
CZECH NATIONAL
CORPUS

Introduction to Text Corpora and Their Applications

Corpora in sociolinguistics linguistics

Lucie Chlumská, Ph.D.

lucie.chlumska@korpus.cz

LECTURE

Sociolinguistics

- traditionally based on empirical data; however, the use of standard corpora in this field has been rather limited

- three issues:
  1. the operationalization of sociolinguistic theory into measurable categories suitable for corpus research
  2. the lack of sociolinguistic metadata encoded in available corpora
  3. the lack of sociolinguistically rigorous sampling in corpus construction

Sociolinguistics: areas of interest

- sociolinguistics traditionally studies variation and change
  - sociolingistic variation:
    - *demographic variation*
      - gender, region, age, education, class...
    - *variation across registers*
      - Biber's dimensions (involved v. informational, narrative v. non-narrative...)
    - *phonetic/prosodic variation*
      - dialects (Cockney etc.)

Example: Biber's dimensions

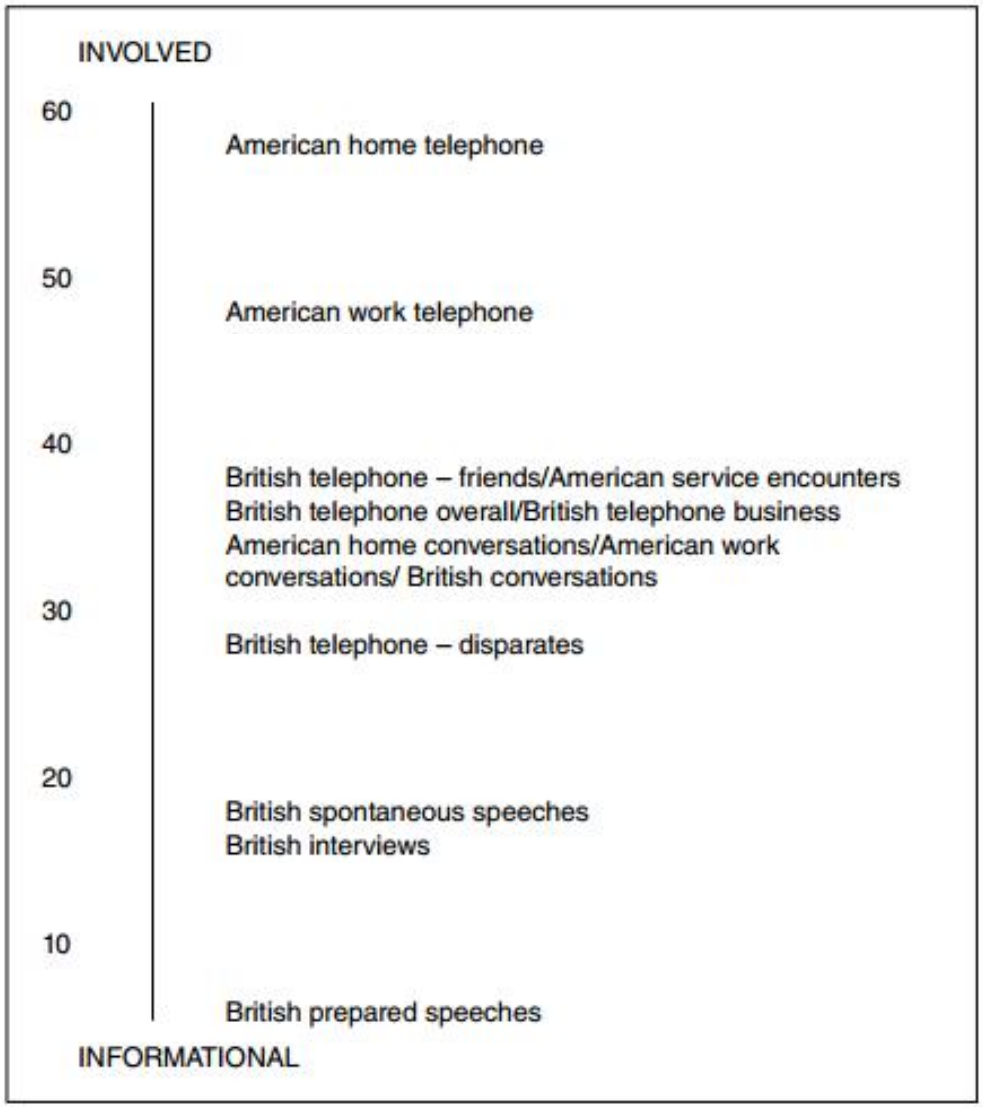


Figure 4.4 Mean scores for American and British spoken genres along Dimension 1, 'involved vs. informational production' (Helt 2001: 175)

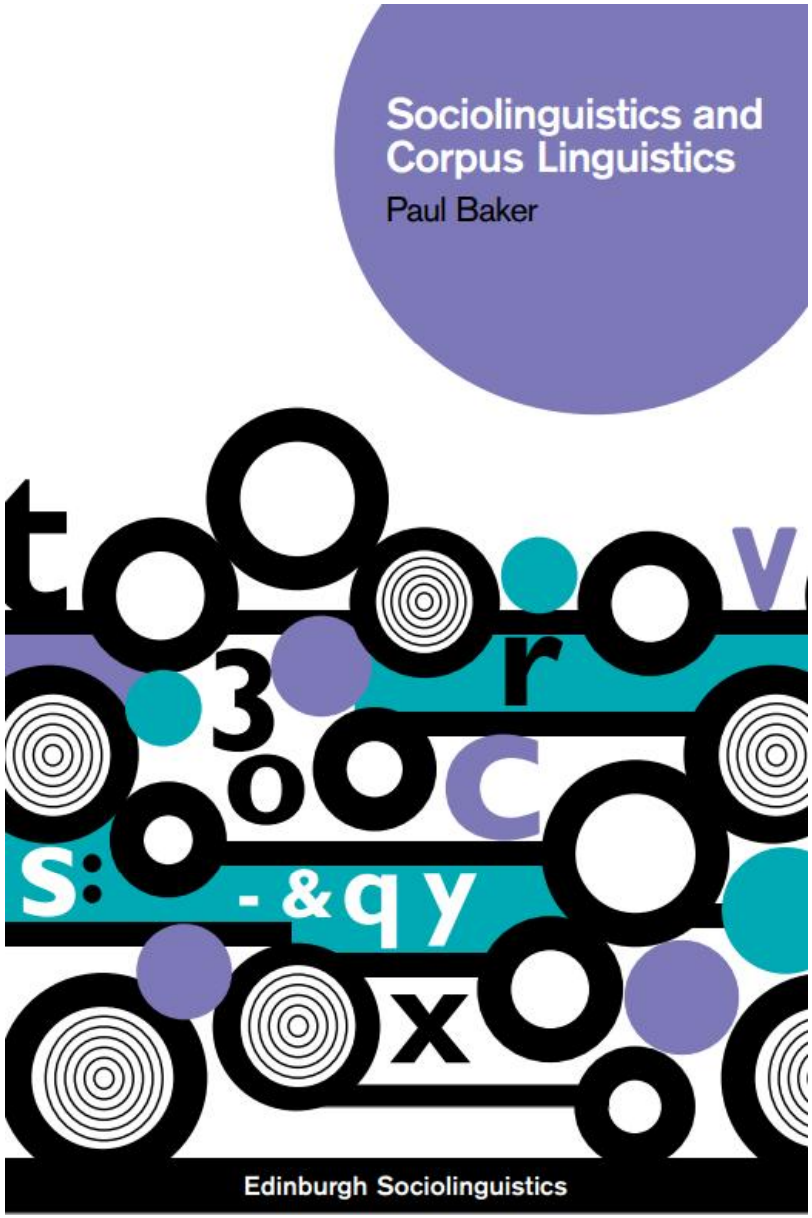
Topics and resources

- diachronic variation:
  - Brown family corpora (Brown/Frown, LOB/FLOB)
  - historical corpora (Helsinki Corpus, ARCHER)
- synchronic variation:
  - global varieties of English (GloWbE corpus)
  - spoken v. computer-mediated varieties (CORE corpus)
- corpus-based sociolinguistics: so far restricted mostly to the area of gender studies at the lexical level
  - sexism, feminism, sexual identity
- Beaugrande (1998: 131): “some interesting prospects” (ongoing collocational approximation v. traditional focus on phonetics, grammar)

Case studies in sociolinguistics

References

Paul Baker (2010):
Sociolinguistics and Corpus Linguistics



Lexical differences in speech

- young speakers: *no but, yeah but* (McEnery's research on modern speech)

Little Britain: "teenager" Vicky Pollard

https://www.youtube.com/watch?v=5pcFcnJKWlg

Lexical differences between sexes

- Rayson, P. et al. (1997): Social differentiation in the use of English vocabulary. *IJCL* 2:1, 133–152.

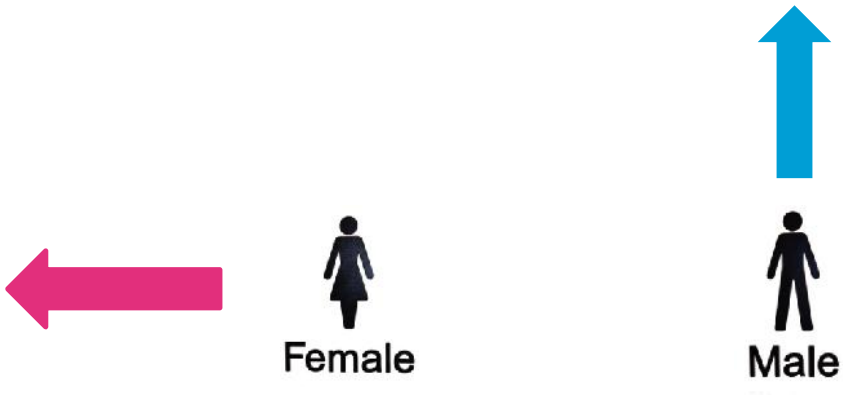
Table 2.4 Main lexical differences between sex, age and social class categories in the BNC (adapted from Rayson et al. 1997)

Sex		Age		Social class	
Male	Female	Under 35s	Over 35s	ABC1	C2DE
fucking, er, the, yeah, aye, right, hundred, fuck, is, of, two, three, a, four, ah, no, number, quid, one, mate, which okay, that, guy, da, yes	she, her, said, n't, I, and, to, cos, oh, Christmas, thought, lovely, nice, mm, had, did, going, because, him, really, school, he, think, home, me	mum, fucking, my, mummy, like, na, goes, shit, dad, daddy, me, what, fuck, wan, really, okay, cos, just, why	yes, well, mm, er, they, said, says, were, the, of, and, to, mean, he, but, perhaps, that, see, had	yes, really, okay, are, actually, just, good, you, erm, right, school, think, need, your, basically, guy, sorry, hold, difficult, wicked, rice, class	he, says, said, fucking, ain't, yeah, its, them, aye, she, bloody, pound, I, hundred, well, n't, mummy, that, they, him, were, four, bloke, five, thousand

Lexical differences between sexes

Table 2.1 Statistically significant lexical differences according to sex in the BNC spoken section (adapted from Schmid 2003)

	Category	Used more by females	Used more by males
Categories believed to be more typically used by females	Adjectives/adverbs	*handsome, lovely, sweet, horrible, dreadful, awful*	-
	Hesitators/hedges	*well, really, you see, you know, I mean*	*erm, perhaps, er, sort of, I guess, in fact*
	Minimal responses	*mm, aha, yes but, no, mhm, yeah, yes*	*okay*
	Questions	*aren't you, can you, are you, isn't it, wouldn't you*	*could I*
	Clothing	*tights, bra, coat, socks, shirt, clothes, sweater, jacket*	-
	Colours	*orange, pink ,grey, brown, white, purple, black, green*	-
	Home	*kitchen, bed, carpet, door, home, garden, phone, chair*	-
	Food and drink	*dinner, tea, lunch, eggs, wine, milk, steak, butter, toast*	-
	Body and health	*breast, hair, headache, legs, doctor, sick, ill, leg, eyes*	-
	Personal reference and relationships	*I, you, she, he, boy, girl, baby, husband, mother, friend, father, brother, sister*	*people, person, man, men, we, son, wife, parents*
	Time	*yesterday, tomorrow, tonight, today*	-
Categories believed to be more typically used by males	Swear-words	*gosh, bloody, shit, damn*	*fuck, fucking*
	Car and traffic	*bus, train, car*	*traffic, crane, windscreen, miles per hour*
	Work	*holiday*	*boss, job, office, meeting, file, colleague*
	Sport	*tennis*	*football, ball, shot, rugby, referee, darts, match, sports*
	Public affairs	-	*reform, government, council, election, Tories, tax, war, Labour*
	Abstract concepts	-	*idea, difference, option, problem, fact, focus, quality*



Lexical differences between sexes

Table 2.5 Frequencies per million words of *lovely* in the BNC tabulated according to sex, age and social class

Age	Males				Females			
	AB	C1	C2	DE	AB	C1	C2	DE
0–14	74.76	121.8	197.27	207.73	501	309.6	113.5	0
15–24	139.55	0	194.46	232.86	220.29	660.81	127.07	380.48
25–34	0	655.72	121.87	199.41	312.51	901.76	599.07	393.08
35–44	62.23	230.72	64.1	276.4	599.85	408.49	288.57	207.75
45–59	233.7	415.57	295.45	401.39	589.31	473.36	462.31	557.84
60+	347.23	465.69	270.15	309.63	1216.7	714.63	369.76	803.77

SOCIAL CLASS IN THE BNC

AB	Higher management: administrative or professional
C1	Lower management: supervisory or clerical
C2	Skilled manual
DE	Semi-skilled or unskilled

Written v. spoken variation

Table 5.4 Discourse management features in the spoken and written sections of the BNC (frequencies per million words)

Word	Spoken	Written
hence	6.15	52.55
it follows	4.42	10.07
furthermore	2.79	32.44
moreover	1.15	47.85
actually	1277.01	143.79
I mean	1956.22	43.9
you see	732.29	45.69

Gender in a diachronic perspective

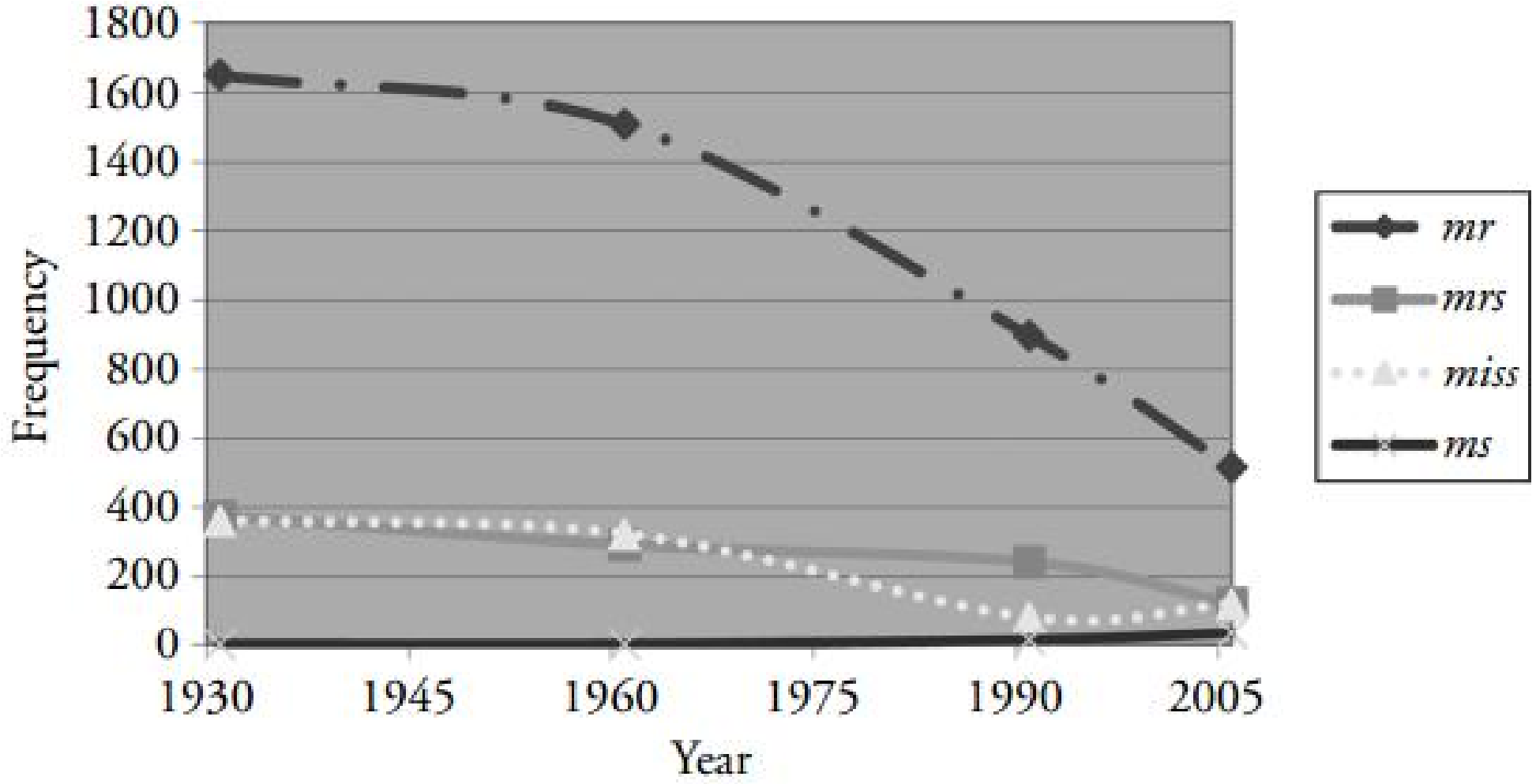


Figure 3.3 Frequencies for gendered terms of address over time

Diachronic variation

Table 3.3 Change in modal verbs in British and American English (Leech 2002)

Verb	British English		Difference (%)	Log likelihood	American English		Difference (%)	Log likelihood
	LOB 1961	FLOB 1991			Brown 1961	Frown 1992		
could	1740	1782	+2.4	2.4	776	1655	–6.8	4.1
can	1997	2041	+2.2	0.4	2193	2160	–1.5	0.2
will	2798	2723	–2.7	1.2	2702	2402	–11.1	17.3
would	3028	2694	–11.0	20.4	3053	2868	–6.1	5.6
should	1301	1147	–11.8	10.1	910	787	–13.5	8.8
might	777	660	–15.1	9.9	635	635	–4.5	0.7
may	1333	1101	–17.4	22.8	1298	878	–32.4	81.1
must	1147	814	–29.0	57.7	1018	668	–34.4	72.8
need	78	44	–43.6	9.8	40	35	–12.5	0.3
shall	355	200	–43.7	44.3	267	150	–43.8	33.1
ought	104	58	–44.2	13.4	70	49	–30.0	3.7
Total	14667	13272	–9.5	73.6	13962	12287	–12.2	68.0

Thank you for your attention!

Questions?

SEMINAR

Reading

common reading:

McEnery, T., Xiao, R. & Tono, Y. (2006): Swearing in modern British English. In *Corpus-Based Language Studies*, Routledge, pp. 264–286.

Discussion

- What are the most frequent methods to be used in corpus-based sociolinguistics?
- Can you see a difference between CADS and sociolinguistics?
- What sociolinguistic variables are usually encoded in spoken corpora?
- How would you sample a spoken corpus of your language?
- Is there a clear social class distinction in your country?

- How can swearing be analyzed from a sociolinguistic perspective?
- Is there a similarly “universal” swear word as *fuck* in your mother tongue?