様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２７　日現在

機関番号：　３２６０６
研究種目：　挑戦的萌芽研究
研究期間：　2014～2017
課題番号：　２６５８００７７

研究課題名（和文）Investigating a Learner Corpus of Computer-mediated Communication

研究課題名（英文）Investigating a Learner Corpus of Computer-mediated Communication

研究代表者
ＭＡＲＣＨＡＮＤ　Ｔｉｍ（Marchand, Tim）

学習院大学・国際社会科学部・准教授

研究者番号：２０６４５１９７

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：本研究は、大学生によるコンピューター・メディア・コミュニケーション（CMC）から学習者コーパスを構築・分析し、CMCを一ジャンルとして定義することを目的とした。データ収集は、大学英語クラスで行われ、4年間で約54万語の学習者コーパスを構築した。さらに、英国放送協会（BBC）ウェブサイト上の読者コメント欄から英語ネイティブスピーカコーパスを構築し、英語学習者コーパスとの比較を試みた。CMCは他のコミュニケーション形式に比べて、説得力が増すコミュニケーションスタイルであること、また、英語学習者の方がCMCを用いたコメントにおいて、より個人的な関わりを示す傾向があることが明らかとなった。

研究成果の概要（英文）：This project investigated a new type of learner corpus research, the computer-mediated communication (CMC) of Japanese university students. The data comes from a course where lesson materials were provided on a news-based blog. Each week, students wrote their reactions to news stories on the blog, and these comments formed the basis of the learner corpus. After 4 years of data collection, the learner corpus reached 540,000 tokens in size.
The learner corpus together with a reference corpus from a BBC news website were analysed to see how the genre of CMC may be described relative to more traditional forms of communication such as speech and writing. Results showed that CMC tends to be more overtly persuasive in form than the other modes. The difference between learner and native-speaker CMC was also analysed to reveal that learners exhibited much greater personal involvement in their texts. Finally attempts to account for learner proficiency in their CMC were met with mixed results.

研究分野：　Learner corpus research

キーワード：　learner corpus　CMC　genre　learner proficiency

１．研究開始当初の背景

Recent reviews of the current state of learner corpus research (LCR) have called for the expansion of the types of tasks and genres of learner data collected, some of which may better reflect the real-world forms of nativeproduced data often found in reference corpora (Granger, 2009).

In this project, we researched one example of a new genre and task type, computermediated communication (CMC). CMC has been recognized as an effective way of connecting with the current population of students, since blogging and social networking are modes of communicating that many language learners use in their daily lives. However there are few examples of LCR using CMC data in the literature, and what examples there are tend to be small-scale and focused on synchronous "chatting" (e.g. Belz and Thorne, 2006; Belz and Vyatkina, 2008). This project proposed to rectify that by building a corpus of learner CMC and analyse its contents with reference to native-speaker norms of CMC as a genre.

２．研究の目的

The learner corpus was based on data of computer-mediated communication (CMC) between university students taking English language courses with a focus on studying news and current affairs. In short, the objectives were:

1) build a learner corpus of computermediated communication data to replicate in task design the user contributions to the BBC website used as a reference corpus;

2) define learner and native-speaker CMC as genres through multidimensional analysis;

3) compare and contrast the learner and reference corpora;

4) create a systematic method for assigning proficiency to corpus texts and grade the corpus by proficiency level;

5) identify interlanguage development and correlate it with certain learner variables

３．研究の方法

The project adopted two approaches that have become mainstays in corpus linguistics research to meet its objectives: multidimensional analysis (Biber 1988) and contrastive interlanguage analysis (Granger 1996, 2015). Specifically, the methodology for achieving the 6 objectives was as follows:

1) The learner corpus has been collected from four cohorts at three different private universities in Japan. Each cohort consists of between 2-3 classes, and all the participants in a class were from the same year group of (mainly) non-English majors studying a required English course as part of their degree programme. The data are the online texts that the learners submitted to the class blog. These comments are grouped together in threads following a news article that was usually the focus of study during the class. The comments were exported from the blog into XML files, and stripped and pre-treated for character coding errors using a text editor.

The data for the expert corpus comes from the comments I collected as part of my previous study of the CMC on the BBC news website (Marchand 2013). At the time, I considered the collection of texts as a native-speaker corpus, but for the purposes of this project, it will be regarded as an expert corpus since the first language (L1) background of all the contributors to the website cannot be confirmed, although their competency is apparently sufficient enough to both pass through the BBC's moderation, and interact with other users who are L1 English speakers.

2 & 3) Defining the genre of CMC for the two corpora was conducted by using the Multidimensional Analysis Tagger (MAT), which offers an approximate replication of the tagger used in Biber's original study (Nini 2015). The tagger was used to analyse the two CMC corpora and the four components of the BNC Baby corpus (SPOK, ACAD, FICT, NEWS) to check for its reliability. A second approach was tried using POS ngrams as the unit of analysis in a new exploratory factor analysis (EFA).

4, 5 & 6) In order to assign proficiency ratings to the learner comments, we sought to create a text-centred approach that led to the design of a proficiency rating measurement tool. The tool incorporates the key concept of complexity, accuracy and fluency (CAF) and uses a Performance Decision

Tree(PDT) for each component of the CAF construct. Two raters used the PDT on a random sample of comments from the corpus, and the results were assessed for validity and reliability. In addition to this text-based method some learner metadata has also been collected from each cohort, in the form of a questionnaire containing information on the learners' linguistic background (e.g. the length of time they have studied English, whether they have lived overseas, scores on any English language proficiency tests) and engagement with their (language) learning (including interest in the lesson topics, self-reporting on their motivation to study, and self-assessment of some transferable skills).

4．研究成果

The scale and scope of the project's objectives turned out to be more challenging than at first anticipated, so not all of the goals were met. However, the results for each ob follows:

1) The amount of data collected can be seen by looking at the size of the two (Table 4).

Table 4.1:Size of the CMC corpora

Corpus	Code	Number of texts	Number of tokens
BBC news website	BBC	346	1.5 mil
News-based English	NBE	146	540,000

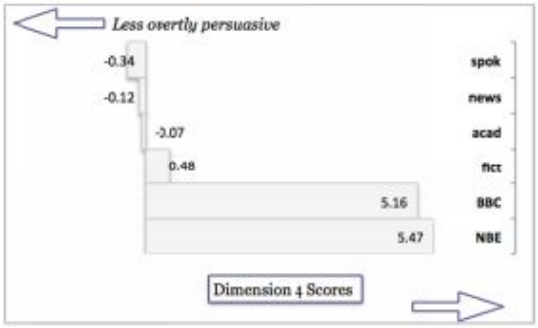
2) The Multidimensional Analysis Tagger (MAT) produced plots for the 4 BNC sub-corpora in line with expectation, proving its general suitability as a substitute to replicate the Biber model. Tabel 4.2 shows the resulting approximate replication of Biber's study, giving five Dimension (D1 - D5) scores for each corpus.

Tabel 4.2:Results of the MAT analysis

Code	D1	D2	D3	D4	D5
SPOK	28.46	1.14	2.06	0.34	2.13
ACAD	11.14	2.02	6.09	0.07	4.54
FICT	0.07	5.61	0.15	0.48	0.91
NEWS	13.48	0.27	3.54	0.12	0.64
BBC	0.33	1.63	2.44	5.16	1.08
NBE	11.54	1.70	3.44	5.47	2.49

Of special interest was how the two CMC corpora align along Dimension 4 in comparison with the other corpora, as can be seen in Figure 4.2.

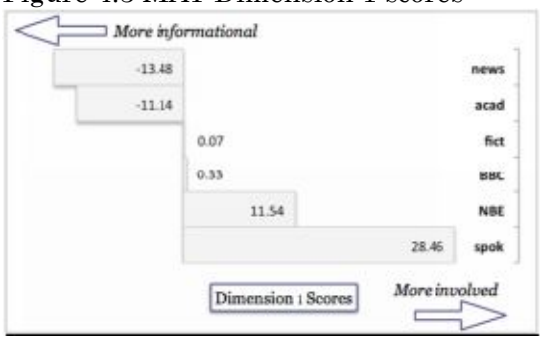
Figure 4.2:MAT Dimension 4 scores



The result reveals that CMC as a genre has much more overtly persuasive language compared to other modes of communication, further confirmed by the MAT tagger identification of both CMC corpora as "Involved persuasion" text types.

3) The multidimensional analysis also revealed a contrast between the expert and learner CMC corpora in terms of Dimension 1 in Biber's model (Figure 4.3).

Figure 4.3:MAT Dimension 1 scores



Low scores on this variable indicate that the text is informationally dense, presenting many nouns, long words and adjectives(among other features). Meanwhile, high scores along this Dimension indicate that the text is affective and interactional, with many verbs and pronouns, for example. As figure 4.3 shows, the BBC CMC corpus demonstrates an even balance between these two poles, and has a similar overall Dimension score as the BNC Baby Fiction corpus. The learners, on the other hand, exhibit much greater personal involvement in their texts.

4) The measurement tool was tested on 1,200 comments. The number of observed agreements between the two raters was 66.5%, which is below the 70% threshold that is generally considered to be an adequate level of

agreement. However, when using Cohen's weighted kappa to establish the reliability of the ratings, the resulting score of k=0.691 suggests that the strength of agreement between the two raters may be considered "good", but this was still deemed insufficient to justify expanding the use of the tool to the whole corpus as using the tool itself was very labour-intensive. An analysis of the questionnaire data, especially focusing on language learning histories and reported scores in proficiency tests revealed that the 4 cohorts in the study can be placed into levels of CEFR equivalent proficiency (Table 4.4).

Table 4.4: CEFR levels of learners

CEFR	A1	A2	B1	B2	C1	NA
	13%	19%	39%	16%	2%	10%

5) After the project abandoned a text-centred approach to assigning proficiency in favour of a learner-centred approach, it has been necessary to recalibrate the method for identifying learner interlanguage development and correlating it with certain learner variables. This is an ongoing concern, and so there are no results to present at this time. However, it is expected that further research using mixedeffects regression models will bear fruit, with Initial testing suggesting that the most significant correlations occur between learner engagement variables and development, rather than proficiency level.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）
① The genre analysis of expert and learner corpora of news-based computer-mediated communication
Tim Marchand
International Computer Archive of Modern and Medieval English 37 May 2016

②The genre classification of texts from expert and learner corpora of computer-mediated communication
Tim Marchand
Learner Corpus Research 2015 Sep 2015

③Computer-mediated communication as source and resource in an EFL course for university students in Japan
Tim Marchand and Sumie Akutsu
17th World Congress of the International Association of Applied Linguistics Aug 2014

④Motivation and the Learner Corpus
Tim Marchand
11th Teaching & Language Corpra Conference 2014 Jul 2014

⑤ Course Design and Material Development Using Computer-Mediated Communication in a Japanese EFL Context
Tim Marchand and Sumie Akutsu
International Symposium on ICT for Language Learning and Teaching Mar 2014

〔学会発表〕（計 3 件）
① Akutsu and Marchand (2015) ComputerMediated Communication for Course Delivery and Teaching Materials Development: A Case Study. International Journal of ComputerAssisted Language Learning and Teaching 5(3) 1-19 Aug 2015

② Marchand and Akutsu (2015) The Compilation and Use of a CMC Learner Corpus for Japanese University Students. In Castello et al. (eds) Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment. Frankfurt: Peter Lang.

③ Marchand and Akutsu (2015) First steps in assigning proficiency to texts in a learner corpus of computer-mediated communication. In Callies & Gotz (eds) Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment. Amsterdam: John Benjamins

〔図書〕（計　　件）

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
MARCHAND Tim（Marchand, Tim）
学習院大学・国際社会化学部・准教授
研究者番号：20645197

(2)研究分担者
阿久津　純恵（AKUTSU,Sumie）
東洋大学・ライフデザイン学部・講師
研究者番号：20460024

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）