Discourse Structures in Instant Messaging:
The Case of Utterance Breaks

Naomi S. Baron
American University

Abstract

Both users of CMC and the popular press commonly assume that online platforms such as email and instant messaging (IM) mirror informal spoken language. The present study investigates the validity of this assumption by examining discourse structures in IM conversations between American college students. Linguistic features of spoken and written language were first compared both paradigmatically and empirically, drawing particularly on research on intonation units by Chafe (1980, 1994). A subsequent fine-grained analysis of the grammatical points at which subjects chunked their IM turns into multiple transmissions revealed that while IM conversations between male dyads tended to resemble spoken discourse according to this dimension, IM conversations between females bore more similarities to traditional written language.

Introduction

Structurally, computer-mediated communication (CMC) can be defined in terms of two basic parameters. The first is synchronicity. In synchronous CMC, transmission is essentially instantaneous, and interlocutors are assumed to be physically present to read and respond to messages, whereas in asynchronous CMC, neither of these assumptions holds. The second parameter is whether the communication is one-to-one (i.e., between two people) or many-to-many (i.e., multiple participants' messages being broadcast to multiple potential interlocutors).[1] Table 1 illustrates the resulting four classes of CMC.

	Synchronous	Asynchronous
one-to-one	instant messaging (IM)	email, texting on mobile phones
many-to-many	chat, computer conferencing	bulletin boards, listservs, blogs, social networking sites

Table 1. Types of computer-mediated communication

In practice, users often cross category lines. For example, contemporary email (an asynchronous technology) often transmits messages in near-real time, and many users reply immediately, rendering the technologically asynchronous medium effectively synchronous.[2] Conversely, in the case of instant messaging (IM), although the technology is designed to work synchronously, it often is used asynchronously, especially by college-aged students (Baron, 2005). Similar caveats apply to the paradigmatic distinction between one-to-one and many-to-many communication. With media designed for dyadic message exchange such as email and IM, duplicate copies of messages are commonly

sent to third parties through use of the copy, blind copy, forwarding, or cut-and-paste functions. Conversely, many-to-many formats such as listservs or blogs sometimes effectively become one-to-one exchanges that take place before audiences of non-participatory readers.

The technology of CMC has generally dictated that computer-mediated messages are written.[3] However, end users often comment on the informal speech-like quality of the medium (e.g., Lee, 2002). A persistent question intriguing Internet researchers has been whether the stylistic features of CMC are more like those of informal speech or paradigmatic writing (e.g., Baron, 1998, 2000, 2003; Collot & Belmore, 1996; Crystal, 2001; Danet, 2001; Ferrara, Brunner, & Whittemore, 1991; Hård af Segerstad, 2002; Herring, 2002, 2003; Maynor, 1994; Yates, 1996).

One challenge in interpreting some of the CMC literature (e.g., Crystal 2001; Ferrara, Brunner, & Whittemore, 1991) is that it generalizes across different genres of CMC (e.g., email, chat, web pages), while usage patterns across genres may show considerable disparity. Moreover, results have typically been stated ahistorically, for a single point in time (one exception is Herring, 1998), while usage patterns sometimes evolve. A case in point is the growing tendency for younger users to see email as a relatively formal medium compared with IM (Shiu & Lenhart, 2004), a perception that contrasts with the 1990s view of email as stylistically informal (Crystal, 2001; Hale & Scanlon, 1999). Third, many forms of CMC reveal considerable stylistic variation, reflecting such considerations as user age and gender, level of user experience, communication function, and medium for composing and receiving a CMC message (e.g., full computer keyboard versus mobile phone keypad) (Herring, 2007). Given the concomitant variance found in traditional spoken and (off-line) written language (Tannen, 1982a, 1982b; Chafe & Tannen, 1987), it is often unclear against what norm CMC data are being compared. Finally, the preponderance of CMC research has focused on many-to-many communication, rather than on dyadic CMC. This is understandable, in that many-to-many communication is often publicly available and hence easier to access and study; however, private CMC is more commonly used and hence important to understand, as well.

Another problem with existing characterizations of CMC as a spoken or written modality is that few empirical attempts have been made to evaluate CMC data against comparable spoken or written corpora. Exceptions are Collot and Belmore (1996) and Yates (1996). However, the CMC data they examined were of many-to-many communication (from a bulletin board system and a computer conferencing system, respectively), not of one-to-one genres such as email and IM. Moreover, their CMC corpora predate the explosion in CMC activity over the past decade, and CMC language patterns may have evolved in the interim.

In recent years, CMC research designs have become increasingly fine-tuned. Herring (2003) has investigated the linguistic effects of gender and synchronicity in American CMC. Hård af Segerstad (2002) has compared the language of Swedish email, chat, IM, and text messaging on mobile phones. Ling (2004) has considered the effects of age and gender on the linguistic character of Norwegian text messages.

This article contributes to the empirical study of the spoken or written nature of CMC by analyzing the discourse structure of English-language instant messaging. Specifically, it examines a common stylistic convention found in the way adolescents and young adults compose and transmit instant messages. That convention is to break down ("chunk") single utterances into several components, which are then transmitted seriatim, rather than typing out the entire utterance and then transmitting it all of a piece. For example, rather than send the whole utterance 'that must feel nice to be in love in the spring with birds chirping and frogs leaping' as a single transmission, a user might send a sequence of short transmissions, e.g.:

IM Transmission 1:   that must be nice
IM Transmission 2:   to be in love
IM Transmission 3:   in the spring
IM Transmission 4:   with birds chirping
IM Transmission 5:   and frogs leaping

The study reported in this article analyzes the linguistic character of utterance chunks used by a group of American college students in their IM conversations with peers. The eventual goal of this analysis is to enable researchers to compare IM utterance breaks with the kinds of chunking found in informal face-to-face communication, thereby furthering the ongoing discussion of whether IM discourse more closely resembles speech or writing.

The article begins with a short overview of CMC as spoken or written discourse. The following two sections summarize relevant IM research from the current literature and introduce the IM corpus used for the utterance break analysis. After examining the utterance breaks themselves, we consider the implications of our findings for broader questions concerning IM as a spoken or written medium, as well as questions pertaining to gender. The article closes with suggestions for future research.

Speech, Writing, and CMC

There is a considerable literature analyzing the relationship between spoken and written language (e.g., Baron, 2000; Biber, 1988; Chafe & Danielewicz, 1987; Chafe & Tannen, 1987; Crystal, 1995; Tannen, 1982a, 1982b). Research suggests that the two modalities often differ in relatively predictable ways. For example, written language is generally more structurally complex than speech, while speech typically has more contractions and more first and second person pronouns than prose.

However, schematic comparisons of speech and writing fail to reveal many of the related properties that are particular to each modality. Given the common propensity to describe IM as a version of spoken discourse, albeit in textual format, I focus here on the related properties of one-to-one spoken conversation, which prototypically occurs face-to-face.

Face-to-Face Spoken Conversation

One such property is the overall structuring of the conversation: What is the length and the speed of the discourse? What is the length (in words or time) of individual turns?[4] What is the lexical and grammatical composition of each turn? A second conversational property is turn-taking and overlaps between turns (Sacks, Schegloff, & Jefferson, 1974). Among the relevant questions here are: How do interlocutors take and hold the floor? What constitutes a felicitous reply to an interlocutor's utterance? When may one interrupt an interlocutor's turn? How does one keep track of multiple conversational threads that overlap? Third is the question of how interlocutors open and close conversations (Schegloff & Sacks, 1974): Is special language used? How many turns (and how much time on the clock) does it take to initiate or terminate a conversation?

Finally, there is the question of whether (and, if so how) speakers divide their turns into smaller units. As Chafe (1980, 1994, 2001), Crystal (1975), Halliday (1967), Swerts and Geluykens (1994), and others have observed, speakers involved in either monologic or dialogic discourse commonly divide their utterances into smaller intonation or breath groups, separated by falling or rising intonation or by a pause. Within the scope of a single turn, a speaker might utter a sequence of smaller chunks, such as:

Chunk 1:   I was wondering
Chunk 2:   whether you're coming to dinner tonight
Chunk 3:   or you need to work.

Division of speaker turns into what Chafe calls 'intonation units' is directly pertinent to the present study, since interlocutors in IM conversations commonly chain together sequences of shorter IM transmissions while continuing to hold the floor. Are these sequences of IM transmissions analogous to intonation units in spoken face-to-face conversation? If so, does the analogy support the argument that IM tends to be a speech-like form of communication?

Over time, Chafe has refined his notion of an intonation unit (see Chafe, 1980, pp. 14-15; Chafe & Danielewicz, 1987, p. 95; Chafe, 1994, ch. 5). However, the primary features of the intonation unit remain these:

• ends with a clause-final intonation contour (i.e., a rising or a falling pitch)
• begins with at least a brief pause
• begins with a conjunction (typically and, although alternatively but or so)
• syntactically, the intonation unit is likely to be a single clause, although some clauses extend over several intonation units

A speech segment need only have one of these characteristics to qualify as an intonation unit, although, by itself, the presence of a pause in the speech stream is not sufficient to signal a new intonation unit. For example, in their coding of a conversation between members of academia, Chafe and Danielewicz (1987, p. 95) indicate that clause structure takes precedence over pauses in determining intonation units:

Intonation Unit 1:   [pause] I just this year have [pause] dropped down to
                                teaching half time.
Intonation Unit 2:   [pause] Which is what I've always wanted.

Thus, Chafe's criteria are not fully consonant with the way IM transmissions work. In analyzing IM, the closest analogue to a pause is the physical transmission of an IM message. Moreover, while Chafe's speech samples include both monologue and dialogue, the IM data are overwhelmingly dialogic.

CMC as Spoken or Written Discourse

Does CMC more closely resembles face-to-face speech or paradigmatic written language? Drawing upon the existing literature, which looked predominantly at data from email, bulletin boards, and computer conferencing, Baron (1998) concluded that as of the late 1990s, CMC was essentially a mixed modality. That is, it resembled speech in that it was largely unedited; it contained heavy use of first and second person pronouns, present tense, and contractions; its level of formality was generally low; and CMC language could be rude or even obscene. At the same time, CMC looked like writing because interlocutors were physically separated, and that separation fostered personal disclosure and helped level the conversational playing field between interlocutors at different points on a social hierarchy. Moreover, CMC resembled writing in that the medium was durable, and interlocutors commonly employed a wide range of lexical choices and complex syntax.

Crystal (2001) constructed a detailed analysis of distinct types of CMC – including web pages, email, chat, and virtual worlds (e.g., MUDs and MOOs)[5], comparing each with his own paradigmatic investigation of spoken and written language. Applying the term "Netspeak" to the collective forms of language used in CMC, Crystal concluded that "Netspeak has far more properties linking it to writing than to speech … Netspeak is better seen as a written language which has been pulled some way in the direction of speech than as spoken language which has been written down" (2001, p. 47).

Both Baron's and Crystal's conclusions are based on findings by other researchers, and none of the studies referenced involved instant messaging. The present article uses original empirical data to study IM as a specific form of CMC.

IM as a Form of CMC

While instant messaging began in the 1980s, the medium gained popularity with the introduction of ICQ in 1996 and then of America Online Instant Messenger (AIM) in

1997 (Baron, 2003; Herring, 2002). In the United States, AIM was the predominant IM platform among teenagers and young adults in the early 2000s, although MSN Messenger and Yahoo!Messenger were commonly used as well.

In writing about IM, it is necessary to be clear about the users whose behavior we are describing. The popularity of IM in the late 1990s was overwhelmingly due to adoption by teenagers and young adults of college age. Later, IM became increasingly common in the workplace (either for conducting business or for personal communication). However, one should not assume that the IM conversations of, say, young teenage girls are linguistically isomorphic with those of mid-level business managers.

Most studies of IM (e.g., Boneva, Quinn, Kraut, Kiesler, & Shklovski, 2006; Grinter & Paylen, 2002; Issacs, Walendowski, Whittaker, Schiano, & Kamm, 2002; Lenhart, Rainie, & Lewis, 2001; Nardi, Whittaker, & Bradner, 2000; Schiano, Chen, Ginsberg, Gretarsdottir, Huddleston, & Isaacs, 2002; Shiu & Lenhart, 2004) have looked at the social dimensions of the medium (e.g., who uses it, how often, for what purposes). With a few exceptions, there has been little detailed empirical analysis of the linguistic character of IM. Hård af Segerstad (2002) studied a limited IM system known as WebWho, which was designed for a Swedish university computing lab to indicate the presence of other users logged onto the system. In analyzing her corpus, Hård af Segerstad looked at word frequency and message content, but not at other linguistic characteristics of the conversations. Baron's (2004) analysis of American college student IM conversations is described later in this article.[6]

Several other IM studies are suggestive as to the linguistic nature of IM. Randall (2002) commented on the use of emoticons, acronyms, and abbreviations, along with reduced attention to grammar in IM conversations. Jacobs (2003) noted that the American teenaged girl she observed sometimes edited her IM messages before transmitting them, a practice that Nardi et al. (2000) also reported among adults using IM in the workplace.

American University IM Corpus: General Findings on Discourse Scaffolding

I assembled a corpus of IM conversations constructed by college-aged young adults in the United States.[7] Before introducing the corpus, I define some terminology below.

Terminology

Since there is no conventional linguistic terminology for describing IM data, I began by identifying appropriate units for analysis. The most important of these divisions is what I call the IM transmission unit. An IM transmission unit constitutes a segment of text that is composed and sent by one member of the IM conversational dyad. Linguistically, a transmission unit may correspond to one or more sentences (each composed of at least

one independent clause) or to a sentence fragment. Each of the following examples from the corpus constitutes an IM transmission unit:

full sentence: whatever the case, i suggest that you not worry about sitting on
                        the couches
multiple sentences: and the prof left—he forgot something in his office
sentence fragment: if the walls could talk

IM transmission units are somewhat analogous to Chafe's (1980, 1994) notion of intonation units in speech, in that both units divide larger utterances into smaller segments, although, as we will see later in the article, the comparison is not precise. Table 2 delineates the terminology adopted for analyzing the IM corpus in the present study.[9]

Transmission Unit: composition (by typing) and transmission of an IM
     e.g., Max: hey man

Utterance: a sentence in IM consisting of at least one independent clause or clause fragment
     e.g., Susan: Somebody shoot me!

Sequence: one or more IM transmissions sent seriatim by the same interlocutor
     e.g., Max: hey man
           Max: whassup
           [this sequence equals 2 IM transmission units]

Utterance Chunking: breaking a single IM utterance into two or more transmissions
     e.g., Joan: that must be nice
           Joan: to be in love
           Joan: in the spring

Utterance Break Pair: two sequential transmissions that are grammatically part of the same utterance
     e.g., Joan: that must be nice
           Joan: to be in love

Break Point (<BR>): the location of the utterance break between two members of an utterance break pair
     e.g., Joan: that must be nice

Utterance Chunk: one transmission in a break pair transmission sequence
     e.g., Joan   to be in love

Closing: series of transmissions (between interlocutors) at the end of an IM conversation, beginning with one party initiating closure of the conversation and ending with termination of the IM connection
     e.g., Sam: Hey, I gotta go [first indication that will terminate conversation]
                  ... [subsequent conversational transmissions]
           Sam: I'm outta here [final transmission in conversation]

Table 2. Terminology for IM analysis

The IM Corpus

The instant messaging corpus was collected in April 2003 from 22 college-aged undergraduate students who were attending school or had graduated the previous semester. Using America Online's freely downloadable program AIM (AOL Instant Messenger), IM conversations were initiated by a cohort of current students (or recent graduates) at American University in Washington, DC. These conversational initiators ("student experimenters") were asked to IM peers on their AIM Buddy List.[10] At the beginning of each conversation, initiators requested permission of their interlocutor to save the IM conversation that was to follow. Formal consent forms were distributed electronically to all parties (student experimenters and their conversational partners) at the end of the IM conversation. Both interlocutors were then given the opportunity to edit out any words or transmissions they wished removed from the corpus (an option rarely taken),[11] and screen names were anonymized. Once the consent forms were completed, student experimenters electronically forwarded the IM conversation files (and consent forms) to a project computer site. Initial discussion in the IM conversations regarding the research project was eliminated from the data, thus precluding an analysis of conversational openings.

The corpus consisted of 23 distinct IM conversations. Nine conversations took place between females (FF) and nine between males (MM). An additional five conversations involved female/male dyads (FM). In a number of the FF and FM conversations, a single student experimenter conversed with several people on his or her Buddy List. Due to last-minute attrition among student experimenters, most of the MM conversations were between the same two interlocutors.

The 23 IM conversations contained a total of 2,185 IM transmission units, made up of 11,718 words. Some of the analyses discussed below were performed on the entire corpus, while others were restricted to comparison of the nine FF and nine MM conversations (together totaling 1,861 IM transmission units).

Linguistic Variables Analyzed

In the general study, three sets of linguistic variables were examined: discourse scaffolding (essentially, how conversations were constructed), lexical issues, and gender. I begin here with discourse scaffolding, since it is the foundation upon which the present study of utterance breaks is based. At the end of the article, I will summarize previously reported findings regarding lexical and gender issues, with the goal of creating a fuller profile of the linguistic nature of IM. Table 3 delineates the discourse scaffolding issues that were examined.

Transmissions:	average transmission length (in words)
	% of one-word transmissions (out of total transmissions)
	transmissions per minute
	longest transmission (in words)
Sequences:	average number of transmissions per sequence
	% of sequences including >1 transmission
	longest sequence per conversation
	% of utterance breaks
Conversation Length:	entire conversations
	closings

Table 3. Issues involving discourse scaffolding

Transmissions

Table 4 summarizes the use of individual transmission scaffolding.

Average transmission length (in words):[a]	5.4
Percent of one-word transmissions (out of total transmissions):[b]	21.8%
Longest transmission (in words):[b]	44
Transmissions per minute: [a]	4.0

[a] Calculated over the entire data base of 2,185 transmissions.
[b] Calculated over just the FF and MM conversations, containing a total of 1,864 transmissions.

Table 4. Transmissions summary

The aggregated data indicate that average IM transmissions are fairly short (5.4 words per transmission). Measured against Chafe and Danielewicz's findings (1987, p. 96) that spoken conversational intonation units averaged 6.2 words while the 'punctuation unit' for written academic papers averaged 9.3 words, the average IM transmission unit more closely resembles face-to-face speech.

Transmission-length averages obscure some of the important characteristics of contemporary college-student IM conversations, including the high proportion of one-word transmissions (21.8% of the FF and MM combined corpus). Similarly, some IM transmissions are quite lengthy, the longest in the corpus being 44 words. AIM has a technical limit of 1,024 characters per IM transmission, of which about 950 are available for the message (the remaining characters are used for formatting). If we assume an average word length of six characters, with each word then followed by one additional

character (a space or punctuation mark), the 950-character limit permits roughly 136 words per IM transmission, which far exceeds the lengths appearing in the corpus.

Using AIM's time-stamping feature, I calculated both how long each conversation lasted and transmission rate. On average, there were 4.0 transmissions per minute. Given that the mean transmission length was 5.4 words, users averaged 21.6 words per minute.

Sequences and Utterance Chunking

I defined a sequence as one or more consecutive IM transmissions from the same interlocutor. If a sender's multiple transmission sequence was interrupted by a message from his or her interlocutor but the initial sender did not attend to the interruption, all consecutive transmissions dealing with the same theme were considered part of a single sequence. Table 5 summarizes the findings regarding sequence length and overall frequency of utterance chunking in the data.

Number of sequences in total data base:	1,292
Longest IM sequence per conversation (i.e., sequential transmissions from same interlocutor)	18
Average number of transmission units per IM sequence:	1.7
% of IM sequences including >1 transmission unit:	42%
% of transmissions involved in utterance break pairs:	16.2%

Table 5. Sequences and utterance chunking

The first step in the analysis was to cluster the data into sequences. There were 1,292 sequences, built from the total 2,185 transmissions. Sequences ranged in length from 1 to 18 transmission units. The average number of transmission units per sequence was 1.7. Out of 1,292 sequences, 42% contained more than one transmission unit.

Almost one-sixth of the 2,185 transmissions (16.2%) constituted utterance break pairs, that is, consecutive transmissions that split an utterance into grammatical chunks. Later in this article, I examine the grammatical nature of these utterance breaks.

Conversation Length

Table 6 presents findings on overall conversation length and length of closing sequences.

ENTIRE CONVERSATIONS	
Average transmissions per conversation:	93.8 transmissions
Average time per conversation:	23.7 mins.
CLOSINGS	
Average number of transmissions:	7.0 transmissions
Average time to close:	31.9 secs.

Table 6. Conversation length

As an aggregate, IM conversations are fairly lengthy, averaging over 93 transmission units apiece and lasting nearly 24 minutes. Individually, however, these conversations show enormous variety, ranging from quick three or four transmission volleys to discussions stretching over more than 200 transmissions and well over an hour. Moreover, although the communication channel may remain open for an extended period, interlocutors are not necessarily engaging with one another throughout that time. For example, one lengthy FF conversation (142 transmissions, 88 minutes) included a 15-minute gap when no transmissions occurred.[12]

Much as in face-to-face spoken encounters (e.g., Sacks, Schegloff & Jefferson, 1974), IM interlocutors often take a while before terminating a conversation. In this example of an IM conversational closing between female interlocutors, the sequence took 19 seconds:

Gale: **hey I gotta run**
Sally: Okay.
Sally: I'll ttyl?
Gale: **gotta do errands.**
Gale: **yep!!**
Sally: Okay.
Sally: ?
Gale: **talk to you soon**
Sally: Alrighty

On average, closing sequences involved seven transmissions and took nearly 32 seconds.

IM Utterance Break Analysis

I now turn to the question of how users chunk IM utterances into sequences of two or more IM transmission units. Specifically, I consider the syntactic composition of the second chunk in IM utterance break pairs with the goal of better understanding the spoken or written character of IM. I then compare these pair breaks with Chafe's notion of spoken intonation units and also investigate whether gender is a relevant variable in IM break pair usage.

The simplest case of a multi-turn IM transmission sequence is one that does not involve syntactic breaks. In the following sequence, the message sender continues to hold the floor, but the transmissions themselves are syntactically distinct:

    Jill: awww…;-(
    Jill: i'm sorry
    Jill: if it makes you feel any better, i'm being held captive by two of Julie's papers

Contrast this example with sequential IM transmissions in which each transmission constitutes a chunk of a single utterance:

    Joan: that must be nice
    Joan: to be in love
    Joan: in the spring

The IM Break Pair Corpus

To study the syntactic nature of IM break pairs, a subset of the larger corpus was analyzed, consisting of all conversations between female interlocutors (FF) and between male interlocutors (MM) (i.e., eliminating mixed gender conversations). The nine FF conversations contained 1,097 transmissions, comprising 272 multi-transmission sequences. Within the nine MM conversations, there were 767 transmissions, composed of 182 multi-transmission sequences. The analysis had three components:

  • tabulating the proportion of multi-transmission sequences containing utterance breaks
  • tabulating the number of total transmissions involved in utterance breaks
  • syntactically coding and analyzing the pairs of transmissions between whose chunks a syntactic break point occurred

Table 7 summarizes the proportion of multi-transmission sequences containing one or more utterance breaks.

Number of Multi-Transmission Sequences Containing One or More Utterance Breaks		Total Multi-Transmission Sequences in Corpus	% of Total Multi-Transmission Sequences with Breaks
FF	61	272	22.4%
MM	71	182	39.0%
Total	132	454	29.1%

Table 7. Multi-transmission sequences with utterance breaks

Out of 454 multi-transmission sequences in the combined FF and MM corpus, nearly 30% contained utterance breaks. To illustrate:

Transmission 1:   he just played here in DC last weekend
Transmission 2:   and by doing so he violated NCAA rules which gaurantees [sic]
                   he is gonan [sic] go pro
Transmission 3:   it's really sorta cool

An utterance break (<BR>) occurs between transmissions 1 and 2, but not between transmissions 2 and 3. Utterance breaks were more common among MM dyads (39%) than among FF dyads (22.4%). That is, if one tallies all sequences that contain at least two IM transmission units, male multi-turn transmissions were significantly more likely to contain a syntactic break than were female multi-turn transmissions (x^2 (1) = 14.54, *p* <.0001).

The second step in the analysis focused on the break pairs themselves. Table 8 summarizes the total number of break pairs in the corpus and the relationship of IM transmissions involved in break pairs to the total number of transmissions in the combined FF and MM corpus.

	Total Break Pairs	Total Transmission Units Involved in Break Pairs[a]	Total Transmission Units in Corpus	% of Total Transmission Units in Corpus Involved in Break Pairs
FF	84	145	1097	13.2%
MM	105	176	767	22.9%
Total	189	321	1864	17.2%

[a] Numbers in this column do not equal double the number of Total Break Pairs, since the same transmission unit is sometimes involved in sequential break pairs.

Table 8. Total transmissions involving utterance breaks

The FF and MM corpus contained 189 utterance break pairs. Since each break pair is composed of two transmission units, one might have anticipated the total transmissions involved in break pairs to be 189 times two, i.e., 378. However, the same transmission unit was sometimes involved in two sequential break pairs. For example, in the multi-turn sequence

Transmission Unit 1:   that must be nice
Transmission Unit 2:   to be in love
Transmission Unit 3:   in the spring

the second transmission unit ('to be in love') is a member of two different break pairs, i.e.,

Break Pair 1: that must be nice
              to be in love

Break Pair 2: to be in love
in the spring

In the corpus, the total number of transmission units involved in break pairs was not 378 but 321.

Consistent with findings reported in Table 7, males tended to have more transmissions involved in break pairs than did females (22.9% of total transmission units vs. 13.2%). That is, if all transmission units involved in utterance break pairs are tallied and those totals are compared with the total number of IM transmissions in the corpus, males used significantly more transmission units involved in break pairs than did females (x^2 (1) = 29.68, p < .00001).

The third part of the analysis entailed syntactically coding the second component of each of the 189 break pairs in the corpus. For example, in the break pair

to be in love
in the spring

the second transmission ('in the spring') serves as an adverbial prepositional phrase, modifying the verb phrase of the previous transmission. In the next section, I present the grammatical coding scheme used for analyzing the utterance break data.

Break Pair Coding Scheme

A grammatical coding scheme was developed for analyzing break pairs. With one exception (addressed below), the coding scheme considered both members of the break pair in coding the syntactic function of the second member of the pair. Nearly all of the coding system followed standard definitions of grammatical categories (e.g., adjective, direct object) and sentential functions (e.g., and is a coordinating conjunction, while because is a subordinating conjunction). The overall grammatical model reflects a simplified subset of early transformational grammar (e.g., Chomsky, 1965) combined with terms from traditional grammatical models (e.g., 'independent clause,' 'subordinating conjunction'), with minor adaptations made to fit the data set. For reference purposes, each of the break pair sets was numbered within the FF and MM data sets, respectively (e.g., FF33 denotes the 33rd break pair set in the female-female corpus).

Table 9 displays the grammatical coding scheme, including examples of each subcategory.[13] Coding the 189 break point sets was generally straightforward. However, a few explanations of decisions made during the coding process render the data analysis more transparent.

Abbreviation Key:
Adj   = Adjective   NP   = Noun Phrase
Adv   = Adverb   Part   = Verb Particle
App   = Appositive   Rel Cl = Relative Clause
Coord   = Coordinating Conjunction   S   = Sentence
DO   = Direct Object   Subord= Subordinating Conjunction
IC   = Independent Clause   VP   = Verb Phrase
IC Frag = Independent Clause Fragment
  = break between first and second
                                 transmission in break pair (used when
                                 analysis of syntactic coding is clarified
                                 through reference to the structure of the
                                 first member of the break pair)

Grammatical Type	Subtype		Example
Conjunctions and Sentences or Phrases Introduced by Conjunction	Coord + S	FF33	and she never talks about him
	Subord + S	FF22	if I paid my own airfare/
	Coord + NP	MM8	or circleville
	Coord + VP	MM41	and had to pay back the bank
	Coord + Adj	MM48	or more
	Coord + Adv	MM20	but all the time
	Coord	FF6	so
	Subord Adj	FF53	though more Japanese than anything else
Independent Clauses	IC	MM27	that's all I'm saying
	IC Frag	FF2	monitor democratic process
Adjectives and Adjectival Phrases	Adj	FF38	completely harmless
	PP	FF47	on Japanese rearmament
	Rel Cl	MM26	that laura even has on her profile
Adverbs and Adverbial Phrases	V + NP PP	FF25	what are you bringing to the party on saturday?
	V + Adv PP	FF20	Kathleen is back in town for a month or so
	V + NP Adv	MM44	but I don't get their mail not here
	V Adv	FF44	i can't read clearly
Noun Phrases	NP App	FF70	thats who you remind me of like garfield
	V DO	MM18	[expletive] you
	Adj Noun	MM73	"THE FAN" radio station
Verb Phrasesª	V App	MM54	I have to at least try at least see if they can stop charging me everyday
	VP Part	FF36	at least not what I know of
	NP VP	MM66	and then Pat McGee Band perform like 7

ª The category Verb Phrase refers to second transmission units in break pairs that begin with an actual verbal element (i.e., verbal appositive, verb particle, or main verb), not with other elements that might occur in a verb phrase (e.g., adverbs, prepositional phrases, direct objects).

Table 9. Grammatical coding scheme

Conjunctions. Traditional grammars identify coordinating conjunctions as the lexical items *and*, *but*, *or*, *nor*, *for*, *so*, *and so*, and *yet*. However, a number of adverbs serving as sentence modifiers play a similar role in linking two independent clauses. These adverbs include *however*, *therefore*, *then*, as well as *and then*, e.g.,

> MM2   they put and [sic] sport, a food, and a numbers [sic] in a big hat
>   then pick one of each

For coding purposes, the second transmission in the MM2 break pair was classified Coord + S.
The data coding system generally analyzed the syntax of the second transmission in each break pair. The only exception to this principle occurred in break pair FF53:

> FF53   year, only ¼ though
>   more japanese than anything

FF53 was coded Subord [= Subordinating Conjunction]
 Adj and listed under the general category of conjunctions, even though the conjunction occurs in the first transmission. This is the only break pair (out of 189) in which the first transmission in the pair ends in a conjunction.

Independent clauses. Two kinds of constructions were coded as independent clauses. The most obvious was independent clauses appearing after a break point where the first transmission in the break pair contained either an adverb or a dependent clause introduced by a subordinating conjunction such as *if* or *since*, e.g.,

> FF26   but if you are only coming for a little bit
>   don't feel pressures

The second type was stand-alone independent clauses appearing after a break point where the first transmission in the break pair was also an independent clause, but the second transmission was clearly linked to the first, e,g.,

> Transmission 1: come on
> Transmission 2: give me a freaking break

Distinguishing between sequences containing utterance break pairs and sequences constituting discrete sentential units was sometimes subjective. However, familiarity with

the data strongly indicated a difference between transmissions within a sequence that were clearly linked and those that were not, e.g.,

Transmission 1:   yeah, i am obsessed with the italians myself
Transmission 2:   maybe i should change my icon to the italian flag instead of
                  this dreaded tweety bird

Appositives. Traditional grammars define an appositive as a noun, noun phrase, or noun clause that refers to the same entity as a preceding noun, noun phrase, or noun clause. For example, in the sentence

My best friend, the town mayor, gave me a tour of City Hall.

the appositive 'the town mayor' refers to the same entity as 'my best friend.' The IM corpus contained a number of examples of nominal appositives, e.g.,

FF42   well, not a real job
       more like an internship

These break pairs were coded NP
 App. However, there was also an instance of what one might call a verb phrase (VP) appositive, in which the second transmission of a break pair was a VP referring to the same action as the VP in the first transmission, i.e.,

MM54 I have to at least try
      at least see if they can stop charging me everyday

This break pair was coded VP
 App.

Grammatical Analysis of Utterance Break Points

After coding each of the 189 break pair sets, totals were tabulated for each subtype and summed for each broad grammatical category. The results are presented in Table 10.

Grammatical Type	Subtype	FF (N=84)	MM (N=105)	Total (N=189)
Conjunctions and Sentences or Phrases Introduced by Conjunction	Coord + S	25	45	70
	Subord + S	6	13	19
	Coord + NP	3	5	8
	Coord + VP	3	4	7
	Coord + Adj	1	1	2
	Coord + Adv	0	4	4
	Coord	1	0	1
	Subord Adj	1	0	1
	Total:	40 (47.6%)	72 (68.6%)	112 (59.3%)
Independent Clauses	IC	9	10	19
	IC Frag	10	0	10
	Total:	19 (22.6%)	10 (9.5%)	29 (15.3%)
Adjectives and Adjectival Phrases	Adj	3	2	5
	PP	1	1	2
	Rel Cl	2	6	8
	Total:	6 (7.1%)	9 (8.6%)	15 (7.9%)
Adverbs and Adverbial Phrases	V + NP PP	8	4	12
	V + Adv PP	1	0	1
	V + NP Adv	0	2	2
	V Adv	1	0	1
	Total:	10 (11.9%)	6 (5.7%)	16 (8.5%)
Noun Phrases	NP App	7	2	9
	V DO	1	3	4
	Adj Noun	0	1	1
	Total:	8 (9.5%)	6 (5.7%)	14 (7.4%)
Verb Phrases	V App	0	1	1
	VP Part	1	0	1
	NP VP	0	1	1
	Total:	1 (1.2%)	2 (1.9%)	3 (1.6%)

Table 10. Tabulation of break pair subtypes

Conjunctions and independent clauses. The most striking result of the grammatical coding of the break pair data is the dominance of conjunctions as the basis for utterance chunking. Out of 189 break pair sets, 112 (59.3%) began the second member with a conjunction. The majority (89 out of 112, i.e., 79.5%) used a coordinating or subordinating conjunction to introduce a sentence (e.g., Coord + S: 'and she never talks about him' or Subord + S: 'if I paid my own airfare/'). Other types of break pairs involving conjunctions (e.g., Coord + NP or Coord + VP) were less frequent (23 out of 112, i.e., 20.5%).

Regrouping the conjunction data with regard to whether the initial word was a coordinating conjunction (e.g., *and*, *but*) or a subordinating conjunction (e.g., *because*, *although*) reveals a strong asymmetry in the distribution, as shown in Table 11.

Conjunction Type	FF	MM	Total
Coordinating	33 (82.5%)	59 (81.9%)	92 (82.1%)
Subordinating	7 (17.5%)	13 (18.1%)	20 (17.9%)
Total:	40	72	112

Table 11. Break pairs beginning with coordinating or subordinating conjunctions

More than four out of five (82.1%) of IM transmissions that appeared as the second member of an utterance break pair and that began with a conjunction used a coordinating conjunction. Only 17.9% of such transmissions began with a subordinating conjunction. A comparison of the 92 transmissions beginning with a coordinating conjunction against the entire corpus of 189 break pairs reveals that 48.7% of all break pairs in the corpus began with a coordinating conjunction.

After conjunctions, the next most prevalent grammatical type for beginning the second transmission in a break pair was independent clauses. Of the 189 break pairs, 15.3% constituted independent clauses. Grammatically, independent clauses are also sentences (or sentence fragments). If one adds the break pairs coded Coord + S or Subord + S (i.e., the combined conjunction category) together with break pairs coded IC or IC Frag (i.e., the combined category of independent clauses), the total (89 + 29 = 118) constitutes 62.4% of the 189 utterance breaks in the IM corpus. Clearly sentence units (whether or not preceded by a conjunction) comprise a significant pattern in constructing IM conversations.

To summarize, the largest grammatical categories for second members of utterance break pairs were conjunctions and independent clauses. If all the conjunction data (N=112) are combined with all the independent clause data (N=29), these two categories account for nearly three-quarters (74.6%) of the 189 break pairs in the IM corpus.

Adjectives, adverbs, noun phrases, verb phrases. The remaining grammatical categories fall into two clusters. In the first (adjectives, adverbs, and noun phrases), each subcategory accounts for less than one-tenth of the data (adjectives: 7.9%; adverbs: 8.5%; noun phrases: 7.4%). Collectively, these three clusters represent 23.8% of the 189 utterance break pairs analyzed. If this 23.8% for adjectives, adverbs, and noun phrases is combined with the 74.6% composed of conjunctions or independent clauses, 98.4% of the total utterance breaks are accounted for.

The second cluster (containing only verb phrases or elements within the verb phrase) included only three examples, i.e., 1.6% of the 189 break pairs:

V <u>App</u>		I have to at least try at least <u>see if they can stop charging me everyday</u>
VP <u>Part</u>		at least not what I know <u>of</u>
NP <u>VP</u>		and then Pat McGee Band <u>perform like 7</u>

In the first example ('see if they can stop charging me everyday'), the phrase constitutes an appositive, modifying the verb 'try.' The second example ('of') most likely is a self-correction to the initial member of the break pair, i.e., 'at least not what I know of.' In the last example, the second member of the pair is an entire verb phrase ('perform like 7') [i.e., perform about seven songs]. Intuitively, an utterance break between the two main constituents of a sentence (i.e., noun phrase and verb phrase) seems a natural place to anticipate finding chunking in IM conversation. However, only one such break occurred in the entire corpus, leading to the surmise that chunking utterances into noun phrases and verb phrases is unnatural in IM conversations.

Another way of looking at adjectives, adverbs, noun phrases, and verb phrases is in terms of more general grammatical function. Of the 48 instances of Adj, Adv, NP, and VP, 41 constitute modifiers. For example:

Adj: modifies noun	FF 38	a really decent dude <u>completely harmless</u>
Adv: modifies verb	FF 44	i can't read <u>clearly</u>
NP: noun appositive	FF 70	thats who you remind me of <u>like garfield</u>
VP: verb appositive	MM 54	I have to at least try at least <u>see if they can</u> stop charging me everyday

If these 41 modifiers are combined with the 112 conjunctions and 29 independent clauses, 182 (i.e., 96.3%) of the total 189 utterance break pairs in the corpus are accounted for.

Gender Variation in Instant Messaging

As shown above, the discourse scaffolding data (Tables 7 and 8) indicate that male interlocutors used significantly more utterance breaks than did females. The data on conjunctions and independent clauses (Table 10) also reveal significant gender differences. Males were more likely than females (68.6% versus 47.6%) to begin the second transmission in a break pair with a conjunction (x^2 (1) = 8.49, $p < .004$). However, females were more likely than males (22.6% versus 9.5%) to chain together related sentences (x^2 (1) = 6.16, $p < .014$). While FF conversations contained roughly equal numbers of full independent clauses (e.g., 'that's all I'm saying') and independent clause fragments (e.g., 'monitor democratic process'), the MM corpus only contained examples of full independent clauses.

Discussion of IM Utterance Breaks

IM as a Spoken or Written Medium

The present data offer suggestive empirical evidence regarding the spoken or written nature of IM. With respect to general discourse scaffolding issues, IM is characterized by relatively short turn length, common use of one-word utterances, and prolonged conversational closings. Moreover, interlocutors often hold the conversational floor by transmitting sequences of short messages. When these IM transmission sequences involve utterance break pairs, the second member of the pair frequently begins with a coordinating conjunction. All of these findings suggest a more spoken than written style.
In an earlier analysis of the IM corpus, I reported on several lexical features, including use of contracted forms (e.g., *don't*) versus uncontracted expressions (e.g., *do not*) (Baron, 2004). Generally, contractions are more common in spoken language, and uncontracted structures more common in writing. In the IM data, out of 763 instances in which a contracted or uncontracted form might appear, contractions were used 65.3% of the time. However, given the popular characterization of IM as an informal medium laden with typing shortcuts, I had anticipated contracted forms would be even more frequent.

Baron (2004) also calculated use of abbreviations, acronyms, and emoticons in the IM corpus. While abbreviations and acronyms are typically devices appearing in writing, the emoticons found in CMC can be thought of as filling some of the prosodic or kinesic functions associated with spoken language. Abbreviations, acronyms, and emoticons all appeared in the IM corpus, although they were relatively infrequent. Out of 11,718 words in the corpus, only 31 were abbreviations specific to CMC (e.g., *cya = see you*), only 90 were CMC-specific acronyms (e.g., *lol = laughing out loud*), and only 49 were emoticons (e.g., ? = smiley).

Table 12 summarizes the discourse scaffolding, utterance break, and lexical findings in the IM corpus with regard to their spoken or written character. On balance, IM more closely resembles face-to-face speech than paradigmatic written language.

	Similar to Face-to-Face Speech	Similar to Paradigmatic Writing
GENERAL DISCOURSE SCAFFOLDING		
• average turn length	yes	no
• one-word utterances	yes	no
• conversational closings	yes	no
UTTERANCE BREAKS (UB)		
• frequency of chunking utterances into multiple sequential transmissions	yes	no
• 2nd member of UB pair begins with conjunction	yes	no
• 2nd member of UB pair begins with coordinating conjunction	yes	no
• 2nd member of UB pair begins with independent clause	no	yes
LEXICAL ISSUES		
• use of contractions	somewhat[a]	somewhat[a]
• CMC abbreviations, acronyms	no[b]	somewhat[b]
• emoticons	yes[c]	no[c]

[a] Some contractions appeared in the corpus, albeit fewer than anticipated.
[b] Some CMC abbreviations and acronyms appeared in the corpus, albeit fewer than anticipated.
[c] Emoticons are typically used in lieu of (spoken) prosodic or kinesic cues available in face-to-face communication.

Table 12. IM as a spoken or written medium

IM Utterance Breaks and Chafe's Intonation Units

In looking for a spoken analogue to chunking IM turns into shorter transmission units, I turned to Chafe's model of discourse analysis, even though the frameworks are not wholly comparable. Recall that in Chafe's model, intonation units are identified in one of four ways: ends with a rising or falling pitch, begins with at least a brief pause, begins with a conjunction (especially a coordinating conjunction), and syntactically, the intonation unit is often a clause, although some clauses extend over several intonation units. Also recall that in Chafe's analysis, the clause criterion outweighs the pause criterion. Chafe judges a speech segment such as 'I just this year have [pause] dropped down to teaching half time' to be a single intonation unit, despite the pause. Moreover, much of Chafe's data derive from monologues rather than from interactive discourse. In

the case of the present IM analysis, the sole criterion for identifying a transmission unit was the fact that a segment of writing was sent to an interlocutor. Moreover, as noted earlier, nearly all of the IM data were dialogic.

Both the IM transmission units and Chafe's intonation units turn out to be relatively short (IM: 5.4 words; intonation units: roughly 6 words; Chafe, 1980, p. 14). In both cases, coordinating conjunctions commonly initiate a new transmission or intonation unit. Moreover, in both instances, new units are sometimes made up of independent clauses. An example from Chafe (1980, p. 15) follows:

Intonation Unit 1:   This time I saw a statue
Intonation Unit 2:   it looked like it was in a park

At the same time, comparison of the IM data with Chafe's intonation units reveals several points at which either the analogy fails or we have inadequate data. We have seen that syntactic breaks between adjectives and nouns, or between noun phrases and verb phrases, were rare in the IM corpus (only one case appearing of each). Chafe, however, reports multiple instances in which a pause occurred between an adjective and a noun, or between a noun phrase and a verb phrase, e.g.:

adjective[14] ('the') – noun ('road'):
        and spilled the pears all over the [pause] road (Chafe 1980, p. 20)

noun phrase ('the picker') – verb phrase ('was picking the pears'):
        where the picker,
        was picking the pears (Chafe 1980, p. 46)

In some instances (e.g., 'where the picker, was picking the pears'), Chafe codes the utterance as two distinct intonation units, presumably because of the falling intonation (indicated with a comma) after 'picker.' In other instances (e.g., 'and spilled the pears all over the [pause] road), Chafe considers the sequence as a single intonation unit, presumably because it constitutes a single clause. Coding system aside, break points occur at syntactic positions in Chafe's spoken corpus where transmission unit breaks rarely appear in IM.

In making an analogy between pauses in Chafe's spoken discourse analysis and transmissions in the IM corpus, I have not taken into consideration ellipses, dashes, and hyphens appearing within individual IM transmission units, e.g.,

haha, ok -- you're right…i just wanted to be in bed like 5 hours ago

Such dashes and ellipses chain together independent clauses, rather than indicating breaks within syntactic structures. Analysis of punctuation in the IM corpus suggests that the vast majority of such pause markers function as in the example above.

In discussing IM utterance breaks that were coded as adjectives, adverbs, noun phrases, or verb phrases, I observed that the majority of such structures constituted modifiers, e.g.:

> Transmission 1:   he is actually <u>a really decent dude</u> [Noun Phrase]
> Transmission 2:   <u>completely harmless</u> [Adjective modifying noun *dude*]

It would be instructive to have comparable data from spoken corpora to determine the extent to which such modifiers appear as distinct intonation units following the nouns or verbs they modify.

Finally, in comparing IM data with speech, one should remember that although both forms of discourse are referred to as conversations, the temporal pacing of IM and spoken conversations may differ significantly. Spoken discourse is prototypically a foregrounded activity. Interlocutors are continuously in one another's physical presence, with politeness conventions requiring them to attend to spoken utterances and reply, as appropriate, in a timely fashion. On the contrary, college student IM conversations often reside in the background, while the interlocutors are immersed in additional activities (see, e.g., de Siqueira & Herring, 2009).

Research regarding online and off-line activities in which American college students are engaged while doing IM (Baron, 2008) suggests that interlocutors contribute to IM conversations at their convenience as they multitask their way through a diversity of computer-based and off-line activities (e.g., using the Web, doing word processing, speaking with another person face-to-face, watching television, or simultaneously conducting multiple IM conversations). An analysis of the time gaps (in minutes and seconds) between transmission units in the IM corpus indicates that while the majority of new transmissions (66.9%) followed within 10 seconds of the prior transmission, 24.5% of new transmissions came after a lag of between 11 and 30 seconds, and 8.2% followed a gap ranging from 31 seconds to over 5 minutes. Admittedly, although most corpora of spoken discourse record interlocutors who are focused on the conversation at hand, real-world conversation is often characterized by participants being distracted, weaving in and out of a conversational stream, or falling silent for periods of time. Thus, while face-to-face speech and IM conversations are, in principle, both synchronous forms of communication, detailed comparative analysis is needed of what else in going on while college students IM one another (or speak face-to-face). Only then can we determine the extent to which the media are comparable, and the degree to which either or both are synchronous or asynchronous.

The Significance of Gender in IM Discourse

The data reported in this article suggest some clear distinctions in the ways male and female college students construct IM conversations. The males in this study are significantly more likely to use multi-turn IM transmissions than are females; males are significantly more apt than females to begin the second member of a syntactic utterance break pair with a conjunction; and females are significantly more prone than males to

begin the second member of a syntactic utterance break pair with an independent clause. Earlier analyses of the IM corpus (Baron, 2004) revealed other gender differences in discourse scaffolding. Females exchanged significantly more IM transmissions to close a conversation and took significantly more time to do so than males (for number of IM transmissions, $t(14) = 2.37$, $p < .05$; for amount of time, in seconds, to close a conversation, $t(14) = 2.56$, $p < .05$).

In addition, Baron (2004) reported significant gender differences regarding contractions and emoticons. Males used contractions in 77.1% of the instances in which a contracted or uncontracted form might appear, while females only utilized contracted forms 57% of the time ($x^2 (1) = 32.8$, $p < .0001$). With regard to emoticons, 12 out of 16 female interlocutors used emoticons, while only one out of six male interlocutors did so ($x^2 (1) = 6.14$, $p < .015$). Similarly, in her studies of chat (which, like IM, is a synchronous medium), Herring (2003) found that females used three times as many representations of smiles or laughter as did males.

Are there gender differences regarding the spoken or written nature of IM? I have reported that college-student IM conversations more closely resemble face-to-face speech than paradigmatic writing. Yet this general conclusion masks important gender distinctions, as shown in Table 13.

	Similar to Face-to-Face Speech		Similar to Paradigmatic Writing	
	Males	Females	Males	Females
GENERAL DISCOURSE SCAFFOLDING				
• conversational closings		longer	shorter	
UTTERANCE BREAKS (UB)				
• frequency of chunking utterances into multiple sequential transmissions	more frequent			less
• 2nd member of UB pair begins with conjunction	more frequent			less
• 2nd member of UB pair begins with independent clause	less frequent			more
LEXICAL ISSUES				
• use of contractions	more frequent			less
• emoticons			more frequent[a]	less frequent[a]

[a] However, as with CMC abbreviations and acronyms, there were relatively few emoticons in the corpus, and most of these were a smiley, i.e., ☺.

Figure 13. The role of gender in rendering IM a spoken or written medium

While the male IM conversations have a great deal in common with prototypical descriptions of face-to-face speech, the female IM conversations more closely approximate paradigmatic writing patterns. The only two exceptions to this generalization are conversational closings and use of emoticons, both of which are more pronounced among the females than among the males, and both of which are more analogous with spoken than written communication.

Future Research

Studies are needed that look at a much wider range of data, including larger and more balanced samples from college-aged subjects; IM conversations of younger users (e.g., middle school and high school students); and findings from adults doing IM in the workplace or communicating with family and friends. Anecdotal evidence suggests that the conversations of younger teenagers are more stylized (e.g., use more abbreviations, acronyms, and emoticons; have poorer spelling; use shorter transmission units) than those of college students, and that the older the users, the more likely their IMs will resemble formal email messages. It will also be of interest to see whether the linguistic characteristics of IM differ between users in the United States and elsewhere in the world. Given that many people in Europe and Asia were avid users of text messaging before taking up IM (the reverse has been true in the U.S.), it remains to be seen if American IM patterns will be reflected in cultures where mobile phone texting is already deeply entrenched.[15]

Second, both deeper and broader linguistic analyses of IM corpora are needed. Among the structural features remaining to be analyzed are grammatical parameters (e.g., questions versus declaratives, grammatical complexity) and sentence mechanics (including punctuation, spelling errors, and self-corrections). With regard to discourse issues, there is a need to perform content analyses and study how interlocutors keep track of multiple conversational threads (cf. Herring, 1999).

Third, it will be important to compare IM conversations with informal speech and informal writing generated by the same demographic cohort. The author is presently analyzing informal speech and writing samples from American college students. The next step will be to collect a new college student IM corpus whose discourse topics match those of the speech and writing sample, thereby reducing the number of variables that currently make comparison of speech, writing, and IM problematic.

Fourth, while this article has reported suggestive differences in the ways male and female college students construct IM conversations, and has tentatively concluded that male IMs more closely resemble speech, while female IMs more generally look like written discourse, it has not explored reasons behind these differences or their implications. Are males, for example, more abrupt or disjointed in face-to-face informal spoken conversations than females – and, if so, why? Do females perceive themselves as putting

more effort into the construction of their IM conversations than do males – and, again, if so, why? Is sexual orientation or personality type, not just biological gender, a relevant variable?

Finally, there is a need to compare empirical findings about the linguistic nature of IM with the growing body of research on other forms of computer-mediated communication, both with respect to specific linguistic properties (e.g., Herring, 1999, 2001; Rintel, Mulholland, & Pittam, 2001; Werry, 1996) and the issue of gender differences (cf. Boneva & Kraut, 2002; Herring, 2003; Ling, 2005). Only then will researchers be in a position to conclude how to characterize the linguistic properties of CMC in general and IM in particular.

Acknowledgments

Notes

1.   Elsewhere (Baron, 2004; Baron & Ling, 2003) I have described the situation of many-to-many CMC as "one-to-many" CMC, reflecting the fact that the message formulated by a single interlocutor ("one") is simultaneously broadcast to multiple recipients ("many").

2.   One established example of essentially synchronous email is Dartmouth College's Blitzmail system (Hafner, 2003).

3.   This article does not consider voice-recognition technology or voice-over-Internet protocols.

4.   In conversational analysis, a "turn" is defined as the language used by a speaker while he or she holds the floor before ceding it or being interrupted by another interlocutor.

5.   MUDs (Multi-User Dungeons/Dimensions), created in the late 1970s, were synchronous environments in which multiple players interacted within a textually-created imaginary setting. MOOs (MUDs, Object Oriented), originating in 1990, used object-oriented programming to model real-world common space in which individuals could interact, often in social or educational contexts.

6.   More recently, Tagliamonte and Denis (2008) analyzed linguistic characteristics of an IM corpus from Canadian teenagers.

7.   The author is grateful to Lauren Squires, Sara Tench, and Marshall Thompson for gathering this IM corpus.

8. In other discussion of this corpus (Baron, 2004), the word "turn" was used in place of what here I call a transmission unit. The terminology employed here reflects the fact that in most discussions of discourse analysis, the word "turn" refers to everything a speaker says while holding the floor. In IM, the equivalent of such spoken "turns" may extend over multiple IM transmissions.

9. Throughout the article, user names have been deleted or altered to preserve anonymity.

10. A Buddy List is a set of IM screen names that users choose to associate with their account. When these users log on to AIM, they can see which people on the Buddy List are currently online.

11. Since these deletions were indicated with "xxx," it was possible to see from the editing that few words or entire transmission units were removed.

12. Baron (2008) reports results regarding additional online and off-line activities in which college students engage while doing IM.

13. Examples maintain original punctuation and spelling. However, some transmissions have been shortened when the omitted words or phrases were not relevant to the syntactic structure being illustrated.

14. Technically, *the* is a determiner, not an adjective. However, structurally, both determiners and adjectives modify or define the scope of the noun that follows.

15. For a comparison between the linguistic characteristics of American IM conversations and text messaging on mobile phones, see Ling and Baron (2007).

References

Baron, N. S. (1998). Letters by phone or speech by other means: The linguistics of email. *Language and Communication*, *18*, 133-170.

Baron, N. S. (2000). *Alphabet to email: How written English evolved and where it's heading*. London and New York: Routledge.

Baron, N. S. (2003). Language of the Internet. In A. Farghali (Ed.), *The Stanford handbook for language engineers* (pp. 59-127). Stanford, CA: CSLI Publications.

Baron, N. S. (2004). See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology*, 23, 397-423.

Baron, N. S. (2008). Adjusting the volume: Technology and multitasking in discourse control. In J. Katz (Ed.), *Handbook of mobile communication studies* (pp. 177-193). Cambridge, MA: MIT Press.

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

Boneva, B., & Kraut, R. (2002). Email, gender and personal relations. In B. Wellman & C. Haythornthwaite (Eds.), *The Internet in everyday life* (pp. 372-403). Oxford: Blackwell.

Boneva, B., Quinn, A., Kraut, R., Kiesler, S., & Shklovski, I. (2006). Teenage communication in the instant messaging era. In R. Kraut, M. Brynin, & S. Kiesler (Eds.), *Computers, phones, and the Internet* (pp. 201-218). Oxford: Oxford University Press.

Chafe, W. (1980). The deployment of consciousness in the production of a narrative. In W. Chafe (Ed.), *The pear stories: Cognitive, cultural, and linguistic aspects of narrative production* (pp. 9-50). Norwood, NJ: Ablex.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Chafe, W. (2001). The analysis of discourse flow. In D. Schriffrin, D. Tannen, & H.E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 673-688). Malden, MA: Blackwell.

Chafe, W., & Danielewicz, J. (1987). Properties of spoken and written language. In R. Horowitz & S. J. Samuels (Eds.), *Comprehending oral and written language* (pp. 83-113). San Diego: Academic Press.

Chafe, W., & Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology, 16*, 383-407.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Collot, M., & Belmore, N. (1996). Electronic Language: A new variety of English. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 13-28). Amsterdam: John Benjamins.

Crystal, D. (1975). *The English tone of voice*. London: St. Martins.

Crystal, D. (1995). *The Cambridge encyclopedia of the English language*. Cambridge: Cambridge University Press.

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

Danet, B. (2001). *Cyberpl@y: Communicating online*. London: Berg.

de Siqueira, A., & Herring, S. C. (2009). Temporal patterns in student-advisor instant messaging exchanges: Individual variation and accommodation. *Proceedings of the Forty-Second Hawai'i International Conference on System Sciences (HICSS-42)*. Los Alamitos, CA: IEEE Press. Retrieved September 12, 2009 from http://ella.slis.indiana.edu/~herring/desiqueira.herring.2009.pdf

Ferrara, K., Brunner, H, & Whittemore, G. (1991). Interactive written discourse as an emergent register. *Written Communication*, 8, 8-34.

Grinter, R. E., & Palen, L. (2002). Instant messaging in teen life. *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW '02) (pp. 21-30). New Orleans, LA, November 16-20. New York: ACM Press.

Hafner, K. (2003). All quiet on campus save the click of keys. *New York Times*, August 14, Circuits Section p.1.

Hale, C., & Scanlon, J. (1999). *Wired style: Principles of English usage in the digital age*. New York: Broadway Books.

Halliday, M. A. K. (1967). Notes on transitivity and theme in English. Part II. *Journal of Linguistics*, 3, 199-244.

Hård af Segerstad, Y. (2002). *Use and adaptation of written language to the conditions of computer-mediated communication*. Department of Linguistics, Göteborg: Göteborg University.

Herring, S. C. (1998). Le style du courrier électronique: variabilité et changement. *Revue d'aménagement linguistique* (formerly *Terminogramme)*, *84-85* (March), 9-16. French version: http://ella.slis.indiana.edu/~herring/terminogramme.1998.pdf . English version: http://ella.slis.indiana.edu/~herring/term.html

Herring, S. C. (1999). Interactional coherence in CMC. *Journal of Computer-Mediated Communication, 4*(4). Retrieved July 1, 2005, from http://jcmc.indiana.edu/vol4/issue4/herring.html

Herring, S. C. (2001). Computer-mediated discourse. In D. Schriffrin, D. Tannen, & H. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 612-634). Malden, MA: Blackwell.

Herring, S. C. (2002). Computer-mediated communication on the Internet. *Annual Review of Information Science and Technology*, *36*, 109-168.

Herring, S. C. (2003). Gender and power in online communication. In J. Holmes & M. Meyerhoff (Eds.), *The handbook of language and gender* (pp. 202-228). Oxford: Blackwell.

Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet*, *4*, article 1. Retrieved September 12, 2010 from http://www.languageatinternet.de/articles/2007/761

Isaacs, E., Walendowski, A., Whittaker, S., Schiano, D., & Kamm, C. (2002). The character, functions, and styles of instant messaging in the workplace. *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW '02) (pp. 11-20). New Orleans, LA. New York: ACM Press.

Jacobs, G. (2003). Breaking down virtual walls: Understanding the real space / cyberspace connections of language and literacy in adolescents' use of instant messaging. Paper presented at the American Educational Research Association, Chicago, April.

Lee, J. (2002). I think, therefore IM. *New York Times*, September 19, Circuits Section p.1.

Lenhart, A, Rainie, L, & Lewis, O. (2001). Teenage life online: The rise of the instant-message generation and the Internet's impact on friendships and family relationships. *Pew Internet & American Life Project*. Retrieved July 1, 2005, from http://www.pewinternet.org/pdfs/PIP_Teens_Report.pdf

Ling, R. (2004). *The mobile connection: The cell phone's impact on society*. San Mateo, CA: Morgan Kaufman.

Ling, R. (2005). The sociolinguistics of SMS: An analysis of SMS use by a random sample of Norwegians. In R. Ling & P. Pedersen (Eds.), *Front-stage / back-stage: Mobile communication and the renegotiation of the social* (pp. 335-350). London: Springer-Verlag.

Ling, R., & Baron, N. S. (2007). Text messaging and IM: A linguistic comparison of American college data. *Journal of language and social psychology*, 26, 291-298.

Maynor, N. (1994). The language of electronic mail: Written speech? In G. Little & M. Montgomery (Eds.), *Centennial usage studies* (pp. 48-54). Tuscaloosa, AL: University of Alabama Press.

Nardi, B., Whittaker, S., & Bradner, E. (2000). Interaction and outeraction: Instant messaging in action. *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW '00) (pp. 79-88). Philadelphia, PA. New York: ACM Press.

Randall, N. (2002). Lingo online: A report on the language of the keyboard generation. Department of English, University of Waterloo and msn.ca. Retrieved July 1, 2005, from http://www.arts.uwaterloo.ca/~nrandall/LingoOnline-finalreport.pdf

Rintel, E. S., Mulholland, J., & Pittam, J. (2001). First things first: Internet relay chat openings. *Journal of Computer-Mediated Communication* 6(3). Retrieved July 1, 2005, from http://www.ascusc.org/jcmc/vol6/issue3/rintel.html

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language*, *50*, 696-735.

Schegloff, E., & Sacks, H. (1973). Opening up closings. *Semiotica*, 7, 289-327.

Schiano, D., Chen, C., Ginsberg, J., Gretarsdottir, U., Huddleston, M., & Isaacs, E. (2002). Teen use of messaging media. *Proceedings of the ACM Conference on Human Factors in Computing Systems* (CHI '02) (pp. 594-595). Minneapolis, MN. New York: ACM Press.

Shiu, E., & Lenhart, A. (2004). How Americans use instant messaging. *Pew Internet & American Life Project*. Retrieved July 1, 2005, from http://www.pewinternet.org/pdfs/PIP_Instantmessage_Report.pdf.

Swerts, M., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech*, 37, 21-43.

Tagliamonte, S., & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83, 3-34.

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

Tannen, D. (1982b). The oral/literate continuum in discourse. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 1-16). Norwood, NJ: Ablex.

Werry, C. (1996). Linguistic and interactional features of Internet relay chat. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 47-63). Amsterdam: John Benjamins.

Yates, S. (1996). Oral and written linguistic aspects of computer conferencing. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 29-46). Amsterdam: John Benjamins.

Biographical Note

Naomi S. Baron [nbaron@american.edu] is a Professor of Linguistics at American University in Washington, DC. Author of *Always On: Language in an Online and Mobile World*, she is presently studying differences between reading in hard copy versus onscreen.