DE GRUYTER
OPEN

ICAME Journal, Volume 40, 2016, DOI: 10.1515/icame-2016-0008

Talk talk, not just small talk. Exploring English contrastive focus reduplication with the help of corpora

Bianca Widlitzki, Justus Liebig University Giessen

Abstract

Contrastive focus reduplication (CR) is a type of reduplication in English which picks out a prototypical or intensified reading of the reduplicated element and shows contrastive stress on the reduplicant: for instance, speakers may use talk talk *to indicate that a 'real talk' – as opposed to e.g. 'just small talk'– took place. The present paper pursues an empirical, corpus-linguistic approach to CR: Based on three mega-corpora of contemporary English, the following aspects in particular are investigated: the importance of the co-text of CR, the possibility of emerging default interpretations for some frequent CRs, and the function(s) CR serves in discourse. In addition, it contains the first analysis of the sociolinguistics of the phenomenon, based on a corpus of blogs. It emerges that contrasts and/or synonyms are commonly employed to clarify the meaning of CR – most frequently in the form of the unreduplicated base (*not talk, but talk talk*) or an explanatory phrase (*talk talk, by which I mean a serious conversation*). CR is most frequent in blogs maintained by women and by young speakers. Its presence in blogs shows that CR is not limited to (fictional representations of) spoken dialogue. Though generally rare, it is also found in other genres (such as fiction, news, and even academic prose). Apart from its disambiguating function, CR is also used for creative purposes (as a kind of wordplay) and apparently serves to build rapport between interlocutors (or bloggers and readers) via reference to common ground.*

1 *Introduction*

Especially in colloquial English, you may come across reduplications like e.g. *talk talk* as in (1) or *like him like him* (2).

(1) my dad and i actually talked. like, not just small talk.. but **talk talk**. it was very nice, indeed. (BC = Blog Corpus, file no. 825029)

2) Now it is this guy […] I don't **like him like him**, I like him as a friend. (BC, 2272189)[1]

These reduplications pick out prototypical or intensified readings of the reduplicated elements: *talk talk* may be taken to mean 'real, serious talk', *like him like him* to signify 'like him in a romantic sense'. As the first element bears contrastive stress and contrasting elements are often part of the immediate co-text (here: *talk talk – small talk*, *like him like him – like him as a friend*), this kind of reduplication is called 'contrastive focus reduplication', abbreviated CR (Ghomeshi *et al.* 2004: 308). The characteristic intonation is sometimes even approximated in writing via capitalization of the focussed element: *I don't just like him, I LIKE HIM like him*.

Research on CR, which has especially picked up in the last 15 years, is largely focussed on its semantics (described as prototypical in some way by most scholars) and the restrictions holding for words and phrases to be reduplicated (e.g. in terms of size and parts of speech involved). For both aspects, sophisticated models have been proposed (cf. Ghomeshi *et al*. 2004 on the scope of CR and e.g. Whitton 2006 or Song and Lee 2014) on its meaning. These accounts are largely based on examples that linguists came across either in personal conversations or scripted dialogue in various media. These 'participant observer' data have been invaluable in determining the properties of CR. Yet, I agree with Hohenhaus (2004: 299), who conducted the (to my knowledge) first and only corpus analysis of CR, that it is "intriguing to try to back up these participant observer impressions by [...] means of corpus linguistics". Moreover, a corpus approach can provide frequency information with regard to the use of CR in general, frequent bases and use among groups of speakers.

The present paper intends to provide a corpus-based perspective on CR – especially on aspects that have yet to be discussed in greater detail. In particular, this study is the first to systematically include a sociolinguistic perspective in order to establish to what extent CR is present among different (groups of) speakers. The use of a corpus of blogs (referred to as the Blog Corpus, abbreviated BC, in the following) as my primary source of data extends the analysis of the phenomenon beyond face-to-face dialogic interaction. Data from two corpora of contemporary English, i.e. the Corpus of Contemporary American English (COCA) and the Soap Opera Corpus (SOAP), refine the picture further. Other aspects to be discussed are the distribution of CR across parts of speech, its co-text (and the role it plays in disambiguating the meaning of CR) and its functions in use (clarification in case of misunderstandings being one). From a methodological perspective, the benefits and challenges of exploring the phenomenon with corpus methods are addressed.

In the following, Section 2 summarizes previous work on the semantics and the formal properties of CR. While the present work refers only intermittently to these well-researched aspects, they are crucial to identifying CR in corpora. Next, Section 3 introduces the methodology of the present study: the corpora are described and important methodological issues are addressed. This includes a discussion of how CR can be distinguished from other types of full reduplication or identical syntactic repetition in English. Section 4 discusses the results of the corpus analysis. After a first overview of the results in terms of formal and semantic aspects of CR 4.1, Section 4.2 examines in particular the role of the co-text for the interpretation of CR, including an excursus into functions of CR. Finally, Section 4.3 seeks to shed light on the users of CR by exploring its distribution across gender and age in the Blog Corpus. Section 5 offers some concluding remarks.

2 Contrastive focus reduplication

Contrastive focus reduplication has been discussed under a number of labels, including “the double construction” (Dray 1987), “lexical clones” (Horn 1993, 2008, Huang 2009), ‘identical constituent compounding’ (Hohenhaus 2004), “reduplicative compounding” (Lieber 2009) or “real-X-TR” (TR = total reduplication; Stolz *et al*. 2011). This terminological variety is a reflection of different theoretical approaches, different kinds of data examined, and ultimately of the fact that CR is very versatile in terms of both its meaning and structure. This section summarizes the most important insights gained from previous research on CR.

In context, the meaning of a particular CR is usually easily interpretable: after all; CR always “signals that one meaning of the [reduplicated] word is being contrasted with other possible meanings” (Ghomeshi *et al*. 2004: 317). What is considerably more difficult is to formulate a precise description of this semantic effect. Horn (1993, 2008: 37), who was one of the first to tackle the issue, distinguishes three possible meanings of a CR: prototypical meaning (especially with nouns), literal meaning (as opposed to figurative use), and finally value-added or intensifying meaning. These possibilities are illustrated in (3) – (5) based on examples from Ghomeshi *et al*. (2004):

(3) prototypical:
I'll make the tuna salad and you make the **SALAD–salad**.
(Ghomeshi *et al*. 2004: 311)

(4) literal:
[Dialogue between a married couple, recently separated and now living apart.]
A: Maybe you'd like to come in and have some coffee?
B: Yeah, I'd like that.
A: Just **COFFEE-coffee**, no double meanings.
(Ghomeshi *et al*. 2004: 315)

(5) value-added/intensifying:
A [to B, who is about to give a recital]: Are you nervous?
B: Yeah, but, you know, not **NERVOUS–nervous**.
(Ghomeshi *et al*. 2004: 315)

Ghomeshi et al. (2004: 316) also notice varying senses of CR. Most of their examples fit a prototype interpretation (Ghomeshi *et al*. 2004: 312), i.e. are seen to pick out the conceptual centre of a category (Taylor 2003: 64). However, the authors also report instances where a CR signals "merely very high salience with no hint of prototypicality, ambiguity, or contrast" (Ghomeshi *et al*. 2004: 316), as in (6):

(6) A: Did you check out the leak in the bathroom?
B: What leak?
A: The **LEAK–leak**. [drags her into the bathroom]
(Ghomeshi *et al*. 2004: 316)

Due to such examples, Ghomeshi *et al*. (2004: 316) declare that they are "uncertain whether CR is itself polysemous or whether it can pick out contextually salient readings in addition to objectively prototypical ones" and conclude that describing the meaning of CR as solely prototypical is too simplistic. Where the phenomenon is discussed under the label 'identical constituent compounding' (ICC), its semantics are described very similarly: Hohenhaus (2004: 301) characterizes the reduplication as carrying a prototype reading or one of extreme degree; Lieber (2009: 364) mentions prototypical or intensified meaning.

This brief overview demonstrates that scholars acknowledge the variability in the semantics of CR, and that all of them include the notion of prototypicality in their explanations – though some state that they only "continue to use it for lack of a better characterization" (Ghomeshi *et al*. 2004: 316). It is especially the debate on the applicability of this notion and the desire to find a more succinct description of the semantics of English CR that prompted two further models: Whitton's (2006) scalar approach and Song and Lee's (2011) dynamic prototypes.

Whitton (2006) points out that prototype-based explanations fall short in several regards. Citing five different readings of *drink drink*, including 'an alcoholic drink', 'hard liquor as opposed to other alcoholic drinks' and even 'a non-alcoholic drink', she shows that CRs of the same lexical item may have different meanings in different contexts (Whitton 2006: 17–21). This is considered incompatible with a prototype-based analysis, which should yield a clear interpretation out of context for all CRs. Furthermore, it is unclear which exemplar(s) of a category should be considered prototypical and where the boundary between prototypical and non-prototypical lies (Whitton 2006: 14). Does *doctor doctor* refer to medical professionals as opposed to other doctors (e.g. doctors of philosophy), to general practitioners as opposed to specialists (such as dentists or epidemiologists), or to other members of the category?

In the end, Whitton (2006: 13) argues that CR does not always signify the most prototypical member of a category but "simply a member that is stronger in some relevant way than other salient alternatives" and proposes a scalar analysis: possible interpretations of a source item (such as *doctor*) are ordered on a relevant scale, which takes into account both the context and the interlocutors' common ground (Whitton 2006: 7). Whichever item is the strongest on that context-dependent scale represents the meaning of an instance of CR.

A different solution is proposed by Song and Lee (2011): instead of abandoning the notion of prototype as an explanatory device, they expand on it. So-called dynamic prototypes, the authors argue, can account for the flexibility in the semantics of CR without making reference to a "large number of ad-hoc dimensions and inconsistent scales" (Song and Lee 2011: 461), for which they criticize Whitton's (2006) proposal. Most importantly, they argue that prototypes of a category co-vary as the contexts vary, which allows a CR like *drink drink* to have varying denotations in varying contexts of use (Song and Lee 2011: 446). This represents a smart 'compromise solution': it retains the comparative simplicity of a prototype-based explanation, but does not assume that a CR is interpretable out of context, which many scholars (e.g. Whitton 2006: 13, Huang 2009: 139, Rossi 2011: 164) consider "impossible" in most cases.

Moreover, Song and Lee (2011: 446) maintain that a CR may refer not only to the (dynamic!) prototype of a category but also to a subcategory within a category and to a category itself. CRs referring to the subcategories in a category may be called 'polysemous' CRs. Based on the example *drink-drink*, they explain: “The complementary attribute ‘alcoholicity’ divides [the] category into two subcategories”, of which a CR may denote either the subcategory ‘alcoholic drink’ or the subcategory ‘nonalcoholic drink’ (Song and Lee 2011: 451–452). One of these subcategories may be the ‘default’ subcategory because it is more frequently referred to and has greater conceptual strength in the mind of the speakers (Song and Lee 2011: 452). Whitton (2006: 36) also mentions default construals of CRs for frequent and relevant distinctions, but is ultimately of the opinion that defaults may always be overridden by context. Finally, CR may refer to a category as such “where the category membership is at issue” (Song and Lee 2011: 454); as an example, the use of *dog dog* to mean a genuine dog instead of an unattractive person is cited. Crucially, Song and Lee (2011: 460) remark that prototypes and defaults are culturally conditioned; what constitutes a *salad-salad* or which of the interpretations of *drink drink* represents the default may vary significantly across speakers.

It is not only in terms of meaning that CR is flexible. Formally, it shows quite a bit of variability as well (Ghomeshi *et al*. 2004: 320): in addition to applying to whole phonological words (e.g. *We talked-talked*), it can apply to items smaller than a phonological word (*we talk-talked* – no copy of the inflectional ending) and also to units bigger than a word (e.g. *I like her like her*). For the latter option, the restriction seems to hold that “in addition to a single lexical contentful item, the scope of CR may only include non-contrastive ‘functional/grammatical’ morphemes” (Ghomeshi *et al*. 2004: 332).[2] It is further assumed that “status as a stored lexical unit plays an important role in defining the scope of CR” (Ghomeshi *et al*. 2004: 325); consequently, idioms like *kick the bucket* and compounds can only be copied in full.

The variability of the scope of CR is partly responsible for the different labels it has been given. Hohenhaus (2004), who extracted reduplications of single words (*talk talk*, but not *hate it hate it*) from corpora, described the phenomenon as a compound whose two elements are the same (Hohenhaus 2004: 299) and labelled it identical constituent compounding (ICC). Ghomeshi *et al*. (2004: fn. 8), however, also had access to examples involving reduplicated phrases, which led them to reject a compounding analysis as a “nonstarter” because English compounding does not involve phrases. As a consequence, they settled on the term CR, which is also adopted in the present paper[3].

2 Data and methodology

CR is notoriously difficult to find in reference corpora. Hohenhaus (2004) only found five instances in the entire British National Corpus (100 million words, ten million spoken), and Blauth-Henke (2008) points out the scarcity of CR in (French) reference corpora. More promising are sources of data that take CR's affinity for dialogic and/or informal settings into account. Hohenhaus (2004), for instance, was able to identify substantially more CRs in self-compiled corpora of TV and film scripts, blogs and online chat conversations (i.e. 35 tokens in ca. 19 million words).

The material for the present study is taken from three corpora: the Blog Authorship Corpus (Blog Corpus/BC; Schler *et al*. 2006, 138 million words), the Corpus of Contemporary American English (COCA; Davies 2008–, 450 million words) and the Corpus of American Soap Operas (SOAP; Davies 2012, 100 million words). BC, which contains texts created between 1999 and 2004, is the main source of data. Its design is illustrated in Table 1:

Table 1: Design of the Blog Authorship Corpus: word counts by age and gender of bloggers

	men	**women**	**sum**
aged 13–17	22,462,376	22,351,444	44,813,820
aged 23–27	32,313,980	33,850,008	66,163,988
aged 33–48	13,016,145	13,799,143	26,815,288
sum	67,792,501	70,000,595	137,793,096

Table 1 contains roughly equal numbers of words produced by men and women across three age groups. As the compilers only provide numbers of texts, the word counts were computed with Wordsmith Tools (Scott 2012).

The Blog Corpus has several advantages. To begin with, personal blogs represent a rather informal genre, which should improve the chances of finding CRs. Blogs allow for "written discourses of sometimes substantial length which have had no editorial interference" and "privilege linguistic idiosyncrasy" (Crystal 2006: 245, 246). The corpus size of almost 140 million words should facilitate retrieving a decent number of examples of such a rare phenomenon as CR. Finally, the availability of bloggers' gender and age (based on their online profiles) makes it possible to conduct a first study on the influence of social factors.

COCA and SOAP are included to research the use of CR in other genres. SOAP allows a closer look at what is considered the most typical context of use for CR, i.e. (scripted) dialogic face-to-face conversations. As both corpora contain American English material and CR is "quite common" in North America

(Ghomeshi *et al*. 2004: 308), these corpora should make good additional sources. For the Blog Corpus, no regional information is available: the bloggers come from all over the world. They may be native speakers of English, or use it as a second or foreign language or as a lingua franca.

A decisive advantage of the Blog Corpus is that it can be downloaded free of charge in its entirety and can thus be searched for reduplications via specialized software employing regular expressions. The programme *search-search* (Garretson 2015), which retrieves repeated strings of letters separated either by a space or a hyphen, was specifically created for the present study.[4] The freely accessible versions of COCA and SOAP are only available via an online search interface that does not support queries of that kind. This makes it impossible to extract all CRs from these corpora. As a workaround, the following procedure was applied: first, BC was searched for reduplications with the aid of *search-search*. Afterwards, I created a list of in total 291 CRs based on the results from BC and from an online list of CRs (Russell 2014). COCA and SOAP were then searched for these 291 items. Although this allows no direct comparison with the BC results, it nevertheless provides worthwhile information.

The search in BC was restricted to reduplications that contain either one-word or two-word bases for expediency, as identifying CR in the output of *search-search* requires a significant amount of manual post-processing. The majority of results are only formally identical to CR, such as the items in boldface in (7) – (10):

(7) in repetition of modifier:
you guys, i'm **so so** frustrated. (BC, 1046946)

(8) adjective modifying NP which contains same adjective as first element:
shes not only the best girlfriend in the world, but shes also the **best best** friend. (BC, 2102033)

(9) formal identity of object and object complement:
[…] people don't call **albums albums** anymore. Now they just call them CDs. (BC, 1538911)

(10) lexicalized word / expression including repetition:
I like creating **win win** situations. (BC, 3598030)

In addition to such commonly found repetitions, other types of reduplication need to be separated out, such as baby-talk reduplication (*night-night*), expressive sound words (*hush-hush*) or depreciative reduplication (e.g. *the 'neigh-*

bours' are suddenly eager to become buddy-buddy; BC, 843566). In the blogs, full reduplication may also be used to express activities or 'emotes' (*cough cough*, *sob sob*; cf. Crystal 2006: 190), and repeated elements are also found in words with multiple identical prefixes (*counter counter culture*, *post post post script*). These need to be filtered out.

In the end, identifying CR in written texts does not hinge on one criterion. Sometimes orthographic clues can help. Some writers capitalize or bold the reduplicant to represent the characteristic intonation of CR (e.g. *TIRED tired*). Hyphenation (e.g. *TIRED-tired*) may serve to separate CR from syntactic repetition, which is not hyphenated as a rule (Hohenhaus 2004: 307). Counter-examples do however exist, i.e. CRs without hyphenation and hyphenated words that are best analyzed as syntactic repetition. Of greater importance is therefore the context, which is only available in the form of co-text for written sources. To decide whether a CR reading is plausible and likely in a specific instance, we need to ascertain whether a contrast is (implicitly or explicitly) established or whether clarification may be needed or wanted. When there was not enough contextual evidence to make that decision, I ignored the example.

4 CR in contemporary English corpora

4.1 Overview: form and meaning

The search in the Blog Corpus yields 194 tokens of CR, with such diverse bases as adjectives (11), adverbs (12) and APs (13), nouns (14), prepositions (15) and PPs (16), pronouns (17), verbs (18) and VPs (19):

(11) my head hurts like hell and idk why im not **sick sick** like doctor sick i think [My head hurts like hell, and I don't know why. I'm not sick-sick, like doctor-sick, I think.] (BC, 4021779)

(12) Please note for future reference: I solemnly swear NEVER to read your blog again. I mean it. I really do. **Really-really**. (BC, 1855313)

(13) […] and he asks dazily "You want to go now? Like.. **right now right now**?" (BC, 1151815)

(14) after playing video games for like the entier afternoon […] i lazily put on some clothes [no i wasnt naked but i put on y'know, **CLOTHES CLOTHES**] and went to church for praise practice. (BC, 1637100)

(15) Incase things are getting too confusing in my life for some of you, I'm not **WITH WITH** Allison. I am pretty sure it was just a one night stand thing. (BC, 192879)

(16) i got all crazy panicky last night when he held me.[…] i freaked out. […] i realize that i'm not really **in love in love** with him. (BC, 664485)

(17) something that has been bothering me is me brother. I mean not **him him** but what he is doing with his life. (BC, 2874698)

(18) I got the links from this girl I know (well, not really **know know** as much as met online at a forum […]) (BC, 3288394)

(19) whats the point of a major that you hate? ok maybe i dont **hate it hate it**, but i find so little about it that i actually like. (BC, 152151)

The vast majority of CRs (170) were formed with a single word as a base (such as 11: *sick sick*); the remaining 24 examples, i.e. 12.4 per cent of all tokens, have two-word bases (such as 13: *right now right now*).

Table 2 shows the distribution of CR across different parts of speech / phrases:

Table 2: CRs in the Blog Corpus by type of base

Adj	**49**	25.3%	a lot, angry, awful, busy (2), cheap, close (2), crazy, cute, dead (2), depressed, different, dirty, done, drunk, dumb, fat, funny (2), gone, good, great, hot, last (2), late, official, old (2), poor (2), real, scary, serious (2), shallow, short, sick, single (2), skinny, stupid, sure, tired, weird, white, young
Adv(P)	**38**	19.6%	a lot (8), at all, away (2), back, for real, here, home (9), literally, a little, officially, really (8), together (3), right now
N	**67**	34.5%	apartment, biker (2), boss (2), break, cake, camping, class, clothes, cone, crush, date (2), fight, friend, friends (2), ghetto, God, group, home (2), house, journal, Kara, Korean, laundry, life, look, love (2), mail, man (3), meat, movie (2), office (2), party (3), patchwork, people, plan, relationship, Ruth, school (2), shit, shopping, snow, study group, talk, traffic, virginity, whore, work (7)
P(P)	**4**	2.1%	in, in love, over, with
Pro	**2**	1.0%	him, nothing
V(P)	**34**	17.5%	admitted, fight, flirt, focus, kissed, know (3), like (8), talk (3), talked (2), watched, work, hate it, kiss him, like her, like him (2), like me, like them, like us, likes me, make it, see them
sum	**194**	100.0%	

Nouns and adjectives represent the most frequent bases, with respectively 34.5 per cent and 25.3 per cent. This corroborates the findings in Hohenhaus (2004). His corpora of scripted film and TV dialogue, blogs and chat conversations yielded 35 CRs that were distributed across parts of speech as follows: 18 nouns (51.4%), 11 adjectives (31.4%), five verbs (14.3%) and one pronoun (2.9%). A striking difference can be seen in the figures for adverbs: while no adverbial CRs are found in Hohenhaus's (2004) corpora, the blogs boast 38 adverbs and adverb phrases, which makes them the third-largest group with a little under 20 per cent of all tokens. In part, this is a result of the searches that were conducted: the software used in Hohenhaus (2004) only isolated CRs with one-word bases, which would have excluded 12 of the tokens. Even considering that, though, there are still more adverbial bases than expected. It is, however, true that just a few common types make up the strength of this category (*a lot*, *home* and *really*). Prepositions and pronouns are rare.

Partially copied bases, such as in *talk-talked,* were not targeted by the search software. Yet, some partial copies of compounds or idiomatic phrases are recovered because parts of them happen to fit the requirements of the queries, i.e. consist of repeated identical character strings. They are shown in (20)–(22):

(20) In he mentions Gabe a lot. **A lot lot**. (BC, 1681913)

(21) do you think you would be able to study better in a study group (and i mean **STUDY STUDY group**) at someone else's house or by yourself at home? (BC, 1903669)

(22) can i borrow someone's dog so i can do the love scene for my dream...wait...i didn't mean **love 'love'** scene. (BC, 3604560)

In (20), only a part of *a lot* is reduplicated.[5] This occurs only twice in my data, while *a lot a lot* is found six times. Examples (21) and (22) represent the only two partially reduplicated compounds found in the Blog Corpus. While earlier research described copies of only parts of stored lexical units as unacceptable (cf. Section 2; Ghomeshi *et al*. 2004: 324–325), their presence in the blogs suggests that such CRs may be exceptional but nevertheless in use. In a recent talk by Laurence Horn (2015), the prohibition on partial copies is also questioned. As for the present examples, it is possible that the written medium played a role and that the presence of spaces between the orthographic words that form a lexical unit supported the partial copies.

By and large, the data from the other two mega-corpora support the results from BC. Table 3 depicts the distribution of CR across parts of speech in COCA and SOAP:

Table 3: CRs in COCA and SOAP by part of speech

	COCA	**SOAP**
Adj	30	27
Adv(P)	14	48
N	46	44
Pro	2	4
V(P)	12	20
Prep	0	3
Sum	**104**	**146**

As these figures were arrived at based on pre-existing lists and do not constitute an overview of all CRs in these corpora, direct comparisons with BC are not possible. Still, they suggest that the use of CR is dependent on genre: in the Blog Corpus, 1.39 CRs are encountered per million words. Based on the above figures (which must be assumed to be conservative), 1.46 CRs pmw are found in the soap dialogues, and just 0.24 CRs pmw in the mixed-genre COCA. The majority of the examples in COCA (57 of 104) come from the FICTION subcategory, which includes TV and film dialogues. While it is theoretically possible that an unfortunate selection of search terms misrepresents the actual conditions, I take these results as indications of a preference of CR for dialogues and blogs over other genres.

The COCA and SOAP data also provide further evidence that adverbial CRs are an important group: the most frequent bases in this group are *a lot*, *here* and *home* in COCA (with three occurrences each) and *together* (17 hits), *now* (10 hits) and *here* (six hits) in SOAP. In BC, *home*, *here* and *really* were in the lead, which shows some overlap but also differences. These similarities and differences across corpora in terms of ‘popular’ CRs deserve a closer look: Table 4 shows the most frequent CRs in the three corpora, i.e. all those that occur more than five times:

Table 4: Most frequent CRs in the three corpora surveyed (min. five occurrences)

BC	**COCA**	**SOAP**
like (+ pronoun) (15)	movie (6)	date (17)
home (11)	dead (5)	together (17)
a lot (9)	job (5)	friend(s) (11)
really (8)	like (5)	like (+ pronoun) (11)
work (8)	white (5)	now (10)
talk(ed) (6)		here (7)
		over (5)

There are more differences than similarities, which I take to be a reflection of two things. First of all, different topics predominate in these corpora. The soap dialogues are obviously very concerned with interpersonal relationships, which gives rise to high frequencies for *date*, *together*, *friend(s)* and *like*. Even *over* occurs predominantly as part of the phrase 'to be over someone'. Topics are more mixed in the multi-genre COCA and also in the Blog Corpus. Secondly, the type of communication engaged in (or fictionally represented) plays a role. It is noticeable that the adverbs *now* and *here* are only frequent in the soap transcripts – most likely because in dialogic face-to-face interactions, it is easiest to negotiate and interpret the meaning of CRs based on deictics, which are by nature context-sensitive. The adverbs *a lot* and *really*, which are frequent in the blogs, behave differently: their interpretation is always an intensified reading. One might argue that there is no other option than a default reading (in Whitton's sense) for these. At any rate, they work with little to no additional information.

Apart from the adverbs *a lot* and *really*, there are other candidates for emerging defaults in the corpora. *Like*, the only item on the list for all corpora (cf. grey shading in Table 4, nearly always conveys the same meaning: in 30 of the 31 cases, it refers to romantic and/or sexual attraction. Only one counter example is found in a blog discussing dress:

(23) Personally, I like wearing ties but I also like wearing skirts (not **LIKE LIKE** but it's nice to wear....) (BC, 4113926)

For *together together*, which is frequent in SOAP, similar observations can be made; it always refers to being involved in a romantic and/or sexual relationship. Since sexual and romantic relationships enjoy a special status in our society and are regularly distinguished from other interpersonal relationships such as friendships, the distinctions invoked by CRs involving *like* and *together* are frequent and relevant. Such "frequent and relevant" distinctions may lead to default construals (Whitton 2006: 18, 36). In these particular cases, sexual/ romantic *liking* and being *together* seem to have emerged as default readings of these CRs. In addition, CR can also be a vehicle for euphemism (Horn 1993: 49–51), which is common in conversation about intimate relationships.

In other cases, there seem to be no frequent, relevant and generally agreed-on distinctions that become defaults. For instance, *movie movie*, the most frequent CR in COCA, is applied to six different films for very different reasons. Three shall suffice to present the range of meaning covered here: *Good Morning Vietnam* is called a *movie movie* because it was produced for cinematic release as opposed to TV (COCA: SPOKEN, 1995); *Star Wars: A New Hope* because it is a "film that foregrounds the sheer pleasure of watching movies" (COCA: NEWS, 1999), and *Dead Again* because it is "a classic Hollywood thriller" (COCA: NEWS, 1991). There is apparently too much diversity here to arrive at a default for *movie movie*.

4.2 CR, its co-text and functions

To facilitate a correct reading, CRs are often part of a "specific type of co-text frame" (Hohenhaus 2004: 301) which involves either negation of the CR (*not home home, but…*) or the unreduplicated base (*not home, but …*), or an interrogative co-text (*do you mean home or home home?*) (cf. also Ghomeshi *et al.* 2004: 336). Less formulaic structures are also found, for example, the paraphrase *Mom and Dad home* for *home home* is mentioned in Hohenhaus (2004: 301).

In the Blog Corpus, 134 of the 194 CRs, i.e. about 69 per cent, are accompanied by a contrasting element in their vicinity to aid interpretation. In 75 cases (39 %), a synonym or paraphrase is used to achieve this aim. For 24 CRs (12.4%), both strategies are employed, as illustrated in (24):

(24) [I] am constantly there instead of here (but only not **here here** at this site and at the puter, i mean at my home) […]. (BC, 1835882)

In this example, the meaning of *here here* is clarified by a) paraphrasing it as *at this site and at the puter* [i.e. *computer*], b) contrasting it with *here*, the base in isolation, and c) contrasting it with *at my home*, which paraphrases *here*. This neatly illustrates that speakers may use more than one synonymous or contrasting element. It is no coincidence that (24), which represents the only occurrence of the deictic *here* in the written blog data, contains such copious information in the co-text. As the readers lack extralinguistic contextual information, the co-text clarifies the meaning of CR.

Different types of contrasts and synonyms are used to specify the meaning of CR. After inspecting all examples, I identified four categories, which are illustrated in the following examples: lexicalized compounds containing the base of CR (25), ad-hoc compounds (usually phrasal) containing the base of CR (26), the base of CR as a free lexeme (27), and larger explanatory phrases or clauses (28).

(25) Lexicalised Compound (here used as a contrast to CR):
I thought this was **school school**, not vocational school. I was ready to read. Not so eager to write cover letters. (BC, 780903)

(26) Ad-hoc compound (here used as synonymous expression for CR):
the webcam was **cheap cheap** - 15 bucks cheap. (BC, 72355)

(27) Base of CR (here used as a contrast to CR):
And I'm angry. Not **Angry-angry**, just angry. I don't want to like him. It's not fair that I should decide not to like him and then he treats me like he does and I can't help liking him. (718851)

(28) Phrase/clause (here used as a synonymous expression for CR):
I worked tonight […] word was that the boss wanted to have a word with me about my availability. I'm talking **boss boss**, the store director, the headiest honcho in the building. (BC, 1939766)

Their distribution in the Blog Corpus is shown in Table 5 and Table 6:

Table 5: Types of contrasts used with CR

contrast type	**freq.**	**% of all**
base	51	38.1%
phrase	50	37.3%
ad hoc compound	20	14.9%
compound	13	9.7%
SUM	134	100.0%

Table 6: Types of synonyms used with CR

synonym type	**freq.**	**% of all**
phrase	34	45.9%
base	33	44.6%
ad hoc compound	5	6.8%
compound	2	2.7%
SUM	75	100.0%

Whether used as a contrasting or synonymous element, unreduplicated bases and explanatory phrases are the most frequent options. Of course, there are also instances where no overt synonymous expressions or contrasts are used, as in (29):

(29) Other scary thing: Jon kind of coughed some thing about will you go out with me or some thing and I'm quite scared! Why is it that if you get close to some one they think that you **like them like them**? (BC, 1784456)

The extended context clarifies what is meant in this case. In addition, *like* has a widely accepted default construal Section 4.1.

Concerning the compounds, it is noteworthy that the ad-hoc compounds are formed explicitly to specify the meaning of CR: for instance, *15 bucks cheap* in (26) illustrates what *cheap cheap* means. There also seems to be an opposite trend, i.e. speakers employing a CR because they wish to create a parallel with a lexicalized compound or a phrase they are using. A first example, in which the compound in question is underlined, is provided in (30). At times, this is driven to extremes in the sense that CR seems to be chiefly employed for a playful or marked effect. Examples include (31) from an interview in COCA, and (32) from SOAP. The 'parallel structures' are underlined:

(30) i only get a two week break between summer school and **school school**. (BC, 780903)

(31) And we wrote a screenplay in a relatively short period of time, a fabulous screenplay. […] And then a **play play**. (COCA: SPOKEN, 2000)

(32) Well, you're spending more time thinking about your ex-wife than your actual **wife-wife**. (SOAP: Young and Restless, 2010-03-19)

It seems unlikely that the CRs in these examples primarily serve to disambiguate between denotations of a lexical item, as there is very little ambiguity here; using *school*, *play* and *wife* instead of their reduplications would not have impeded comprehension. This is evidence that CR has an additional, creative function and can serve as an "attention-seeking device" (Hohenhaus 2007: 23, cf. also Lipka 2000). This function seems to be especially in evidence when there is time to plan an utterance / a text: the examples above (30–32) come from written texts and prepared / scripted conversations. That is probably why the most blatant example of CR employed as wordplay (33) comes from a blog post:

(33) No, I am not talking about Bernard Laundry, the Quebec person. I am talking about **laundry laundry**. You know the cleansing of soiled garments by mean of water and detergent or dry chemicals.
(BC, 3186819)

The blogger then goes on to recount his adventures in unclogging the fabric softener dispenser of a washing machine.

Finally, CR contributes to building rapport. Where a CR is used, it is implicitly assumed that the interlocutors have sufficient shared common ground to establish the intended meaning/reference (Horn 2015). Horn (2015) argues that this indexes closeness. This function might be one of the reasons for the comparatively high frequency of CRs in blogs (cf. Section 4.1). Research by Stefanone and Jang (2007: 135) suggests that "blogs have been adopted as a mode of communication for strong tie network contacts", i.e. people the writers are close to. Among many others, CR seems to be one of the linguistic cues that index such relationships and even help create them.

4.3 CR and its users

Figure 1 and Table 7 provide an overview of the use of CR by age group and gender (based on how users self-report this information in their profiles) in BC:

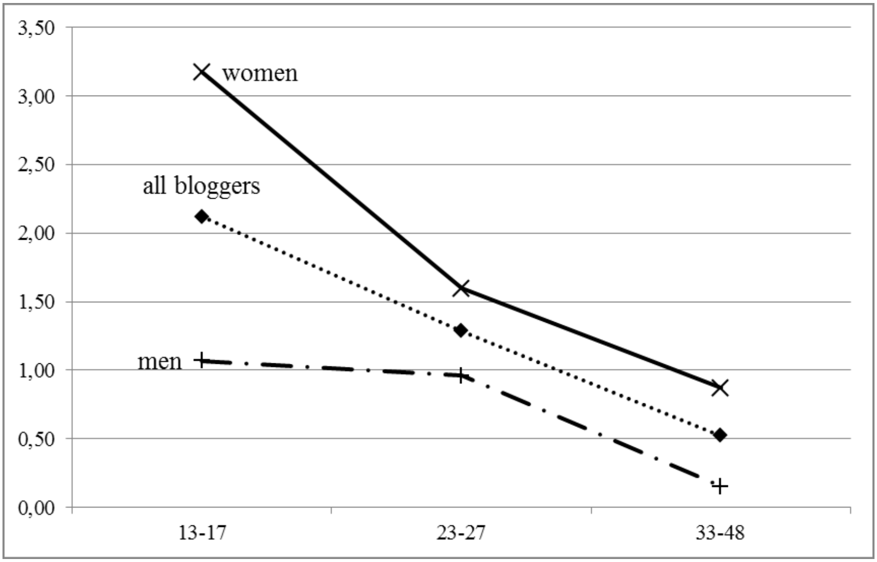

Figure 1: CRs pmw in the Blog Corpus by gender and age

Table 7: CRs in the Blog Corpus by gender and age

	men		women		sum	
	abs.	pmw	abs.	pmw	abs.	pmw
13–17	**24**	1.07	**71**	3.18	**95**	2.12
23–27	**31**	0.96	**54**	1.60	**85**	1.28
33–48	**2**	0.15	**12**	0.87	**14**	0.52
Sum	**57**		**137**		**194**	

At this point, it should be mentioned that individual bloggers may use more than one CR during the five years that are chronicled in BC. This circumstance is rare enough not to skew the data[6], though, which is why the numbers in Table 7 and Figure 1 are an adequate basis for discussion.

It emerges that CR is correlated with age group and gender. The younger the bloggers, the more instances of CR are reported: among the 13–17 year-olds, 2.12 CRs pmw are found. This figure drops to 1.28 for bloggers in their twenties and to 0.52 for bloggers between 33–48. An χ^2 test based on the raw frequencies and the word counts in the age-stratified subcorpora reveals that the differences between the age groups are significant at $p<0.001$ (χ^2 32.95, df = 2). CR, an informal feature (cf. e.g. Hohenhaus 2007: 26), is likely to be more accepted among younger writers, who often exhibit a very informal style in other respects as well, such as with regard to spelling (34):

(34) i noe of sumone hu likes cookieZ. its actualli sumone i dun realii like ...ok not H8 but i jus can barely tok to her up to 2 sentences? (CR, 3896217)

Especially in written language, older speakers may shy away from using features that are considered very informal.

CR is significantly more frequent in women's blogs than men's blogs ($p<0.001$, χ^2 29.70, df = 2). This pattern is not only true globally but also across all ages. In each age group the frequency per million words among women is higher than that among men. This might be a reflection of the types of blogs women tend to write. There is research to indicate that women bloggers favour "more personal content and orientation towards the social aspects of blogging, as opposed to a male emphasis on information" (Pedersen and Macafee 2007: 1487). This 'personal content' leaves more leeway for the kind of informality and linguistic creativity found e.g. in CR.

In the end, I do not think that CR is associated with a particular gender or age group as such. However, I believe that most speakers only consider it acceptable for use in informal contexts. This impression is reinforced by a negative metalinguistic comment on CR: while only five instances of explicit commentary on CR were found (two in the blogs, one in SOAP and two in COCA), it is noteworthy that four of them are negative. Two of them are shown in (35) and (36):

(35) Lily: Oh, so you [and Karen] are, like, uh, a hang out thing rather than a **thing-thing**?
Neil: College has done wonderful things for your vocabulary.
(SOAP, YR 2007-09-17)

(36) You are doing the dishes Friday when the phone rings. Wipe your hands on your jeans and answer it. "I'm here," he says. Say: "You're here? **Here-here**?" **Here-here**? You think. What a retard you are. Do I like him or **like him-like him**? It's **bad-bad**.
(COCA: FICTION - Bochan, *How to have a visitor*, 2011)

The ironic remark in (35), which may just as well have been prompted by the dummy compound *hang out thing* as by the CR *thing-thing*, can be considered gentle mocking, but the character's self-talk in (36) is much harsher and even makes use of the slur 'retard'. Such comments make it obvious that CR is not accepted by all speakers, and not deemed appropriate for all contexts.

5 Conclusion

This study investigated CR in three English mega-corpora (BC, SOAP, COCA) with the objective of exploring in particular its formal properties, the co-text frames it occurs in, its interpretation and its functions in discourse.

On the basis of data from the Blog Corpus (Schler *et al.* 2006), it was shown that CR mostly targets nouns, adjectives and adverbs. Adverbial CRs are more frequent than assumed, representing almost a fifth of all tokens. The preferred strategy to specify the meaning of a particular CR via the co-text is to use contrasting elements. It is particularly interesting, though, that ca. 29 per cent of CRs were accompanied by both contrasting elements and synonyms at the same time. This shows that bloggers take care to provide clues to the interpretation of CR. This may at least partly be due to the nature of blog entries: in contrast to spoken face-to-face interaction, a blog lacks extralinguistic clues and immediate feedback.

As for the interpretation of CR, it appears that very few CRs are candidates for emerging default construals which allow an out-of-context interpretation of a CR. In a list of CRs that are frequent in all three corpora, three groups can be distinguished: 1) only one interpretation possible: *a lot a lot* and *really really* (intensified reading as default); 2) emerging defaults found: *like like* and *together together;* 3) no defaults identified (example: *movie movie*). The latter is the case for the majority of CRs. This makes sense in so far that default construals can only be expected for very frequent and relevant distinctions (Whitton 2006: 36). On the whole, it seems that its versatility and flexibility makes CR useful to speakers in the first place.

Functionally, CR is not limited to disambiguation and euphemism. It is used in wordplay and to build rapport. The users of CR are mainly young bloggers, especially young women. These writers seem to focus most on the social component of blogging and on personal narrative. Among all groups, they are also most open to using informal style in the written medium. Like other informal features, CR is not accepted in all texts and contexts. In blogs, especially personal blogs, it is quite frequent, though, compared with other genres.

Methodologically, the present study has some interesting implications. I believe that the analysis of CR benefits greatly from the use of corpus methods. Most importantly, automatic searches for reduplicated strings can uncover CRs that may go unnoticed otherwise, and eventually refine our understanding of CR. It is crucial, though, that corpus methods are accompanied by a qualitative component involving close reading of the examples (cf. Curzan and Palmer 2006, where the importance of a combined approach is discussed with regard to historical linguistics). Although CR has been used at least since the 1950s (Whitton 2006: 8), it is especially its presence in a growing body of informal written text, e.g. in computer-mediated genres like blogs, in recent years that has opened up new possibilities of researching the phenomenon.

Finally, English is not the only language with contrastive reduplication, as the German example in (37), retrieved from an online forum, illustrates (cf. also Finkbeiner 2014):

(37) [I]ch habe irgendwie 2 zu Hause. So vor Wochenenden, wenn ich sage, ich gehe jetzt nach Hause, fragen meine Arbeitskolleginnen dann häufig: "Hier heim oder **Heim heim**?
[Somehow, I have two homes now. Before the weekends, when I say “I’m driving home now”, my colleagues often ask “here home or **home home**?”][7]

In future work, it would be interesting to systematically examine the phenomenon in a cross-linguistic perspective, ideally based on corpora similar to the Blog Corpus.

Notes

1. In the present paper, all examples are represented in the original (i.e. including nonstandard spellings, hyphenation, capitalization etc.), with the exception that all CRs are in boldface.
2. For a very detailed account of the scope of CR, the reader is referred to Ghomeshi et al. (2004).
3. For reasons of space, I pass over the related discussion whether CR is to be considered a morphological or a syntactic phenomenon (cf. e.g. Travis 2001: 11 for arguments for a syntactic interpretation, and Hohenhaus 2004 or Ghomeshi et al. 2004 for differing opinions).
4. Alternatively, the Reduplication Finder (Fessl 2006), which is available online free of charge, may be used. However, it is quite slow when processing large amounts of data.

5. It is of course possible that speakers considered *lot* (not *a lot*) as the base for reduplication here.
6. If just the first CR uttered by each blogger is considered, the overall picture that emerges is the same. In the age group 13-17, we find 62 CRs produced by women, 21 by men; among the bloggers aged 23-27, 42 CRs were written by women and 27 by men; and the last group (33-48) yields ten CRs by women and two by men.
7. retrieved from http://www.spin.de/forum/msg-archive/3/2014/03/556367?page=3; comment by user 'kägifrettli' on 14 March 2014, 21:51

Acknowledgements

I am very grateful to Gregory Garretson for creating the *search-search* programme that greatly facilitated the search for CR. My thanks also go to the delegates at ICAME 36 in Trier who were kind enough to discuss this project with me, and to the anonymous reviewer, who provided very helpful comments. Special thanks are due to Barbara Güldenring and Magnus Huber for their valuable feedback on earlier versions of this paper. All remaining mistakes are mine.

References

Blauth-Henke, Christine. 2008. Reduplikation in Korpora. Zum Zusammenhang von Methodenreflexion und Forschungsgegenstand. In S. Buch, Á. Ceballos and C. Gerth (eds.). *Selbstreflexivität: Beiträge zum 23. Forum Junge Romanistik (Göttingen, 30.05.–2.6.2007)*, 35–50. Bonn: Romanistischer Verlag.

Crystal, David. 2006. *Language and the Internet.* 2nd edn. Cambridge: Cambridge University Press.

Curzan, Anne and Chris C. Palmer. 2006. The importance of historical corpora, reliability, and reading. In R. Facchinetti and M. Rissanen (eds.). *Corpus-based studies of diachronic English*, 17–36. Bern and New York: Peter Lang.

Davies, Mark. 2008– . The corpus of contemporary American English: 450 million words, 1990–present. Available at: http://corpus.byu.edu/coca. Accessed 4.6.2015.

Davies, Mark. 2012. The corpus of American soap operas: 100 million words, 2001–2012. Available at: http://corpus2.byu.edu/soap. Accessed 2.5.2015.

Dray, Nancy. 1987. Doubles and modifiers in English. Unpublished MA thesis, University of Chicago.

Fessl, Angela. 2006. Reduplication finder 1.1. Available at: http://reduplication.uni-graz.at/. Accessed 1.10.2014.

Finkbeiner, Rita. 2014. Identical constituent compounds in German. *Word Structure* 7 (2): 182–213.

Garretson, Gregory. 2015. Search-search: Reduplication finder. (Software).

Ghomeshi, Jila, Ray Jackendoff, Nicole Rosen and Kevin Russell. 2004. Contrastive focus reduplication in English (The salad-salad paper). *Natural Language & Linguistic Theory* 22 (2): 307–357.

Hohenhaus, Peter. 2004. Identical constituent compounding – a corpus-based study. *Folia Linguistica* 38 (3–4): 297–332.

Hohenhaus, Peter. 2007. How to do (even more) things with nonce words (other than naming). In J. Munat (ed.). *Lexical creativity, texts and contexts,* 15–38. Amsterdam and Philadelphia: John Benjamins.

Horn, Laurence R. 1993. Economy and redundancy in a dualistic model of natural language. In S. Shore and M. Vilkuna (eds.). *SKY – Yearbook of the Linguistic Association of Finland*, 33–72. Helsinki.

Horn, Laurence R. 2008. Pragmatics and the lexicon. In P. van Sterkenburg (ed.). *Unity and diversity of languages*, 29–42. Amsterdam and Philadelphia: John Benjamins.

Horn, Laurence R. 2015. The lexical clone: Pragmatics, prototypes, and productivity. Presentation at 37. Jahrestagung der DGfS, Leipzig. 5.3.2015.

Huang, Yan. 2009. Neo-Gricean pragmatics and the lexicon. *International Review of Pragmatics* 1 (1): 118–153.

Lieber, Rochelle. 2009. Identical-constituent compounds. In R. Lieber and P. Štekauer (eds.). *The Oxford handbook of compounding*, 364–365. Oxford: Oxford University Press.

Lipka, Leonhard. 2000. English (and general) word-formation – the state of the art. In B. Reitz and S. Rieuwerts (eds.). *Anglistentag 1999 in Mainz: Proceedings*, 5–20.Trier: Wissenschaftlicher Verlag.

Pedersen, Sarah and Caroline Macafee. 2007. Gender differences in British blogging. *Journal of Computer-Mediated Communication* 12 (4): 1472–1492.

Rossi, Daniela. 2011. Lexical reduplication and affective contents: A pragmatic and experimental perspective. *Belgian Journal of Linguistics* 25: 148–175.

Russell, Kevin. 2014. Corpus of English contrastive focus reduplications. Available at: http://home.cc.umanitoba.ca/~krussll/redup-corpus.html. Accessed 20.7.2015.

Schler, Jonathan, Moshe Koppel, Shlomo Argamon and James Pennebaker. 2006. Effects of age and gender on blogging. In N. Nicolov, F. Salvetti, M. Liberman and J. H. Martin (eds.). *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 199–205. Available at: http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-039.pdf. Accessed 1.9.2015.

Scott, Mike. 2012. WordSmith Tools version 6. Stroud: Lexical Analysis Software.

Song, Myoung Hyoun and Chungmin Lee. 2011. CF-reduplication: Dynamic prototypes and contrastive focus effects. In N. Ashton, A. Chereches and D. Lutz (eds.). *Proceedings of the 21st Semantics and Linguistic Theory Conference (SALT 21)*, 444–462. Available at: http://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/2590.
Accessed 1.9.2015.

Stefanone, Michael A. and Chyng-Yang Jang. 2007. Writing for friends and family: The interpersonal nature of blogs. *Journal of Computer-Mediated Communication* 13 (1): 123–140.

Stolz, Thomas, Cornelia Stroh and Aina Urdze. 2011. *Total reduplication*: *The areal linguistics of a potential universal.* Berlin: Akademie Verlag.

Taylor, John R. 2003. *Linguistic categorization.* 3rd edn. Oxford: Oxford University Press.

Whitton, Laura. 2006. The semantics of contrastive focus reduplication in English: Does the construction mark prototype-prototype? Unpublished manuscript, Stanford University.