Chapter 3. Material and method

3.1 Introductory remarks

In the present study, two genres of conversational writing are contrasted with spoken and written genres to investigate the relationships between conversational writing, speech and writing. A corpus of conversational writing was collected for the purpose, and a corpus of face-to-face conversations was sampled to create a complementary corpus of spoken conversations. As the study uses Biber's (1988) methodology, the chatted texts will also be contrasted with the two corpora studied by Biber, i.e. LLC for speech and LOB for writing (see Appendix I). The corpus of conversational writing consists of two components: one of SCMC, viz. Internet relay chat, and one of SSCMC, viz. split-window ICQ chat, both annotated for the purposes of the present study. The complementary, spoken corpus consists of a subset of face-to-face conversations sampled from the Santa Barbara Corpus of Spoken American English, part 1 (Du Bois et al. 2000). The sub-corpus, henceforth called the "SBC subset," was essentially brought into in the study to supplement Biber's (1988) genres from LLC with more recent spoken material. Table 3.1 presents briefly, in numbers, the corpora compiled/sampled and annotated, corpora on which most of this chapter focuses (for Biber's 1988 corpora, see Appendix I). (The corpus of ACMC, with which findings will also be contrasted, is introduced in section 4.1.)

Table 3.1: Size of corpora compiled/sampled and annotated for the present study

	Corpus	Number of texts	Min. length	Max. length	Average length	Corpus size
Conversational writing (UCOW)	Internet relay chat	10	961	999	984	9,841 words
	Split-window ICQ chat	12	459	1150	772	9,261 words
Face-to-face conversations	SBC subset	14	673	720	712	9,962 words

The next two sections of this chapter, 3.2 and 3.3, describe the collection, adaptation and annotation of the conversational writing corpus, UCOW (Uppsala Conversational Writing Corpus). Each of the two UCOW components, Internet relay chat and split-window ICQ chat, constitutes a corpus in itself, meaning that it may be referred to both as a corpus and as a UCOW component throughout this study. The ensuing section, 3.4, describes the sampling and annotation of the

SBC subset. The chapter proceeds, in section 3.5, to explain how Biber's (1988) MD methodology was applied to the data, focusing on the frequency standardizations and dimension score calculations. Section 3.6 then turns to the two corpora studied in Biber (1988), summarized in Appendix I. The section describes how average figures for writing and speech were obtained from these, for speech in combination with the SBC subset data, and how the results will be presented and analyzed in this study. The final section, 3.7, sums up the chapter.

3.2 Creating and annotating a corpus of Internet relay chat

The corpus of Internet relay chat was recorded in 2002 from five public Internet relay chat channels: #20_something, #30_something, #Chat_world, #Family and #USA. Each channel was logged for several hours, yielding log files of several hundred thousand words in total for the five channels. From these log files, two sample texts were drawn from each chat channel (texts 1a and 1b from #20_something, texts 2a and 2b from #30_something, etc.). Each sample text consisted of thousands of words but was subject to a purging procedure aimed at sifting out only the linguistic messages explicitly keyed in by the human conversationalists. The raw samples were all long enough to ensure that the resulting texts, containing the full log of user-generated messages, would comprise approximately one thousand words each; cf. table 3.1. Exemplified in (1) are the first few minutes of an unpurged sample from the channel #Family, which ultimately contributed the first twelve lines to Internet relay chat text 4a in UCOW. Sample (1) includes a session start message, time stamps, bracketed nickname turn indicators and server-generated messages such as join- and quit-information (indicated by ***), of which all except the bracketed nickname turn indicators were purged to sift out the twelve user-generated turns. The twelve turns of current interest are then re-presented as (2).[37]

(1) Session Start: Mon Mar 25 18:01:47 2002
[18:01] *** Now talking in #family
[18:01] <River> woohoo,
[18:01] <Genie500> Laughing Out Loud
[18:01] <River> my hair is almost as long as yours
[18:02] <Genie500> now ya know who to look for honking across the street
[18:02] <River> yep
[18:02] <Genie500> really?? lol

37 Examples of material excluded from (2) and other passages of text, e.g. channel operator interference, action commands and foreign language turns, are found in Appendix IV.

[18:02] *** edi-tr has joined #family
[18:02] <River> well just in the back
[18:02] <Genie500> Laughing Out Loud
[18:02] *** edi-tr has quit IRC (Killed (NickServ (Nickname Enforcement)))
[18:03] <Genie500> and what color is yours??
[18:03] *** EmeL_ has joined #family
[18:03] *** Guest_984 has joined #Family
[18:03] *** edi- has joined #family
[18:03] <lookingforagirl> blue
[18:03] *** blue-ice has joined #Family
[18:04] <Genie500> oh river just a sec I gotta turn something off for you to send okay
[18:04] <River> this one is from 95 without the glasses .

(2)

<River>	woohoo,
<Genie500>	Laughing Out Loud
<River>	my hair is almost as long as yours
<Genie500>	now ya know who to look for honking across the street
<River>	yep
<Genie500>	really?? lol
<River>	well just in the back
<Genie500>	Laughing Out Loud
<Genie500>	and what color is yours??
<lookingforagirl>	blue
<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 .

Internet relay chat text 4a (UCOW)

Example (2) represents the default format in which examples from the Internet relay chat texts are presented in the ensuing chapters of this study.[38] It retains the bracketed nickname turn indicators to mark turn boundaries, although, needless

38 Nicknames are pseudonyms in themselves, and not real-life identities, and have traditionally been retained in published research on texts deriving from public CMC domains, such as public IRC channels, with non-sensitive content (cf. Werry 1996, Danet et al. 1998, Schulze 1999, Ooi 2002, Waldner 2009). As their real-life identities are disguised by such pseudonyms, the IRC chatters were not asked for their informed consent to be recorded in the present study. This is in line with e.g. Rafaeli et al. (1998), Liu (1999), Cameron (2001) and Waldner (2009), the second of whom notes that public IRC conversations are "acts deliberately intended for public consumption" (Liu 1999: no page number available) and may be regarded as exempt from practices of obtaining informed consent. Moreover, it was felt that intruding in conversations to obtain chatters' consent was not only impracticable, but also would have disrupted the natural

to say, the indicators were not subject to linguistic annotation or feature counts. Rather, the texts for annotation contain only the linguistic messages explicitly keyed in by participants; see example (3). It is in a state such as (3) that the corpus is reflected in table 3.1; the Internet relay chat component of UCOW contains a total of 9,841 words of such linguistic messages exchanged between interlocutors, drawn from 10 texts averaging 984 words each.

(3) woohoo,
Laughing Out Loud
my hair is almost as long as yours
now ya know who to look for honking across the street
yep
really?? lol
well just in the back
Laughing Out Loud
and what color is yours??
blue
oh river just a sec I gotta turn something off for you to send okay
this one is from 95 without the glasses.

Internet relay chat text 4a (UCOW)

Once the texts had been distilled to the format exemplified in (3), the linguistic annotation began. The purpose of the annotation procedure was to mark up all occurrences of Biber's (1988) 65 linguistic features likely to distinguish among spoken and written texts; see table 2.1 (the features type/token ratio and word length do not require markup). Attempts were made to run a few texts through automatic part-of-speech tagging software, e.g. the CLAWS tagger, for an annotational starting point. Because of the irregular orthography of the chatted texts, however, the automatic taggers repeatedly failed or achieved insufficient accuracy, which made manual annotation the only remaining option.[39]

The manual annotation proceeded in a series of time-consuming steps, each involving meticulous consideration.[40] First, the texts were annotated for parts

flow of conversations to the point of destroying the data (cf. Sveningsson 2001, Waldner 2009).

39 Developing software for the purpose of tagging the UCOW texts was beyond the time allotted for the study. Manual annotation was instead carried out straight onto the texts in a word processor, which eventually conveniently enabled feature counts by way of the program's "find all" option.

40 The demanding task of manual annotation explains the relatively limited size of the corpora compiled and sampled for the present study (see table 3.1). However, although 10,000-word corpora may appear small, few corpus linguistic studies have been based

of speech roughly in accordance with the tagset devised in the Penn Treebank project (Santorini 1990, Marcus et al. 1993). The part-of-speech tagging provided useful basic classification, as some of the Penn Treebank tags directly correspond to features in Biber (1988), e.g. VBD for past tense verbs and NN for nouns. In the annotation process, however, a great number of tags had to be modified to denote Biber features, and new tags invented ad hoc, to eventually cover all of the 65 features. Example (4) shows the resulting annotation of the example considered here, in which a great number of custom-made tags mark up the linguistic items.

(4) woohoo,/IJ
Laughing/VX Out/ADV Loud/ADV
my/PRP1s hair/NN is/VBP/-BE almost/46/47 as long/-PADJ/+ADJ as yours/+PRP2s
now/TADV ya/PRP2s know/VBP/PRV who/23 to/TO look/VBI for/PP honking/-28 across/PADV the street/NN
yep/IJ
really??/IJ/ADV lol/EM
well/IJ/ADV/DP just/49 in/PP the back/NN
Laughing/VX Out/ADV Loud/ADV
and/CONJ/65 what/?WHQ color/NN is/VBP/BE yours??/+PRP2s
blue/-PADJ/+ADJ
oh/IJ river/NPna just/49 a sec/NN I/PRP1s gotta/+NM/?CT turn/VBPi
something/PN off/ADV for/PP you/PRP2s to/TO send/VBI okay/-PADJ/+ADJ
this/DET one/NN is/VBP/BE from/PP 95/CD without/PP the glasses/NN .
Internet relay chat text 4a (UCOW)

Table 3.2 provides a key for interpreting the tags in (4). The annotation was tailored to comply closely with the algorithms provided in Biber (1988: 211–245) for the detection of linguistic features. Matching the algorithms in the annotation was a *sine qua non* for ensuring the material maximum comparability with the spoken and written texts in Biber's study. As an example, verbs were tagged as infinitives only when following infinitive marker *to* and an optional adverbial element (that is, whenever identified by the algorithm "to+(adv)+vb"), as *look* and *send* in example (4), both tagged /VBI (Biber's feature no. 24). Occasionally,

on such large manually, and single-handedly, lexico-grammatically annotated datasets. Three of Biber's (1988) (automatically annotated) genres are of a similar size, viz. science fiction, personal letters and professional letters (see Appendix I), which is seen as unproblematic as Biber (1990) convincingly demonstrates and establishes that ten 1,000-word text samples (i.e. a total of 10,000 words) suffice for the adequate lexico-grammatical representation of a genre.

the close compliance with Biber's algorithms meant that an item's affiliation with a particular linguistic category had to be ignored. Consider, for instance, the occurrences of *be* as main verb in example (4), one tagged /-BE and two tagged /BE. Biber's algorithm for *be* as main verb (1988: 229) detects only the instances in which the verb is followed by a determiner, possessive pronoun, address title, preposition or adjective. The algorithm thus would exclude the first instance of *be* as main verb in (4), as it is followed by a downtoner adverb, but properly detect the other two instances, which are followed by a possessive pronoun and a preposition, respectively. Accordingly, in the manual annotation, the first item had to be marked as deviant, /-BE, and only items tagged /BE eventually counted as instances of Biber's feature no. 19, as inferable from table 3.2.

Table 3.2: Tags used in the annotation of the first twelve turns in Internet relay chat text 4a (UCOW)

Tag	Description	Feature no.	Explanation
+ADJ	adjective not identified as predicative	40	
ADV	adverb	42	
BE	BE as main verb	19	
–BE	not identified as BE as main verb		see Biber's algorithm for 19
CD	cardinal number		not a Biber feature
CONJ	conjunction		not a Biber feature
?CT	not identified as contraction		see Biber's algorithm for 59
DET	demonstrative	51	
DP	discourse particle	50	
EM	emotive		genre-specific feature
IJ	interjection and/or insert		genre-specific feature
+NM	not identified as necessity modal		see Biber's algorithm for 53
NN	noun	16	
NPna	proper noun, nickname address term		genre-specific feature
–PADJ	not identified as predicative adjective		see Biber's algorithm for 41
PADV	place adverbial	4	

Tag	Description	Feature no.	Explanation
PN	indefinite pronoun	11	
PP	preposition	39	
PRP1s	first person singular pronoun	6	
PRP2s	second person singular pronoun	7	
+PRP2s	not identified as second person singular pronoun		see Biber's algorithm for 7
PRV	private verb	56	
TADV	time adverbial	5	
TO	infinitive marker		not a Biber feature
VBI	infinitive	24	
VBP	present tense verb	3	
VBPi	present tense verb, base form	3	
VX	progressive verb		not a Biber feature
?WHQ	not identified as WH-question		see Biber's algorithm for 13
23	WH clause	23	
−28	not identified as present part. WHIZ deletion		see Biber's algorithm for 28
46	downtoner	46	
47	hedge	47	
49	emphatic	49	
65	non-phrasal coordination	65	

Table 3.2 indicates which items in example (4) count as instances of Biber's (1988) features (by feature numbers in the third column in accordance with table 2.1 of the present study) and explains why others do not. Items not counted as instances of Biber features diverged for one of two reasons: either they did not conform to detection by Biber's algorithms (like/-BE) or they simply constitute a feature not studied by Biber (e.g. /CD, cardinal numbers); see the explanations column in table 3.2. Moreover, three tags were developed in the present study to denote instances of genre-specific features, none of which was studied in Biber (1988): the tag for nicknames used as address terms, /NPna, typical of IRC; the tag for inserts, /IJ, for e.g. interjections, frequent in conversational genres; and the tag for emotives, /EM, marking up emoticons ("smileys") and sentiment initialisms, typical of chatted texts. Nicknames used as address terms (e.g. *river* in

example 4) are not regarded as nouns in this study (unlike other proper nouns), to avoid skewing the noun count of the IRC texts, but their frequency of occurrence nevertheless affects the relative incidence of other features in IRC, as will be seen in section 5.2.1. Inserts and emotives will be expounded upon in section 4.6, in which the annotation of inserts is explained further.

Even though table 3.2 lists only the tags used for the example text given here, it is indicative of the procedure for detecting Biber items, the items to be included in the linguistic feature counts. Among the more than one hundred tags used for the annotation, several were interim, and a slightly more limited set was eventually subjected to feature counts – the tags that mark up Biber features. The final step of the annotation process involved summing up the latter items. Example (5) illustrates the incidence of Biber features, as such items are here numbered in accordance with the third column of table 3.2 (based on the feature list in table 2.1).

(5) woohoo,/IJ
Laughing Out/42 Loud/42
my/6 hair/16 is/3 almost/46/47 as long/40 as yours
now/5 ya/7 know/3/56 who/23 to look/24 for/39 honking across/4 the street/16
yep/IJ
really??/IJ/42 lol/EM
well/IJ/42/50 just/49 in/39 the back/16
Laughing Out/42 Loud/42
and/65 what color/16 is/3/19 yours??
blue/40
oh/IJ river/NPna just/49 a sec/16 I/6 gotta turn/3 something/11 off/42 for/39 you/7 to send/24 okay/40
this/51 one/16 is/3/19 from/39 95 without/39 the glasses/16.

Internet relay chat text 4a (UCOW)

The sample in (5) thus contains, for instance, seven adverbs (feature no. 42), two first person pronouns (feature no. 6), seven nouns (feature no. 16), etc. As mentioned above, the genre-specific tags /NPna, /IJ and /EM are of further interest (the first in section 5.2.1, and the latter two in section 4.6), but are not numbered in (5), as they are not among Biber's (1988) features.

As seen in table 3.1, the UCOW Internet relay chat texts contain 984 words on average. To make feature counts comparable across texts and genres, Biber's (1988) methodology prescribes the normalization of counts to occurrences per 1,000 words. Internet relay chat text 4a contains 975 words, which means that the occurrences had to be multiplied by 1,000/975 to attain normalized frequencies. The raw frequency for adverbs in text 4a, 82, for instance, was normalized to 84.1 (as 82 × 1,000/975 is 84.1). The resulting normalized frequencies for all

the features in every Internet relay chat text are found in Appendix II table 5, and the normalized frequencies for the whole Internet relay chat corpus are found in Appendix II table 1 (based on the sum of average frequencies in the individual texts, divided by the total number of texts). The tables in Appendix II thus sum up results of fundamental importance for the Internet relay chat corpus, as well as those for the other corpora studied, results upon which this study is based.[41]

Two of Biber's (1988) features, type-token ratio (TTR) and word length, are used to measure the lexical diversity and specificity of texts. These two features did not require annotation, but necessitated some other processing of the texts. For the purpose of TTR and word length calculations, the texts in the present study were purged of all regular punctuation except apostrophes within words, emoticons and simple imagery,[42] rendering texts the appearance exemplified in (6). In compliance with Biber (1988: 238), TTR was calculated "by counting the number of different lexical items that occur in the first 400 words of each text, and then dividing by four." Three pieces of lexical analysis software were used for computing the TTR, viz. KWIC 5.0, AntConc 3.2.4 and WordSmith Tools 5.0.0.334, which all yielded congruent results. For the average word length count, only WordSmith Tools was used; the full texts were used as input and no normalization was needed. Section 4.3 explains the procedures further and discusses the results. The figures for TTR and word length in the texts studied are given among the features in the tables of Appendix II (features 43 and 44), as numbers equally central to the present study as those of the annotated features.

(6) woohoo
Laughing Out Loud
my hair is almost as long as yours
now ya know who to look for honking across the street
yep
really lol
well just in the back
Laughing Out Loud
and what color is yours
blue
oh river just a sec I gotta turn something off for you to send okay
this one is from 95 without the glasses

Internet relay chat text 4a (UCOW)

41 The raw frequencies are provided in Appendix III, but no reference is made to them in the study.

42 "Simple imagery" is explained further in section 4.3.

Besides TTR and word length, section 4.3 of the present study also considers the lexical density of texts. Lexical density was introduced in section 2.4 as Halliday's (1985a) only quantitative means for distinguishing between spoken and written texts, indicating a low lexical density for spoken and a high lexical density for written texts. The lexical density calculation in the present study was carried out on texts in a state such as in example (3) above, i.e. with the punctuation retained. Single stranded punctuation marks were not counted, but emoticons (e.g. *:), ;)*) were counted as words (non-lexical). Lexical density measures the ratio of lexical words (i.e. content words) to the total number of words. The measurement is insensitive to text length, which means that it was computed on the full corpus, and was not normalized. The full corpus size, however, was increased slightly before the lexical density calculation, to compensate for abbreviations and contractions, the former typical of conversational writing. Abbreviations such as *idk* meaning "I don't know" and *nm* meaning "not much" were thus tokenized, that is, considered as if their constituents were spelled out, i.e. *idk* as four words (*I, do, n't, know*), *nm* as two, except for sentiment initialisms, e.g. *lol*, *rofl* and *lmao*, which were considered as uniform words.[43] The lexical density count further considered accidentally conjoined words (such as *guessyea*) as separate tokens, and accidentally separated words (such as *out side*) as single tokens. To account for these irregularities, a total of 165 tokens were added to the Internet relay chat corpus size, most of the tokens deriving from the tokenization of abbreviations and contractions. Section 4.3 details which of all the tokens were then taken to be lexical words. As mentioned, lexical density is a measurement of the ratio of lexical words to the total number of words, the total number of words in the Internet relay corpus being 9,841+165, i.e. 10,006, for the lexical density calculation only. Section 4.3 further explains how the measurement of lexical density was applied to the material and discusses the findings.

43 Determinant for the treatment of chat abbreviations was their propositional content. Whereas the tokenized abbreviations typically convey propositional content, the sentiment initialisms, just like emoticons, typically are non-propositional and rather may be regarded as "textual indicators of illocutionary force" (cf. Dresner & Herring 2010: 260). Moreover, the first two sentiment initialisms, *lol* and *rofl*, in effect have become lexicalized in the English language; see further section 4.6. *Lol* means "laughing out loud," *rofl* "rolling on the floor laughing" and *lmao* "laughing my ass off" (Crystal 2004a). Sentiments spelled out in the original text, e.g. *Laughing out loud* in examples (2)–(5), of course, remained several tokens in the lexical density calculation.

3.3 Creating and annotating a corpus of split-window ICQ chat

Collecting a corpus of split-window ICQ chat between individuals in private conversations demanded a greater effort on the part of the present researcher than did the recording of the Internet relay chat discourse, which is readily available in public chat channels (cf. section 3.2). The split-window ICQ chat component of UCOW was collected in 2004 by logging conversations between high school seniors in the northeastern USA. The present author involved two high schools in the project, which yielded 23 informants' conversations, and two high school students were recorded in their home. All in all, 12 texts of split-window ICQ chat were compiled, as indicated in table 3.1. Out of these texts, eleven were conversations between dyads and one was a conversation among three people. Eighteen students were male and seven were female. Most conversations took place between males or in mixed-sex dyads; only one conversation involved two females.

About a week before the recording, the subjects were informed about the study, both orally and in writing, and asked to bring home an informed consent form for their guardians' review and signature, if the student was underage (below the age of 18). All students interested in participating brought home the form and brought it back signed, regardless of their age. For the recording in the home setting, informed consent was obtained orally from the subjects' parents. The recording event took place during one lesson in each high school, and for the equivalent period of time in the home setting. A computer classroom was set up for the purpose in the high schools, and a home office for the latter event. The students formed dyads, and a triad, on their own before entering the classroom and were assigned computers as distant from their conversational partner as possible upon entering the classroom. Computers had been pre-set with the required software, the ICQ chat Pro 2003b program as well as the HyperCam screen capturing software, the latter intended to capture the split-window ICQ chat action as a video file. After all students had been seated, they were introduced to the software and allowed ten minutes to practice. As they were all apt and avid online chatters from before (most with more than ten hours of chatting experience, typically in chat rooms or on AIM, although none had used ICQ), they immediately caught on and managed the ICQ program. The following four instructions were given in a sheet taped to the physical desktop beside each computer:

> Do not move, close, cover up or disrupt the chat window while recording
> Immediately close any popup-windows
> Do not follow links or advertisements
> Would you like to save this Chat session? Click OK and save to desktop

Because of the limited time allowed in the classroom, the time for recording conversations was restricted to about 20 minutes. Students were informed that the content of their discourse would not be subject to assessment and were explicitly encouraged to converse freely on any topic of their choice. Nevertheless, before the students were instructed to initiate the recording, they were given a few topics to resort to in the event that their conversations might run dry. The suggested topics were written on the board: "Plans for the weekend," "Plans for the summer," "Plans for next year"; in the home setting the same topics were suggested orally.[44] As it turned out, only a few utterances in four of the texts may have sprung from these suggestions. As will be seen in textual examples throughout this study, most split-window conversations revolve around other topics and appear remarkably uninhibited and diverse, considering that the discourse was recorded in a situation of elicitation. Example (7), the first 15 turns of split-window ICQ chat text 8, indicates typical split-window ICQ discourse produced, with very few traces of the situation of elicitation (such traces are discernible in lines 6–10, i.e. only in four out of 71 lines in the full text). In the recordings, the interlocutors typically carried on eagerly with the conversation initiated while practicing and were all noticeably unperturbed by the experimental setting, as evident in lines 11 ff. of example (7). Upon finishing their recording, the students saved their conversations both as a screen-captured video clip and as a textual log file. Before leaving the room, they were further instructed to retrieve a minor remuneration for their participation from underneath the taped instruction sheet, which they did gladly (as no remuneration had been mentioned before).

(7)	<9>	YES
	<9>	!!!!!!!
	<I>	
	<I>	hey baby
	<9>	we suck at this
	<I>	well there ya go… uits time to record our 20 minutes sessions
	<9>	did u press record
	<9>	yep
	<I>	yeah did u?
	<I>	ok good
	<I>	so question…
	<9>	who said i hooked up with her

44 Future plans was one of several productive topics given to informants in Renouf's (1986) elicitation of spoken English.

<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.

Split-window ICQ chat text 8 (UCOW)

The corpus of split-window ICQ chat was then collected from the classroom computers. Due to the varying quality of the video clips, it was decided that the textual log files would constitute the material for lexico-grammatical analysis (the corpus is made up of the entire collection of logs, i.e. the chats were not sampled). Unlike the IRC logs, the raw textual log files of the split-window chats are readily legible; see example (7), in effect derived straight from such a file.

As seen in (7), participants' turns are preceded by bracketed turn indicators. Just as in IRC, these contain a nickname, even though, for practical reasons, the nicknames in the split-window ICQ recordings were pre-set on computers and not invented by the participants. Unlike in IRC, the split-window chatters were able to personalize their messages with variable fonts, font size and font color, which they did, even though none of this is reproduced in the text samples in this study. This variability in ICQ, nevertheless, is seen as one of the paralinguistic devices available to chatters and is further discussed in section 4.5. The ICQ program also offers a set of graphic emoticons and pre-programmed textual action tropes. Participants did not use the graphic emoticons, but experimented somewhat with the action tropes. A trope is realized as a line of text, which is not explicitly keyed in by the participants, but assigns an action to his/her nickname (e.g. *9 picks a flower and hands it to you*), imitating an action command in IRC. Just as for the IRC material, however, the action lines of the split-window ICQ chat component were removed before the annotation began, along with the bracketed nickname turn indicators and a few foreign language turns (see Appendix IV). No other purging was needed to adapt the ICQ material for the lexico-grammatical annotation and analysis. The average text of the analyzed split-window ICQ chat component is 772 words long, as indicated in table 3.1, and the whole corpus contains 9,261 words.

The annotation of the split-window ICQ texts followed the same procedure as did the annotation of the IRC texts (cf. section 3.2), i.e. the same tags were applied, and Biber's features were eventually identified and counted and the frequencies normalized. The normalized frequencies of the 65 features, as well as the TTR and word length of the individual split-window ICQ chat texts, are found in Appendix II table 6, and the figures for the whole split-window ICQ corpus are found in Appendix II table 2. For the lexical density calculation, as described for

IRC in section 3.2, the size of the split-window ICQ corpus was increased slightly to compensate for items contained in abbreviations (e.g. for *ic*, meaning "I see") and contractions, as well for accidentally conjoined words (e.g. *lastnight*) minus accidentally split-up words (e.g. *any way*). A total of 51 tokens were added to the corpus size, thus making the denominator for the split-window ICQ corpus in the lexical density computation 9,312 words. The lexical density of split-window ICQ chat, just as that for IRC, will be further discussed section 4.3.

Examples from the split-window ICQ chats will be given in a format such as that in (7) throughout this study, i.e. with bracketed turn indicators retained. Informants' textual references to personal names, locations, etc., have been carefully masked in order to preserve informants' anonymity; accordingly, *laurie* in (7) is fictitious. The annotation and the TTR, word length and lexical density calculations, however, were carried out on the original texts with original verbatim references.

3.4 The Santa Barbara Corpus subset

The present section briefly introduces LLC's conversational genres and compares them to the UCOW genres, in order to explain the motives for studying SBC as a supplementary corpus to LLC. The section further describes the sampling of SBC and the annotation of the SBC subset texts, concluding with a remark on how the SBC subset results are treated in the study.

In section 2.3 of the present study, Biber's (1988) MD methodology for studying textual variation was introduced. As mentioned there, Biber (1988) discovered six dimensions of variation among written and spoken texts and positioned six genres of speech and 17 genres of writing on each of them. The present study intends to position the two UCOW genres of conversational writing on the same dimensions. By using the established positions of Biber's genres on the dimensions, especially those of oral conversations (face-to-face and telephone conversations), it should be possible to determine the level of orality in conversational writing, i.e. its similarity to oral conversations. The positions of the conversational writing genres will also help to address, for instance, the two hypotheses stated in section 1.2, suggesting different levels of orality in SCMC and SSCMC. Throughout this study, samples from conversational writing will be contrasted with textual samples of spoken conversations, as well as with samples from other genres, to exemplify for instance the distribution of lexico-grammatical features. The spoken genres in Biber's (1988) study derive from LLC and the written genres from LOB, as well as two collections of letters (see Appendix I for a list of genres studied in Biber 1988). As the conversational genres are of particular interest

in the present study, the comparability of UCOW to LLC conversations is an important concern.

The conversations in LLC are face-to-face and telephone conversations recorded among speakers of British English in the 1960s and 1970s, most of whom were academics (Greenbaum & Svartvik 1990). UCOW, as described in the two sections above, was recorded among random chatters in various chat channels, and among high school seniors in private conversations, in the 2000s. The ages of the IRC chatters are unknown, and their varieties of English are unpredictable (ranging from the "global" English of EFL speakers, to the subtle regional variants of native speakers). The split-window ICQ chatters, however, are a fairly homogeneous group of adolescent American English speakers (most from middle-class suburban neighborhoods). To improve the quality of comparisons between UCOW and oral conversations it was thus desirable to study a corpus of spoken American English, alongside LLC, preferably one recently collected. The corpus opted for, to fulfill these requirements, was Part 1 of the Santa Barbara Corpus of Spoken American English, here SBC, recorded in the late 1980s to mid-1990s (Du Bois et al. 2000).[45] In addition to being regionally and temporally closer to UCOW, SBC also represents the spoken conversations of "a wide variety of people of different regional origins, ages [inter alia teenagers], occupations, genders, and ethnic and social backgrounds"[46] and is thus possibly socially more diversified than LLC.

SBC part 1 was released in 2000 and consists of 14 face-to-face conversations. To limit the burden of annotation in the present study, it was decided that the SBC conversations be sampled to obtain a corpus of a size similar to the UCOW components, i.e. approximately 10,000 words. Consequently, the first 712 words (on average) from each of the 14 conversations were sampled, to form an SBC part 1 subset corpus; see table 3.1. This SBC subset was first purged of its time-stamps and stripped of its original prosodic mark-up, before it was annotated with the tags used in the present project; cf. section 3.2.[47]

Compared to the collection, adaptation and annotation of the two UCOW components (sections 3.2 and 3.3), the sampling, adaptation and annotation of the SBC subset was a fairly straightforward task. Unlike the UCOW texts, the

45 The Santa Barbara Corpus of Spoken American English is currently part of the International Corpus of English (ICE) project, directed by Gerald Nelson; see <http://ice-corpora.net/ice/index.htm> (accessed 2015-10-13).

46 Cf. the Santa Barbara Corpus web site <http://www.linguistics.ucsb.edu/research/santa-barbara-corpus> (accessed 2015–10–13).

47 The SBC texts were obtained prior to their part-of-speech mark-up in the ICE project.

texts of the SBC subset have regular spelling and punctuation and lack abbreviations denoting several words, which make them comparatively easy to annotate. The few foreign language turns found (one of which is exemplified in Appendix IV) were removed from the subset, but no other purging of the raw texts was needed. Example (8), the first ten turns of "face-to-face conversations SBC" text 2, shows how SBC subset texts will be presented in this study. Speaker names indicating turns, e.g. "Jamie" and "Harold" in (8), are carried over from the original corpus, but were naturally not subject to annotation.

(8)
Jamie:	How can you teach a three-year-old to tap dance.
Harold:	I can't imagine teaching a
Jamie:	Yeah,
	really.
Miles:	Who suggested this to em.
Harold:	I have no idea.
	It was probably my sister-in-law's idea because,
	I think they saw that movie.
Jamie:	Tap?
Harold:	What was the,
Miles:	They had
Harold:	the movie with that really hot tap dancer.
Jamie:	Oh that kid.

Face-to-face conversations SBC text 2

The annotation and summing-up of Biber's (1988) features in the SBC subset followed the same procedure as for the UCOW texts; see sections 3.2 and 3.3. The resulting normalized frequencies of the features in the individual SBC subset texts, as well as the TTR and word length of texts, are found in Appendix II table 7 and the frequencies, TTR and word length of the whole SBC subset are found in Appendix II table 3. The lexical density of the SBC subset was also computed, for the eventual comparison with the UCOW genres (to be explored in section 4.5). Unlike for the UCOW components, however, the lexical density calculation for the SBC subset required no tokenization of the texts (cf. section 3.2) beyond the tokenization of contractions. The regular orthography of the linguist-transcribed spoken texts means that no words have been accidentally split-up or conjoined, and in the spoken texts no words are "hidden" in abbreviations.

Crucially, the SBC subset corpus provides a supplementary point of reference, besides LLC, for comparisons between conversational writing and face-to-face conversations in the present study. The SBC subset supplements LLC in three valuable ways: it is regionally comparable with the split-window ICQ chats, as both represent American English; the SBC subset is also temporally adjacent to

UCOW, as the two were recorded in successive decades; and, finally, the SBC subset possibly represents the English of a socially more diverse set of speakers, including adolescents, making it slightly better suited than LLC for comparisons with UCOW. This is not to say that LLC is ruled out in the analyses to come. On the contrary, LLC feature counts will be referred to on a regular basis, as they are integral to Biber's (1988) investigation. Textual examples of face-to face conversations, however, will mostly be drawn from the SBC subset. Moreover, "face-to-face conversations LLC" and "face-to-face conversations SBC" will be positioned as separate genres on Biber's dimensions (in chapter 5).

Spoken English, nonetheless, consists of more than face-to-face conversations. Besides face-to-face-conversations, LLC contains texts from telephone conversations, interviews, broadcasts, spontaneous speeches and prepared speeches; see Appendix I. Of the six genres in LLC, the two conversational genres (face-to-face conversations and telephone conversations) are of primary concern in the forthcoming comparisons with conversational writing. The vast majority of LLC's telephone conversations were recorded in the 1970s (only five are from the 1960s), which may put this genre in less urgent need of supplementing with updated corpus data than the LLC face-to-face genre (which has only slightly more texts from the 1970s than from the 1960s). Even so, a newer telephone conversations corpus admittedly would have been desirable. Annotating such a corpus, however, was beyond the time scope of the study. Consequently, no corpus of telephone conversations, beyond LLC, was sampled or otherwise brought into the present study.

One final remark needs to be made here in connection with LLC and the SBC subset. As mentioned in chapter 1, the present study intends to contrast the distribution of salient features in conversational writing with the distribution of the same features in speech, writing and ACMC. When it comes to the medium of "speech," the SBC subset will be merged with the LLC genres to constitute a uniform point of reference. Accordingly, in chapter 4, the medium of speech is represented by LLC's six genres of speech (cf. Appendix I) combined with the SBC subset face-to-face genre. Just how average figures for this combined set of "speech" were obtained will be further described in section 3.6. This chapter now turns to a description of how Biber's (1988) MD methodology was applied to the feature count data from UCOW and the SBC subset to obtain standardized scores and, eventually, dimension scores for the genres under study.

3.5 Standardization and dimension score computation

This section follows up on section 2.3, in which Biber's (1988) methodology for computing dimension scores for the genres of speech and writing was described. As outlined there, a complete MD analysis involves eight methodological steps. Of the eight steps outlined in section 2.3, the present study has, by this point, implemented steps 1–5. Three corpora have been "designed" in the present study, namely the two UCOW components "Internet relay chat" and "split-window ICQ chat," as well as the "SBC subset" (step 1). The linguistic features chosen for the study are the same as those identified in Biber's (1988) original study, i.e. the 67 features listed in table 2.1 (step 2). The three corpora have been manually tagged for their occurrences of all of these features, and the average TTRs and word lengths of all texts have been computed (steps 3 and 4), and finally, frequency counts have been computed and compiled into tables in Appendix II, along with TTRs and word lengths (step 5). As the present study relies on Biber's pre-defined dimensions, it leaves out steps 6 and 7 of a complete MD analysis. This means that only step 8, the final step, remains, in which dimension scores are computed for the texts/genres on each dimension. As mentioned in section 2.3, however, the present investigation lingers in step 5 for a while, devoting considerable space to the discussion of salient results obtained. The present section explains what this means, before homing in on the dimension score calculations.

To understand the calculations to be surveyed here, readers are advised to first review the tables in Appendix II, which are central to most results to be presented in this study. Appendix II contains seven tables. Tables 1–3 sum up the normalized feature counts, i.e. the descriptive statistics, for each of the corpora annotated in the present study: table 1 for Internet relay chat, table 2 for split-window ICQ chat and table 3 for the SBC subset. Tables 5–7, furthermore, present the equivalent normalized frequency counts for each text in the three corpora: table 5 for the IRC texts, table 6 for the split-window chats and table 7 for the SBC subset texts. The tables mentioned have all been introduced in the sections above. Now, readers are recommended to turn to Appendix II table 4.

Appendix II table 4 constitutes the zero point (i.e. the baseline) for the standardization of frequencies and for the dimension score calculations, in Biber's (1988) study as well as in the present one. The table, drawn from Biber (1988: 77–78), gives the normalized frequencies of the features in Biber's *full* corpus of speech and writing, that is, the average figures for *all* of the LLC and LOB texts, as well as the two collections of letters (cf. Appendix I). The table thus forms the backdrop against which the individual texts and the individual genres in Biber (1988) were measured, as well as those studied in other MD analyses following

Biber (1988) (e.g. in Conrad & Biber 2001, as exemplified in section 2.2), and now it constitutes the baseline for calculations in the present study. To obtain standardized scores for individual features, and eventually dimension scores for individual texts and genres, the features, texts, and genres are all contrasted with Appendix II table 4.

As touched upon in section 2.3, the normalized frequencies are first contrasted with Biber's full corpus mean, i.e. the left-most column in table 4, and then with Biber's full corpus standard deviations, the rightmost column in table 4, to obtain standardized scores, henceforth "standard scores," for the features. More specifically, a standard score for a feature is obtained by performing the following calculation.

$$\text{standard score} = \frac{(\text{frequency in text} - \text{mean frequency in Biber's full corpus})}{(\text{standard deviation in Biber's full corpus})}$$

For example, consider first person pronouns (feature 6) in Internet relay chat text 1a (Appendix II table 5). The text has a normalized frequency of 48.0 first person pronouns. By applying the above calculation, the standard score arrived at for first person pronouns in IRC text 1a is 0.8; that is, the frequency of first person pronouns in IRC text 1a is 0.8 standard deviations higher than Biber's (1988) full corpus mean. Once the standard scores for first person pronouns have been computed for all the IRC texts, the mean of these, viz. 1.1, may be taken to be the standard score for first person pronouns in the whole genre of "Internet relay chat."

Next, consider the same feature in split-window ICQ chat text 1 (Appendix II table 6). First person pronouns appear to be more common in split-window chats than in IRC. Accordingly, by the above calculation, the normalized frequency of 110.5 first person pronouns yields the standard score 3.2 for the feature in split-window ICQ text 1. The average standard score arrived at for first person pronouns in the whole genre of split-window ICQ chat, once computed, is 2.4. In other words, first person pronouns are more than two standard deviations more frequent in split-window ICQ chat than in Biber's corpus as a whole. Features with absolute standard scores above 2.0 deviate markedly from Biber's mean for speech and writing, i.e. they are markedly frequent. First person pronouns may, consequently, be regarded as one of the most salient features in conversational writing, by virtue of their high frequency in split-window ICQ chat (despite their not being as salient in IRC).

Chapter 4 of the present study explores the features that deviate from Biber's mean by more than two standard deviations ($|s.d.|>2.0$) in either, or both, of the

conversational writing genres.[48] These features are seen to characterize conversational writing by their high relative frequency, or relative infrequency, in the chats and are the most influential ("most salient") contributors to the dimension score(s) of the relevant chat genre(s). As mentioned in section 2.3, chapter 4 also considers other salient features of conversational writing, those studied in previous accounts of CMC discourse (e.g. modal auxiliaries, paralinguistic features, emoticons and abbreviations) as well as previously understudied aspects of conversational writing, such as its lexical density and inserts. *That* is how the present study "lingers" in step 5 of the MD methodology (cf. section 2.3), before moving on to present the results of the final step in the MD methodology, the dimensions scores of the new genres, in chapter 5.

The computation of dimension scores for the new genres "Internet relay chat," "split-window ICQ chat" and "face-to-face conversations SBC" followed the procedure described for Biber's (1988) genres in section 2.3. A genre's dimension score is found by averaging the dimension scores for all texts in a genre. This means that dimension scores are first computed for the individual texts. As mentioned in section 2.3, the dimension score of a text is computed by summing the standard scores for the features co-occurring on the dimension, except on Dimensions 1 and 3, on which the sum of the "negative" features' standard scores is subtracted from the sum of the "positive" features' standard scores (cf. table 2.2). Section 2.3 exemplified the dimension score calculation for a general fiction text on Dimension 2 (as explained in Biber 1988: 94–95). The present section briefly considers Internet relay chat text 1a on the same dimension, to further illustrate the procedure.

The dimension score for IRC text 1a is calculated by summing the standard scores of the features co-occurring on Dimension 2 (cf. table 2.2). In the chatted text, the features display a much lower incidence than in the general fiction text exemplified in section 2.3, an incidence generally even lower than the mean for Biber's full corpus. The standard scores for the features on Dimension 2 in IRC text 1a are -0.7 for past tense verbs, -0.8 for third person pronouns, -1.3 for perfect aspect verbs, -1.1 for public verbs, 0.2 for synthetic negation and -0.6 for present participial clauses, the negative numbers indicating that the features are rarer than in Biber's mean. The resulting dimension score for the text is thus -4.2,[49] which reflects the sparsity of the features on this dimension. While a high incidence of Dimension 2 features indicates a narrative concern in a text, as in

48 Appendix V lists the features with a |standard score| >2.0 in the genres studied.

49 The value -4.2 is the sum of unrounded standard scores.

the general fiction text exemplified in section 2.3, a low incidence indicates a that a text is unmarked for narrative concern. As it turns out, most IRC texts, like text 1a, display a considerable paucity of the lexico-grammatical markers of narration co-occurring on Dimension 2. The average dimension score for the genre, consequently, turns out to be very low, positioning Internet relay chat on the non-narrative extreme of the dimension scale, opposite to Biber's (1988) fiction texts. The dimension scores of the conversational writing genres, as well as those of the SBC subset, will be further presented and discussed alongside Biber's (1988) genres, on all dimensions, in chapter 5.

As astute readers may have noticed upon review of the tables in Appendix II, a dimension score for a genre may equally well be computed directly by summing the standardized scores for features in the whole genre (e.g. for IRC, those computed by contrasting Appendix II table 1 with table 4), without considering the dimension scores of the individual texts. This is feasible because the descriptive statistics for the whole genre (cf. Appendix II table 1) in reality simply reflect the average frequencies of the genre's constituent texts (cf. Appendix II table 5). The roundabout way for computing dimension scores (via individual texts) was, nevertheless, taken in the present study for the sake of adherence to Biber's (1988: 94–95) description of the procedure. Moreover, besides presenting the dimension scores for the genres, chapter 5 will also present the spread of dimension scores across the individual texts (e.g. as minimum and maximum values), results that inevitably rely on the computation of dimension scores for individual texts. Lastly, statistical tests also rely on the availability of such scores.

3.6 Average figures for writing and speech, respectively

As mentioned in chapter 1, the primary purpose of the first results chapter (chapter 4) is to document the features that are salient in conversational writing. The chapter expounds on the incidence in SCMC and SSCMC of such features and contrasts this with the distribution of the same features in writing, speech and ACMC, i.e. at the level of medium (cf. section 1.2). The media to be contrasted in chapter 4, as described, are "writing," "speech," "ACMC," "SCMC" and "SSCMC" (even though the latter three, of course, comprise only one prototypical genre each, BBS conferencing, Internet relay chat and split-window ICQ chat). For all the features to be investigated, normalized frequencies will be contrasted, rather than standard scores. The most salient features, i.e. those that in conversational writing deviate by more than two standard deviations ($|s.d.|>2.0$) from Biber's (1988) mean, will be treated in sections 4.2 (viz. personal pronouns) and 4.4 (the other most salient features); section 4.1 explains why they are presented in

separate sections. Chapter 4 also considers several other Biber (1988) features, as well as features typical of conversational writing that were not studied in Biber (1988). Whenever possible, the quantitative findings are contrasted with the quantitative findings in Collot's (1991) corpus of ACMC, as well as with the findings for writing and speech, respectively. The figures for ACMC are derived directly from Collot (1991), and the figures for SCMC and SSCMC derive from the feature counts in this investigation. The present section is dedicated to describing how average figures for the media "writing" and "speech" were obtained.

In the section above (3.5), Appendix II table 4 was seen to represent the "zero point" for comparisons of all genres, as it summarizes the mean frequencies for the features in Biber's (1988) full corpus of written and spoken texts. Appendix II table 4 (from Biber 1988: 77–78) can thus be regarded as representing the mean frequencies of the features in the English language overall. In the present study, by analogy, the written genres studied in Biber (1988) are taken to represent the medium of writing. Average figures for writing, accordingly, were obtained by considering the average normalized frequencies for the written texts studied by Biber (1988). Biber (1988: 247–263) details the average normalized frequencies for the 17 genres of writing (to save space, Biber's tables are not reproduced here, although Appendix I presents a list of the genres). The following algorithm was employed to compute the normalized frequency in a medium.

$$\text{normalized frequency in medium} = \frac{\sum((\text{normalized frequency in genre}) \times (\text{no. of texts in genre}))}{\sum(\text{no. of texts in genre})}$$

To exemplify the computation, we will consider the occurrence of first person pronouns in "writing." As indicated in Biber (1988: 247–263), the genre "press reportage" has a normalized frequency of 9.5 first person pronouns, "press editorials" 11.2, "press reviews" 7.5, etc. As seen in Appendix I here, Biber's (1988) written corpus consists of 44 texts of press reportage, 27 texts of press editorials, 17 texts of press reviews, etc. The normalized frequencies are occurrences per 1,000 words of running text and, for the current purpose, each text in a genre may be seen to consist of the average normalized frequency for a feature in the genre. To the full "corpus" to represent the normalized frequency in writing, press reportage thus contributes 44 × 9.5 first person pronouns, as the genre consists of 44 texts, each containing on average 9.5 first person pronouns. The genre of press editorials, by inference, contributes 27 × 11.2 first person pronouns, as each text contains on average 11.2 first person pronouns. By summing the first person pronouns calculated thus in all the genres of writing, and dividing the sum by

the total number of written texts (i.e. 340; cf. Appendix I), the average normalized figure for first person pronouns in writing is obtained, that is 17.0.[50]

Average normalized frequencies for all of Biber's (1988) features to represent "writing" here were obtained by the same procedure, even for TTR and word length. As mentioned, the normalized frequencies for salient features will be presented and discussed in chapter 4. For TTR, the presentation of average figures in chapter 4 will be complemented with standard deviations. The standard deviation for TTR in writing was computed by considering the standard deviations for TTR in Biber's individual genres, given in connection with the normalized frequencies in Biber (1988: 247–263). The calculation was carried out by applying the following equation, in which x is the TTR in each genre and n is the number of texts involved.

$$\text{TTR standard deviation in medium} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}}$$

Turning now to the medium of "speech," readers may recall (from section 3.4) that the medium of speech in the present study is represented by Biber's six genres of speech, deriving from LLC, combined with the SBC subset face-to-face conversations genre. Average normalized frequencies for the features in Biber's (1988) spoken genres are presented in Biber (1988: 264–269), and Appendix I here lists the number of texts in each of Biber's spoken genres. The normalized frequencies for the same features in the SBC subset of face-to-face conversations are given in Appendix II table 3. In the present study, the average normalized frequencies for Biber features in "speech" were computed by considering the distribution in a combined "corpus" of speech consisting of the 141 texts from LLC studied by Biber (Appendix I), and the 14 texts from the SBC subset, once more applying the algorithm for "normalized frequency in medium" given above.

As an example for "speech," consider again the distribution of first person pronouns. Biber (1988: 264–269) tabulates the mean normalized frequencies for first person pronouns in the LLC genres studied. In "face-to-face conversations" (LLC), there are 57.9; in "telephone conversations," 70.7; "interviews," 50.5; "broadcasts," 11.8; "spontaneous speeches," 60.4; and in "prepared speeches,"

50 This method draws upon Biber's (1988) procedure for computing the frequencies in the full corpus of writing and speech (1988: 77–78), the results of which (see Appendix II table 4 here) appear to average the normalized frequencies in all the individual genres (1988: 246–269), for each genre taking into account its number of texts.

41.8. Appendix II table 3 in the present study correspondingly presents the normalized frequency for first person pronouns (feature no. 6) in the "face-to-face conversations SBC" genre, viz. 61.0. As mentioned, Biber's LLC spoken genres comprise a total of 141 texts (cf. Appendix I) and the SBC subset consists of 14 texts (cf. table 3.1). In both corpora, each text in a genre may be considered to contain the normalized frequency for the feature in the genre. To the full "corpus" to represent speech in the present study, face-to-face conversations in LLC accordingly contribute 44 × 57.9 first person pronouns, telephone conversations contribute 27 × 70.7, interviews 22 × 50.5, etc., and, finally, face-to-face conversations from the SBC subset contribute 14 × 61.0 first person pronouns. After all these contributions are summed, the sum is divided by the total number of texts (i.e. 155, cf. Appendix I and table 3.1) to obtain the average normalized figure for first person pronouns in speech, that is, 52.8. The same procedure was then used to compute average frequencies in speech for all of Biber's features, as well as the average figures for speech as regards TTR and word length. The TTR standard deviation for speech was calculated by using the formula "TTR standard deviation in medium" above, as explained for "writing."

In sum, the normalized frequencies, TTR and word length for writing to be presented in chapter 4 represent Biber's (1988) 17 genres of writing, and the normalized frequencies, TTR and word length for speech represent Biber's (1988) six genres of speech, supplemented with the SBC subset face-to-face conversations genre. In the comparisons between conversational writing and speech, however, recurring reference will be made not just to the whole "corpus" of speech (Biber's six genres + the SBC subset), but also, more importantly, to the conversational genres it contains. Whereas tables and figures in chapter 4 by default present the average figures for "speech" overall, explanations and discussions frequently indicate its constituent average figures for the individual conversational genres, that is, for face-to-face conversations and telephone conversations from LLC, and for face-to-face conversations from the SBC subset. The present study, after all, is concerned not just with the comparison of conversational writing to writing, ACMC and speech, but also, more specifically, with the similarities, or differences, between conversational writing and the spoken conversational genres.

3.7 Chapter summary

Chapter 3 has outlined the methodology for obtaining the data to be investigated in the present study – from the collection of the textual material to the quantitative results. After presenting the size of the three corpora compiled for the study, each corpus was treated in a separate section. First, the recording, adaptation and

annotation of the Internet relay chat corpus was described; second, the collection and annotation of the split-window ICQ chat corpus was explored; and third, the sampling of SBC was motivated and explained, as well as the annotation of the resulting SBC subset corpus of face-to-face conversations. For each corpus, two important tables in Appendix II were highlighted – one summarizing the normalized frequency counts for Biber's 67 features in the corpus and one detailing the normalized frequencies in individual texts. Next, the chapter described the process of standardizing the normalized frequencies. The standard scores then provided the requisite input for computing dimension scores for the genres under investigation, dimension scores that will be presented and discussed alongside textual examples in chapter 5. The penultimate section, finally, explained how average figures for the media writing and speech, respectively, were computed, for comparisons with the CMC media in the ensuing chapter. Chapter 4 is now at hand, which will present the salient features found in conversational writing.