To appear in: Andrew Wilson, Paul Rayson and Tony McEnery, eds., *Corpora by the Lune: a festschrift for Geoffrey Leech*, Peter Lang, Frankfurt.

Do women and men really live in different cultures? Evidence from the BNC

Hans-Jörg Schmid, University of Bayreuth, Germany

1. Introduction

In her bestseller *You just don't understand* Deborah Tannen tried to show that "talk between women and men is *cross-cultural* communication" (1990: 18; my emphasis). A little earlier, she had argued that

> male-female conversation is always cross-cultural communication. Culture is simply a network of habits and patterns gleaned from past experience, and women and men have different past experiences. From the time they're born, they're treated differently, talked to differently, and talk differently as a result. Boys and girls grow up in different worlds, even if they grow up in the same house. And as adults they travel in different worlds, reinforcing patterns established in childhood. (Tannen 1986: 60)

As in the work of her main forerunner, Robin Lakoff (1975), Tannen's claims concerning women's and men's speech styles are based on evidence of a rather unsystematic kind. Transcripts of everyday conversations, stories of and by friends, relatives and students, extracts from fiction and drama, and other pieces of more or less anecdotal evidence are interspersed with references to experimental studies from developmental psychology and sociology. That notwithstanding, the huge number of sold copies indicates that Tannen certainly managed to strike a chord with linguistically (or psychologically) inclined laypersons.

Two years after Tannen's book came out, Geoffrey Leech and Roger Fallon (1992) published their paper "Computer corpora - what do they tell us about culture?". They showed that the frequencies of words from a dozen everyday domains in the Brown and the LOB corpora mirror the importance of certain concepts in American and British culture. Words concerned with firearms like *bullet(s), gun(s), rifle(s)* or *shot,* for example, were found to occur much more frequently in Brown than in LOB (Leech and Fallon 1992: 40, 49), and this can certainly be said to reflect the greater interest in this domain in the USA. Closer to my present concerns, Leech and Fallon pointed out (with reference to earlier comparisons carried out by Hofland and Johansson 1982: 32-40) that "the American corpus appears to be more extreme in its 'masculinity' than the British corpus: *he, boy* and *man* are more fully represented in Brown, whereas *she, girl* and *woman* are more fully represented in LOB" (1992: 30f.). In a note, Leech and Fallon expressed their hope that "by the year 2000, it will be possible to make use of these corpora [i.e. BNC and COBUILD] for *cross-cultural* studies on a much larger scale than is now possible on the limited basis of the Brown and LOB corpora" (1992: 47; my emphasis). I am not sure whether what they had in mind were studies across the male and female cultures, but it

is certainly to a large extent due to Geoff Leech's own contribution to corpus linguistics that their hopes were not in vain and studies of this kind have now become feasible.

In 1997, after the publication of the BNC, Leech did, in fact, look at the social differentiation in the use of English vocabulary with regard to the parameters gender, age and social group (Rayson, Leech and Hodges 1997). The focus of this joint paper, however, is less on the cultural implication of the usage of vocabulary of different social groups than on opening new avenues of research in corpus-based research in this field and illustrating some of the possibilities.

Combining Tannen's claims with Leech and Fallon's simple but groundbreaking method, we arrive at an obvious challenge: can corpora tell us whether women and men indeed live in different cultures - at least as far as their conversational styles are concerned? I will take up this challenge in the present paper.

2. Methodological issues

The method used for this study is borrowed from Leech and Fallon (1992). I am going to compare frequencies of words and collocations in two different corpora. The two corpora used are both taken from the spoken section of the BNC: they consist of all utterances that are marked up as being spoken by either a woman or a man respectively. According to the Zurich BNC*web* Query System (Lehmann, Hoffmann and Schneider 1996-1998), with which all searches reported here have been carried out, these two corpora consist of 4,918,075 words spoken by men and 3,255,533 spoken by women. To my knowledge, these two parts of the BNC are not only by far the largest but also the most contextually and demographically balanced samples of women's and men's spoken language available at present.

Rayson, Leech and Hodges (1997) did not use the same set of data from the BNC for their research but restricted their attention to the demographically-sampled part of the BNC (the "Conversational Corpus"), presumably because this is the most reliable part as far as the mark-up of social parameters is concerned, and because it consists of everyday spontaneous interactive discourse and excludes other spoken genres, especially more formal ones like radio interviews, public speeches, committee meetings, or face-to-face and telephone conversations at work. The difference between the two data sets used in Rayson, Leech and Hodges (1997) and here leads to interesting divergences in the results which will be discussed in Section 7 below. One observation worth mentioning at this point is the overall amount of data contributed by men and women to the two subcorpora. In Rayson, Leech and Hodges' Conversational Corpus, male speakers account for 1,714,443 of the total of 4,552,255 words and women for 2,593,452. Thus "for every 100 word tokens spoken by men in the demographic corpus, 151 were spoken by women" (Rayson, Leech and Hodges 1997: 137), and this is true even though the number of male and female speakers in the Conversational Corpus is almost identical. The skewage is due to two facts: women contribute a larger number of turns, and, on average, their turns are a little longer than those of men. As the numbers given in the preceding paragraph indicate, in the 8,173,608 words used in this study this relation is precisely reversed: for every 100 words spoken by women, there are 151 spoken by men. Since according to Aston and Burnard (1998: 120), the numbers of utterances spoken by women and men in the whole spoken section of the BNC are roughly the same (307,539 female utterances as opposed to 304,278 male ones), the overrepresentation of men can only be due to the fact that their average turn is considerably longer than the women's. It is probably a quite safe guess that this reversal reflects the well-known claim that women are linguistically more active and productive in the private domain, while men tend to contribute a larger amount of talk in public situations (Tannen 1990: 76ff.).

The Zurich *BNCweb* Query System gives, in addition to concordances and other common display features, both absolute frequency scores and scores per million words (relative to the respective extract from the whole corpus) for all words and collocations queried. Both of these scores will be used in this study, but for different purposes.

The normalized scores per million words are used as input into a coefficient formula which is taken over from Leech and Fallon (1992), who in turn borrowed it from Holland and Johansson (1982). The application of the formula to the present question is given in the following figure:

$$\frac{\text{Frequency per million words}_{\text{MEN}} - \text{Frequency per million words}_{\text{WOMEN}}}{\text{Frequency per million words}_{\text{MEN}} + \text{Frequency per million words}_{\text{WOMEN}}}$$

Figure 1: Difference coefficient formula (based on Leech and Fallon 1992: 30)

The values for this coefficient range from 1.00 to -1.00. If a word is equally frequently used by women and men in the two sections of the BNC, the score is 0.00. Negative scores mean that a word occurs more frequently in utterances attributed to women, positive ones that it is more often used in male utterances. The hypothetical value 1.00 - which is never reached in the actual data - means that a word only occurs in utterances marked up as male, and the value -1.00 that it only occurs in utterances attributed to women.

The absolute frequencies of occurrence, which cannot be used for the coefficient because the two corpora differ in size, are used to calculate the significance level of the differences with the hypergeometrical approximation of the binomial distribution (see e.g. Hartung 1999: 202-209). I have decided to choose this statistical measure rather than the much more widely used chi-square test because strictly speaking, the latter must only be applied when it is guaranteed that the individual data are independent from each other. Since speakers in both corpora have supplied more than one single occurrence of certain words or expressions, this precondition for the use of the chi-square test is

not met.[1] It must be emphasized that the binomial test imposes stricter requirements on significance than the chi-square test, especially when the observed frequencies of items are fairly low. Had the chi-square test been applied to the data presented here, almost all observed differences would have turned out significant on the 99% level.

In their study with the Brown and LOB corpora Leech and Fallon (1992: 34f.) overcame the problem of multiple meanings of lexemes by introducing a two-stage procedure. In the first stage, they collected frequency lists of graphic forms. In order to make sure that the forms were comparable from a semantic point of view as well (i.e. that only the intended senses of polysemous lexemes were contrasted in the two corpora), they checked all occurrences in KWIC-concordances before fixing the final comparative scores. This procedure was not feasible with the material for the present study. For one thing, the raw corpora amount to more than 8 million words, more than four times as many as the Brown and LOB corpora taken together. While this has the welcome effect that the material is more representative and reliable, it also renders the manual inspection of concordances quite time-consuming. Indeed, many of the forms investigated are so frequent in the 8 million words that manual sense-differentiation would have turned into a major research project in its own right. Therefore, a more practical way out was chosen for this study: words with several fairly equally-distributed senses were excluded from the list of test items, while monosemous lexemes, and those with one clearly predominant sense, were favoured. Since most scores for the latter type of words were fairly high, it could be assumed that the unintended (and rare) senses would not distort the results too much. The only kind of prior differentiation that was carried out was not a semantic but a grammatical one: word-class tags were added to all grammatically ambivalent graphic forms in the queries (e.g. *look=NN1* vs. *look=VVB* and *look=VVI*).

3. The domains investigated

When the first ideas for the present study were born, my aim was to investigate some of the well-known examples of linguistic gender-markers compiled for example by Lakoff in her classic and much-quoted description of "women's language" (1975: 53ff.). 'Women's words' like *lovely, charming, divine, adorable,* men's alleged predilection for swearwords, and linguistic signs of the alleged uncertainty of women like the hedges *sort of, maybe* and many others were obvious starting-points for the intended comparison. Some of these words had already been investigated by Rayson, Leech and Hodges (1997) and, as in their paper, with a few notable exceptions my corpus findings clearly confirm the expectations of the gender-linguistic literature.

I would like to thank my colleagues from Bayreuth University, Prof Wiebke Putz-Osterloh (Psychology) and Prof Helmut Rieder (Mathematics), for their advice on the appropriate test of significance, and Dipl. Math. Matthias Kohl for his help with its application and implementation.

For two reasons, however, this did not seem satisfactory. For one thing, this procedure would have exploited the corpus data for nothing more than a confirmation of what was to be expected anyway. How much more exciting did it seem to utilize the two subcorpora to discover something new! On the other hand, a strange feeling was beginning to creep up on me that the differences in frequencies of usage by women and men that I found could be artifacts of some unknown feature of the BNC and that, therefore, they would be found for perfectly normal everyday words, too. When random words were spot-checked, the latter suspicion was in fact confirmed: it turned out that even perfectly innocuous words like *house, breakfast* and *car* were not equally distributed across the two subcorpora either. However, when larger numbers of hypothetically neutral words were tested, it soon transpired that the observed differences were neither due to mere chance nor did they simply seem to be a result of the composition of the BNC. They appeared to represent the tip of a much more exciting iceberg, whose precise nature will be discussed further down (see Section 7). It was this recognition that sparked off a massive extension of the scope of this study. As a result, findings on words and collocations from the following domains can be reported on:

- Conversational behaviour: 'women's words', hesitation and hedges, minimal responses, questions
- Domains with expected female preponderance: clothing, colours, home, food and drink, body and health, personal reference, personal relationships, temporal deixis
- Domains with expected male preponderance: swearwords, car and traffic, work, computing, sports, public affairs, abstract notions

The words and collocations queried for each of these domains were selected on the basis of gut feeling. In the present exploratory stage of large-scale gender-cultural corpus linguistics, principled decisions on the choice of words did not yet seem to be necessary. My domain-related method complements that of Rayson, Leech and Hodges (1997) who looked for high chi-square values in order to select those words that are particularly good markers of gender and other social differences.

4. Data on conversational behaviour

This section, just like Section 6, consists mainly of tables representing the scores of words and collocations. All tables have the same design: the five columns give the words, their relative frequencies per million words for MEN,[2]

[2] At this stage, I am only presenting the findings from the two subcorpora and not mounting any claims concerning the linguistic behaviour of women and men as such (whatever that might be; see Section 7 for a discussion). In order to avoid the danger of making statements like "women use word *X* Y times as often as men" for the time being, I am referring

their relative frequencies per million words for WOMEN, the value of the difference coefficient and the significance level. The significance levels are 99% (indicated by an *a)* and 95% (indicated by *b)*. To save space, two tables will always be juxtaposed. The tables will only be accompanied by short comments explaining the reasons why certain domains or expressions were chosen and drawing attention to particularly interesting aspects of individual words or scores. More general conclusions will be drawn in Sections 5 and 7.

4.1 'Women's words'

The list of adjectives and adverbs that have traditionally been attributed to women (Jespersen 1922: 249, Lakoff 1975: 11-13, 53) clearly meets the expectations raised by the literature (see Table 1). The favourite in WOMEN is undoubtedly *lovely*, which boasts quite a high frequency of occurrence and is found more than three times as often in WOMEN than in MEN. Jespersen's example of a typical female intensifier, *vastly*, is obviously fairly rare these days and is more frequently found in MEN than in WOMEN. It must be added, however, that the frequencies of *horribly*, *tremendously* and *vastly* are very low and therefore not reliable; the differences are statistically not significant. *Pretty* is much more frequent as an adverb (tag AVO) than as an adjective (AJO) and, interestingly, it is only as an adjective that it is used more frequently in WOMEN than in MEN; as an adverb, it is more frequent in MEN.

Table 1: 'Women's words'

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
handsome	2.03	6.76	-0.54	a
lovely	131.35	421.13	-0.52	a
sweet	18.50	56.62	-0.51	a
horrible	32.74	91.23	-0.47	a
dreadful	9.15	24.27	-0.45	a
awfully	6.51	16.89	-0.44	a
pretty=AJ0	7.12	18.34	-0.44	a
disgusting	14.44	36.25	-0.43	a
awful	77.67	137.30	-0.28	a
terrible	68.43	108.43	-0.23	a
charming	2.85	4.30	-0.20	a
appalling	9.76	11.67	-0.09	
terribly	23.59	24.57	-0.02	
horribly	1.22	1.23	0.00	
pretty=AV0	115.40	104.74	0.05	
tremendously	5.49	3.69	0.20	
vastly	3.46	2.15	0.23	

Table 2: Hesitators and hedges

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
well=ITJ	120.98	427.27	-0.56	a
really	1406.65	2098.58	-0.20	a
, you see	112.24	152.97	-0.15	a
, you know	837.73	1060.35	-0.12	a
I mean	1751.50	2142.20	-0.10	a
I mean#you know/<s>	263.11	320.68	-0.10	a
well#I mean/<s>	273.89	317.61	-0.07	b
I think	2349.50	2398.07	-0.01	
maybe	302.15	279.22	0.04	
erm	6.011.91	4.480.99	0.15	a
perhaps	460.14	327.75	0.17	a
sort of	709.63	474.27	0.20	a
I guess	18.91	10.75	0.28	a
er	9.970.16	5.036.96	0.33	a
in fact	337.94	164.03	0.35	a

to the corpora by the use of small capitals (WOMEN, MEN) and using these like references to separate corpora such as Brown or LOB.

4.2 Hesitators and hedges

The items collected in Table 2 cover a range of clear examples of hesitators *(er* and *erm)* over functionally ambiguous discourse markers like *well, I mean* and *I think* to fairly clear cases of hedges *(sort of, maybe, perhaps)*. Good candidates for a common motivation behind the use of all these expressions are tentativeness and uncertainty. As is well known, these conversational traits are usually attributed to women (cf. e.g. Lakoff 1975: 53-55, Coates 1986: 102). The actual dataset, however, does not confirm this admittedly simplistic approach (see Coates 1996: 152ff. for a more differentiated view on hedges).

The first striking observation in the present data is that the clear hesitators *er* and *erm* occur much more frequently in MEN than in WOMEN. This finding supports one aspect of an otherwise highly dubious remark by Jespersen on the articulatory and rhetorical skills of women and men:

> In language we see this very clearly: the highest linguistic genius and the lowest degree of linguistic imbecility are very rarely found among women. The greatest orators, the most famous literary artists, have been men; but it may serve as a sort of consolation to the other sex that *there are a much greater number of men than women who cannot put two words together intelligibly, who stutter and stammer and hesitate,* and are unable to find suitable expressions for the simplest thought. Between these two extremes the woman moves with a sure and supple tongue which is ever ready to find words and to pronounce them in a clear and intelligible manner. (Jespersen 1922: 249; my emphasis)

The other items that occur significantly more often in MEN than in WOMEN, viz. *in fact, I guess, sort of* and *perhaps*, are of a fairly mixed kind. *In fact* has a rather factual and objective ring to it, while *I guess* carries precisely the opposite tone of subjectivity and uncertainty. *Perhaps* is fairly formal while *sort of* is colloquial. A tendency, let alone a coherent pattern, does not emerge from this section of the data, partly because of the distinct context-dependence and polyfunctionality of these items. More detailed research using the concordances must be carried out here before a clearer picture can emerge.

The same is of course even truer of the discourse marker *well* with its multiple functions (see Schiffrin 1987: 105ff). If the main function of *well* is indeed to mark dispreferred seconds in adjacency pairs and other potentially face-threatening utterances, as Schiffrin claims, the enormous overrepresentation in WOMEN is indeed remarkable and illuminating.

The markers *you see* and *you know* are clearly addressee-oriented. The fact that they are found more often in WOMEN than in MEN ties in with the data on minimal responses and questions (see Sections 4.3, 4.4 and 5). *I mean* and its combined occurrences with *you know* and *well* may presumably be interpreted as fairly clear evidence of a relatively higher linguistic uncertainty in WOMEN.

4.3 Minimal responses

Minimal responses are means of lubricating conversations, of showing the other discourse participant(s) that one is paying attention to what they are say-

ing and willing to continue listening. According to Tannen (1990: 142), women and men tend to interpret minimal responses in fundamentally different ways, but this cannot be tested in the corpus. Long before Tannen, Zimmermann and West (1975; see also Coates 1986: lOOff.) had claimed that men tend to be more parsimonious in providing this type of conversational support. This behaviour, they argued, helps them to signal their lacking enthusiasm for topics chosen by the other discourse participant(s) and to thereby control or even dominate the choice of topics.

At first sight, the data on minimal responses collected in Table 3 are not all that coherent. On closer inspection of the individual items, this inconsistency can be resolved, however. The only three expressions with preponderance in MEN, *yep, you 're right* and *okay* differ qualitatively from the rest (perhaps with the exception of *that's right)* insofar as they can be used to close down rather than carry on topics and may thus curb the other speakers' enthusiasm to speak rather than encourage them. They can be used to acknowledge what the other person has said but in contrast to the supportive markers *mm, yeah* or even *no,* they convey the feeling that one regards the matter at hand as settled and wants to discuss, or even do, something else.

All other types of minimal responses listed in Table 3 have been found to occur more often in WOMEN than in MEN. These scores appear to confirm the claim (cf., e.g., Coates 1986: 116f., 1989: 95ff, Tannen 1990: 195ff. et passim) that women tend to behave in a more cooperative and supportive way in conversation than men, especially in all-female conversations.

Table 3: Minimal responses

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
mm	2101.23	4.500.65	-0.36	a
aha	177.92	326.83	-0.30	a
yes, but	38.43	64.51	-0.25	a
no=ITJ	3530.45	5547.79	-0.22	a
mhm	553.67	797.41	-0.18	a
yeah	6.712.99	9.309.07	-0.16	a
yes=ITJ	3354.16	4124.67	-0.10	a
that's right	519.11	548.60	-0.03	
yep=ITJ	125.66	117.34	0.03	
you're right	28.06	23.34	0.09	
okay	1414.58	696.35	0.34	a

Table 4: Questions

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
aren't you	55.51	126.86	-0.39	a
why don't you	28.47	60.51	-0.36	a
couldn't you	14.84	26.42	-0.28	a
are you	486.37	828.44	-0.26	a
can you	289.34	468.13	-0.24	a
isn't it	403.61	617.41	-0.21	a
wouldn't you	29.28	44.23	-0.20	a
would you	251.52	283.52	-0.06	
can I	287.51	281.37	0.01	
could you	95.16	82.01	0.07	
could I	47.99	34.40	0.16	a

4.4 Questions

Questions are a notoriously multi-facetted conversational domain. Not only are there many different kinds of questions from a syntactic point of view, but we are faced with the additional problem that most types of questions can serve a wide variety of different functions, some of which even oppose each other (cf. Coates 1986: 105f., 152, Cameron et al. 1989, Tsui 1992, Coates 1996: 176ff.). Obvious functions are asking for information, making a request for action, an offer or an invitation, asking for confirmation, agreement or permission to do something, initiating a story, criticising people or telling them off. A thorough comparison of male and female usage of different types of questions in an 8-million word corpus would clearly make up a research project at the level of a PhD thesis.

However, like minimal responses, questions are an important indicator of a speaker's willingness to foster linguistic interaction. Being the first part of an adjacency pair, a question will almost never be a topic-closing turn in a conversation, no matter which particular function it may have. (An obvious exception is a directive like *will you shut up.)* It is precisely because of this property of questions that they are worthy of our attention here. To reduce the domain to a manageable size, I have investigated a small number of interrogative constructions which can function syntactically either as yes/no-questions or tag-questions (see Table 4). The assumption behind this move was that questions of this type clearly tend to have the effect of promoting rather than stifling a conversation.

The data collected in Table 4 are fairly clear. The list is topped by questions which are both distinctly addressee-oriented and comparatively indirect. *Why don't you* and *couldn't you* in particular can best be imagined functioning as indirect suggestions. The only question form that is overrepresented in WOMEN and not addressee-oriented is *isn't it,* and this may be attributable to its function in indirect statements and confirmation-seeking tag-questions. The two types of questions more frequently found in MEN, *could you* and *could I,* have a relatively narrow range of functions, with the former mainly being used as a little-hedged request, and the latter as a request for permission.

5. Discussion

It is always dangerous to summarize findings which have been made on an already fairly general level of abstraction. Nevertheless, a few general trends seem to be reliable enough to allow for an intermediate discussion. The data on women's words have more or less confirmed what the literature has predicted: a number of adjectives and adverbs that are felt to be typical of women's speech by native speakers of English were indeed found more often in WOMEN than in MEN. Perhaps a little less predictably, in spite of Jespersen's remarks, the two major audible markers of hesitation, *er* and *e r m ,* occur much more frequently in MEN than in WOMEN. The classic examples of hedges, on the other hand, were indeed used more often by the women in the BNC than by the men. This is particularly true of the addressee-oriented ones, *you know* and *you see,* and of *well,* many of whose uses also tend to be motivated by interpersonal considerations. Similarly, a preponderance of minimal responses and certain interrogative clause fragments could be found in the scores for women in the spoken part of the BNC.

Taken together, these findings provide converging evidence for the claim that women tend to behave more cooperatively in conversation than men in the

sense that they show more interest in the other discourse participant(s), in their topics and their contributions, and that they invest more effort in keeping the other speakers involved. The data also indicate that women have a stronger tendency than men to hedge utterances and use indirect interrogative patterns. These linguistic gestures have traditionally been interpreted as signs of uncertainty and tentativeness. It remains open to question whether this interpretation is correct, or Coates (1996: 156ff.) is right in claiming that hedges are also a sign of cooperation and considerateness because they leave room for disagreement.

6. Data on semantic fields

6.1 Domains with expected female preponderance

6.1.1 Clothing

The first semantic domain for which an overrepresentation in WOMEN was expected on the basis of remarks in the literature (e.g. Jespersen 1922: 248f.) and everyday stereotypes is the domain *clothing*. The scores for the terms investigated are clearly in line with these expectations (see Table 5). It is interesting that even words for men's clothes (e.g. *shirt*), are more often found in WOMEN than in MEN. Words for women's garments *(tights, bra)* are hardly ever used by the men in the corpus.

Table 5: Clothing

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
tights	1.42	16.28	-0.84	a
bra	0.61	6.14	-0.82	a
coat	26.64	105.67	-0.60	a
socks	10.98	36.86	-0.54	a
blouse	1.22	3.99	-0.53	a
skirt	7.32	23.34	-0.52	a
sweater	1.02	3.07	-0.50	a
jacket	14.84	41.16	-0.47	a
clothes	37.21	92.15	-0.42	a
shoe =NN1	9.96	21.50	-0.37	a
trousers	23.79	51.30	-0.37	a
shirt	15.86	33.17	-0.35	a
shoes	39.85	76.18	-0.31	a
hat	24.81	47.00	-0.31	a
jeans	11.59	16.59	-0.18	

Table 6: Basic colours

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
orange=AJ0	4.68	12.59	-0.46	a
pink	25.01	58.06	-0.40	a
grey	18.91	42.39	-0.38	a
brown	30.91	61.13	-0.33	a
white	124.44	203.65	-0.24	a
purple	9.76	15.36	-0.22	
black	137.25	209.49	-0.21	a
green	105.12	146.21	-0.16	a
red	130.13	164.95	-0.12	a
blue	98.21	113.35	-0.07	
yellow	54.09	61.13	-0.06	

6.1.2 Basic colours

Ever since Lakoff's research (1975: 8ff.), there has been a common assumption in linguistics that women have a wider vocabulary in the domain *colour* than men and, further, know and use a far larger number of rare or even exotic terms for colours than men. Unfortunately, this claim cannot be tested in the spoken part of the BNC because the frequencies of words like *mauve, aquamarine* or *magenta* are too low to be reliable. What is possible, however, is to compare the frequencies of the eleven basic colour terms. The result is that all of them occur more frequently in WOMEN than in MEN (see Table 6), not all of them with a significant difference, however.

It can be noted in passing that those colour terms that are known to occur fairly late in the evolution of languages (see Berlin and Kay 1969), viz. *orange, pink, grey, brown* and *purple,* are found at or towards the top of the list, which means that the difference between WOMEN and MEN is particularly large here. On average, these terms are also rarer than the more 'basic' basic colour terms.

6.1.3 Home

The list of terms related to the domain *home* is headed by three words denoting rooms, *sitting room, dining room* and the much more frequent *kitchen* (Table 7). These are followed by words for pieces of furniture. It should be noted that the word *home* is ambiguous, having several meanings in different word classes. Consequently, the scores for this lexeme should not be overestimated.

Table 7: Home

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
sitting room	2.44	13.82	-0.70	a
dining room	5.90	22.73	-0.59	a
kitchen	38.02	135.77	-0.56	a
curtain(s)	15.05	49.45	-0.53	a
cupboard	19.52	54.37	-0.47	a
bed	115.49	297.03	-0.44	a
carpet	22.98	50.68	-0.38	a
door	198.86	360.00	-0.29	a
home	390.60	665.64	-0.26	a
garden	73.00	120.10	-0.24	a
(tele)phone	223.26	344.03	-0.21	a
house	389.58	572.26	-0.19	a
chair=NN1	72.59	96.14	-0.14	a
book(s)	297.07	359.08	-0.09	a
table	152.70	164.64	-0.04	

Table 8: Food and drink

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
chocolate	24.20	78.94	-0.53	a
crisps	8.54	27.54	-0.53	a
dinner	61.61	169.00	-0.47	a
biscuit	10.78	29.18	-0.46	a
coffee	60.19	149.90	-0.43	a
tea	120.58	280.45	-0.40	a
cheese	30.09	68.81	-0.39	a
lunch	50.22	108.74	-0.37	a
beans	11.79	24.27	-0.35	a
eggs	27.45	54.06	-0.33	a
milk	49.21	94.61	-0.32	a
steak	6.91	13.21	-0.31	a
butter	20.54	36.86	-0.28	a
toast	17.69	31.33	-0.28	a
breakfast	30.91	54.06	-0.27	a
lager	5.08	8.60	-0.26	
bread	47.58	79.86	-0.25	a
wine	32.13	51.60	-0.23	a
whisky	15.66	21.81	-0.16	
food	110.00	147.13	-0.14	a
beer	22.57	22.42	0.00	
pizza	17.49	15.97	0.05	
pint	20.94	18.12	0.07	

6.1.4 Food and drink

Only three terms in the list of words from the domain food and drink are balanced *(beer)* or used more frequently in MEN *(pizza* and *pint)* (Table 8). All other words, even *lager, wine* and *whisky,* are more often found in WOMEN than

in MEN, although it must be said that *lager* is so rare that the difference is not statistically significant.

6.1.5 Body and health

In this list (Table 9), especially the domain of words for body parts deserves much more detailed scrutiny, since lexemes like *leg, finger, eye* and *hand* have a large number of metaphorical and/or metonymic senses. For a proper assessment of possible differences in the uses of these words in WOMEN and MEN, it will therefore be necessary to resort to concordances and carry out sense differentiations. As it stands, the domain is skewed towards WOMEN, with the words from the domain health exhibiting a quite unequivocal preponderance in WOMEN.

Table 9: Body and health

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
breast	4.47	15.97	-0.56	a
hair	55.71	195.67	-0.56	a
headache	4.47	14.13	-0.52	a
legs	32.94	79.86	-0.42	a
sore throat	2.03	4.91	-0.41	
doctor	71.17	139.76	-0.33	a
sick	48.39	90.31	-0.30	a
ill	30.30	55.90	-0.30	a
leg	40.06	65.12	-0.24	a
eyes	49.21	79.56	-0.24	a
fingers	30.91	44.85	-0.18	b
finger=NN1	26.03	29.49	-0.06	
eye=NN1	53.27	58.67	-0.05	
body	101.00	103.52	-0.01	
hands	96.99	98.29	-0.01	
hand=NN1	231.39	214.40	0.04	

Table 10: Personal reference

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
she	2,266.94	7,842.65	-0.55	a
girl	91,91	274,92	-0.50	a
8 female names	261.48	590.07	-0.39	a
boy	127.49	232.22	-0.29	a
woman	108.17	180.92	-0.25	a
he	6,437.68	9,820.51	-0.21	a
I	26,298.91	36,701.21	-0.17	a
8 male names	866.40	1,067.41	-0.10	a
you	25,124.76	30,555.67	-0.10	a
women	173.65	198.74	-0.07	
they	9,134.26	10,563.21	-0.07	a
person	236.27	229.46	0.01	
man	425.57	403.01	0.03	
we	11,549.23	9,032.93	0.12	a
men	232.00	179.39	0.13	a
people	2,116.07	1,600.05	0.14	a
persons	13.62	5.53	0.42	a
the people	122.41	42.08	0.49	a

6.1.6 Personal reference and personal relationships

The common stereotype that women tend to talk more about people than men is also borne out by the corpus data. The possibilities for referring to people that were investigated are proper names, personal pronouns and general nouns (see Table 10), as well as lexemes denoting kinship and other personal relations (Table 11). The proper names queried were the eight most frequent female and male first names in the corpus, *Jane, Ann, Mary, Jean, Margaret, Sarah, Sue* and *Charlotte,* and *John, David, Paul, Michael, Peter, Richard, Chris* and *Dave.*

In Table 10, there is a clear overrepresentation in WOMEN. It should be noted that all expressions with a skew towards MEN (except *we)* are either masculine, general and/or impersonal and detached in nature. On the whole, the men in the corpus thus exhibit a rather impersonal way of referring to persons. It would be a matter of closer scrutiny of concordances to decide whether *we* is also more often used in MEN than in WOMEN in generic reference (Quirk et al. 1985: 353f.) comparable to *the people, they* and *one.*

The list of words denoting personal relationships (Table 11) is also clearly dominated by WOMEN. While the two exceptions *wife* and *my wife* are hardly in need of special explanations, it should be added that the word *parents* occurs strikingly frequently in spoken conversations of a fairly formal or institutional type, often with no determiner (as in *parents have to be asked as well,* text FYB, Methodist Church meeting). Interestingly, the men in the corpus used the word *son* twice as often as the word *daughter.* The women used *daughter* more frequently than the men used *son,* and *son* just a little less frequently than the men.

Table 11: Personal relationships

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
my husband	6.10	38.09	-0.72	a
baby	39.04	183.07	-0.65	a
boyfriend	5.29	16.89	-0.52	a
girlfriend	4.47	12.29	-0.47	a
husband	46.56	118.57	-0.44	a
sister	39.45	100.75	-0.44	a
mother	126.47	291.50	-0.39	a
daughter	37.01	79.18	-0.36	a
dad	206.58	410.07	-0.33	a
mum	341.19	575.33	-0.26	a
kids	96.38	157.58	-0.24	a
friend	68.32	111.50	-0.24	a
father	113.87	160.96	-0.17	a
brother	65.47	90.61	-0.16	a
family	138.88	181.84	-0.13	a
children	315.16	402.08	-0.12	a
son	74.01	72.18	0.01	
wife	116.51	87.24	0.14	a
parents	105.94	79.25	0.14	a
my wife	28.47	4.30	0.74	a

Table 12: Temporal deixis

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
yesterday	136.44	257.10	-0.31	a
tomorrow	199.87	353.55	-0.28	a
last week	73.00	127.78	-0.27	a
tonight	151.28	251.26	-0.25	a
this morning	153.72	229.15	-0.20	a
today	469.90	577.79	-0.10	a
next week	86.42	98.60	-0.07	
last year	127.90	129.32	-0.01	
next year	66.08	52.53	0.11	

6.1.7 Temporal deixis

A final domain with female preponderance, which is perhaps altogether not so expectable, is that of *temporal deictic expressions.* This domain is added here because it ties in quite nicely with other observations on women's and men's concerns and interests that will be discussed in Section 7 below. All expressions listed - except *next week* and the more 'distant' *last year* and *next year* - are found significantly more often in WOMEN than in MEN (Table 12).

6.2 Domains with expected male preponderance

6.2.1 Swearwords

The classic examples of typical 'men's words' are swearwords or expletives. Jespersen, for example, states that

> there can be no doubt that women exercise a great and universal influence on linguistic development through their instinctive shrinking from coarse and gross expressions and their preference for refined and (in certain spheres) veiled and indirect expressions. [...] Among the things women object to in language must be specially mentioned anything that smacks of swearing. (Jespersen 1922: 246)

But as the data show, we are in for a surprise in this domain (Table 13). Only the very strong four-letter words are indeed found more frequently in MEN than in WOMEN. Beginning with *damn* and moving upwards in the wordlist, the tide turns towards female preponderance with quite astonishing scores indeed especially for *bloody hell* and *bloody*. Even more surprisingly, when we look closer at the age pattern of the usage of *bloody*, we find that by far the highest relative frequency is found with WOMEN in the 45-to-59 age bracket (1095 occurrences per million words). While it must be said that this finding is to some extent influenced by a small number of texts with outrageously high frequencies (e.g. KB1, KB7, KBE, KCN), the frequency in this age band would still be high even if these texts were neglected. More in line with intuition, the peak of the usage of *fucking* is found with the MEN in the 14-to-25 age bracket (2670 occurrences per million words).

Table 13: Swearwords

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
gosh=ITJ	15.45	41.16	-0.45	a
bloody	272.46	527.10	-0.32	a
bloody hell	36.80	69.11	-0.31	a
shit	56.73	78.02	-0.16	a
damn	35.58	36.55	-0.01	
fuck	68.52	32.56	0.36	a
fucking	283.44	98.60	0.48	a

Table 14: Car and traffic

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
bus	74.42	126.55	-0.26	a
train	65.68	90.00	-0.16	a
car	360.10	482.26	-0.15	a
bike	34.57	40.85	-0.08	
underground	6.91	7.06	-0.01	
motorway	23.38	20.89	0.06	
4 car brands	28.06	21.81	0.13	
tyres	12.40	9.22	0.15	
crane	12.40	5.22	0.41	b
traffic	156.97	63.58	0.42	a
windscreen	3.05	1.23	0.43	a
miles per hour	7.73	1.54	0.67	a

6.2.2 Car and traffic

This domain is not as clearly skewed towards MEN as one might have believed (Table 14). In fact, the more general words for means of transport, *bus, train, car* and *bike* are more often found in WOMEN than in MEN (*bike* not significantly more often). When we turn to more specific lexemes, however, we see the MEN gaining in weight. It must be noted that the frequencies of the four car brands *(BMW, Ford, Rover* and *Vauxhaul)* and of *tyres, crane, windscreen* and *miles per hour* are fairly low, which results in lacking significance in some cases. The only relatively frequent word in the bottom half of the table is *traffic* with a distinct skewage towards MEN.

6.2.3 Work

The list for the domain *work* is short and fairly homogeneous because it is difficult to come up with words that can be related unambiguously to this field (Table 15). *Appointment,* for example, is clearly a term that crops up in many everyday circumstances outside the workplace, and the same is true of *holiday(s), job* and *office.* The words *file* and *colleague* are cases of distinct male preponderance.

Table 15: Work

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
appointment	32.33	41.16	-0.12	
holiday(s)	124.24	152.36	-0.10	b
boss	27.04	24.88	0.04	
job	394.26	355.09	0.05	b
office	175.88	148.67	0.08	a
meeting=NN1	112.44	81.40	0.16	a
file	71.57	19.35	0.57	a
colleague	118.34	23.34	0.67	a

Table 16: Computing

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
floppy	5.90	4.91	0.09	
desktop	4.68	3.38	0.16	
server	29.48	19.35	0.21	b
computer	105.94	64.81	0.24	a
monitor	4.47	2.15	0.35	
printer	15.86	6.45	0.42	a
Windows	17.28	0.92	0.90	a

6.2.4 Computing

The low frequencies in this list undoubtedly reflect the fact that at the end of the 80s and beginning of the 90s, when the conversations for the BNC were recorded, not nearly as many people as now had come into contact with computers (Table 16). What the list also indicates, however, is that at this stage it was predominantly men who talked about the new technology and were well-versed enough to use (at that time) novel and specific terms like *Windows.*

6.2.5 Sports

With the three notable exceptions of *tennis, soccer* and *snooker* the field of sports meets our intuitive expectations (Table 17). The concordances for *soccer* and *football* give the impression that MEN use the term *football* more frequently than WOMEN to refer to the same kind of activity. The words *ball* and *shot=NN1* are problematic because of their fairly wide range of meanings.

Table 17: Sports

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
tennis	13.62	16.59	-0.10	a
soccer	3.66	4.30	-0.08	
snooker	7.73	7.68	0.00	

Table 18: Public affairs

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
Prime Minister	22.16	20.58	0.04	
Conservatives	14.84	11.67	0.12	
the Queen	4.07	3.07	0.14	

sport(s)	62.22	47.30	0.14	b
football	87.23	59.90	0.19	a
match=NN1	45.95	26.11	0.28	a
darts	7.12	3.99	0.28	
game	149.25	82.94	0.29	a
rugby	28.67	15.67	0.29	a
cricket	16.47	8.29	0.33	a
ball	127.69	47.00	0.46	a
shot=NN1	38.84	9.83	0.60	a
referee	11.39	1.54	0.76	a

Labour	125.05	77.71	0.23	a
Tory	36.40	21.50	0.26	a
war	189.30	103.82	0.29	a
tax	135.62	74.95	0.29	a
Party	93.94	47.00	0.33	a
parliament	44.33	21.81	0.34	a
Tories	19.52	9.22	0.36	a
election	53.48	23.04	0.40	a
council	476.41	204.57	0.40	a
European	82.15	34.71	0.41	a
government	319.64	131.16	0.42	a
reform	15.86	2.46	0.73	a

6.2.6 Public affairs

Tannen relates women's interest in other people, and their inclination to gossip, to men's interest in news and sports. For her, both types of interests satisfy similar needs but bring with them different dangers:

> Men's interest in the details of politics, news and sports is parallel to women's interest in the details of personal lives. If women are afraid of being left out by not knowing what is going on with this person or that, men are afraid of being left out by not knowing what is going on in the world. And exchanging details about public news rather than private news has the advantage that it does not make men personally vulnerable. The information they are bartering has nothing to do with them. (Tannen 1990: 1 l0f.)

That the women in the corpus do speak more about people than the men was shown in Tables 10 and 11. That the men in the corpus speak more about sports can be gleaned from Table 17, and Table 18 clearly confirms that the men used words from the domain public affairs more frequently than the women in the corpus.

6.2.7 Abstract notions

The last domain on which I have collected data is that of abstract nouns. Some of these nouns are fairly rare and their scores therefore not particularly reliable. Others, however, for example *idea, problem* and *fact* belong to the most frequent nouns in English, and nouns like *quality* and *development* are not really rare either. The general picture for all these nouns is very consistent: they are used significantly more often in MEN than in WOMEN, but for the more common ones, the difference seems to be less marked.

Table 19: Abstract notions

Word	Freq. MEN	Freq. WOMEN	Coeff.	Signif.
idea	280.19	209.18	0.15	a
difference	156.97	111.20	0.17	a
option	29.89	18.12	0.25	b
problem	422.32	223.93	0.31	a
fact=NN1	568.52	285.36	0.33	a
compensation	16.06	7.99	0.34	b
focus	19.11	8.91	0.36	a
consideration	38.43	15.36	0.43	a
proportion	28.06	11.06	0.43	a
alternative	34.16	13.21	0.44	a
reflection	9.15	3.38	0.46	b
quality	123.63	39.62	0.51	a
tendency	14.44	4.61	0.52	a
development	167.34	53.14	0.52	a
contrast=NN1	9.15	2.46	0.58	a
probability	9.96	2.15	0.64	a
assumption	12.61	2.46	0.67	a
correlation	13.01	2.15	0.72	a
democracy	28.47	4.30	0.74	a
ratio	14.23	1.84	0.77	a

7. General discussion

It is fairly obvious that virtually every single one of these tables cries out for more detailed research. In most cases, one feels that it would be necessary to include more words, preferably on the basis of some objective criterion, and to differentiate multiple meanings and/or functions. And it would be illuminating to take other social parameters like social class and education into consideration as well. A11 this, however, is not possible here for reasons of space. Research is under way with the aim of delving deeper into some of these areas.

On the whole, the data represented in Tables 5 to 19 have shown that even perfectly innocuous-looking words are not used with the same frequency by the women and men recorded in the BNC. Not all of the differences are statistically significant; some of them are not because the overall frequency of the words in the two subcorpora is too low. This indicates that even larger collections of spoken language must be gathered to get a better picture of gender-differences in the usage of words. What would also be desirable is an even more extensive coverage and mark-up of other demographic factors and of information on speech situations, topics and relations between speakers, even though it must be said that the BNC constitutes a major step forward in this field.

In most domains, the frequency scores that were found were in line with widespread stereotypes about favourite female and male topics. An overrepresentation in WOMEN was confirmed for the domains *clothing, basic colours, home, food and drink, body and health* as well as *people*. Words and expressions from the domains *work, computing, sports* and *public affairs* tended to be found more often in MEN than in WOMEN. In the domains of *sports* and *public affairs,* the data suggest that male preponderance tends to increase together with the specificity of the items investigated. This is also in line with the findings from the domain *car and traffic,* where the more general terms *bus, train* and *car* were found more often in WOMEN. The data on swearwords are somewhat astonishing, since four of the items investigated *(gosh, bloody, bloody hell* and *shit)* occurred more often in WOMEN than in MEN.

While it is evident, at least in hindsight, that most of the findings were expected and predictable, I do not think they are trivial. To begin with, one must not forget that linguistic data that are based on more than 8 million words of authentic conversation had previously not been available, and that the scores

can therefore be seen as strong and comparatively objective confirmations of long-standing intuitions and gut feelings concerning typical female and male topics.

Furthermore, given the distinct differences in many domains it is only natural to assume that the scores represent more than just differences in the use of language. If a person talks more about, say, food than another person, one will conclude that the former person is also more concerned with food, perhaps even more interested in food, than the latter. Similarly, if one group of people talk more about football than another, we assume the first are more concerned with it than the second. With groups as large and heterogeneous as women and men, one tends to find sweeping statements of this type a little irritating, and I think this is justified, because many other factors besides the speakers' gender have an influence on their choice of words and topics, most notably the classic demographic factors education, age and social class. As a consequence, we are never at a loss for good counterexamples, for example women who never discuss clothes, or men who are not interested in sports. As the data show, however, such people are not really counterexamples at all, because no word was found which was restricted exclusively to female or male usage; there was not a single word with a coefficient of 1, -1, or anything close to them. The closest we came was 0.90 for *Windows,* a score that has certainly changed in the meantime, -0.84 for *tights,* and -0.82 for *bra.* Thus, what the scores indicate are only statistically significant tendencies concerning the linguistic behaviour of these heterogeneous sections of society. Arguably, however, they reflect more than that: they reflect trends about women's and men's concerns, to use a very neutral term, for certain domains. Going one step further, one can argue that the differences in frequency scores actually reflect women's and men's interests, hobbies, worries and problems. This could clearly mean that the corpus data demonstrate some sort of cultural difference between women and men, in the same way as Leech and Fallon's (1992) data reflected differences between American and British cultures.

An obvious objection to this claim would be that the differences in word frequencies are not caused by different concerns and interests, but by the social roles of the women and men who were recorded for the BNC: their jobs, their daily routines, obligations and activities. After all, it is a sociological fact that more women stay at home to take care of children or other relatives and more men go to work. This will also be reflected in the composition of the corpus and can explain the score differences, for example those from the domains *clothing, home, personal relationships* and *personal reference, work, computing* and even *abstract nouns.* But this is in actual fact not really an objection to the claim that the corpus can tell us something about male and female culture. Instead it shows that Chomsky was right after all when he argued that corpora mirror extra-linguistic facts (cf. Kennedy 1998: 23; needless to add that he was wrong in claiming that corpora have no relevance for linguistic analysis and description). What the BNC mirrors is the state of British society at the beginning of the 1990s. So Lakoff was right, too, when she wrote that "the speaker of English who has not been raised in a vacuum *knows* that all of these disparities exist in English for the same reason: *each reflects in its pattern of usage the difference between the role of women in our society and that of men."* (1975: 49; original emphasis). This study has shown that these patterns of usage can be observed in a corpus. It has thus provided evidence that there is not just a link between corpora and the linguistic system of the language collected (as Halliday 1993: 3ff. has argued), and a link between corpora and cognition (as I have argued elsewhere, cf. Schmid 2000: 38ff.), but also a link from corpora to culture.

From the gender-cultural corpus-linguistic perspective that I have been taking here, it would be a particularly exciting prospect to create a corpus as closely parallel to the composition of the spoken part of the BNC in 2020 or so and compare data from this corpus to find out about changes in the place of women and men in British society. At the moment, the question whether gender differences in linguistic usage are ultimately caused by the speakers' gender or by their place in society could only be settled with several large parallel corpora of sociologically comparable women and men, but corpora of this type are not yet available at present.

Yet another objection arises from the composition of the corpus itself. It is clear that the findings collected here can only mirror society insofar as the corpus itself mirrors society in its composition. This is probably a much more serious objection. For one thing, we have seen in Section 2 above that the demographically sampled part of the corpus used by Rayson, Leech and Hodges (1997) contains a larger proportion of data spoken by women, while the corpus used here includes more words originally produced by men. The demographically sampled corpus consists mainly of spontaneous casual everyday conversations, while the context-governed part adds to this samples of discourse of a more official and formal, and less interactive and involved nature.

Given this difference, it is interesting to compare the data collected here - which derive from what could be called the "Spoken Corpus" - to those presented in Rayson, Leech and Hodges (1997) based on the demographically-sampled Conversational Corpus. This comparison is possible for 18 words which were investigated in both studies. If the gender-differences in vocabulary frequency observed in this study were exclusively determined by the parameter gender - a very unlikely hypothesis, indeed - then they should stay the same, even if only one part of the corpus used here is investigated. The comparison is summarized numerically in Table 20, where columns A/B and C/D give the relative scores per million words in this study and Rayson, Leech and Hodges (1997)[3] for men and women, respectively. Columns E and F give the coefficient for the scores found here and for the scores given in Rayson, Leech and Hodges, while column G gives the difference between the coefficient scores. The table is sorted according to column G.

[3] The relative scores for Rayson, Leech and Hodges' data have been calculated using on the absolute scores given in their tables on pages 136 to 139 and the overall frequencies given on page 136.

Table 20: A comparison to some data from Rayson, Leech and Hodges (1997)

	A	B	C	D	E	F	G
	Relative scores per million words				Coefficient		
	Men		Women				
	Spoken Corpus	Convers. Corpus	Spoken Corpus	Convers. Corpus	Spoken Corpus	Convers. Corpus	Differ-ence
okay	1,414.58	765.85	696.35	500.49	0.34	0.21	-0.13
er	9,970.16	5,593.07	5,036.96	3,588.65	0.33	0.22	-0.11
father	113.87	67.08	160.96	118.38	-0.17	-0.28	-0.11
home	390.60	428.13	665.64	640.84	-0.26	-0.20	0.06
brother	65.47	79.33	90.61	88.30	-0.16	-0.05	0.11
I	26,298.91	32,381.36	36,701.21	35,838.33	-0.17	-0.05	0.11
daughter	37.01	40.83	79.18	65.94	-0.36	-0.24	0.13
mother	126.47	158.65	291.50	241.76	-0.39	-0.21	0.19
she	2,266.94	4,161.12	7,842.65	8,723.12	-0.55	-0.35	0.20
sister	39.45	61.24	100.75	99.10	-0.44	-0.24	0.20
lovely	131.35	241.48	421.13	468.10	-0.52	-0.32	0.21
yeah	6,712.99	12,861.32	9,309.07	10,983.43	-0.16	0.08	0.24
fucking	283.44	817.18	98.60	125.32	0.48	0.73	0.25
son	74.01	99.74	72.18	57.45	0.01	0.27	0.26
mm	2,101.23	4,193.20	4,500.65	4,970.60	-0.36	-0.08	0.28
fuck	68.52	195.40	32.56	41.26	0.36	0.65	0.30
dad	206.58	548.87	410.07	491.62	-0.33	0.06	0.39
mum	341.19	960.66	575.33	715.65	-0.26	0.15	0.40

In view of the variegated nature of this set of words, it is no surprise that there are few general tendencies to be observed here. The following remarks can be made concerning the various perspectives that this table opens up:

- Except for three words, all words listed keep the same sign (plus or minus) in both corpora. This is a reassuring indication that the differences between men and women tend to be of the same kind in the Spoken Corpus and the Conversational Corpus. Only *yeah, dad* and *mum* change from minus to plus when only casual conversation is taken into consideration. This is due to the fact that they are relatively more often used by men than by women in casual speech as opposed to casual plus formal speech. Therefore, these three items can be considered markers of men's private speech.
- The differences between the Spoken Corpus and the Conversational Corpus tend to be more pronounced for the men's utterances than for the women's; the scores in columns C and D tend to be closer to each other than those in columns A and B. However, I do not think that it would be right to trace this back to the claim that men show a greater situation-dependent speech adaptation than women. A more likely reason for this finding lies in the different proportions of the two corpora (cf. the numbers given in Section 2 above): the Spoken Corpus includes 3.2 million more words spoken by men than the Conversational Corpus, but only 660,000 more words spoken by women. Given that much more new material is added on the men's side, it is only natural that there are more pronounced differences in the men's than in the women's parts of the two corpora.
- For most words, the difference between the frequencies of women and men, as indicated by the coefficient, is smaller in the Conversational Corpus than in the Spoken Corpus; most scores in column F are closer to zero than those in column E. Arguably, this reflects the fact that the demographically sampled corpus is indeed more homogeneous in its composition than the whole spoken subsection. The tendency is counterevidence to the hypothesis mentioned above, since it indicates that factors other than gender must play a role. Notable exceptions to this tendency are the words *father, fucking, son* and *fuck*. For *father*, there is an even stronger skewage towards WOMEN in the Conversational Corpus, presumably because men tend to use *d a d* rather than *father* in the private domain, something they do much less often in public speaking (cf. the scores for *dad*). *Fucking, son* and *fuck*, on the other hand, exhibit a more pronounced skewage towards MEN in the Conversational Corpus. Not surprisingly (at least for the four-letter words), these three items are apparently relatively more often used by men in private or casual conversations than in public speech. They can thus be included in the set of markers of men's private speech, which, then, consists of the motley, but in a way not so surprising, collection *yeah, dad, mum, son, fucking* and *fuck*.
- Focusing on the data for the men, it is interesting to note that only *er, okay* and *father* have a higher relative frequency in the Spoken Corpus than in the Conversational Corpus. All other words occur relatively more frequently in the men's section of the Conversational Corpus. So *er, okay* and *father* can be seen as markers of men's public speech.
- For the women, the differences are more balanced: *yeah, she, mm, mum, dad, lovely, fucking* and *fuck* - presumably all markers of interactive and involved style - occur relatively more frequently in the more casual Conversational Corpus. *Sister, brother, daughter, son, home, father, mother, okay, I* and *er* are relatively more frequent in the Spoken Corpus, which includes public and more formal speech genres. The group of terms from the field of family relations suggests that women might actually talk just as much about people outside their immediate private domain as within it. This confirms Tannen's claims (1990: 91) that women show a greater tendency than men to approach situations in the public domain as an extension of the private domain. If nothing else, these findings are an indication that there are topic preferences that are indeed mainly determined by the speaker's gender and not so much by the situation or other social parameters.

Many other interesting observations could be added to this but this comparison is not the main purpose of this paper. What should be emphasized, however, is the tendency that words belonging to a colloquial register are relatively more frequent in the Conversational Corpus, and this finding in turn supports the hypothesized distinction between the two samples.

Two further remarks concerning the findings of this paper can be ventured with all signs of caution that should accompany such wild generalizations as I am going to offer now. The first brings the data on *temporal deictics* and *abstract nouns* back into consideration, about which I have not yet said much. Taking into account also the findings on *clothes, colours, home, food and drink,* and *people,* one the one hand, and *sports* and *public affairs* on the other, one can claim that women are indeed more concerned with concrete things in their immediate environment than men, while men are more concerned with remote events and abstract ideas. As before, it is more than likely that this difference is ultimately caused by the traditional roles of women and men in British society, but again as before, this does not cast doubt on the relevance of the corpus evidence. In his dubious chapter on women's language, on which I have already drawn, Jespersen quotes a passage from a scholar called Havelock, who in turn reports on a study on male and female vocabulary carried out by an American professor named Jastrow. In this study, university students had been asked to write down as rapidly as they could one hundred words. From the lists obtained and their frequency analysis, Jastrow, and after him Havelock concluded the following:

> In general the feminine traits revealed by this study are an attention to the immediate surroundings, to the finished product, to the ornamental, the individual, and the concrete; while the masculine preference is for the more remote, the constructive, the useful, the general and the abstract. (Havelock 1904: 189, quoted after Jespersen 1922: 249)

While not all of these traits receive confirmation in the present study, some clearly do.

The second daring generalization is related to the first, since it also has to do with proximity and distance. Much more than men, women seem to be engaged - presumably again because of their social roles - in what is usually regarded as prototypical spontaneous speech. According to specialists in the field (see e.g. Koch and Oesterreicher 1985, Biber 1986, Chafe and Danielewicz 1987), this genre is marked among other things by high involvement in the interaction and little spatial, temporal and emotional distance between the speech participants. In the present study, these characteristics show up in the data on minimal reponses, supportive discourse markers and questions, but also in the overrepresentation in WOMEN of words that are either clearly or possibly related to the immediate speech situation (e.g. *home, people, temporal deictics).* Further corroborative evidence for this claim, on which I have not reported here (but see Schmid in preparation), is that in WOMEN we find a smaller number of postmodified noun phrases than in MEN, a smaller number of prepositions, especially of the 'grammatical' prepositions *of* and *in,* fewer tokens of the most frequent types of nouns but more of the most frequent types of verbs, and more occurrences of personal and demonstrative determiners. In MEN, we find more markers of written, detached and 'distant' language like larger numbers of nouns and noun-postmodifiers, which combine to create a much more condensed and compact style. What this comes down to ultimately is that the more intense involvement of women and the higher degree of detachment of men is not only reflected in their discourse behaviour, but also in the frequencies with which they use certain words and words of certain word classes. Perhaps even more than the lists of differences in the usage of single words, this suggests that women and men actually live in different cultures. It is patently obvious, however, in the data presented here that to a very large extent these two cultures overlap. Ironically (and iconically), probably the best pictorial representation of this kind of overlap is the well-known image of two intersecting wedding rings.

References

Aston, G, Burnard L 1998 *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Berlin B, Kay P 1969 *Basic color terms. Their universality and evolution*. Berkeley – Los Angeles: University of California Press.

Biber D 1986 Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* 62: 384–414.

Cameron D, McAlinden F, O'Leary K 1989 Lakoff in context: the social and linguistic functions of tag-questions. In Coates J, Cameron D (eds), *Women in their speech communities*. London – New York: Longman, pp 74–93.

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

Coates J 1986 *Women, men and language*. London – New York: Longman.

Coates J 1989 Gossip revisited: language in all-female groups. In Coates J, Cameron D (eds), *Women in their speech communities*. London – New York: Longman, pp 94–122.

Coates J 1996 *Women talk. Conversation between women friends*. Oxford – Cambridge/Mass.: Blackwell.

Halliday M A K 1993 Quantitative studies and probabilities in grammar. In Hoey M (ed), *Data, description, discourse. Papers on the English language in honour of John McH. Sinclair*. London: HarperCollins, pp 1-25.

Hartung J 1999 *Statistik. Lehr- und Handbuch der angewandten Statistik*. 12th ed., München – Wien: Oldenbourg.

Havelock E 1904 *Man and woman*, 4th ed., London.

Hofland K Johansson S 1982 *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Jespersen O 1922 *Language. Its nature and development*. London etc.: George Allen & Unwin.

Kennedy G 1998 *An introduction to corpus linguistics*. London – New York: Longman.

Koch P, Oesterreicher W 1985 Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36: 15-43.

Lakoff R 1975 Language and women's place. New York etc.: Harper Colophon Books.

Leech G, Fallon R 1992 Computer corpora – What do they tell us about culture? *ICAME Journal* 16: 29–50.

Lehmann H-M, Hoffmann S, Schneider P 1996–1998 *The Zurich BNC*web *Query System*. (http://escorp.unizh.ch)

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

Rayson P, Leech G, Hodges M 1997 Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics* 2(1), 133-152.

Schmid H-J 2000 *English abstract nouns as conceptual shells. From corpus to cognition*. Berlin – New York: Mouton de Gruyter.

Schmid H-J in prep Ten reasons why men use the definite article more frequently than women.

Schiffrin D 1987 *Discourse markers*, Cambridge: Cambridge University Press.

Tannen D 1986 *That's not what I meant*. London – Melbourne: Dent.

Tannen D 1990 *You just don't understand*. New York: William Morrow and Company.

Tsui A 1992 A functional description of questions. In Coulthard M (ed), *Advances in spoken discourse analysis*, London – New York: Routledge, pp. 89-110.

Zimmermann D, West C 1975 Sex roles, interruptions and silences in conversations. In Thorne B, Henley N (eds), *Language and Sex*, Rowley/Mass: Newbury House, pp 105–129.