

Chapter 4. Salient features in conversational writing

4.1 Introductory remarks

This chapter presents the salient features of conversational writing, both those that have become conspicuous through the standard score calculations using Biber's (1988) methodology, explained in chapter 3, and those that are salient for other reasons. The chapter serves both as a prelude and a complement to the final results of the application of Biber's (1988) methodology to be presented in chapter 5. The principal aim of the present chapter is to point out, describe and discuss the salient features and the functions they serve in conversational writing. Firstly, we will investigate the use of modal auxiliaries and personal pronouns in conversational writing. Modal auxiliaries and personal pronouns are two of the main carriers of interpersonal meaning in language, defined in Halliday's system of semiotics (1985a, 2004), and therefore will be discussed under one and the same heading in the second section of this chapter (4.2). Their distribution in the conversational writing genres reveals a great deal about the modality of the discourse and the presentation of self, enabling informed contrastive analysis of the chatted and spoken texts. The third section of the chapter, 4.3, investigates the lexical properties of conversational writing by contrasting measures of word length, type-token ratio and lexical density in writing, speech and the conversational writing genres. Sections 4.2 and 4.3 largely draw on the choice of features in Yates' (1993) application of Halliday's semiotics to asynchronous CMC, and thus, chiefly, serve to complement the field of CMC variation studies with the analogous documentation of synchronous data. The two sections are kept together to facilitate for readers to compare the results with those in Yates' 1993 study. The fourth section, 4.4, departs from Yates, but stays closely tuned to Biber's (1988) methodology in that it presents the most salient features annotated in the conversational writing corpora. In the present study, ten features altogether have been found to deviate from Biber's (1988) mean for speech and writing by more than two standard deviations. Two of these are first and second person pronouns, which are addressed in section 4.2. The fourth section of this chapter, section 4.4, presents the remaining eight of these features, and what each of them reveals about the kind of communication going on in the chats. The fifth section, 4.5, goes on to survey the paralinguistic cues and extra-linguistic features found in the conversational writing corpora, and the penultimate section of the

chapter, 4.6, presents two salient linguistic features that are not among Biber's list of features, but that nevertheless serve important functions in computer-mediated conversational writing: inserts and emotives. The last section, 4.7, then sums up the results presented in the chapter.

The genres of conversational writing, IRC and split-window ICQ chat, are subsumed into their respective media categories in the present chapter: the media of synchronous and supersynchronous CMC (SCMC and SSCMC), as explained in chapters 1 and 2. The distributions of the linguistic features in SCMC and SSCMC are contrasted with the distributions in three other media: writing, speech and ACMC (asynchronous computer-mediated communication). ACMC is included in this chapter mostly as a reference point, and will receive rather cursory treatment (the focal concern of the study being synchronous and supersynchronous chat), but its inclusion here serves as a useful reminder of the inherent variability of computer-mediated texts. The five media to be compared in this chapter are represented by the following corpora:

Writing:	LOB + private and professional letters (as sampled by Biber 1988; see Appendix I)
Speech:	LLC (as sampled by Biber 1988; see Appendix I) + SBC subset, i.e. first c. 712 words of each text in SBC part 1 (annotated by Jonsson)
ACMC:	"ELC other" corpus of BBS conferencing (recorded and annotated by Collot 1991)
SCMC:	UCOW, the IRC component (recorded and annotated by Jonsson in 2002)
SSCMC:	UCOW, the split-window ICQ component (recorded and annotated by Jonsson in 2004)

The corpus of ACMC, called "ELC other" ("Electronic Language Corpus other"), was collected and annotated by Milena Collot (Collot 1991, Collot & Belmore 1996). It is not available as raw texts, but was annotated with the Biber tags in Collot's original study and represented as feature count data (in Collot 1991), from which the figures are derived for the present comparison. Collot's corpus consists of messages posted to an international bulletin board system, a BBS, located in Canada. It comprises 115,618 words and was collected from nine conferences, their topics ranging from "Chit-Chat" to "Medical" (Collot 1991: 45). The designation "other" implies that the messages were composed online, as opposed to messages composed offline, which were compiled into a separate corpus ("ELC off-line," not to be considered here). Collot was able to positively identify offline messages as they contained a software-generated marker, and those which lacked the marker were assumed to be written online. Collot, however, notes that

“there is always the possibility that certain messages were pre-written using an ordinary word processor or editor” (Collot 1991: 45), which would not add the marker. By labeling the resulting corpus “other” instead of “on-line” she implies that it contains, but is not necessarily limited to, online texts (Collot 1991: 46).

As mentioned in section 2.5, various modes of ACMC have been studied by linguists and communication scholars over the years, including computer conferencing systems (Korsgaard Sorensen 1993, Yates 1993, 1996, Davis & Brewer 1997), listservs (Herring 1996b), newsgroups (Severinson Eklundh 2010), BBSs (Collot 1991, Collot & Belmore 1996), web fora (LeBlanc 2005), e-mail (Yates & Orlikowski 1993, Maynor 1994, Baron 2000, Zitzen 2004, Anglemark 2009, Cho 2010, Georgakopoulou 2011b, Rowe 2011), weblogs (Scoble & Israel 2006, Anglemark 2009, Peterson 2011) and Twitter (Petrovic et al. 2010, Pak & Paroubek 2010). Collot’s study, however, appears to be the only one to have applied Biber’s (1988) full multidimensional analytical tool to ACMC. The readily available frequency counts in her study lend themselves conveniently to comparison with the feature frequencies found for the corpora annotated in the present study, and with those presented in Biber (1988) for LOB and LLC. Comparable frequency counts are particularly amenable to graphic, diagrammatic representation, which is why, in this chapter, Collot’s ACMC corpus will receive its own representation in the figures, even though, owing to the unavailability of comprehensive raw ACMC texts, the ACMC figures for some features will be left uncommented.

4.2 Distribution of modal auxiliary verbs and personal pronouns

In Bybee & Fleishmann’s (1995a) co-edited volume on modality, Guo (1995) briefly, but pertinently, considers English modals, positing that “[i]n English, physical ability can be expressed either by the modal auxiliary *can* or by the adjective *able*, as in *be able to*. Similarly, social permission can be expressed by *can* or *be permitted to*” (Guo 1995: 228, original italics). In each case, the two options are referentially interchangeable. However, the options differ in their grammatical status; modals belong to a closed grammatical class and are thus more grammaticalized than adjectives and verbs, which leads Guo (1995) to further argue that:

This grammatical difference has significant consequences with regard to the meanings expressed. With lexical forms such as *able* or *permitted*, the speaker presents a fact without any personal involvement. We interpret the utterance as ‘I’m stating X to you’. But when modal auxiliaries are used, the resulting utterances are colored by speaker involvement in the form of opinion, affect, or personal dynamics. We interpret such

utterances as 'I'm challenging/objecting to/arguing with you by stating X to you'. (Guo 1995: 228)

Modality thus indicate the speaker's evaluation of his/her proposition, for instance the gradience of likeliness (if the speech event is a proposition) or desirability (if it is a proposal) (Halliday 2004: 116).

Halliday (2004) discusses finite verbs in terms of what they bring in to the clause and their functions in the systems of polarity and modality. Finiteness is expressed through a verbal operator, which is either temporal (realized by tense) or modal (realized through modal auxiliaries). In the system of polarity, the operators appear in positive and negative form (as e.g. *it is/isn't, do that/don't do that, you can/can't do it*), whereas in the system of modality there are intermediate degrees (e.g. *it must/will/may be, you must/should/may do that*, etc.). Polarity is the choice between yes and no, whereas modality construes "the region of uncertainty that lies between 'yes' and 'no'" (Halliday 2004: 147). In this way, the modality system of a language is an important functional component carrying interpersonal meaning (Halliday & Hasan 1989, Halliday 2004). In fact, Guo (1995: 229) proposes that language actually "developed the grammatical category of modal auxiliaries to serve the function of regulating interpersonal relations in social interaction."

Several studies have found modals to be more common in speech than in writing (e.g. Coates 1983, Biber et al. 1999, Kennedy 2002). Bybee & Fleischman (1995b: 8) make the points that "many modal functions surface only in face-to-face interactive discourse," that is, they depend on dialogic "speaker-addressee interaction" (ibid.) and that "modals can be viewed as strategic linguistic tools for the construction of social reality" (ibid.). In a similar vein, Kennedy (2002) notes that modals reflect the role of modality in face-to-face conversations: "to hedge and soften utterances and express subtle differences in degrees of certainty, attitudes, value judgements and the truth conditions of propositional content" (2002: 88, also noted by Andersen 2006: 18). Lexico-grammatically, interpersonal meaning is carried by e.g. markers of mood (indicative or imperative, but also by interrogatives, e.g. WH-interrogatives, as we shall see later), the use of personal pronouns and the choice of modal auxiliaries (Halliday 1978, 1985a, Halliday & Hasan 1989, Halliday 2004). Together, these features reflect the semiotic "tenor" of the communication (Halliday & Hasan 1989), as touched upon in section 2.4; that is, they reflect the personal relationships involved in the communication.

In their model of critical linguistics, Fowler & Kress (1979) consider what they call the grammar of modality, concentrating upon, among other things, the last two linguistic items mentioned: personal pronouns and modal auxiliaries

(see also Hodge & Kress 1988, Yates 1993: 106). Following their example, albeit in the reverse order, we will look first at the distribution of modal auxiliaries, then at the use of personal pronouns, in the genres of conversational writing. The purpose of the investigation is to find out to what extent these two are used in conversational writing, and how their distribution in these genres relates to that in writing and speech, as well as to that in the medium of ACMC. Given the interpersonal nature of conversational writing, and the importance assigned to modal auxiliaries as carriers of interpersonal meaning, we should expect a distribution in conversational writing similar to speech, or, more specifically, similar to traditional conversation (face-to-face and telephone conversations).

The modals included in Biber (1988), and therefore tagged in UCOW and the SBC subset, are the following:⁵¹

Possibility modals: *can, may, might, could* (+ negated forms)

Necessity modals: *ought, should, must* (+ negated forms)

Prediction modals: *will, would, shall* (+ contracted and negated forms)

The distribution of modal auxiliaries in each medium is illustrated in figure 4.1, based on table 4.1. (In the present chapter, all figures and tables are based on average, normalized frequencies per thousand words and derive from Biber 1988: 247–263 for writing, Collot 1991: 69–70 for ACMC, Biber 1988: 264–269 and Appendix II table 3 for speech, Appendix II table 1 for SCMC, and Appendix II table 2 for SSCMC, unless otherwise indicated.) For the results of statistical significance tests among SCMC, SSCMC, writing and speech for the features treated in this chapter, see Appendix VI.⁵² In the figures and tables, the media are ordered according to their basic synchronicity of communication (cf. table 1.1), i.e. from most asynchronous on the left (writing) to supersynchronous (SSCMC) on the right (although the conversational genres of speech, of course, may exceed SCMC in synchronicity). Immediately noticeable in figure 4.1 is the elevation of the ACMC and SSCMC bars. The texts of the two media display identical and remarkably high distributions of modals (totals of 20.5 per thousand words, compared to 12.8 for writing and 15.1 for speech). The frequent use of modals suggests that communication participants in both media are interpersonally

51 The generic term “possibility modals” here designates the modals marking possibility, ability or permission; “necessity modals” designates the modals marking necessity or obligation, and “prediction modals” the modals marking prediction or volition (Biber 1988: 241, Quirk et al. 1985: 219, Coates 1983).

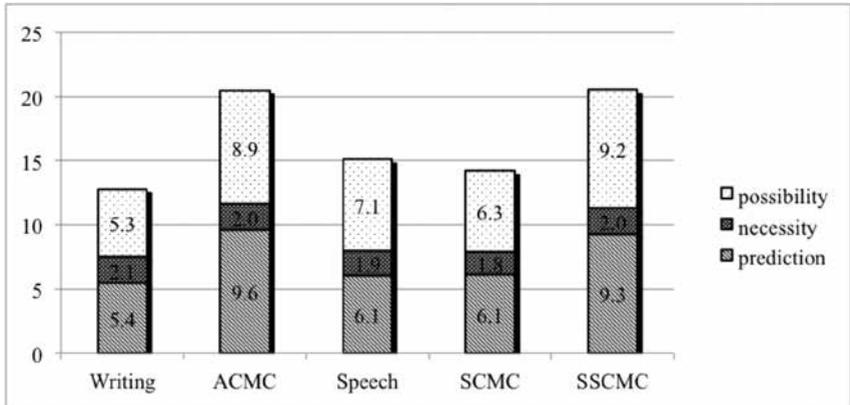
52 No statistical test was carried out on the ACMC frequencies, as the requisite data is not provided in Collot (1991).

involved to a high degree, i.e. they exchange discourse that is “colored by speaker involvement in the form of opinion, affect, or personal dynamics,” to use the words of Guo (1995: 228). With regard to the conversational nature of chatted texts, this seems to be a logical finding, but as for asynchronous texts, it is more unexpected. The overall modal auxiliary use in SCMC (14.2) is also higher than in writing, but slightly lower than in speech, a finding to be discussed later.

Table 4.1: Frequencies of possibility, necessity and prediction modals per 1,000 words (normalized values)

	Writing	ACMC	Speech	SCMC	SSCMC
possibility modals	5.3	8.9	7.1	6.3	9.2
necessity modals	2.1	2.0	1.9	1.8	2.0
prediction modals	5.4	9.6	6.1	6.1	9.3
total	12.8	20.5	15.1	14.2	20.5

Figure 4.1: Distribution of possibility, necessity and prediction modals per 1,000 words (normalized values).



To enable functional comparison of the three CMC media, a brief introduction to the interpersonal aspects of the ACMC corpus is needed. The “ELC other” texts of ACMC (Collot’s material) are unavailable for scrutiny, except for a few examples cited in Collot (1991), but Collot describes the texts as discussions about issues pertaining to the conference topics: e.g. “Medical,” “Finance,” “Sports,” “Current Events,” “Science,” “Cooking,” “Chit-chat” and “Film and Music.” Participants in

a conference are primarily joined by their common interest in the topic, and “social and demographic features rarely seem to play any role” (1991: 36). Nevertheless, in the conferences, personal relations inevitably develop. Collot describes the participants’ relationships in the following way:

People who have been using a bulletin board for a while know each other’s nicknames, mannerisms and ideas. They have followed each other’s arguments on many different subjects, and have accumulated a wealth of shared knowledge. Even people who are new to the board know that their audience will be generally sympathetic because they are bound to them by common interests. The BBS makes for a special kind of intimacy, not often found in other varieties. The messages are similar to personal correspondence because of the shared knowledge and friendly tone. (Collot 1991: 36)

The spare examples of corpus text given in Collot (1991) are taken from the “Chit Chat” conference. Messages are about 50 words long and their asynchronous character is evident in their similarity to e-mail messages; participants identify themselves and their addressees by first and last names, posts are date- and time-stamped and messages have a subject line. Some messages further resemble letters in that they begin and end with greetings, and one even ends with the participant’s signature. The examples in (1) are from Collot (1991: 31) and contain one prediction modal (*’ll*) and one possibility modal (*can*).

(1) Date: 02-04-90 (10:11) CHITCHAT Number: 3 (Echo)
To: SKIP BERTSCH Refer#: 679
From: CLIFF WATKINS Read: NO
Subj: HELLO

Hello Skip! Are you from the Riverside area or another one of those beautiful SoCal cities/counties? I’m looking forward to visiting Cal and hope to make it to the southern portion as well. Mostly I’ll be from S.F. and north. Anyway, your testing echo made it to NY. bye...

Date: 02-05-90 (22:33) CHITCHAT Number: 4 (Echo)
To: CLIFF WATKINS Refer#: 3
From: JONATHAN NEAL Read: 02-06-90
Subj: HELLO (12:33) (Has Replies)

I’m on the same board as Skip’s and I *can* say that the cities here are NOT beautiful.... with the exceptions of Riverside (my hometown), Palm Springs, and some mountain towns. The weather is not spectacular, either. It has been raining the last three days, a high of 63. Oh, well.... Also, I am 12.

see ya....

Jonathan

The personal relationships under development in the BBS approximate the relationships of previous acquaintances, such as those between the chatters in the split-window ICQ corpus. To form and develop relationships, interlocutors on a BBS, as well as in ICQ, need to stay sensitive to each other's opinions and propositions. By modalizing their utterances, they maintain the ongoing dynamics of social interaction. In example (2) from split-window ICQ chat, part of a discussion of college plans between the two high school classmates recorded, participants' previous acquaintance outside of the medium shines through. The example contains three possibility modals (*could*, *can*), one necessity modal (*should*) and one prediction modal (*'ll*).

- (2) <K> so did you get a scholarship fir tennis or are u just going
 <11> yea
 <K> did you talk to the coach
 <K> o
 <11> he said that i have a pretty good chance
 <K> is he goign to come see you play
 <11> yea this season it starts in march
 <11> what school are you going to again
 <K> when do u start
 <K> [college name] its down in [city name]
 <11> are you definetly playing there
 <K> yea i went donw and the coach said i *could* start as a true freshmen
 <11> well thats good
 <K> he came up to talk to my parents and we ate dinner and all kinds of shit
 <11> that cool
 <K> so yea i *can* sing my letter of intent when ever
 <11> thatws cool
 <K> you *should* come down
 <11> yea defenetly
 <K> and i'll come up we *can* chill

Split-window ICQ chat text 10 (UCOW)

While the BBS conference participants seem inclined to form fairly long-standing friendships in the medium, and the participants in ICQ chat are previous acquaintances even outside of the medium, the participants in IRC are casual acquaintances forming fleeting relationships. Messages in conversational writing, unlike asynchronous messages, are produced on the fly; they appear briefly on the screen and then scroll off. However, while split-window ICQ chatters can scroll back and edit their turns, IRC participants' turns, once posted, are uneditable. To get a foothold in the jumble of turns, IRC chatters produce very short

messages; many turns are there simply to signal the user's active presence. To detect conversational threads among the turns, participants must manage to untangle the jumble and the constant flow of server-generated messages (cf. Elser & Charniak 2008). Occasionally, conversations involve more than two participants and last for several minutes, but more often they are dyadic, short-lived and ephemeral. Example (3) from the IRC channel #20_something comprises less than a minute of communication. The example is unformatted, for illustrative purposes, and thus retains time stamps (“[22:14]”) and server-generated messages (lines marked by “***”).⁵³

- (3) [22:14] <^^katy^> wbbb crash
 [22:14] <Princess> i meant shuuuuu. i am hiding from lizard
 [22:14] <^^Crash^^> ty hun hugzz
 [22:14] *** Amike-USA has joined #20_something
 [22:14] <chanel> well you *can* write...but i have a bf...u *should* know that
 [22:14] *** Sweetpea-Soup is now known as ^Sara^
 [22:14] <iowachick> ne one from iowa or illinois?
 [22:14] *** Lara2002-117553 has quit IRC (Connection reset by peer)
 [22:14] <Chaser> chanel babe *can* i have your emailaddress
 [22:14] <^^Crash^^> ty katy babes
 [22:14] *** Pet-Ratty has left #20_something
 [22:14] *** Pablo is now known as Argentino
 [22:14] <chanel> um princess...hate to bust your bubble but chaser is lizard
 [22:14] <^^katy^> np crash..lol
 [22:14] <chanel> you had it chaser
 [22:14] <Princess> oh
 [22:14] <Farkles> princess? tisha?
 [22:14] <Argentino> holaaaaaaaaaaaa
 [22:14] *** stalesgr has quit IRC (<http://ircqnet.icq.com/>)
 [22:14] <^^Crash^^> hi chanel
 [22:14] *** canadiangirl has left #20_something
 [22:14] <chanel> hiya crash;))
 [22:14] <Princess> yes
 [22:14] *** iowachick has quit IRC (<http://ircqnet.icq.com/>)
- Internet relay chat text 1b (UCOW)

53 All subsequent examples of IRC have been purged of time stamps, join and quit messages and other server-generated text. As described in chapter 3, only user-generated conversational text was annotated and included in the linguistic feature counts.

Example (3) illustrates the typically superficial relationships among the IRC participants, a feature that becomes strikingly evident from the full corpus. Short-lived conversations take place between e.g. ^^katy^ and ^^Crash^^, Princess and Farkles, and chanel and Chaser, while iowachick and Argentino simply signal their presence/entrance. Large portions of the IRC corpus, like example (3), consist of greetings and phatic devices whereby users announce their own and others' entrance (*holaaaaaaaaaaaaa*, *hiya* and *wbbb*, i.e. welcome back, where letter *b* is repeated for emphatic endorsement). Politeness terms abound (*ty*, meaning "thank you," *np*, meaning "no problem"), as is often the case in spoken discourse. Considering the rarity of substantial discussion, the high ratio of greetings and phatic devices, and moreover, the impact of altogether verbless turns, it is rather unexpected that modal auxiliaries should find their way into the discourse at all. Judging from figure 4.1, however, modals in SCMC are almost as frequent as in spoken discourse (although no significant difference obtains between the distribution of modals in SCMC, as compared to either writing or speech; see Appendix VI). Nevertheless, it seems that, as messages become more lengthy, or rather, whenever a turn contains a full clause, i.e. a subject and a main verb, the main verb is often preceded by a modal auxiliary, as in *well you can write...but i have a bf...u should know that* (in example 3).

Example (3) contains two instances of the possibility modal *can* and one necessity modal, *should*. On Biber's (1988) dimensions of textual variation, possibility modals load on Dimension 1, as markers of involved production, whereas necessity and prediction modals load on Dimension 4, as markers of overt expression of persuasion. The frequencies illustrated in figure 4.1 indicate higher values than in writing for possibility and prediction modals in all three CMC media, but lower values for necessity modals. The media of ACMC and SSCMC surpass both writing and speech with regard to their distribution of possibility and prediction modals, whereas SCMC displays no significant difference in the distribution of modals compared to either writing or speech (Appendix VI). The division of modals into their semantic categories and their respective distributions in the five media will not be further explored in this section, however, as the network of modal meanings is too complex for the brief analysis intended here. The annotation of their semantic categories was done primarily to enable the positioning of the conversational writing genres on Biber's textual dimensions; see chapter 5, where their respective functions will be explored. Moreover, the modals were not annotated for root and epistemic meanings in all five media (as described in Coates 1983 or Coates 1995), rendering more detailed exploration

of their values impossible. Worthy of notice, nevertheless, is that in Yates' (1993) study of, inter alia, modals in ACMC (from a computer conferencing system), possibility modals were divided into their root and epistemic meanings (by analogy with Coates 1983). Yates' results show that the ACMC discourse makes more frequent use of modals than either speech or writing overall, except for the use of modals of epistemic possibility (*may, might*), which show a distribution similar to writing (higher than speech). In the corpora of SCMC and SSCMC, however, modals of epistemic possibility are found in fewer than every third IRC text, and as less than one instance per ICQ text, which makes them rarer than what Yates found for either writing, speech or ACMC, giving the conversational writing texts a more spoken, than written, character.

Among the genres amalgamated into the mean score for speech in figure 4.1 are the genres of face-to-face conversations, with 15.6 (in LLC) and 16.2 modals (in the SBC subset) per thousand words, and telephone conversations, with 18.3 modals per thousand words (Biber 1988: 264–265). In the conversation genre of the Longman Spoken and Written English Corpus (LSWE), Biber et al. (1999: 486) and Biber (2004) find approximately 22 core modal verbs per thousand words – a rate slightly higher than in any of the corpora studied here. This means, nevertheless, that the figures for ACMC and SSCMC (20.5), as shown in figure 4.1, are more in keeping with the results for core modals in conversation presented in Biber et al. (1999) and Biber (2004), approximately 22, than they are with those for conversations in Biber (1988), 15.6 and 18.3. Biber et al. (1999: 489) find both core modal and semi-modal verbs to be more common in conversation than in any of three written registers (fiction, news and academic prose).⁵⁴ The corpora of ACMC and SSCMC, in figure 4.1, show an overall frequency of core modals that is approximately 60 percent higher than in traditional writing.

Judging from the figures in Collot (1991: 74–75), the BBS conferences (represented as ACMC in figure 4.1) vary in their use of modals; the conferences “Medical,” “Finance” and “Sports” show a greater frequency of modals, with

54 In Biber et al. (1999: 486) semi-modals (*have to, (have) got to, (had) better* and *be going to*) add approximately another five modals to the count for the LSWE conversation genre, yielding a total frequency for modals of approximately 27 per 1,000 words. To relate to this figure, the IRC, ICQ and SBC subset corpora were annotated for semi-modals in a complementary study (unpublished), in which approximately three semi-modals per 1,000 words were found in IRC, four in ICQ and seven in the SBC subset. More precisely, when semi-modals are included, the total normalized count for modals in IRC is 17.0, in ICQ 24.8 and in the SBC subset 23.6.

“Medical” contributing significantly to the high overall frequency, whereas participants in the “Chit Chat” conference use modals to an intermediate degree. The interlocutors in the split-window ICQ chat corpus (SSCMC in figure 4.1) are also interpersonally involved to a high degree, seeing as they are classmates and know each other in the real-life context. As seen in the discussion of example (3), IRC chat (SCMC in figure 4.1) is highly interactive and interpersonal, even though the rate of modals in this fleeting communication is the lowest of all three CMC media. The modal auxiliary use of IRC sits between that of face-to-face conversations from LLC (Biber 1988) and the average for writing, but to the extent that verb phrases do appear in IRC, they seem to contain no fewer modals than those in ICQ. Returning to Guo’s (1995: 229) statement that language actually “developed the grammatical category of modal auxiliaries to serve the function of regulating interpersonal relations in social interaction,” it can be concluded that in all three CMC media, such regulation is going on, even though only the ACMC and SSCMC users employ modal auxiliaries to the extent of speakers in the more recent accounts of conversation (those in Biber et al. 1999 and Biber 2004).

Turning now to a survey of personal pronoun use, it will be seen that the genres of CMC differ from writing and speech in other ways, but in ways that further highlight their functions as media for social interaction.

A subheading in Chafe’s (1982) article on involvement and detachment in literature proclaims that “speakers interact with their audiences, writers do not” (1982: 45, original is in all capital letters). The subheading follows upon Chafe’s characterization of speech and writing into “fragmented” vs. “integrated” discourse and sets the tone for his further delineation of speech and writing into the qualities representing “involvement” vs. “detachment.” Speakers are typically involved with their audience, a trait manifested, *inter alia*, in speakers’ more frequent reference to themselves, i.e. through their frequent use of first person pronouns (henceforth 1PP). Writers, on the other hand, are detached from their audience and more concerned with presenting “logically coherent,” “consistent and defensible” text which “will stand the test of time” (1982: 45). In his corpora, Chafe finds a ratio of approximately thirteen 1PP in speech to one in writing, the actual numbers being 61.5 and 4.6 per thousand words respectively (1982: 46). A few pages later, Chafe admits that his categorical statements regarding speech and writing apply to extremes on the continuum; his figures are from maximally differentiated samples, spontaneous conversation vs. academic prose. (The ratio in Biber (1988) for the equivalent genres,

face-to-face conversations vs. academic prose, is roughly the same: ten 1PP to one (1988: 264, 255).) Unknown to Chafe in 1982, however, was that in the next two decades genres were to appear, with texts in which the ratio at hand is challenged or augmented further. In SSCMC, more specifically in split-window ICQ chat (see figure 4.2, based on table 4.2), for instance, the ratio of 1PP is that of nearly sixteen to one, compared to academic prose (Appendix II table 2 vs. Biber 1988: 255), or more than nineteen to one, compared to Chafe's corpus of writing (Appendix II table 2 vs. Chafe 1982: 46). Moreover, these CMC genres represent *writing*, rather than speech. Or do they? This idiosyncratic, confounding finding is one of many that suggest the definition of SCMC and SSCMC as something other than either writing or speech, hence warranting the term "conversational writing."

The first, second and third person pronouns included in Biber (1988), and therefore tagged in UCOW, are the following:⁵⁵

- First person (1PP): *I, me, my, myself, we, us, our, ourselves*
(+ contracted forms)
- Second person (2PP): *you, your, yourself, yourselves*
(+ contracted forms)
- Third person (3PP): *she, her, herself, he, him, his, himself, they, them, their, themselves*
(+ contracted forms)

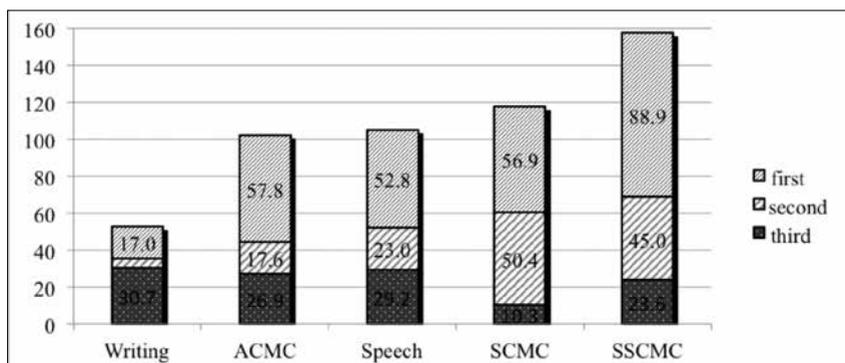
Table 4.2 and figure 4.2 present the distribution of personal pronouns in the media investigated.

55 Biber (1988: 225) subsumes personal, possessive and reflexive pronouns under the heading "personal pronouns." None of the forms for "it" are included, nor are the independent possessive pronouns (*mine, yours*, etc.) (cf. Quirk et al. 1985: 361).

Table 4.2: Frequencies of first, second and third person pronouns per 1,000 words (normalized values)

	Writing	ACMC	Speech	SCMC	SSCMC
first person pronouns	17.0	57.8	52.8	56.9	88.9
second person pronouns	5.0	17.6	23.0	50.4	45.0
third person pronouns	30.7	26.9	29.2	10.3	23.6
total	52.7	102.3	105.0	117.6	157.5

Figure 4.2: Distribution of first, second and third person pronouns per 1,000 words (normalized values).



First and second person pronouns are two of the features that, in either or both of the conversational writing genres, deviate from Biber's mean of writing and speech (Appendix II table 4) by more than two standard deviations ($|s.d.| > 2.0$). They are taken up in this section chiefly because, together with modal auxiliaries, they constitute important carriers of interpersonal meaning in language (Halliday 1985a, 2004, Yates 1993). All in all, ten features deviate thus; section 4.4, below, will explore the other eight: WH-questions, analytic negation, demonstrative and indefinite pronouns, present tense verbs, predicative adjectives, contractions and prepositional phrases. By their sheer relative frequency (or infrequency in the case of prepositional phrases), these features can be said to epitomize the linguistic character of conversational writing. As the word frequency lists (Appendix VII) show the first person singular pronoun (*I*) to be in a distinguished first position (i.e. as the most frequent lexeme) in all three corpora annotated in the present study, and the second person pronoun (*you*) among the top three in all, it seems befitting that our exploration begins with these.

As mentioned, Chafe (1982) and Chafe & Danielewicz (1987) claim that speakers' involvement with their audience is manifested in speakers' frequent reference to themselves. Writers, who rarely see their audience, typically use fewer first and second person pronouns. Chafe (1982) and Chafe & Danielewicz (1987) find that the relationships between speakers/listeners and writer/readers are encoded in language by the varying levels of involvement and detachment in speech and writing. Chafe (1982: 45) argues that the involvement typical of speech arises from the fact that:

It is typically the case that a speaker has face to face contact with the person with whom he or she is speaking. That means, for one thing, that the speaker and listener share a considerable amount of knowledge concerning the environment of the conversation. It also means that the speaker can monitor the effect of what he or she is saying on the listener, and that the listener is able to signal understanding and to ask for clarification. It means furthermore that the speaker is aware of an obligation to communicate what he or she has in mind in a way that reflects the richness of his or her thoughts [...] with the complex details of real experiences [...]. (Chafe 1982: 45)

Chafe (1982: 45) goes on to contrast the experiential involvement typical of speech with the typically detached nature of written discourse:

The situation of the writer is fundamentally different. His or her readers are displaced in time and space, and he or she may not even know in any specific terms who the audience will be. The result is that the writer is less concerned with experiential richness, and more concerned with producing something that will be consistent and defensible when read by different people at different times in different places, something that will stand the test of time. (Chafe 1982: 45)

Fowler & Kress (1979) also find that first person pronouns are rare in writing but regard this as an effect of "appropriate" attendant social practices rather than an effect of the medium (see also Yates 1993: 109). In other words, the "impersonal, generalizing tone of newspapers, textbooks, scientific articles" (Fowler & Kress 1979: 201) calls for a "[r]emoval of the pronoun associated with personal speech" (ibid.). Fowler & Kress, however, take note of varying subjectivity in different genres, observing slightly higher frequencies of first person pronouns in e.g. self-centered articles and eye-witness accounts (1979: 201) than in other writing. In a like manner, Chafe & Danielewicz (1987: 107) investigate first person pronoun use in four genres: conversations, lectures, academic papers and informal letters, finding informal letters to contain the highest number (57 per thousand words) – despite the written mode. Their finding underscores the purported significance of an identified audience, present *or* remote, for the formation of involved discourse, and leads Chafe & Danielewicz to conclude, like Fowler & Kress (1979), that other factors than the medium itself may be at play:

The use of first person pronouns is thus not necessarily a feature which differentiates spoken from written language, but rather a feature which the absence of a direct audience may even foster when the circumstances are right. (Chafe & Danielewicz 1987: 107)

Two circumstances that “foster” involved communication are thus 1) an identifiable, attentive and responsive audience (present or remote) and 2) a medium in which social and cultural practices permit the discussion of self. A third circumstance, as will be seen, is the synchronicity factor. That synchronicity is a predictor of high first and second person pronoun incidence is clearly illustrated in figure 4.2.

Each of the synchronous and supersynchronous genres of conversational writing, in figure 4.2, displays a combined usage of first and second person pronouns (1PP and 2PP) that surpasses either of the asynchronous media, as well as speech. In IRC, 1PP are about as common as in APMC, but 2PP are more frequent than in any of the other media. Furthermore, interlocutors in IRC are the least concerned with third person reference, as seen in the extremely low frequency of third person pronouns (3PP). Example (3) above, from IRC, characteristically contains four 1PP (*i*) and seven 2PP (*ty, you, u, your*), but no instance of a 3PP. “[T]he more first and second person pronouns chatters use, the more involved they are with their fellow interlocutors,” says Freiermuth (2003: 74) in his account of public America Online political chat channel data. In his data, 1PP and 2PP make up about 77 percent of the total personal pronoun use (2003: 128).⁵⁶ In the IRC data in figure 4.2, from the channels #20_something, #30_something, #Chat-World, #Family and #USA, the sum of 1PP and 2PP constitutes more than 90 percent of the personal pronouns (see further figure 4.3). This result seems to indicate an extremely high degree of involvement on behalf of the interlocutors in IRC. The communication indeed tends to center around the self and the second person, as in example (4), but rather than being used to express subjective opinions (as in political chat), the first and second person pronouns in IRC are mostly used for addressing others upon entering and leaving the channel (e.g. *i will be back, c ya, c u*), or for polite speech-act formulae (e.g. *ty*, meaning “thank you,” *yvw*, meaning “you’re very welcome”).

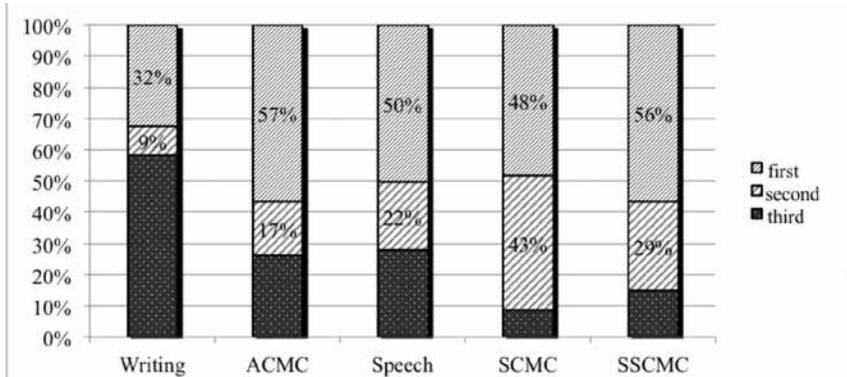
- (4) <Cheeky1> *i will be back*
 <|mad_max|> *ok*
 <|mad_max|> *take care*
 <Cheeky1> *gotta go for 5 minutes*

56 Unlike Biber (1988), Freiermuth (2003) includes the independent possessive pronouns (*mine, yours*, etc.), but not the reflexive pronouns (Freiermuth 2003: 127).

<Cheeky1> u 2 max sweety
 <Cheeky1> c ya in a sec u hunk of spunk
 <Cheeky1> hehehe
 <|mad_max|> c u
 <Cheeky1> cya

Internet relay chat text 3a (UCOW)

Figure 4.3: Proportions for first, second and third person pronouns of total personal pronoun use.



Returning to Chafe's (1982) description of involved discourse, quoted above, the IRC communicators, despite the written medium, "share a considerable amount of knowledge concerning the environment of the conversation" (1982: 45). Fellow IRC participants are identifiable, attentive and responsive, and participants know that the medium is intended for social interaction. The communication in the shared window is immediate and responses appear in seconds, as in face-to-face and telephone conversation. To the extent that responding turns appear, the interlocutor in IRC, as in spoken conversations, "can monitor the effect of what he or she is saying on the listener" (Chafe 1982: 45) and "the listener is able to signal understanding and to ask for clarification" (ibid.). Chafe's description of involved spoken discourse therefore for the most part holds true for IRC. Chafe's (1982) description of writing, however, is not applicable to conversational writing. IRC chatters are not "displaced in time and space" (1982: 45), as their communication is synchronous and appears in a shared virtual context. Through the list of logged-on participants, chatters have a notion of their audience, and their discourse therefore, like speech, appears to be more concerned with experiential richness than with the objective sharing of information.

In the supersynchronous medium of split-window ICQ chat, SSCMC in figures 4.2 and 4.3, interlocutors, to an even greater degree than other chatters, appear to be concerned with expressing subjectivity and personal opinion. First person pronouns abound in the corpus, as in (5) which contains nine 1PP (*i, me*) and three 2PP (*you, u*), but no 3PP.

- (5) <J> how come *you* didnt take bio II?
 <10> last year *i* started to like it after a bio class and *i* enjoyed it a lot
 <10> *i* did *i* had a good teacher
 <10> r *u* taking bio 2 lol
 <J> no.. to tell the truth.. *i* hate bio.. to *me*.. its all like studying things and not much creativity like calculus or physics.. were *you* really have to think to solve problems.. *i* guess *i* 0just like math in general
 <10> *i* just hate doing those long problems

Split-window ICQ chat text 9 (UCOW)

The chatters in the ICQ data, high school classmates, are slightly less concerned than the IRC chatters with the second person (judging from second person pronoun use). In the ICQ chats, second person pronouns are used less in greetings and politeness terms (*see you, thank you, you're welcome*) than they are in IRC, but more as parts of committed questions; see example (5). Used thus, the second person pronouns in ICQ, unlike those in IRC, reveal interlocutors' real-life acquaintance and the genuinely involved character of their communication (*how come you didnt take bio II?*, *r u taking bio 2*). Split-window ICQ chatters, furthermore, use more third person pronouns than do IRC chatters – another result of ICQ chatters' acquaintance outside of the medium and their exposure to the same human referents. (On the other hand, the human referents shared in IRC, i.e. the fellow chat participants, are mostly referred to by their nicknames and not by third person pronouns, to avoid deictic confusion.) In conclusion, split-window ICQ chat, like IRC, is more in harmony with the involved discourse typical of speech as defined by Chafe (1982) than with his definition of writing. Neither chatter has “face to face contact with the person with whom he or she is speaking” but the chatters “share a considerable amount of knowledge concerning the environment of the conversation”; they “can monitor the effect of what [they are] saying on the listener”, and “the listener is able to signal understanding and to ask for clarification” (1982: 45 for all four quotes). In the words of Chafe (1982) this means for split-window ICQ chatters, furthermore, that, like speakers, they are “aware of an obligation to communicate what [they have] in mind in a way that reflects the richness of [their] thoughts [...] with the complex details of real experiences” (Chafe 1982: 45).

Above, we identified three circumstances that foster involved communication: 1) an identifiable, attentive and responsive audience (present or remote),

2) a medium in which social and cultural practices permit the discussion of self, and 3) the synchronicity factor, which enables dialogic communication. Seeing the effect that previous acquaintance has upon ICQ chatters' discourse, a fourth circumstance might be added: 4) close personal acquaintance. Certainly, more factors could be added, but for the present purpose this collection will do. In combination, these factors all contribute to proximity and directness between interlocutors and increase the personal reference among them. In the SCMC corpus, the first three factors are at work and in the SSCMC corpus all four.

What about the asynchronous modes of CMC, then? Judging from figure 4.2, ACMC implements pronominal reference to about the same degree as speech, both as regards the combined use of first and second person pronouns, and as regards overall use. Judging from figures 4.2 and 4.3, ACMC users employ first person pronouns more than speakers and second person pronouns slightly less than speakers. Collot (1991), whose counts underlie the ACMC bars in figures 4.2 and 4.3, notes that first and second person pronouns in her corpus, among other features, "indicate a highly verbal, and personally involved style" (1991: 80) but does not delve further into their use (the ACMC example (1) above contains six 1PP and three 2PP). Yates' (1993) study of asynchronous computer conferencing texts, however, discusses pronominal reference at length, much of which inspired the above account for the synchronous chats. Yates finds first and second person pronouns to constitute 64 percent of all personal pronouns in his ACMC corpus.⁵⁷ In Collot's ACMC corpus "ELC other," the same proportion is 74 percent, slightly more than in Biber's (1988) genres of speech. In Biber's (1988) genres of writing, however, 1PP and 2PP together constitute only 41 percent of all personal pronouns (see figure 4.3). ACMC, despite being a written, asynchronous medium, therefore clearly deviates from the other genres of writing; for another thing, the overall use of personal pronouns in ACMC is nearly twice the number of traditional writing, as measured in normalized frequencies (see table 4.2).

What is it about the ACMC medium that makes for personal, involved communication of this kind? To answer this question, we must look at the written genre that most closely resembles the ACMC genre here: personal letters. Both ACMC and personal letters are produced under at least two of the circumstances that foster involved communication: they are directed at a presumably responding, albeit remote, audience and their attendant social practice is of an interactional

57 Yates (1993, 1996) does not specify whether possessive and/or reflexive pronouns are included in the count of personal pronouns, or exactly which personal pronouns are counted. Collot (1991), however, follows Biber's (1988) feature annotation scheme, which makes her figures ideally suited for comparison with the other media.

kind that permits, expects or condones the discussion/presentation of self. Personal letters are, moreover, exchanged between previous acquaintances. Casual ACMC messages, as seen in example (1) above, assume a similar personal tone as private letters, especially messages exchanged among ACMC users who consciously seek lasting friendships through the BBS. As mentioned before, Chafe & Danielewicz (1987) find first person pronouns to be more slightly more common in informal letters than in conversation. In seeking and maintaining friendship through asynchronous written media such as letters and ACMC, the presentation of self is evidently central, expected and culturally sustained. Among the four factors identified as triggers of involved communication, the synchronicity factor is the only one not at play in the asynchronous discourse.

A few references with regard to traditional asynchronous communication – viz. letters – are in place here. Besnier (1988, 1991, 1995) notes for Nukulaelae Tuvaluan letters (see also Biber 1995, Yates 1993) that they “include phatic communion” and are “heavily affective” toward the addressee (Besnier 1988: 714). He therefore criticizes linguists who regard writing as a medium in which emotional content and self-expressions are minimized (Besnier 1988). Biber’s (1995: 175) multidimensional analysis of Besnier’s letters positions them beyond Nukulaelae Tuvaluan conversations as regards interpersonal reference, noting that they “make frequent reference to the author (‘I’) and receiver (‘you’), even though [the] direct interaction through [the] letters is extended over long periods of time” (Biber 1995: 174). Biber’s (1988, 1995) own collection of personal letters in English does not assume a position beyond English face-to-face or telephone conversations as regards involved production (on his Dimension 1 distinguishing between informational and involved production), but a position second only to conversations, beyond all other genres of writing and speech (1988: 128; see also section 5.2.1 here). First and second person pronouns constitute 61 percent of the total personal pronoun use in Biber’s personal letters (1988: 262). The letters contain 62.0 1PP and 20.2 2PP per thousand words, respectively. In Collot’s ACMC “ELC other” corpus there are 57.8 1PP and 17.6 2PP per thousand words, respectively; see figure 4.2. There is thus a close affinity between personal letters and ACMC, as well as between ACMC and the average for speech, considering their nearly identical overall personal pronoun use; see figure 4.2: 102.3 and 105.0 pronouns per thousand words in respective medium. The spoken genres most akin to ACMC, as regards personal pronominal reference, are spontaneous speeches and interviews (Biber 1988: 268, 266, also noted by Collot 1991 with regard to personal reference as well as to other features).

Turning now to the intermediate bar in figures 4.2 and 4.3 – speech – a few remarks are called for. Numerous linguistic authorities, such as Chafe (1982),

Chafe & Danielewicz (1987), Wales (1996), Biber (1988, 1995) and Biber et al. (1999), have drawn attention to the overall high numbers of personal pronouns observed in speech, as opposed to their numbers in writing (with the exception of personal/informal letters). Explaining message structure in English, Halliday (1985a) declares, in functional grammatical terms, that the Theme in spoken language, “the peg on which the message is to hang,” is often a pronoun, “most typically *I* or *you*” (1985a: 73, original italics). Halliday (2004) expounds:

In everyday conversation the item most often functioning as unmarked Theme (Subject/Theme) in a declarative clause is the first person pronoun *I*. Much of our talk consists of messages concerned with ourselves, and especially with what we think and feel. Next after that come the other personal pronouns *you*, *we*, *he*, *she*, *it*, *they*; and the impersonal pronouns *it* and *there*. (Halliday 2004: 73, original italics)

Wales (1996) notes that the first personal singular pronoun (*I*) occurs most frequently in speech, and that it is the second most common word in the spoken part of the British National Corpus, second only to *the* (1996: 68). Among the spoken genres, Biber (1988) finds personal pronouns most common in telephone conversations (totaling 126.7 per thousand words; 1988: 265), closely followed by face-to-face conversations (totaling 117.9 per thousand words; 1988: 264). Biber et al. (1999) explain that first and second person pronouns, referring to the speaker and the addressee, are “naturally very common in conversation because both participants are in immediate contact, and the interaction typically focuses on matters of immediate concern” (1999: 333). None of the linguists mentioned, however, has investigated conversational writing, such as IRC and split-window ICQ chat. That personal pronouns are “by far most common in conversation” (c. 135 per thousand words in the LSWE corpus; Biber et al. 1999: 333) is a statement that can be qualified. Not only does Biber (1988) find them equally common in personal letters (135.0 per thousand words; 1988: 262), but the investigation of personal pronouns, in the present section, has proved that they are even more common in supersynchronous conversational writing (157.5 per thousand words). Moreover, it is in the conversational writing genres that the ratios of first and second person pronouns to all personal pronouns are the highest. Chafe’s initially striking finding of 61.5 first person pronouns per thousand words in spoken discourse, with which he introduces the concept of “involvement” (Chafe 1982: 46), pales by comparison with the finding of 88.9 first person pronouns in split-window ICQ chat in the present study; see figure 4.2. “Involvement,” instantiated through first person reference, thus epitomizes the character of supersynchronous conversational writing more than it does the character of any other genre.

4.3 Word length, type/token ratio and lexical density

In Biber's (1988) multidimensional study of written and spoken language, "word length" and "type/token ratio" (the ratio between the number of different words, "types," and the total number of words, "tokens,"⁵⁸ per text) are the two features intended to measure the lexical specificity and diversity of texts. They are powerful tools in the study, as differences in lexical specificity and diversity truly are found to correlate with production differences between writing and speaking. Longer words have been found to convey "more specific, specialized meanings than shorter ones" and words tend to "become shorter as they are more frequently used and more general in meaning" (Biber 1988: 238, referring to Zipf 1949). Zipf (1949: 65) finds an "inverse relationship between the lengths of words and their frequency" in language, not just in English, but in several other languages (including Peipingese Chinese, two American Indian languages and the main Western European languages), i.e. that the short words in these languages tend to recur. Zipf (1949), Drieman (1962), DeVito (1965) and Gibson et al. (1966) all consider measures of word length in their studies (as seen in chapter 2), finding longer words more frequent in writing than in speech. The latter three studies, furthermore, employ the measurement of type/token ratio, henceforth TTR, in distinguishing between written and spoken texts, finding higher TTR values in writing. In the present section, the lexical properties of conversational writing are explored, facilitated by the measurements of word length and TTR, as well as by the more revealing measurements for our purposes, those of lexical density.

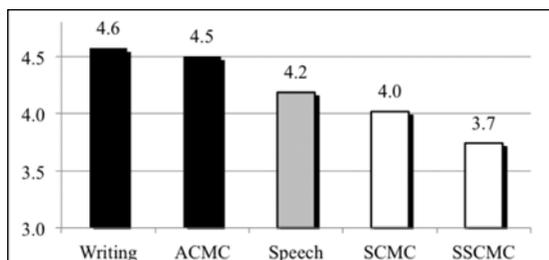
Drieman (1962) and Gibson et al. (1966), by measuring numbers of syllables, and Blankenship (1974), by measuring word length per se (most likely by characters), all find word length to be a distinguishing factor between writing and speech, observing shorter words in speech. The difference in word length is attributed to the different production circumstances of writing and speech, less encoding time in speech and the consequent need for the speaker to select "easy, short, and familiar" words (DeVito 1970: 11). Longer words usually entail higher levels of lexical specificity, and are typically produced under circumstances that permit editing and longer contemplation. When writing, "[w]e can take hours,

58 "Token" by default denotes a string of orthographic keystrokes set apart from other strings by a blank space or a new line. Tokens in conversational writing mostly constitute words semantically, but also for instance initialisms (e.g. *hb* meaning "hurry back"), and emoticons (e.g. :)), not traditionally referred to as words. (Accordingly, even in other sections of the present study, the term "token" is occasionally preferred over "word" when discussing the data.)

if we need to, to find an appropriate word,” say Chafe & Danielewicz (1987: 88), and “...we are free to revise [the words] again and again until they satisfy us” (ibid.). Halliday (1985a) prudently establishes that the distinction between long and short words in reality reflects the continuum from lexis into grammar. The distinction is simply embodied in the spelling system: lexical items typically require a minimum of three letters, whereas grammatical items may comprise only one or two letters. Halliday incidentally points out that most prepositions belong in the grammatical class, “because of words like *at, in, to, on*, which otherwise would have to be spelt *att, inn, too, onn*” (1985a: 63, original italics). The distinction between long (mostly lexical) and short (mostly grammatical) words can thus be seen as fundamental to the difference between writing and speech.

Figure 4.4 indicates the average word length of texts in the five media: writing, asynchronous CMC, speech, synchronous and supersynchronous CMC, respectively. The figure shows a neatly declining scale of word length, from 4.6 orthographic letters per word in writing to 3.7 in split-window ICQ chats. In figure 4.4, as well as in all subsequent diagrams, the written genres are represented in black, spoken genres in gray and conversational writing genres in white. For the p-values from statistical tests of findings in SCMC and SSCMC, as compared to writing and speech, see Appendix VI.

Figure 4.4: Average word length in the five media.



Word length entails “the mean length of the words in a text, in orthographic letters” (Biber 1988: 239). In conversational writing, this is indicated as the number of orthographic keystrokes found between blanks, after texts were purged of all regular punctuation, except apostrophes within words, emoticons and simple imagery.⁵⁹ Example (6) is a part of a text ready for the word-length count, a text

59 See chapter 3 for a description of the purging and adaptation procedure, and section 4.5 for examples of retained imagery.

which exemplifies a few remaining, albeit rare, instances of such imagery (:0), i.e. a smiley, and <===== (=0, a sword). The example illustrates the irregular length of tokens typically found in chats (repeated xxxxxx... etc., meaning “kisses” vs. *c u*, meaning “see you”). To avoid skewing the word length results, extremely long tokens were truncated at 50 keystrokes; five such long tokens existed in the IRC component, and two in the split-window ICQ component.

(6) i dont know who he really is
 yeah women!
 lol
 true
 be careful
 that i am
 hi all
 <===== (=0
 <===== (=0
 any girl wanna chat?
 <===== (=0
 nice sword
 lol
 u have been practising a lot
 he has
 now he is ready
 saba 20 where are you
 alot of work put into that piece of artwork
 to impress the ladies
 lol
 i will be back
 ok
 take care
 gotta go for 5 minutes
 u 2 max sweety
 c ya in a sec u hunk of spunk
 hehehe
 c u
 cya
 xxx
 :0)
 hi

Internet relay chat text 3a (UCOW)

For maximum economy of typing, either for minimum effort or minimum production time, or both, or for mere adherence to genre conventions, IRC

interlocutors abbreviate and contract words and expressions in various ways (e.g. *lol, u, wanna, alot, u 2, c ya, sec, c u, cya* in example (6)). Such abbreviation schemes naturally render short words pervasive. On the other hand, resting a finger on a key for an entire turn, as in *xxxxxx...* etc. in (6), and posting precomposed imagery, such as the sword in (6), are also devices available to chatters, devices which increase the average word length. Nevertheless, from figure 4.4 it is evident that IRC chatters on average operate with shorter words than do speakers. For one thing, speakers cannot abbreviate words, e.g. *see, you*, into their corresponding homophonous letters, *c, u* (or rather, transcribers of speech do not).

Comparing the average *turn* length of IRC (4.3 tokens/turn) with split-window ICQ (7.0 tokens/turn), in connection with the average word length displayed for these genres in figure 4.4, gives the impression that, in conversational writing, longer turns entail shorter words. As seen, ICQ chatters indeed employ shorter words than IRC chatters; the average word length in split-window ICQ is only 3.7 orthographic keystrokes. On the other hand, as seen in example (6), a great number of IRC turns consist of very short messages, e.g. greetings (*hi, cya*), with very short word length. Also, “turn length” in split-window ICQ is a rather artificial concept as it is determined by the logging feature of the software, more than by the actual user. The cut-off point between turns is not always clear-cut in the supersynchronous chats, where simultaneous typing frequently occurs and where users do not hit enter to post their turn. For this reason, turn length is not a reliable construct for comparisons of word length. The results in figure 4.4, nevertheless, underscore that split-window ICQ chatters operate with shorter words than IRC chatters. Example (7), from the split-window ICQ word length count, shows that the ICQ chatters, for the economy of typing, use similar abbreviations as the users in example (6) (*r, u*), as well as apostrophe-less contractions (*thats, im, its, wasnt, didnt*), although the abbreviations are less frequent in split-window ICQ than in the IRC chats.

- (7) what r u doing this weekend
 i'm going to be sitting at home watching love movies by myself :(
 awwwwwwww
 thats cute
 im not sure what im doing but it will probably be just as boring
 yea its going to suck
 well at least you're allowed to go out and stuff
 ya thats true
 no offence u didn't have to get a speeding ticket
 yea thanx

well yea but i wasnt even doing 75 but i just said i did because i didnt want to
fight the case
why u could have won
yea but that means iw ould have to miss a couple days of school just to go to court
Split-window ICQ chat text 1 (UCOW)

The subject matter discussed, of course, could be a factor influencing word length. The topics in the split-window ICQ chats are more tangible and diverse, and the discussions more vivid, than in IRC. On the other hand, both ICQ and IRC chatting are leisure-time activities for casual social interaction, and neither communication requires well-reasoned exposition or highly explicit lexical choices from users. The short word length of conversational writing therefore, more than anything, seems to be determined by the same factor that renders short words in speaking, briefly considered in the beginning of this section: their stronger affiliation with the grammatical rather than lexical classes of words. The lexical density of conversational writing will be further investigated below, but first we must briefly touch upon the classic measurement of TTR.

The type/token-ratio measure is regarded as a useful tool for exploring the vocabulary variety of a given text. To arrive at the TTR, the number of different words (“types”) in a text is divided by the number of words (“tokens”) in that text. Consider the split-window ICQ example below, from example (7), which serves to illustrate the procedure:

“well yea **but i** wasnt even doing 75 **but i** just said i did because i didnt want to fight the case”

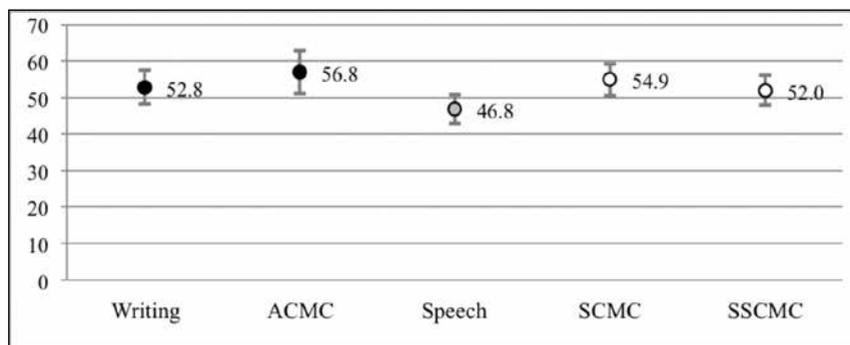
The example contains 22 words (tokens), but there are only 18 different words, as *i* is used four times and *but* twice. The type/token ratio of this sentence is consequently $18/22$, i.e. 0.818, or by convention, expressed as a percentage, 81.8. In order for the TTR to reliably represent the diversity of a text, however, samples must be of substantial length, though not too long as the relation of types to tokens is not linear. Biber (1988) finds the ideal sample size for measuring TTR to be 400 words. In Biber (1988), the ratio is computed “by counting the number of different lexical items that occur in the first 400 words of a text, and then dividing by four” (1988: 238, as explained in section 3.2). All texts in the five media to be compared here have undergone this computation method for TTR, and the results are shown in table 4.3 and figure 4.5, along with the standard deviations among texts.⁶⁰

60 See section 3.6 for an explanation of the procedure for calculating the TTR standard deviation of the texts in the media writing and speech.

Table 4.3: Type/token ratio, with standard deviation

	Writing	ACMC	Speech	SCMC	SSCMC
type/token ratio	52.8	56.8	46.8	54.9	52.0
standard deviation	4.7	5.9	3.9	4.4	4.1

Figure 4.5: Type/token ratio, with standard deviation.



As mentioned in section 2.2 and in the beginning of this section, linguists have consistently found higher TTRs in writing than in speech (Drieman 1962, Gibson et al. 1966, Blankenship 1974, Chafe & Danielewicz 1987, Biber 1988, 1999). Chafe & Danielewicz (1987) explain:

[S]peakers tend to operate with a narrower range of lexical choices than writers. Producing language on the fly, they hardly have time to sift through all the possible choices they might make, and may typically settle on the first words that occur to them. The result is that the vocabulary of spoken language is more limited in variety. (Chafe & Danielewicz 1987: 88)

The linguists mentioned above have all dealt with texts that have been ideally suited to represent their respective media in written format. The written texts mostly derive from published sources and have thereby undergone careful editorial scrutiny, which has rendered misspellings and other irregularities extremely rare in them. The spoken texts for the most part have been transcribed by linguists, who have devoted considerable time and effort to correctly representing speech, with detailed attention to spelling, regularity and consistency. Writing and speech are consequently reliably represented as regards vocabulary variety in figure 4.5.

The texts of CMC are of a different kind. None of them has undergone careful editorial scrutiny or been transcribed by linguists. Instead, they are taken

straight from their respective media and represent authentic user-generated material. APMC texts may well contain carefully prepared exposition. Yates (1993) notes for his corpus that the APMC medium does provide the opportunities for redrafting that according to Chafe & Danielewicz (1987) bring about a greater vocabulary in written texts, but notes that these opportunities may not be taken by all CMC users. Yates observes a TTR for APMC which, like Collof's (1991) APMC plot in figure 4.5, is closer to writing than to speech. With regard to the synchronously and supersynchronously mediated texts, however, the TTR representation in figure 4.5 is more problematic. To simplify, here is a speculative, but viable, analogy: if the conversational writing texts were writing, they would be a very first draft, produced under severe time constraints, with no chance for editorial or self-revision; if they were speech, and therefore transcribed by linguists, they would be transformed into a format as regular and consistent as spoken texts, and likely attain a TTR similar to speech, or even face-to-face and telephone conversations.⁶¹ The TTR of the SBC subset (face-to-face conversations), for instance, is 44.2.

Speculation aside, why are the TTRs for conversational writing so high? Texts with high TTR display a great number of types, i.e. different words. The heterogeneity of words in conversational writing is immediately noticeable upon studying the texts. Firstly, besides abbreviations, emoticons and certain imagery, the texts bristle with other irregularities: misspellings (*sence* for *sense*), slips of keys (*wel ive* for “we live”), missed keystrokes (*jus* for “just”), contractions with omitted apostrophes (*dont, im, thats, shes*), letters repeated for effect (*desire for fooooooooooooooooood*), graphic re-representation of letters (*\Velcome*), simplified, phonological spelling (*proolly* for “probably,” *sleepin, cuz* for “because,” *kinda* for “kind of”) and multitudes of renderings of one and the same lexeme (*u, ya, yu, you, yah, yôÛ, jÛÛ, Yôù*, to mention but a few representations of the second person pronoun). Secondly, chatters in IRC, unlike speakers in most spoken conversations, repeatedly address each other by nicknames to designate the recipient of an utterance. Greetings, for instance, are frequently followed by nicknames, which serve to designate the recipient as well as to signal that the new user's presence has been noticed (Werry 1996, Anglemark 2009). Nicknames serving as address terms also facilitate the untangling of threads in the communication.

61 Despite the feasibility of such a task, it goes against the grain for the present researcher to transcribe or manipulate this kind of unique user-generated conversational writing data to attain more comparable figures. Worthy of notice, however, is Ko's (1996) study, which arrives at a TTR of only 33.7 for classroom setting SCMC with seemingly regular and consistent user-generated orthography (rating from examples given).

“Such a high degree of addressivity is imperative on IRC, since the addressee’s attention must be recaptured anew with each utterance,” says Werry (1996: 52). The designatory nicknames add to the number of types in a text, especially since they are frequently changed and as users tend to invent their own pet names out of them (e.g. } melons{{ is addressed *melons*, *mels*, Rich23 *rich*, *rick*, [mad_max|mad_max, mad_max, etc., each variant counted as a separate type in the TTR calculation). Thirdly, chatters frequently emulate spoken communication to add emphasis to utterances. In a spoken language corpus, an “utterance” like laughter follows the same regularized transcription convention throughout the corpus, whereas in conversational writing users invent their own “transcriptions” ad hoc (HAHAHAHAHAHAh, ha ha, haha, hahahahahahaha) – in which, as regards TTR, a thirteen-character laughter token counts as a different “type” than a sixteen-character one. This equally applies to chatters’ alternative transcriptions of word stress (*yeesssssss*, *this suxxxxxxxxxxxxxxx*, *yummm....*). Moreover, several of the emoticons may contain repetition of a character for emphasis (whereby :) counts as one “type,” :) as another, etc.). All in all, this user-generated orthographic heterogeneity results in a multitude of types in the type/token-calculation, rendering inordinately high ratios for the conversational writing genres (also noted by e.g. Freiermuth 2003, Forsyth 2007, Forsyth & Martell 2007, who similarly discover type/token ratios in SCMC that are closer to writing than to speech, and who explain this inter alia by the variable spelling and nickname usage).⁶² The representation of conversational writing in figure 4.5 must, therefore, be taken with a grain of salt, and we must find alternative ways to approach and explore the lexical complexity, or lack of complexity, in conversational writing.

This is when we turn to the two more revealing methods for measuring the lexical properties of conversational writing: the measures of *lexical density* and *lexical density per clause* (cf. Ure 1971, Halliday 1985a, 1987, 2004). Unlike TTR, these measures distinguish between lexically complex (“most likely to be written”) and grammatically complex (“most likely to be spoken”) texts (Halliday 1987: 59). While the TTR measure, rather mechanically, indicates the ratio of new types among the tokens, the lexical density measures take into account the lexical properties of the words. Moreover, the lexical density measures are not

62 By contrast, neither Collot (1991) nor Yates (1993) problematizes the high TTR of their ACMC texts as being the result of irregular spelling or other orthographic anomalies. Judging from corpus examples in both computer conferencing studies, participants’ spelling is consistent and appears to follow the norms of writing. The TTR for Collot’s (1991) ACMC corpus in figure 4.5 thus justifiably indicates a vocabulary variety in asynchronous computer-mediated texts above that of writing.

sensitive to text length (Yates 1993). In the discussion of word length above, Halliday (1985a) was shown to have drawn attention to the distinction between lexical and grammatical items in discourse. The short average word length in conversational writing was suspected to be due to there being more grammatical than lexical words in the discourse. It is now time to find out whether this is the case.

The lexical density of a text is the proportion of lexical items (content words) to the total discourse (Halliday 1985a, 1987). It can be measured in at least two ways: the ratio of lexical items to the total number of running words in a text, or to the total number of clauses, with or without weighting for relative frequency in the language.⁶³ In our consideration of the lexical density of conversational writing, no weighting will be employed. To understand the measurement, consider Halliday's (1985a: 61) classic example, which contrasts a written sentence with its "translation" into a likely spoken equivalent:

Investment in a rail facility implies a long term commitment (L:7; G:3)

If you invest in a rail facility, this implies that you are going to be committed for a long term (L:7; G:13)

The first of these sentences (more typical of writing) contains a ratio of seven lexical items (L:7) to three grammatical (G:3), the lexical items being *Investment*, *rail*, *facility*, *implies*, *long*, *term* and *commitment*. A ratio of seven lexical items to a total of ten words yields a lexical density of 7/10, i.e. a lexical density of 70%. The second sentence (more typical of speech) contains more grammatical items and therefore yields a lexical density of 7/20, i.e. 35%. Relative to each other, written language is lexically dense and spoken language is lexically sparse, or put differently: spoken language is grammatically dense; it displays "grammatical intricacy" (Halliday 1985a: 87, 1987: 62ff, 2004: 655).

To calculate the lexical density of a text, all orthographic items (tokens) must first be identified as either belonging to the closed sets of grammatical items, or to the open-ended classes of lexical items – a fairly cumbersome but, as we shall see, worthwhile task. Halliday (1985a: 61) identifies the grammatical items in English to be "determiners, pronouns, most prepositions, conjunctions, some classes of adverb, and finite verbs." He goes on to give a number of example sentences indicating finite full verbs, such as the third person present tense verb *implies* in the above example, as lexical items. In light of the examples, his

63 In weighted lexical density calculations, low-frequency lexical items are given a higher "score" (or "weight") than high-frequency ones. In unweighted lexical density calculations, all items are treated alike (Halliday 1985a, Yates 1993).

definition of “finite verbs” as being grammatical items must therefore be re-interpreted as “auxiliary verbs.” Furthermore, his example sentences indicate all forms of the verbs *be*, *have* and *do* as grammatical items. In the present study, lexical items were consequently taken to be all non-auxiliary, i.e. full verbs (except *be*, *have*, *do*), as well as nouns (including nominalizations, nominal gerunds and proper nouns),⁶⁴ adverbs (except discourse particles, adverbs *all*, *as*, *here*, *how*, *then*, *there*, *when*, *where*, *why*, *anywhere*, *everywhere*, *nowhere*, *somewhere*, *so*, synthetic negation *no*, *neither*, *nor*, analytic negation *not*) and adjectives, in agreement with the examples given in Halliday (1985a: 61–62).⁶⁵ This means that lexical items were found among Biber’s (1988) features (full verbs among e.g. features 1–3, 17, 18, 24–26, 55–58; nouns among features 14–16; adverbs among e.g. features 4, 5, 42, 45–49, and adjectives among features 40–41; see table 2.1 for the numbered features), but also had to be found outside of this list of features, as for instance the main verbs of progressive verb phrases are not identified by it. The identification of lexical items therefore required a separate round of annotation, beyond the annotation of Biber’s features.

With the identification of the lexical items completed, the calculation of lexical density for each corpus was fairly straightforward. As mentioned, the lexical density measurement simply indicates the ratio (percentage) of lexical items to the total number of running words. The results are presented in table 4.4, along with the lexical densities calculated by Yates (1993) for LOB, to represent writing, and for LLC, to represent speech. No lexical density was calculated for ACMC in Collof (1991); therefore, to represent ACMC in table 4.4 is the figure for Yates’ (1993) computer conferencing corpus. The results are not presented graphically here since previous graphs, and graphs to come, indicate figures for Collof’s (1991) ACMC corpus and incorporating Yates’ (1993) ACMC figure for only this feature would interrupt the consistency across graphs.⁶⁶

64 Nicknames used as address terms (very common in IRC) were not included in the count for proper nouns, to avoid skewing the data, but nicknames used about a third person were included, as well as all other proper nouns.

65 Numerals, infinitive markers, inserts (except e.g. *Shit*, *God*; see table 4.9) and emotives were considered to be “grammatical” words.

66 Note that Yates’ (1993) lexical density figure for ACMC is only indicative here, as it might be that Yates’ computer conferencing corpus deviates lexico-grammatically from Collof’s (1991) corpus of BBS communication. As Yates’ ACMC corpus texts are unavailable, the lexical density of ACMC will not be discussed further here, apart from concisely corroborating, in this footnote, Yates’ (1993: 94) conclusion that ACMC and writing are close on this measure.

Table 4.4: *Unweighted lexical density for five corpora (LOB writing, ACMC and LLC speech from Yates 1993)*

Unweighted lexical density	
LOB writing	50.3
ACMC	49.3
LLC speech	42.3
Face-to-face SBC subset	36.6
SCMC	38.7
SSCMC	39.6

Judging from table 4.4, conversational writing ranks lower than LLC speech, but higher than face-to-face conversations from the SBC subset, as regards lexical density. Ure (1972) conducted a study of the lexical density of 30 written and 34 spoken texts, finding most written texts to have a lexical density of over 40% and most spoken under 40%.⁶⁷ Halliday's (1985a) example sentences contrasting written and spoken versions of the same messages display lexical densities above 45% for the written, and below 45% for the spoken versions. Halliday (1987), moreover, experiments with a passage of formal written English, rewording it in two steps into a "less written" and a "more spoken" version and finds the lexical density to dramatically decrease with increased "spokenness." His formal "written" version has a lexical density of 55%, his "less written" 47% and his "more spoken" version 39%. Even though no explicit dividing line is drawn in Halliday (1985a, 1987), one around 45% seems relevant. Stubbs (1996) finds a large overlap in lexical density among the genres of writing from LOB (with a range of 40 to 65 percent) and those of speech from LLC (with a range of 34 to 58 percent), and therefore no absolute difference between writing and speech, but establishes that the lexical density measurement is a "robust method of distinguishing genres" (1996: 76). In the LSWE corpus, Biber et al. (1999: 61) find "conversations" to have the lowest (41%) and "news" the highest lexical density (63%). Bringing the implicit dividing lines of these studies to bear on the results in table 4.4, we find conversational writing well settled on the spoken side of the continuum.

Interestingly, however, LLC speech and face-to-face conversations (SBC subset) diverge from each other slightly. As mentioned in section 3.4, LLC contains spoken texts of not just dialogs (e.g. face-to-face and telephone conversations),

67 Ure (1971) gives no account of what word classes were included among those with "lexical properties."

but also spoken texts of monologic character (e.g. broadcasts and speeches). In a discussion of LLC's genres, Stubbs (1996) discloses what might be suspected here, namely that the more monologic genres slightly boost the lexical density for LLC overall. Ure (1971) likewise notes higher lexical density among prepared than among unprepared spoken texts. More importantly, Ure (1971) makes a penetrating remark regarding texts with low lexical densities. She finds spoken texts with the lowest lexical densities to exclusively derive from sources where there is verbal response to the speaker, or some perceptible nonverbal response that would make the speaker adjust their language. This kind of response, known as *feedback*, she identifies as "an even more powerful factor in determining lexical density than the spoken/written choice" (1971: 448).

That feedback contributes to lower lexical density is borne out, in table 4.4, also in that conversational writing approximates the SBC subset face-to-face conversations more than does speech overall, or writing for that matter. Whereas the written genres of LOB contain monidirectional texts, the texts of the conversational writing genres, just like face-to-face conversations, are by default bidirectional. Besides feedback, Ure (1971) considers the influence of personal and social relations to have a bearing on lexical density, arguing that when impersonal texts coincide with those without feedback, the lexical density is increased. The face-to-face conversations from the SBC subset and the conversational writing texts, in table 4.4, all contain personal communication, some between previous acquaintances ("familiar" as opposed to "distant" relations, in Ure's 1971: 449 terms), which implies that their lexical density is loosened up.

What features in the face-to-face conversations from the SBC subset, then, account for giving the genre a lexical density below conversational writing? The answer lies not with the lexical items, but rather with a few of the grammatical ones, among which four stand out: face-to-face conversations contain more third person pronouns (as seen in section 4.2), more prepositions (to be explored in section 4.4), more of the impersonal pronoun *it*, and slightly more discourse particles, than does conversational writing. Example (8) from SBC serves to illustrate the abundance of grammatical items in face-to-face conversations, a discussion among friends cooking a meal together.

- (8) Roy: I could eat one of those.
Marilyn: You could?
Pete: Hm.
Roy: Well,
but I won't.
Pete: Then I guess
Roy: I mean,

Marilyn: Okay.
 Pete: Divide it in half.
 Roy: well don't
 Marilyn: Then I'll
 Roy: Y-
 What you oughta do though Mar,
 cook all the fish.
 Marilyn: Hm.
 Roy: Cause
 well,
 we won't use it,
 if you don't cook it.
 Now.
 Marilyn: Well I was gonna make ceviche with the leftovers.
 Roy: Oh alright,
 that sounds good.

Face-to-face conversations SBC text 3

Example (8) contains 65 words, only 17 of which are lexical items (*eat, guess, mean, divide, half, though, Mar, cook, fish, use, cook, make, ceviche, leftovers, alright, sounds, good*), yielding a lexical density of merely 26.2 for the passage. Among the grammatical words, there are three prepositions (*of, in, with*), three pronouns *it*, and as many as five discourse particles (*Well, well, Now*).

Pronoun *it* is often used as prop-*it*⁶⁸ in oral conversations but also substitutes for a range of referents, for “nouns, phrases, or whole clauses” (Biber 1988: 226). Referents in conversations are frequently tangible objects, as the fish in example (8). Chafe & Danielewicz (1987) explain the high frequency of pronoun *it* in spoken conversations thus:

Speakers not only have less time to choose vocabulary, but they also cannot or do not take the time to be as explicit about what they are referring to. A symptom of this kind of vagueness is the use of third person neuter pronouns, usually *it, this, or that*. Typically, the antecedent of a pronoun has been spelled out in an earlier noun phrase. (Chafe & Danielewicz 1987: 90, original italics)

Chafe & Danielewicz note that in conversations the antecedent is typically spelled out first, and then referred to by inference from the textual or situational

68 Prop-*it* is a dummy pronoun used as “‘empty’ or ‘prop’ subject, especially in expressions denoting time, distance, or atmospheric conditions” (Quirk et al. 1985: 348), e.g. “What time is *it*? *It*’s half past five,” but also for instance as “nonreferring” *it* with “vague implications of ‘life in general’, etc.,” e.g. “How’s *it* going?” (Quirk et al. 1985: 349, original italics).

context. In conversational writing, objects are rarely shared, or tangible, and deictic *it* therefore emerges less frequently than in face-to-face conversations. Instead, chatters by necessity refer to objects as nouns, i.e. as lexical items, which in turn contribute to the slightly higher lexical density figure for conversational writing.

The discourse particles annotated in the present study are *well*, *now*, *anyway*, *anyhow* and *anyways* (Biber 1988: 241), the first one by far the most frequent. Discourse particles are used to maintain conversational coherence (Biber 1988, Aijmer 2002). *Well* helps speakers in involved discourse monitor the information flow to the listener, and to ascertain that the communication is functioning smoothly (Chafe 1985, Schiffrin 1985); *now* is closely related to *well*, but also has a discourse-organizing function (Aijmer 2002: 57) and *now*, *anyway*, *anyhow* and *anyways* also function as emphatic topic changers. The moderate incidence of discourse particles in conversational writing implies that users find other ways to monitor the information flow, and to introduce new topics (ways expounded in Zitzen 2004). Another likely reason for their relative rarity (3.3 in IRC and 4.9 in ICQ, compared to 7.7 per thousand words in the SBC subset) is that the behavior of many users is governed by economy of typing. In spoken conversations, *well* frequently occurs in brief sequences of overlapping speech, when both speakers attempt to make their voices heard. Conversational writers do not encounter such situations, as every typed word is assumed to be read, and, consequently, for economy of typing users more often leave out discourse particles and cut straight to their message. Interestingly, however, Ko (1996) finds more discourse particles in his chat corpus than in face-to-face conversations, and attributes this to chatters' "increased need to monitor the flow of information in a situational context where there are multiple participants and no simultaneous feedback cues available to show listenership" (1996: no page number available). In the conversational writing corpora here, discourse particles are about as common as in the medium of speech overall (i.e. in Biber' spoken genres + the SBC subset, which together contain on average 4.2 discourse particles per thousand words).

The messages in IRC, i.e. the *turns*, contain only 4.3 tokens on average, while the turns in the annotated SBC subset contain 8.1 words on average (no equivalent figure for LLC speech was computed or found in the literature). Split-window ICQ turns, with 7.0 tokens on average, also appear to be shorter than in speech. As mentioned, however, comparing the turn length of split-window ICQ with those of IRC or speech is not practicable as, in split-window ICQ, turns are determined by the logging feature of the software, more than by the users. Consequently, for the analysis of textual complexity here, the *turn* is not an altogether reliable

construct. All the same, it must be recognized that the perceived complexity of texts relies not only on the overall lexical density of texts, how closely packaged the information is in general terms. The perceived complexity also depends on the packaging of the information into the constituent grammatical structures of the text. Halliday (1985a) identifies the most relevant of these structures to be the *clause*; “The clause is the grammatical unit in which semantic constructs of different kinds are brought together and integrated into a whole” (Halliday 1985a: 66). The clause is also seen as the most reliable construct upon which to carry out comparative investigations into the genre variation of language. For comparative purposes, the main requirement is consistency, and the clause is recognized as “perhaps the most fundamental category in the whole of linguistics,” as well as “critical to the unity of spoken and written language” (1985a: 67). Therefore, to relate the perceived complexity of texts to the discrepancy of clauses in spoken and written discourse, Halliday introduces the next measure to be considered, lexical density per clause.

The perceived complexity of a text depends not just on the lexical density overall, but also upon the composition of the text’s clauses, especially the length of clauses. The average clause in the annotated face-to-face conversations from SBC is 5.7 words long; in IRC it is only 3.9 and in split-window ICQ 4.6. “Lexical density per clause” indicates the number of lexical items per clause.⁶⁹ Consider again the ICQ turn from example (7), which will help to explain the calculation procedure:

“well yea but i wasnt even doing 75 | but i just said | i did | because i didnt want to fight the case”

The number of lexical items in the above turn is six (*even, just, said, want, fight, case*) and the turn consists of four non-embedded clauses (separated by vertical lines).⁷⁰ There are consequently, on average, 6/4, that is, 1.5 lexical items per clause, in this turn. Now contrast the chatted turn with a sentence from the biography genre of LOB:

69 In this study, Halliday’s (1985a) definition of the clause was observed, i.e. both finite and non-finite clauses were counted, whether independent (in “parataxis”) or dependent (in “hypotaxis”), but not restrictive relative clauses (which Halliday 1985a: 84 calls “embedded”). For further description of what constitutes a clause; see Halliday (1985a: 67ff).

70 In the example, “i did” is an instance of hypotaxis, but not “embedding,” in Halliday’s (1985a: 83) terms.

- (9) The story of the resplendent premiere, the gradual disintegration and eventual catastrophic debacle of this first French production of Don Giovanni can be followed in detail through the reviews in the contemporary press.

Biographies LOB G: text 44

The sentence from LOB contains 18 lexical items in one single clause, yielding an extremely high lexical density per clause: 18.0. Stubbs (1996: 75) finds the particular text from LOB (G: text 44, i.e. the full text) to have among the highest lexical densities of the written texts (58%), but does not carry the investigation further to the clausal level. Halliday (2004), however, explains that the complexity of spoken and written language is two-fold right down to the clausal level: the complexity of spoken language is grammatical, while that of written language is lexical. He describes the different complexities thus:

In spoken language, the ideational content is loosely strung out, but in clausal patterns that can become highly intricate in movement: the complexity is dynamic – we might think of it in choreographic terms. In written language, the clausal patterns are typically rather simple; but the ideational content is densely packed into nominal constructions: here the complexity is more static – perhaps crystalline. (Halliday 2004: 656)

Spoken language becomes complex by being grammatically intricate. Just as in spoken conversations, the ideational content of the split-window ICQ-turn above is “loosely strung out” (cf. Halliday 2004: 656), but the chatter “builds up elaborate clause complexes out of parataxis and hypotaxis” (cf. 2004: 654) (e.g. paratactical *but i just said* and hypotactical *i did* in the turn from example (7)). Written language, on the contrary, typically “becomes complex by being lexically dense: it packs a large number of lexical items into each clause” (2004: 654), even though the clausal pattern overall is rather simple (e.g. only one verb, *can be followed*, in example (9)). Halliday notes that the total number of lexical items in written texts usually just have “fewer clauses to accommodate them” (2004: 655). What typically happens in writing is that the lexical items are incorporated into nominal groups, as in example (9) (e.g. the long subject *The story of the resplendent premiere... Giovanni*). The nominal group is grammar’s primary resource for “packing in lexical items at high density” (Halliday 2004: 655).

Halliday (1985a), however, admits that the term “lexical density” is semantically loaded and repeatedly cautions against thinking of written texts as more complex: the measurement equally could have looked at the same phenomenon from the grammatical end; “[w]e could [say] that the difference between spoken language and written language is one of [grammatical] intricacy, the intricacy with which [spoken] information is organised” (1985a: 62). Halliday (1985a, 1987,

2004) therefore consistently calls spoken language more intricate than written. While spoken language represents phenomena as “processes,” written language represents phenomena as “products” (Halliday 1985a: 81); complex relationships are expressed “clausally” in spoken language and “nominally” in written language (Halliday 2004: 655; see also Castello 2008). Both kinds of complexity, nevertheless, can be accounted for under a single generalization, the notion of lexical density, which measures the different kinds of complexity, grammatical and lexical, that arise “in the deployment of words” (Halliday 1985a: 63). With that, we now turn to the measurement of lexical density within clauses with regard to the corpora annotated in the present study.

In the discussion of lexical density per clause here, only the results for the corpora annotated in the present study are tabulated, as no comparable average results were found for LOB, ACMC or LLC.⁷¹ As the numbers of lexical items had been identified already in the general lexical density calculation, the calculation of the new measure merely required the identification of the total number of clauses in each corpus. The total number of lexical items was then divided by total number of clauses for each corpus. Recall from the discussion of average turn length, in connection with example (6) above, that IRC turns are frequently very short (occasionally consisting of no more than a token, e.g. turns *true*, *ok*, *:0*, *hi*). Similar short turns are found in the SBC subset face-to-face conversations (*So*, *Well*, *No*, *Yeah*), and, although to a lesser extent, in SSCMC (*awwwwwwww* in split-window ICQ example (7)). No matter how short, a turn was always counted as, at least, one clause. The resulting average numbers of lexical items per clause are presented in table 4.5. As mentioned, this measure is known as “lexical density per clause” (Halliday 1985a, 2004), and it is found in the first column of the table.

Table 4.5: *Unweighted lexical density per clause and related measures*

	Lexical density per clause	Proportion of lexical items per clause	Average clause length
Face-to-face SBC subset	2.1	36.6%	5.7
SCMC	1.5	38.7%	3.9
SSCMC	1.8	39.6%	4.6

71 The calculation of lexical density per clause for LOB and LLC is beyond the scope of the present study. Yates (1993) presents no unweighted lexical density per clause for his ACMC corpus and, as mentioned in section 2.5, Collot (1991) did not study lexical density at all.

From table 4.5 it is clear that SCMC, i.e. IRC, contains the fewest lexical items per clause (1.5), but SSCMC, i.e. split-window ICQ, also has fewer lexical items per clause than face-to-face conversations. On the basis of various samples, Halliday (1985a: 80) notes that “a typical average lexical density [per clause] for spoken English is between 1.5 and 2, whereas the figure for written English settles down somewhere between 3 and 6.” Given Halliday’s well-established measure of lexical density per clause, then, conversational writing is definitely not typical writing, but rather shares an important defining characteristic of speech – a low lexical density.

Chatted and spoken texts are made up of large numbers of interrelated short clauses, whereas traditional writing contains longer integrated clauses. This means that any vital interpretation of the lexical density per clause, in table 4.5, must be accompanied by the consideration of average clause length in each of the three media (tabulated in the third column of table 4.5). Furthermore, to explain the utility of the lexical density per clause measure, a provisional measure is interspersed into table 4.5: the proportion of lexical items to total items in the average clause, termed “proportion of lexical items per clause.” From this measure, found in the second column, we can deduce a one-to-one correspondence with the figures presented in table 4.4 for lexical density overall (e.g. SCMC’s proportion of 38.7% lexical items in the clause, in table 4.5, is reflected in its corresponding overall lexical density of 38.7, in table 4.4).⁷² The provisional measurement is provided here to demonstrate the one-to-one relationship between Halliday’s measures of lexical density and lexical density per clause, that the measures in reality are the same. Halliday’s application of the lexical density measure on the clausal level simply underscores the variability of clause length in different genres. Lexical density per clause is a more sensitive measure of lexical density, one that takes into account the number of clauses in texts of equal length and generates more explicit differences in score.

Comparing numbers of lexical items in the clause is a straightforward task; as seen in table 4.5, a typical spoken clause contains more than two lexical items, whereas a chatted clause contains fewer than two. More intriguingly, as turns in conversational writing most frequently consist of a single clause, the measure of lexical density per clause provides a glimpse into the typical turn of chatted interaction. The measure thus enables us to capture the special properties of

72 The percentages indicated in the “proportion of lexical items per clause” column in table 4.5 are based on the unrounded figures for lexical density per clause divided by unrounded average clause length.

chatted language, and their relationship to face-to-face conversations. From table 4.4 we rated that conversational writing ranges slightly higher than face-to-face conversations as regards lexical complexity overall, whereas in table 4.5 the lexical density per clause measure renders a slightly nuanced picture, one which draws attention to the short average clause length in conversational writing.

Calculating the average clause length in LOB writing and LLC speech, or in any other written or spoken corpus, is unfortunately beyond the scope of this study, even though such a project, for contrastive purposes, is highly recommended and anticipated. Until such figures are obtained, the analysis of the lexical complexity of conversational writing is bound to remain a preliminary one. Nevertheless, given Halliday's finding that a typical average lexical density per clause for written English "settles down somewhere between 3 and 6" (1985a: 80) and the concurrent general findings of lexical densities around 50% for writing (see table 4.4), a reasonable deduction is that writing on average contains more than six words per clause, possibly up to twelve. Based on the discussion of numerous invented examples, Halliday assumes that the lexical density per clause for writing is "likely to be of the order of twice as high as that for speech" (1985a: 80). Chafe & Danielewicz (1987) discuss clause construction in spoken and written language, finding "intonation units" (the majority of which are clauses) to vary in length, from 6.2 words per unit in conversation to 9.3 in academic papers. "[U]nder normal conditions," they explain, "a speaker does not, or cannot, focus attention on more than can be expressed in about six words" (1987: 95). Chafe & Danielewicz (1987) point out that writing frees writers from the constraint of production time that keeps down both the lexical variety of spoken language and the size of spoken intonation units. Although their argument holds true for traditional writing, it is inapplicable to conversational writing; chatters are highly constrained in time; they produce even shorter clauses than speakers, and chatted clauses contain fewer lexical words than spoken clauses. Given Halliday's (1985a) assumption that the lexical density for writing is likely to be twice as high as that for speech, it will be interesting then, in the future, to find out the lexical density per clause relationship between writing and conversational writing. A plausible assumption is that it will be of the order of three times as high in writing as in conversational writing.

In conclusion, the measures applied in the analysis of the lexical diversity and specificity in conversational writing have yielded a number of important findings. Firstly, the average word in the computer chats is shorter than in any other medium. Short words are seen as an effect of the immediacy of the online

medium, the short encoding time and users' economy of typing, but also as an effect of the words belonging to the grammatical classes – findings which in turn accentuate the similarity between conversational writing and spoken conversations. Secondly, the type/token ratio of conversational writing is by definition a high one; the tokens in conversational writing display a striking lexico-orthographic heterogeneity, an abundance of types. This heterogeneity is explained by the particularities and irregularities of the uniquely user-generated material, in sharp contrast to the nature of corpora of traditional edited writing and the consistently transcribed corpora of speech. TTR is ultimately deemed an inadequate tool for determining the nature of conversational writing as regards its relationship to traditional writing and speech. Thirdly, a more reliable tool for measuring the lexical complexity of the chatted texts was the measure of lexical density, which finds the ratio of lexical items to all items in the texts, thereby reflecting the relationship between lexical and grammatical items. Conversational writing presents itself with a lexical density intermediate between the average for speech from LLC and face-to-face conversations from the SBC subset. The measure of lexical density per clause, finally, reveals the character of typical clauses in conversational writing, presenting their fewer lexical items per clause than in the face-to-face spoken texts. Complemented with average clause length, the measure was used to determine the proportion of lexical items in clauses. The preliminary results, pending the calculation of average clause length in writing, give the impression that conversational writing is slightly more lexically dense than face-to-face conversations, even on the clausal level, even though the lexical density per clause measure better than the overall lexical density measure manages to accentuate and reflect the short average turns of conversational writing. Taken together, the results point in the direction that conversational writing is a variant of spoken communication, or more precisely: a means of communication in which users package information in grammatically intricate ways that are definitely more speech- than written-like.

4.4 The most salient features

In section 4.2, first and second person pronouns were mentioned to be the first two features, in either of the conversational writing genres, that deviate by more than two standard deviations ($|s.d.| > 2.0$) from the average of Biber's spoken and written genres (Appendix II table 4 from Biber 1988: 77–78). They were taken up in connection with modal auxiliaries as these grammatical categories together are two of the carriers of interpersonal meaning (Fowler &

Kress 1979, Halliday 1985a, Hodge & Kress 1988). It was mentioned in the section, however, that altogether ten out of Biber's 67 features deviate in such a way, and that the present section is dedicated to the other eight: direct WH-questions, analytic negation, demonstrative and indefinite pronouns, present tense verbs, predicative adjectives, contractions and prepositional phrases. Table 4.6 summarizes the frequencies per thousand words of these salient features. By their sheer frequency in the chatted texts (or infrequency, in the case of prepositional phrases), these features together give an intimation of the linguistic character of conversational writing. Out of the features to be taken up below, the first two (direct WH-questions and analytic negation), like modal auxiliaries and personal pronouns, also form part of the interpersonal system in language: direct WH-questions as markers of mood, and analytic negation (*not*, *n't*) as a marker of polarity within the modality system (Halliday & Hasan 1989). The remaining six features do not conform as clearly as these to any one of Halliday's metafunctions in language, but will be surveyed on their own terms, as their distributions reveal important patterns. The two ensuing sections, 4.5 and 4.6, will give an account of other salient features of conversational writing, features not found through Biber's (1988) methodology, which nevertheless are instrumental for determining the nature of the communication. Once all of these features have been considered, we will be ready to apply the final step of Biber's (1988) methodology, to position conversational writing on Biber's six dimensions of variation (in chapter 5).

Table 4.6: *Frequencies per 1,000 words for the most salient linguistic features (i.e. normalized frequencies). "N.a." means that the figure is not available*

	Writing	ACMC	Speech	SCMC	SSCMC
first person pronouns	17.0	57.8	52.8	56.9	88.9
second person pronouns	5.0	17.6	23.0	50.4	45.0
direct WH-questions	0.1	2.9	0.8	3.5	3.9
analytic negation	6.4	15.1	13.9	13.1	29.7
demonstrative pronouns	2.3	6.5	10.6	6.6	16.4
indefinite pronouns	0.9	4.6	3.1	11.7	6.0
present tense verbs	64.6	67.6	112.3	147.2	168.5
predicative adjectives	4.8	n.a.	4.9	8.4	15.3
contractions	4.6	16.6	36.1	30.8	55.0
prepositional phrases	117.3	116.9	91.1	47.0	42.0

Halliday (1985a, 2004) and Halliday & Hasan (1989) propound the theory of metafunctions in language, because “it helps us to interpret the features that we actually find in the text” (Halliday & Hasan 1989: 35–36). The variables of “field,” “tenor” and “mode” “collectively determine the functional variety, or register, of the language that is being used” (1985a: 44). The interpersonal metafunction, the tenor of the communication, reflects the personal relationships involved and is realized in texts through e.g. modal auxiliary use (the hedging of statements), personal pronouns (the presentation of self), both dealt with in section 4.2, as well as through mood (declarative, imperative or interrogative) and the system of polarity (the use of negation). As it turns out, the last two grammatical categories, just like the first two, contain features that distinguish the five media contrasted in this chapter from each other. With regard to the grammatical category of mood, only the interrogative mood is annotated in the texts, in the form of direct WH-questions (detected as WH-pronoun, e.g. *what, where, when, how, why*, + auxiliary), but its distributional pattern reveals the inherently communicative function of conversational writing. Analytic negation (*not*, including the contracted form) is found in previous research to correlate with spoken, communicative texts, and by analogy conversational writing could be expected to follow and display a similar distribution. The five media contrasted, as before, are writing, ACMC, speech, SCMC and SSCMC. ACMC is included for reference in the diagrams but, as the corpus is unavailable, no ACMC text examples will be given. Rather, the survey of all features below focuses on the distributions of the features in writing, speech and the conversational writing genres. Figures 4.6 and 4.7 present the distribution of interrogative WH-questions and analytic negation in the five media. Figures 4.6–4.13 in the present section all reflect table 4.6, representing occurrences per thousand words (i.e. normalized frequencies). All figures and tables in the present and ensuing sections of this chapter are based on average numbers from Biber 1988: 247–263 for writing, Collot 1991: 69–70 for ACMC, Biber 1988: 264–269 and Appendix II table 3 for speech, Appendix II table 1 for SCMC, and Appendix II table 2 for SSCMC, unless otherwise indicated, and the results of statistical tests between SCMC, SSCMC, writing and speech, as before, are found in Appendix VI.

Figure 4.6: Direct WH-questions.

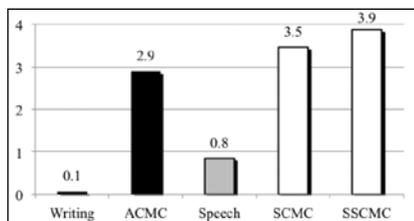
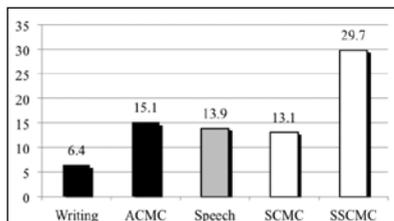


Figure 4.7: Analytic negation.



Questions, both yes/no questions and WH-questions, indicate “a concern with the interpersonal functions and involvement with the addressee” (Biber 1988: 227). Yes/no-questions cannot easily be identified by automatic analysis, and were therefore not included in Biber’s (1988) methodology, but WH-questions, which are more easily identified, were tagged and counted in all of Biber’s genres (in Biber 1988: 247–269), as well as in Collot’s genre of ACMC (in Collot 1991), and, in the present study, in the SBC subset face-to-face conversations (amalgamated into the bar for speech in figure 4.6; see Appendix II table 3 for the frequency in the SBC subset) and the conversational writing genres (SCMC and SSCMC; cf. Appendix II tables 1–2). Biber et al. (1999: 203) point out that “interrogative clauses tend to occur in dialogue situations,” and that “they are frequent only in conversation and (to a lesser extent) in fiction” (ibid.). Judging from figure 4.6, however, Biber et al.’s statement is up for qualification; direct WH-questions are used to an even higher degree in CMC than in spoken conversations (also noted by Ko 1996).⁷³ Figure 4.6 underscores the interpersonal and involved character of computer-mediated discourse: while direct WH-questions are nearly absent in traditional writing, they are slightly more common in speech, and very common in conversational writing. Among the genres amalgamated into the speech bar in figure 4.6, are face-to-face conversations from LLC and from the SBC subset, which contain 0.7 and 2.7 WH-questions per thousand words, respectively, and telephone conversations with 1.1 – genres that contribute to raising the overall figure for speech, but that nevertheless are surpassed by all modes of CMC. Typical WH-questions in IRC are *Where do you come from?*, *What do you do?*, *How are you doing?*, and in ICQ *What are you doing this weekend?*, *How did that go over?*. The slightly different nature of the questions in IRC (more general) and in split-window ICQ (more specific), moreover, reveals the status of the relationships in the two corpora of conversational writing; the IRC chatters are beginning their acquaintance, whereas the ICQ

73 Freiermuth (2003) finds questions overall many times more frequent in chat than in speech and writing, but his results do not specify the occurrence of WH-questions.

chatters inquire into mutually known circumstances – revealing their previous acquaintance with each other. Ko (1996) aptly explains chatters' frequent questions as partly a cohesive strategy; as participants' physical separation obstructs them from coherent and orderly patterns of turn-taking, frequent WH-questions helps them to structure the interaction, "in compensation for the unavailability of other turn-taking cues such as intonation, gesture, and gaze" (1996: no page number available).

Analytic negation (*not*, including the contracted form) and synthetic negation (*no*, *neither*, *nor*) are devices grammaticalized in language for speakers and writers to express negative opposition. Analytic negation typically occurs in conjunction with finite verbs (Biber et al. 1999, Halliday 2004), e.g. *doesn't*, *isn't*, *can't*, and realizes "an essential concomitant of finiteness": polarity, i.e. "the choice between positive and negative" (Halliday 2004: 116). Tottie (1981, 1983b, 1991) finds negation overall to occur twice as often in speech as in writing. Tottie (1991), furthermore, finds "*not*-negation to prevail in spoken language" and "*no*-negation to dominate in written language" (1991: 140, original italics). Biber (1988), like Tottie (1983a), distinguishes between analytic and synthetic negation, finding analytic negation (e.g. *she didn't write any letters that day*) to be more colloquial and fragmented, and synthetic more literary and integrated (e.g. *she wrote no letters that day*). In Biber (1988), accordingly, analytic negation is found to be more than twice as frequent in communicative, spoken interaction, than in written discourse, a finding reflected in figure 4.7. Similarly, in the LSWE corpus, Biber et al. (1999) find negative forms overall to be many times more common in conversation than in writing, with analytic negation most common in conversation and synthetic most common in news. Given the conversational nature of computer chat, analytic negation, as might be expected, turns out to be prevalent in the chats, making the feature deviate markedly in SSCMC from spoken and written language overall. In figure 4.7, SCMC shows a distribution of analytic negation similar to speech, although notably lower than the face-to-face conversations in the SBC subset, which contain 18.9 occurrences per thousand words (Ko 1996, however, finds more analytic negation in his SCMC corpus than in face-to-face conversations). SSCMC, by contrast, contains more than twice as many occurrences of analytic negation as does speech overall.

Upon studying the occurrences of analytic negation in both genres of conversational writing, a few functional distinctions can be made. In IRC, *not* frequently occurs in answers to questions like *How are you?* and *What's up?*: e.g. *not too bad*, *not much*, *not much, u?* and in other generally mitigated, friendly expressions like *don't miss me too much*, *don't mean to sound ungrateful*, and *you have a good day*

now, won't you? The nature of negated expressions in IRC thus reveal the ephemerality, or tentativeness, of relationships formed in the channels. In split-window ICQ, by contrast, analytic negation is often found in connection with adversarial discussions, such as in *DON'T EVEN START WITH ME!!!!!!*, *that wasn't me, it's not funny*, but also in connection with involved, supportive discourse: *so how come you don't talk to mike anymore?*, *I can't take when he is in a bad mood* – which reveal participants' close relationships in real life, outside of the medium. In the split-window ICQ corpus, moreover, turns are occasionally hedged with the abbreviation *idk* (meaning “I don't know”),⁷⁴ a mitigating “marker of uncertainty” (Tsui 1991: 619, Diani 2004: 162) such as in *like idk i'm one of those scarcastic girls...*, *idk he's confusin*, a typically spoken feature not found in the IRC corpus. The ICQ communication thus, more than IRC, serves as an extension of the face-to-face interaction that takes place regularly between interlocutors – involving both adversarial and supportive discourse, as well as mitigation. Tottie (1982, 1983b) attributes the greater frequency of analytic negation in spoken than in written language to the greater frequency of denials, rejections, questions, supports, repetitions and mental verbs in speech. Several of Tottie's (1983b) fundamental categories of negative sentences (e.g. denials, rejections and supports) appear to be more frequent in split-window ICQ than in IRC. In addition, the distribution of what Tottie calls mental verbs (e.g. *know*, *think*, *mean*), largely private verbs in Biber's (1988) methodology, is much higher in split-window ICQ than in IRC (cf. Appendix II tables 1 and 2). Split-window ICQ contains more affective discussions than IRC, with expressions of denial, rejection, support and opinion that ICQ chatters recurrently modalize by means of negative polarity.

The modality system of language, manifested inter alia in modal auxiliary use, choice of mood, negation, and the insertion of a mitigator like *idk*, is available to speakers for encoding attitude towards a statement or the content of an utterance (Hodge and Kress 1988, Yates 1996, Halliday 2004). Hodge & Kress (1988) explain the effect of modality thus:

Modality expresses affinity – or lack of it – of speaker with hearer via an affirmation of their affinity about the status of the mimetic system. Affinity is therefore an indicator of relations of solidarity or of power [...] A high degree of affinity indicates the expression of solidarity between participants. A low degree of affinity indicates that power difference is at issue. (Hodge & Kress 1988: 123)

74 In the conversational writing annotation, initialisms like *idk* and *nm* (meaning “not much”) were tagged as if their constituents were spelled out, finding analytic negation in them. (This treatment, however, was not applied to sentiment initialisms such as *lol* and *lmao*, to be treated in section 4.6, as explained in section 3.2.)

The present chapter has revealed a high degree of modality in SSCMC, e.g. a great number of modal auxiliaries (figure 4.1), frequent switches into the interrogative mood (figure 4.6), prevalent use of the polarity indicator *not* (figure 4.7) and insertion of a hedge such as *idk*. The findings all highlight a significant situational circumstance of the ICQ communication: the ICQ chatters are interpersonally involved in not just the online medium, but also in the offline world, and experience close affinity in both modes. This high degree of affinity, expressed through highly modalized language, indicates, in Hodge & Kress' terms, "solidarity between participants" (1988: 123). Chatters in IRC modalize their utterances to about the same degree as speakers as regards modal auxiliaries, but slightly less than speakers as regards analytic negation. On the other hand, IRC chatters switch into the interrogative mood more than speakers, which reveals that they, too, are interpersonally focused, even though the relationships formed in the public IRC channels tend to be of a more superficial nature. In conclusion, relating Halliday's metafunction of tenor to relevant features annotated in the conversational writing texts has shed light on the relationships among interlocutors and yielded insights into the functions served by the respective media. In what follows, Halliday's metafunctions are left aside for a while, but we will find reason to return to Halliday in other respects shortly.

The next two features that deviate from the mean for all of Biber's spoken and written genres (Appendix II table 4) by more than two standard deviations, in either of the conversational writing genres, are demonstrative pronouns and indefinite pronouns. Biber (1988) subsumes these two features, together with "pronoun *it*," under the heading "impersonal pronouns," in contrast to "personal pronouns" (Biber 1988: 225–226). In the present chapter, personal pronouns have been discussed at length, as first and second person pronouns are clear markers of involved discourse. Pronoun *it*, furthermore, was mentioned in connection with the finding of slightly more grammatical items in the face-to-face conversations from the SBC subset than in the conversational writing genres, which rendered a lower lexical density for face-to-face conversations. Pronoun *it* was found to occur more often in face-to-face conversations partly because of the deictic function it can serve there (cf. the discussion of example (8) above), in addition to the function of substituting for nouns, phrases and whole clauses. Demonstrative pronouns (*that, this, these, those*)⁷⁵ and indefinite pronouns (e.g.

75 "Demonstrative pronouns" here constitute Biber's (1988) feature no. 10, that is (a) *that/this/these/those* followed by verbs, clause-punctuation, tone-unit boundaries, wh-pronouns or conjunction *and*, (b) *that's* and (c) *that* immediately after a tone unit boundary; see Biber (1988: 226) for algorithms. *That* as relative pronoun is not included. Note

everyone, somebody, anything, nothing), as we shall see, can serve similar functions in conversational writing.

Biber (1988: 226) notes that demonstrative pronouns can refer to “an entity outside the text, an exophoric referent, or to a previous referent in the text itself.” Biber et al. (1999: 349) find demonstrative pronouns “far more common in conversation than in the written registers” and demonstrative pronoun *that* in conversation “by far the single most common demonstrative pronoun” (ibid.). As regards indefinite pronouns, they find all groups (the *every-*, *some-*, *any-* and *no-* groups) to be most common in conversation and fiction, and least common in academic prose (1999: 353). The distributions of demonstrative and indefinite pronouns in the five media contrasted in the present chapter are given in figures 4.8 and 4.9.

Figures 4.8 and 4.9 are best understood by studying sample occurrences of the features in the texts. Among the demonstrative pronouns, *that* is the most frequent one in the face-to-face conversations from SBC and in conversational writing genres, and it is used in analogous ways in the three genres, that is, to denote a previous referent in the text itself; see italicized phrases in examples (10) from SBC, (11) from IRC and (12) from split-window ICQ.

Figure 4.8: Demonstrative pronouns.

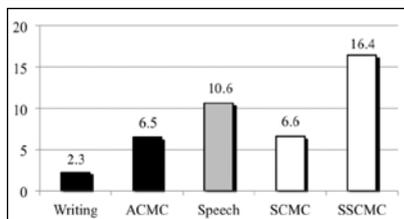
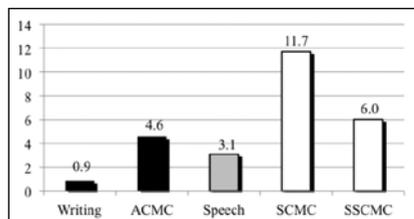


Figure 4.9: Indefinite pronouns.



- (10) Phil: they asked me *to meet with them about ... Teresa's thing.*
 Brad: Mhm
 Phil: ... *that* .. I find v- really,
 ... nothing,
 ... to be honest,
 nothing of any validity.

Face-to-face conversations SBC text 10

that feature no. 10 differs from feature no. 51 “demonstratives” *that/this/these/those*, in that the demonstratives in feature 51 are followed by nouns (e.g. *this thing*).

- (11) <furryman> so will it be a *long interview* blondii
 <blondii> *that* depends on you
 Internet relay chat text 2b (UCOW)

- (12) <1> So *what do you think about Joey?*
 <A> what kind of question is *that*
 Split-window ICQ chat text 1 (UCOW)

Pronouns *that* in examples (10) through (12) refer to events, states, or phrases, rather than to nominal referents. Chafe (1985) finds demonstrative pronouns referring to events and states to occur predominantly in speaking, prescriptively claiming that they are among the several “grammatical devices that are not accepted in written English” (1985: 114). The demonstrative pronouns *that*, *this*, *these* and *those* are naturally inherently deictic in all three corpora, referring to phrases or whole utterances, but also referring to specific nominal referents, such as the italicized noun phrase in (13).

- (13) <AdamSxy35> oups why dont you try a *business chat room on yahoo?*
 <_oups> hm...well do they have *that*..
 Internet relay chat text 5b (UCOW)

Demonstrative pronouns are typically found in passages of involved discussion in all three sampled corpora (see examples 10–13). However, since such affective, involved passages are more rare in IRC than in the face-to-face conversations from SBC or in split-window ICQ, the overall incidence of demonstrative pronouns drops for SCMC in figure 4.8. In speech, demonstrative pronouns can also refer to nominal referents *outside* of the text, e.g. *this is cream soda* (although one might argue that the referent is cataphoric here). Such use is frequent in, for instance, a minor part of LLC, a physics demonstration, but occurs only marginally more in the annotated face-to-face conversations from SBC than in conversational writing. The number of demonstrative pronouns in the SBC face-to-face conversations is 16.0 per thousand words – roughly the same as in ICQ. Chatters in split-window ICQ, in other words, well manage to bridge over the spatial distance between themselves and put the demonstrative pronouns to text-internal deictic use. Their conversations via the written online medium largely follow the same pattern as face-to-face conversations, as regards demonstrative pronouns.

Indefinite pronouns (e.g. *anybody*, *everyone*, *something*) are another feature that split-window ICQ chatters and face-to-face conversationalists employ to an approximately equal extent: 6.0 per thousand words in ICQ, see figure 4.9, and 6.6 in the face-to-face conversations from SBC. Examples are *do you like someone else* and *its great being someone who can be mentor*, in ICQ, and *you have something on your tooth*, in SBC. The SBC example reveals a usage not found in

the conversational writing texts – a reference to something specific, visible to the speaker, but not to the listener. *Something* in the chats, just like the other indefinite pronouns, always refers to a general idea, concept or phrase, or an indefinite person or thing, not immediately visible. Indefinite pronouns are “markers of generalized pronominal reference, in a similar way to *it* and the demonstrative pronouns” (Biber 1988: 226, original italics). The split-window ICQ and face-to-face conversations display functionally analogous usage of indefinite pronouns, numerically on a par, but as can be seen in figure 4.9, indefinite pronouns are almost twice as common in IRC as in ICQ.

What brings about the high frequency of indefinite pronouns in IRC? The answer to the question is very simple, and it is immediately discernible in the various occurrences sampled from IRC in (14 a–f).⁷⁶

- (14) a. *anyone* wanna chat
b. *anyone* from sydney
c. hello *everyone*
d. how old's *everyone*?
e. *Anybody* here???
f. wassup with *everyone* today

Internet relay chat (UCOW)

IRC chatters employ indefinite pronouns when angling for conversational partners in the channel, but also in greetings and questions intended for indefinite recipients. This kind of usage accounts for half, or more, of the indefinite pronouns in the IRC texts, which wholly explains the high frequency of indefinite pronouns found for SCMC in figure 4.9. Noting similar results for his chat corpus, Ko (1996) relates the high frequency of indefinite pronouns to the situational context; “[u]sers do not know for certain who their audience is at any given moment” (1996: no page number available).

As mentioned, Biber et al. (1999) find in LSWE approximately the same number of indefinite pronouns in fiction as in conversation (a rough estimate is 5 per thousand words, in each genre, for the same indefinite pronouns that Biber 1988 considers), a number significantly higher than in the other genres they studied: news and academic prose. Biber et al.'s (1999: 352) finding for fiction, however, does not tally with Biber's (1988) figures for fiction. Among Biber's (1988) written genres in figure 4.9, a fiction genre, adventure fiction, contains the highest number of indefinite pronouns (2.7 per thousand words) but most other genres, including

76 As explained in section 1.4, no text numbers are given for examples that contain turns sampled from several texts.

the other fiction genres, contain fewer than 2 per thousand words. Thus, no parallel similar to Biber et al.'s (1999) finding can be drawn between fiction and conversations, or conversational writing, in the present study. Among Biber's (1988) spoken genres in figure 4.9, the highest number (3.9) is recorded for face-to-face conversations, closely followed by telephone conversations (with 3.6). All three corpora annotated in the present study thus display a usage of indefinite pronouns beyond previously recorded findings. In conclusion, indefinite pronouns in conversational writing are worthy of mention not just for contributing to the obviously oral character of the online communication, but also for distinguishing functionally among the conversational writing genres. Their use in IRC reveals one of the main functions of the public medium: finding conversational partners.

Out of the ten features in conversational writing that deviate by more than two standard deviations from Biber's mean for speech and writing, only two, in themselves, may constitute lexical items: present tense verbs and predicative adjectives, both to be taken up next (direct WH-questions and prepositional phrases, of course, may also contain lexical items). However, while predicative adjectives are lexical by default, a vast number of present tense verbs are forms of the verbs *be*, *have* and *do*, which are grammatical items (cf. Quirk et al. 1985: 67 and section 4.3 here). Judging from the low lexical density found for conversational writing in the previous section, the prevalence of grammatical items among the most salient features tallies with possible expectations; conversational writing indeed contains many more grammatical than lexical items. The distributions of present tense verbs and predicative adjectives in the five media are shown in figures 4.10 and 4.11.⁷⁷

Figure 4.10: Present tense verbs.

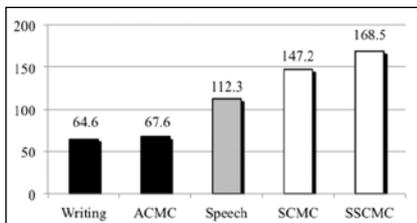
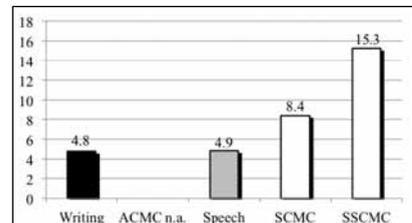


Figure 4.11: Predicative adjectives.



77 The figure for predicative adjectives in ACMC is not available (Collot 1991: 69).

Present tense verbs and predicative adjectives do not share pragmatic-functional properties like the pairs of features treated above (direct WH-questions and analytic negation as parts of the modality system, and demonstrative and indefinite pronouns as markers of impersonal pronominal reference). The ten features in this account are all, naturally, entirely unforeseen, as they have crystallized from their sheer frequency in the conversational writing genres and not by kinship or the author's choice. The pairing of features in the present section is thus mostly incidental and applied for practical, rather than necessarily linguistically motivated, reasons. All the same, when studying the textual occurrences of one of the ten features, very often another one pops up. Such is the case for nearly all of the occurrences of predicative adjectives in the annotated corpora; examples from SBC are *That's not bad, Tha- that's right, They're cool*; from IRC *this is slow, i'm lost, your welcome* and from ICQ *its ok, that's cute!!, That's pretty cool...*⁷⁸; in which the predicative adjectives tend to co-occur with present tense verbs, very often with demonstrative pronoun that, and sometimes with analytic negation (not).

Present tense verbs is one of the features that carry the largest weight on Biber's (1988) first dimension, distinguishing texts with highly involved, interactive discourse from texts with more informational content. Figure 4.10 illustrates the pervasiveness of present tense verbs in speech, as opposed to writing. Present tense verbs "deal with topics and actions of immediate relevance" (Biber 1988: 224), whereas past tense and perfect aspect verbs are typically markers of narrative or descriptive, mostly written, texts (Biber 1988, Biber et al. 1999). On Biber's (1988) first dimension ("Informational vs. Involved Production," to be discussed in chapter 5), present tense forms indicate a verbal (involved), as opposed to nominal (informational), style. Spoken language is typically verbal, interactional and affective, whereas written language is nominally elaborated (cf. Wells' 1960 verbal and nominal styles). Judging from figure 4.10, the verbal, involved style found in speech is augmented further in conversational writing – in split-window ICQ chat, practically every sixth word is a present tense verb, in writing, by contrast, only every sixteenth.

Among the spoken genres in figure 4.10, the face-to-face conversations from SBC contain 141.6 per thousand words, face-to-face conversations from LLC 128.4 and telephone conversations 142.6; the three highest figures for speech. Sample occurrences of present tense verbs in SBC and the conversational writing genres can be found in any one of the numbered text examples given in this chapter, requiring only a few examples to be given here. Furthermore, in section

78 Biber's (1988: 238) algorithm 41 (b) for finding predicative adjectives was interpreted "be+adv+adj+xxx (where xxx is not adj or n)."

5.2.1, the impact of present tense verbs for distinguishing among genres will be discussed, in conjunction with examples from the annotated corpora. Clearly, present tense verbs contribute to a sense of orality in conversational writing, a sense that is further born out in that they are frequently also private verbs (e.g. *feel, know, think, guess*). Present tense verbs are about as indicative of speech as nouns are of writing. In fact, if asked to distinguish among genres by one word class alone, opting for verbs might prove a felicitous choice, as by their character, tense and frequency, verbs reveal a great deal about a text's genre affiliation. Private verbs, for instance, on average occur twice as often in Biber's (1988) spoken, as compared to his written genres. Present tense verbs, moreover, occur as private verbs twice as often in split-window ICQ as in IRC; in ICQ to about the same extent as in face-to-face conversations and in IRC almost as infrequently as in writing. Examples of present tense private verbs in ICQ are *I know, i think i'll just..., that's cool I guess, I mean if you like him* – typically used to introduce evaluative and emphatic utterances. The relative rarity of private verbs in IRC is likely to be due to the superficial character of relationships in the public channels; interlocutors simply do not know each other well enough to discuss preferences, express evaluation or give supportive advice. The low frequency of private verbs in IRC, in turn, is a likely explanation for the slightly fewer present tense verbs found for IRC (SCMC), as compared to split-window ICQ (SSCMC), in figure 4.10. What this means for IRC, with regard to Biber's first dimension, will be further explored in the next chapter.

Turning now to predicative adjectives (figure 4.11), a few remarks are in order. Firstly, "predicative adjectives" is not a factor that distinguishes among Biber's (1988) genres of writing and speech. As seen in figure 4.11, writing and speech contain approximately the same number of predicative adjectives (4.8 and 4.9 per thousand words, respectively). Consequently, predicative adjectives, as a linguistic feature, did not load on any of Biber's (1988) dimensions of genre variation (unlike all other features discussed in this section).⁷⁹ Collot (1991), therefore, decided not to count the feature, which is why no result is available for ACMC in figure 4.11. For IRC and split-window ICQ, however, identifying and summing the predicative adjectives has proved highly valuable: figure 4.11 indicates that SCMC contains nearly twice as many, and SSCMC more than three times as many predicative adjectives as writing and speech, respectively. This means

79 Predicative adjectives loaded tentatively on Biber's (1988) fifth dimension, but their low weight (0.31) was below the cut-off point (0.35) for the feature to be considered in dimension score calculations.

that, if a new factor analysis was carried out with the inclusion of chatted texts, predicative adjectives may turn out to load on one of the resulting dimensions.⁸⁰

While predicative adjectives do not distinguish between written and spoken genres in Biber's (1988) study, nor in Chafe's (1982) account, attributive adjectives do. In Biber's (1988) methodology, attributive adjectives are identified as all adjectives preceding nouns, or otherwise "not identified as predicative" (1988: 238). Attributive adjectives are used to elaborate nominal information, and thus highly integrative in their function (Chafe 1982, Chafe & Danielewicz 1987, Biber 1988), while predicative adjectives "might be considered more fragmented" (Biber 1988: 237). Chafe (1982) notes that the use of attributive adjectives "allows states to be expressed as modifiers rather than assertions," e.g. "the *old* house," as opposed to "the house was *old*," and calls them integrative devices and a prevalent feature of written language (1982: 41–42, original italics). To delve beyond the equal figures for writing and speech in figure 4.11, therefore, we will consider the ratio of predicative to attributive adjectives in writing, speech, SCMC and SSCMC. The results of such a calculation reveal that, in Biber's (1988) genres of writing, only every fifteenth adjective is predicative; in speech, every tenth (although in the SBC subset as many as every fifth); in SCMC, every seventh; and in SSCMC, as many as every third adjective is a predicative adjective. If attributive adjectives are typical of nominal, written discourse, then the relative rarity of attributive, and the prevalence of predicative adjectives, is typical of conversational writing. Although previous studies have not found predicative adjectives to be typical of speech, the present study finds predicative adjectives to be highly typical of conversational writing.

Biber (1988) notes that predicative adjectives are frequently used for marking stance. Biber et al. (1999) find that, "(s)emantically, the most frequent predicative adjectives of conversation tend to be evaluative and emotive, e.g. *good*, *lovely*, and *bad*" (1999: 516, original italics). The examples of predicative adjectives from SBC, IRC and split-window ICQ given above (in connection with the discussion of present tense verbs) confirm Biber et al.'s finding for conversation, for the conversational writing genres: the predicative adjectives in conversational writing are also largely evaluative and supportive responses to statements made by partners in the online conversations. Two additional such examples conclude the account of predicative adjectives here: example (15) illustrates a typical occurrence in IRC and example (16) one in ICQ, both evaluative and/or supportive.

80 Interestingly, face-to-face and telephone conversations from LLC contain 4.2 and 6.0 predicative adjectives per thousand words, respectively, whereas face-to-face conversations SBC contain 8.2, possibly suggesting that they are becoming more frequent in conversations, or simply are more frequent in American than in British English conversations.

- (15) <yazzie^> I'm coming to Aussie next Xmas 404!!
 ...
 <Guest22> wow yazzie, thats *great*...for how long
 Internet relay chat text 4b (UCOW)
- (16) <6> but if practice makes perfect and no ones perfect then y practice?
 <F> gotta practice! It makes perfect ya know
 <6> i try
 <F> thats *deep*
 Split-window ICQ chat text 5 (UCOW)

The final two features, out of the ten that collectively epitomize the character of conversational writing (by deviating from Biber's mean for speech and writing by more than two standard deviations), are contractions and prepositional phrases. Contractions deviate by their high frequency in split-window ICQ, and prepositional phrases by their striking *infrequency* in both conversational writing corpora. The two features are entirely independent of each other in the texts, but they share the ability to distinguish among writing, speech and conversational writing. The distributions of the features are shown in figures 4.12 and 4.13.

Figure 4.12: Contractions.

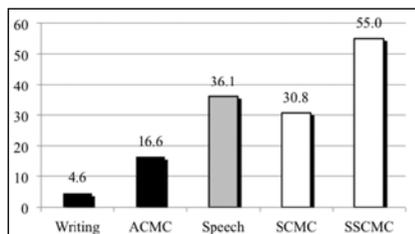
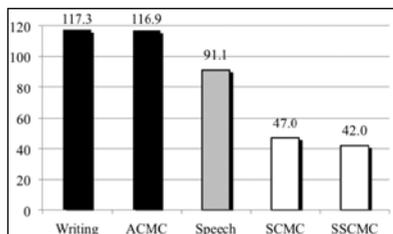


Figure 4.13: Prepositional phrases.



Chafe & Danielewicz, in their 1987 account of the properties of spoken and written language, find contractions (e.g. *it's*, *I'm*, *don't*), as well as prepositional phrases, to be distinguishing factors between speech and writing. "Spoken language commonly employs contractions" whereas "[s]uch items are rare in academic written language" (Chafe & Danielewicz 1987: 93). Finegan and Biber (1986) find contractions to be distributed as a cline: most frequently used in conversation, least frequent in academic journals, and with intermediate frequencies in broadcast, public speeches and press reportage. Biber (1988) presents very similar findings, except that he also finds official documents to be virtually void of contractions; see figure 4.12 for the average figures for speech and writing found here (based on Biber 1988, the former supplemented with the SBC subset). In the face-to-face conversations from LLC, there are 46.2 per thousand words;

in the SBC subset, there are 48.5; and in telephone conversations, as many as 54.4 per thousand words, whereas there is only one (1) contraction in the 28,000-word official documents component of LOB studied. “Contractions are the most frequently cited example of reduced surface form,” says Biber (1988: 243). Biber et al. (1999) separate the reduced surface form into verb contractions (e.g. *she’s going*) and *not*-contraction (e.g. *couldn’t go*), but find the distribution of both types to follow the same decreasing cline among their four genres: conversation > fiction > news > academic writing, in order of frequency. In the present study, contractions were found to be slightly more frequent in SSCMC than in spoken conversations, but less frequent in SCMC than in speech overall; see figure 4.12.⁸¹

Detecting contractions, as well as most other features, in conversational writing requires meticulous manual annotation. For written and spoken texts (that is: texts transcribed by linguists), automatic detection of contractions is usually possible; one simply queries the text for apostrophes and then sorts out irrelevant hits (e.g. to exclude genitive inflections). Most contractions in chat, by contrast, do not contain apostrophes, and like other chatted words, they are frequently misspelled. Chatters, governed by economy of typing, by leaving out apostrophes take the reduced form one step further. Below are various occurrences of contractions that illustrate the intricacies of annotational detection. The examples in (17), from IRC, and (18), from ICQ, also illustrate the results of chatters’ economy of typing, i.e. some ultra-reduced surface forms.

- (17) *your* (you’re), *yvw* (you’re very welcome), *their*, *there* (they’re), *where* (we’re), *dunno*, *tis*, *whatcha* (what’re you), *lets*, *lits* (let’s), *wassup*, *wasssup*, *sup* (what’s up), *whats*, *whast*, *whts*, *whxx* (what’s)

Internet relay chat (UCOW)

- (18) *their*, *there* (they’re), *were*, *where* (we’re), *ur*, *your* (you’re), *no ones* (noone’s), *souldnt* (shouldn’t), *dunno*, *dunnp*, *donno* (don’t know), *idk*, *idjk* (I don’t know), *whos*, *itz*, *lets*, *thas*, *ain’*, *shoulda* (should’ve), *can’s*, *caznt* (can’t), *whats*

Split-window ICQ chat (UCOW)

81 Contractions in Biber’s (1988: 243) terms are those on pronouns, on auxiliary forms (negation) and suffixed on nouns (except possessive forms). The following are examples of contractions accordingly left unannotated in the corpora annotated in the present study: in IRC: *wheres*, *where;s* (“where has/is”), *where*, *WR* (“where are”), *whered* (“where did”), *theres* (“there is”), *heres* (“here is”), *old’s* (“old is”), *hows* (“how is”), *kinda*, *flirt’n*, *dutch’n*; in ICQ: *pound’n*, *laugh’n*, *where* (“where are”), *theres* (there is), *kinda*, *tell’n*, *turn’n*, *pay’n*, *wait’n*, *frik’n*, *look’n*, *let’n*, *kid’n*; and in the SBC subset: *how’s*, *kinda*, *where’s*, *there’s*.

As mentioned in section 2.5, Freiermuth (2003) uses Chafe & Danielewicz's (1987) methodology to contrast linguistic features in synchronous political chat data, with spoken discussions from a political television talk show, and written (political) newspaper editorials. In his SCMC data, Freiermuth finds chatters to use contractions less frequently than speakers, but more often than writers – a finding that is corroborated in the present study, but that does not lend itself to easy explanation. What partly limits the number of contractions in IRC, compared to split-window ICQ, is the relative rarity of analytic negation in IRC (as this includes the contracted form *n't*; see discussion of analytic negation above), but compared to speech this explanation is insufficient (as IRC contains only marginally fewer instances of analytic negation than speech). A different, more likely, explanation for the rarity of contractions in IRC is presented in section 5.2.1; in the present section, we acquiesce in Freiermuth's and the concurrent finding, for the genre of SCMC, and simply observe that the conversational writing genres diverge from each other as regards frequency of contractions. As seen in the examples in (17) and (18), however, contractions in both genres show analogous composition, and both groups deviate from writing and transcribed speech in that they are occasionally realized as ultra-reduced forms. The SSCMC users employ contractions to about the same degree as speakers in conversations (55.0 per thousand words in ICQ vs. 48.5 in the SBC subset), whereas the SCMC users employ them less.

The final linguistic feature that deviates from Biber's mean for speech and writing by more than two standard deviations ($|s.d.| > 2.0$) is prepositional phrases; see figure 4.13. This feature deviates negatively for both conversational writing genres; SCMC and SSCMC both display a remarkable paucity of prepositional phrases. The frequency of prepositional phrases is nearly three times as high in writing as in conversational writing. Chafe & Danielewicz (1987, as well as Chafe 1982, 1985) find prepositional phrases, and sequences of them, to be factors that distinguish written discourse from spoken discourse, as represented by academic papers and conversations respectively. In connection with the lexical density discussion in the previous section (4.3), Chafe & Danielewicz's concept of the "intonation unit," roughly equivalent to a clause, was touched upon. Chafe & Danielewicz (1987), in their discussion of the intonation unit, expound on linguistic devices that writers, more than speakers, employ to increase the size of the unit. One of the devices is prepositional phrases (other devices are attributive adjectives, also discussed above, and e.g. nominalizations). Prepositional phrases thus typically elaborate the nominal information and expand the length of clauses. Biber (1988) postulates that prepositional phrases are "important device[s]"

for packing high amounts of information into academic nominal discourse” (1988: 237), but in his study they are also found to be frequent in other kinds of written discourse and, actually, most frequent in official documents. In the LSWE corpus, Biber et al. (1999) find prepositional phrases most common in academic prose and least common in conversation. The results of Biber’s (1988) study with regard to prepositional phrases in LOB writing and LLC speech, are reflected in figure 4.13: writing contains on average nearly 30 percent more prepositional phrases per thousand words than speech. Among the genres amalgamated into the speech bar in figure 4.13 are face-to-face conversations (LLC with 85.0, and the SBC subset with 61.1) and telephone conversations (with 71.8). Spoken American English (the SBC subset) somewhat restricts the elevation of the speech bar; yet, conversational writing contains significantly fewer prepositional phrases than the SBC subset. Apparently, very little clausal elaboration by way of prepositional phrases (or e.g. attributive adjectives and nominalizations; cf. Appendix II) takes place in conversational writing. Ko (1996) and Freiermuth (2003) both find a similar sparsity of prepositional phrases in their chat corpora, Ko making the observation that the chatted clauses “tend to be stripped down to their obligatory core, minus optional adjuncts such as prepositional phrases” (Ko 1996: no page number available).

In the lexical density discussion in the previous section, conversational writing was found to display more grammatical than lexical items. A prepositional phrase is initiated by a preposition (a grammatical item), and in written texts the phrase typically contains at least one nominal (lexical) item. Prepositional phrases, as a feature, therefore, are practically neutral in the lexical density calculation for written texts (as 1 grammatical + 1 lexical item “cancel” each other out). In spoken language and in conversational writing, however, the composition of the prepositional phrase is usually different. In these media, a typical prepositional phrase contains just a stranded preposition (grammatical) or a preposition followed by other grammatical items (such as pronouns). Prepositional phrases thus typically contribute to *lowering* the lexical density for spoken and conversational writing texts. On the other hand, prepositional phrases are extremely rare in the latter genres, as shown in figure 4.13. The effect of prepositional phrases and other elaborating devices on the mean length of written clauses, however, is palpable. It was seen in the discussion of lexical density per clause (section 4.3) that the average “intonation unit” (roughly: clause) in academic writing is 9.3 words long (Chafe & Danielewicz 1987). Table 4.5, furthermore, revealed that the average clause length in face-to-face conversations is around six words (also found by Chafe & Danielewicz 1987 for conversation), and that the average

conversational writing clause only is about four words long. Considering average clause length in conjunction with figure 4.13, consequently, we find what is partly missing in conversational writing clauses: clause-extending devices, such as prepositional phrases.

A few textual examples will shed further light on average clauses, and the effect of prepositional phrases in them. The excerpts from academic prose (19), face-to-face conversation SBC (20), IRC (21) and split-window ICQ (22), below, serve to illustrate the typical distribution of prepositional phrases in respective genres. The prepositional phrases are marked by their preposition in bold script.⁸²

- (19) It is not clear that the growth **of** the spread **between** earnings and wage rates **in** the UK over the period **of** our sample can be plausibly explained **in** cost terms. If it is argued that such a gap is automatically opened **by** the rise **in** piece-workers' earnings as productivity increases, or **by** changes **in** the amount **of** overtime worked, such changes may themselves be traced back **to** the existence **of** a high level **of** demand.

Academic prose LOB J: text 44

- (20) Jamie: Aren't you guys gonna stick up **for** me?
and beat up **on** him or something?
Miles: He's bigger **than** I am.
Pete: (laughter)
Miles: He's not bigger **than** you.
Pete: No.
Harold: But he's my –
Miles: (laughter)
Harold: he's my friend
Pete: Tha- that's right.
Miles: (laughter)
Jamie: (laughter)
Pete: You know who I'll stick up **for**
Miles: (laughter)
Pete: ... I stuck up **for** you today **at** that store.
Harold: That's true.
Jamie: ... You did.
You made me get the,

82 Biber's list of prepositions (used to detect prepositional phrases) is taken from Quirk et al. (1985: 665–667), but excludes prepositions “that have some other primary function, such as place or time adverbial, conjunct, or subordinator (e.g., *down*, *after*, *as*)” (Biber 1988: 236–237) as well as, for instance, *over*. Examples (20) and (21) contain one and two stranded prepositions, respectively, which also count as instances of feature no. 61 (stranded prepositions; see Appendix II).

Pete: Mhm,
 Jamie: um,
 Pete: that's right.
 Jamie: the green scarf.
 ... That's right.
 ... He was my fashion consultant today.

Face-to-face conversations SBC text 2

- (21) <Guest_258> wassup **with** everyone today
 <SoulSearchR> Be Back Later
 <italan> hi lily lily
 <bored> hey P where you **from**?
 <furryman> still least your still be young when they grow up.lol
 <darth> well do it again barbiegirl
 <blondii> yeah, thats how i look **at** it
 <Guest_258> oh isee
 <Lilly_Lilly> hi iatalan
 <furryman> so whens the next one.
 <blondii> the youngest is 3, so i dont know
 <carrots35ca-bbl> hb SoulSearchR
 <blondii> no more!!
 <italan> where you **from**
 <barbiegirl> cool rock
 <nane> Mart??
 <brokenwing-ange> if one day you dont see me anymore...it means i given
 up **of** my life

Internet relay chat text 2a (UCOW)

- (22) <9> who said i hooked up **with** her
 <I> if u dont wanna be **with** laurie anymore, why did u just hook up **with**
 her **on** saturday???
 <9> we were both lying there and i kissed her but i wouldnt say we hooked up
 <I> i asked her yesterday when th elast time u hooked up and she told me
 satruday. but dont tell her that im telling u this.
 <9> cause she thought katie was still awake
 <9> i dunnp

Split-window ICQ chat text 8 (UCOW)

Excerpts (19) through (22) are approximately equally long (c. 75 words), but whereas the academic prose example (19) contains 13 prepositional phrases, the sampled SBC face-to-face conversation (20) contains seven, and the conversational writing excerpts, (21) and (22), only five and four, respectively. The sloping cline for prepositional phrases across written and spoken genres, from academic prose to conversations, found by other scholars (Chafe 1982, 1985, Chafe &

Danielewicz, Biber 1988, Biber et al. 1999) thus continues its descent across conversational writing, as seen in figure 4.13.

That prepositions “serve to integrate high amounts of information into a text” (Biber 1988: 104) is distinctly shown in example (19) from academic prose. In (19), the prepositional phrases each contain at least one lexical item (*spread*, *earnings*, etc.) and the phrases extend and elaborate clauses to make the text extremely integrated. Moreover, prepositional phrases are stacked upon each other (*by changes in the amount of overtime*) in sequences, which Chafe & Danielewicz (1987) find typical of academic writing. In the other three examples, (20) through (22), however, the prepositional phrases display an entirely different distribution. Not only are the prepositions here often left stranded, which Chafe (1985: 115) cites as examples of “errors” typical in speech due to production constraints, but also the prepositional phrases contain mostly grammatical items, and therefore less clearly serve the function of elaborating clauses. Halliday (1987) calls the complexity of written language “crystalline,” “whereas the complexity of spoken language is choreographic” (1987: 66). He explains the latter thus:

The complexity of spoken language is in its flow, the dynamic mobility whereby each figure provides a context for the next one, not only defining its point of departure but also setting the conventions by reference to which it is to be interpreted. (Halliday 1987: 66–67)

Consequently, the difference between writing and speech lies not just in the presence vs. absence of prepositional phrases, or in the relations between lexical and grammatical items, but also in the usage of these items. Halliday (1987) criticizes Chafe (1982) for describing both writing and speech “using a grammar of writing” (Halliday 1987: 67). Halliday instead proposes a kind of choreographic grammar, one that recognizes the intricacy of spoken language; that “its mode of being is as process, not as product” (1987: 67). For Halliday spoken language has:

[...] a considerable degree of intricacy; when speakers exploit this potential, they seem very rarely to flounder or get lost in it. In the great majority of instances, expectations are met, dependencies resolved, and there are no loose ends. (Halliday 1987: 67)

Halliday explains that the intricacy of spoken language is of a grammatical kind; it has multiply linked clause structures. This intricacy requires the use of grammatical items, as they provide the glue that connects the parts of a spoken utterance together (Halliday 1987, Yates 1993).

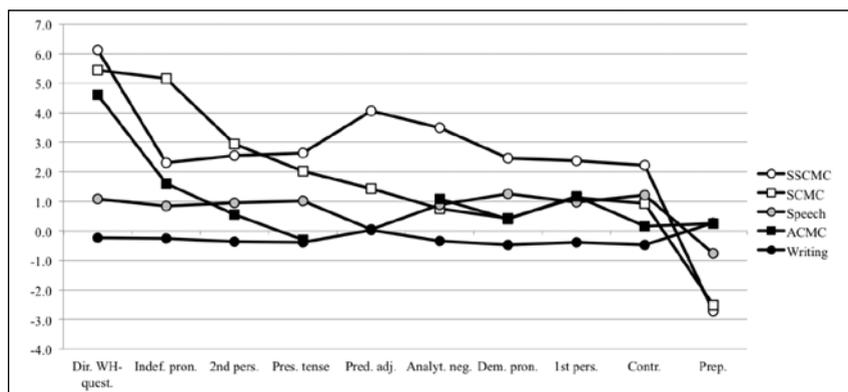
Whether we side with Chafe or Halliday is of secondary importance in the account of conversational writing here. Chafe and Halliday, of course, have both

developed their stances over the years; Halliday into his (choreographic) functional grammar (e.g. Halliday 2004) and Chafe along a more cognitive linguistic track (e.g. Chafe 1994); though both constantly in tune with natural language data. Their interpretational quibble apart, the excerpts of face-to-face conversation and conversational writing, (20) through (22), have managed to elucidate the important primary finding here: the striking similarity of the three conversational genres. Prepositional phrases are distributed in analogous ways in conversations and in both genres of conversational writing, ways that sharply distinguish these genres from the most “written” mode of writing, the genre of academic prose (19). With regard to clausal elaboration by means of prepositional phrases, other genres of writing, and speech, are intermediate between these two poles. That academic prose constitutes the “written” end of the pole, as regards prepositional phrases, was a well-established fact. The present study has extended the “spoken” end beyond conversations, finding prepositional phrases in conversational writing not just to be rare, but also possibly to serve other functions than just clausal elaboration.

In conclusion, to sum up the ten salient features of conversational writing, i.e. first and second person pronouns (described in section 4.2) and the eight features explored in the present section, we will take advantage of the standard scores calculated for each feature. Recall from chapter 3, section 3.5, that a standard score was computed for each feature, which equals the feature’s number of standard deviations from Biber’s mean for speech and writing (Appendix II table 4, from Biber 1988: 77–78). These standardized scores are ideal for enabling the comparison of features across texts and genres, and crucial for the calculation of comparable dimension scores. The present chapter has exploited the fact that the features with the highest standard deviation in conversational writing are the features that collectively epitomize the nature of conversational writing. The cut-off point for a feature’s inclusion as a salient feature in the present section was two standard deviations, which meant that a convenient number of ten features crystallized. (Modal auxiliaries, word length, TTR and the lexical density measures are included in this chapter for other justified reasons.) The ten most salient features are not necessarily the most frequent features, but the features that together distinguish English chatted texts, on average, from English written and spoken texts, on average. Naturally, the make-up of conversational writing is more complex and many-faceted than what the ten most salient features depict, and in the next chapter, therefore, all of Biber’s 67 features will be taken into account to more accurately describe the chatted material. It was decided, nevertheless, that the features that deviate from Biber’s mean by more than two

standard deviations would be of statistical interest, and that the account of them in the present chapter serves well as an introduction to the more all-round investigation of the conversational writing genres in chapter 5. Figure 4.14, finally, sums up the ten most salient features of conversational writing, or rather: those features, which, in either SCMC or SSCMC (or both), deviate from Biber's mean by more than two standard deviations, for each showing its distributions in the other three media as well. The zero point in figure 4.14, by inference, constitutes Biber's (1988) mean for speech and writing. The standard scores are based on numbers from Biber 1988: 247–263 for writing, Collot 1991: 69–70 for ACMC, Biber 1988: 264–269 and Appendix II table 3 for speech, Appendix II table 1 for SCMC and Appendix II table 2 for SSCMC, contrasted with the mean numbers and standard deviations for Biber's speech and writing overall (Appendix II table 4, from Biber 1988: 77–78); see section 3.5 for a description of the procedure of standard score calculation.

Figure 4.14: Standard score distribution of the linguistic features that, in SCMC or SSCMC, deviate by more than 2 s.d. from Biber's (1988) mean.⁸³



In the next two sections, other important features of conversational writing will be taken up – features that are characteristic of conversational writing, but not identified through Biber's (1988) methodology; firstly, the paralinguistic cues and extra-linguistic features of chat, the latter most common in IRC, and lastly, two salient linguistic features: inserts and emotives.

83 The standardized score for predicative adjectives in ACMC is unavailable (Collot 1991: 69).

4.5 Paralinguistic features and extra-linguistic content

Before the advent of computer-mediated conversational writing among the general public in the late 1980s, linguists justifiably concluded that writing is unable to incorporate all the features of speech. Halliday (1985a), for instance, points out that:

There are various aspects of spoken language that have no counterpart in writing: rhythm, intonation, degrees of loudness, variation in voice quality ('tamber'), pausing, and phrasing – as well as indexical features by which we recognise that Mary is talking and not Jane, the individual characteristics of a particular person's speech. (Halliday 1985a: 30)

The features that writing typically leaves out are what in spoken language are known as prosodic and paralinguistic features. Prosodic features are part of the linguistic system; they extend across long stretches of speech (e.g. rhythm, intonation, pausing and phrasing) as systematic phonological realizations, as in an intonation contour (Halliday 1985a: 30–31). Paralinguistic features can also extend across varying stretches of speech, but they are “not systematic – they are not part of the grammar, but rather additional variations by which the speaker signals the import of what he is saying” (1985a: 30), as by the degree of loudness, variation in voice quality (“tamber”), tempo and facial/bodily gestures. Halliday (1985a: 31) considers “prosodies” and paralinguistic to be of linguistic status, but calls a third group of features non-linguistic, “indexical.” Indexical features are not part of the language at all, but rather “properties of the individual speaker” (1988: 30), such as individual preferences for certain prosodic and paralinguistic patterns. The prosodic, paralinguistic and indexical features are difficult to represent in writing, says Halliday (1985a: 30) “because they do not belong at any particular point.” Yet, Halliday proceeds to challenge and partly dismiss the notion that these features are entirely missing from writing. Spacing and punctuation (comma, semicolon, full stop, question mark, parenthesis, etc.), he claims, are used in writing to overcome the omission of prosodic features. Spacing marks off words, and punctuation marks off grammatical units, or prosodic units, giving written text systematic variation similar to the intonation contour in speech. Nevertheless, Halliday inevitably resorts to the conclusion that “[w]ritten language never was, and never has been, conversation written down” (1985a: 41). Except for the linguistic transcription of natural spoken recordings (for linguistic research) the task of writing down speech is not what writing is about. “Why?” Halliday asks rhetorically, and answers:

—because in its core functions, writing is not anchored in the here-and-now. The particular conditions that obtain at the time of writing are not going to be present to the

reader anyway, who is usually at some distance from the writer both in time and place; so much of the message that is contained in the rhythm and tamber of speech would simply be irrelevant. (Halliday 1985a: 32)

Having made this case for writing, it is easy to see how conversational writing differs from writing: conversational writing, in its core function, *is* anchored in the here-and-now (cf. Ooi 2002). The particular conditions that obtain on the computer screen, the ideational “field” (Halliday 1985a) in the “ideational metafunction” (Halliday 2004), are present to both interlocutors at once. The text is presented to them dynamically – it happens, much like airwaves traveling through the air in speech. Linguists inquiring into SCMC therefore generally concur in describing computer chat as speech-like communication. Dresner (2005) goes so far as to propound that the visual perception of the transmitted text is analogous to auditory reception:

In a simple (i.e., single-window [...]) chatroom situation all participants sit in front of their computer screens. All of them are seeing the same thing—the text lines accumulating in front of them. As opposed to visual perception in spoken conversation, where each participant sees a completely different picture, in textual conversation vision functions somewhat like hearing in auditory discourse—it enables mutual focus on the buffer on which communication takes place. We see that the affinity between ordinary and textual chat goes beyond (or, rather, deeper) than synchronicity. The structure of mutual visual perceptual intake in computer mediated textual chat is topologically similar to its auditory counterpart. (Dresner 2005: 15–16)

This means that computer chatters, like what Halliday claims for listeners in conversations, are “predisposed to take a dynamic view of what [the text] means” (1985a: 81). Conversational writing thus turns text from “product” into “process” and writers from authors into interlocutors, that is, almost into speakers and listeners.

Interlocutors in conversational writing use a number of prosodic, paralinguistic and indexical devices, here generically called “paralinguistic features,” to enrich their writing with cues that assimilate speech, or at least to assimilate a situation similar to face-to-face interaction. Conversational writing by no means incorporates all the paralinguistic features of speech, but several of the devices employed, as we shall see, are passable attempts to bridge the gap to face-to-face spoken discourse. In conversational writing, the paralinguistic cues are applied to written text, and not spoken, and therefore differ somewhat from Halliday’s definition. Paralanguage, in this section, is used as a broad term covering several salient aspects of conversational writing that Biber’s (1988) features fail to include, aspects ranging from nicknames, personalization tropes and self-imposed

spoken language transcription, to abbreviations, graphic devices, “leet,”⁸⁴ interlanguage and code-switching. The paralinguistic of conversational writing is realized in the messages of the communication. The paralinguistic devices therefore also provide clues to the role language plays in online communication, the semiotic mode of conversational writing, i.e. what the language is being used to achieve, as regards, for instance, conscious self-representation (Halliday 1985a, Halliday & Hasan 1989), reflecting the textual metafunction of language (Halliday 2004). Besides paralinguistic, this section will also cover extra-linguistic factors in the communication that are not always realized in the user-generated messages, such as pictures and music shared among the chatters. Extra-linguistic factors form important parts of interlocutors’ shared time and space in conversational writing (their field), influencing their communication. The survey of paralinguistic devices and extra-linguistic factors is essentially brought in here to complement the comprehensive linguistic analysis to be undertaken in chapter 5.

Paralinguistic features, innovative orthography and neologisms in textual computer and cellphone communication have been pet areas for linguistic researchers over the past few decades as witnessed by a host of publications dealing with these (inter alia Wilkins 1991, Yates & Orlikowski 1993, Werry 1996, Jonsson 1998, Schulze 1999, Crystal 2001, 2008a, Gajadhar & Green 2003, Baron 2008, Waldner 2009, Rowe 2011, to name but a few). As mentioned, the primary concern of the present study is to apply Biber’s (1988) methodology, with its 67 linguistic features, to the conversational writing data, no feature of which covers paralinguistic and extra-linguistic factors. The chatted texts, as described in chapter 3, were annotated for Biber’s list of linguistic features after the texts had been purged from bracketed nickname turn indicators, server-generated messages, action commands and certain other strings of text (e.g. graphic noise and mass-advertising dumped into the IRC channels) that were impossible to tag and/or apt to skew the results (for examples of excluded material, see Appendix IV). This purging was kept to an absolute minimum, as it was of utmost importance that texts remain as intact as possible, and that all user-generated, i.e. keyed-in, linguistic messages, in English, were retained. The present section, however, is devoted to bringing some of the excluded material temporarily back into the account.

84 Leet, “leet speak” or “1337 5p34k” denotes the language of “elite” chatters, such as online gamers and hackers, who e.g. incorporate symbols and numbers as substitutes for letters in words. It is partly used as a means for experienced users to demarcate themselves from “newbs” or “n00bs” (those new to the medium) (see e.g. Van de Velde & Meuleman 2004, Blashki & Nichol 2005, Nichol & Blashki 2006). LeBlanc defines leet or “l33t” as “elite geek speech” (LeBlanc 2005: 72).

The first paralinguistic device employed by chatters, in both IRC and ICQ, is the choice of a nickname, decided upon before logging in. IRC nicknames (nicks) are usually easily changed, whereas ICQ nicknames (more like user-IDs) are connected to an account. (The ICQ nicknames in the present study, however, were not chosen by participants, but pre-set on lab computers by the present researcher, for practical reasons.) As seen in the IRC text samples in the present chapter, chatters make a conscious choice of nicknames; examples are big-dog, River, Chaser, }melons{{, Sweet_Victoria, Cheeky1, BillClinton and blondii. “The nick is their electronic identity,” says Crystal (2001: 160); “it says something about who they are, and acts as an invitation to others to talk to them” (ibid.). Anglemark (2009: 89) notes that “[t]he nick is often the only identity a chat room participant displays in a chat session.” Indeed, quite frequently, IRC chatters “lurk” in the channel, “eavesdropping” without contributing to the ongoing communication. Occasionally, chatters signal their presence with an empty turn, displaying only their nickname, as <remut> in (23).

- (23) <Heart35> some using Mark
 <remut>
 <dony> c\free
 <biro> hi nuttygrl :o)
 <remut>
 <bergs> ..it is Brad!

Internet relay chat (UCOW)⁸⁵

Other chatters put their nicks to creative use in combination with their turn, as in the example of attempted flooding (dumping repeated jabberwock) in (24), and with the graphic feature in (25).

- (24) <Can[You]Handle[This]> %%f8738200%%
 <Can[You]Handle[This]> %%f8738200%%
 ...
 <Can[You]Handle[This]> %%f3738003bf9c129fec%%
 *** Can[You]Handle[This] was kicked by Sheila (flood)

Internet relay chat (UCOW)

- (25) <dj_19_m_uk> <===== any girls with pic message me!!!

Internet relay chat (UCOW)

85 Unnumbered Internet relay chat texts in examples (23) through (45) are from the part of the corpus that exceeds the texts sampled for annotation; see description of corpus creation in chapter 3.

The chatter's nickname is indicated within angle brackets, by the software, in the chatter's every turn. These bracketed nickname turn indicators are not part of the annotated IRC corpus, yet it must be recognized that in the ongoing communication they have a certain discourse value. Crystal (2001) points out that "they provide a crucial means of maintaining semantic threads in what is otherwise a potentially incoherent situation" (2001: 161). Moreover, the nicknames that are used as address terms in messages provide invaluable links in the conversational threads. Crystal considers the function of these links "analogous to the role of gaze and body movement in face-to-face conversation involving several people" (2001: 162).

The second paralinguistic device available to chatters in ICQ and web-based chats (not in IRC) is the choice of font style, font color and font size. The chatters in the split-window ICQ corpus employ this device diligently, with constant changes, to personalize their messages in a way comparable to the vocal variation of intonation found in speech. The changes in font style, color and size are retained in the corpus, though not reproduced in textual examples given here. Other personalization schemes are exemplified in (26) and (27), whereby individuals attract attention in the flow of IRC turns.

(26) <^mekrisi^> hi guys does any one wanna
chat ???
Internet relay chat (UCOW)

(27) <}}melons{{> \V/elcome Back angeldelight
Internet relay chat text 1b (UCOW)

Chatters typically mark their entrance into the chat room/channel/program/site by a greeting, e.g. *Hello All*, *hii all*, *hey room imback*, in which the first element is an interjection. (Interjections are particularly pervasive in the IRC texts, and all interjections were tagged in the corpora in the present study, but as they are not among Biber's (1988) list of features, they will be treated separately, among "inserts" in section 4.6.) Greetings, like other turns, can be personalized; see the two alternative enthusiastic responses to electrolite's modest general greeting in (28), in which BK is trying to attract electrolite's attention.

(28) <electrolite> hi all
<BK> _,-*^'^*-,_,-*' electrolite _,-*^'^*-,_,-*'
<BK> □□□□□□□□ Hello electrolite □□□□□□□□
Internet relay chat text 4b (UCOW)

The keystrokes in BK's turns in (28) are combined into iconographic effects, making up sets of decorative strings. While the IRC interface used by the chatters in

UCOW has no readily available supply of graphic icons, the ICQ program (and web-based chats) provide users with a choice of graphic emoticons, e.g. ☺. Moreover, as mentioned in section 3.3, ICQ has a supply of graphic action tropes for users to employ ad hoc by a simple click. A graphic action trope is realized in the text as e.g. “B picks a flower and hands it to you”; see Appendix IV. The chatters in the split-window ICQ corpus used these readily available graphic devices to a moderate extent. However, as the inclusion of a graphic icon or an action trope implies no conscious linguistic typing on the part of the chatter, neither device was retained in the purged material for annotation. For consistency then, in IRC, action commands were also purged away before the annotation of Biber’s (1988) features (see section 3.2 for a description of the purging process and Appendix IV for examples). Both the IRC and the split-window ICQ chatters, however, used textual emoticons (e.g. :, ;) which were preserved in the texts and tagged as emotives – and therefore to be treated separately, along with inserts, in section 4.6.

In their messages, chatters employ a vast number of paralinguistic devices to assimilate spoken interaction, i.e. to transcribe their own texts as if into speech. Enthusiasm, surprise, anger, or mere emphasis, is signaled through repeated exclamation marks (*Sweetie!!!!!!!, whooops!!!!, NOOOOOO!!!!, i know!!!!!!!!!!!!!!!*) and puzzlement through repeated question marks (*uhhhh.....?????, when???????*). Punctuation is also used to signal pauses (*for sure chanel...can't match up to ours huh...lol*). Capital letters mark off text expressed in a loud voice, sometimes as if it was screamed (*i'm very ANGRY!, Well i DO like those skateboarders... especially MATT!, THAT IS SOOOO MEAN, CAPS ARENT COOL THEY WILL GET U KICKED*). Repeated letters denote added emphasis, e.g. *this suxxxxxxxxxxxxxxxx* or, for instance, long vowel sounds (*oooooooooooooooooooo u didnt say that before well then thats a whole different ball game, YOU'VE GOT WORMS EWEWWWWWWWWWW*). As seen in the latter, the two devices can be combined for increased effect (capital, repeated letters). Capital letters are fairly common in split-window ICQ, but very rare in IRC – their use in IRC is regarded as screaming, for which the channel operator may “kick” the user from the channel. Two passages from the conversational writing texts serve particularly well to highlight interlocutors’ sense of being on the verge of an auditory medium (as suggested by Dresner 2005): example (29) from IRC and (30) from ICQ.

- | | |
|------|--|
| (29) | <p>< mad_max > missing my voice
 < mad_max > to scream a little ...
 <Cheeky1> hahaha
 < mad_max > screaaaaaaaaaaaaammmmmmm!!!!!!!!!!!!!!!!!!!!!!!</p> |
|------|--|
- Internet relay chat text 3a (UCOW)

- (30) <Pilot1> yo, did you read that capian underpants bok
 <Pilot1> dude i'm not reading when i'm typing. i,m outof practice, i haven't typed any school paper's or e-mails in a while. yeah ne way....
 <esoteric> hi dude you can't spell. dude why are your eyes brown? you are boring to talk to so i have to get someone else to type. you are slow. yes. yes YES!!! dang you are slow. just use 2 fingers neither am i i am looking at the keyboard. duh! oh yeah whatever shut up yeah yeah i'm not listening..... la la la la la la la la ooooh saaaaay can you seeeeeeee!!!!!!!!!

Split-window ICQ chat text 12 (UCOW)

Example (30) ends in the transcribed equivalent of the user esoteric singing the first stanza of the US national anthem. It is part of a turn in which the user is trying to “make his voice heard” over the conversational partner in the split ICQ window (as if they were speakers in the auditory medium). The users were new to SSCMC and were at once intrigued and annoyed by the supersynchronicity, which entailed that most of esoteric’s turn in (30) was overlapped by the other interlocutor, Pilot1’s, turns. That there is overlapping “speech” in (30) illustrates the similarity between split-window ICQ and face-to-face conversations. On the other hand, the supersynchronous mediation of text in split-window ICQ goes beyond speech, in that it does not require interlocutors to “stop and listen” at the same points as in the auditory medium. Experienced interlocutors in split-window ICQ can carry on with their interaction and simultaneously listen (read) and speak (write), only pausing in strategic moments to maintain a certain consecutiveness in the communication. If this were done in the auditory medium (in long completely overlapping passages), the communication would be rendered incomprehensible. Supersynchronous CMC thus not only resembles auditory conversation; it surpasses it.

IRC surpasses auditory conversation by another type of simultaneous “speech,” in that a vast number of online chatters can engage in a conversation at once. Chat channels “provide virtually unlimited access to people who want to chat on a particular channel in a moment in time” (Freiermuth 2003: 31). However, chat channels rarely contain only one conversation; rather, several conversational threads are interlaced, requiring untangling skills from users in order for threads to be followed. Elsner & Charniak (2008, 2010) find an average of 2.75 conversations active at a time in their IRC corpus. If IRC were an auditory situation, it would be a cocktail party (Crystal 2001). Dresner (2005) notes for the auditory situation that a person can “admittedly catch his name in a conversation going on in another part of the room, but the rule is that we do not, and cannot, follow more than one conversation line for a substantial period of time” (2005: 20). IRC

chatters, on the other hand, are “continually perceptually aware of more than one conversation line” (2005: 21). Dresner goes on to explain that it is the “visual spatiality” of the synchronous texts that enables chatters to untangle conversations; “(p)ictorial processing abilities seem to help us sort out the entanglement of conversation lines” (2005: 21). Following Dresner’s reasoning, then, computer chat approximates auditory face-to-face interaction; yet, it is only through the visual medium that simultaneous speech, as in split-window ICQ, and simultaneous threads, as in IRC, can be perceived. In either format, to be sure, chatters must be apt typers to keep up with the simultaneous reception and production of text.

IRC chatters, possibly more than split-window ICQ chatters, are concerned about keeping pace with the conversations at hand. Certainly, IRC chatters have slightly more processing time than speakers, but in order to stay abreast of the unfolding conversation, they must construct text quickly. When typing in on-line chat, “it becomes imperative to use precious construction time efficiently” (Freiermuth 2003: 171). Werry (1996) points out that:

The language produced by users of IRC demands to be read with the simultaneous involvement of the ear and eye. One can discern an intensified engagement with the sounds of language, with the auditory and iconographic potential of words. (Werry 1996: 59)

This “intensified engagement with the sounds of the language,” with the auditory potential of words, brings chatters to impose spoken language transcription schemes upon their discourse, such as those discussed among the paralinguistic features above. The “iconographic potential of words” (Werry 1996: 59) is further explored below.

Earlier in this chapter, we observed the short clause length of conversational writing, which indicates brief turns. The brief turns, moreover, consist of very short words. Occasionally, the short words constitute abbreviations (initialisms such as *idk* (I don’t know), *brb* (be right back), *lol* (laughing out loud), *lmao* (laughing my ass off), *a/s/l?* (age/sex/location?)), which chatters employ to speed up typing, but which really represent several longer words in themselves. The answer to the latter initialism (*a/s/l?*), for instance, might be almost as brief as the question, and yet, impart a great deal of information (e.g. *31/blk m/usa tx, 20/m/syd*). (As mentioned in section 3.2, abbreviations were retained in the present study and annotated for their constituent linguistic items; *idk*, for instance, was tagged with Biber’s (1988) features nos. 3, 6, 56, 59 and 67.) Werry (1996) observes a general tendency for IRC words to be stripped down to “the fewest possible letters that will enable them to be meaningfully recognized” (1996: 55). The same tendency is observed in both the IRC and the split-window ICQ

beyond the turn, as in (32), however, were removed before the annotation (this particular instance was found to occur five times, but it was the only graphic feature to extend beyond one turn in the corpus annotated).

The examples of conversational writing in this section illustrate that chatters are masters of their keyboard. They exploit its every key to enliven their textual interaction, rendering turns spoken-like to the point of their being sung, and graphic to the point of their being art. Just like in spoken discourse, there are slips of the tongue in computer chat, or rather, slips of the key. A famous slip of the key from 21st-century CMC, more specifically from computer gaming lingo, is found in the verb *pwn*, meaning “own.” In “elite” computer-mediated chat lingo (so-called “leet,” “l33t” or “l337”), used by e.g. gamers and hackers, *pwn-ing* stands for “owning” (Pichlmair 2010). A computer gamer taking over an enemy base, or a hacker taking over a server, would say that they pwn it. A slip of the key thus perpetuated in this sub-language and eventually became a symbol of how leet-speakers, advanced chatters, “pwn” the English language (Pichlmair 2010). Moreover, at the beginning of the 21st century, *pwn* (pronounced /pəʊn/ in British, /poʊn/ in U.S. English) and other leet terms (e.g. *noob*, meaning “newbie,” i.e. inexperienced users; *lol*; *irl*, as in “meeting in RL” and *leet* as a term in itself) passed over into spoken language, mostly among adolescents (Bennett 2007). The persistence of these terms in the spoken medium, of course, remains to be proved. If *pwn* is given the persistence of the term *qwerty*, which denotes a standard for keyboards introduced in the 19th century, also derived from the adjacency of keys, the term *pwn* is likely to stay in the language outside of CMC for some time. Unfortunately, no matter how intriguing the subject matter, a more thorough analysis of the lexis of leet is beyond the scope of this study; instead, interested readers are referred to e.g. Van de Velde & Meuleman (2004), Blashki & Nichol (2005) and LeBlanc (2005).

The lexis of the conversational writing corpora in the present study, in both IRC and ICQ, is English (in which leet is reflected, of course). English is the only language allowed in the recorded chat channels, and it was the only language allowed in the recording of the split-window ICQ chats. In IRC, language rules are often displayed automatically upon the user’s entrance and channel operators are particularly quick to enforce them. Nevertheless, users in IRC are globally dispersed, and English is not the native language of all of them, which means that a few instances of interlanguage, code-switching and non-English fonts inevitably surface in the IRC corpus, as in examples (34) through (36).

- (34) <DJ-XNS|Vs|DJ_RMX> halooow are ther some one ho will talk
with a swedich boy?
Internet relay chat text 5a (UCOW)

(35) <CLAUDIAA> si somos muchos lo que hablamos espanolporque/
Internet relay chat (UCOW)

(36) < mouad__> □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
...
*** mouad__ was kicked by AussieDino (You have been kicked
for using non-english fonts. Please speak in English next time.)
Internet relay chat (UCOW)

Naturally, the interlanguage in (34) causes no action from the channel operator. The user in (35) is politely reminded of the language rule, whereas the channel operator's enforcement upon the use of non-English font in (36) is severe (and included here as an illustration of channel operator interference). IRC communication is ASCII-based, and therefore unable to render non-English fonts correctly. (The non-English orthography in (36), consequently, was not correctly represented in the log, either.) As partly touched upon in section 3.2, instances such as (34) through (36) were treated in the following way in the present study: all interlanguage, such as in (34), was retained and annotated, whereas all foreign language turns, such as those in (35) and (36), were removed, as well as all foreign language items within English turns (extremely few). The same procedure applied to the ICQ and SBC texts, whereby a few foreign turns and words were removed.

The omniscient presence of a vigilant channel operator is perceptible to all users in IRC. The operator's and interlocutors' nicknames are displayed as a list in the software, which is constantly updated upon users joining and leaving. The chat channel is thus the virtual equivalent of a room, in which people mingle, chat and act upon each other's actions. In sharp contrast to rooms in real life, however, the chat channel is textual – the mingling, chatting and acting is carried out via written characters. Or rather, they are carried out in written characters for the most part. A few extra-linguistic factors make their way into the communication. These will be tended to shortly.

The linguistic environment – the co-text, or context – of any linguistic item is crucial for the interpretation of the item. Halliday & Hasan (1976, 1989) call text-internal reference “endophoric” (co-textual) and situational reference “exophoric” (contextual). Reference is one of the cohesive devices that enact the textual metafunction in language, reflecting the semiotic mode of the interaction. Endophoric reference makes a text cohere within itself and exophoric makes it cohere with the context of situation (Halliday & Hasan 1976). Endophoric reference is realized, for instance, in the use of personal and demonstrative pronouns (referring to antecedents) and other text-internal deictic devices. Another cohesive

device is ellipsis – the omission of words that are recoverable in an earlier passage of text. Ellipsis is frequent in the conversational writing corpora, for instance in answers to questions; see B's answers to 2 in (37).

- (37) <2> do u know were ur going to college yet?
 umm i dont know
<2> were do u want to go
 umm 7 places

Split-window ICQ chat text 2 (UCOW)

Endophoric reference and ellipsis are thus co-textual (Halliday & Hasan 1989), i.e. inferable from the surrounding (here: preceding) text. Example (37) illustrates what typically happens in spoken interaction, as well as in conversational writing; Hughes (1996) postulates that there is far more ellipsis in speech than in writing as “speakers can assume that listeners will ‘fill in’ the gaps from their shared knowledge” (1996: 155). Exophoric reference, by contrast, typically depends on the speaker “pointing to” something in the text-external, situational context, explicitly or implicitly (as in a nod or a gaze), “for example, ‘she’s nice’ said with a nod towards a person in the vicinity” (Hughes 1996: 155). Exophoric reference, per definition then, is more common in speech than in writing (cf. Coleman 1996: 43). In writing, exophoric reference, expressed in for instance dialogs, needs to be explicitly explained by the author in order for the reader to understand what is referred to. In face-to-face conversations, the extra-linguistic content is evident in the surrounding environment in which speakers are situated. The extra-linguistic situation is thus often brought to bear on typical spoken discourse, or, rather, speech typically depends on the extra-linguistic, contextual situation.

Conversational writing is carried out in a contextual situation distinct from both speech and writing. The interlocutors’ physical surroundings may be vastly, that is globally, separated, but the communication takes place in a shared, virtual space on the interlocutors’ computer screens. This virtual space (the chat window itself, or an adjacent window) can carry shared extra-linguistic information, or content that affects both interlocutors at once. The IRC protocol preceded the hypertext protocol (the World Wide Web), and was in its earliest form a mere textual affair (although occasionally complemented with file transfer via protocols like ftp and gopher). Over the years, however, IRC users increasingly complemented their communication with information shared via other protocols (the direct client-to-client protocol, for instance). The IRC chat corpus in the present study gives proof of, or suggests, a few instances in which extra-linguistic content is being shared. Posting web addresses into the public

IRC channel is rarely tolerated, and the interlocutors in the IRC corpus are not found to discuss web content. The direct client-to-client contacts and the private chats, however, may involve the sharing of web sites from which, for instance, scripts can be obtained. Scripts obtained for free may surreptitiously program a user's leave-message to display an advertisement of the free script site, as in (38). The high frequency of leave-messages like (38) in the corpus therefore suggests that users engage in the sharing of scripts, which in the direct client-to-client protocol or private chat (outside of the public channel) most likely yields instances of exophoric reference. The chatter in (39) is using a script that automatically detects what music is being played on the user's computer and displays this as an action in the public channel, an action which may lead to the user being asked to share the file. In example (40) a brief sound is played into the channel (audible only in a few chat clients, i.e. IRC "programs"), and in example (41) a chatter initiates a trivia game to be played with fellow chatters in the public channel.

- (38) *** Tina^^B has quit IRC (»;« Scoop Script 2001 »!« The best script ever seen! Get yours copy at www.scoopsite.com)
Internet relay chat (UCOW)
- (39) * I_C_Triple is now playing: Artist: Tukan | Title: Light A Rainbow [CJ Stone Rmx] | Genre: Trance | Year: 2001 | Comment: <http://mphase.6x.to> | Quality: 160kbps 44kHz | Position: 1:31 | Length: 8:02
Internet relay chat (UCOW)
- (40) [SittingBull SOUND]
Internet relay chat (UCOW)
- (41) <sOLDierZ____> 04Starting the trivia. Round of 035
04questions. 03!strivia 04to stop. Total: 037841
Internet relay chat (UCOW)

Text such as that in (38) through (41) was not retained in the annotated corpus; (38) is a server-generated join- and quit-message; (39) and (40) are action commands, and the turn in (41) is not consciously keyed in as a linguistic message by the user, but rather produced through strict programming. Nevertheless, the examples provide clues to what extra-linguistic content might be shared in pending private chat windows or via the client-to-client protocol. Shared music files are more common than parlor games in the IRC corpus (the game in (41) is the only instance). On the whole, the sharing of extra-linguistic content leaves remarkably few imprints on the discourse in the public channel. Example (42) is a rather amusing exception, in which |mad_max| hums a song being played

(recites its lyrics) and eventually asks Brutal_Beauty for a dance, and in (43) a chatter expresses his/her enthusiasm over another song played. Overall, the most commonly shared extra-linguistic content seems to be photographs. In example (44) two chatters share photos via file transfer and discuss these, and (45) exemplifies another turn with exophoric reference to a shared photo.

- (42) <|mad_max|> looking back over my shoulder
 <Tha-Kappo-tan> hey what up people
 <|mad_max|> i can se e that look in ur eyees
 <|mad_max|> eyes*
 <Brutal_Beauty> Tha-Kappo-tan, Nothings up here. :)
 <Tha-Kappo-tan> any people from the land of Oz msg me
 <|mad_max|> hey, bartender gimme some more of that!!!
 <Brutal_Beauty> |mad_max| :S
 <|mad_max|> wow
 <|mad_max|> hi, beauty
 <|mad_max|> u wanna dance?
 Internet relay chat text 3a (UCOW)
- (43) <yazzie^> !BK I-Will-Survive.wav
 WooHoooo!!!...like taking candy from a baby!!!
 ...
 <yazzie^> can you send that song to me plz BK
 Internet relay chat text 4b (UCOW)
- (44) <Genie500> oh river just a sec I gotta turn something off for
 you to send okay
 <River> this one is from 95 without the glasses .
 <River> ok
 ...
 <Genie500> okay try again
 ...
 <River> but the hair is almost the same now as then
 <River> plus a wee bit more grey in it
 <Genie500> Laughing Out Loud ok
 Internet relay chat text 4a (UCOW)
- (45) <SittingBull> [Bahamut] i need to send a newer pic that one
 was in england and from 2 years ago
 Internet relay chat (UCOW)

Whereas there is a relative paucity of exophoric reference to shared audible and physical (i.e. virtual) extra-linguistic devices in the public channels, more subtle kinds of exophoric reference permeate throughout. It is evident in the corpus, for instance, that the IRC chatters experience their software window, and the textual

flow, as a confined, shared space, much like a room in real life. Spatial pro-forms and other exophoric references to the room abound (*here, where, back, on*, cf. Quirk et al. 1985: 514ff); see the various turns in (46). Chatters *look for* people in different rooms, *see* each other in rooms, or refer to other, private, rooms, as in the various turns in (47), and they refer deictically to both the room, and the ongoing interaction, (as *this*) in (48).

- (46) a. Anybody *here*???
- b. I'll be *here* for a while...
- c. *where* shes not answering me
- d. *where* have yu been
- e. i will be *back*
- f. \\elcome *Back* ^xelle^
- g. matt is *on*...lmao

Internet relay chat (UCOW)

- (47) a. h0rnyale you just *missed* ann...she was *lookin for* ya
- b. *looking for* saba
- c. Hi MOM_OF_3_BRATZ! I'm just so happy to *see* you today! :)
- d. *see* ya barbie
- e. she is *in my room* hcmk

Internet relay chat (UCOW)

- (48) a. well now *this* is fun isnt it
- b. just getting use to *this*
- c. is *this* slow tonite or what?

Internet relay chat (UCOW)

Moreover, chatters refer exophorically to the shared time in the room (*while ive been away, 2night, tonight, later*), as in (49), incidentally ignoring that, in their global dispersion, a time adjunct like *tonight* may be perceived differently in a different time zone.

- (49) a. ah been talking *while ive been away* have you ?
- b. no ops here *2night*
- c. hm not much talking in here *tonight*
- d. u r really a big help *2night*
- e. tks see you *later* ulsterman
- f. hey i'll talk to ya all *later* i need to jet for a *lil while*

Internet relay chat (UCOW)

Besides spatial and temporal adverbials, Halliday & Hasan (1976) and Halliday (2004) also consider, inter alia, the definite article and personal pronouns to be carriers of exophoric reference; "the definite article is the item that, in English, carries the meaning of specific identity or 'definiteness' in its pure form" (1976:

32) and this definiteness can sometimes be achieved only through an examination of the situational context. The first and second person pronouns “do not normally refer to the text at all” but rather are “normally interpreted exophorically” (1976: 48), whereas the third person essentially refers to the text, but also “may refer exophorically to some person or thing that is present in the context of situation” (1976: 49). In (50), two chatters are exchanging files and experiencing trouble opening the files because of an unknown file format. From their use of the definite article (in *the first one, the extention*), and the subsequent pronoun *it (di it open)*, it is evident that both chatters from their situational context can infer which file and which extension are referred to. All the while, their exophoric reference is obscure to other chatters, who do not have access to the same extra-linguistic material. In (51), the definite article (in *the server*) signals shared common knowledge among all chatters on the same server, but to an outsider, reading this log, it is not evident which server is referred to. Thus, extra-linguistic information plays an important role in both cases.

- (50) <River> oops Genie500, *the first one* you may not be able to
 open, forgot to look at *the extention*.
 <Genie500> back
 <big-dog> `WeLCoMe BaCK.Genie500`WeLCoMe BaCK.
 <River> wb Genie500
 <River> di *it* open for you ?
 <Genie500> Thank You River big-dog
 <Genie500> not yet I froze up when I tried

Internet relay chat text 4a (UCOW)

- (51) <River> looks like big troubles on *the server* today

Internet relay chat text 4a (UCOW)

The various turns in (52), finally, exemplify exophoric reference whereby chatters refer to other chatters in the room, almost as if they were nodding or gazing at the intended referent. A plural second person pronoun (*u girls, you 2, u*) is used to address two participants, or the *ladies* identified in the room. Third person pronouns (*he, she*) refer to foregoing speakers, and it is evident to all chatters that pronoun *they* (in *theyd boot you*) refers to the rigid channel operator. In neither case does the pronoun refer to an explicitly stated, anaphoric, referent, but rather to persons simply identified as present in the room, inferred from the extra-linguistic context (for instance, from the list of logged-in participants). The first person plural *us* (in *let's sing him*) is also clearly exophoric, including all chatters as referents and a foregoing speaker (*him*) as the recipient of the intended action.

- (52) a. *u girls* are from the uk right
 b. hey *you 2* gonna quit fighting and talk to me or what?
 c. hello ladies *any of u* care too chat with me
 d. chanel *he's* here...lmao
 e. *he's* not tooking with you
 f. *she* is here hcmk23
 g. didnt know *theyd* boot you for saying s\$cks
 h. let's sing *him*

Internet relay chat (UCOW)

To sum up, the *explicit* extra-linguistic content shared in connection with the IRC communication (e.g. music, pictures, a game) is found to leave remarkably few traces in the discourse, whereas the *implicit* extra-linguistic content (the shared space, the shared time, the turns themselves, and the people apparent in the room) gives rise to prevalent exophoric reference. Naturally, defining the latter extra-linguistic content as exophoric is an intricate matter, as the content is indeed reflected in users' messages (as if endophoric) – nevertheless, the reference to it is contextual, not co-textual, as examples (46) through (52) have shown. As mentioned, Halliday & Hasan (1989) defined endophoric reference as co-textual, referring to the surrounding text, and exophoric reference as contextual, referring to the shared situation. In conversational writing, the shared situation is largely made up of text, and yet, this mass of text and the shared window, together, make up an extra-linguistic environment, a room, in which people interact.

The present section has explored the paralinguistic features of conversational writing, finding the account of them to elucidate the semiotic mode of conversational writing, the “particular part that the language is playing in the interactive process” (Halliday & Hasan 1989: 24). Chatters' nicknames are conscious choices for self-representation, and chatters' personalization tropes and self-imposed spoken language transcriptions all tinge their turns, just like their abbreviations, graphic devices and instances of “leet,” interlanguage and code-switching. Most turns in the chat carry a clue to the identity of their producer, regardless of whether chatters consciously exploit their major means at hand, the keyboard, to construe the identity or not. A majority of the section has tended to the circumstances of IRC, but naturally, several of the features equally apply to the medium of split-window ICQ. In split-window ICQ, however, the virtual room is usually shared by only two participants, who know each other outside of the medium, which means that more intense chatting goes on, and less action, joining, leaving, and conscious self-representation. Moreover, the ICQ chatters in the present study were instructed not to leave their chat window, and were therefore unable to share extra-linguistic content, such as music, graphics, or web sites.

The split-window ICQ corpus has fewer references to the shared space and time in the chat, but more to the shared real-life environment (*how bout we bounce outta here, you shoudl come down, yesterday, satruday, last night, last weekend, next year*). In both IRC and split-window ICQ, exophoric reference is made to the shared contextual situation, but whereas the IRC chatters share only the virtual room, the split-window ICQ chatters share both the virtual and the real life “room” (cf. section 3.3), and this is reflected in their chats.

In the next section, two salient *linguistic* features of conversational writing are discussed: inserts and emotives. They are not found among Biber’s (1988) list of linguistic features, but emerged in the annotation process as decidedly characteristic of chatted texts.

4.6 Inserts and emotives

Neither of Biber’s (1988, 2006) two major multidimensional analyses of the English language considers the use of interjections, or “inserts” overall, in the spoken and written genres studied. Yet, linguistic intuition suggests that inserts are one of the most immediate discriminating markers of spoken discourse, apt to be an influential factor in any analysis distinguishing among spoken and written registers. At an early stage, therefore, it was decided that the corpora annotated in the present study should be tagged for their inserts. In the annotation of the IRC corpus (SCMC), moreover, it soon became evident that without this feature, nearly every tenth word would have been left untagged (typically greetings). Biber et al. (1999) describe “inserts” as a class of words typically found in conversations, recognizing that “[i]f we are to describe spoken language adequately, we need to pay more attention to them than has traditionally been done” (1999: 56). Accordingly, Biber et al. (1999) devote a subsection of the chapter entitled “The grammar of conversation” to inserts, grouping them into nine major functional types: interjections (e.g. *oh, ah, wow*), greetings and farewells (e.g. *hello, bye*), discourse markers (e.g. *well, right, now*), attention signals (e.g. *hey, yo*), response elicitors (e.g. *right?, huh?*), response forms (e.g. *uh huh, mhm*), hesitators (e.g. *uh, erm*), polite-speech formulae (e.g. *thanks, sorry*) and expletives (e.g. *shit, geez*) (1999: 1082–1099).

The annotation of inserts in the corpora in the present study, UCOW and the SBC subset, proceeded in three steps. First, all occurrences of interjections were manually annotated (i.e. those classified as interjections in OED). This annotation ran parallel with the annotation of Biber’s (1988) 67 linguistic features and

was essentially done in an effort to assign a tag to every token.⁸⁶ Without a tag for interjections, approximately every twentieth to every tenth word would have been left unannotated in the texts (e.g. *oh, wow, hi, hello, hey, yah, no, uh, um*), even if certain interjections also received a tag, or two, from Biber's features (e.g. *well*, tagged as both adverb and discourse particle). After the annotation of Biber's (1988) features, and interjections, was complete, the second step was taken. In the second step, Biber et al.'s (1999) definition of inserts was used, which meant that approximately ten percent additional occurrences, in each corpus, were found to belong to the category, all words that rightfully had been assigned Biber tags (and that, naturally, also keep those, e.g. *well*). In the third and final step, all interjections were renamed "inserts" and the total occurrences were summed up. The number of inserts per thousand words in the three corpora is shown in figure 4.15 (based on table 4.7). No equivalent annotation of inserts was carried out for writing, or speech overall. Unlike previous diagrams, the speech bar in figure 4.15 thus represents face-to-face conversations from the SBC subset only.⁸⁷

Table 4.7: Frequencies of inserts

	Writing	ACMC	f-t-fSBC	SCMC	SSCMC
inserts	n.a.	n.a.	60.0	97.3	66.6

Table 4.8: Frequencies of emotives

	Writing	ACMC	Speech	SCMC	SSCMC
emotives	0.0	n.a.	0.0	23.4	3.8

Figure 4.15: Inserts (normalized freq.).

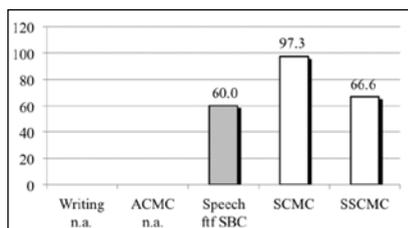
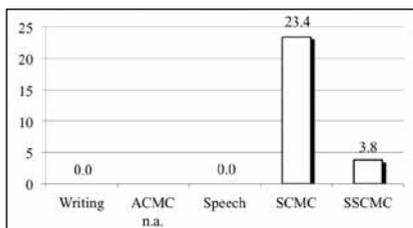


Figure 4.16: Emotives (normalized freq.).



The annotation of "emotives" (a new linguistic feature, introduced in the present study; see section 1.5) was also begun alongside the annotation of Biber's (1988) features, but completed after the annotation of all inserts. The new linguistic

86 Several tokens (such as abbreviations and contracted words), of course, were assigned two or more tags.

87 The results of statistical tests of the frequencies in the relevant media are found in Appendix VI. In tables 4.7 and 4.8 and figures 4.15 and 4.16, "n.a." means that the figure is "not available," as the texts were not annotated for the feature.

feature assigned tags to a few more tokens otherwise ignored, ultimately rendering practically all tokens bestowed with tags. Emotives are items typically found in conversational writing whereby chatters add an emotional zest to their utterances, e.g. :), ;), :(, :P, lol, rofl, lmao (partly taken up as emoticons or smileys in the previous literature; see e.g. Werry 1996, Jonsson 1998, Schulze 1999, Mar 2000, Crystal 2001, 2011a, Ooi 2002 and Baron 2008). Emotives thus comprise both emoticons *and* the initialisms that typically denote the sentiment in which an utterance is produced or intended to be received. Both emoticons and such sentiment initialisms illustrate chatters' intention to ensure that their message, produced on the fly, is correctly interpreted by the recipient. The number of emotives in the corpora is shown in figure 4.16 (based on table 4.8). No figure for the corpus of APMC is available. It was mentioned in the discussion of initialisms, in section 4.5, that all abbreviations in the chatted corpora were annotated for their constituent linguistic items (*idk*, for instance, was tagged with Biber's (1988) features nos. 3, 6, 56, 59 and 67). The initialisms that constitute emotives, however, did not receive this treatment, but rather were assigned the emotive tag only. Emotives will be discussed further, shortly.

Inserts and emotives can both be regarded as operators within the interpersonal metafunction, the tenor of communication, enacting social relationships. Previous sections of the present chapter explored how interpersonal meaning is carried lexico-grammatically by modal auxiliaries and personal pronouns, but also by e.g. markers of mood (WH-interrogatives) and negation (Halliday 1978, 1985a, Halliday & Hasan 1989, Halliday 2004), all part of the modality system of language. In the present section we will explore the ways in which inserts and emotives also, among other things, serve as lubricants in the social machinery.

Hodge & Kress (1988) introduce their discussion of the modality system of language thus:

In every day communication it manifestly matters a great deal what weight we are to attach to an utterance. A statement may be said emphatically, without qualifications, and we know that we are being asked to believe that it is true. Or it may be hedged with 'I think', 'it may be that'. Perhaps it is spoken with a rising intonation like a question, and we know that a speaker is offering the statement more tentatively. Or it may be said with a laugh or an ironic sarcastic tone, and we know that the speaker does not believe in the statement at all. (Hodge & Kress 1988: 121)

Inserts comprise discourse markers and hesitators, which, like the hedges Hodge & Kress mention, construct relations between the communicating parties, signaling their tentative, pending attitudes to messages. Emotives modalize utterances

by indicating the tone in which a “prosodic” unit might be read. Modality is at play in the semiotic act of chatting, as well as in face-to-face interaction, and inserts and emotives may be regarded as important carriers of modality in synchronous CMC, reflecting the tenor of the communication. The primary focus of the present section, however, is not to bear out the modality status of these features, but rather to point to their salience in conversational writing and to contrast their distributions in the annotated corpora (that these features act within the modality system of language is merely background information, implicitly understood).

Biber et al. (1999) note that inserts “comprise a class of words that is peripheral, both in the grammar and in the lexicon of the language” (1999: 1082). They are “stand-alone words” that are generally unable to “enter into syntactic relations with other structures” (ibid.). Nevertheless, they tend to “attach themselves prosodically to a larger structure, and as such may be counted as part of that structure” (ibid.). The inserts found in the annotated corpora are exemplified in table 4.9, below, along with the “larger structures,” i.e. the turns, in which they appear. Inserts either stand alone in the corpora (and comprise a turn in themselves), or else typically introduce larger “prosodic” units. Whereas Biber et al. (1999) classify as interjections only inserts that have an “exclamatory function, expressive of the speaker’s emotion” (1999: 1083), inserts classified as interjections in OED are represented among all the insert types in table 4.9. The few additional inserts found in annotation step two mainly belong to the types “discourse markers” and “polite speech-act formulae.”

The quantitative distribution of each type of inserts in the three corpora investigated is largely depicted by the proportions of exemplified turns in table 4.9. As seen in figure 4.15, SCMC contains the greatest number of inserts; in fact, inserts rank as the third most prevalent linguistic feature in IRC, if seen from the perspective of Biber’s (1988) list (next to present tense verbs and nouns, cf. Appendix II table 1). Table 4.9 reveals the insert type to which the most abundant SCMC inserts belong: greetings and farewells. The abundance of greetings and farewells in SCMC fully accounts for the higher number of inserts in SCMC (i.e. in IRC) overall, as compared to the other annotated corpora. Approximately half of the inserts in IRC are greetings, farewells and attention signals. IRC communication is a textual cocktail party involving the circulation of dozens of participants who, at any given moment, enter and leave rooms, continually greeting each other, calling for attention or bidding each other farewell. Greetings are the most

common initiators of social contact in face-to-face situations and conversational writing alike. In the chat room environment, the initiators often incorporate the nickname of a new participant and serve to confirm that the participant entering the room has been noticed (Anglemark 2009). Biber et al. mention that greetings are usually “reciprocated in a ‘symmetrical’ exchange” (Biber et al. 1999: 1085). In IRC, the reciprocation is not symmetrical (if it was, the quantity of greetings would be intolerable). The split-window ICQ communication contains symmetrical exchanges of greetings and farewells, although considerably fewer than IRC as each ICQ conversation for its full duration here involves only two participants (one conversation involves three). The SBC subset face-to-face conversations contain no greetings or farewells exchanged between informants; the instances found are reported speech. Apart from the disproportion of greetings and farewells, inserts are distributed fairly equally in the three corpora (see table 4.9), except for response forms and hesitators, which appear to be more common in split-window ICQ than in IRC.

The largely similar distribution of inserts in the three corpora makes a strong case for conversational writing as regards orality. Chatters, like face-to-face conversationalists, express emotional involvement by way of interjections. They readily accept the effort it takes to not just produce the conventionalized *oh* and *ah*, but also to create phonological spelling equivalents of other exclamations; see examples in table 4.9 (a finding echoing Ooi’s in 2002). Interjections convey chatters’ and conversationalists’ intensity of feeling alike: their surprise, their sympathy, their laughter – as well as their disgust, and their pain, among other sentiments. Chatters use slightly fewer discourse markers than oral conversationalists, for the reasons adduced in section 4.3 (with regard to Biber’s 1988 discourse particles). The discourse markers used, however, just like in speech, signal transitions in the evolving conversations, as well as “an interactive relationship between speaker, hearer, and message” (Biber et al. 1999: 1086).⁸⁸

88 Biber et al. (1999) suggest the finite verb formulae *I mean*, *you know* and *you see* as discourse markers, admittedly “open to debate” (1999: 1086), but these were not tagged as inserts in the present study.

The chatters (mostly those in IRC) use slightly more attention signals than the SBC speakers, but fewer response elicitors. The paucity of response elicitors briefly reminds us of the textuality of the medium; whereas spoken turns may need repetition to be correctly overheard, turns in conversational writing linger long enough to be re-read. Response forms in conversational writing array themselves in approximately the same orthography as in transcribed face-to-face conversations, with canonical *yeah* overriding the less frequent *yes*, for instance, but differ with regard to backchannels. Speech includes a variety of vocalized sounds as backchannels (transcribed *mhm*, *uh huh*, *unhunh*, etc.), which are not found in the chats. Chatted response forms tend to array themselves in variants of *yeah*, even when used as backchannels and, as mentioned, they are more common in the split-window ICQ chats than in the IRC chats. In both media, response forms, including backchannels, nonetheless serve the same functions as in spoken conversations; they provide answers to yes/no questions, responses to statements, or simply signal feedback to the conversational partner that the messages are understood and accepted – all in order to further lubricate the social machinery and ensure that the communication is functioning well. Backchannels were also found in IM texts in Nuckolls' (2005) study, although fewer than in face-to-face conversations recorded in the same study.

Hesitators are “pause fillers, whose main function is to enable the speaker to hesitate, i.e. to pause in the middle of a message, while signaling the wish to continue speaking” (Biber et al. 1999: 1092). Hesitators are very common in the SBC subset, and interestingly, these “pause fillers” to some extent also occur in conversational writing, despite users' inability in the textual CMC media to audibly hold the floor over their conversational partners. Whereas, in IRC, the hesitator merely signals that the message required some contemplation from its producer, in ICQ, it potentially signals the interlocutor's intention to keep or take over the conversational floor. The higher frequency of response forms and hesitators in split-window ICQ, compared to IRC, thus indicates a certain supersynchronicity effect in ICQ. Just as in oral conversations, while the conversational partner is producing their turn, the ICQ chatters may interpose these inserts to signal simultaneously their understanding, puzzlement or intention to “speak,” whereas this is not possible in IRC. To investigate whether the higher incidence of response forms and hesitators is an effect of the supersynchronicity, however, would require a close examination of the overlapping sequences in ICQ, which unfortunately is unfeasible due to the varying quality of the video recordings of the split-window ICQ material at hand.

The penultimate type of insert, polite speech-act formulae, provides another interesting contrastive finding in the corpora. These inserts are used in conventional speech acts, such as thanking, apologizing and requesting, and are interestingly found to be much more common in IRC than in face-to-face conversations or split-window ICQ. Possibly, the IRC users' lack of acquaintance with each other, and their tentative, forming relationships, trigger a higher degree of politeness among users, a desire to appear polite. Finally, expletives are rare in all three corpora, with taboo expletives non-existent in the IRC chats (in the channels recorded, users were immediately "kicked" upon their use).

All in all, the use of inserts in conversational writing distinctly resembles the use in spoken conversations, both as regards quantity (except for the abundance of greetings and farewells in IRC) and as regards functional quality. Chatters are not just chatters, but also (presumably) experienced speakers and, to further their human relationships, they bring their conversational routines to bear on both social media alike (face-to-face as well as computer-mediated conversations). Inserts provide valuable links between utterances in both forms of social exchange. The distribution of inserts in the written genres remains to be expounded, but is expected to contrast sharply with the corpora annotated here, for which reason future studies of the variation among written, spoken and computer-mediated genres are encouraged to take inserts into account. Halliday (1985a) makes the point that "[t]he spoken language is every bit as highly organised as the written, and is capable of just as great a degree of complexity. Only, it is complex in a different way" (1985a: 87). Whereas written language is "static and dense," spoken language is "dynamic and intricate" (*ibid.*). The present study regards Halliday's (1985a) claim regarding speech equally applicable to conversational writing, and finds inserts to be some of the most central markers of this "spoken language" complexity.

Turning now to emotives, the first point to make is with regard to their "linguistic" status adopted here. Emotives in their current form have been around in the English language since the 1980s.⁹⁰ Common emoticons (e.g. :) ;(;) :-))

90 Scott Fahlman in 1982 suggested the use of :-)) and :-(in a messageboard (ACMC) exchange as "joke markers" and "to mark things that are not jokes," respectively, widely recognized as the original use of what later came to be called emoticons (see <<http://www.cs.cmu.edu/~sef/Orig-Smileys.htm>> for a clip of the original messageboard thread). Lol, meaning "laughing out loud" is claimed to originate from messages in a bulletin board system (ACMC) in the "early-to-mid-80s" (<<http://pages.cpsc.ucalgary.ca/~crwth/LOL.html>>) (cf. Morgan 2011).

and sentiment initialisms (e.g. *lol*), are used and understood by a wide Anglophone, and international, audience. In fact, in 2011, *lol* (meaning “laughing out loud”) entered into OED, as both interjection and noun, with the pronunciations /,ɛləʊ'ɛl/, /lɒl/ in British, and /,ɛl,ou'ɛl/, /lɒl/ in US English. Walther & D'Addario (2001: 329) state that “[a]lthough emoticons may be employed to replicate non-verbal facial expressions, they are not, literally speaking, nonverbal behavior.” They go on to explain that in face-to-face interaction a person may smile unconsciously, whereas in CMC “it is hard to imagine someone typing a :-) with less awareness than of the words he or she is selecting” (ibid.). Marvin (1995) similarly recognizes that smiles in face-to-face conversations can be strategic, spontaneous, or unintentional, whereas in SCMC (more specifically in the mode of MOO that she studied, a text-based online virtual reality system) every smile is consciously indicated: “a conscious choice must be made to type it out” (Marvin 1995: no page number available). Moreover, an SCMC participant “might frown at the keyboard” and yet “decide to type a strategic smile” (ibid.). An emoticon can thus be both strategic and spontaneous, but rarely unintentional (except as a slip of the key). Smileys are not just appended to statements that are ironic or ambiguous; they are also incorporated as “friendly gestures, indications of approval or appreciation” (Marvin 1995: no page number available), much like smiles in face-to-face interaction.

The conscious typing of emotives in the conversational writing corpora in the present study yields a nearly finite set of types, almost as if they belonged to a closed grammatical class. On the other hand, individual emoticons display something like morphological inflections, as :((((((((is a variant of :(. Emotives are at once paralinguistic (indicating the tone of the utterance) and linguistic, constituting tokens in their own right (usually set apart from other words orthographically). They resemble other paralinguistic features of chat (like repeated exclamation marks appended to words for emphasis), but are not appended to other words – rather, more like inserts, stand-alone words or appended to prosodic units, like the laughter particles identified as inserts above (e.g. *hehe*). On the other hand, emotives do not lend themselves easily to phonology; only pronounceable ones (*lol* and *rofl*) have crossed over into speech and thereby become lexicalized. CMC studies to date have typically regarded emoticons and sentiment initialisms as paralinguistic features of the communication, substituting for the lack of non-verbal cues (e.g. Dery 1993, Thompson & Foulger 1996, Werry 1996, Schulze 1999, Derks et al. 2007, Waldner 2009). Crystal (2001),

however, hesitates to call emoticons paralinguistic, emphasizing that “they have to be consciously added to a text” (2001: 34). Dresner & Herring (2010) also extend the function of emoticons beyond substituting for non-verbal cues, construing them as “textual indicators of illocutionary force” (2010: 260). The present study recognizes the paralinguistic denotation of emotives; they are chatters’ own ways of transcribing their “speech.” The present study, nevertheless, is an investigation into the variation between genres of writing and speech, and such a study needs to recognize every token of the texts. Once- or twice-occurring graphological tokens, like `_,-^*^*-,__,-*'` (see example 28 in section 4.5), are easily dismissed as hapax legomena or as void of meaning, whereas emotives carry modal meaning and can be expected to recur in texts. After all, the most common ones have recurred in texts for thirty years, to date. It is, consequently, high time that emotives be given linguistic status as markers of CMC discourse. In variation studies, they effectively set computer- and cellphone-mediated texts apart from other texts, and thus, clearly, constitute a linguistic feature to take into account in future multidimensional studies of the variation of the English language. The remainder of this section presents the distribution of emotives in the annotated conversational writing corpora.

Recall from figure 4.16 that SCMC (that is, IRC) contains far more emotives than SSCMC (that is, split-window ICQ). In IRC, emotives are the ninth most common feature, more common than for instance past tense verbs, third person pronouns and pronoun *it*. *Lol* is the predominant marker of emotional involvement in both modes of CMC; in IRC *lol* accounts for 56 percent of all emotives, in ICQ it accounts for as much as 73 percent. In spite of this, the use of *lol* is much more rare in split-window ICQ than in IRC. The distribution of the individual emotives in the conversational writing corpora is illustrated in figure 4.17, detailing their overall distribution from figure 4.16, per thousand words.


```

<Cheeky1>      lol
<Cheeky1>      i will be back
<|mad_max|>    ok .....
<|mad_max|>    take care
<Cheeky1>      gotta go for 5 minutes
<Cheeky1>      u 2 max sweety
<Cheeky1>      c ya in a sec u hunk of spunk
<Cheeky1>      hehehe
<|mad_max|>    c u
<Cheeky1>      cya
<Cheeky1>
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
<Cheeky1>      :0)
<USA__MALE>    hi

```

Internet relay chat text 3a (UCOW)

One and the same user (Cheeky1), first signaling his/her appreciation of a fore-going joke about women, and later signaling his/her continued sympathetic presence, produces all *lol*-turns in the example. At the end of the example, the user announces his/her exit, and before leaving flashes a brief :0) "grin." Stand-alone emotives (constituting a turn in themselves) are found in both IRC and split-window ICQ, but in IRC, stand-alone *lol* appears to function more often as a jovial presence marker than as a transcription of actual laughing. Chatters in IRC are in initial stages of contact and are concerned about appearing congenial. *Lols* and smiles are therefore sprinkled into the IRC conversations much as friendly smiles would be in face-to-face first encounters. Such use of emotives seems to account for much of the discrepancy in the emotives distribution between the two modes of CMC. In the split-window ICQ chats (SCMC, right bars in figure 4.17), the *lols* seem more co-textually motivated, as for instance in the amusement 10 expresses over the comment J makes about his sister in example (54).

```

(54) <J>      to tell the truth.. i dont think i've ever seen my sister go 10
        feet away from the shore.. let alone anywhere else in a big body of water
<10>  lol

```

Split-window ICQ chat text 9 (UCOW)

In both modes of CMC, initialisms appended to turns appear in both initial and end positions, with a few rare instances in medial position. Emoticons in IRC turns appear in medial and end positions, whereas in split-window ICQ they are exclusively appended at the end. It seems as though IRC chatters are more concerned than the ICQ chatters to set the tone of their message as early as possible. Smiles in both media represent friendly smiles more than laughter, and in IRC

they are typically appended to turns close to greetings and farewells; see the various IRC-turns in (55). The winking smiley is, surprisingly, more common in IRC than in split-window ICQ, possibly because in IRC it also appears at the end of greetings and farewells as in the last two turns in (55). The winking smiley otherwise prototypically signals tongue-in-cheek comments, as in (56), and given the ICQ chatters' previous acquaintance, they could be expected to use them more.

- (55) a. hi again rainman19 :)
 b. puck....hello to you too..:))
 c. AdamSxy35 :)
 d. REVOLI, Im fine, how are you? :)
 e. Raha,take care, bye :)
 f. hiya CityWoman and y'all;)
 g. Ta ta Adam...;)

Internet relay chat (UCOW)

- (56) <AdamSxy35> oups why dont you try a business chat room on yahoo?
 <_oups> hm...well do they have that..
 <AdamSxy35> it works for me when i cant fall asleep ;)

Internet relay chat text 5b (UCOW)

In general, the IRC corpus displays a wider emoticon repertoire than the split-window ICQ corpus. IRC chatters are presumed to be experienced emotives-users, and often thought to proliferate emoticons. In a large-scale emoticons study, however, Schulze (1999) plays down the need for smiley dictionaries. His 28,345 "line" long IRC chat corpus contains no more than eight major types of emoticons (with several minor variations) (1999: 76). Ten years later, Waldner (2009) finds no more than 15 emoticons used regularly in IRC (2009: 81). The IRC corpus in the present study is about ten percent the size of Schulze's, but can be said to proportionally agree with his findings. Out of Schulze's (1999) eight major types of emoticons, the present study finds representatives of four: the "smiley" :), the "frowney" :(, "sticking out tongue" :P and "slight frown" :/, but also two additional major types: the "winking smiley" ;) and the "indifferent" one :I, i.e. altogether six types. In the split-window ICQ corpus, the emoticon repertoire is even more limited, with representatives of only four major types.

A few writers have investigated linguistic gender differences in computer-mediated communication, putting e.g. Coates (1993) and Tannen's (1990, 1994) findings of gender-differentiated conversational styles to the test on empirical CMC data. Herring (1996b) finds women and men to present different styles of interaction and information exchange on two Internet mailing lists (ACMC), styles that she terms the "aligned variant" (supportive, mostly used by women)

and the “opposed variant” (more insulting or aggressive, mostly used by men). Echoing this finding, Herring (1998, 2003) notes that, in SCMC, women type three times as many representations of smileys or laughter as do men. Wolf (2000) finds women to use more emoticons in same-gender newsgroups (ACMC), but finds no significant difference between women’s and men’s use in mixed-gender newsgroups. Baron (2004) describes a study of instant messaging (IM) data, collected among college students, in which she found differences e.g. in women’s and men’s use of emoticons (women used more) and contractions (men used more). Replicating these studies on the UCOW IRC data is not feasible, as no record of the Internet relay chatters’ gender exists, but for the ICQ data a comparable investigation yields interesting results with regard to emoticons (no comparable investigation was carried out on contractions).

Baron’s (2004) IM corpus is approximately the same size as the UCOW split-window ICQ corpus and thus comes in handy for a comparison. A total of 49 emoticons were used in Baron’s data. Females were found to be the prime users of emoticons; out of the 16 female participants three-quarters used one or more emoticons. Of the 6 male participants only one used emoticons (2004: 415). The results for the comparable analysis of the UCOW split-window ICQ data are presented in the first row of table 4.10.

Table 4.10: *Individuals’ emotives usage in the split-window ICQ corpus, by gender; f=female (7), m=male (18). N.B. raw figures*

	f f f f f f f	m m m m m m m m m m m m m m m m m m m
Emoticon	3 0 0 0 0 0 0	2 2 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Sentiment initialism	0 0 0 0 1 2 8	0 2 2 0 0 0 0 0 0 0 0 0 0 0 1 1 2 2 3 4

A total of nine emoticons are used in the split-window ICQ corpus. *Males* are here found to be the prime users of emoticons; out of the 18 male participants, four used emoticons, whereas out of seven females, only one did the same. All the while, however, 28 sentiment initialisms were used (all of them *lol*, except one *lmao*; see the second row in table 4.10); 43 percent of the females used sentiment initialisms, and 44 percent of the males. About half of the males used emotives overall, whereas fewer than half of the females did. In other words, the findings for the UCOW split-window ICQ corpus do not corroborate the findings in Baron (2004) with regard to emoticon use. On the other hand, the average number of emotives used by males in the ICQ corpus is only 1.3, whereas for females the same number is 2. One of the females produced the highest number

4.7 Chapter summary

The present chapter has expounded on the salient features in conversational writing. The bulk of the chapter zoomed in on the ten linguistic features which, in either mode of CMC, synchronous or supersynchronous, or in both, deviate from Biber's (1988) mean for spoken and written language overall by more than two standard deviations. The chapter set out from two of the important carriers of interpersonal meaning in language: modal auxiliaries and personal pronouns, the latter of which reveals salient traits in the chatted texts, the pervasive first and second person pronouns. Next, the lexical properties of the conversational writing genres, writing, and speech were investigated, through the employment of contrasted measurements of word length, type-token ratio and lexical density, essentially revealing the latter to be most appropriate for capturing the grammatical intricacy (or lack of lexical density) of the chatted texts. The fourth section presented the salient features annotated in the corpora and what each of them reveals about the communication, as regards, quantity, quality, orality and Halliday's metafunctions, most notably about the tenor of the discourse. Paralinguistic cues and extra-linguistic features were then surveyed, which further incorporated consideration of the textual and ideational metafunctions (the semiotic mode and field of the discourse). Finally, the last section proposed two linguistic features to be incorporated into future accounts of the variation of the English language, inserts and emotives, which both serve important functions in computer-mediated communication. In the next chapter, the more granular, yet all-round, results of the application of Biber's (1988) methodology will be presented and the positions of the conversational writing genres on Biber's dimensions of linguistic variation revealed. It will be seen there, that most of the salient linguistic features presented above load on one and the same dimension of variation (Dimension 1), distinguishing involved production from informational. As mentioned in chapter 3, inserts and emotives are not linguistic features in Biber's (1988) methodology and consequently have no bearing on the dimension scores to be presented in chapter 5 (except for a few items contained within inserts, which are also tagged in Biber's methodology). The present chapter subsumed numerous written genres into mean figures for writing, and numerous spoken genres into mean figures for speech. In chapter 5, these two multitudes of genres will be split up to further diversify the contrastive analysis, but in the end to make for a unified picture as regards the nature of synchronous and supersynchronous computer-mediated communication.