VARIATION IN LINGUISTIC CHARACTERISTICS BETWEEN THE TYPES OF COMPUTER-MEDIATED ACADEMIC DISCOURSE

NATAĻJA CIGANKOVA

University of Latvia, Latvia

Abstract. The article presents the results of empirical research investigating the specific linguistic characteristics of the types of Computer-Mediated Academic Discourse (CMAD) – the English language use by language teaching professionals in academic computer-mediated seminars (webinars), synchronous conferences (chats), asynchronous discussion fora, e-mails, weblogs and hypertexts. Six specialised corpora were complied to represent each type of CMAD. Multidimensional analysis of the variance of linguistic features (Biber, 1988) was applied as the main quantitative research method. Considerable differences have been revealed in the use of fifty-five types of linguistic features in the sub-corpora. The results of Scheffé's test show that there is a significant statistical difference between at least one pair of the mean values on each dimension. This indicates that the studied types of CMAD are rather similar on one dimension but different on another. The author demonstrates that each type of CMAD has specific linguistic characteristics distinguishing it from other types. The findings obtained in the research may be of interest to researchers investigating varieties of computer-mediated language, language educators and other specialists in applied linguistics.

Key words: applied linguistics, language in use, computer-mediated academic discourse, linguistic variation, multidimensional analysis

INTRODUCTION

The role of technology in human communication has been constantly increasing recently, which has triggered a dramatic rise in interest in the study of computer-mediated language (Shortis, 2000; Crystal, 2001; Herring and Paolillo, 2006; Grieve et al., 2011). Moreover, many scholars emphasise the growing significance of the role of computer-mediated discourse in academic communication (e.g., Stuart, 2006). However, despite the importance of this type of discourse and noticeable scholars' attention to it, the specific characteristic features of computer-mediated discourse types that occur in academic settings have not yet been systematically investigated. Therefore, the aim of this study is to distinguish and compare the characteristics of the English language use by language teaching professionals in six types of computer-mediated academic discourse and to reveal where exactly the differences between them lie.

BACKGROUND

Among many contemporary approaches to the study of linguistic variation in discourse, the presented below approach that relates the use of linguistic means to their functions and applies corpus linguistic methodology to the study of frequency of linguistic features in different text types stands out in terms of reliability and robustness of obtainable results.

According to Halliday and Matthiessen (2004), the choice of linguistic means in a text depends on language functions. The functions of language are the factors that cause linguistic variation. For example, the choice of linguistic features used by a language user depends on whether the text is meant to inform or to persuade the reader, to maintain social relationship or to express disagreement or complaint. Moreover, the choice of linguistic means depends on the linguistic features already used in the text.

On the basis of the co-occurrence of linguistic features in different text types, Biber (1988) has developed a method of multidimensional factor analysis that groups the linguistic features in a text into a limited number of factors – functional dimensions – according to the functions they perform in texts. The assumption underlying the methodology is that linguistic features do not randomly co-occur in textual realisations of discourse. If persistent co-occurrence of some linguistic features is observed in a group of texts, it is reasonable to suppose that there is an underlying functional relationship between the features that makes them co-occur. Thus, the patterns of co-occurrence mark underlying functional dimensions. According to Biber, it is not possible to analyse linguistic variation in discourse along one dimension, e.g., speaking/writing. A multidimensional approach is necessary.

Having used a multivariate analysis statistical method, Biber (1988) identified which linguistic features typically co-occur in different types of texts. He selected 59 linguistic features, but reduced the number of variables to a small set of factors to find out the co-occurring linguistic features. The researcher revealed that the linguistic features that serve similar discourse functions tend to appear in similar text types. Different groups of co-occurring features constitute different dimensions. Thus, the linguistic dimensions are the continua along which register variation occurs and the types of discourse differ from one another in the English language. Biber has applied this methodology in his studies of academic discourse, e.g. to the analysis of spoken and written academic discourse in American universities (Biber et al., 2004).

Other researchers have also used Biber's method for investigating CMD. Having applied multidimensional analysis research methodology to the study of the corpus of computer-mediated language, Collot and Belmore (1993) made a distinction between the messages composed at the moment of communication and the messages that have been pre-written, carefully thought over before being sent. The different ways in which the messages are produced are now known as *synchronous* and *asynchronous* modes of CMC correspondingly. The researchers

argue that the situational constraints by which the 'electronic language' is characterised make it different from other varieties of English, the main difference being that 'electronic language displays some of the linguistic features which have been associated with certain forms of written language and others which are more usually associated with spoken language' (Collot and Belmore, 1993: 48).

Many scholars have applied Biber's methodology to investigate linguistic characteristics of popular and academic texts (Conrad, 2001; Gries, 2003). Biber and Kurjian (2007), for example, have applied the multidimensional analytic method to the study of text categorisation in Google searchers on the Web and suggested some changes and improvements in the taxonomy of texts. However, texts representing computer-mediated academic discourse have not been included in their research.

CMAD is defined as a computer-mediated process of functional language use for communicative purposes in academic contexts that is realised in semantically connected, verbal instances of spoken or written language longer than a sentence (texts), which are meaningful to the communicating language users (Cigankova, 2009). Linguistically competent communicators choose different linguistic means available in the English language in different types of CMAD. Therefore, quantitative linguistic characteristics may vary across text types. The author aims to support this claim by providing objective statistical data obtained in the corpus-based quantitative research applying Biber's multidimensional view of variation in discourse.

RESEARCH PROCEDURE

The CMAD corpus (60,000 words) contained computer-mediated texts representing the disciplinary domain of education, produced in 2007- 2009 by the members of three European online communities of university teachers (Online 1, 2, 3). The data comprised 1350 participants representing a wide range of the first language backgrounds (42 languages). The percentage of native speakers of English was not higher than 0.3%. The corpus was divided into six specialised sub-corpora (10,000 words each), representing the following CMAD types: computer-mediated seminars (webinars), synchronous conferences (chats), asynchronous discussion fora, e-mails, weblogs and hypertexts. Each type of CMAD is the result of a unique combination of transactional or interactional type of discourse, synchronous or asynchronous mode of interaction, spoken or written mode of discourse and the type of software used for communication.

The main quantitative research method applied in this study was the method of multidimensional analysis of variance of linguistic features (Biber, 1988). To reveal the patterns in the use of linguistic features in texts as realisations of CMAD, the author identified fifty-five linguistic features in representative samples of each CMAD type and computed the frequency of their occurrence. The system proposed by Biber (1988) for coding linguistic features in the corpus was applied.

For the comparability of the results, the frequency counts in CMAD texts were normalised to 1000 (except type/token ratio and word length) and standardised. The variation of the frequency of linguistic features in each text from the mean in the whole corpus was measured in standard deviations, applying the following formula (McEnery et al., 2006: 303)**:**

$$k = \frac{F - \mu}{\sigma}$$

In the formula, k is the computed standard value, F stands for the frequency of the linguistic feature in the text, μ is the mean value, and σ is standard deviation (SD).

The *Factor Analysis* procedure was repeated five times, each time with a different number of extracted factors set (from 9 to 5), in order to make a decision on the optimal number of factors. The best result, in terms of Kaiser-Meyer-Olkin Measure of Sampling Adequacy, Chi-Square and the level of significance, was received for five extracted factors. Hence, a five-factor model was applied in the research, including the following dimensions that had been previously identified by Biber (1988: 13): Dimension 1 (*Involved/ Informational production),* Dimension 2 (*Narrative/ Non-narrative concerns*), Dimension 3 (*Explicit/ Situation-dependent reference*), Dimension 4 (*Overt expression of persuasion*), Dimension 5 (*Abstract/ Non-abstract information*).

General Linear Models (ANOVA) was applied to calculate the statistical differences between CMAD types along the textual dimensions identified in the *Factor Analysis*. The author aimed to find sufficient sample evidence to reject the null hypothesis (H0) which stated that there was no difference between CMAD types at significance level α=0.05. Essential probability statistic F ratio (Fisher's *Six Sigma data set comparison*) was calculated with the help of IBM SPSS Statistics 19 programme.

To reject the null hypothesis with a higher degree of confidence, it was necessary to prove that at least one of the mean values was not the same as the other mean values in the group. To reveal that, Scheffé's test was conducted to analyse the pairs of mean values to see if there were differences between them and reveal where exactly the differences lay. Multiple comparisons were conducted between the mean standardised frequency values for each one type of CMAD and the mean standardised frequency values for the other types along each dimension.

RESULTS AND DISCUSSION

The standard values were calculated for each CMAD type on each dimension and compared. Table 1 presents the mean frequency counts per 1000-words and the standard values of lexico-grammatical features co-occurring together in different CMAD text types. As some linguistic features are more frequent in

English than other features, mean frequency counts may not show their actual frequency in texts in comparable values. Therefore, standard values are given to show the frequency of occurrences in the number of standard deviations from the mean in the corpus. The linguistic features are listed in the table in the descending order of their standard values.

Table 1 **Mean frequency counts and standard values of lexico-grammatical features co-occurrence in CMAD text types**

Synchronous conferences	Wh-questions (3.1*; 1.9**), present tense verbs (24.2; 1.7), time (5.2; 1.6) and place (1.2; 0.9) adverbials, amplifiers (5.1; 1.1), first- (51.1; 0.9) and second- (8.8; 0.9) person pronouns, discourse particles (4.3; 0.8), contractions (13.3; 0.6), indefinite pronouns (11.8; 0.6), final prepositions (0.1; 0.5)
Academic e-mails	Suasive verbs (0.3; 1.6), wh- relative clauses on object position (3.0; 1.5), discourse particles (11; 1.4), adverbials (5.5; 1.1), private verbs (11.1; 1.1), first- (45.1; 0.4) and second- (25.2; 1.0) person pronouns, general emphatics (13.9; 0.9), pronoun *it* (12.8; 0.7), possibility modals (11.9; 0.6), indefinite pronouns (10.7; 0.5), perfect aspect verbs (9.7; 0.5), conditionals (4.9; 0.4), time adverbials (11.5; 0.3)
Discussion fora	Infinitives (53.4; 1.6), possibility modals (15.5; 1.4), conditionals (7.7; 1.4), hedges (5.0; 1.3), gerunds (19.2; 1.2), high type/token ratio (43.97; 1.2), demonstratives (8.4; 1.1), general emphatics (14.5; 1.0), analytic negation (8.6; 0.9), phrasal coordination (25.6; 0.8), suasive verbs (0.2; 0.8), indefinite pronouns (11.8; 0.8), necessity modals (2.5; 0.6), public verbs (5.7; 0.6), prediction modals (12.5; 0.5)
Academic weblogs	past tense verbs (37.5; 2), adverbs (42.4; 2), that-deletion (10.5; 1.7), predicative adjectives (46.3; 1.4), third-person pronouns (20.7; 1.1), clausal subordination (12.9; 1.0), conjunctions (8.2; 1.0), contractions (20.1; 0.9), amplifiers (7.8; 0.9), present tense verbs (28.1; 0.9), that relatives (14.9; 0.9), sentence relatives (2.7; 0.9), perfect aspect verbs (11.1; 0.8), general emphatics (13.0; 0.8)
On-line seminars	Predictions modals (20.7; 1,7), wh- relative clauses on subject position (4.0; 1.6), public verbs (8.6; 1.4), wh-clauses (10.3; 1.2), demonstrative pronouns (12.2; 1.0), pronoun *it* (14.0; 0.9), suasive verbs (0.7; 0.8), place adverbials (5.6; 0.8), infinitives (41.4; 0.7), *that* relatives (13.6; 0.6)
Academic hypertexts	passive voice constructions (40.7; 3.4), past and present participle clauses (32.3; 3.8), attributive adjectives (103.5; 2.1), nominalisations (45.7; 1.7), nouns (239.1; 1.6), phrasal coordination (32.3; 1.6), mean syntactic length (28.8; 1.6), mean word length (5.5; 1.6), adverbials (9.6; 1.0), conjunctions (7.8; 0.9), present tense (40.9; 0.6) and perfect aspect (10.2; 0.6) verbs

* The first figure represents the mean frequency value per 1000 words
** The second figure shows the calculated standard value

For each type of CMAD a mean standard frequency value was calculated for each dimension. The results of simultaneous multiple comparisons between the means (ANOVA) revealed a significant statistical difference (F) in the

frequency of occurrence of lexico-grammatical features between CMAD types in all the dimensions. As it is evident from Table 2, the level of significance α for all cases was found to be less than 0.05; therefore, in no one of the cases the H0 hypothesis could be accepted. The results of Scheffé's test confirmed that there was a significant statistical difference between at least one pair of the mean values in each dimension. For this reason, H0 hypothesis was rejected in favour of the alternative statistical hypotheses in all the cases. Thus, the applied method with a 95% level of confidence provided evidence that the type of CMAD was a possible reason for the variance in the frequency of co-occurrence of linguistic features in the specialised CMAD corpora.

Table 2 **The results of the analysis of variance (ANOVA) between the mean frequency values of six CMAD types on five dimensions**

		Sum of Squares	df	Mean Square	F	Sig.
Dim 1	Between Groups	2222,995	5	445,999	28,144	,000
	Within Groups	855,752	54	15,847		
	Total	3085,747	59			
Dim 2	Between Groups	272,647	5	54,529	46,137	,000
	Within Groups	63,822	54	1,182		
	Total	336,469	59			
Dim 3	Between Groups	102,150	5	20,430	11,947	,000
	Within Groups	92,340	54	1,710		
	Total	194,491	59			
Dim 4	Between Groups	473,497	5	94,699	34,906	,000
	Within Groups	146,501	54	2,713		
	Total	619,998	59			
Dim 5	Between Groups	313,956	5	62,791	20,762	,000
	Within Groups	163,315	54	3,024		
	Total	477,271	59			

For a meaningful interpretation of the results, it was important to find out where exactly the differences lay. For this reason, the results of Scheffé's test for the mean values for all five dimensions were analysed and presented in the frequency polygons.

Fig. 1 demonstrates that the mean difference values of five out of six types of CMAD hold a high position on Dimension 1 – *Involved/ Informational Production*. The lower the position of the CMAD type in the frequency polygon, the less similarity it has with other CMAD types. The lowest position on this dimension is held by academic hypertexts, which means that they are the most informational in the type of information production of all the other types of CMAD. Other CMAD types have the degree of similarity above zero, except synchronous conferences, which are close to zero. This finding suggests that the CMAD type marked 'academic hypertext' differs significantly in the frequency of specific linguistic features from other five types of CMAD, having the highest frequency

of negative linguistic features on Dimension 1. The mean difference figures of other types of CMAD hold the position above or close to zero, implying a very small difference between them.
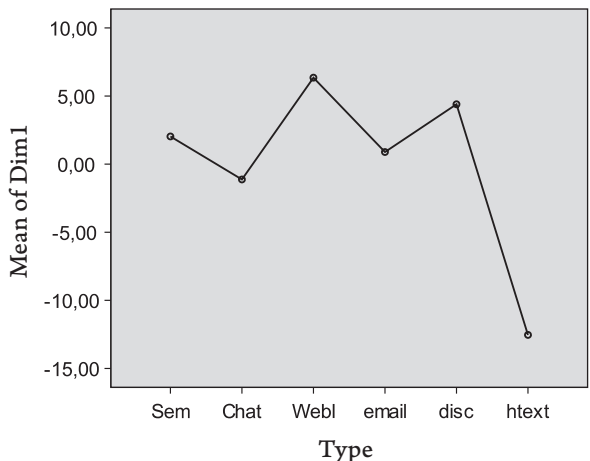

Figure 1 **Mean difference values on Dimension 1 for six types of CMAD**

The frequency polygon in Fig. 2 presents the mean difference values of CMAD types on Dimension 2 – *Narrative/ Non-narrative Concerns.*
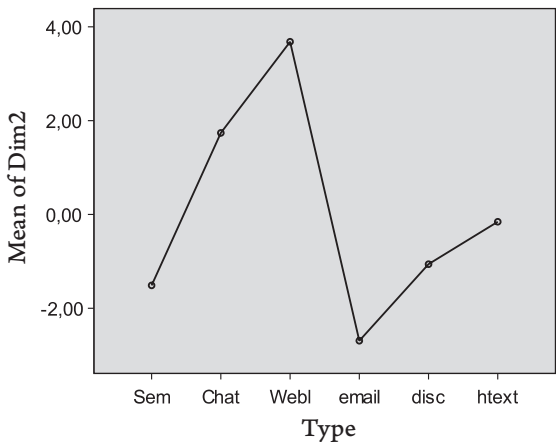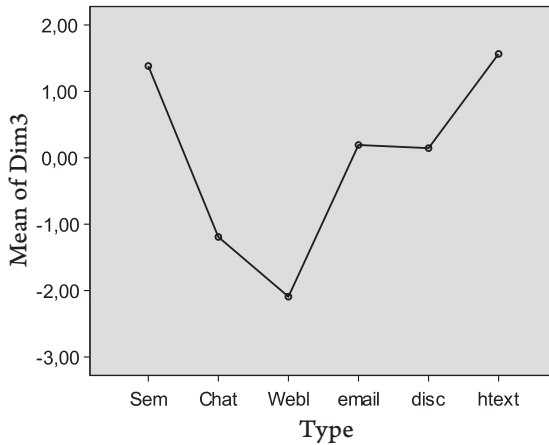

Figure 2 **Frequency polygon presenting the mean difference values for six types of CMAD on Dimension 2**

Academic weblogs show the biggest mean difference from other types of texts on this dimension, followed by synchronous conferences. E-mails show the smallest mean difference from other types of CMAD. On-line academic discussion fora hold almost the same position as academic seminars and are close to academic hypertexts. This confirms the previously made inference that these types of CMAD have similar linguistic characteristics on this dimension.

As the frequency polygon in Fig. 3 demonstrates, academic weblogs, closely followed by synchronous conferences, show the biggest mean difference from other types of texts on Dimension 3 (*Explicit versus Situation-Dependent Reference*). On-line academic seminars (spoken mode) and academic hypertexts (written mode) hold a high position on this dimension, which signals the high degree of similarity between them. The mean values for academic e-mails and discussion fora are almost identical, which implies a high degree of similarity in linguistic characteristics between these two types of CMAD on this dimension.

Figure 3 **Frequency polygon presenting the mean difference values for six types of CMAD on Dimension 3**

The frequency polygon in Fig. 4 shows the position of CMAD types on Dimension 4 – *Overt Expression of Persuasion*.
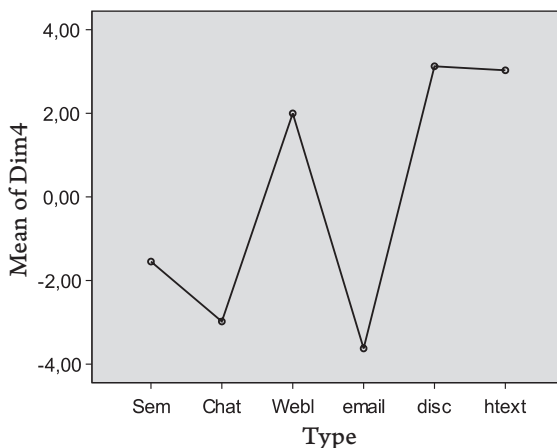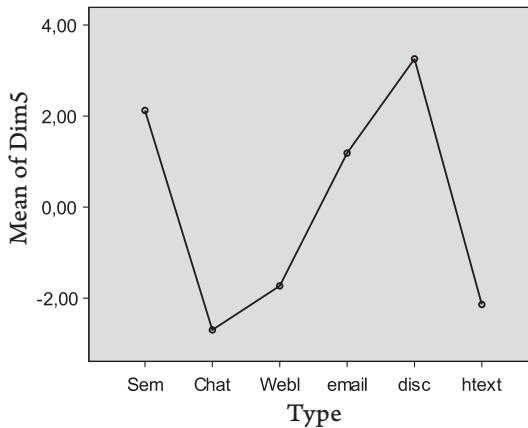

Figure 4 **Frequency polygon presenting the mean difference values for six types of CMAD on Dimension 4**

On-line seminars, e-mails and discussion fora show similar mean figures on this dimension, implying that they contain the linguistic features signalling explicitly expressed persuasion. In contrast, synchronous conferences, weblogs and academic hypertexts do not demonstrate high frequency of such linguistic features. This fact indicates that there are two distinct groups of CMAD types in respect of the explicitness of persuasion.

One more frequency polygon in Fig. 5 presents the mean difference values of CMAD types on Dimension 5 – *Abstract versus Non-abstract Style*. In respect of the abstractness of the information in the texts, the CMAD types fall into two distinct groups. Academic weblogs, discussion fora and hypertexts demonstrate a high degree of abstractness while on-line seminars, synchronous conferences and e-mails convey non-abstract information.

Figure 5 **Frequency polygon presenting the mean difference values for six types of CMAD on Dimension 5**

Academic discussions and hypertexts show the highest degree of similarity, as they both render the most abstract information. In contrast, e-mails and synchronous conferences are rather similar in that they both convey non-abstract information.

CONCLUSIONS

The differences along five textual dimensions in the frequency of occurrence of linguistic features among the six specialised corpora investigated in the research reveal multidimensional linguistic variation in the English language use across the investigated types of CMAD. Strong patterns of co-occurrence of linguistic features have been discovered that are regarded as different underlying functional dimensions along which the variation in CMAD occurs and the types of CMAD differ from one another. The results of Scheffé's test show that there is a significant statistical difference between at least one pair of the mean values

on each dimension. This indicates that the studied types of CMAD are rather similar on one dimension but different on another. The distinguished patterns of the co-occurrence of linguistic features are specific to each type of CMAD. Thus, each type of CMAD has been found to possess specific linguistic characteristics distinguishing it from other types.

REFERENCES

Biber, D. and Kurjian, J. (2007) Towards a taxonomy of web registers and text types: A multi-dimensional analysis. In M. Hundt, N. Nesselhauf, and C. Biewer (eds.) *Corpus Linguistics and the Web* (pp. 109-132). Amsterdam: Rodopi.

Biber et al. (2004) *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus.* Princeton, New Jersey: Educational Testing Service.

Biber, D. (1988) *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Cigankova, N. (2009) *Linguistic Variation in English Computer-Mediated Academic Discourse.* Synopsis of Ph.D. Thesis. Riga: University of Latvia.

Collot, M. and Belmore, N. (1993) Electronic language: a new variety of English. In J. Aarts, P. Haan and N. Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation* (pp. 41-55). Amsterdam: Rodopi.

Conrad, S. (2001) Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In S. Conrad and D. Biber (eds.) *Multi-dimensional Studies of Register Variation in English* (pp. 94-107). Harlow: Pearson Education.

Crystal, D. (2001) *Language and the Internet.* Cambridge: Cambridge University Press.

Gries, S.T. (2003) *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placemen*t. New York, London: Continuum.

Grieve, J., Biber, D., Friginal, E., and Nekrasova, T. (2011) Variation among Blogs: A Multi-dimensional Analysis. In A. Mehler, S. Sharoff and M. Santini (eds.) *Genres on the Web.* Springer.

Halliday, M.A.K. and Matthiessen, C.M.I. (2004) *An Introduction to Functional Grammar.* London: Hodder Arnold.

Herring, S. C. and Paolillo, J. C. (2006) Gender and genre variation in weblogs. *Journal of Sociolinguistics.* Vol. 10 (4): 439–459.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-Based Language Studies: An Advanced Resource Book.* London, New York: Routledge.

Shortis, T. (2000) *The Language of ICT.* London: Routledge.

Stuart, K. (2006) Towards an Analysis of Academic Weblogs. *Revista Alicantina de Estudios Ingleses* 19: 387-404.

INTERNET SOURCES

1) *European Association for the Teaching of Academic Writing* (EATAW) homepage. Available from http://www.eataw.eu/listserv/ [Accessed August 12, 2009].

2) *Academici Knowledge Network.* Available from http://www.academici.net/ [Accessed August 12, 2009].

3) *The Community of Teachers Using WebCT/ Blackboard Technology Homepage.* Available from http://www.blackboard.com/Communities.aspx [Accessed August 12, 2009].

Nataļja Cigankova (Dr. Philol., Assist. Prof.) is currently working at the University of Latvia. Her research interests include corpus-based studies of English in use, computer-mediated academic discourse and new applications of cutting-edge technologies in computer-assisted language learning. Email: *natalja.cigankova@lu.lv*