# Fleshing out your text: How elaboration and contextualization moves differentially predict writing quality

Henk Pander Maat[1], Kees de Glopper[2], Kay Raaijmakers, Joris Veerbeek[1] & Dennis Vermeulen[1]

[1] Utrecht University | The Netherlands
[2] University of Groningen | The Netherlands

**Abstract**: This study explores the relation between writing quality and contextualization and elaboration moves, two kinds of textual expansion devices crucial for building common ground between writers and readers. We ask whether elaboration and contextualization features differentially predict writing quality and whether their quality contributions differ between genres. We also ask to what extent elaboration and contextualization are tied to individual writers, and can be explained by writer characteristics. To examine these issues, we annotated descriptive and argumentative texts of Dutch adolescents. Text quality was rated holistically, using benchmark scales.

As regards elaboration, depth affects quality more than breadth does. It also contributes across genres, whereas breadth only contributes in argumentations. Depth shows a large individual consistency across tasks, which is substantially related to students' school type, grade and gender. Breadth shows weaker links to individual writers and their characteristics.

With regard to contextualization, opening and closing moves play a modest role in text quality. Initial support moves contribute to quality across tasks; concluding moves contribute more in argumentations. Concluding moves are most consistent within writers; however, for all contextualization moves, the writer variance is substantially explained by writer characteristics. This study opens up new avenues for explicating writing quality and writing skill.

**Keywords**: writing quality; annotation; text length; move structure; elaboration; contextualization

## 1. Introduction

The text-focused tradition in writing research aims at explicating the concept of writing quality in terms of observable text features. Regarding local features, human ratings of L1 writing quality have been shown to correlate with features such as lexical frequency (e.g. McNamara et al., 2010; Olinghouse & Wilson, 2013; McArthur et al., 2019), syntactic complexity (Myhill, 2008; Beers & Nagy, 2009; McNamara et al., 2010; Uccelli et al., 2013; Dobbs, 2014), lexical diversity (McNamara et al., 2010; Olinghouse & Wilson, 2013; Salas et al., 2016) and grammar, usage and mechanics (Chen et al., 2017). Beyond the sentence level, referential cohesion (Weston et al., 2011; Varner et al., 2013; MacArthur et al., 2019) and markers of organization and coherence (Uccelli et al., 2013; Dobbs, 2014) show relations to text quality judgments. Many of the features mentioned so far lend themselves to automatic analysis, at least for English, which opens exciting opportunities both for research and for the development of tools providing writing analytics.

At the same time, we should ask whether the features mentioned so far might miss essential components of the writing quality concept. One indication that they may be doing so is that text length remains the single best predictor or writing quality (e.g. Espin et al., 2005; Gregg et al., 2007; Berman & Nir-Sagiv, 2009; Myhill, 2009b; Bae & Bachman, 2010; Crossley et al., 2015). Even studies addressing micro-features report that text length carries more predictive weight than the features focused on in the particular study (Weston et al., 2011; Varner et al., 2013; Uccelli et al., 2013; Dobbs, 2014; Salas et al., 2016 and MacArthur et al., 2019). Length-quality rating correlations of over .5 are common.

As it seems implausible that essay raters judge text quality simply by eyeballing text length, we need to ask which textual virtues are driving the length-quality correlation. In other words, how can writers 'flesh out' their texts? In what follows, we will look at two expansion operations in particular, elaboration and contextualization.

## 2. Background and research questions

The elaboration operation has been conceptualized in various ways and has likewise received various labels. Enright and Quinlan (2010, p. 325) observe: "A certain amount of development is necessary in most written communication, calling for a minimum number of words to convey the writer's intended meaning." According to Myhill (2009a, p. 409), an important aspect of writing development can be summarized as "from declaration to elaboration". Elaboration may take place on various levels: a writer may elaborate on the clause level by adding details in modifiers, on the sentence level by adding subordinate clauses and on the text level by adding sentences. Myhill (2009a) illustrates the first two levels; in this paper, we will concentrate on the text level.

Text level elaborations take different shapes depending on the text genre. Most studies focus on argumentative elaboration. McCann (1989) and Knudson (1992) examine the argumentative sophistication in writing products using the Toulmin model. Crammond (1998) uses the same model to show how expert writers' argumentations differ from children's argumentations (e.g. by using rebuttals and qualifications). Stavans, Seroussi and Zadunaisky Ehrlich (2019) define five stages of text structural sophistication for argumentative essays written by second to fifth graders. In the final stage, the essay contains both 'peripheral' elements (intro and conclusion) and a fully elaborated argumentation involving a claim, argumentative support, as well as a counter-claim. Vilar and Tolchinsky (2022) examine the development of analytical writing going from elementary school through high school to university level, using a three-fold move classification: expository, assertive (bare claims) and argumentative (claim pus support). They show that the two older groups outscore the youngest group in structures combining expository and argumentative elements.

Besides showing clear developmental trends, argumentative elaboration also affects text quality perceptions and persuasiveness. For instance, Ferretti et al. (2000) had two argumentations from by fourth and sixth graders rated for persuasiveness and found that the presence of argumentative elements accounted for 39% and 44% of the rating variance respectively. Ferretti et al. (2009) find that prompting writers to elaborate by providing argumentative subgoals leads to more elaborate and more persuasive text. Similar results are presented by Nussbaum et al. (2005), Goldstein et al. (2009), and Klein et al. (2017).

Crossley and McNamara (2016) gave their participants 15 minutes to add at least two paragraphs to the first draft of their essays, to further illustrate its main idea. Both text versions were then further revised for cohesion by writing experts, thus creating four text versions. Both the elaboration and the cohesion intervention improved text quality ratings. Nevertheless, essays with both elaboration and cohesion were rated better than the original versions with added cohesion, but not better than the elaborated essays without added cohesion; this suggests that the elaboration dimension is crucial.

Elaboration is not a monolithic feature. In argumentative text, one may distinguish between elaboration breadth (how many different arguments are provided on the highest level) and elaboration depth (how many arguments for arguments are provided). The regression analyses of Ferretti et al. (2009) suggest that depth is a stronger predictor than breadth.

Myhill (2009b) presents a descriptive study on the development of paragraphing skills. Her analysis of both personal narratives and persuasive texts of year 8 and year 10 writers shows that text length, number of paragraphs per text, paragraph length and topical unity of paragraphs increase with writing quality, irrespective of genre. Both elaboration breadth and depth seem at issue here, although it remains

unclear to what extent. In an experimental study on descriptive elaboration prompts, Graham et al. (1995) asked pupils to add three pieces of information to a story, more specifically "things that happened, description of things, or details" (o.c. p. 232); this intervention focuses on elaboration depth. They found that the intervention group produced better texts than a control group that was only told to make their texts better. The intervention group did not write longer texts, but made a greater number of meaning-changing revisions than the control group. In a study among 3rd grade writers, Tracy et al. (2009) compared the effects of teaching strategies and knowledge for story writing with those of traditional writing instruction that focuses on spelling and grammar. The stories of strategy-instructed children were more elaborate as they were longer and schematically stronger, i.e. they contained more story grammar elements. The effects of the intervention transferred from writing a story prompted by pictures to writing a personal narrative with a verbal prompt.

Argumentative and descriptive elaborations extend the central message of the text. Another device for expanding the text is contextualization, our term for adding supporting discourse moves that serve to anchor the central message in the communicative situation. Such moves enable the text to function as a self-sustained, interactionally appropriate unit (Pander Maat, 2002). In decontextualized argumentative text, such elements are often restricted to introducing the contentious issue at the start and recapitulating the conclusion at the end (Stavans et al. 2019). For many real-world genres however, there is a larger set of contextualizing moves, some of which have become conventional text components. One subclass contains moves identifying discourse participants. For instance, letters generally need to contain salutations ('dear Mr. Smith'), closing moves ('yours faithfully') and signatures (compare Yunxia (2000) for business letters and Henry and Roseberry (2001) for application letters). Other types of contextualization are politely inviting readers to consider the text (Henry & Roseberry, 2001) or getting their attention by surprising openers (compare Upton (2002) on fundraising letters). Still another subclass of contextualizing moves refers to the situation motivating the writing of the text. In school writing assignments, this situation is often a fictitious one, sketched in the writing assignment. By invoking this situation, the writer may explain why the text is written and what it should accomplish for the reader. The supporting move may also refer to the text itself, i.e. it may announce what the text will be about, or has been about. At the end of the text, it may also be appropriate to conclude with pleasantries, as regularly happens in fundraising letters (Upton, 2002), or to welcome a response and to thank the reader, moves Henry and Roseberry (2001) observe in application letters.

Evidently, many supporting moves are only relevant for writing assignments that specify a rhetorical situation, typically by outlining the audience, the communicative purpose, the topic, and/or recent events the text responds to. This is what writing assignments in educational environments often do, at least when

they can be placed in the 'social practice approach' to writing (Jeffery, 2009). Such assignments often require the writers to produce an exemplar of a particular genre, but do not spell out how this genre generally looks like. Assignments thereby invite a certain degree of role play: the writers need to mentally transport themselves into the scenario, and provide a plausible performance of their part in it (Clark, 2005). We hypothesize that successful 'genre performances' tend to use discourse moves referring to the genre- and scenario-specific rhetorical situation.

There is a fundamental reason why elaboration and contextualization are important for success in written communication. The central challenge for beginning writers is that writing is typically used in 'long-distance' situations in which there is little common ground between the interlocutors, while at the same time written text is a medium which imposes high costs for the process called grounding, i.e. checking addressee understanding. Consider the difference between writing and face-to-face conversation. As Clark and Brennan (1991) have explained, conversation is an extremely effective grounding machine, due to the co-presence, visibility and audibility of the participants as well as to the co-temporality and sequentiality of their contributions. Thus, the conversation context enables speakers and hearers to reduce the total communicative effort by presenting provisional utterances which can be amended by hearers, or present larger units in installments in order to assess hearer reactions along the way. These affordances of conversation lead other authors to dub it the most 'natural' (Kock, 2004) or 'synchronous' medium (Dennis et al., 2008). In contrast, exchanging continuous written texts is a much less straightforward way of communicating. It involves hard work for the participants, especially for the sender (Kock, 2007). As shown by Clark and Brennan (1991), writing does not only involve high costs for communication start-up and speaker change, but also high formulation, fault and repair costs: hence writers do well to fine-tune their utterances before sending them, as misunderstandings and disagreements can better be prevented than remedied.

We posit that contextualizing and elaborating moves are central discourse devices for servicing far-away readers. They pre-emptively respond to reader queries such as 'why are you approaching me?', 'what do you expect me to do with this text?', 'what do you mean by that?', 'can you tell me a bit more?' and 'why should I believe you?' As beginning writers are new to the long-distance communication context, they need to learn to anticipate such questions. This is cognitively demanding, as they now need to add a representation of the reader to the representations of their own thoughts and of their text, and combine the three representations (Kellogg, 2008).

Against this background, our aim in this paper is to decompose the text quality effect of textual expansions in terms of elaboration and contextualization moves. Conceivably, every extra word contributes equally to the prediction of text quality;

in that case, the text length effect would be non-decomposable. But alternatively, various aspects of elaboration and contextualization may contribute differentially. This is our first research issue. Secondly, we probe whether the quality contributions of elaboration and contextualization features differ between text genres (see Olinghouse and Wilson (2013) for a similar study on vocabulary features).

A next aim of our paper is to further explore the generalizable components of elaboration and contextualization. It has been established that writer performance is variable: within individuals, text quality varies to a considerable extent, depending on factors such as topic, genre and assessment procedures (Schoonen, 2005; Bouwer et al., 2015). Hence our third research question is to what extent elaboration and contextualization across tasks are determined by the individual writer. Finally, we examine the extent to which the contribution of this writer-bound variance can be further explained by writer characteristics, specifically from their education level, grade and gender, variables that tend to correlate with writing achievement. Success in prediction supports an interpretation of elaboration and contextualization as the outcomes of general skills in composing.

To address these issues, we annotated two collections of text by adolescent writers for elaboration and supporting moves. The text features thus extracted were used to predict text quality assessments.

## 3.   Method

### 3.1    Writing assignments

We used two writing assignments: one is descriptive, the other one argumentative. Both assignments come from the so-called Schrijfmeterscorpus (De Glopper en Prenger, 2013). Our descriptive task asked the students to write a letter to a Swedish girl who will move to The Netherlands next month, and will then become a classmate. The letter should prepare her for the transition by answering the question 'what is typically Dutch'. The assignment suggests three topic domains to draw from (sports, landscape, food). The student is asked to describe at least four Dutch peculiarities as clearly as possible, and to format the text as a letter. Clearly, this task is set in a situation of low common ground, as the addressee lacks elementary prior knowledge on the text topic.

The argumentative task introduces the following scenario. The school will get funding to 'improve the school building'. The school board has opted for a reconstruction of the gym, as it is old and hardly provides gym equipment. The student does not agree, as he/she thinks that renovating the canteen is a more pressing concern. The student is then asked to write an opinion piece for the school paper in which this view is argued for, directed both at the school board and the school paper readers in general. The writers are provided with a subgoal in that they

are told to give at least three reasons. In this task, there is more common ground in terms of prior knowledge, but less in terms of the stance on the text topic.

As discussed earlier, these assignments invite writers to adopt the perspective of a particular communicative situation, and 'perform' a response to it. Using a different artistic metaphor, one could say that our writers should be able to verbally improvise on a theme, in the sense of smoothly integrating new ideas with the material given in the prompt.

## 3.2    Writers, texts and quality ratings

Our corpus of letters and opinion pieces was collected among 440 students from five Dutch secondary schools in Groningen, Haren, Meppel, Utrecht en Winschoten. Table 1 provides the education and grade levels for the 435 writers of the letters. Education level in the Netherlands corresponds closely to academic achievement: by the end of primary school grade 6, students are referred to different levels of secondary education on the basis of their scores on nation-wide tests of scholastic achievement. Disabled students tend to be in special schools and are not included in the sample. The writers are 12 to 15 years old. Across school and grade levels 50% of the students identified themselves as girls and 48% as boys; data on gender are missing in 2% of the cases. L2 learners make up 4% of the sample.

*Table 1.* Levels of education and grade levels of the writers in the letter task

| Level of education | Grade | | | |
|---|---|---|---|---|
| | 7 | 8 | 9 | Total |
| Vocational secondary | 36 | 42 | 59 | 137 |
| General secondary | 61 | 54 | 44 | 159 |
| Academic secondary | 43 | 47 | 49 | 139 |
| Total | 140 | 143 | 152 | 435 |

Our corpus contained 434 argumentative texts on canteen reconstruction. We annotated and analyzed 418 texts, since, upon closer inspection, 16 of them did not follow the instruction in that they opted for other claims (e.g. spend money both on the gym and the canteen). As we were unsure how this would affect our analysis, we discarded those texts. Table 2 provides school and grade levels for the writers of the argumentative texts. The above remarks on students' disability, gender and L2 hold for this collection as well.

The letters and the argumentative texts were rated holistically, using rating scales with five benchmark or anchor texts (Blok 1985; Schoonen, 2005; Pollmann et al., 2012). The rating scales were constructed by teams of five raters each.

*Table 2.* Levels of education and grade levels of the writers in the argumentative task

| Level of education | Grade | | | |
|---|---|---|---|---|
| | 7 | 8 | 9 | Total |
| Vocational secondary | 33 | 41 | 50 | 124 |
| General secondary | 57 | 53 | 43 | 153 |
| Academic secondary | 43 | 48 | 50 | 141 |
| Total | 133 | 142 | 143 | 418 |

All raters scored a sample of 30 letters and 30 argumentative texts on a five-point scale with the instruction to focus on textual content, global structure, paragraph structure and the use of coherence markers. Those five letters and five argumentative texts were selected for inclusion in the rating scales that were graded with most unanimity and had mean scores that were low, below average, average, above average and high. The benchmark texts exemplified scores of 70, 85, 100, 115 and 130 respectively on a scale that ranged from 50 (i.e. lower than the weakest benchmark) to 150 (higher than the strongest benchmark). Both scales aimed to reflect a quality distribution with a mean of 100 and a standard deviation of 15.

For both assignments, eight trained students from bachelor programs in language and communication were involved in rating the 853 texts. Texts were distributed across individual raters in such a way that each text was rated by three raters. Each possible combination of raters, 56 three-person juries in total, rated an equal number of essays. The raters assigned every text a score between 50 and 150. Before rating, the texts were anonymized, i.e. all information about the writer was deleted. The average reliability (average measures ICC) across 56 juries was 0.85 (SD = .12) for the letters and 0.83 (SD = .12) for the argumentative texts. The mean score for the letters was 103.7 and the standard deviation 14.3. For the argumentative texts a mean score of 95.4 was obtained and a standard deviation of 14.8.

### 3.3 Annotations and automatically derived text features

We coded elaboration and contextualization features for both tasks, using task-specific annotation schemes (Pander Maat et al., 2018). While some moves are task-specific, most move types apply to both assignments. Illustrations of annotations are presented in Appendix A (for the descriptive task) and Appendix B (for the argumentative task).

For the descriptive letter, elaboration breadth and depth are indicated by topic management features, while contextualization is indicated by opening, closing and supporting moves. The following features were used.
- *Elaboration features.* The first feature is the number of different topics dealt with, indicating elaboration breadth. Besides sports, landscape and food, the

students bring up topics such as feast days, cities and the weather. The first sentences of new topics were coded as topic transition sentences. Topics are not labeled; they are identified by topic transitions, and counted by adding one to the number of transitions. Elaboration depth for this task is determined as follows. The number of topic words for each text was calculated by subtracting the words for conventional opening and closing moves and supporting moves (see below) from the text length. The result is our approximation of elaboration depth, to be used in combination with (and thus correcting for) the number of topics.

- *Conventional opening and closing moves outside the body text*. This includes salutations, date and place information, and closing greetings; all these elements are placed above and under the body text.

- *Supporting moves in the body text*. As the literature contains no canonical move structure for assignments such as these, the supporting moves were inductively identified during the annotation of the first quarter of the corpus. Every sentence was separately coded. The seven types of supporting moves are:
  1. Sketching the situation motivating the text (*I heard you will be joining our class next year*).
  2. Introducing the writer (*I am Trisha and I live in Heerenveen, in the North of the country*).
  3. Announcing the topic (*I will tell you some typically Dutch things*).
  4. Welcoming the new classmate (*Welcome to our school!*).
  5. Comparing The Netherlands with Sweden (*Unlike Sweden, The Netherlands does not have lots of snow*); while this move has no fixed place in the text structure and is not realized in a separate utterance, is clearly supports the processing of the letter in relating the information to the background knowledge of the reader.
  6. Recapitulating the letter (*I hope that by now you know how things are over here*).
  7. Closing sentence of the body text (*Hope to see you soon; hope you will have a nice time over here*).

The introductory moves (1-3) mostly appear at the beginning, and the concluding moves (6-7) at the end (6-7). Moves 4 and 5 were found all over the text. Eventually, they were left out of the data analysis, since they have no parallel in the argumentative task and can therefore not contribute to the analyses across genres that we aim for.

Opening, closing and supporting moves reflect the writer's efforts to contextualize his contribution, both in terms of establishing a connection with the reader and facilitating information processing. They were coded sentence by sentence. For every move, the number of words was added to the dataset.[1]

Table 3 in the next section presents an overview of the variables for the descriptive task.

Coding reliability between the first two coders was tested for 315 coding units: 166 move types and 149 topic transitions. For contextualization move types, there were ten labels for utterances or text elements (e.g. 'introducing the writer', or 'greeting'). The text element first needs to be identified as a unit to be coded, and is subsequently labeled. Counting elements only identified by one of the coders as disagreements, the Kappa was .93 (N = 166). For topic transitions, the second coder identified 96% of the transitions seen by the first coder. (Kappa could not be calculated as there were no transitions only seen by the second coder.) The coded text (topic words, supporting words, opening and closing words) covers 96% of the text words.

For the school paper opinion piece, elaboration breadth and depth are indicated by arguments and concessions; contextualization is again indicated by opening, closing and supporting moves. Fifteen task-specific moves were annotated, using sentences as coding units (see the illustration in Appendix B).

- The *main claim* (i.e. the canteen should be renovated, not the gym); this could be present at various places in the text, or presented several times.
- *Primary arguments* directly supporting the main claim. The number of different primary arguments (and concessions, see below) is the main indicator of argumentative breadth, in analogy to the number of topics for the descriptive text. Arguments could focus both on the importance of renovating the canteen and on the relative unimportance of renovating the gym, although the first focus was more common.
- *Secondary arguments* that support primary arguments; more than one secondary argument may be added to one first order argument. We assessed depth of argumentative elaboration in terms of the numbers of words dedicated to secondary arguments
- *Concessions*, i.e. statements acknowledging that there may be reasons to renovate the gym (concessions were typically followed by the more compelling reasons to renovate the canteen). As concessions can be seen as a 'another kind' of argument and are unrelated to the pro-arguments, our argumentative breadth measure added number of concessions (typically only one) to the number of primary arguments.

These features are comparable to, but somewhat less detailed than the annotations used in earlier studies of argumentative writing, such as Crammond (1998) and Ferretti et al. (2000). We forego the finer distinctions offered by the extended Toulmin model, including various kinds of counter-argumentation. However, our annotations do enable us to distinguish between argumentation depth and breadth, and contain an indication of explicit counterargumentation.

- *Conventional opening and closing moves.* Our writers framed the piece in two genres: some adopted letter conventions and provide (1) salutations, (2) dates and (3) greetings; others choose an article frame and provide (4) titles and (5) author names (at the top). As the instructions not explicit about the choice between the two genre frames, we coded all five moves, later collapsing them into one set.
- *Supporting moves in the body text,* which could appear at the beginning (1-3) or at the end (4-7). Again, the supporting moves were collapsed in two sets, introductory moves and concluding moves:
  1. introducing the writer;
  2. sketching the situation motivating the text;
  3. announcing what the text will do;
  4. recapitulating the piece (*these are the reasons I think …*);
  5. persuasive appeals directed at the school board (*hope I have convinced you to …*);
  6. calls to action directed at fellow students (*if you agree, sign our petition*);
  7. general closing sentences (*lots of fun this school year/ thanks for reading this*).

Table 4 in the next section presents an overview of the variables for the argumentative task. Reliability was determined for a sample of 234 moves. Kappa was .75. The majority of the sample consisted of primary and secondary arguments (N = 133). As this distinction is a crucial but subtle one, we separately checked the reliability for this subsample, and found it to be sufficient (.66).

The coded text (claim words, primary argument words, secondary argument words, supporting move words, opening and closing words) covers 87% of the text words. In order to see what kinds of moves were left uncoded, we manually inspected five texts with unusually large proportions of uncoded sentences. One of these texts could be classified as 'off-topic', i.e. as a non-serious, jocular approach of the topic. The other four show various kinds of digressions (e.g. a long stretch of text devoted to the details of a signature campaign for subsidizing a canteen renovation) and more-or-less-relevant passages that did not fit the coding scheme (exhortations such as 'let us spend this money wisely', or 'we personally inspected the gym in order to see whether it needs renovation'). In other words, our coding scheme was fairly restrictive, in the sense of focusing to moves in the 'argumentative core'.

Both corpora were entirely coded by two coders. Disagreements were resolved, if needed in discussion with a third coder.

### 3.4 Descriptive statistics for the text features

Recall that our research aim is to assess whether specific text elements are responsible for the ubiquitous correlation between text length and text quality. For the text features involved, Table 3 and 4 provide their descriptive statistics, along with those for text quality and text length.

Table 3 shows that the letters contain 4 to 5 topics on average. Topic words constitute on average 81% of the text's words. Opening and closing moves tend to be short and formulaic. Supporting moves make larger contributions to text length, with relatively large variances. The texts contain about five times as many topical words as words in supporting sentences.

*Table 3.* Descriptive statistics for the descriptive task (N = 435)

| Variable | M | SD | Min | Max |
|---|---|---|---|---|
| Text quality rating | 103.72 | 14.35 | 61.67 | 146.67 |
| Text length (n of words) | 206.29 | 78.97 | 33.00 | 498.00 |
| *Elaboration* | | | | |
| Breadth (n of topics) | 4.56 | 2.00 | 1.00 | 12.00 |
| Depth (n of topic words) | 167.62 | 70.48 | 7.00 | 442.00 |
| *Contextualization* | | | | |
| Opening & closing moves (n of words) | 5.06 | 2.66 | 0.00 | 21.00 |
| Introductory supporting moves (n of words) | 13.13 | 12.33 | 0.00 | 76.00 |
| Concluding supporting moves (n of words) | 11.59 | 9.85 | 0.00 | 46.00 |
| Supporting moves elsewhere (n of words) | 8.89 | 13.02 | 0.00 | 95.00 |

Table 4 shows that the opinion pieces contain 4 to 5 primary arguments and concessions on average. Words in secondary arguments constitute on average 27% of the text's words. Opening and closing moves are very short in most cases, while the share of supporting moves is more or less similar to the letter task. Outside the scope of our analysis but nevertheless interesting is the number of words in main claims and primary arguments: they make up 13% and 31% of the words in the text, respectively.

### 3.5 Statistical analysis

Our data have a nested or hierarchical character: they concern quality ratings and characteristics of 853 texts (level 1) originating from 435 writers (level 2). For this reason, we used multilevel regression analysis in MLwiN (Rasbash et al., 2000).

*Table 4.* Descriptive statistics for the argumentative task (N = 435)

| Variable | M | SD | Min | Max |
|---|---|---|---|---|
| Text quality rating | 95.47 | 14.90 | 50.00 | 138.67 |
| Text length (n of words) | 167.31 | 74.96 | 11.00 | 477.00 |
| *Elaboration* | | | | |
| Breadth (n of primary arguments and concessions) | 4.77 | 2.17 | 0.00 | 12.00 |
| Depth (n of words in secondary arguments) | 44.77 | 41.63 | 0.00 | 242.00 |
| *Contextualization* | | | | |
| Opening & closing moves (n of words) | 1.75 | 2.76 | 0.00 | 17.00 |
| Introductory supporting moves (n of words) | 12.12 | 10.01 | 0.00 | 76.00 |
| Concluding supporting moves (n of words) | 10.77 | 12.58 | 0.00 | 46.00 |

The scores for text quality and for aspects of elaboration and contextualization are at level 1, along with the variable task. At level 2 we have predictors for school level and year. The assumptions for multilevel regression analysis are met: analysis shows that all cells in a crosstabulation of the independent variables contain more than 5 (expected) observations. Text quality scores are standardized across both tasks. This allows us to estimate effects of the task on text quality ratings. The scores for the predictors are standardized within tasks, which ensures that the predictors are comparable across tasks, despite the differences in the mean length of the letters (206 words) and the opinion texts (167 words).

In the analyses, we first estimate the intercept and variance for text quality at level 1 only (Model 1) in MLwiN. We then determine whether there is also statistically significant variance in text quality at the writer level (Model 2). In Model 3 we add task as a predictor at level 1 (with the letter task as the reference category); with this we check whether there is a difference in average text quality between tasks. In Model 4, at level 1 we add the text features as predictors with the restriction that their regression coefficients are identical. Model 4 corresponds to the idea that it doesn't matter what writers spend their words on, because the effect of text length is monolithic. In Models 5a to 5e, we release these restrictions one by one, for successively elaboration breadth, elaboration depth, opening and closing moves, introductory supporting moves and concluding supporting moves. With these models we start to decompose the effect of text length. In Model 6, the decomposition of the effect of text length is complete: here we freely estimate the regression coefficients for all text features. Finally, in Model 7 we add five interaction terms, one for each combination of task and the text characteristics. This

model explores the generalizability of the relationship between text features and text quality. Because each of the tested models is nested in a previous model, we compare the fit of the models using the -2loglikelihood ratio (Hox et al., 2018). We report the regression coefficients for the best fitting model, the one that is optimal in terms of fit and parsimony.

In a next step we estimate three subsequent models in which elaboration breadth, elaboration depth, opening and closing moves and introductory supporting moves and concluding supporting moves are dependent variables. Their intercepts and variances are estimated in Model 1 at the text level, and in Model 2 at the text and writer level, and predicted in Model 3 by writer characteristics: school level (with dummy variables for general secondary and academic secondary education), grade (with dummy variables for grade 8 and grade 9) and gender (with male as the reference category). The fit of the models is assessed in the manner described directly above. We again report the regression coefficients for the best fitting model.

## 4. Results

### 4.1 Predictive contributions of elaboration and contextualization features

Table 5 below examines seven ways of modeling the contributions of elaboration and contextualization to text quality ratings. Here and below, we refer to the different models by their number and by abbreviating their distinctive features. All abbreviations are explained below Table 1. Model 2 [TL-WL] shows that adding the writer level to the text level improves the model: text quality is associated with individual skills. Model 3 [TL-WL-T] is an improvement over Model 2 [TL-WL] and indicates that the level of text quality varies between tasks. Model 4 [TL-WL-T-EQ-TF], which introduces the five text features with equality constraints, shows a major improvement of fit compared to Model 3 [TL-WL-T]. Model 4 [TL-WL-T-EQ-TF] can be further improved upon by releasing these constraints for elaboration breadth, elaboration depth and opening and closing moves in Models 5a [TL-WL-T-EB], 5b [TL-WL-T-ED], and 5c [TL-WL-T-OC] respectively. This does not hold for Models 5d [TL-WL-T-IS] and 5e [TL-WL-T-CS]. In Model 6 [TL-WL-T-EQ-IS-CS] the equality constraint is therefore retained for introductory and concluding supporting moves. When compared to Model 4 [TL-WL-T-EQ-TF], Model 6 [TL-WL-T-EQ-IS-CS] demonstrates superior fit. (An alternative version of Model 6 without any equality constraints is no improvement over the model reported in the table ($X^2 = 0.007$; df = 1; p = .933). The final Model 7 [TL-WL-T-T*TF], with five interaction terms, one for each combination of task and the text characteristics, gives another improvement, now over Model 6.

Table 6 reports the parameter estimates for Models 1 [TL], 2 [TL-WL], 3 [TL-WL-T] and 7 [TL-WL-T-T*TF] and the variances Model 7 explains at the text and writer

levels, when compared to the two-level model where task is the only predictor of text quality (Model 3 [TL-WL-T]). According to Model 7 [TL-WL-T-T*TF], task and all aspects of elaboration and contextualization contribute to the prediction of text quality. Task has a negative coefficient, which corresponds to what the descriptive statistics in Tables 3 and 4 already showed: the mean quality rating for the letter is higher than the mean for the opinion piece. Elaboration breadth is slightly negative in the letter task. Its interaction effect with task is positive and large, which indicates that breadth is not rewarded in the letter task, but is important in the argument task. Elaboration depth has the largest effect; it is positive and it does, interestingly, not interact with task. The coefficient for opening and closing moves is positive and relatively small; its interaction with task is not significant. The two types of supporting moves show positive contributions of comparable sizes. For concluding supporting moves the main effect is qualified by a negative interaction with task which indicates that this move type is more important in the letter task. Finally, in comparison to Model 3 [TL-WL-T], Model 7 [TL-WL-T-T*TF] explains 28% of the variance at the text level and no less than 90% of the variance at the writer level. This striking discrepancy reflects the fact that the main effects in the model – the predictors that are general across tasks – carry the bulk of the predictive power.

## 4.2    Prediction of skill in elaboration and contextualization

We now turn to the prediction of the elaboration and contextualization features themselves, across tasks; see Table 7. Again, we refer to the different models by their number and by abbreviating their distinctive features. First, we note that significant proportions of variance are associated with the writer level, since Model 2 [TL-WL] fits better than Model 1 [TL] for all aspects of elaboration and contextualization. In Model 2 [TL-WL] the second level variance makes up 17% of the total variance for elaboration breadth, 51% for elaboration depth, 11% for opening and closing moves and for introductory supporting moves, and 20% for concluding supporting moves. Clearly, the amount of writer-bound variance in elaboration and contextualization is quite variable. It is interesting to see that it is largest for elaboration depth, the most important predictor of text quality.

*Table 5.* Fit of successive multilevel regression models for the prediction of text quality

| Model | TL | WL | I | T | EB | ED | OC | IS | CS | T*EB | T*ED | T*OC | T*IS | T*CS | Deviance | MC | ΔX2 | Δdf | *p* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 TL | + | - | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2419.709 | - | - | - | - |
| 2 TL-WL | + | + | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2309.844 | 2 vs. 1 | 109.865 | 1 | <.001 |
| 3 TL-WL-T | + | + | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2156.472 | 3 vs. 2 | 153.372 | 1 | <.001 |
| 4 TL-WL-T-EQ-TF | + | + | 1 | 2 | 3 | 3 | 3 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 1665.886 | 4 vs. 3 | 490.586 | 1 | <.001 |
| 5a TL-WL-T-EB | + | + | 1 | 2 | 4 | 3 | 3 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 1615.110 | 5a vs. 4 | 50.776 | 1 | <.001 |
| 5b TL-WL-T-ED | + | + | 1 | 2 | 3 | 4 | 3 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 1541.112 | 5b vs. 4 | 124.774 | 1 | <.001 |
| 5c TL-WL-T-OC | + | + | 1 | 2 | 3 | 3 | 4 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 1656.938 | 5c vs. 4 | 8.948 | 1 | .003 |
| 5d TL-WL-T-IS | + | + | 1 | 2 | 3 | 3 | 3 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 1665.729 | 5d vs. 4 | 0.157 | 1 | .692 |
| 5e TL-WL-T-CS | + | + | 1 | 2 | 3 | 3 | 3 | 3 | 4 | 0 | 0 | 0 | 0 | 0 | 1665.814 | 5e vs. 4 | 0.072 | 1 | .788 |
| 6 TL-WL-T-EQ-IS-CS | + | + | 1 | 2 | 3 | 4 | 5 | 6 | 6 | 0 | 0 | 0 | 0 | 0 | 1500.667 | 6 vs. 4 | 165.219 | 3 | <.001 |
| 7 TL-WL-T-T*TF | + | + | 1 | 2 | 3 | 4 | 5 | 6 | 6 | 7 | 8 | 9 | 10 | 11 | 1413.764 | 7 vs. 6 | 86.903 | 5 | <.001 |

*Note.* TL = text level; WL = writer level; I = intercept; T = task; EQ = equality constraint; TF = all five text features; EB = elaboration breadth; ED = elaboration depth; OC = opening & closing moves; IS = introductory supporting moves; CS = concluding supporting moves; T* = interactions with task; Deviance = -2 loglikelihood; MC = model comparison; +/- = level present or absent; 1-12 = indicators of predictors with distinct values

*Table 6.* Parameters estimates for successive multilevel regression models for the prediction of text quality (SE's between brackets)

| Model | I | T | EB | ED | OC | IS | CS | T*EB | T*ED | T*OC | T*IS | T*CS | Var. TL | Var. WL | $R^2$ TL (3) | $R^2$ WL (3) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 TL | -0.000ns | | | | | | | | | | | | 0.999 | | | |
| | (0.034) | | | | | | | | | | | | (0.048) | | | |
| 2 TL-WL | 0.000ns | | | | | | | | | | | | 0.519 | 0.477 | | |
| | (0.041) | | | | | | | | | | | | (0.036) | (0.054) | | |
| 3 TL-WL-T | 0.274*** | -0.565*** | | | | | | | | | | | 0.357 | 0.564 | | |
| | (0.046) | (0.041) | | | | | | | | | | | (0.025) | (0.052) | | |
| 7 TL-WL-T-T*TF | 0.271*** | -0.552*** | -0.112*** | 0.516*** | 0.137*** | 0.236*** | 0.236*** | 0.359*** | -0.045ns | -0.016ns | -0.001ns | -0.088** | 0.256 | 0.056 | .282 | .901 |
| | (0.027) | (0.035) | (0.027) | (0.028) | (0.027) | (0.018) | (0.018) | (0.038) | (0.040) | (0.039) | (0.033) | (0.033) | (0.018) | (0.016) | | |

*Note.* TL = text level; WL = writer level; I = intercept; T = task; TF = all five text features; EB = elaboration breadth; ED = elaboration depth; OC = opening & closing moves; IS = introductory supporting moves; CS = concluding supporting moves; T* = interactions with task; Var. TL = variance at text level; Var. WL = variance at writer level; $R^2$ TL (3) = explained variance at text level in comparison to Model 3; $R^2$ WL (3) = explained variance at writer level in comparison to Model 3; ** $p < .01$; *** $p < .001$

In Model 3 [TL-WL-WC] the writer characteristics level of education, grade level and gender are added as predictors. These three writer characteristics explain 20% of the writer level variance for elaboration breadth, 51% for elaboration depth, 100% for opening and closing moves, 61% for introductory supporting moves, and 46% for concluding supporting moves. We take these outcomes as support for our interpretation of aspects of elaboration and contextualization as outcomes of skills in composing.

Of the two dummy variables for education level, the one for academic secondary education consistently predicts the writer level variance in the text characteristics. The effect of general secondary education, the middle level, does not come through consistently. But taken together it is clear that level of education has an impact on elaboration and contextualization. The same holds for grade level. Here the pattern is similar, with mostly clearcut effects for grade 9. Gender contributes to the prediction of elaboration depth and concluding supporting moves.

*Table 7.* Prediction of skill in elaboration and contextualization: model comparison and parameter estimates (SE's between brackets)

| Model fit | EB | ED | OC | IS | CS |
|---|---|---|---|---|---|
| **Model 1 TL** | | | | | |
| Deviance | 2418.707 | 2418.707 | 2418.707 | 2418.707 | 2418.707 |
| **Model 2 TL-WL** | | | | | |
| Deviance 2 vs. 1 | 2406.720 | 2290.054 | 2414.260 | 2413.211 | 2401.522 |
| $X^2$ 2 vs. 1 | 11.987 | 128.653 | 4.447 | 5.496 | 17.185 |
| df | 1 | 1 | 1 | 1 | 1 |
| P | <.001 | <.001 | 0.035 | 0.019 | <.001 |
| **Model 3 TL-WL-WC** | | | | | |
| Deviance | 2340.481 | 2079.913 | 2255.069 | 2321.802 | 2292.628 |
| $X^2$ 3 vs. 2 | 66.239 | 210.141 | 159.191 | 91.409 | 108.894 |
| df 3 vs. 2 | 1 | 1 | 1 | 1 | 1 |
| p | <.001 | <.001 | <.001 | <.001 | <.001 |
| Parameter estimates | EB | ED | OC | IS | CS |
| **Model 1 TL** | | | | | |
| Var. TL | 0.998 (0.048) | 0.998 (0.048) | 0.998 (0.048) | 0.998 (0.048) | 0.998 (0.048) |
| **Model 2 TL-WL** | | | | | |
| Var. TL | 0.830 (0.057) | 0.483 (0.033) | 0.893 (0.062) | 0.884 (0.061) | 0.799 (0.055) |
| Var. WL | 0.167 (0.050) | 0.514 (0.055) | 0.105 (0.049) | 0.113 (0.049) | 0.198 (0.050) |

| Model 3 TL-WL-WC | | | | | |
|---|---|---|---|---|---|
| Var. TL | 0.826 (0.058) | 0.487 (0.034) | 0.858 (0.042) | 0.890 (0.062) | 0.797 (0.056) |
| Var. WL | 0.133 (0.048) | 0.252 (0.039) | 0.000 (0.000) | 0.040 (0.046) | 0.107 (0.045) |
| $R^2$ WL (2) | .204 | .510 | 1.00 | .646 | .460 |
| | | | | | |
| **Model 3 TL-WL-WC** | | | | | |
| Intercept | -0.126 | -0.697 | -0.564 | -0.482 | -0.693 |
| | $(0.095)^{ns}$ | $(0.091)^{***}$ | $(0.085)^{***}$ | $(0.090)^{***}$ | $(0.092)^{***}$ |
| General secondary | -0.091 | 0.159 | 0.070 | 0.198 | 0.268 |
| | $(0.088)^{ns}$ | $(0.084)^{***}$ | $(0.078)^{ns}$ | $(0.083)^{*}$ | $(0.085)^{**}$ |
| Academic secondary | 0.279 | 1.035 | 0.562 | 0.522 | 0.539 |
| | $(0.091)^{**}$ | $(0.086)^{***}$ | $(0.080)^{***}$ | $(0.085)^{***}$ | $(0.087)^{***}$ |
| Grade 8 | -0.005 | 0.047 | 0.380 | 0.124 | 0.338 |
| | $(0.090)^{ns}$ | $(0.085)^{ns}$ | $(0.080)^{***}$ | $(0.085)^{ns}$ | $(0.086)^{***}$ |
| Grade 9 | -0.004 | 0.335 | 0.481 | 0.438 | 0.472 |
| | $(0.089)^{ns}$ | $(0.085)^{***}$ | $(0.079)^{***}$ | $(0.084)^{***}$ | $(0.086)^{***}$ |
| Gender | 0.140 | 0.315 | 0.098 | 0.090 | 0.282 |
| | $(0.072)^{ns}$ | $(0.069)^{***}$ | $(0.064)^{ns}$ | $(0.068)^{ns}$ | $(0.069)^{***}$ |

*Note.* TL = text level; WL = writer level; WC = writer characteristics; EB = elaboration breadth; ED = elaboration depth; OC = opening & closing moves; IS = introductory supporting moves; CS = concluding supporting moves; Deviance = -2loglikelihood; Var. TL = variance at text level; Var. WL = variance at writer level; $R^2$ WL (2) = explained variance at writer level in comparison to Model 2; $^{*}$ p < .05; $^{**}$ p < .01; $^{***}$ p < .001

## 5. Conclusions and discussion

Our analysis has explored two devices of expanding a text: elaboration and contextualization. By annotating texts for two elaboration and three contextualization features and linking these features to text quality, we have attempted to decompose the ubiquitous link between text quality and text length. Our analysis addresses four questions.

1. To what extent do the five features differentially contribute to text quality?
2. To what extent do these contributions differ by text genre?
3. To what extent are the features linked to individual writers, i.e. share variance across tasks?
4. To what extent can this writer-bound variance be further explained by writer characteristics, specifically from their level of education, grade and gender?

Table 8 summarizes our results.

*Table 8.* Results summary

| Research question | Elaboration breadth | Elaboration depth | Opening & closing moves | Initial support moves | Concluding support moves |
|---|---|---|---|---|---|
| Contributes to quality across genres* | - | +++ | + | ++ | ++ |
| Contribution is genre-specific** | ++ | - | - | - | + |
| Is linked to individual writers*** | + | ++ | - | - | + |
| Writer characteristics explain variance | + | ++ | ++ | ++ | ++ |

*Note. On this line and the next one b's > .5 are considered large (+++), b > .2 substantial (++).
**Note. The elaboration breadth interaction is substantial and cancels the main effect, the concluding support moves interaction qualifies it.
***Note. On this line and the next one, explained variance proportions of over 15% get a single +, proportions of over 40% get ++.

As regards elaboration, depth is clearly more important to text quality than breadth. It strongly contributes across genres, whereas breadth only contributes in the argumentative context: i.e. having more arguments is a virtue, while discussing more topics is immaterial to descriptive text quality. Depth is also interesting in that it shows a large writer-bound component across tasks, which is also substantially related to factors such as school type, grade and gender. Of course, it remains unclear whether these background factors themselves point to writer maturation, or to the effect of instruction variables, or both. The point here is that elaboration depth is a consistent factor across tasks, both in its text quality contributions and its link to individual writers.

In contrast, elaboration breadth does not contribute to text quality across tasks, and shows a weaker link to individual writers. These two results may be two sides of the same coin, in that elaboration breadths need to be adjusted to writing tasks, and hence can be expected to show less writer-bound variance.

With regard to contextualization, our analysis distinguishes between peripheral elements (opening and closing moves outside the body text) and support moves integrated in the text. Extra-textual contextualization plays a modest role for text quality. It is not a very consistent component within individual writers; but where it is, the writer variance is strongly explained by school type and grade level.

For intra-textual support, we need to distinguish initial and concluding moves. Initial moves strongly contribute to quality across genres; concluding moves also contribute in both genres, but more so in argumentative text. Concluding moves have more writer variance; but for both kinds of support moves, the writer variance is substantially explained by writer characteristics.

Our study helps to further develop our notion of writing well: expanding your text in ways that strengthen your message as well as serve your readers. In the introduction, we have related elaboration depth with reader awareness, i.e. the ability to anticipate and address the needs of readers with different prior knowledge levels (for the descriptive task) or opinions (for the argumentative task). The effectiveness of encouraging writers to elaborate seems to indicate that such 'perspective taking' has to be learned; this is understandable, given the central writing challenge of communicating in a low-common-ground-situation while using a medium with high grounding costs. In reviews, such elaboration instructions have been called 'setting product goals' (e.g. Graham & Perin, 2007). Our own writing assignments do encourage writers to present several topics and arguments, i.e. to aim for elaboration breadth; but they do not talk about elaboration depth. This may help explain why there is more writer variance in elaboration depth than in elaboration breadth.

But importantly, our data also learn that elaboration is not a universal writing virtue; its contribution to text quality depends on whether there is synergy between the elements added to the text: elaboration breadth may weaken the focus in descriptive context by adding new topics, while in argumentative contexts adding new arguments tend to strengthen your point.

The other main expansion operation in our study is contextualization: relating your message to the communicative situation it springs from. This seems especially important when dealing with writing assignments presenting specific rhetorical situations. It has been pointed out that writing assignment are role-playing games, requiring writers to actively construct rhetorical persona's (Clark, 2005). This goes for all writing prompts, including essay prompts only providing a topic. But it holds even more for our contextually rich assignments. They especially call for support moves, i.e. moves in which writers show that they have absorbed the contextual information, and actually place themselves in the rhetorical situation. As these moves carry more information than opening and closing moves, it is only natural that their contribution to text quality is stronger.

To be sure, this work has its limitations. For each task, we present only one specimen, in which genre properties (letter vs. school paper article) are conflated with task type (description vs. argumentation). This raises the question whether our task effects can be generalized, and whether our text annotation system is re-usable for other descriptive and argumentative writing tasks. Further ahead, there is the issue of manual annotation as such: we have shown that text structure is crucial to text quality, but to what extent can the annotation of text structure be automated in order to make it more efficient?

In this study, elaboration and contextualization emerge as crucial components of text quality. We suggest to study them further, in various ways. An obvious next step

is to study the operations in a third important type of writing task: narrative. Conceivably, elaboration is important for narrative quality, but its definition is not self-evident in this context: does it involve fleshing out the 'actions', the descriptions, or the characters?

Another angle is to look at elaboration and contextualization as topics of writing instruction. For instance, goal setting is an important part of the Self-regulated Strategy Development approach to writing instruction (Harris & Graham, 2009). At numerous occasions, the students graph the structure of example texts or their own texts (e.g. premise – three reasons – conclusion). It remains to be seen how much of the learning gains made in an SRSD cycle can be attributed to the goal setting component, as opposed to other SRSD activities such as exploring your current self-regulation abilities, teacher modeling of self-regulation strategies and their memorization. As discussed earlier, the effectiveness of short-term goal setting interventions suggests that goal setting may be a considerable factor all by itself. We know of no study on goal setting prompts referring to contextualizing moves. This may be understandable as such moves are primarily relevant for genres situated in particular 'real-life' writing scenarios with little common ground.

We also ask whether beginning writers always need to be told to elaborate and contextualize. Perhaps it is preferable to use other means to heighten their reader awareness. For instance, Traxler and Gernsbacher (1993) had writers prepare for instruction writing by first carrying out the task themselves. This led to better instructions; presumably it helped them see how readers would process their text (see Holliway & McCutchen, 2004 for similar results). In another 'user experience' intervention, Cho and Cho (2011) asked undergraduate students to comment on texts written by peers, while the control group just read those texts; the commenting group afterwards wrote better texts themselves. Further work on reader awareness is needed to see what text features specifically profit from it, but it seems plausible that elaboration is among these.

Finally, we would welcome work on where elaboration and contextualization moves are produced in the writing process. Kellogg (2008, p. 10) points out that "the capacity to see the text from the perspective of the reviewer can be put to use during the composition of a first draft rather than delayed until revising an initial effort". As an example, it would be interesting to look at the writing process of high-contextualizing and high-elaborating writers. To what extent do they produce larger, elaborated units (roughly, paragraphs) in a single or a series of P-bursts, and to what extent do their elaborations arise from reviewing text and finding it insufficiently fleshed out? Such process data may help us to better understand how writers learn to solve their central problem: reaching their readers.

## Note

[1] For the letters and the opinion pieces we used number of words rather than number of moves as the operationalization of opening & closing moves and supporting moves. In multilevel regression analyses (see the specification in section 3.4 below) number of words is a stronger predictor of text quality than number of moves for both opening & closing moves, introductory supporting moves and concluding supporting moves. Also, models that predict text quality from task and number of words improve only marginally or not at all (for introductory supporting moves) when number of moves is added to the model: $R^2$ increases from 0.154 to 0.167 for opening & closing moves, and from 0.282 to 0.285 for concluding supporting moves.

## References

Bae, J., & Bachman, L. F. (2010). An investigation of four writing traits and two tasks across two languages. *Language Testing, 27*(2), 213-234. doi: 10.1177/0265532209349470

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing, 22*(2), 185-200. doi: 10.1007/s11145-007-9107-5

Berman, R., & Nir-Sagiv, B. (2009). Cognitive and linguistic factors in evaluating text quality: Global versus local. In Vyvyan Evans & Stéphanie Pourcel (eds.), *New directions in cognitive linguistics*, 421–440. Amsterdam: John Benjamins. doi: 10.1075/hcp.24.26ber

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement*, 22, 41–52. doi: 10.1111/j.1745-3984.1985.tb01048.x

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32*(1), 83-100.
doi: 10.1177/0265532214542994

Chen, J., Zhang, M., & Bejar, I. I. (2017). An Investigation of the e-rater® Automated Scoring Engine's Grammar, Usage, Mechanics, and Style Microfeatures and Their Aggregation Model. *ETS Research Report Series, 2017*(1), 1-14. doi: 10.1002/ets2.12131

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science, 39*(5), 629-643. doi: 10.1016/j.learninstruc.2009.08.006

Clark, I. (2005). A genre approach to writing assignments. *Composition Forum, 14*(2). https://compositionforum.com/issue/14.2/clark-genre-writing.php

Clark, H.H., & Brennan S.E. (1991) Grounding in Communication. In L. Resnick, J. Levine & S.Teasley (Eds.), *Perspectives on Socially Shared Cognition* (127-149). Hyattsville, MD: American Psychological Association. Doi: 10.1037/10096-006

Crammond, J. G. (1998). The uses and complexity of argument structures in expert and student persuasive writing. *Written Communication, 15*(2), 230-268.
doi: 10.1177/0741088398015002004

Crossley, S. A., & McNamara, D. S. (2016). Say More and Be More Coherent: How Text Elaboration and Cohesion Can Increase Writing Quality. *Grantee Submission, 7*(3), 351-370. doi: 10.17239/jowr-2016.07.3.02

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment, 8*. http://journalofwritingassessment.org/article.php?article=80.

Dennis, A. R., Fuller, R. M., & Valacich, J. S. (2008). Media, tasks, and communication processes: A theory of media synchronicity. *MIS quarterly, 32*(3), 575-600. doi: 10.2307/25148857

Dobbs, C. L. (2014). Signaling organization and stance: Academic language use in middle grade persuasive writing. *Reading and Writing, 27*(8), 1327-1352. doi: 10.1007/s11145-013-9489-5

Donovan, C. A., & Smolkin, L. B. (2002). Children's genre knowledge: An examination of K-5 students' performance on multiple tasks providing differing levels of scaffolding. *Reading Research Quarterly*, 37(4), 428-465. Doi: 10.1598/rrq.37.4.5

Donovan, C.A. (2001). Children's development and control of written story and informational genres: Insights from one elementary school. *Research in the Teaching of English, 35*(3), 452-497.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing, 27*(3), 317-334. doi: 10.1177/0265532210363144.

Espin, C. A., De La Paz, S., Scierka, B. J., & Roelofs, L. (2005). The relationship between curriculum-based measures in written expression and quality and completeness of expository writing for middle school students. *The Journal of Special Education, 38*(4), 208-217. doi: 10.1177/00224669050380040201

Ferretti, R. P., Lewis, W. E., & Andrews-Weckerly, S. (2009). Do goals affect the structure of students' argumentative writing strategies?. *Journal of Educational Psychology*, 101(3), 577-589. doi: 10.1037/a0014702

Ferretti, R. P., MacArthur, C. A., & Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology, 92*(4), 694. doi: 10.1037/0022-0663.92.4.694

Glopper, K. de, & Prenger, J. (2013). Schrijfmeters maken [Making the Schrijfmeters Corpus]. In A. Mottart & S. Vanhooren (Red.), *Zevenentwintigste conferentie Onderwijs Nederlands* (pp. 86–89). Gent: Academia Press.

Goldstein, M., Crowell, A., & Kuhn, D. (2009). What constitutes skilled argumentation and how does it develop? *Informal Logic*, 29, 379–395. Doi: 10.22329/il.v29i4.2905

Graham, S., MacArthur, C., & Schwartz, S. (1995). Effects of goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology*, 87(2), 230. doi: 10.1037//0022-0663.87.2.230

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of educational psychology, 99*(3), 445-476. doi: 10.1037/0022-0663.99.3.445

Gregg, N., Coleman, C., Davis, M., & Chalk, J. C. (2007). Timed essay writing: Implications for high-stakes tests. *Journal of Learning Disabilities, 40*(4), 306-318.
doi: 10.1177/00222194070400040201

Harris, K. R., & Graham, S. (2009). Self-regulated strategy development in writing: Premises, evolution, and the future. *British Journal of Educational Psychology, 6,* 113-135. Doi: 10.1348/978185409X422542

Henry, A., & Roseberry, R. L. (2001). A narrow-angled corpus analysis of moves and strategies of the genre: 'Letter of Application'. *English for Specific Purposes, 20*(2), 153-167. doi: 10.1016/s0889-4906(99)00037-x

Holliway, D. R., & McCutchen, D. (2004). Audience perspective in young writers' composing and revision: Reading as the reader. In L. Allal, L. Chanquoy, & P. Largy (), *Studies in Writing: Vol. 13. Revision: Cognitive and instructional processes* (pp. 87-101). Amsterdam, the Netherlands: Kluwer. doi: 10.1007/978-94-007-1048-1_6

Hox, J.J., Moerbeek, M., & Schoot, R. van de (2018). *Multilevel analysis. Techniques and applications*. New York: Routledge. Doi: 10.4324/9781315650982

Jeffery, J. V. (2009). Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing, 14*(1), 3-24.
doi: 10.1016/j.asw.2008.12.002

Kamberelis, G. (1999). Genre development and learning: Children writing stories, science reports, and poems. *Research in the Teaching of English, 33*, 403-463.

Kellogg, R.T. ( 2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, 1, 1-26. doi: 10.17239/jowr-2008.01.01.1

Klein, P.D., Haug, K.N. & Arcon, N. (2017). The effects of rhetorical and content subgoals on writing and learning, *The Journal of Experimental Education, 85*(2), 291-308. doi: 10.1080/00220973.2016.1143795

Knudson, R. E. (1992). Analysis of argumentative writing at two grade levels. *The Journal of Educational Research, 85*(3), 169-179. doi: 10.1080/00220671.1992.9944434

Kock, N. (2004). The psychobiological model: Towards a new theory of computer-mediated communication based on Darwinian evolution. *Organization science, 15*(3), 327-348. doi: 10.1287/orsc.1040.0071

Kock, N. (2007). Media naturalness and compensatory encoding: The burden of electronic media obstacles is on senders. *Decision Support Systems, 44*(1), 175-187. doi: 10.1016/j.dss.2007.03.011

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction?. *Reading and Writing, 32*(6), 1553-1574.
doi: 10.1016/j.asw.2008.12.002

McCann, T. M. (1989). Student argumentative writing knowledge and ability at three grade levels. *Research in the Teaching of English, 23*(1), 62-76.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86. doi: 10.1177/0741088309351547

Myhill, D. (2008). Towards a linguistic model of sentence development in writing. *Language and Education, 22*(5), 271-288. doi: 10.2167/le775.0

Myhill, D. (2009a). Becoming a designer: Trajectories of linguistic development. In Bear, R., Myhill, D., Riley, J., Nystrand, M. *The Sage Handbook of Writing Development*. (pp.402-13). London: Sage Publications. Doi: 10.4135/9780857021069.n28

Myhill, D. (2009b). Developmental trajectories in mastery of paragraphing: Towards a model of development. *Written Language & Literacy, 12*(1), 26-51. doi: 10.1075/wll.12.1.02myh

Nussbaum, E. M., Kardash, C. M., & Graham, S. E. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology, 97*(2), 157. doi: 10.1037/0022-0663.97.2.157

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing, 26*(1), 45-65. doi: 10.1007/s11145-012-9392-5

Pander Maat, H. (2002). *Tekstanalyse. Wat teksten tot teksten maakt* [Text analysis. What makes a text a text]. Coutinho, Bussum.

Pander Maat, H., Raaijmakers, K., Veerbeek, J. & Vermeulen, D. (2018). *Handleiding Utrechts systeem voor annotatie van leerdertekst*. [Manual Utrecht System for Annotating Learner Text]. Internal publication Utrecht University.

Pollmann, E., Prenger, J., & Glopper, K. de (2012). Het beoordelen van leerlingteksten met behulp van een schaalmodel [Rating learner texts using a scale model]. *Levende Talen Tijdschrift, 13*(3), 15-24.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M. et al. (2000). *A user's guide to MLwiN*. London: Multilevel Models Project Institute of Education, University of London.

Salas, N., Llauradó, A., Castillo, C., Taulé, M., & Martí, M. A. (2016). Linguistic correlates of text quality from childhood to adulthood. In J. Perera, M. Aparici, E. Rosado & N. Salas (eds.), *Written and spoken language development across the lifespan. Essays in honour of Liliana Tolchinsky*. 307- 326. New York: Springer International Publishing. doi: 10.1007/978-3-319-21136-7_18

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing, 22*(1), 1–30. doi: 10.1191/0265532205lt295oa

Stavans, A., Seroussi, B., & Zadunaisky Ehrlich, S. (2019). Literacy-related abilities' effects on argumentative text quality structure. *Journal of Literacy Research, 51*(3), 315-335. Doi: 10.1177/1086296x19859515

Stein, N. L. & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In Freedle, R. O. (ed.), *New directions in discourse processing*, 53–120. Norwood, NJ: Ablex.

Tracy, B., Reid, R., & Graham, S. (2009). Teaching young students strategies for planning and drafting stories: The impact of self-regulated strategy development. *The Journal of Educational Research, 102*(5), 323-331. Doi: 10.3200/joer.102.5.323-332

Traxler, M. J., & Gernsbacher, M. A. (1993). Improving written communication through perspective-taking. *Language and Cognitive Processes, 8*(3), 311-334. doi: 10.1080/01690969308406958

Uccelli, P., Dobbs, C. L., & Scott, J. (2013). Mastering academic language: Organization and stance in the persuasive writing of high school students. *Written Communication, 30*(1), 36-62. doi: 10.1177/0741088312469013

Upton, T. A. (2002). Understanding direct mail letters as a genre. *International Journal of Corpus Linguistics, 7*(1), 65-85. doi: 10.1075/ijcl.7.1.04upt

Varner, L. K., Roscoe, R. D., & McNamara, D. S. (2013). Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis. *Journal of Writing Research, 5*(1). doi: 10.17239/jowr-2013.05.01.2

Vilar, H., & Tolchinsky, L. (2022). The rhetorical structure of analytical writing: a developmental approach. *Text & Talk*, 42(1), 131-152. Doi: 10.1515/text-2019-0228

Weston, J. L., Crossley, S. A., McCarthy, P. M., & McNamara, D. S. (2011). Number of words versus number of ideas: Finding a better predictor of writing quality. In R. C. Murray, & P. M. McCarthy (Eds.), Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference(pp. 335–340). Menlo Park, CA: AAAI Press.

Yunxia, Z. (2000). Structural moves reflected in English and Chinese sales letters. *Discourse Studies, 2*(4), 473-496. doi: 10.1177/1461445600002004004

**Appendix A. Descriptive letter, annotated**
Spelling errors in the Dutch original text are approximately retained in the corresponding English words

| Dutch text | English gloss | Annotation legend |
|---|---|---|
| [Utrecht, 26 november 2012]PD<br>[Beste Annika,]S | [Utrecht, November 26th 2012]PD<br>[Dear Annika,]S | PD: place and date<br>S: salutation |
| [Gisteren is aan mij verteld dat ik aan jou een brief moest schrijven over typisch Nederlandse dingen, leuk!]STR<br>[In deze brief zal ik dus proberen wat over Nederland te vertellen.]AN Hopelijk zul je alles begrijpen!<br>Nederland is een plat land, zonder reliëf.<br>De wind waait er het hele jaar door en met regen zijn wij Nederlanders zeer vertrouwd. | [Yesterday I was told to write you a letter about typically Dutch things, nice!]STR [So in this letter I will try to tell something about the Netherlands.]AN Hopefully, you will understand everything!<br>The Netherlands is a flat country, without height differences. The wind blows all year long and being Dutch we are very familiar with rain. | STR = situation the text responds to<br>AN: announcing the text topic<br><br>Topic 1 ('flat country')<br>Topic 2 ('weather') |

| | | |
|---|---|---|
| Dat het Nederlandse weer niet vaak aangenaam is, is een minpuntje, maar je zult in niet veel landen zulke gezellige en knusse markten vinden zoals wij die hebben! Zeker rond de Kerst is het daar erg aangenaam.<br>In ons land hebben we, [vast ook zoals in jouw land]CNS, bepaalde lekkernijen, zoals stroopwafels, pannenkoeken en haring! | That the Dutch weather is not often agreeable, is a small disadvantage, but in many countries you won't find such cozy marketplaces like we have them.<br>Especially in around Christmas they are very nice.<br>In our country, [probably like in yours]CNS, we have certain delicacies, such as 'stroopwafels', pancakes and herring! | Topic 3 ('marketplaces')<br><br>Topic 4 ('delicacies')<br>CNS = comparing The Netherlands with Sweden |
| Bijna elke Nederlander vind dit wel lekker.<br>Bovendien eten wij bij Sinterklaas (een feest in december) pepernoten.<br>Dat zijn een soort kleine koekjes waaraan je je, als je niet oppast, helemaal miselijk kunt eten!<br>Naast veel eten kunnen Nederlandes ook goed sporten en er met name goed naar kijken.<br>Nu heb ik het over voetbal, dé sport in Nederland. Als je met een Nederlander een gesprekje wilt | Almost every Dutchman like these.<br>Furthermore, for 'Sinterklaas' (a festivity in December) we eat spice nuts. Those are a kind of small biscuits, which you will eat until you are very sik when you don't watch yourself!<br>Apart from eating much the [Duth are also good at sports and especially in watching sports.<br>I am referring to football here, which is the nr. 1 sport in The Netherlands. When you want to have a conversation with a | Topic 5 ('Sinterklaas')<br><br><br>Topic 6 ('sports') |

| | | |
|---|---|---|
| houden, kun je het daar altijd over hebben!<br><br>[Ik hoop dat je alles in mijn brief hebt begrepen en ik wens jou en je ouders een heel goede tijd in Nederland toe!]CS<br><br><br>[Hartelijke groeten,<br><voornaam>]CG | Dutchman, you can always discuss that.<br>[I hope that you were able to follow everything in my letter and I wish you and your parents a very good time in The Netherlands] CS<br><br>[Cordial greetings,<br><Christian name>]CG | CS: closing sentence<br><br><br><br>CG: closing greeting |

**Appendix B. Argumenative text, annotated**

| Dutch text | English gloss | Annotation legend |
|---|---|---|
| [De directeur van de school heeft een geldbedrag van de overheid gekregen om de school te verbeteren.]STR | [The school director has received a sum of money from the government to improve the school.]STR | STR = situation the text responds to |
| [Omdat de gymzaal oud is en er nauwelijks toestellen staan]PARG1 wil de directie het geld in de gymzaal besteden. | [Because the gym is old and barely has gym equipment]PARG1 the board wants to spend the money to the gym. | PARG1 = first primary argument |
| [Ook al is de gymzaal er slecht aan toe, [de kantine is nog slechter in staat.]CON | [Although the gym is in bad shape, the canteen is even worse in condition.]CON | CON = concession |
| [Daarom denk ik dat de directie het geld beter in de kantine kan besteden.]C | [That's why I think the board had better spend the money in the canteen.]C | C = claim |
| [Wanneer je in de kantine komt na de les is het meestal al helemaal vol met kinderen.]PARG2 | [When you enter the canteen after the lesson he is most of the time already entirely filled with children.]PARG2 | PARG2 = second primary argument |
| [Daarom staan kinderen soms ook te eten in plaats van dat zitten.]SARG2.1 | [As a result, the children stand eating instead of sit down.]SARG2.1 | SARG2.1 = further support of second argument |
| [De kantine ziet er ook erg onverzorgd uit.]PARG3 | [The canteen also looks very unkempt.]PARG3 | PARG3 = third primary argument |
| [De stoelen vallen bijna uit elkaar en de bankjes kraken.]SARG3.1 | [The chairs almost fall apart and the benches crack.]SARG3.1 | SARG3.1 = further support of third argument |

| | | |
|---|---|---|
| [Vaak vallen de prullenbakken om, omdat ze te veel wiebelen.]SARG3.2 | [Often the wastebaskets fall over, because they wiggle too much.]SARG3.2 | SARG3.2 = further support of third argument |
| [Ook is er te weinig eten in de kantine.]PARG4 | [Also, there is not enough food in the canteen.]PARG4 | PARG4 = fourth primary argument |
| [Er zijn maar 3 verschillende smaken frisdrank en [er is altijd eentje op.]SARG4.1 | [There are only 3 kinds of fizzy drinks and [is always one of them sold out.]SARG4.1 | SARG4.1 = further support of fourth argument |
| [Bij de snoepautomaat is hetzelfde geval.]SARG4.2 | [With the candy machine the same is case.]SARG4.2 | SARG4.2 = further support of fourth argument |
| [Als de directie het geld in de kantine besteedt in plaats van de gymzaal, dan zouden meer kinderen het leuk vinden om pauze te houden.]PARG5 | [When the board would spend the money in the canteen instead of the gym, more children would enjoy having a break.]PARG5 | PARG5 = fifth primary argument |
| [De gymzaal heeft al genoeg spullen om een fatsoenlijke gymles te hebben, maar de kantine niet om pauze te houden.]PARG6 | [The gym has enough equipment to have a decent gym class, but the canteen has not enough equipment to have a break.]PARG6 | PARG6 = sixth primary argument |