Do we talk (or write?) differently over the Net?
- A lexical enquiry into ‘a’ Net-EN -

JUN ARATA TAKAHASHI
UNIVERSITY OF ESSEX

1. Introduction

Emails, Internet discussion fora, chat rooms...a growing number of people use network connected computers for their personal, educational, and job-related correspondence. Some are very formal and some others are very casual. When people are in such computer mediated communication (CMC), do they talk (or write) differently from conventional written or oral communications? In lodging an on-line application form, for instance, it is easy to imagine that the form can be identical to a printed hard copy. The question is, however, what would it be that ones specific in computer mediated communications such as those emails, Internet fora, and chat rooms? Are they used in the same way as personal or professional letters (writing) or telephone conversations (speaking)? Herring (1996) defines the characteristics of the language in CMC as:

> ...it is typed, and hence like writing, but exchanges are often rapid and informal, and hence more like spoken conversation...CMC is not homogeneous, but like any communicative modality, manifests itself in different styles and genres, some determined by the available technologies (e.g., real-time “chat” modes, as opposed to asynchronous e-mail), others by human factors such as communicative purpose and group membership.
>
> (pp. 3-4)

She also identifies that users of CMC often use unique characters and acronyms conveying special meaning or shortening the time of composing.

2. Previous research

2.1 Linguistic enquiries and language in CMC

Due to its anonymity and unique presentation in its communication, Herring (1996) identifies two major foci of research aspects in this field: socio-linguistic aspects (e.g., gender, identity, personality, and mode of their interactions) and linguistic aspects (e.g., lexis, syntax, phrase, mode, discoursal and genre functions).

As is mentioned in the introduction, the issue of linguistic mode (speaking and writing) needs to be considered. Baron (2000) identifies some competing agenda and models in linguistic enquiry: ethnographic (function of form), technological (medium and message), and pedagogic (grammar) agenda and opposition, continuum, and crossover models between writing and speaking language. In terms of ethnographic agenda or descriptive language analysis, Tannen (1982), from her research on oral and literate narrative strategies, suggests that the features distinguishing one discourse from another are not only written or spoken mode but also genre and register. From another perspective on genre, Crystal (1991), inspired from a previous study with his colleagues on grammars of language impaired patients, shows an example of a stylistic language profiling analysis between newspaper articles and legal documents. The profiles include five categories: graphetics, graphology, phonology, grammar, and semantics. These categories are rated scales in three aspects, frequency, distinctiveness, and precision. He reports that legal documents’ distinctiveness is in nominal group and clause structure while newspaper articles’ is in sentence structure and connectivity. In his conclusion, Crystal (1991) foresees the invaluable outcomes of the research into stylistic aspect of language across different genres.

Biber (1988 and 1995) has developed the notion of Multi-Dimensional analysis over the years, which categories the nature of genre into six dimensions. He argues for forms and styles are more genre specific in English language rather than medium oriented. In that sense, his research findings support the crossover model between writing and speaking language. However, does it apply to the emerging new mode, computer mediated communication (CMC) or it has something different from the cousins?

As for such language models, Baron (2000) sees language in CMC through the continuum model with the idea of the crossover view of language: posting web version of printed articles such as online journal can be seen as the products of writing which are static while Internet discussion forum, chat rooms, and email correspondence can be called a process and thus dynamic communication. Even within this dynamic communication, however, there can be further two types of interactions, one is synchronous such as Internet relay chat (IRC) and chat rooms and the other, asynchronous (spontaneous responses though not synchronous) such as emails and BBS (Herring, 1996 and Simpson, 2002). Simpson (2002) investigates the former for discoursal features (turn taking and unique utterances) in IRC and instant messaging. He has found that like face-to-face conversation, interaction in IRC maintains strong coherence in turn taking despite the fact that multiple conversations may occur between more than two groups of participants in one message screen. For instance, interlocutors employ repetitions of style like the ones Tannen (1989) has found in face-to-face conversation. The present study will look into one of the latter form of interaction, which is dynamic and spontaneous but not synchronous communication, Internet forum.

Not from the discoursal ones but from lexical perspectives, the present research will deal with the texts in Internet discussion forum. English language nowadays can be regarded as a multi-national language (EML), rather than inter-national language, which belongs to every user of English regardless weather it is mother tongue or not. Because of this, the present study will focus on the language change in the effect of the medium used among Japanese EML users. The concept of English as a multi-national language (EML), in Brutt-Griffler's (2002) term 'World English', can be described as "World English is not simply made *through* speakers of other languages but *by* them" (p. ix). In terms of the impact of entire speech community on second language acquisition (SLA), she further introduces a new concept, "macroaqcuisition"; a new English is acquired through and nurtured by the speech community. Brutt-Griffler (2002) categorises "macroaqcuisition" into two types: *Type A* macroaqcuisition occurs in a multi-lingual setting that has adopted another unifying language and will develop an entirely new speech community while *Type B* occurs as transformation of a monolingual community into a bilingual one (p. iix). In the case of English as foreign language (EFL) setting, however, it does not fit into either of those but it can be categorised as another type, Type C, which introduces another (international) language as a tool of communication with other parts of the world rather than within the country. A publicly opened Internet discussion forum using English as its medium can be one of such *macroaqcuisition* settings falling in this Type C and it can be one of the gateways to English as multi-national language (EML), which belong to everyone who uses it.

The present research will seek if texts in such a unique bleed of communication and acquisition setting reveal another side of English as multi-national language today.

2.2 Lexical and functional research into Net-EN

Lexical functionality in CMC English (hereafter Net-EN)[1], general written English (LOB), and spoken English (London-Lund) are compared in Yates (1996). He has found that the Net-EN shows significantly different lexical choices from the other

[1] The Net-EN investigated in Yates (1996) is English language in Open University on-line discussion groups in U.K. See Yates (1996) for further details.

two modes and Net-EN can be seen as a unique mode of medium. He adopts Halliday's (1978) concept of language function (textual, interpersonal, and ideational) in analysing Net-EN and comparing it with written and spoken English from lexical functional perspectives. In texts, the textual function deals with information and surrounding context, the interpersonal function realises the relationship of the participants in the texts, and the ideational function expresses the speakers'/writers' external experience and/or internal reality (Halliday, 1978). Yates (1996) also adapts methods of other researchers' previous studies to analyse these three functions of Net-EN.

First, for the textual function, the three modes of communication (written, spoken, and Net-EN) are compared against type/token ratio and then lexical density. In simplest term, lexical density is "the number of lexical items [content words] as a proportion of the number of running words [total words]" (Halliday, 1985: 64). In short, the higher the lexical density, the more information is included in the texts. Referred as a previous investigation in Yates (1996), Ure (1971) has investigated the difference of the lexical density in both written and spoken texts with a focus on their register variances:

> The distinction between spoken and written texts is one of medium, and with it can be grouped all those other distinctions due to the physical conditions of the delivery of the text, in sound, in space and in time. The time available to prepare a text is also a factor that could be important on the dimension of physical circumstances.
>
> (p. 447)

From the lexical aspect, Yates (1996) conducts a similar investigation with Ure's in written (LOB), spoken (London-Lund), and Net-EN. Yates reports that the level of lexical density is highest in written texts, then Net-EN and spoken with the lowest and the differences of the levels between them are statistically significant ($p<0.05$). He suggests that the way delivering information in Net-EN is more like written texts than spoken ones[2].

Yates (1996) then counts personal pronouns as a barometer of interpersonal function of texts[3]. Total number of first, second, and third pronoun use in written, spoken, and Net-EN are all statistically different from each other, however, he notices that the relative distributions of these three kinds of pronouns are distinctively different among the three modes of communication: Net-EN's proportion is very different from written one but somewhat similar to the one for spoken, but Net-EN uses more first and second pronouns (64%) than the other two (27% for written and 58% for spoken) (Yates, 1996: pp.42-3).

Finally, Coates's (1983) model of analysing five categories of modal auxiliaries is adapted to Yates's (1996) investigation to the ideational function[4]: the five categories are obligation and necessity (e.g., must), ability and possibility (e.g., can), epistemic possibility (e.g., may), volition and prediction (e.g., will), and hypothetical modals (e.g., would). He has found that the use of all modal auxiliaries in Net-EN is greatest among the three modes and significantly different in every category from the other two modes of communication except in epistemic possibility against written mode. Again, the relative distribution of modal auxiliaries in those five categories in Net-EN is similar to spoken one (Yates 1996).

From the findings of Yates (1996), although the actual values of lexical frequency in the three Hallidayan linguistic functions are significantly different from the other two modes of communication, one can interpret that the users of Net-EN deliver information (textual) as if they are writing texts but use personal

[2] See Yates (1996) for further discussions about the ease or difficulty in understanding the texts and about the repetition of the same or similar lexical items.

[3] See Chafe (1982) and Chafe and Danielewicz (1987) and Fowler and Kress (1979) for the original and further discussions on this issue.

[4] See Coates (1983) for detail.

pronouns (interpersonal) and modal auxiliaries (ideational) as if they are speaking.

As Ure (1971) notified, Collotte and Belmore (1996) on the other hand use categories of on-line and off-line language (a categorisation on the bases of the time available) rather than conventional written and spoken language. They apply Biber's (1995) Six Multi-Dimensional linguistic variation analysis model for informativity, narrativity, explicitness, persuasion, abstraction, and elaboration in analysing on-line and off-line international Internet bulletin board (BBS) English. Biber (1988) describes the notion of Multi-Dimensional linguistic variation analysis in three ways:

1) Similar or the same type of texts shares particular types of linguistic features;
2) Those features are measured in six bi-polar scales (dimensions) from positive to negative weights (except Dimension 4 and 5) rather than existence or non-existence of those features; and
3) Conceptualisation of dimensions is not predetermined but they are the results of quantified empirical findings.

A summary of textual features of the six dimensions (Biber, 1988 and 1995) and Collotte and Belmore's (1996) findings is as follows:

Dimension 1: Involved versus informational production

Nouns, word length, prepositions, high type-token ratio, and attributive adjectives are negative features in the scale of this dimension. Carefully elaborated long sentences with long words and frequent use of nouns are typical in texts in informational processing. In this Dimension 1 continuum, Collotte and Belmore's Net-EN falls between romantic fiction and personal letters.

Dimension 2: Non-narrative versus narrative

Past tense, third person personal pronouns, perfect aspect verbs, and public verbs are strong positive markers of narrative discourse while present tense verbs and attributive adjectives are negative features which make non-narrative discourse. In this continuum, their Net-EN falls between press reviews and interviews.

Dimension 3: Situation-dependent versus elaborated reference

Time and place adverbials and some adverbs are positive weights of situation-dependent texts while relative constructions, such as WH-relative clauses, are seen as features of elaborated reference. In the continuum, their Net-EN has scored very similar to humour and press reportage.

Dimension 4: Overt expression of argumentation

This dimension only has positive features. Infinitives, prediction modals, suasive verbs, conditional subordination, necessity modals, and split auxiliaries. These features are commonly found in professional letters and editorials but virtually non-existence in press reviews and broadcasts. In this dimension, Collotte and Belmore's Net-EN appears between personal letters and editorials which are close to the end of positive scale. However, both 'off-line' and others[5]in the Net-EN score closer to hobby and romantic fiction which fall in-between 'off-line' and others.

Dimension 5: Abstract versus non-abstract styles

This dimension only has negative features. Conjuncts, agentless passives, past participial adverbial, postnominal clauses, and by-passives are most important features of abstract styles, such as technical and engineering prose and

[5] For some 'on-line' texts, it is difficult to identify whether they are really on-line or pre-constructed off-line ones or not, therefore Collot and Belmore (1996) group their texts 'off-line' and 'other'.

official documents. In this continuum, their Net-EN appear between religion and official documents, however the Net-EN scores are not close to these genre and Collotte and Belmore further report that their Net-EN does not have the features which play a major role of this dimension.

Dimension 6: On-line informational elaboration marking stance

Demonstratives, some *that*-clauses (as verb complements, in object position, and as adjective complements), final prepositions, and demonstrative pronouns are typical markers of informational spoken ('on-line') discourse, e.g., prepared speeches and public conversations, while edited or non-informational texts like most fictions have phrasal co-ordination as their feature. In this last dimension, Collotte and Belmore's Net-EN appear close to editorials and professional letters.

In sammary, Collotte and Belmore (1996) suggest that the genre across Net-EN, such as BBS, Internet forum, and on-line news bulletin, will likely show different textual features to each other but the genres which most resemble their Net-EN are public interviews and personal and professional letters.

All those researches support that Net-EN can be a third variety of language mode. In the present study, another kind of Net-EN is investigated, which is a Net-EN by Japanese EML (English as multi-national language) users.

3. Data Gathering and Method of Analysis

3.1 The source and the nature of the Net-EN corpora (NC)

The texts are gathered from an internet BBS maintained by Keio University, Shonan Fujisawa Campus (SFC) in Japan, which is a part of their EFL curriculum. This aims to develop students' interactive writing and communicative skills in computer mediated communication. This BBS, called IWC[6] (Interactive Writing Community), has various sub-fora discussing from serious social and political issues (e.g., Euthanasia) to hobbies and sports. It also invites not only other university students but also participants from other countries without any status restrictions.

The present research uses Net-EN corpora (NC) containing two sub-corpora, Base Text Corpus (BC) and Response Text Corpus (RC), of approximately 564,000 token from about 3050 base and response articles. The average length of an article in BC is about 370 words and for RC, 160 words. In order to achieve as general analysis as possible, selected 190 most common words are used. In the Net-EN corpora (NC), the 190 words occur more than the most commonly used proper noun that is specific in IWC, *sfc*, which is the abbreviated campus name of the university organising this forum. The 190 words cover 68.6%, 67.0%, and 70.2% of NC, BC, and RC respectively[7]. These words occur at least either 100 times in BC or 89 times in RC per 100,000 words. Any comparison between corpora and between given words are conducted within these 190 words.

As is advocated in the introduction, the present research investigates English as multi-national language (EML) in the Internet discussion forum that is computer mediated *communication*. In this principle, any initial posting which has not attracted any response is regarded as neither EML in use nor a part of communication. Any articles, therefore, fall in the criteria below are excluded from the corpora:

[6] http://www.sfc.keio.ac.jp/iwc/IWC/index_2002f.html for the current fora. Previous IWC discussion fora can be accessed by replacing the part after the underscore (_) in the URL address to the year and semester looked for, e.g., for Spring semester in Year 2000 will be ".../index_2000s.html".

[7] Only as a guide, the 190 words cover approximately 59.4% of BNC spoken corpus and 48.6% of BNC written corpus.

1) Base articles that have not received any response;
2) Any articles posted by other than Japanese participants; and
3) Any articles posted by high school age Japanese.

This maintains the NC, BC, and RC as corpora of EML in use by Japanese English users with certain linguistic level which enable the participants to communicate with others in English. Further figures of the NC, BC, and RC are in Table 1 below.

Table 1: A profile of the Net-EN Corpora (NC, BC, and RC)

Profile of the Net-EN Corpora (NC, BC, RC)					
Name of the BBS Forum	No. of articles in BC	Size of BC (token)	No. of articles in RC	Size of RC (token)	Size of NC (token)
IWC	409	151,000	2636	413,000	564,000

3.2 Method of analysis

As a preliminary comparison, the 50 most frequently used lexis are listed and compared between written and spoken English from British National Corpus (BNC) and the Net-EN corpus (NC). As its main analysis, it applies the same approach above in comparing the Base Text Corpus (BC) and Response Text Corpus (RC). A selected word with unique uses will then be investigated in the light of senses, collocates, and patterns[8] of use. *Collins COBUILD English Dictionary* (CCED) (1995) and Francis, Hunston and Manning (1996) will be referred to for this purpose.

4. Results

4.1 A cross comparison of frequency information between written, spoken, and Net-EN.

Like the findings in Yates (1996) reported in the earlier section, in the frequency list, a largest selection of prepositions and functional words (e.g., determiners and conjunctions) appears in written English, next Net-EN, and then the smallest in spoken English though some functional words specific to spoken language appear. Various types of content words are most used in Net-EN, next spoken English, and then least in written English. Like spoken English, Net-EN appears with high use of the first and second personal pronoun while written English the third person comes most frequently used pronoun. Further in the Net-EN, the word *I* is actually the most commonly used lexis while both written and spoken English have the in their first rank (See Table 2 below).

Table 2: List of 50 most frequent word from in three corpora[9] [10]

	Written (BNC)	Spoken (BNC)	Net-EN (NC)		Written (BNC)	Spoken (BNC)	Net-EN (NC)
1	the	the	I	26	but	well	be
2	of	I	to*	27	from	so	very

[8] Hunston and Francis (2000) *Pattern Grammar* for theoretical detail of patterns.

[9] Frequency ranks of both written and spoken English in British National Corpus (BNC) are obtained and adapted from Leech, Rayson, and Wilson (2001) and thus it is a guide only.

[10] Word forms with '*' include all the possible part of speech they are used in the corpora.

3	and	you	the	28	which	oh	can
4	to*	and	and	29	she	got	as
5	a	it	is	30	they	not	people
6	in	that*	of	31	or	are	they
7	is	to*	you	32	an	if	like*
8	was	a	in	33	were	with	do
9	it	of	a	34	as	no	about
10	for	in	that*	35	we	she	what
11	that*	we	it	36	their	at	if
12	with	is	my	37	been	there	on
13	he	do	for	38	has*	think	essay
14	be	they	we	39	that	yes	when
15	on	er	think	40	will	just	school
16	I	was	have*	41	would	all	many
17	by	what	but	42	her	can	because
18	at	he	your	43	there	then	from
19	you	but	so	44	all	get	at
20	are	for	was	45	can	did	by
21	had*	erm	not	46	if	or	there
22	his	be	are	47	who	like*	also
23	not	on	this	48	said	would	life
24	this	this	me	49	do	mm	don't
25	have*	know	with	50	what	them	want

A list of frequency rank of written, spoken, and Net-EN is provided in Table 2 above. The 50 most common word forms store up to 47.2% for Net-EN corpus (NC), 46.3% for Base Text Corpus (BC), and 48.1% for Response Text Corpus (RC). This shows a brief picture of the differences of the lexical use across the three different mediums.

The frequencies of most common verbs are also noticeable. For instance in the first 50 most common words, *had*, *have*, *has*, and *said* are most common verb forms in written English, *know*, *got*, *think*, and *get* for spoken, and *have*, *think*, *had*, and *thank* for Net-EN. These lexis appear in the very different ranks in the frequency list.

In addition, the frequencies of the top 190 words show statistically significant differences between the Net-EN corpora (NC, BC, NC) and BNC (spoken and written) ranging from p<0.001 to p<0.032 except between the Base Text Corpus (BC) and BNC spoken corpus (p<0.083) although it is a guide only. However, no significant differences have been found between the sub-corpora of Net-EN.

5. Findings and Implications for Further Research

5.1 A close investigation of a selected word in the use and patters

One of most noticeable findings is that major verbs with high utility and in use are used in unique ways in the research corpora (Net-EN). The difference is not only the frequency of the word (including inflections) but also the ways (senses and patterns) they are used. The word *make* in the NC, for instance, is most commonly used as a causative use while the sense, carrying an action (e.g., *make a suggestion*), is the first in CCED which is based on the frequency information in Bank of English (See Table 3 below).

Table 3: Frequencies of uses of *make* in Net-EN corpora (per 100,000 tokens).

	Order of uses in CCED	Net-EN (NC)	Base Text (BC)	Res. Text (RC)
1	carrying an action	69	64	73
2	causative	140	137	143

3	creating or producing	89	83	95
4	others	16	22	10
	TOTAL	314	306	321

As it can be seen, about half the uses of *make* is causative in both BC and RC. Table 4 below further shows the frequency of each pattern of those causative uses. Again, the most common pattern of use is not the same as the one found in a corpus based on general English by native speakers.

Table 4: Raw frequencies of patterns of causative uses of *make* in Net-EN corpora[11].

	Order of appearances in CCED	NC	BC	RC
1	V n inf.	309	81	228
2	be V *to* inf.	8	2	6
3	V n n	28	9	19
4	V n adj.	394	100	294
5	V *it* adj. (that)/to inf.	30	7	23
*	V n be adj.	12	2	10
*	V n to inf.	22	6	16
*	V n –ing	2	2	0

This is possibly an example showing the unique use of English produced by Japanese users of English as multi-national language (EML) in today's context, computer mediated communication (CMC). Further research of this kind can reveal such unique use of English in CMC among Japanese EML users. This will contribute to capture a diverse use of English language today. Looking into the root of their unique use of lexical items can be a valuable and promising further research topic.

6. References

Baron N 2000 *Alphabet to email*. London, Routledge.

Biber D 1988 *Variation across speech and writing*. Cambridge, Cambridge University Press.

Biber D 1995 Dimensions of register variation. Cambridge, Cambridge University Press.

Brutt-Griffler J 2002 *World English: A study of its development*. Clevedon, Multilingual Matters Ltd.

Coates J 1983 *The semantics of the modal auxiliaries*. London, Croom Helm.
Collins COBUILD English Dictionary 1995 London, HarperCollins.

Collot M, Belmore O 1996 Electronic language: A new variety of English. In Herring S 1996 (ed) *Computer-mediated communication: Linguistics, social and cross-cultural perspectives*. Amsterdam, John Benjamins. pp 13-28.

Crystal D 1991 Stylistic profiling. In Aijmer K, Altenberg B (eds) *English Corpus Linguistics*. London, Longman.

Francis G, Hunston S, Manning E 1996 *Collins COBUILD Grammar Patterns 1: Verbs*. London, HarperCollins.

[11] Patterns with '*' do not appear in CCEC as patterns for *make*.

Halliday M A K 1978 *Language as social semiotic: The social interpretation of language and meaning*. London, Edward Arnold.

Halliday M A K 1985 *Spoken and written language*. Oxford, Oxford University Press.

Herring S 1996 Introduction. In Herring S (ed) *Computer-mediated communication: Linguistics, social and cross-cultural perspectives.* Amsterdam, John Benjamins. pp 1-10.

Hunston S, Francis G 2000 *Pattern Grammar.* Amsterdam, John Benjamins.
Leech G, Rayson P, Wilson A 2001 *Word frequencies in written and spoken English: based on the British National Corpus*. Harlow, Pearson Education Ltd.

Simpson J 2002 Discourse and synchronous computer-mediated communication: Uniting speaking and writing? In Spelman Miller K, Thompson, P (eds.) *Unity and diversity in language use: Selected papers from the Annual Meeting of the British Association for Applied Linguistics held at the University of Reading, September 2001*. London, Continuum. pp. 57-71.

Tannen D 1982 Oral and literate strategies in spoken and written narratives. *Language,* (58): pp 1-21.

Tannen D 1989 *Talking voices: Repetition, dialogue and imagery in conversational discourse*. Cambridge, Cambridge University Press.

Ure J 1971 Lexical density and register differentiation. In Perren, G E, Trim L M (eds.) Applications of Linguistics: selected papers of the second International Congress of Applied Linguistics, Cambridge, 1969. Cambridge, Cambridge University Press.

Yates S J 1996 Oral and written linguistic aspects of computer conferencing: A corpus based study. In Herring S (ed.) *Computer-mediated communication: Linguistics, social and cross-cultural perspectives*. Amsterdam, John Benjamins. pp 29-46.