Formality in Digital Discourse: A Study of Hedging in CANELC

Dawn Knight, Svenja Adolphs, and Ronald Carter

1 Introduction

Technology has transformed the way we communicate in the modern digital age. No longer do we simply rely on speech and writing but also on a range of different forms of 'e-language'. E-language is defined here as any communicative, interactive and/or linguistic stimulus that is digitally based and 'incorporates multiple forms of media bridging the physical and digital' (Boyd and Heer 2006: 1): from e-mails to discussion board threads, SMS messages and so on ('e-language' is also known as Computer Mediated Communication, CMC: see Walther 1996; Garcia and Jacobs 1999; Herring 1999 and Thurlow et al. 2004, and 'netspeak', Crystal 2003: 17). As a relatively new 'genre' of communication (Herring 2002), the definition and description of the features of e-language and how it compares and contrasts with spoken and written genres of communication is an on-going concern in studies of CMC, Applied Linguistics, Corpus Linguistics and beyond. This is something that will be examined in more detail in the current chapter.

Based on Crystal (2003: 17), there is a suggestion that spoken and written language effectively exist on a 'continuum' of formality (also see Condon and Cech 1996; Ko 1996; Herring 2007 for further discussions on the differences between spoken and written discourse). The 'more' formal language structures exist on the

D. Knight (✉)
School of Education, Communication and Language Sciences,
Newcastle University, Newcastle NE1 7RU, UK
e-mail: Dawn.Knight@ncl.ac.uk

S. Adolphs • R. Carter
School of English Studies, The University of Nottingham,
University Park, Nottingham NG7 2RD, UK

J. Romero-Trillo (ed.), *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies*, Yearbook of Corpus Linguistics and Pragmatics 1,
DOI 10.1007/978-94-007-6250-3_7, © Springer Science+Business Media Dordrecht 2013

left of the continuum, where written language is conventionally positioned, and the least formal exists towards the right end of the continuum, where spoken language is conventionally perceived to be positioned (although obviously their positioning is somewhat fluid as no absolute positioning in this abstract notion can ever exist – it is a theoretical continuum not a static classification system).

Considered as a distinct genre of communication, Crystal suggests that 'netspeak' is perhaps somewhere in the middle, between spoken and written language (2003: 17). He suggests that there is essentially a blurring of traditional characteristics of spoken and written language, in digital communication, making it a combination of both of the more 'traditional' genres (also Biber 1993; Collot and Belmore 1996; Yates 1996; Crystal 2001 for further discussion). Others have added to this notion, instead suggesting that each e-language 'mode' (Murray 1988) is structurally, semantically and pragmatically different from one another as well as spoken and written language types, making their relative positioning along this continuum of formality highly variable (see Murray 1988; Baym 1995; Cherny 1999; Herring 1996).

Levels of formality in *specific* modes of e-language have already received attention from researchers (see works by Sutherland 2002; Hard af Segersteg 2002; Shortis 2007; Crystal 2008 for further details). For example, Tagg (2009) and Ling (2003) both report on the tendency for SMS messages to be immediate and personal, written in the first person and directed to specific recipients. Tagg adds to this, underlining that 'the informal and intimate nature of texting encourages the use of speech-like language' in this e-language mode (2009: 17, also see Crystal 2003; Oksman and Turtianen 2004). Similarly, Baron highlights that although email, as with texting and other common forms of e-language, is typed or 'written' rather than spoken, 'participants exploit it for typically spoken purposes' (1998: 36), and it therefore shares more similarities with communication situated at the spoken rather than written end of the continuum.

Levels of formality across e-language as a specific *genre* and the relationships that exist between individual *modes*, however, is something that remains under-explored in corpus-based analyses of real-life data. Initial developments in this area of research have been made by Knight et al. (forthcoming, 2012) who provided some preliminary observations about the frequency of pronouns and deictic markers in e-language, compared to written and spoken excerpts from the BNC.[1] This study is extended in the present chapter but with a focus, instead, on the use of forms of hedging in e-language. The corpus used in this chapter is CANELC, the Cambridge and Nottingham e-language Corpus, a one-million-word corpus of digital discourse taken from British contributors or those posting to British websites in 2010–2011. It includes data from discussion boards, blogs, tweets, emails and SMS messages, distributed according to Fig. 1 (word counts for each mode are included in this figure).

[1] The British National Corpus, BNC, is a 100 million word corpus of written and spoken discourse in English. For more information see: http://www.natcorp.ox.ac.uk/

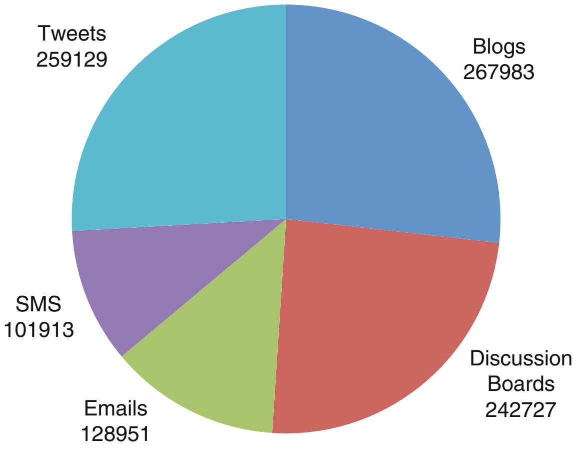


Fig. 1 The contents of the CANELC corpus

CANELC was built to allow for the querying of data at the general level of the *genre* of interaction as well as at the level of individual the communicative *mode*. So, using results from corpus-pragmatic based enquiries of CANELC, we will aim to create a deeper understanding of how different modes of e-language relate to Crystal's notion of the 'continuum' of formality.

2 Corpus Pragmatics

2.1 Overview

The study of the pragmatics of language use has traditionally concentrated on spoken registers rather than written language because the latter tends to be 'referentially explicit' (McEnery et al. 2006: 104) while the former allows for a more 'extensive reference to the physical and temporal situation of discourse' (Biber 1988: 144) in the construction of meaning. Spoken interaction is, in other words, highly context specific, and meaning is not only determined by the specific spoken or written 'sign' (Morris 1946: 287) used, but by a range of other 'extrinsic'; 'social, cultural and interactive' factors, and 'intrinsic', 'cognitive, affective and conative' factors that exist (Kopytko 2003: 45; also see Labov 1972; van Dijk 1977; Duranti and Goodwin 1992; Eckert and Rickford 2001; Fetzer 2004, for further discussion on language and context).

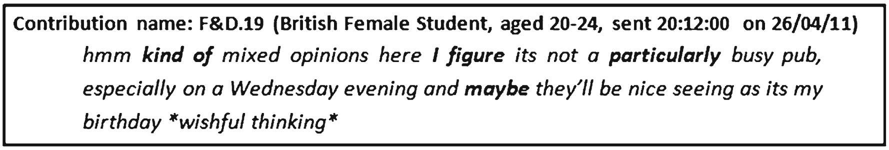
There is no one-to-one relationship between language form and function as the interpretation of a given message is highly dependent on the communicative function of a word or utterance, in a specific discursive context (for discussions of language and context see Labov 1972; Bates 1976; Nelson et al. 1985; Brown 1989; Halliday and Hasan 1989; Duranti and Goodwin 1992; Widdowson 1998; Green 2002; Scollon and Scollon 2003). In spoken communication, much of the discursive context is 'shared' (McEnery et al. 2006: 105) between a speaker and an interlocutor.

This affects the type of language used as there is a temporal and/or physical closeness in spoken discourse between the individuals as well as a shared knowledge about the immediate communicative context. This provides a 'clear advantage in using contextual expressions such as *I*, *there*, or *now*, [for example,] which are shorter and more direct' (Heylighen and Dewaele 2002: 301). Depending on the relationship and social distance between the speaker and interlocutor, speakers can thus use less formal expressions and a larger number of pronouns and deictic markers in this shared communicative space (see Fowler and Kress 1979; Chafe and Danielewicz 1987; Biber 1992; Biber et al. 1999; Leech 2000; Carter and McCarthy 2006; Atkins 2011). There is more of a gulf in spatial distance and time between writers and readers of written texts as there is no guarantee of when a text may be read or by whom. Written texts are not as contextually bound and thus often lack the shared knowledge and understanding between writer and reader, which often correlates with a decrease in the use of contextual (deictic) expressions in these texts.

While not necessarily true of all forms of e-language (instant messaging, IM, for example), the different modes of data included in CANELC are somewhat similar to one another in the fact that they do not 'require that users be logged on at the same time in order to send and receive messages' (Herring 2007: 13). The content sent via these different modes are 'stored at the addressee's site until they can be read' by the recipient (Herring 2007: 13). They are not forms of communication which necessarily require an instant response as, again, IMs do and face-to-face (spoken) interaction does. They are, therefore, asynchronous (for more detailed discussion of synchronicity see Condon and Cech 1996; Ko 1996; Herring 2007).

This asynchronicity means that the data in CANELC is arguably structurally organised in a way that is more consistent with written than spoken language (which is also asynchronous). It is interesting, then, to note that it is actually often the case that only a few seconds or minutes passes between the time when a message is sent and attended to across different e-language modes, despite this asynchronicity. There may in fact only be a short delay between the time a message is composed and read/responded to (although there is likely to be some inconsistency in the average time taken across the different modes of e-language). This is likely to reduce the temporal and social distance between sender and receiver as highly context-specific information about the message (related to time) is more likely to be shared and understood.

As a consequence of this, as outlined in Knight et al. (forthcoming, 2012), there is often a frequent use of 'temporal referents….deictic marking (as with the prolific use of personal pronouns)' in e-language. These discursive features again hint at forms of communication that are potentially allowing for an immediate or near-immediate information exchange, a forum for communicating reports of events and incidents in near real-time, as the understanding of the temporal referent is shared'. There is a shared digital space rather than physical space, within which 'the social, physical and temporal context is frequently changeable' (Knight et al. forthcoming, 2012). This is contrary to what is expected from asynchronous



Contribution name: F&D.19 (British Female Student, aged 20-24, sent 20:12:00 on 26/04/11)

*hmm **kind of** mixed opinions here **I figure** its not a **particularly** busy pub, especially on a Wednesday evening and **maybe** they'll be nice seeing as its my birthday *wishful thinking**

Fig. 2 An example of hedging, taken from the discussion board data in CANELC

communicating, aligning e-language more closely to more informal, spoken discourse, despite the fact it is not synchronous and is typed/written rather than spoken.

2.2 Hedging

In addition to pronouns and deictic markers, another pervasive feature that relates to levels of formality in discourse is the use of hedging (first coined by Lakoff 1972: 195). In pragmatics, hedges are 'expression[s] of tentativeness and possibility' (Hyland 1996: 433) which operate to 'mitigate the directness of what we say and so operate as face-saving devices' (O'Keeffe et al. 2007: 174 – for more information on politeness theory and the notion of 'face', see Brown and Levinson 1978, 1987). They are 'pragmatic markers' (Carter and McCarthy 2006: 223) which can be used 'to downtone…..the force of an utterance for various reasons e.g. politeness, indirectness, vagueness and understatement' (Farr et al. 2004: 13). The specific form, frequency and functions that hedges adopt also 'vary relative to context' (O'Keeffe et al. 2007: 174). Examples of hedging are seen in Fig. 2:

We see the use of four hedges (in bold) in this discussion board thread. The contributor is making plans for her birthday evening, discussing the possibility of inviting a party of friends to a local pub to celebrate. *Kind of* operates as an inexact stance adverb, softening the content of the thread. As with *maybe*, *kind of* acts almost as a 'downtoner', as instead of saying '*it would be nice to go the pub*, *especially since it is my birthday*', the use of this hedge provides an approximate reflection of what the contributor really means (Hübler 1983: 68). *I figure* also functions in a similar way, acting as a verb with a modal meaning, used to soften the meaning of the assumption about the pub, in order to mitigate against a potential face threat for the sender or receiver of the message, while *particularly* also has a similar effect as an omission of the adverb in this context would result in the utterance seeming blunt.

As face-saving devices, 'softeners' (Nikula 1997: 188), the frequent use of hedges is often linked to formal rather than informal contexts of communication (this is true of both spoken and written discourse, but given the tendency for written to be 'more' formal, the level of hedging is generally higher for written discourse vs. spoken discourse). Farr and O'Keeffe's (2002) study of hedging in the spoken

LCIE corpus (Limerick Corpus of Irish English[2]) best illustrates this pattern (2002). In this study, hedges were found to be most frequently used in institutional settings including teacher training contexts and radio discourse, with their use reducing in conversations between family and friends (see Farr et al. 2004) where there 'fixed relationships' (Clancy 2002), a closeness between speakers and listeners (creating less of need for participants to save face). The context where the fewest hedges were used in the corpus was in shop encounters. This is 'perhaps explained by the lesser need to protect face in service encounters, where a customer and a server do not know each other, and where they are interacting within transactional roles' (O'Keeffe et al. 2007: 176). The potential face threat is lower so the use of the mitigating hedging devices is not as essential in such discursive contexts.

Having said this, other studies have suggested that since it is performed in 'real-time' (Leech 2000), spoken 'conversation is [often] more vague than written genres' (McEnery et al. 2006: 105), so an increase in the frequency of certain forms of hedging functioning as vague language markers is often seen. For example, based on queries of the World Edition of the BNC (British National Corpus), Gries and David (2007) discovered that *kind of* and *sort of* were both forms of hedges functioning as vague stance adverbs that are frequently used in spoken discourse, in comparison to written discourse. Although, of these two clusters, *sort of* was significantly more common in written mode than *kind of*, while the reverse was found to be true of the spoken mode. Of written communication specifically, Biber et al. reported that the clusters *kind of* and *sort of* are both used more frequently in formal, academic prose than in other written registers (based on a study of the Longman Spoken and Written English Corpus, 1999: 560–561, other studies of these clusters have been carried out by Crystal and Davy 1975 and Quirk et al. 1985 – comparing their frequency of use between British and American English).

This pattern is inversely true of more private and personal forms of communication as opposed to more public forms (Carter and McCarthy 2006: 9–16). So written interaction, for example, that is most public (professional) and formal in nature (a government policy document for example), will likely see an increase in the number of vague stance adverbs used, when compared to a more personal expression of feelings, for example as this 'softening' function is unlikely to be required with close or intimate relationships.

Numerous other studies have been carried out on hedging in written discourse (Dubois 1987; Channell 1990; Drave 1995; Allison 1995), spoken interaction (see Crystal and Davy 1975; Brown and Yule 1983; McCarthy 1991; Cheng and Warren 1999; Jucker et al. 2003 for examples) and individual modes of e-language including SMS messages (Crystal 2001; Tagg 2009), Blogs (Myers 2010), Instant Messaging (IMs – Brennan and O'Haeri 1999), Discussion Boards (Atkins 2011) and Twitter (Benjamin 2011). More large scale corpus-based, studies have also examined vague language (arguably a sub-set of hedging) in both written and written discourse

[2] The Limerick Corpus of Irish English, LCIE, is a one million word corpus of spoken interaction from a range of different speech genres in Irish.

(Channell 1985, 1994; Kennedy 1987). To date, however, no studies offer an insight into hedging use across these different communicative genres. The current study aims to fill this research 'gap'.

3 Analysis

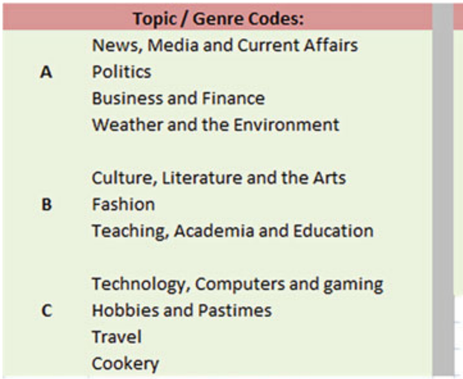
3.1 Study Questions

To build on the foundations of what was previously discovered about levels of formality in e-language (using CANELC – Knight et al. forthcoming, 2012), the following sections focus on the use of hedges in more detail. The analyses address the following research questions:

- Is there a significant difference in the frequency of hedging used:
  - Between all modes of e-language in CANELC, compared with data from the spoken and written BNC?
  - Between the different topic categories of data included in CANELC?
- What do the frequency and use of this phenomenon reveal about the levels of formality within and across the different modes of e-language in CANELC?

To answer these questions, the following sections present results from an analysis the use of hedges in e-language compared to one-million-word samples from the written and spoken BNC samples (which contain 968,267 and 982,712 words respectively). Given that the size of the corpora used are slightly inconsistent, the results are normalised using statistical measures so accurate comparisons can be made. The analyses are conducted out using Rayson's WMatrix software (2003) which includes utilities for carrying out word, cluster and parts of speech queries (centring around the production of key word lists and key-word-in-context, KWIC, outputs), and allows researchers to explore the patterned use of these features in a corpus. With the use of the WMatrix semantic tagger, common themes and semantic associations connected with corpora can also be queried using the software.

In addition to the 'data' taken from communication performed across the different e-language modes, CANELC also contains detailed metadata records: data about the data. Metadata is critical to a corpus as without it 'the investigator has nothing but disconnected words of unknowable provenance or authenticity' (Burnard 2005) to examine. As outlined by Knight (2011: 31, based on Burnard 2005) 'the inclusion of this information assists in identifying the name of the corpus (administrative metadata), who constructed it, and where and when this was completed (editorial metadata), together with details of how components of the corpus have been tagged, classified (descriptive metadata), encoded and analysed (analytic metadata)'. Collectively, this information allows us to reconstruct aspects of the reality of the discursive context in which specific e-language messages were sent, allowing us to



Topic / Genre Codes:	
A	News, Media and Current Affairs
	Politics
	Business and Finance
	Weather and the Environment
B	Culture, Literature and the Arts
	Fashion
	Teaching, Academia and Education
C	Technology, Computers and gaming
	Hobbies and Pastimes
	Travel
	Cookery

Topic / Genre Codes:	
D	Music
	Sports
E	Celebrity news and gossip
	TV
	Humour
F	Health and Beauty
	Parenting and Family Life
	Personal and Daily Life

Fig. 3 Topics featured in CANELC

frame the language in a more contextually accurate way. The following metadata is included in CANELC:

- Author's (and receivers) name, age, gender, nationality
- Date and time composed
- Intended recipient
- Content
- General topic of content
- Follow up comments/responses
- 'Other' relevant information

Regarding 'general topic of content', it is viable to note that in addition to the metadata information, data in CANELC is also broadly categorised by topic. This is based on the schema presented in Fig. 3.

Topics in category 'A' are aligned with more public concerns such as news, politics and current affairs, while those in category 'F' are more aligned with personal issues such as personal and daily life (with B-E existing almost on a continuum between these poles). The distribution of the CANELC data, by number of words, across these different topic categories is represented in Fig. 4.

Figure 4 illustrates that across the entire corpus there is a dominance of contributions in categories 'F' and 'A'. The majority of data in category 'F' is included in the SMS messages and personal emails included in the corpus, which primarily contain language discussing topics concerning aspects of personal and daily life. More public, outward facing, topics such as business, finance and the news are frequently featured in the language of the blogs, tweets and discussion boards, although the tweet and blog sub-corpora have the most balanced distribution of contributions/word count across each of the thematic categories. Finally, CANELC also includes a number of business emails, which contribute to the high frequency of data type 'A'.

While the assignment of the content to these thematic groupings was fairly transparent in some cases, other messages were slightly more 'fuzzy' and flexible, insofar as they discussed multiple topics ranging across the different categories. In these

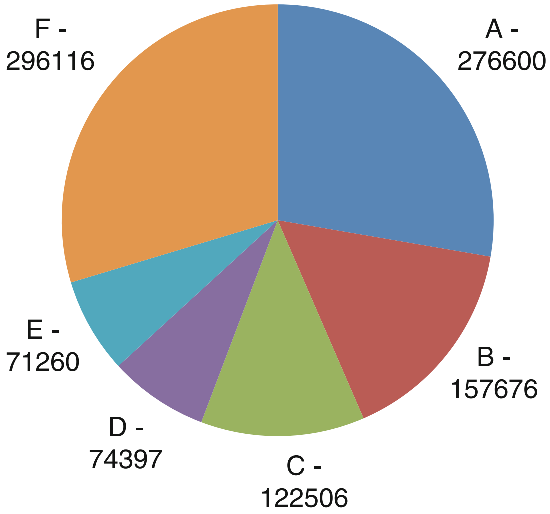


Fig. 4 Approximate distribution of words across the 6 topic categories of CANELC (refer to Fig. 3 for data key)

instances, when compiling CANELC, the data was given a range of category codes, so A/B/C rather than simply 'A'. For the purpose of Fig. 4 and the analysis seen in Sect. 3.3, individual contributions are counted once across these groupings, so they are classified according to, crudely, their 'best fit'. That is, even in instances where multiple categories were assigned, only one single category was counted. This was, subjectively, the category which is descriptively the 'most' appropriate for these contributions, that is, the one that is approximately the most representative/ appropriate of that data. In other words if data was assigned the categories A/B/C, for example, and the content was described as being most dominantly 'business related' [i.e. category A], content was re-labelled as being category 'A' only.

The inclusion of this categorisation scheme provides a helpful way-in to querying levels of formality in CANELC as, in parallel with previous comments, the division of public vs. private can affect the levels of formality in a text. So comparisons of hedging within and across both the modes of data in CANELC and these different topics, can help us to assess how closely e-language compares with more formal (akin to the written end of the continuum) and informal discourse (positioned toward the spoken end of the continuum).

Given the level of contextual specificity, 'hedging can be achieved in indefinite numbers of surface forms' (Brown and Levinson 1987: 146), making it potentially difficult to draw up a 'list of hedges' (Clemen 1997: 236, 243; Nikula 1997: 190) to use as a basis of a study of this phenomenon. Despite this, across the literature there are specific words or expressions that are *often* used as hedges. For example, as outlined by Farr et al. (2004: 13–14) the most salient hedges are 'core modal verbs' and 'verbs with modal meaning' (O'Keeffe et al. 2007: 175 – e.g. *might*, *may*), 'clausal items' (e.g. *I think*, *you know*), 'noun based expressions' (e.g. *the thing is*), 'degree adverbs' (e.g. *really*, *necessarily*) and 'stance adverbs' (e.g. *of course*, *sort of*)

Actually	Generally	Likely	Only	Really	Surely
Apparently	Guess	Maybe	Partially	Relatively	Thing
Arguably	I think	Necessarily	Possibility	Roughly	Typically
Broadly	Just	Normally	Probably	Seemingly	Usually
Frequently	Kind of	Of course	Quite	Sort of	You know

Fig. 5 Some common hedges in spoken and written discourse

and so on. The hedges that the present study will focus on are some of the most common forms that have been examined in past studies of this topic (based on Biber et al. 1999; Carter and McCarthy 2006; O'Keeffe et al. 2007: 175), and are forms which are frequent in the CANCODE[3] (Cambridge and Nottingham Corpus of Discourse in English), BNC, CEC[4] (Cambridge English Corpus) and CANELC corpora. These are listed in Fig. 5. These terms were queried in the CANELC data.

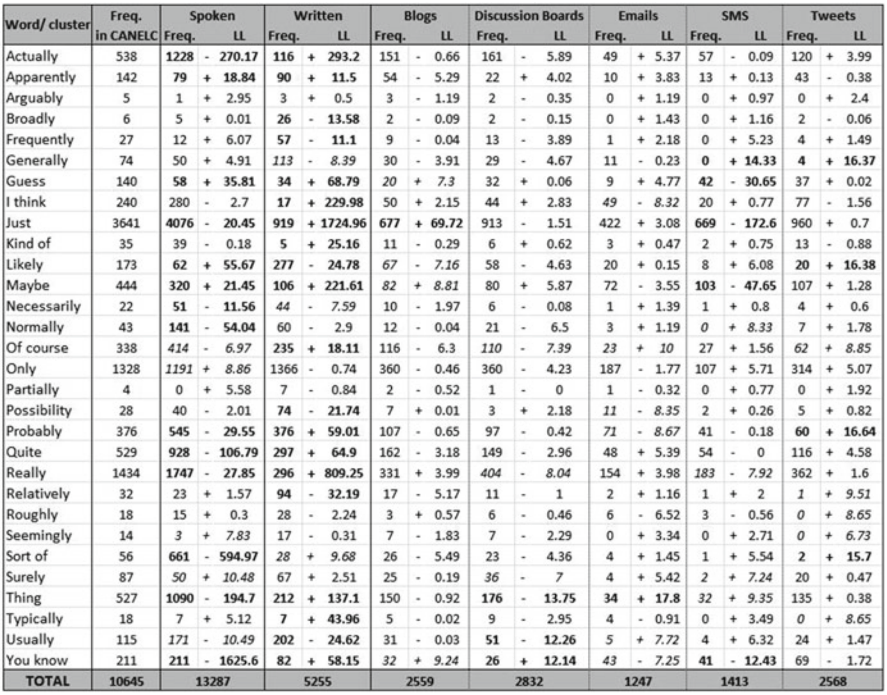
Some of the adverbs listed here, such as *just*, have the softening hedging function, but are also often used with intensifying and specifying functions in discourse. *Just do it*; *it's just about five o'clock* and *we'll only be a couple of minutes late* are examples of this. Of course is another examples of this, this cluster can be used as a hedge when it has a pragmatic function but it can also be emphatically and directly; *Are you coming*? *Of course*. So although we can define some frequent forms of hedges, a more qualitative screen by screen study is needed if we are to drill down into specific functions. The current study undertakes a more quantitative approach, but a more qualitative assessment of the data would be welcomed in future studies of this nature and are, indeed, necessary.

3.2 Frequency of Hedges

The frequency of use of the terms in Fig. 5 were queried across the entire corpus as well as each mode is presented and compared, along with the frequency of use seen in the written and spoken BNC sub-corpora. Results are shown in Fig. 6. Log-likelihood scores are also presented in this figure. These provide a statistical measure of the relationship between the frequencies, indicating whether specific patterns of significant differences are likely to exist by chance or not. In this figure, a '+' log-likelihood score indicates that a particular rate of use is statistically higher in the CANELC corpus compared to the other parameter defined,

[3] CANCODE stands for *Cambridge and Nottingham Corpus of Discourse in English*. This corpus has been built as part of a collaborative project between The University of Nottingham and Cambridge University Press with whom sole copyright resides. CANCODE is comprised of five-million words of (mainly casual) conversation recorded in different contexts across the British Isles.

[4] CEC stands for Cambridge English Corpus, a corpus of over one billion written and spoken words in English. For more information visit: http://www.cambridge.org/



Word/ cluster	Freq.	Spoken			Written			Blogs			Discussion Boards			Emails			SMS			Tweets		
	in CANELC	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL
Actually	538	**1228**	-	**270.17**	**116**	+	**293.2**	151	-	0.66	161	-	5.89	49	+	5.37	57	-	0.09	120	+	3.99
Apparently	142	**79**	+	**18.84**	**90**	+	**11.5**	54	-	5.29	22	+	4.02	10	+	3.83	13	+	0.13	43	-	0.38
Arguably	5	1	+	2.95	3	+	0.5	3	-	1.19	2	-	0.35	0	+	1.19	0	+	0.97	0	+	2.4
Broadly	6	5	+	0.01	**26**	-	**13.58**	2	-	0.09	2	-	0.15	0	+	1.43	0	+	1.16	2	-	0.06
Frequently	27	12	+	6.07	**57**	-	**11.1**	9	-	0.04	13	-	3.89	1	+	2.18	0	+	5.23	4	+	1.49
Generally	74	50	+	4.91	*113*	-	*8.39*	30	-	3.91	29	-	4.67	11	-	0.23	**0**	+	**14.33**	**4**	+	**16.37**
Guess	140	**58**	+	**35.81**	**34**	+	**68.79**	*20*	+	*7.3*	32	+	0.06	9	+	4.77	**42**	-	**30.65**	37	+	0.02
I think	240	280	-	2.7	**17**	+	**229.98**	50	+	2.15	44	+	2.83	*49*	-	*8.32*	20	+	0.77	77	-	1.56
Just	3641	**4076**	-	**20.45**	**919**	+	**1724.96**	**677**	+	**69.72**	913	-	1.51	422	+	3.08	**669**	-	**172.6**	960	+	0.7
Kind of	35	39	-	0.18	**5**	+	**25.16**	11	-	0.29	6	+	0.62	3	+	0.47	2	+	0.75	13	-	0.88
Likely	173	**62**	+	**55.67**	**277**	-	**24.78**	*67*	-	*7.16*	58	-	4.63	20	+	0.15	8	+	6.08	**20**	+	**16.38**
Maybe	444	**320**	+	**21.45**	**106**	+	**221.61**	*82*	+	*8.81*	80	+	5.87	72	-	3.55	**103**	-	**47.65**	107	+	1.28
Necessarily	22	**51**	-	**11.56**	*44*	-	*7.59*	10	-	1.97	6	-	0.08	1	+	1.39	1	+	0.8	4	+	0.6
Normally	43	**141**	-	**54.04**	60	-	2.9	12	-	0.04	21	-	6.5	3	+	1.19	*0*	+	*8.33*	7	+	1.78
Of course	338	*414*	-	*6.97*	**235**	+	**18.11**	116	-	6.3	*110*	-	*7.39*	*23*	+	*10*	27	+	1.56	*62*	+	*8.85*
Only	1328	*1191*	+	*8.86*	1366	-	0.74	360	-	0.46	360	-	4.23	187	-	1.77	107	+	5.71	314	+	5.07
Partially	4	0	+	5.58	7	-	0.84	2	-	0.52	1	-	0	1	-	0.32	0	+	0.77	0	+	1.92
Possibility	28	40	-	2.01	**74**	-	**21.74**	7	+	0.01	3	+	2.18	*11*	-	*8.35*	2	+	0.26	5	+	0.82
Probably	376	**545**	-	**29.55**	**376**	+	**59.01**	107	-	0.65	97	-	0.42	*71*	-	*8.67*	41	-	0.18	**60**	+	**16.64**
Quite	529	**928**	-	**106.79**	**297**	+	**64.9**	162	-	3.18	149	-	2.96	48	+	5.39	54	-	0	116	+	4.58
Really	1434	**1747**	-	**27.85**	**296**	+	**809.25**	331	+	3.99	*404*	-	*8.04*	154	+	3.98	*183*	-	*7.92*	362	+	1.6
Relatively	32	23	+	1.57	**94**	-	**32.19**	17	-	5.17	11	-	1	2	+	1.16	1	+	2	*1*	+	*9.51*
Roughly	18	15	+	0.3	28	-	2.24	3	+	0.57	6	-	0.46	6	-	6.52	3	-	0.56	*0*	+	*8.65*
Seemingly	14	*3*	+	*7.83*	17	-	0.31	7	-	1.83	7	-	2.29	0	+	3.34	0	+	2.71	*0*	+	*6.73*
Sort of	56	**661**	-	**594.97**	*28*	+	*9.68*	26	-	5.49	23	-	4.36	4	+	1.45	1	+	5.54	**2**	+	**15.7**
Surely	87	*50*	+	*10.48*	67	+	2.51	25	-	0.19	*36*	-	*7*	4	+	5.42	*2*	+	*7.24*	20	+	0.47
Thing	527	**1090**	-	**194.7**	**212**	+	**137.1**	150	-	0.92	**176**	-	**13.75**	**34**	+	**17.8**	*32*	+	*9.35*	135	+	0.38
Typically	18	7	+	5.12	**7**	+	**43.96**	5	-	0.02	9	-	2.95	4	-	0.91	0	+	3.49	*0*	+	*8.65*
Usually	115	*171*	-	*10.49*	**202**	-	**24.62**	31	-	0.03	**51**	-	**12.26**	*5*	+	*7.72*	4	+	6.32	24	+	1.47
You know	211	**211**	-	**1625.6**	**82**	+	**58.15**	*32*	+	*9.24*	26	+	**12.14**	*43*	-	*7.25*	**41**	-	**12.43**	69	-	1.72
TOTAL	10645	13287			5255			2559			2832			1247			1413			2568		

Fig. 6 The frequency of common forms of hedges used in CANELC, compared to the spoken and written sub-corpora from the BNC

while a '–' log-likelihood indicates a statistically lower frequency of use in CANELC. Numbers in **bold** indicate that there is a statistical difference (measured using a log-likelihood score) in the frequency of usage across specific modes/ genres to a *p* value of <0.01 (with a critical value range of 6.63–10.82) while those in *italics* mark a significant to *p* value <0.001 (critical value of 10.83). So an '+' indicates an overuse in CANELC compared to the listed parameter and thus an underuse in the given category.

In Fig. 6 we see that, for the terms *actually*, *just*, *you know*, *probably*, *quite*, *really*, *thing*, there is a significant underuse in CANELC compared to the written BNC corpus, while there is a significant overuse compared to the spoken BNC sub-corpus (to $p<0.001$). *Probably* is significantly underused in the twitter data and overused in the email data (to $p<0.01$ and $p<0.001$) while *really* is overused in the discussion boards and SMS messages compared to rate of use across CANELC (to $p<0.001$). *Just* is significantly underused in the blog data and overused in the SMS data, while *you know* is underused in the blog and discussion board data but overused in the email and SMS data and *just* is underused in the email but overused in the discussion board data. Finally, there is no real significant difference in the rate of use of *quite* and *actually* across the different e-language modes.

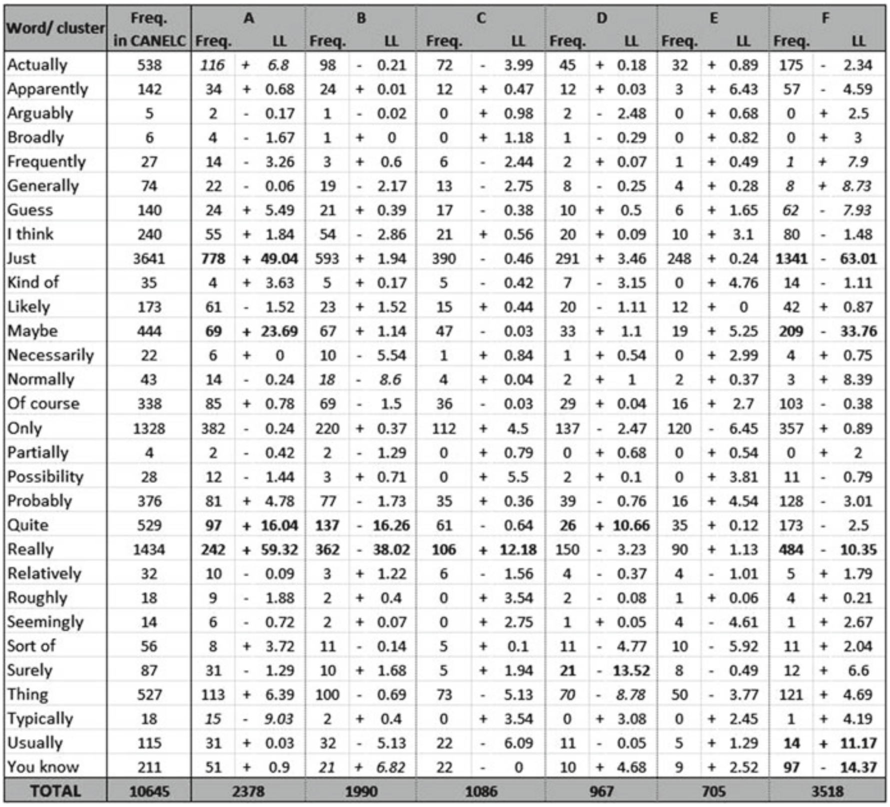
The only item that is significantly overused, at $p < 0.01$, in the spoken BNC **and** underused in the written compared to CANELC is *likely*. There are, however, some terms which are overused in CANELC, compared to both sub-corpora. These include *apparently*, *guess* and *maybe*. Of these terms, *apparently* is used at a near-consistent rate across all of the modes in CANELC, while *guess* is underused (to $p < 0.001$) in the blogs and significantly overused in the SMS (to $p < 0.01$) when compared to the other modes. *Maybe* and *likely*, on the other hand, are both underused in the blogs (to $p < 0.001$ respectively) but the former is overused in the SMS messages and the latter in the tweets (both to $p < 0.01$).

I think, *kind of*, *broadly*, *typically* and, to some extent *of course* are used at a significantly higher rate in CANELC than the written BNC (to $p < 0.01$), but no significant difference exists between the rate that they are used in the spoken BNC (aside from *of course* where the difference is to ($p < 0.001$)). Conversely, there is an underuse of the expression *normally* in CANELC compared to the spoken data (to p<0.01) while there is no significant difference between the use of this term when compared to the written corpus. *Kind of* is used at a consistent rate across all modes in the corpus, while *typically* and *normally* are used at consistent rates across all modes aside from tweets and SMS messages where a slight underuse occurs when compared to CANELC respectively (to $p < 0.001$). Similarly *of course* is slightly underused in the SMS messages but slightly overused in the discussion board data (to $p < 0.001$) and *I think* is slightly overused in the email data, but used consistently across the other modes in CANELC.

Figure 6 also indicates that there is a slight overuse of *only*, *seemingly* and *surely* compared to the spoken BNC (to $p < 0.01$) while no difference exists between the rate of use of these words in CANELC versus the written BNC.

Frequently, *possibility*, *relatively* and, to some extent, *generally* are all underused in CANELC compared to the written BNC, while there is a near-consistent rate of use of these terms when compared to the spoken BNC data (to $p < 0.01$ aside from *generally* which is to $p < 0.001$). The rate at which *frequently* is used across each of the modes in CANELC is near-consistent while there is an overuse of *possibility* in the email data, an underuse of *relatively* in the tweets (both to $p < 0.001$) and a significant underuse of *generally* in the SMS and tweet data (to $p < 0.01$). Similarly, *only* is used at a near-consistent rate across the different modes while *seemingly* is slightly underused in the twitter data and *surely* is underused in the SMS data but overused in the discussion board data (to $p < 0.001$).

Necessarily, *usually* and *sort of* are all underused in CANELC when compared to the spoken BNC (to $p < 0.01$, $p < 0.01$ and $p < 0.01$ respectively) and, similarly, the first two of these terms are also underused compared to the written data (to $p < 0.001$ and $p < 0.01$ respectively) while *sort of* is slightly overused compared to the written BNC (to $p < 0.001$). *Necessarily* and *sort of* are used at consistent rates across all modes aside from the tweets, where a significant underuse *of sort* of can be seen when compared to CANELC (to $p < 0.01$). Comparatively, *usually* is significantly overused in the discussion board data and underused in the email data compared to the other modes included in CANELC (to $p < 0.01$ and $p < 0.001$ respectively).



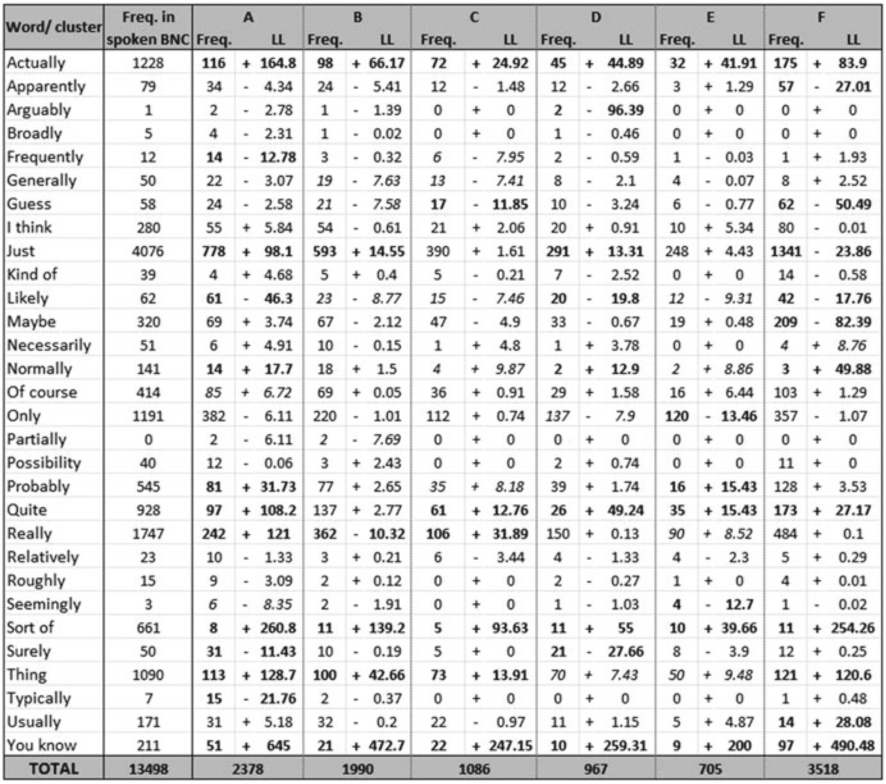
Word/ cluster	Freq. in CANELC	A			B			C			D			E			F		
		Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL
Actually	538	*116*	*+*	*6.8*	98	-	0.21	72	-	3.99	45	+	0.18	32	+	0.89	175	-	2.34
Apparently	142	34	+	0.68	24	+	0.01	12	+	0.47	12	+	0.03	3	+	6.43	57	-	4.59
Arguably	5	2	-	0.17	1	-	0.02	0	+	0.98	2	-	2.48	0	+	0.68	0	+	2.5
Broadly	6	4	-	1.67	1	+	0	0	+	1.18	1	-	0.29	0	+	0.82	0	+	3
Frequently	27	14	-	3.26	3	+	0.6	6	-	2.44	2	+	0.07	1	+	0.49	*1*	*+*	*7.9*
Generally	74	22	-	0.06	19	-	2.17	13	-	2.75	8	-	0.25	4	+	0.28	*8*	*+*	*8.73*
Guess	140	24	+	5.49	21	+	0.39	17	-	0.38	10	+	0.5	6	+	1.65	*62*	*-*	*7.93*
I think	240	55	+	1.84	54	-	2.86	21	+	0.56	20	+	0.09	10	+	3.1	80	-	1.48
Just	3641	**778**	**+**	**49.04**	593	+	1.94	390	-	0.46	291	+	3.46	248	+	0.24	**1341**	**-**	**63.01**
Kind of	35	4	+	3.63	5	+	0.17	5	-	0.42	7	-	3.15	0	+	4.76	14	-	1.11
Likely	173	61	-	1.52	23	+	1.52	15	+	0.44	20	-	1.11	12	+	0	42	+	0.87
Maybe	444	**69**	**+**	**23.69**	67	+	1.14	47	-	0.03	33	+	1.1	19	+	5.25	**209**	**-**	**33.76**
Necessarily	22	6	+	0	10	-	5.54	1	+	0.84	1	+	0.54	0	+	2.99	4	+	0.75
Normally	43	14	-	0.24	*18*	*-*	*8.6*	4	+	0.04	2	+	1	2	+	0.37	3	+	8.39
Of course	338	85	+	0.78	69	-	1.5	36	-	0.03	29	+	0.04	16	+	2.7	103	-	0.38
Only	1328	382	-	0.24	220	+	0.37	112	+	4.5	137	-	2.47	120	-	6.45	357	+	0.89
Partially	4	2	-	0.42	2	-	1.29	0	+	0.79	0	+	0.68	0	+	0.54	0	+	2
Possibility	28	12	-	1.44	3	+	0.71	0	+	5.5	2	+	0.1	0	+	3.81	11	-	0.79
Probably	376	81	+	4.78	77	-	1.73	35	+	0.36	39	-	0.76	16	+	4.54	128	-	3.01
Quite	529	**97**	**+**	**16.04**	**137**	**-**	**16.26**	61	-	0.64	**26**	**+**	**10.66**	35	+	0.12	173	-	2.5
Really	1434	**242**	**+**	**59.32**	**362**	**-**	**38.02**	**106**	**+**	**12.18**	150	-	3.23	90	+	1.13	**484**	**-**	**10.35**
Relatively	32	10	-	0.09	3	+	1.22	6	-	1.56	4	-	0.37	4	-	1.01	5	+	1.79
Roughly	18	9	-	1.88	2	+	0.4	0	+	3.54	2	-	0.08	1	+	0.06	4	+	0.21
Seemingly	14	6	-	0.72	2	+	0.07	0	+	2.75	1	+	0.05	4	-	4.61	1	+	2.67
Sort of	56	8	+	3.72	11	-	0.14	5	+	0.1	11	-	4.77	10	-	5.92	11	+	2.04
Surely	87	31	-	1.29	10	+	1.68	5	+	1.94	**21**	**-**	**13.52**	8	-	0.49	12	+	6.6
Thing	527	113	+	6.39	100	-	0.69	73	-	5.13	*70*	*-*	*8.78*	50	-	3.77	121	+	4.69
Typically	18	*15*	*-*	*9.03*	2	+	0.4	0	+	3.54	0	+	3.08	0	+	2.45	1	+	4.19
Usually	115	31	+	0.03	32	-	5.13	22	-	6.09	11	-	0.05	5	+	1.29	**14**	**+**	**11.17**
You know	211	51	+	0.9	*21*	*+*	*6.82*	22	-	0	10	+	4.68	9	+	2.52	**97**	**-**	**14.37**
TOTAL	10645	2378			1990			1086			967			705			3518		

Fig. 7 The use of hedges in the topic categories in CANELC

Finally, we see no statistical difference in the use of *arguably* and *partially* when comparing CANELC to the spoken and written BNC, or across the individual modes of e-language.

3.3 *Patterns of Use Across Topics*

In addition to exploring the use of the hedges across the different modes in CANELC, we are able to look in more detail at differences in use across the topic categories detailed in Fig. 3. Figure 7 documents the frequency of word use across the different topic categories and provides a log-likelihood score of difference in use for each category compared to CANELC (note – a '+' indicates an overuse in CANELC compared to a category, thus an underuse in the given category), while Figs. 8 and 9 tabulate the frequency of use across these topics compared to the spoken and written BNC (note – a '+' indicates an overuse in the BNC compared to a category).



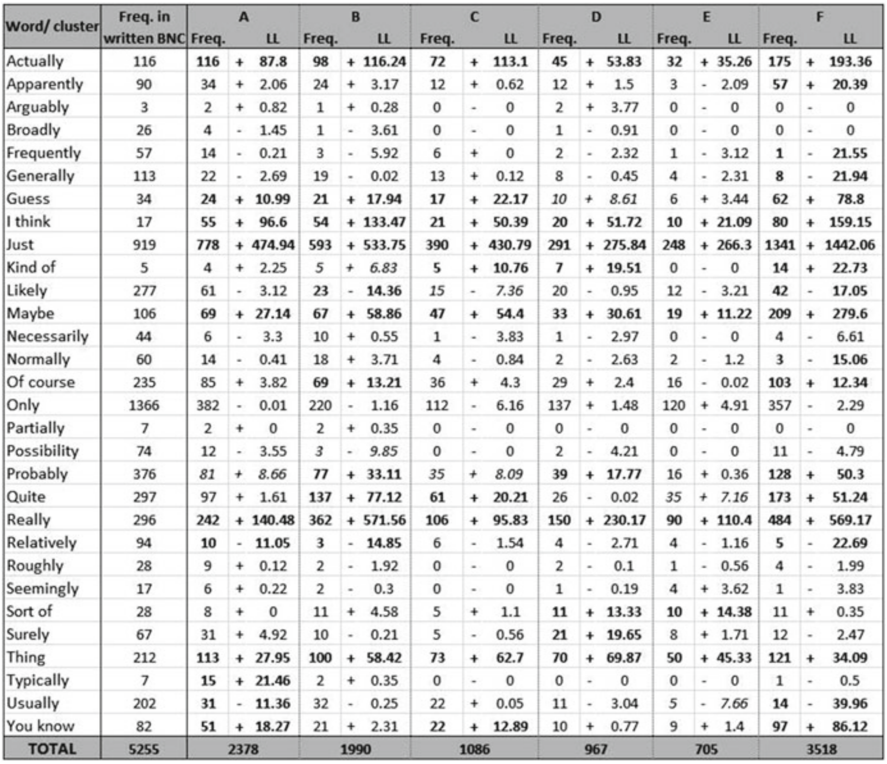
Word/ cluster	Freq. in	A			B			C			D			E			F		
	spoken BNC	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL
Actually	1228	116	+	164.8	98	+	66.17	72	+	24.92	45	+	44.89	32	+	41.91	175	+	83.9
Apparently	79	34	-	4.34	24	-	5.41	12	-	1.48	12	-	2.66	3	+	1.29	57	-	27.01
Arguably	1	2	-	2.78	1	-	1.39	0	+	0	2	-	96.39	0	+	0	0	+	0
Broadly	5	4	-	2.31	1	-	0.02	0	+	0	1	-	0.46	0	+	0	0	+	0
Frequently	12	14	-	12.78	3	-	0.32	6	-	7.95	2	-	0.59	1	-	0.03	1	+	1.93
Generally	50	22	-	3.07	19	-	7.63	13	-	7.41	8	-	2.1	4	-	0.07	8	+	2.52
Guess	58	24	-	2.58	21	-	7.58	17	-	11.85	10	-	3.24	6	-	0.77	62	-	50.49
I think	280	55	+	5.84	54	-	0.61	21	+	2.06	20	+	0.91	10	+	5.34	80	-	0.01
Just	4076	778	+	98.1	593	+	14.55	390	+	1.61	291	+	13.31	248	+	4.43	1341	-	23.86
Kind of	39	4	+	4.68	5	+	0.4	5	-	0.21	7	-	2.52	0	+	0	14	-	0.58
Likely	62	61	-	46.3	23	-	8.77	15	-	7.46	20	-	19.8	12	-	9.31	42	-	17.76
Maybe	320	69	+	3.74	67	-	2.12	47	-	4.9	33	-	0.67	19	+	0.48	209	-	82.39
Necessarily	51	6	+	4.91	10	-	0.15	1	+	4.8	1	+	3.78	0	+	0	4	+	8.76
Normally	141	14	+	17.7	18	+	1.5	4	+	9.87	2	+	12.9	2	+	8.86	3	+	49.88
Of course	414	85	+	6.72	69	+	0.05	36	+	0.91	29	+	1.58	16	+	6.44	103	+	1.29
Only	1191	382	-	6.11	220	-	1.01	112	+	0.74	137	-	7.9	120	-	13.46	357	-	1.07
Partially	0	2	-	6.11	2	-	7.69	0	+	0	0	+	0	0	+	0	0	+	0
Possibility	40	12	-	0.06	3	+	2.43	0	+	0	2	+	0.74	0	+	0	11	+	0
Probably	545	81	+	31.73	77	+	2.65	35	+	8.18	39	+	1.74	16	+	15.43	128	+	3.53
Quite	928	97	+	108.2	137	+	2.77	61	+	12.76	26	+	49.24	35	+	15.43	173	+	27.17
Really	1747	242	+	121	362	-	10.32	106	+	31.89	150	+	0.13	90	+	8.52	484	+	0.1
Relatively	23	10	-	1.33	3	+	0.21	6	-	3.44	4	-	1.33	4	-	2.3	5	+	0.29
Roughly	15	9	-	3.09	2	+	0.12	0	+	0	2	-	0.27	1	+	0	4	+	0.01
Seemingly	3	6	-	8.35	2	-	1.91	0	+	0	1	-	1.03	4	-	12.7	1	-	0.02
Sort of	661	8	+	260.8	11	+	139.2	5	+	93.63	11	+	55	10	+	39.66	11	+	254.26
Surely	50	31	-	11.43	10	-	0.19	5	+	0	21	-	27.66	8	-	3.9	12	+	0.25
Thing	1090	113	+	128.7	100	+	42.66	73	+	13.91	70	+	7.43	50	+	9.48	121	+	120.6
Typically	7	15	-	21.76	2	-	0.37	0	+	0	0	+	0	0	+	0	1	+	0.48
Usually	171	31	+	5.18	32	-	0.2	22	-	0.97	11	+	1.15	5	+	4.87	14	+	28.08
You know	211	51	+	645	21	+	472.7	22	+	247.15	10	+	259.31	9	+	200	97	+	490.48
TOTAL	13498	2378			1990			1086			967			705			3518		

Fig. 8 The rate of use of hedges in the topic categories in CANELC, compared to the spoken BNC

Six sub-corpora of the CANELC data were created (for A–F) to draw these comparisons in the data.

From Fig. 7 we can see that none of the hedging terms are overused in data classified under topic category 'A' compared to CANELC, although *just*, *maybe*, *quite* and *really* are all significantly underused (to $p < 0.01$) and *actually* and *typically* are slightly underused (to $p < 0.001$). Similarly, Fig. 7 shows an underuse of *a bit*, *like* and *stuff* in this category when compared to the corpus as a whole (to $p < 0.01$). As documented in Figs. 8 and 9, *actually*, as used in category 'A' in CANELC occurs at a far less frequent rate than it does in the spoken and written BNC (both to $p < 0.01$) and the converse is true for *relatively* (to $p < 0.01$). While for *frequently*, *likely*, *seemingly* and *partially*, there is a higher rate of use in category 'A' than the spoken BNC, but a near consistent rate of use to the written corpus (to $p < 0.01$, $p < 0.01$ and $p < 0.001$ respectively).

Surely and *typically* are used at a higher rate in the category 'A' data in the spoken BNC data, but while *surely* is used at a near consistent rate to the written BNC, *typically* is far less frequent in A. The converse of this is true for *typically*. While *arguably*, *possibility*, *roughly*, *only* and *generally*, when classified in category 'A'



Word/ cluster	Freq. in written BNC	A			B			C			D			E			F		
		Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL	Freq.		LL
Actually	116	116	+	87.8	98	+	116.24	72	+	113.1	45	+	53.83	32	+	35.26	175	+	193.36
Apparently	90	34	+	2.06	24	+	3.17	12	+	0.62	12	+	1.5	3	-	2.09	57	+	20.39
Arguably	3	2	+	0.82	1	+	0.28	0	-	0	2	+	3.77	0	-	0	0	-	0
Broadly	26	4	-	1.45	1	-	3.61	0	-	0	1	-	0.91	0	-	0	0	-	0
Frequently	57	14	-	0.21	3	-	5.92	6	+	0	2	-	2.32	1	-	3.12	1	-	21.55
Generally	113	22	-	2.69	19	-	0.02	13	+	0.12	8	-	0.45	4	-	2.31	8	-	21.94
Guess	34	24	+	10.99	21	+	17.94	17	+	22.17	10	+	8.61	6	+	3.44	62	+	78.8
I think	17	55	+	96.6	54	+	133.47	21	+	50.39	20	+	51.72	10	+	21.09	80	+	159.15
Just	919	778	+	474.94	593	+	533.75	390	+	430.79	291	+	275.84	248	+	266.3	1341	+	1442.06
Kind of	5	4	+	2.25	5	+	6.83	5	+	10.76	7	+	19.51	0	-	0	14	+	22.73
Likely	277	61	-	3.12	23	-	14.36	15	-	7.36	20	-	0.95	12	-	3.21	42	-	17.05
Maybe	106	69	+	27.14	67	+	58.86	47	+	54.4	33	+	30.61	19	+	11.22	209	+	279.6
Necessarily	44	6	-	3.3	10	+	0.55	1	-	3.83	1	-	2.97	0	-	0	4	-	6.61
Normally	60	14	-	0.41	18	+	3.71	4	-	0.84	2	-	2.63	2	-	1.2	3	-	15.06
Of course	235	85	+	3.82	69	+	13.21	36	+	4.3	29	+	2.4	16	-	0.02	103	+	12.34
Only	1366	382	-	0.01	220	-	1.16	112	-	6.16	137	+	1.48	120	+	4.91	357	-	2.29
Partially	7	2	+	0	2	+	0.35	0	-	0	0	-	0	0	-	0	0	-	0
Possibility	74	12	-	3.55	3	-	9.85	0	-	0	2	-	4.21	0	-	0	11	-	4.79
Probably	376	81	+	8.66	77	+	33.11	35	+	8.09	39	+	17.77	16	+	0.36	128	+	50.3
Quite	297	97	+	1.61	137	+	77.12	61	+	20.21	26	-	0.02	35	+	7.16	173	+	51.24
Really	296	242	+	140.48	362	+	571.56	106	+	95.83	150	+	230.17	90	+	110.4	484	+	569.17
Relatively	94	10	-	11.05	3	-	14.85	6	-	1.54	4	-	2.71	4	-	1.16	5	-	22.69
Roughly	28	9	+	0.12	2	-	1.92	0	-	0	2	-	0.1	1	-	0.56	4	-	1.99
Seemingly	17	6	+	0.22	2	-	0.3	0	-	0	1	-	0.19	4	+	3.62	1	-	3.83
Sort of	28	8	+	0	11	+	4.58	5	+	1.1	11	+	13.33	10	+	14.38	11	+	0.35
Surely	67	31	+	4.92	10	-	0.21	5	-	0.56	21	+	19.65	8	+	1.71	12	-	2.47
Thing	212	113	+	27.95	100	+	58.42	73	+	62.7	70	+	69.87	50	+	45.33	121	+	34.09
Typically	7	15	+	21.46	2	+	0.35	0	-	0	0	-	0	0	-	0	1	-	0.5
Usually	202	31	-	11.36	32	-	0.25	22	+	0.05	11	-	3.04	5	-	7.66	14	-	39.96
You know	82	51	+	18.27	21	+	2.31	22	+	12.89	10	+	0.77	9	+	1.4	97	+	86.12
TOTAL	5255	2378			1990			1086			967			705			3518		

Fig. 9 The rate of use of hedges in the topic categories in CANELC compared to the written BNC

occur at near-consistent rates to the spoken and written BNC data (as seen in Fig. 8) and *relatively*, although nearly-consistent to the spoken BNC, is used at a much higher rate in the topic 'A' data than the written BNC (to $p<0.01$, as seen in Fig. 9).

For topic 'B', that is topics covering 'culture, literature and the arts', 'fashion' and 'teaching, academia and education', Fig. 7 indicates that the only significant differences seen are in the rate of use of *quite* and *really*, both of which are used at a rate higher than the average rate seen in CANELC.

Necessarily, *normally*, *broadly* and *usually* are terms that are most commonly classified under topic category 'B' in CANELC. The rate of use of these terms, in this category are shown to be nearly consistent to the rates of use in the spoken and written BNC, as no real significant differences are outlined in Figs. 8 and 9. There is, however, an underuse of *sort of*, in the category 'B' data compared to the spoken BNC (which is also most commonly classified under category 'B'), while near consistent rates to the written BNC are shown.

Figure 7 indicates that there are no significant differences in the use of the search terms for topic 'E'. There is, however, a significant underuse of *really* in CANELC compared to 'C', and an underuse of *quite* and an overuse of *surely* compared to 'D'.

These are the only real difference seen for these categories (to $p<0.01$). None of the hedges explored were more frequently used in the data classified under topic category 'E' or 'C' than the other topic categories. The only ones frequently used in 'D' were *arguably* and *sort of*. *Arguably* is overused in this category compared to the average use in the spoken BNC, but near-consistent with rates of use in the written BNC, while *sort of* is used at a significantly lower rate in the topic 'D' data than the spoken and written BNC (to $p<0.01$).

Finally, Fig. 7 highlights that *just*, *maybe* and *really* are all used at a significantly higher rate in the data for category 'F' than the CANELC average (all to $p<0.01$) and *usually* is used at a lower rate than the CANELC average (both to $p<0.01$). The first of these terms are also significantly overused compared to the spoken BNC, but significantly underused compared to the written BNC. It is the use of terms in this category that we see the most marked difference in frequency rates when compared to the written and spoken BNC data (Figs. 8 and 9).

Apparently, *guess*, *just*, *maybe*, *stuff*, *or so* and *a bit* are all used at a significantly higher rate in CANELC compared to both the spoken and written data (all to $p<0.01$ aside from *a bit* and *or so* which are to $p<0.001$ for the spoken and written data respectively) while *like*, *quite*, *you know* and *thing* are all underused in the category 'F' data compared to the spoken BNC but overused when compared to the written data (all to $p<0.01$). *Kind of*, *I think*, *probably* and *really* are all significantly overused in the category 'F' data when compared to the written BNC but are used at near consistent rates to the spoken excerpt (to $p<0.01$). Conversely, *sort of* is significantly underused in this data compared to the spoken BNC, but used at near-consistent compared to the written data and *of course* is used at near-consistent rates in the category 'F' data compared to both the written and spoken BNC.

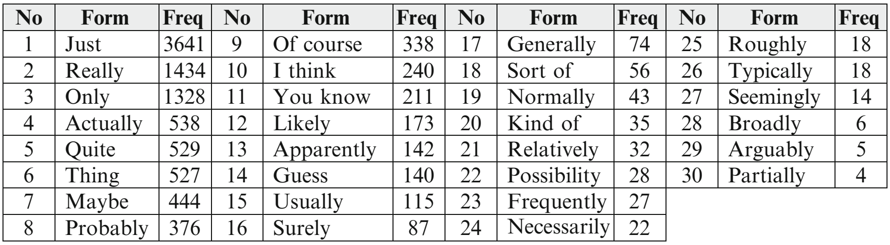
4 Discussion

Of the hedges examined, the most commonly used forms featured in CANELC were:

From this we can surmise that:

1. Of the forms examined, the most frequent hedge used in CANELC is the adverb *just*, followed by *really* and *only*.

Seven of the top ten of these hedges featured in Fig. 10 were shown to be significantly underused in CANELC compared to the spoken BNC but overused compared to the written BNC. The first of these adverbs were also shown to be frequently used in the study of hedging in LCIE (Farr et al. 2004), but none of noted as common hedges in studies of written academic discourse (see Channell 1990; Clemen 1997; Gries and David 2007). As discussed by Atai and Sadr (2006) the use of full verbs, nouns and adjectives as hedges (in that order) are often the most commonly used forms in more formal, written contexts. Although hedges of these forms were common in the data, they were used far less frequently than the adverbial forms.



No	Form	Freq	No	Form	Freq	No	Form	Freq	No	Form	Freq
1	Just	3641	9	Of course	338	17	Generally	74	25	Roughly	18
2	Really	1434	10	I think	240	18	Sort of	56	26	Typically	18
3	Only	1328	11	You know	211	19	Normally	43	27	Seemingly	14
4	Actually	538	12	Likely	173	20	Kind of	35	28	Broadly	6
5	Quite	529	13	Apparently	142	21	Relatively	32	29	Arguably	5
6	Thing	527	14	Guess	140	22	Possibility	28	30	Partially	4
7	Maybe	444	15	Usually	115	23	Frequently	27			
8	Probably	376	16	Surely	87	24	Necessarily	22			

Fig. 10 Rank order of the 30 hedges in CANELC (by frequency of use)

This suggests that, by form alone, the use of hedging in e-language shows some clear similarities with those used in more informal, spoken discourse.

More generally, of the 30 hedges examined, 15 were found to be more frequent in the spoken than written BNC sample than in CANELC. Of these terms, 11 were significantly underused in CANELC compared to the BNC (10 to $p<0.01$ and 1 to $p<0.001$) while only 2 were overused in CANELC. Similarly, there was a higher rate of underuse of the 15 terms most frequently used in the written data, although this was only seen with 7 of the terms (with 2 of these 15 being overused in CANELC). Across all 30 terms, we saw that 12 of them were significantly underused and 7 overused in CANELC compared to the spoken data, while 15 were overused and 8 were underused in CANELC compared to the written data. This can be summarised as follows:

2. Hedges that were most frequently used in the spoken rather than written BNC sample (and vice versa) were used at a significantly lower rate in the e-language data.
3. Of the forms analysed, a higher proportion were significantly overused rather than underused in CANELC when compared to the written data (15 vs. 8).
4. Of the forms analysed, a higher proportion were significantly underused rather than overused in CANELC when compared to the spoken data (12 vs. 7).

These findings suggest that the rate of hedging use in the e-language data is inconsistent with typical rates in spoken and written discourse. While more hedges were used compared to the written data, far fewer were used than in the spoken data. This provides an argument for classifying e-language as its own distinct genre (as suggested in Sect. 2).

When comparing the patterns of use across the different modes of data we also see the following:

5. Emails and discussion boards contained fewer disparities in the rate of under/overuse of specific hedging forms than other modes of e-language (i.e. they were most 'similar').
6. The SMS, discussion board and twitter data contained the most disparities in the rate of under/overuse of specific hedging forms than other modes of e-language (i.e. they were the least 'similar' modes of e-language).

In terms of relative frequencies (calculated as the number of hedges used per word in each of the modes) we see that:

7. Hedges were used at a more frequent rate in the SMS and discussion board data than the other modes (1:72 words and 1:86 words), while they were used at a near consistent rate across the twitter, email and blog modes (1:101, 1:103 and 1:105 respectively).

Again, this is an interesting finding as it is in the 'most immediate' form of e-language, SMS messages (which, from show a shorter delay in the response times to messages in CANELC), there is a tendency for a higher number of hedges to be used. For the SMS messages, given that the relationship between the sender and sendee is often 'fixed', with messages being directed at individuals or groups of people known to the sender, and are often classified as being of the 'personal and daily life' topic, the need for hedging to mitigate against potential face threats is assumed to be reduced, so the reverse of this is interesting here. Similarly, while it is not necessarily the case that discussion board members 'know' each other personally, this mode of e-language often involves a fixed community of contributors who respond to each other regularly, creating a closeness between those involved.

The data also reveals that dramatic differences are seen in frequency rates across the different topic categories, compared to corpus as a whole. Of all the hedges analysed, the most common topic of the content was classified under category 'F'. When compared to the BNC, we saw that those terms in category 'F' were statistically overused in the 'F' data than in both the written and spoken BNC. This was true of 8 of the 17 terms featured under the category 'F' data in Fig. 8 (to $p<0.01$ or $p<0.001$). These patterns can be summarised as follows:

8. Based on frequency, content classified under the topics in categories 'A' and 'F' used more hedging than the other topic categories.
9. Of the hedges analysed, all were, on average, used at a less frequent rate in each of the topic sub-corpora when compared to the written BNC.
10. While all hedges were also used at a less frequent rate in the topic sub-corpora than in the spoken BNC, the difference in rate of use was less significant than when compared to the written BNC.
11. Hedges used in topic categories 'B', 'C' and 'D' were underused and overused a near-consistent rate when compared to the spoken BNC. Hedges used in the category 'A' data were most significantly underused in the data when compared to the spoken BNC.

As is perhaps to be expected, then, the more formal and the more 'spoken' topic categories (i.e. interpersonal contexts, category 'F') witnessed a higher rate of hedging use than was the case with the other topics. As we saw earlier, spoken discourse often utilises more hedges than written discourse, but more formal spoken and written contexts use more hedges than the informal ones. The content which concerns matters related to personal and daily life are more akin to spoken discourse (although at the more informal end) so the more extensive use of hedging in this category is as expected. Similarly, the topics in category 'A' are most akin to 'formal' discursive contexts (both across written and spoken genres) so the frequent use of hedging also aligns with expectations.

If we look at some specific forms of hedging in more detail we see that *kind of* and *sort of* are two hedges which have previously been found to be particularly frequent in formal language contexts, specifically academic discourse (Biber et al. 1999: 560–56; Poos and Simpson 2002: 1). We would thus expect them to be more prevalent in the content classified under category B, in 'teaching, academia and education'. This pattern was not mirrored in the e-language content and, in fact, there was a general underuse of both of these terms across the topics, modes and corpus when compared to the spoken and written data.

5 Summary

This chapter has revealed that there is no clear-cut relationship between the use of hedging in e-language compared to written and spoken genres of discourse. The use of hedging across different communicative contexts (defined by topic categories) and across the different modes of e-language is fluid and not necessarily fixed, although when compared to standard (BNC) written and spoken modes of discourse the forms of hedging isolated for the purposes of this study appear to behave in a way that suggests greater internal similarity across the modes than similarity with the standard (BNC) written and spoken data. As initially suggested by Crystal (2003), there appears to be an argument to conceptualise e-language as its own distinct variety on the continuum of formality: between spoken and written discourse. The more immediate forms of e-language (e.g. SMS messages) are positioned closer to the 'spoken' end while the emails and blogs are better positioned towards the more formal, written end (based on what we have found here).

To build on what has been found here, a more qualitative, screen by screen study of the data would allow us to examine, more closely, specific functions of the common hedging forms analysed here. A closer observation of hedging use between specific contributors (according to gender and relationship, for example) may also help us to create a clearer profile of use across the different modes. Finally, a focus on a wider range of hedging forms and a clearer distinction between the individual functions of forms, in specific contexts, as well as extending the focus to synchronous forms of e-language (e.g. IMs) would add to the discussions. There is scope to carry out such investigations in future studies of this nature.

References

Allison, D. 1995. Assertions and alternatives: Helping ESL undergraduates extend their choice in academic writing. *Journal of Second Language Writing* 4: 1–16.

Atai, M., and L. Sadr. 2006. A cross-cultural genre study on hedging devices in discussion section of applied linguistics research articles. In Proceedings of the 11th conference of Pan-Pacific Association of Applied Linguistics, 42–57. Hong Kong: The Chinese University of Hong Kong.

Atkins, S. 2011. *A cognitive linguistic perspective on social space in online health communities*. Unpublished Ph.D. thesis. Nottingham: The University of Nottingham.

Baron, N. 1998. Writing in the age of email: The impact of ideology versus technology. *Visible Language* 32(1): 35–53.

Bates, E. 1976. *Language and context*. New York: Academic.

Baym, N. 1995. The emergence of community in computer-mediated communication. In *Cybersociety: Computer-mediated communication and community*, ed. S.G. Jones, 138–163. Thousand Oaks: Sage.

Benjamin, J. 2011. Tweets, Blogs, Facebook and the Ethics of 21st- Century Communication Technology. In *Social media: Usage and impact*, ed. Noor Al-Deen, H.S. and J.A. Hendricks, 271–288. Maryland: Lexington Books.

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

Biber, D. 1992. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes* 15: 133–163.

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257.

Biber, D., S. Conrad, G. Leech, J. Svartvik, and E. Finegan. 1999. *The Longman grammar of spoken and written English*. Harlow: Longman.

Boyd, D., and J. Heer. 2006. Profiles as conversation: Networked identity performance on Friendster. In Proceedings of the Hawaii International Conference on System Sciences (HICSS-39), Persistent Conversation Track, 4–7 Jan 2006. Kauai: IEEE Computer Society.

Brennan, S.E., and J.O. Ohaeri. 1999. Why do electronic conversations seem less polite? The costs and benefits of hedging. In Proceedings, International Joint Conference on Work Activities, Coordination, and Collaboration (WACC '99), 227–235, San Francisco.

Brown, G. 1989. Making sense: The interaction of linguistic expression and contextual information. *Applied Linguistics* 10(1): 97–108.

Brown, P., and S.C. Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, ed. E. Goody, 56–311. Cambridge: Cambridge University Press.

Brown, P., and S.C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Brown, G., and G. Yule. 1983. *Discourse analysis*. Cambridge: Cambridge University Press.

Burnard, L. 2005. Developing linguistic corpora: Metadata for corpus work. In *Developing linguistic corpora: A guide to good practice*, ed. M. Wynne, 30–46. Oxford: Oxbow Books.

Carter, R.A., and M.J. McCarthy. 2006. *Cambridge grammar of English*. Cambridge: Cambridge University Press.

Chafe, W.L., and J. Danielewicz. 1987. Properties of spoken and written language. In *Comprehending oral and written language*, ed. R. Horowitz and S.J. Samuels, 83–113. New York: Academic.

Channell, J. 1985. Vagueness as a conversational strategy. *Nottingham Linguistic Circular* 14: 3–24.

Channell, J. 1990. Precise and vague quantities in academic writing. In *The writing scholar: Studies in the language and conventions of academic discourse*, ed. W. Nash, 95–117. Newbury Park: Sage Publications.

Channell, J. 1994. *Vague language*. Oxford: Oxford University Press.

Cheng, W., and M. Warren. 1999. Inexplicitness: What is it and should we be teaching it? *Applied Linguistics* 20(3): 293–315.

Cherny, L. 1999. *Conversation and community: Chat in a virtual world*. Stanford: Center for the Study of Language and Information.

Clancy, B. 2002. The exchange in family discourse. *Teanga* 21: 134–150.

Clemen, G. 1997. The concept of hedging: Origins, approaches and definitions. In *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts*, ed. R. Markkanen and H. Schröder, 235–248. Berlin: Walter de Gruyter.

Collot, M., and N. Belmore. 1996. Electronic language: A new variety of English. In *Computer mediated communication: Linguistic, social and cross-cultural perspectives*, ed. S.C. Herring, 13–28. Amsterdam: John Benjamins.

Condon, S., and C. Cech. 1996. Functional comparison of face-to-face and computer-mediated decision-making interactions. In *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*, ed. S. Herring, 65–80. Philadelphia: John Benjamins.

Crystal, D. 2001. *Language and the internet*. Cambridge: Cambridge University Press.

Crystal, D. 2003. The joy of text. *Spotlight magazine,* 16–17.

Crystal, D. 2008. *Txtng: The Gr8 Db8*. Oxford: Oxford University Press.

Crystal, D., and D. Davy. 1975. *Advanced conversational English*. London: Longman.

Drave, N. 1995. *The pragmatics of vague language*: *A corpus-based study of vagueness in national vocational qualifications*. Unpublished Master's dissertation, University of Birmingham.

Dubois, B.L. 1987. 'Something on the order of around forty to forty-four': Imprecise numerical expressions in biomedical slide talks. *Language in Society* 16: 527–541.

Duranti, A., and C. Goodwin (eds.). 1992. *Rethinking context: Language as an interactive phenomenon*. Cambridge: Cambridge University Press.

Eckert, P., and J.R. Rickford (eds.). 2001. *Style and sociolinguistic variation*. Cambridge: Cambridge University Press.

Farr, F., and A. O'Keeffe. 2002. Would as a hedging device in an Irish context: An intra-varietal comparison of institutionalised spoken interaction. In *Using corpora to explore linguistic variation*, ed. R. Reppen, S. Fitzmaurice, and D. Biber, 25–48. Amsterdam: John Benjamins.

Farr, F., B. Murphy, and A. O'Keeffe. 2004. The limerick corpus of Irish English: Design, description and application. *Teanga* 21: 5–29.

Fetzer, A. 2004. *Recontextualizing context: Grammaticality meets appropriateness*. London: John Benjamins Publishing Company.

Fowler, R., and G. Kress. 1979. Critical linguistics. In *Language and control*, ed. R. Fowler, B. Hodge, G. Kress, and T. Trew, 185–213. London: Routledge Kegan Paul.

Garcia, A.C., and J.B. Jacobs. 1999. The eyes of the beholder: Understanding the turn-taking system in quasi-synchronous computer-mediated communication. *Research on Language and Social Interaction* 32: 337–367.

Green, L.J. 2002. *African American English: A linguistic introduction*. Cambridge: Cambridge University Press.

Gries, S.Th., and C.V. David. 2007. This is kind of/sort of interesting: Variation in hedging in English. In *Studies in variation*, *contacts and change in English 2*: *Towards multimedia in corpus studies*, Vol. 2. Helsinki: Varieng.

Halliday, M.A.K., and R. Hasan. 1989. *Language, context and text: Aspects of language in a social semiotic perspective*. Oxford: OUP.

Hard af Segerstag, Y. 2002. *Use and adaptation of the written language to the conditions of computer-mediated communication*. Unpublished Ph.D. thesis, University of Goteborg.

Herring, S.C. (ed.). 1996. *Computer-mediated communication: Linguistic, social and crosscultural perspectives*. Amsterdam: John Benjamins.

Herring, S. 1999. Interactional coherence in CMC. *Journal of Computer-Mediated Communication* 4(4): 1–13.

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

Herring, S. 2007. A faceted classification scheme for computer-mediated discourse. *Language@Internet 4*(1): 1–37.

Heylighen, F., and J.-M. Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science* 6: 293–340.

Hübler, A. 1983. *Understatements and hedges in English*. Amsterdam: John Benjamins.

Hyland, K. 1996. Writing without conviction? Hedging in scientific research articles. *Applied Linguistics* 17(4): 433–454.

Jucker, A.H., S.W. Smith, and T. Ludge. 2003. Interactive aspects of vagueness in conversation. *Journal of Pragmatics* 35: 1737–1769.

Kennedy, G. 1987. Quantification and the use of English: A case study of one aspect of the learner's task. *Applied Linguistics* 8(3): 264–286.

Knight, D. 2011. *Multimodality and active listenership*. London: Continuum.

Knight, D., S. Adolphs, and R. Carter. 2014. CANELC – Constructing an e-language corpus. *Corpora Journal*. 9(1).

Ko, K. 1996. Structural characteristics of computer-mediated language: A comparative analysis of InterChange discourse. *Electronic Journal of Communication* 6(3).

Kopytko, R. 2003. What is wrong with modern accounts of context in linguistics? *Vienna English Working Papers* 12: 45–60.

Labov, W. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lakoff, R. 1972. Language in context. *Language* 48(4): 907–927.

Leech, G. 2000. Grammar of spoken English: New outcomes of corpus-oriented research. *Language Learning* 50(4): 675–724.

Ling, R. 2003. The socio-linguistic of SMS: An analysis of SMS use by random sample of Norwegians. In *Mobile communications: Renegotiation of the social sphere*, ed. R. Ling and P. Pedersen, 335–349. London: Springer.

McCarthy, M. 1991. *Discourse analysis for language teachers*. Cambridge: Cambridge University Press.

McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based language studies: An advanced resource book*. London: Routledge.

Morris, C. 1946. *Signs, language and behaviour*. Englewood-Cliffs: Prentice Hall.

Murray, D.E. 1988. The context of oral and written language: A framework for mode and medium switching. *Language in Society* 17: 351–373.

Myers, G. 2010. *The discourse of blogs and wikis*. London: Continuum.

Nelson, K., S. Engel, and A. Kyratzis. 1985. The evolution of meaning in context. *Journal of Pragmatics* 9: 453–474.

Nikula, T. 1997. Interlanguage view on hedging. In *Hedging and discourse: Approaches to the analysis of a pragmatic phenomenon in academic texts*, ed. R. Markkanen and H. Schröder, 188–207. Berlin: Walter de Gruyter.

O'Keeffe, A., M. McCarthy, and R. Carter. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Oksman, V., and J. Turtianen. 2004. Mobile communication as a social stage: Meanings of mobile communication in everyday life among teenagers in Finland. *New Media and Society* 6(3): 319–339.

Poos, D., and R.C. Simpson. 2002. Cross-disciplinary comparisons of hedging: Some findings from the Michigan Corpus of Academic Spoken English. In *Using corpora to explore linguistic variation*, ed. R. Reppen, S.M. Fitzmaurice, and D. Biber, 3–23. Amsterdam: John Benjamins.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Rayson, P. 2003. *Matrix*: *A statistical method and software tool for linguistic analysis through corpus comparison*. Unpublished Ph.D. thesis, Lancaster University.

Scollon, R., and S. Scollon. 2003. *Discourse in place: Language in the material world*. London: Routledge.

Shortis, T. 2007. Gr8 Txtpectations: The creativity of text spelling. *English Drama Media Journal* 8: 21–26.

Sutherland, J. 2002. Cn u txt? *The Guardian*, 11 Nov 2002. http://www.guardian.co.uk/technology/2002/nov/11/mobilephones2. Accessed 22 April 2013.

Tagg, C. 2009. *A corpus linguistics study of SMS text messaging*. Unpublished Ph.D. thesis. Birmingham: The University of Birmingham.

Thurlow, C., L. Lengel, and A. Tomic. 2004. *Computer mediated communication: Social interaction and the internet*. London: Sage.

van Dijk, T.A. (ed.). 1977. *Text and context: Explorations in the semantics and pragmatics of discourse*. London: Longman.

Walther, J.B. 1996. Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research* 23: 3–43.

Widdowson, H.G. 1998. Communication and community: The pragmatics of ESP. *English for Specific Purposes* 17(1): 3–14.

Yates, S. 1996. Oral and written linguistic aspects of computer conferencing: A corpus based study. In *Computer mediated communication: Linguistic, social and cross-cultural perspectives*, ed. S.C. Herring, 29–56. Amsterdam: John Benjamins.