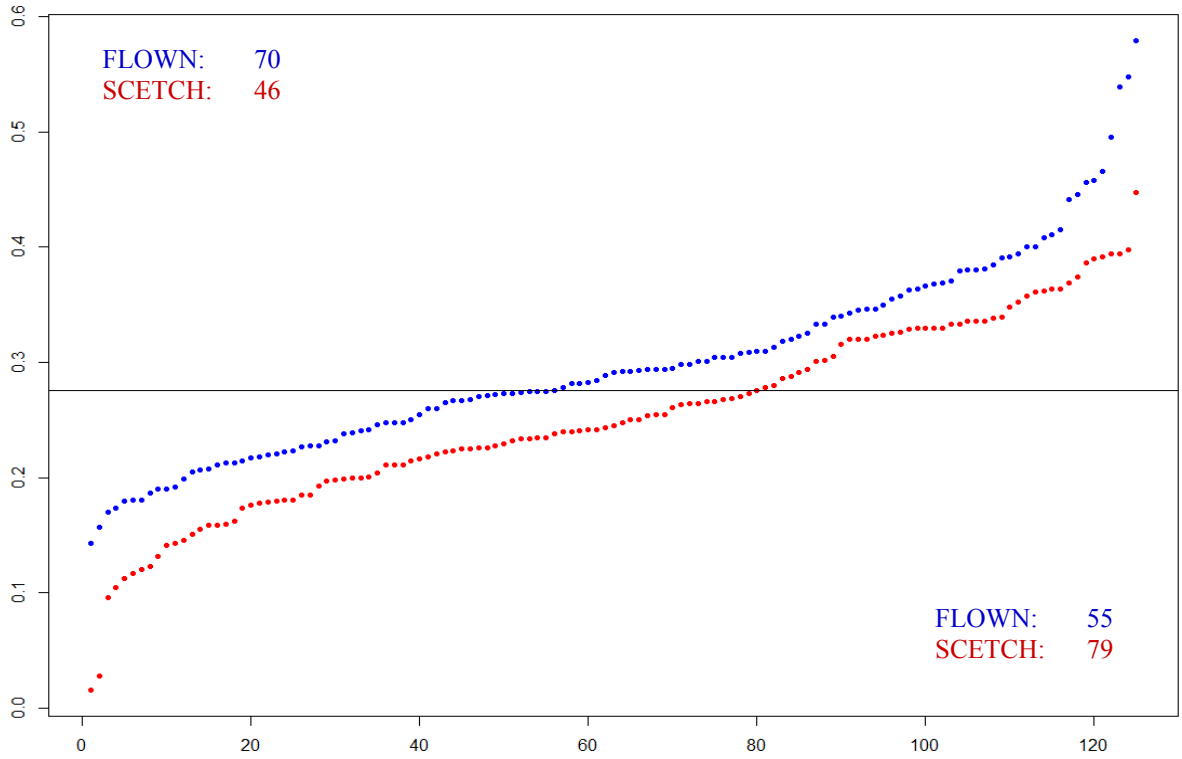
(8) a. CP/C plot: FLOWN vs. SCETCH



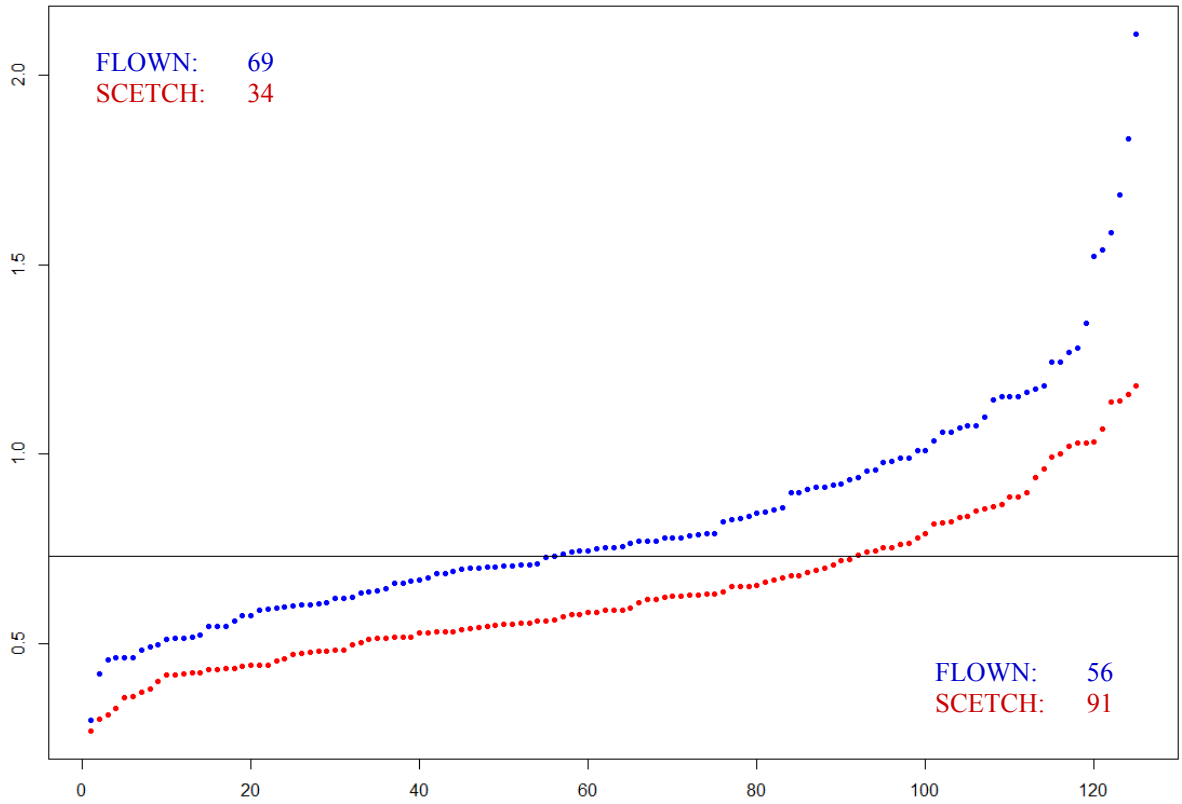
b. T/S plot: FLOWN vs. SCETCH



c. DC/C plot: FLOWN vs. SCETCH



d. CN/C plot: FLOWN vs. SCETCH



The scatterplots consolidate the impression gained from the *t*-tests and the boxplots: There is no big difference on average between FLOWN and SCETCH for CP/C and T/S; but note that with CP/C, in contrast to all the other constructional indices, the number of SCETCH texts that score above the mean is higher ($n = 59$) than for FLOWN ($n = 39$), a situation that is also revealed by the corresponding boxplot in (7) above (and which is similar to LD in the domain of the lexical richness indices). There is a more substantial difference on average between FLOWN and SCETCH for DC/C and CN/C. With DC/C, 70 of the 125 texts from FLOWN score above the mean while 46 of the 125 texts from SCETCH do so. With CN/C, 69 texts from FLOWN score above the mean while only 34 texts from SCETCH do so.

Let us consider the correlation matrix for the constructional syntactic complexity indices.

(9) Spearman correlation matrix for the constructional syntactic complexity indices

	C/S	VP/T	C/T	DC/C	DC/T	T/S	CT/T	CP/T	CP/C	CN/T	CN/C
C/S	1.000	0.876	0.950	0.810	0.886	0.718	0.883	0.412	0.132	0.689	0.456
VP/T	0.876	1.000	0.892	0.889	0.923	0.511	0.908	0.485	0.223	0.757	0.565
C/T	0.950	0.892	1.000	0.840	0.924	0.491	0.896	0.355	0.058	0.701	0.450
DC/C	0.810	0.889	0.840	1.000	0.982	0.422	0.938	0.431	0.180	0.675	0.482
DC/T	0.886	0.923	0.924	0.982	1.000	0.458	0.960	0.423	0.149	0.708	0.488
T/S	0.718	0.511	0.491	0.422	0.458	1.000	0.506	0.393	0.262	0.400	0.294
CT/T	0.883	0.908	0.896	0.938	0.960	0.506	1.000	0.424	0.153	0.683	0.472
CP/T	0.412	0.485	0.355	0.431	0.423	0.393	0.424	1.000	0.941	0.629	0.647
CP/C	0.132	0.223	0.058	0.180	0.149	0.262	0.153	0.941	1.000	0.428	0.527
CN/T	0.689	0.757	0.701	0.675	0.708	0.400	0.683	0.629	0.428	1.000	0.945
CN/C	0.456	0.565	0.450	0.482	0.488	0.294	0.472	0.647	0.527	0.945	1.000

The matrix shows that the indices C/S, VP/T, C/T, DC/C, DC/T and CT/T correlate strongly with one another (for all, $r > 0.800$), partly very strongly ($r > 0.900$). This is not surprising since they all measure the number of clauses (with 'clause' being in part conceptualised in slightly different ways) in relation to the number of clausal units (with 'clausal unit' again in part conceptualised in slightly different ways) in which they are embedded as dependent clauses. This is different for T/S, with which the indices just mentioned, with the exception of C/S, correlate only moderately ($0.422 < r < 0.511$). The relation between two adjoining T-units in a sentence is parataxis. The higher than moderate correlation between C/S and T/S ($r = 0.718$) is very probably due to the fact that a clause (C) is a necessary part of a T-unit (T); on the other hand, it is not necessarily the case that an increase (decrease) in the number of clauses in relation to sentences results in an increase (decrease) of the number of T-units in relation to sentences, since clauses of a sentence need not be paratactically related, but may be hypotactically related or embedded in one another. As can be expected, the two indices that are concerned with the complexity of noun phrases, CN/T and CN/C, correlate very strongly with one another ($r = 0.945$). While CN/C correlates moderately ($0.294 < r < 0.647$) with all the other indices, CN/T correlates moderately to strongly with them ($0.400 < r < 0.757$). Similarly, the two coordination indices CP/C and CP/T correlate very strongly with one another ($r = 0.941$), with CP/T correlating moderately ($0.355 < r < 0.647$) with all the other indices and CP/C correlating clearly less than CP/T with each of the other indices ($0.058 < r < 0.527$)

The statistical observations made so far show that on average FLOWN and SCETCH differ clearly with respect to most, but not all, of the lexical richness and syntactic complexity indices taken into account here. As far as lexical richness is concerned, lexical density (LD) does not distinguish the two corpora. The other lexical richness indices indicate a clear tendency on average towards a higher score for FLOWN compared to SCETCH. On the side of syntactic complexity, the coordination indices CP/T and CP/C as well as T/S, which is partly coordination-based as well (see below), do not distinguish the corpora. The majority of the syntactic complexity indices indicate a tendency on average towards a higher score for FLOWN. All in all,

then, and as was to be expected, FLOWN shows a trend towards more lexical richness and higher syntactic complexity than SCETCH on average.

However, with respect to any of the indices looked at, there is quite some overlap in the scores. For example, if we look at the score regions below and above the mean, we observe that there is no index on which there are less than 23 texts from FLOWN (= 18.4% of the FLOWN texts) below the mean for the whole corpus and no index on which there are less than about 26 texts from SCETCH (= 20.8% of the SCETCH texts) above this mean. This count (23 below mean for FLOWN; 26 above mean for SCETCH) holds for LV, which, on this criterion, is the best single indicator of those looked at here for the distinction between FLOWN and SCETCH. The overlap for the other indices is larger to different extents.

4.3 Linear discriminant analysis (LDA)

Recall the second question raised at the beginning of section 4: How well can texts for children be statistically distinguished from texts for adults on the basis of their indices data gathered from the two corpora? The preceding discussion has already given answers to this question for individual indices. The present section is devoted to finding answers for combinations of indices. The statistical technique to be employed is linear discriminant analysis (LDA). Backhaus et al. (1980/2016: 216) explain the general aim of LDA in the following way (see also Tabachnik & Fidell 1983/2014: ch. 9):

> Die Diskriminanzanalyse ist ein multivariates Verfahren zur Analyse von Gruppenunterschieden. Sie ermöglicht es, die Unterschiedlichkeit von zwei oder mehreren Gruppen hinsichtlich einer Mehrzahl von Variablen zu untersuchen, um Fragen folgender Art zu beantworten:
> *'Unterscheiden sich die Gruppen signifikant voneinander hinsichtlich der Variablen?'*
> *'Welche Variablen sind zur Unterscheidung zwischen den Gruppen geeignet bzw. ungeeignet?'*
> […]
> Während die Analyse von Gruppenunterschieden primär wissenschaftlichen Zwecken dient, ist ein weiteres Anwendungsgebiet der Diskriminanzanalyse von unmittelbarer praktischer Relevanz. Es handelt sich hierbei um die Bestimmung oder *Prognose der Gruppenzugehörigkeit* von Elementen (Klassifizierung). Die Fragestellung lautet:
> *'In welche Gruppe ist ein 'neues' Element, dessen Gruppenzugehörigkeit nicht bekannt ist, aufgrund seiner Merkmalsausprägungen einzuordnen?'*

The conceptual idea behind LDA is explained by Baayen (2008: 154) as follows: "in discriminant analysis, the idea is to choose the linear discriminants such that the means of the groups are as different as possible while the variance around these means within the groups is as small as possible". In the present context, the texts from FLOWN constitute a first group, those from SCETCH a second group; the variables mentioned in the quotation from Backhaus et al. 1980/2016: 216 above are a subset from the lexical richness and syntactic complexity indices. Linear discriminants are those linear combinations (i.e. linear combination functions) of the variables that discriminate best between the groups, with 'best' here to be interpreted as explained in the quotation from Baayen just given. If there are only two groups to be discriminated, as in the present study, there is only one linear discriminant. The variables that form part of a linear discriminant are often called predictor variables or simply predictors.

For carrying out a statistically powerful LDA it is necessary to look at the correlations between potential predictors. According to Norris (2015: 310),

> high correlations (a typical criterion is $r > .70$) between predictor variables [of an LDA] reduce the power of the analysis and may confuse the determination of discriminant functions due to superfluous variables. Where high correlations are identified, a single marker variable should be selected and other correlated variables eliminated from the analysis […].

Let us look at the correlations among the lexical richness indices first. As already pointed out, LV appears to be the best individual lexical richness discriminator between FLOWN and SCETCH texts. CTTR is the second best individual discriminator among the lexical richness indices, but, as shown by the correlation matrix in (6), it correlates very strongly with LV and correlates more strongly with the other lexical richness indices than does LV. Therefore LV will be retained for the LDA while CTTR is excluded. Since LS1 and LS2 are also strongly correlated with one another, it is reasonable to select only one of them in combination with LV. LS1 correlates less strongly with LV than does LS2 while LS2 distinguishes better between the two corpora. This gives the inclusion of LS1 an advantage with respect to the criterion of correlation, but possibly a disadvantage with respect to what can be expected of it in terms of its importance for the discrimination. The reverse situation holds for LS2. As for the constructional syntactic complexity indices, only one of the very strongly correlated nominal indices CN/C and CN/T should be used in the LDA and only one of the (very) strongly correlated clausal indices C/S, VP/T, C/T, DC/C, DC/T and CT/T. As far as CN/C and CN/T are concerned, I will work with CN/C as it is that of the two that correlates less strongly with the other indices.

Whether the inclusion in the LDA of LD, T/S and one of the strongly correlated coordination pairs CP/C and CP/T can produce a discriminating effect is hard to anticipate in advance. As was pointed out above, they do not statistically distinguish between the two corpus sections on their own. As far as the coordination indices are concerned, I assume that the inclusion of CP/C is to be preferred over CP/T since CP/C correlates less strongly with the other indices.

A series of LDAs was conducted, making use of the `lda()` function from the R MASS package, with the predictors from the list of indices in (10). The curly brackets here enclose strongly correlated indices of which only one was selected as a predictor in a given LDA. Moreover, if T/S was used, C/S was not used, since they both correlate strongly ($r = 0.718$).

(10) LD, LV, {LS1, LS2}, {C/S, VP/T, C/T, DC/C, DC/T, CT/T}, T/S, CP/C, CN/C

The series of LDAs comprised all combinations of 4 to 7 predictors from this set that conform to the conditions just pointed out. It was determined which combination resulted in the best classification in terms of the number of texts classified in accordance to the actual membership in the FLOWN or SCETCH group. This is the combination of the four predictors given in (11), which results in a classification according to fact of 106 FLOWN texts (= 84.8%) and 107 SCETCH texts (= 85.6%), that is, of 213 texts in all (85.2%).[24]

(11) LV, CT/T, CP/C, CN/C

The coefficients of the linear discriminant resulting from this LDA applied to the standardised values of the data set (i.e. the values of the original data set scaled to a mean of 0 and a variance of 1) are given in (12) below (rounded to the third decimal place).

[24] a. There is no problematic multicollinearity between the predictors in (11), as shown by applications of `vif()` from the R car package (see Wollschläger 2010/2014 : 202). No variance inflation factor (VIF) for any one of these predictors is higher than 3.08 (with values up to 4 being conventionally taken as uncritical).
b. The worst classification with a combination of the predictors in (11) that conforms to the conditions resulted in 146 texts being classified according to fact.

(12) Standardised coefficients of linear discriminant function

```
LV    0.998
CT/T  0.048
CP/C -0.670
CN/C  0.568
```

The scaling makes it possible to assess the relative importance of each of the predictors for the discrimination by comparing the absolute values for the coefficients of the linear discriminants resulting from the LDA (see Backhaus et al. 1980/2016: 243ff.).

The coefficients tell us that in the LDA based on this selection of indices as predictors, LV has the greatest weight in discriminating between the two corpora. It is followed by CP/C, CN/C and CT/T in this order, with CT/T being only marginally important. What is particularly interesting is the role of CP/C. It is the second most important predictor, with an importance weight of about two thirds of that of LV, but its influence on the discriminant score goes in the direction opposite to that of LV and CN/C.[25] That is, although CP/C does not statistically distinguish between FLOWN and SCETCH on its own, as we saw above, it plays an important role in improving on the discrimination by counteracting the weight of LV and CN/C. Note also that the grammatical conceptual basis of CP/C is parataxis. As pointed out earlier, CP/C is defined as "number of coordinate phrases divided by number of clauses", with coordination being the syndetic manifestation of parataxis. Hence the results of the LDA suggest that, if we want to consider the degree of parataxis as operationalised by CP/C as indices for syntactic complexity, it has to be considered a different kind of syntactic complexity than that operationalised by CN/C and CT/T. For obvious syntactic reasons (parataxis involving concatanation at the same hierarchical level; the structures underlying the operationalisation of CN/C and CT/T involving concatanation at different hierarchical levels), I would like to call the syntactic complexity manifested by CP/C first level syntactic complexity, and that manifested by CN/C and CT/Tsecond level syntactic complexity. I would venture that second level complexity is cognitively more demanding than first level complexity.

There are three combinations of predictors that resulted in the second best classification in terms of the number of texts classified according to fact, namely 212. These together with their coefficients are given in (13)

(13) a. LD, LS1, LV, CP/C, CN/C

Standardised coefficients of linear discriminant function

```
LD   -0.251
LS1  -0.119
LV    0.985
CP/C -0.595
CN/C  0.722
```

b. LD, LV, C/T, T/S, CP/C, CN/C

Standardised coefficients of linear discriminant function

```
LD   -0.302
LV    0.924
C/T   0.077
T/S  -0.089
CP/C -0.579
CN/C  0.716
```

[25] The discriminant score is the value for each text of the corpora that results as the sum of a constant plus the sum of the products of the score achieved on each predictor index multiplied by the corresponding coefficient in the linear discrimant.

c. LV, C/T, CP/C, CN/C

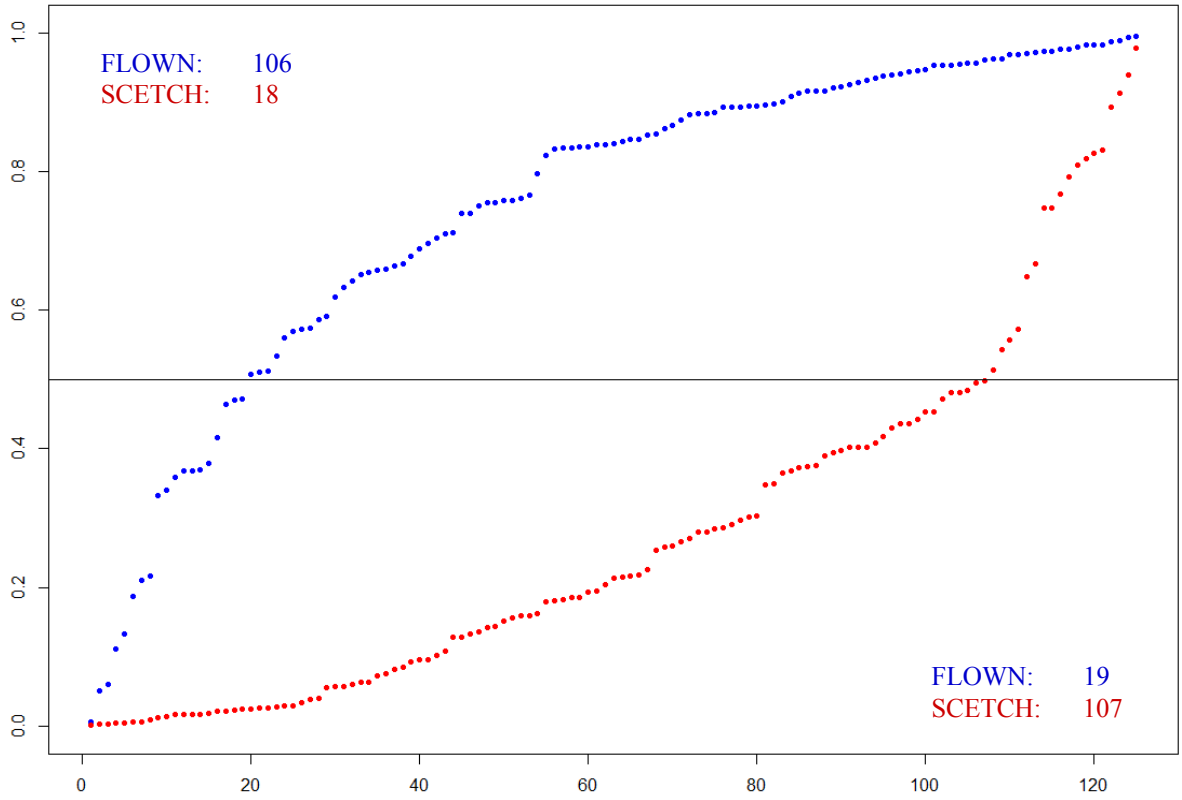
Standardised coefficients of linear discriminant function

```
LV    0.988
C/T   0.105
CP/C -0.649
CN/C  0.532
```

It can be seen that in combination with certain other predictors LD as well can make a contribution to a better fit of the classification to the actual grouping despite the fact that it does not distinguish between FLOWN and SCETCH on its own. This is similar to the case of CP/C mentioned above, with which LD shares in addition that it loads negatively on the discriminant function. Note also that if T/S appears in the set of predictors, it also loads negatively on the discriminant function. This supports the distinction between first and second level syntactic complexity made above, since the conceptual basis of T/S, like that of CP/C, is parataxis. As mentioned earlier, T/S is defined as "number of T-units divided by number of sentences". Coordination is the syndetic manifestation of parataxis, and the manner in which the number of T-units in a sentence can be increased is by combining T-units paratactically, i.e. syndetically (coordination) or asyndetically.

(14) below gives a sorted scatterplot based on the so-called posterior probabilities yielded by the LDA with the predictors in (11).[26] This is the probability for each text of the corpora to belong to FLOWN or SCETCH resulting from an application of the linear discriminant function to the values for each text on each index that is included as predictor in the linear discriminant.

(14) Posterior probabilities plot: FLOWN vs. SCETCH



[26] The so-called 'prior probability' is determined by the number of observations in each group. Since both groups (FLOWN and SCETCH) contain 125 observations each, the prior probability of each text to belong to either FLOWN or SCETCH is 0.5.

The plot is to be interpreted in the following way: The value of a blue or red dot on the *y*-axis provides the posterior probability for the text represented by the dot to belong to the FLOWN group. The horizontal line marks the 0.5 probability. That is, the lower the probability value for a text, i.e. the closer the corresponding dot to *y = 0*, the lower its posterior probability to belong to FLOWN and, correspondingly, the higher its posterior probability to belong to SCETCH; the higher the probability value for a text, i.e. the closer the corresponding dot to *y = 1*, the higher its posterior probability to belong to FLOWN, and, correspondingly, the lower its posterior probability to belong to SCETCH. In categorial terms, the texts below the horizontal line are analysed by the LDA as belonging to SCETCH and the texts above the horizontal line are analysed as belonging to FLOWN. It can be seen that 19 texts (= 15.2%) from FLOWN are assigned to the SCETCH group by the LDA, and 18 texts (= 14.4%) from SCETCH are assigned to the FLOWN group. This is an overall rate of 14.8% of the texts classified contrary to fact, i.e. 85.2% of the texts are classified conforming to fact.

The results of the LDA presented so far cannot be taken at face value without further statistical discussion, though. The problem is that LDA is based on a set of assumptions concerning the structure of the data. According to Norris (2015: 309), these assumptions are "independence of observations on each variable, univariate and multivariate normality, homogeneity of variance and covariance, and no evidence of multicollinearity". Univariate and multivariate normality as well as homogeneity of variance and covariance are not given in my data set. In addition, some of the predictors are characterised by the existence of outliers, a situation that is also commonly mentioned as problematic for LDA (see e.g. Norris 2015: 309). On the other hand, the data structure does not appear to be strongly deviant from normality and homogeneity, and the number of outliers does not appear to be very high. Moreover, as pointed out by Bortz & Schuster 1977/2010: 492),

> [a]uch für die Diskriminanzanalyse gilt, dass Verletzungen der Voraussetzungen […] mit wachsendem Stichprobenumfang weniger folgenreich sind. Unter dem Gesichtspunkt der Stabilität der Kennwerte der Diskriminanzanalyse (insbesondere der Faktorladungen) fordert J. Stevens […] dass *N* [i.e. the size of the sample] 20-mal so groß sein sollte wie *p* [i.e. the number of variables, i.e. indices in our case].

Note that with the four to six variables for the LDAs described above, the criterion of minimally 20 sample points per variable is fulfilled. Still, it is hard to judge what the effects on the LDA are of the deviances from the assumptions in the present data set on the one hand and of the apparently sufficiently large sample size on the other hand.

For situations like the one just described, Plonsky (e.g. 2015) and Larson-Hall (e.g. 2016), among others, suggest the bootstrapping methodology in order to determine confidence intervals for statistical measures. Making use of the `boot()` function of the R boot package, I conducted a run of `lda()` with the 4 predictors mentioned in (11) above over 10,000 random samples from the standardised data set and of the same size as this data set (such samples commonly being called replicates). The result of this bootstrapped LDA is summarised in (15).

(15) Result of LDA based on 10,000 bootstrap replicates

```
      original    bias  std. error      confidence interval (bca)
LV       0.998   0.009       0.127      95%   ( 0.726,  1.227 )
CT/T     0.048   0.006       0.100      95%   (-0.146,  0.242 )
CP/C    -0.670  -0.012       0.140      95%   (-0.925, -0.375 )
CN/C     0.568  -0.004       0.175      95%   ( 0.219,  0.907 )
```

The values given under *original* here are the standardised coefficients of the linear discriminant computed for the standardised data set, that is, the same values as in (12) above. The bias is the mean of the differences between the original value for each predictor and the coefficient com-

puted for each of the 10,000 replicates. It can be seen that the bias is low for each of the variables. By contrast, the standard error, that is, the variance of the values computed for the respective predictor in each of the replicates, tends to be rather high for all predictors, and, correspondingly, the 95% confidence interval for each predictor is rather large. The small bias means that the original LDA represents the *average* for coefficients of linear discriminants yielded by applications of LDA to a large number of random replicates of the original data set quite well; the rather high standard errors (and thus rather large confidence intervals) mean that the coefficients computed for the individual replicates are quite strongly *dispersed around the average value*. Such a strong dispersion was to be expected. The reason is that there is with respect to every single one of the indices that went into the LDA as a predictor quite a substantial number of texts from SCETCH that achieve scores above the mean for the respective index, and quite a substantial number of texts from FLOWN that achieve scores below the mean for the respective index. Consequently, there is quite some chance for several of the 10,000 replicates in the bootstrapped LDAs to be composed in a way that deviates rather strongly from the original composition of the data set. This, in turn, is bound to lead to coefficients for linear discriminants that deviate quite strongly from the mean. However, as far as the discrimination between FLOWN and SCETCH in terms of an average general trend is concerned, the original LDA appears to give a rather accurate picture. Bootstrapped LDAs with the predictors in (13)a-c result in the same pattern: rather low biases (between ca. 0.001 and 0.015); rather high standard errors (between ca. 0.101 and 0.184) and consequently rather large confidence intervals.

The linear discriminant resulting from the LDA with the predictors in (11) was used to calculate the probabilities for belonging to the population of texts represented by FLOWN or SCETCH respectively of 22 additional texts, that is, texts that are not part of FLOWN or SCETCH and, thus, whose complexity indices data were not used for the LDA. These 22 texts consist of 11 narrative texts for adults and 11 narrative texts for children. The texts for adults come from the same pool of texts from which the texts for FLOWN were extracted (i.e. they were randomly selected from FLOB and FROWN) and were modified in the same way as the texts that became part of FLOWN. 10 of the additional texts for children were extracted from the internet, just like the majority of texts in SCETCH, and one was taken from a printed source. The additional texts for children were, of course, also modified in the same way as the texts that became part of SCETCH.

The classification of these additional 22 texts on the basis of the linear discriminant resulting from the LDA with the predictors in (11) yielded the following result: All 11 texts for children were predicted to belong to the SCETCH group with probabilities ranging from 0.876 to 0.993 for 9 of these texts and 0.768 and 0.516 for the remaining two texts. As for the texts for adults, nine of them were predicted to belong to the FLOWN group with probabilities ranging from 0.588 to 0.989 and two were predicted to belong to the SCETCH group with probabilities 0.523 and 0.769. If we assume that the FLOWN group represents narrative texts for adults and the SCETCH group narrative texts for children, then this classification has a conforming-to-fact prediction rate of ca. 91% (20 from 22 texts classified conforming to fact).

In order to give an impression of the text for adults that was classified as belonging to the SCETCH group with a rather high probability of 0.769, I quote here its first eight sentences:

(16) When he noticed that he was being watched, Milt Saunders sank his head between his shoulders so that it appeared momentarily as if he had no neck. He reminded Tommy of a bird. Then Milt straightened up and raised his glass and drank from it. Tommy watched him swallow. He had never before paid much attention to the movement of a man's Adam's apple. By 10:30 the Legion was crowded and noisy. The band continued to play and eve-

> ryone had to talk above the music if they hoped to be heard. Under the lights the thick smoke hung in the air like fog. (from FROWN: P26)

5 Conclusion and some discussion

The statistical analyses in the preceding sections were carried out in order to provide answers to the following questions (see section 1): Is the genre of literature for children distinguishable in terms of lexical and/or syntactic complexity from the genre of literature for adults – and if so, to what extent? As always when statistical analyses of samples rather than whole populations are involved, the answers that were more or less explicitly given above can only be as adequate as the FLOWN and SCETCH corpus sections are representative of the genre of literature for adults and for children respectively. Given that the population of texts is immense and ever growing, the question of the representativeness of text samples is a difficult and well-known one in corpus-based linguistics. The following quotations from Koester 2010, however, suggest that working with a corpus of such dimensions as the present one is not commonly considered worthless:

> By running a number of statistical tests, Biber [1993] discovered that the most common linguistic features (e.g. personal pronouns, contractions, past and present tense and prepositions) are relatively stable in their occurrence across 1,000-word samples. He also looked at how many text samples are needed to adequately represent a register or a genre in a corpus, and found that the linguistic tendencies are quite stable with ten (and to some extent even five) text samples per genre or register (Biber 1990). (Koester 2010: 70)

> [W]e should bear in mind that it is not possible to evaluate 'representativeness' entirely objectively (Tognini Bonelli 2001: 57). We may only discover that a corpus is *not* representative if it turns out that the results are skewed in some way. (Koester 2010: 69)

Thus, just like other corpus-based work with rather small corpora, it is one function of the present study to invite further research on linguistic similarities and differences between (narrative) texts for adults and (narrative) texts for children. However, as already pointed out above, it is also clear that SCETCH can at best be representative of that part of the genre of children's literature that comprises texts that are longer than about 1,000 words.

Against the background of the proviso just mentioned, the present study suggests that the genres of narrative texts for adults and for children are indeed statistically different with respect to most of the lexical and syntactic indices considered here taken individually. On the lexical side, only LD (lexical density) and on the syntactic side only the parataxis-based indices CP/C (coordinate phrases per clause), CP/T (coordinate phrases per T-unit) and T/S (T-units per sentence) do not distinguish them. However, the plots in (4), (5), (7), and (8) as well as the figures provided for the number of texts above and below the respective mean scores (see also footnote 23) show that there is a large overlap of FLOWN and SCETCH texts with respect to the scores above and below the mean for most of the indices. Only for the lexical index LV is the overlap as low as ca. 25% in both regions, i.e. above and below the mean (25.5% above the mean; 23.2% below the mean). Among the syntactic indices, it is CN/T that yields the lowest overlap, but it is much larger than the one for LV, namely 41.7% in the region above the mean and 55.8% in the region below the mean. Hence, among the indices looked at here, LV is the index with respect to which the genres of texts for adults and for children differ most clearly, while they differ much less clearly with respect to any of the syntactic indices. This insight is confirmed by the coefficients of the linear discriminants that classify the texts in a manner that conforms best to actual group membership. LV has clearly the greatest weight here, followed at a considerably distance by CN/C, which weights in the same direction as LV and is a syntactic second level complexity

index, and by CP/C, which weights in the opposite direction and is a syntactic first level complexity index. Thus, the result of the present study that may turn out to be relevant and important for pedagogical and didactic purposes concerned with learning and teaching to read English as L2 is this: Tendentially, texts aimed at adults differ from texts aimed at children considerably more with respect to lexical richness, especially as measured by LV, than with respect to syntactic complexity.

The study also suggests that an appropriately selected set of more than one complexity index can statistically discriminate between texts from the two genres quite successfully. A set of four indices (LV, CT/T, CP/C, CN/C) from among those considered here was shown to discriminate between the texts from FLOWN and SCETCH with a conforming-to-fact rate of 85.2% in an LDA. Three other sets of four to six indices result in a slightly lower conforming-to-fact classification of 84.8%. The four indices LV, CT/T, CP/C, CN/C used as predictors applied to 22 texts from the genres of narrative literature for adults and for children that are not part of FLOWN and SCETCH yielded a conforming-to-fact rate of ca. 91%.

For a text from FLOWN to be classified contrary to fact with a high probability does not necessarily mean that it should give the impression of a linguistically simple text. It may very well be true that the majority of texts from FLOWN that are so classified do give this impression, as the extracts in (17) may be able to suggest, where I set some extracts from such texts from FLOWN beside some extracts from texts from SCETCH that are classified in conformity to fact with a high probability as well.

(17) a. A large crowd had gathered, encircling the fighters. Even Tom Riley was there with his deputies. They seemed to be enjoying the show. 'Knock his teeth down his damn throat, Sam!' a man yelled. 'Who said that?' Pete shouted, looking around him. Sam decked him, and the young man landed hard on his butt. (FLOWN R-N02)

b. Two guards marched in front of him, and two guards marched behind him. They all carried guns. 'There he is!' Rilla cried. 'Commander!' And she began to run along the road to him. Commander Zadak did not stop. A guard opened the door of the train and the Commander got in. Slowly, the train began to move. Kiah ran to the taxi and jumped in. (SCETCH b-fsl)

c. He was a trickster child. Stone, the last born manidoo, seldom moved from his place on the earth. The birds, flowers, and animals came with his birth. Naanabozho and his other brother were together most of the time, but not with their brother Stone. At night the brothers shared their adventures with Stone because he could not travel. (FLOWN R-K07)

d. The zebra looked up as he swallowed a mouthful of water. 'But you're a hippo. Hippos don't need stripes.' Harry grew impatient. 'I know, I know but hippos are so ugly and zebras so pretty. I just wish.' 'Be careful what you wish for.' the zebra said and walked away. Harry sat by the watering hole all day watching the animals come and go. (SCETCH a-005)

e. 'We should've gone with them.' Paula suggested after Mom and David were out of view. 'I don't have any gum, and my ears always bother me on airplanes.' 'Paula.' Christy pointed out, 'you've only been on one airplane in your whole life and that was a few days ago coming out here.' 'I know. And I chewed gum the whole time. Marti, would it be okay if we went to get some gum?' (FLOWN R-P24)

f. We call her Rubbish because she'll eat anything. She ate my shoelaces this morning. When the bucket is full Granny and Lancelot take it to their house and put it in a big plastic tub. They're making it into yoghurt. And you should see our garden too. Dad's planted vegetable seedlings all over the place - potatoes, tomatoes, beans, spinach, lettuce, marrows - all sorts. (SCETCH o-002)

The extract in (18), though, is a FLOWN text that is equally predicted by the LDA with the predictors LV, CT/T, CP/C, CN/C to belong to the SCETCH group with a high probability, but makes the impression of a much more difficult text.

(18) A number of circumstances focused James VI's attention on Lewis in particular. It was at this time owned by a branch of the powerful Clan Macleod, and in the last years of the old Chief Ruari or Rory Macleod, who died c. 1595, Lewis was torn by ferocious family feuding over who was to succeed him on his death. There were a number of claimants, for the old man had been married three times, and had fathered a brood of illegitimate sons besides the legitimate offspring from his lawful marriages. In the last decades of the 16th century Lewis had acquired the reputation of possessing great undeveloped agricultural and fishing wealth. (FLOWN L-N24)

The reason for the contrary-to-fact classification of this text appears to be the following: It has a well above mean score for LV and a very high (second highest) score for CN/C, a slightly below mean score for CT/T, while its CP/C score is extremely high, the by far highest one, actually. I would argue that the scores for LV and CN/C are responsible for the impression of its being a rather difficult text. But its extremely high score for CP/C, which loads negatively on the linear discriminant and is the second most important coefficient for the discrimination, causes it to get a discriminant score well below the mean and thus to be classified as belonging to SCETCH. This is certainly an exceptional case, but one that illustrates potentially surprising details of the LDA quite well. Such exceptions aside, we should be allowed to expect that texts that score similarly with respect to their posterior probabilities tend also to be similar with respect to their linguistic complexity.

That said, the results of the LDA clearly suggest that there are quite some texts from SCETCH that manifest a complexity that makes them indistinguishable from texts from FLOWN. I give a couple of extracts from SCETCH texts that are classified as FLOWN texts with a rather high probability.

(19) a. Forcing his way through a gap in a hedge, he caught his hand on a thorn and saw beads of blood ooze from a long scratch. Setting his face towards the barrow, which was still perhaps half a mile away in dramatic relief against the angry sky, he had the feeling that he was being followed; that wherever he went, and no matter how fast he covered the ground, his every move was shadowed. But who were his pursuers? Were they strange, silent figures, festooned with Christmas baubles, tracking him with measured, ghostly tread; timeless, faceless, without mercy? (SCETCH b-acv)

b. 'Good bye, then.' he said, brandishing his notebook, containing Ashok's identity card number; all the soldiers had claimed that they were never registered for the card, which Ashok really doubted, but which the policeman didn't question, as the air whistled out of his nostrils and he sweated in his uniform. The rains had finally come, the skies opening like floodgates, the rain falling in sheets the color of the pollution they absorbed on their fall from the heavens. (SCETCH w-026)

c. 'Idle recriminations and sweeping statements are a waste of time. Where's the reason in pursuing something that doesn't exist. Threats haven't moved Mr Lambert or the blackamoor. We've one course left open. Put such facts as we have before the Sheriff and the Merchant Venturers and let them sort it out.' No one was sure whether he had deliberately set a trap, or whether it was merely a chance remark that brought about such astounding results. (SCETCH b-c85)

d. Dotty stared at the tiny little figures, perfect in every detail, with their bare feet and dirty faces, and their circular chimney brushes. There could now be no question as to whether her mum knew about the sweeps; the only question that remained was whether she had been the only one in the Calendar House to know about it, or whether the whole

> household knew. Dotty was sure that Pip would be able to tell her. But Pip was nowhere to be found. It occurred to Dotty that if only she could find her way into the courtyard she would perhaps be in with a chance of finding him, or at least in with a chance of discovering some of the answers to her questions. (SCETCH w-021)

It was probably to be expected that there are texts aimed (presumably or allegedly) at a child audience that cannot be distinguished on the basis of linguistic complexity indices from texts aimed (presumably or allegedly) at an adult audience, just as vice versa. This is due to the fact that the concept of genre is largely characterised by non-linguistic, or language external, criteria. Moreover, the membership of a text to a genre may also be determined by not much more than, or even nothing but, what genre it is decreed to belong to by an individual or group of individuals (e.g. the author and/or the publisher). This may be for a variety of reasons, including marketing reasons. In view of this consideration it may appear quite surprising that the LDA applied to FLOWN and SCETCH results in a rather high degree of conforming-to-fact discrimination.

References

Anderson, Celia C. 1985. "Style and language in children's books". *Children's literature association quarterly* 10;3: 113-134.

Baayen, R. H[arald]. 2008. *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Backhaus, Klaus & Erichson, Bernd & Plinke, Wulff & Weiber, Rolf. 1980/2016. *Multivariate Analysemethoden*. 14th edn. Berlin: Springer Gabler.

Biber, Douglas. 1990. "Methodological issues regarding corpus-based analyses of linguistic variation". *Literary and linguistic computing* 5;4: 257-269.

Biber, Douglas. 1993. "Representativeness in corpus design". *Literary and linguistic computing* 8;4: 243-257.

Bortz, Jürgen & Schuster, Christof. 1977/2010. *Statistik für Human- und Sozialwissenschaftler*. 7th edn. Berlin: Springer.

Fitzgerald, Jill & Elmore, Jeff & Koons, Heather & Hiebert, Elfrieda H. & Bowen, Kimberly & Sanford-Moore, Eleanor E. & Stenner, A. Jackson. 2015. "Important text characteristics for early-grades text complexity". *Journal of educational psychology* 107;1: 4-29.

Gressenich, Eva. 2011 "Einführung von Diskursreferenten im Bilderbuch". *Zeitschrift für Literaturwissenschaft und Linguistik* 162 [Special issue: Klein, Wolfgang & Meibauer, Jörg (eds.), *Spracherwerb und Kinderliteratur*]: 74-92.

Hempelmann, Christian F. & Rus, Vasile & Graesser, Arthur C. & McNamara, Danielle S. 2006. "Evaluating state-of-the-art treebank-style parsers for Coh-Metrix and other learning technology environments". *Natural language engineering* 12;2: 131-144.

Hunt, Kellogg W. 1965. *Grammatical structures written at three grade levels*. NCTE research report no. 3. Champaign (IL): National council of teachers of English.

Klein, Dan & Manning, Christopher D. 2003. "Accurate unlexicalized parsing". In: *Proceedings of the 41st annual meeting of the association for computational linguistics*. Vol. 1. Association for Computational Linguistics. 423-430.

Knowles, Murray & Malmkjaer, Kirsten. 1996. *Language and control in children's literature*. London: Routledge.

Koester, Almut. 2010. "Building small specialised corpora". In: O'Keeffe, Anne & McCarthy, Michael (eds.). *The Routledge handbook of corpus linguistics*. London: Routledge. 66-79.

Koirala, Cesar & Jee, Rebecca Y. 2015. "Experimental analyses of the factors affecting the gradience in sentence difficulty judgements". In: Helm, Francesca & Bradley, Linda & Guar-

da, Marta & Thouësny, Sylvie (eds.). *Critical CALL: Proceedings of the 2015 EUROCALL conference, Padova, Italy*. Dublin: Research-publishing.net. 324-329.
Larson-Hall, Jenifer. 2015. *A guide to doing statistics in second language research using SPSS and R*. London: Routledge.
Levy, Roger & Andrew, Galen. 2006. "Tregex and Tsurgeon: tools for querying and manipulating tree data structures". In: *Proceedings of the 2006 conference on language resources and evaluation*. European Language Resources Association. 2231–2234.
Lindgren, C. 2003. "A linguistic approach to children's literature". *Moderna Språk* 97;1: 71-83.
Lu, Xiaofei. 2010. "Automatic analysis of syntactic complexity in second language writing". *International journal of corpus linguistics* 15;4: 474-496.
Lu, Xiaofei. 2011. "A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development". *TESOL quarterly* 45;1: 36-62.
Lu, Xiaofei. 2012. "The relationship of lexical richness to the quality of ESL learners' oral narratives". *The modern language journal* 96;2: 190-208.
Lu, Xiaofei. 2014. *Computational methods for corpus annotation and analysis*. New York: Springer.
McNamara, Danielle. S. & Graesser, Arthur C. & McCarthy, Philip. M. & Cai, Zhiqiang. 2014. *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge (MA): Cambridge University Press.
Mesmer, Heidi Anne & Cunningham, James W. & Hiebert, Elfrieda H. 2012. "Toward a theoretical model of text complexity for the early grades: learning from the past, anticipating the future". *Reading research quarterly* 47;3: 235-258.
Minnen, Guido & Carroll, John & Pearce, Darren. 2001. "Applied morphological processing of English". *Natural language engineering* 7;3: 207-223.
Munat, Judith. 2007. "Lexical creativity as a marker of style in science fiction and children's literature". In: Munat, Judith (ed.). *Lexical creativity, texts and context*s. Amsterdam: Benjamins. 163-185.
Norris, John M. 2015. "Discriminant analysis". In: Plonsky (ed.) 2015: 305-328.
Plakans, Lia & Bilki, Zeynep. 2016. "Cohesion features in ESL reading: comparing beginning, intermediate and advanced textbooks". *Reading in a foreign language* 28;1: 79-100.
Plonsky Luke. 2015. "Statistical power, *p* values, descriptive statistics, and effect sizes: a "back-to-basics" approach to advancing quantitative methods in L2 research." In: Plonsky (ed.): 23-45.
Plonsky, Luke (ed.). 2015. *Advancing quantitative methods in second language research*. New York: Routledge.
Puurtinen, Tiina. 1998. "Syntax, readability and ideology in children's literature". *Translators' journal* 43;4: 524-533.
Stamou, Anastasia G. 2012. "Representations of linguistic variation in children's books: register stylisation as a resource for (critical) language awareness". *Language awareness* 21;4: 313-329.
Stephens, John. 2005. "Analysing texts: linguistics and stylistics". In: Hunt, Peter (ed.). *Understanding children's literature*. London: Routledge. 73-85.
Tabachnick, Barbara G. & Fidell, Linda S. 1983/2014. *Using multivariate statistics*. 6th edn. Edinburgh: Pearson Education.
Thompson, Paul & Sealey, Alison. 2007. "Through children's eyes? Corpus evidence of the features of children's literature". *International journal of corpus linguistics* 12;1: 1-23.
Tognini Bonelli, Elena. 2001. *Corpus linguistics at work*. Amsterdam: Benjamins.
Toutanova, Kristina & Klein, Dan & Manning, Christopher D. & Singer, Yoram. 2003. "Feature-rich part-of-speech tagging with a cyclic dependency network". In: *Proceedings of the 2003 conference of the North American chapter of the Association for Computational Lin-*

guistics on human language technology. Vol. 1. Association for Computational Linguistics. 173-180.

Vajjala, Sowmya & Meurers, Detmar. 2012. "On improving the accuracy of readability classification using insights from second language acquisition". In: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*: 163-173. Association for Computational Linguistics.

Wolfe-Quintero, Kate & Inagaki, Shunji & Kim, Hae-Young. 1998. *Second language development in writing*: *measures of fluency*, *accuracy, and complexity*. Honolulu (HI): University of Hawaii Press.

Wollschläger, Daniel. 2010/2014. *Grundlagen der Datenanalyse mit R: Eine anwendungsorientierte Einführung*. 3rd edn. Berlin: Springer.