Linking Language to Math Success in an On-Line Course

Scott Crossley
Georgia State University
Atlanta, GA 30303
scrossley@gsu.edu

Tiffany Barnes
Collin Lynch
North Carolina State University
Raleigh, NC 27606
tmbarnes, cflynch@ncsu.edu

Danielle S. McNamara
Arizona State University
Tempe, AZ, 85287
dsmcnama@asu.edu

ABSTRACT

This study takes a novel approach toward understanding success in a math course by examining the linguistic features and affect of students' language production within a blended (with both on-line and traditional face to face instruction) undergraduate course (n=158) on discrete mathematics. Three linear effects models were compared: (a) a baseline linear model including non-linguistic fixed effects, (b) a model including only linguistic factors, (c) a model including both linguistic and non-linguistic effects. The best model (c) explained 16% of the variance of final course scores, revealing significant effects for one non-linguistic feature (days on the system) and two linguistic features (*Number of dependents per prepositional object nominal* and *Sentence linking connectives*). One non-linguistic factor (*Is a peer* tutor) and two linguistic variables (*Words related to* self and *Words related to tool* use) demonstrated marginal significance. The findings indicate that language proficiency is strongly linked to math performance such that more complex syntactic structures and fewer explicit cohesion devices equate to higher course performance. The linguistic model also indicated that less self-centered students and students using words related to tool use were more successful. In addition, the results indicate that students that are more active in on-line discussion forums are more likely to be successful.

Keywords

NLP, math, student success, on-line learning

1. INTRODUCTION

Cognitive skills are crucial for student success in the math classroom. While research has primarily focused on skills that strongly overlap with math knowledge including spatial attention and quantitative ability [1], cognitive skills supporting math success such as language ability remain under-researched. At the same time, a number of researchers have argued that language skills are a prerequisite for transferring cognitive operations between math and language domains and that lower language skills can present critical obstacles in math reasoning.

Prior research has examined links between language skills and math success to examine the premise that students with greater language abilities are better able to engage with math concepts and problems. This research is based on the notion that success in the math classroom can be partially explained through language skills that allow students to constructively participate in math discussions as well as to quantitatively engage with math problems [2]. Similarly, math literacy is thought to be not just knowledge of numbers and symbols, but also knowledge of language to understand the discourse of math (i.e., the words surrounding the numbers and symbols) [3].

Despite research that links language skills to math success in the classroom, a major methodological problem in previous studies is the reliance on correlational analyses among standardized tests of math and linguistic knowledge. For instance, several studies have looked at the links between tests of language proficiency (e.g., syntax, knowledge, verbal ability, and phonological skills) and success on tests of math knowledge (e.g. algebraic notation, procedural arithmetic, and arithmetic word problems [4, 5]). Other studies have compared success on standardized math tests between native speakers of English and second language speakers of English with lower linguistic ability [6, 7]. While a few studies have focused on the perceived linguistic complexity of math problems in standardized tests [8, 9], the majority of studies have not analyzed the actual language produced by students and the relationship between language complexity and success on math assessments (see [10] for an exception).

This study builds on the work of Crossley et al. [10] and examines links between the complexity of language produced by students in on-line question/answer forum in a blended math class and their success in the course. To do so, we examine students' forum posts within the on-line tools used in the class for a number of linguistic features related to text cohesion, lexical sophistication, syntactic complexity, and sentiment derived from natural language processing (NLP) tools. The goal of this study is to examine the extent to which the linguistic features produced by students are predictive of their final scores in a blended discrete mathematics course. In addition to the linguistic features, we also examined a number of non-linguistic factors that are potentially predictive of math success including: whether the student was a peer tutor, class section (of two sections), and on-line forum behavior including: how many times they viewed posts, how many posts they made, how many questions they asked, how many answers they provided, and how many days they visited the on-line class forum.

1.1 Language and Math Relationships

Prior studies have investigated the connections between language proficiency and math skills in native speakers (NS) of English. These studies generally demonstrate strong links between math ability and language ability. For instance, Macgregor and Price [5] found that students who scored high on an algebra test also scored well on language tests. A follow-up study using a more difficult algebra test found a stronger relationship between algebraic notation and language ability. Similarly, Vukovic and Lesaux [4] reported links between language and math skills, but that the language skills differed in their degree of relation with math knowledge. For example, general verbal ability was indirectly related through symbolic number skills while phonological skills were directly related to arithmetic knowledge. Other research has focused on the indirect links between math and language skills. For example, Hernandez [11] analyzed students' scores from the reading and math sections of a standardized test and found significant positive correlations between reading ability and math achievement. These findings led Hernandez to recommend that students' reading skills and strategy training should be factored into math instruction in order to increase effectiveness, especially for poor readers. However, not all studies have found strong links between math knowledge and language skills. For instance, LeFevre et al. [1] reported that linguistic skills were related to number naming, that quantitative abilities were related to processing numerical magnitudes, and that spatial attention was related to a variety of numerical and math tests. However, non-linguistic features such as quantitative abilities and spatial attention were stronger predictors of math ability.

In terms of language production, only one study, to our knowledge, has examined the links between the language produced by students and their success in the math classroom. Crossley et al. [10] examined linguistic and non-linguistic features of elementary student discourse while students were engaged in collaborative problem solving within an on-line math tutoring system. Student speech was transcribed and NLP tools were used to extract linguistic information related to text cohesion and lexical sophistication. They examined links between the linguistic features and pretest and posttest math performance scores as well as links with a number of non-linguistic factors including gender, age, grade, school, and content focus (procedural versus conceptual). Their results indicated that non-linguistic factors are not predictive of math scores but that linguistic features related to cohesion, affect, and lexical proficiency explained around 30% of the variance in students' math scores. Specifically, higher scoring students produced more cohesive texts that were more linguistically sophisticated.

1.2 Current Study

A number of studies have demonstrated strong links between students' linguistic knowledge, affect, and their success in math. Studies examining these links have traditionally relied on correlational analyses between linguistic knowledge tests and standardized math tests [1, 3, 4]. In this study, we take a novel approach and examine the linguistic features and affect of students' language production in a blended math class with both on-line and traditional face to face instruction. To derive our variables of interest, we analyzed the linguistic and affective features produced by the students in their forum postings using a number of NLP tools. These tools extract information related to text cohesion, lexical sophistication, syntactic complexity, and sentiment. In contrast to most prior studies (see [10] for an exception), our interest is not on linguistic performance as measured by standardized tests, but on linguistic performance as a function of language production as found in students' forum posts.

Our criterion variables are students' final score in the semester-long blended math class. In addition to examining relations between linguistic features of student language production and math scores, we also examined a number of non-linguistic factors including: whether the student was a peer tutor; how many times they viewed posts in the on-line forum; how many posts they made in the on-line forum; how many answers they provided in the on-line forum; how many questions they asked in the on-line forum; how many days they visited the on-line forum; and class section (there were two sections). Thus, in this study, we addressed two research questions:

1. Are non-linguistic factors significant predictors of math performance in a blended math class?
2. Are linguistic factors related to lexical sophistication, cohesion, syntactic complexity, and affect significant predictors of math performance in a blended math class?

2. METHOD

2.1 The Blended Math Class: Discrete Math

Discrete Mathematics is an undergraduate math course offered by the computer science department at North Carolina State University. Students in the course are provided instruction on the mathematical tools and abstractions that are integral to a general CS education, including logic, truth tables, set theory, graphs, counting, induction, recursion, and functions. Students majoring in CS must complete the course with a grade of C or better in order to remain in their degree program. The course includes 10 homework assignments, 5 lab assignments, 3 midterms, and a final exam.

The discrete math course studied is a *blended course*. In addition to the standard lecture and office hours, students are supported by a range of on-line tools. These include a Piazza question/answer forum, on-line homework assignments through WebAssign, and two labs that are Intelligent Tutoring Systems for logic and probability. Our focus in this analysis is the Piazza data. Piazza is a standard question-answering forum. Students, teaching assistants (TAs), and instructors are allowed to post questions or topic prompts as well as general polls. The members of the class may then respond to these posts with replies and sub-replies. They may also choose to recommend both posts and replies as being particularly informative but clicking on a "good question" or "good answer" button. Question responses are classified in Piazza. The instructors and TAs may post an official "instructor response". If that is done, then these are flagged separately from student replies. Individuals may edit their replies over time in response to users' comments. While Piazza may be configured to permit anonymous posting by students, this function was turned off by default in this course. In addition to the basic thread structure, Piazza requires that posts be categorized by topic and it keeps a running list of threads and supports basic search to help students locate relevant information.

We study data from the Fall 2013 semester of this course. During that semester, the class was divided into two sections with two primary instructors, five teaching assistants, and 250 students. In addition to the instructor and official graduate TAs, the course was supported by a set of peer tutors. These are high-performing students in the course who are given extra credit for acting as mentors. During the Fall 2013 semester, 32 students volunteered to act as peer tutors and roughly 1/3 of them completed the required 10 hours to receive extra credit.

For the purposes of our analysis, we collected Piazza data recording the students' interactions once the course was complete. This data included how many times students viewed posts in the Piazza forum, how many posts students made in the Piazza forum, how many answers students provided in the Piazza forum, how many questions students asked in the Piazza forum, and how many days students visited the Piazza forum.

2.2 Forum Posts

We selected forum posts because they provide students with a platform to exchange ideas, discuss lectures, ask questions about the course, and seek technical help, all of which lead to the production of language in a natural setting. Such natural language can provide researchers with a window into individual student motivation, linguistics skills, writing strategies, and affective states. This information can in turn be used to develop models to improve students' learning experiences [12].

Students in the course were given access to the Piazza forum at the start of the class. Students were encouraged to use Piazza (not email) for course communications by posting their questions to the forum outside of class, and answering questions posed by their peers. The TAs and peer tutors were required to check the forum regularly with the goal of ensuring an average response time of 15 minutes per post, and that no single question would "go stale" by being left for more than 2 hours without a reply. In addition to basic question/reply Piazza interactions, the instructor and TAs posted regular announcements and general comments to the forum, making it the primary vehicle for non-lecture communication in the course.

Student posts were retrieved from a Piazza database that was extracted at the end of the course. The student posts were segmented out to eliminate duplicate content as well as unnecessary markup. Of the 250 students in the course, 169 made posts on the forum. For the 169 students who made a forum post, we aggregated each of their posts such that each post became a paragraph in a text file. We selected only those students who produced at least 50 words in their aggregated posts (n = 158). We selected a cut off of 50 words in order to have sufficient linguistic information to reliably assess the student's language using NLP tools.

2.3 Natural Language Processing Tools

We used several NLP tools to assess the linguistic features in the aggregated posts of sufficient length. These included the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [13], the Tool for the Automatic Analysis of Cohesion (TAACO) [14], the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) [15], and the SEntiment ANalysis and Cognition Engine (SEANCE) [16]. The selected tools reported on language features related to lexical sophistication, text cohesion, and sentiment analysis respectively. The tools are discussed in greater detail below.

2.3.1 TAALES

TAALES incorporates about 150 indices related to basic lexical information (e.g., the number of tokens and types), lexical frequency, lexical range, psycholinguistic word information (e.g., concreteness, meaningfulness), and academic language for both single word and multi-word units (e.g., bigrams and trigrams).

2.3.2 TAACO

TAACO incorporates over 150 classic and recently developed indices related to text cohesion. For a number of indices, the tool incorporates a part of speech (POS) tagger and synonym sets from the WordNet lexical database [17]. TAACO provides linguistic counts for both sentence and paragraph markers of cohesion and incorporates WordNet synonym sets. Specifically, TAACO calculates type token ratio (TTR) indices, sentence overlap indices that assess local cohesion, paragraph overlap indices that assess global cohesion, and a variety of connective indices such as logical connectives (e.g., *moreover, nevertheless*) and sentence linking connectives (e.g., *nonetheless, therefore, however*).

2.3.3 TAASSC

TAASSC measures large and fined grained clausal and phrasal indices of syntactic complexity and usage-based frequency/contingency indices of syntactic sophistication. TAASSC includes 14 indices measured by Lu's [18] Syntactic Complexity Analyzer (SCA), 31 fine-grained indices or clausal complexity, 132 fine-grained indices of phrasal complexity, and 190 usage-based indices of syntactic sophistication. The SCA measures are classic measures of syntax based on t-unit analyses [19]. The fine-grained clausal indices calculate the average number of particular structures per clause and dependents per clause. The fine-grained phrasal indices measure 7 noun phrase types and 10 phrasal dependent types. The syntactic sophistication indices are grounded in usage-based theories of language acquisition [Ellis, 2002] and measure the frequency, type token ratio, attested items, and association strengths for verb-argument constructions (VACs) in a text.

2.3.4 SEANCE

SEANCE is a sentiment analysis tool that relies on a number of pre-existing sentiment, social positioning, and cognition dictionaries. SEANCE contains a number of pre-developed word vectors developed to measure sentiment, cognition, and social order. These vectors are taken from freely available source databases. For many of these vectors, SEANCE also provides a negation feature (i.e., a contextual valence shifter) that ignores positive terms that are negated (e.g., not happy). SEANCE also includes a part of speech (POS) tagger.

2.4 Statistical Analysis

We calculated linear models to assess the degree to which linguistic features in the students' language output along with other fixed effects (e.g., question/note posted, questions answered, site visits) were predictive of students' final math scores. Prior to linear model analysis, we first checked that the linguistic variables were normally distributed. We also controlled for multicollinearity between all the linguistic and non-linguistic variables ($r \geq .900$) such that if two or more variables were highly collinear, all but one of the variables was removed from the analysis. We used R [21] for our statistical analysis. Final model selection and interpretation was based on t and p values for fixed effects and visual inspection of residuals distribution for non-standardized variables. To obtain a measure of effect sizes, we computed correlations between fitted and observed values, resulting in an overall R^2 value for the fixed factors. We developed and compared three models: (a) a baseline linear model including non-linguistic fixed effects, (b) a model including only linguistic factors, (c) a model including both linguistic and non-linguistic effects. We compared the strength of each model using Analyses of Variance (ANOVAs) to examine which models were most predictive.

3. RESULTS

3.1 Non-linguistic Linear Model

A linear model considering of all non-linguistic fixed effects revealed significant effects for whether the student was a tutor or not and number of days spent on the Piazza forum. Table 1 displays the coefficients, standard error, t values, and p values for each of the significant non-linguistic fixed effects. The overall model was significant, $F(3, 154) = 6.116$, $p < .001$, $r = .326$, $R^2 = .107$. Inspection of residuals suggested the model was not influenced by homoscedasticity. The non-linguistic variables explained around 11% of the variance of the math scores and indicated that students who acted as peer tutors and visited the system more often received higher overall grades in the class.

Table 1. Non-linguistic model for predicting math scores

Fixed Effect	Coefficient	Std. Error	*t*
(Intercept)	83.988	1.484	56.603***
Is a peer tutor	5.410	1.995	2.712**
Is not a peer tutor	3.340	2.090	1.598
Days on system	0.038	0.012	3.116**

Note * $p < .050$, ** $p < .010$, **p < .001

3.2 Linguistic Linear Model

A linear model including linguistic fixed effects revealed significant effects for a number of features related to reference self, syntactic complexity, reference to tools, and cohesion. Table 2 displays the coefficients, standard error, *t* values, and *p* values for each of the linguistic fixed effects. The overall model was significant, $F(4, 153) = 9.456$, $p < .001$, $r = .360$, $R^2 = .130$. Inspection of residuals suggested the model was not influenced by homoscedasticity. The linguistic variables explained around 13% of the variance of the math scores and indicated that students who referred to themselves less often, used more complex syntax, referred to words related to the use of tools, and used fewer sentence linking terms received higher final grades in the course. An ANOVA comparison between the non-linguistic model and the linguistic found a significant difference between the models, ($F = 8.120$, $p < .001$), indicating that linguistic features contributed to a better model fit than non-linguistic features.

Table 2. Linguistic model for predicting math scores

Fixed Effect	Coefficient	Std. Error	*t*
(Intercept)	91.089	3.795	24.002***
Words related to self	-67.146	26.024	-2.580*
Number of dependents per prepositional object nominal	6.800	2.478	2.744**
Words related to tools	144.097	62.658	2.300*
Sentence linking connectives	-77.055	33.947	-2.27*

Note * $p < .050$, ** $p < .010$, **p < .001

3.3 Full Linear Model

A linear model considering non-linguistic and linguistic fixed effects revealed significant effects for one of the non-linguistic features (days on the system) and two of the linguistic features (*Number of dependents per prepositional object nominal* and *Sentence linking connectives*). One non-linguistic factor (*Is a peer* tutor) and two linguistic variables (*Words related to* self and *Words related to tool* use) demonstrated marginal significance. Table 3 displays the coefficients, standard error, *t* values, and *p* values for each of the fixed effects. The overall model was significant, $F(7, 150) = 9.295$, $p < .001$, $r = .399$, $R^2 = .159$. Inspection of residuals suggested that the model was not influenced by homoscedasticity. The non-linguistic and linguistic variables explained around 16% of the variance of the math scores and followed the same trends as reported in the first two models. An ANOVA comparison between the full model and the linguistic model found a significant difference between the models, ($F = 2.790$, $p < .050$), indicating that a combination of non-linguistic and linguistic features contributed to a better model fit than linguistic features alone.

Table 3. Full model for predicting math scores

Fixed Effect	Coefficient	Std. Error	*t*
(Intercept)	86.564	4.065	21.296***
Is a peer tutor	3.840	1.974	1.946
Is not a peer tutor	1.516	2.065	0.734
Days on system	0.028	0.012	2.273*
Words related to self	-44.990	26.876	-1.674
Number of dependents per prepositional object nominal	6.156	2.455	2.507*
Words related to tools	120.451	62.545	1.926
Sentence linking connectives	-72.463	33.644	-2.154*

Note * $p < .050$, ** $p < .010$, **p < .001

4. DISCUSSION AND CONCLUSION

Previous research has indicated that language skills are related to math success. Much of this research examined links between standardized tests of language proficiency and success on tests of math knowledge [4, 5] while other research has compared native English speakers to second language speakers of English in terms of success on standardized math tests [6, 7]. In general, these studies have yielded positive relationships between language skills and math success. However, the majority of these studies did not examine links between the language produced by students and math success. A notable exception to this is Crossley et al.'s [10] study that used NLP tools to examine links between language used in an third grade math classroom and success on math assessments. This study reported that linguistic features related to cohesion, affect, and lexical proficiency explained around 30% of the variance in the math scores.

In this study, we take a similar approach to Crossley et al. [10] and use NLP tools to extract a number of linguistic and sentiment features from forum posts found in a blended discrete math undergraduate course. We found that a number of non-linguistic and linguistic features were strong predictors of math success. For instance, peer tutors and students who spent more time on the Piazza forums were more likely to be successful in the class. Linguistically, students who used fewer words related to self, more syntactically complex sentences, more words related to tool use, and fewer connectives were also more successful in the class. The non-linguistic model explained about 11% of the variance in the math scores while the linguistic model explained about 13% of the variance. A model that included both non-linguistic and linguistic variables explained about 16% of the variance in the math scores.

The variance explained by our model was lower than that reported in Crossley et al [10]. However, unlike Crossley et al., our participants were not elementary level students and they were not involved in collaborative discourse. Rather, our participants were college students and the language samples used in this study came from on-line forum posts as compared to natural conversation between students in a classroom. These differences likely explain the disparities reported between the two studies. For instance, in the current study we found a negative correlation between a cohesion index (*sentence linking connectives*) and math scores. This may be the result of linguistic development in which young children develop text cohesion using explicit markers of cohesion while college students use complex syntax to develop cohesive text [22, 23]. This distinction likely indicates that the strong positive correlation between syntactic complexity and math success reported in this study indicates that more skilled writers have greater success in the math classroom.

This study also found that a number of different indices than those reported by Crossley et al. were predictive of math success. These included words related to self, which was negatively associated with math success, and words related to tool use, which was positively associated with math success. The finding for words related to self should likely be interpreted in terms of self-centered behavior such that students who were more self-centered were likely to be less successful in the math class. This may be a result of the collaborative nature of the Piazza forum in which students were encouraged to work together to answer questions and solve problems. In terms of words related to tool use, the findings likely indicate that more successful students used terms that were more strongly related to the domain such as *computer, equipment, file, machine, mechanism,* and *paper*. However, it is notable that neither the use of words related to self or to the use of tools were a significant predictor in the full model that included both linguistic and non-linguistic variables.

In terms of non-linguistic features, this analysis demonstrated that two non-linguistic factors were important indicators of math success: *peer tutoring* and *days on Piazza*. The findings indicate that those students who volunteered to peer tutor were more successful in the class. In addition, those students who spent a greater number of days on the Piazza forum were more successful suggesting that engagement in the class discussion forum led to greater success. However, only the number of days spent on the Piazza forum was a significant predictor in the full model.

The findings from this study have practical implications for understanding math performance in a blended math class at the university level. Specifically, the findings provide additional support that language proficiency is strongly linked to math performance such that more complex syntactic structures and fewer explicit cohesion devices equate to higher course performance. The linguistic model also indicated that less self-centered students and students using words related to tool use were more successful. In addition, the results indicate that students who are more active in on-line discussion forums are more likely to be successful. The study also provides a contrast to early research [10] in that differences are reported between age levels (elementary and college level students) and learning environments (collaborative discussions and forum posts). Future studies can build on these results by continuing to examine language features and math success in a number of different student populations and learning environments.

5. ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation (DRL- 1418378). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

[1] LeFevre J. A., Fast L., Skwarchuk S.L., Smith-Chant B.L., Bisanz J., Kamawar D., Penner-Wilger M. 2010. Pathways to math: Longitudinal predictors of performance. *Child Development 81*(6): 1753–1767.

[2] Vukovic, RK, Lesaux NK (2013) The relationship between linguistic skills and arithmetic knowledge. Learning and Individual Differences *23*: 87-91.

[3] Adams TL (2003) Reading math: More than words can say. The Reading Teacher 56(8): 786–795.

[4] Vukovic, RK, Lesaux NK (2013) The relationship between linguistic skills and arithmetic knowledge. Learning and Individual Differences *23*: 87-91.

[5] MacGregor M, Price E (1999) An exploration of aspects of language proficiency and algebra learning. Journal for Research in Math Education 449–467. doi: 10.2307/749709

[6] Alt M, Arizmendi GD, Beal CR (2014) The relationship between math and language: Academic implications for children with specific language impairment and English language learners. Language, Speech, and Hearing Services in Schools 45(3): 220–233. doi:10.1044/2014_LSHSS-13-0003

[7] Hampden-Thompson G, Mulligan G, Kinukawa A, Halle T (2008) Math Achievement of Language-Minority Students During the Elementary Years. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

[8] Martiniello M (2008) Language and the performance of English-language learners in math word problems. Harvard Educational Review 78(2): 333–368.

[9] Martiniello M (2009) Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. Educational Assessment 14(3–4): 160–179. doi:10.1080/10627190903422906

[10] Crossley, SA, Liu R, McNamara D (2017). Predicting Math Performance Using NLP Tools. Proceedings of the 7th International Learning Analytics and Knowledge (LAK) Conference. New York, NY: ACM.

[11] Hernandez F (2013) The Relationship Between Reading and Math Achievement of Middle School Students as Measured by the Texas Assessment of Knowledge and Skills (Doctoral dissertation).

[12] Ramesh, A., Goldwasser, D., Huang, B., Daume, H., & Getoor, L. (2014). Understanding MOOC Discussion Forums using Seeded LDA. In *9th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 28–33). Baltimore, MA: ACL.

[13] Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757-786. doi:10.1002/tesq.194

[14] Crossley, S. A., Kyle, K., & McNamara, D. S. (in press). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*.

[15] Kyle, K., & Crossley, S. A. (in press). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*.

[16] Crossley, S. A., Kyle, K., & McNamara, D. S. (in press). Sentiment Analysis and Social Cognition Engine (SEANCE): An Automatic Tool for Sentiment, Social Cognition, and Social Order Analysis. *Behavior Research Methods*.

[17] Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM, 38*(11), 39–41.

[18] Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

[19] Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518.

[20] Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 143–188.

[21] R Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013: ISBN 3-900051-07-0.

[22] Crossley, S. A., Weston, J., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. Written Communication, 28(3), 282-311.

[23] Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. Written Communication, 17, 307-352.