Data in Discourse Analysis Conference

Linguistic Data and Social Groups

Paul Baker
p.baker@lancaster.ac.uk

What's a social group?

- Interaction: inter-dependence, structures and roles e.g. a sports team (Platow et al 2011)
- Shared identity characteristics: e.g. Scottish people (Turner 1982)
- Shared goals: the Muslim community (Riecher 1982, Anderson 1983))

- Variationist sociolinguistics
- Analysis of representations

Anderson, B. (1983) *Imagined Communities*. London: Verso.

Platow, M.J.; Grace, D.M.; Smithson, M.J. (2011). "Examining the Preconditions for Psychological Group Membership: Perceived Social Interdependence as the Outcome of Self-Categorization". *Social Psychological and Personality Science*. 3 (1): 5-13.

Reicher, S.D. (1982). The determination of collective behaviour. In H. Tajfel (ed.), *Social identity and intergroup relations*. Cambridge: Cambridge University Press, pp. 41-83.

Turner, J.C. (1982). Tajfel, H. (ed.). "Towards a cognitive redefinition of the social group". *Social identity and intergroup relations*. Cambridge: Cambridge University Press, pp. 15–40.

‘it sometimes seems to be the case that sociolinguists' borrowings from corpus linguistics are shallow and …possibly at times only name-deep)’ (Kendall 2011: 364)

‘Sociolinguistic research often focuses on non-standard varieties of language, and spoken language in particular, and the large "conventional" (i.e. standard language and often primarily written) corpora have simply not been of great use for pursuing sociolinguistic research on non-or less-standard varieties’ (Kendall 2011: 363)

‘it appears to be an impossible task to make replicable and generalizable, especially through corpus-based methods, the ethnographic and instance-specific knowledge a researcher must gain in order to understand the actual creation and negotiation of social meaning "on the ground.’ (Kendall 2011: 370)

Kendall, T. (2011) Corpora from a sociolinguistic perspective. *Revista Brasileira de Linguística Aplicada*, *11*(2), 361-389.

Why analyse the language of social groups?

- Discovery-based research
- To raise the profile of a group, start a conversation or consider power (Baxter 2009)
- Identify linguistic norms of a group, e.g. for teaching materials, to tailor messages for different types of speaker or inform forensic research

---

- Baxter, J. (2009) *The Language of Female Leadership*. Palgrave.

Finding data

1. Build your own corpus, small, specific, lends towards qualitative analyses (Murphy 2010)

Murphy, B. (2010) *Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk*. Amsterdam: John Benjamins.

Finding data

1. Build your own corpus, small, specific, lends towards qualitative analyses (Murphy 2010)

2. Use a publicly available large corpus (e.g. BNC Spoken 1994 and 2014 versions) (Brezina et al 2018)

Brezina, V., Love, R. and Aijmer, K. (eds) (2018) *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*. London: Routledge.

Murphy, B. (2010) *Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk*. Amsterdam: John Benjamins.

Age, Cohort and Period related effects

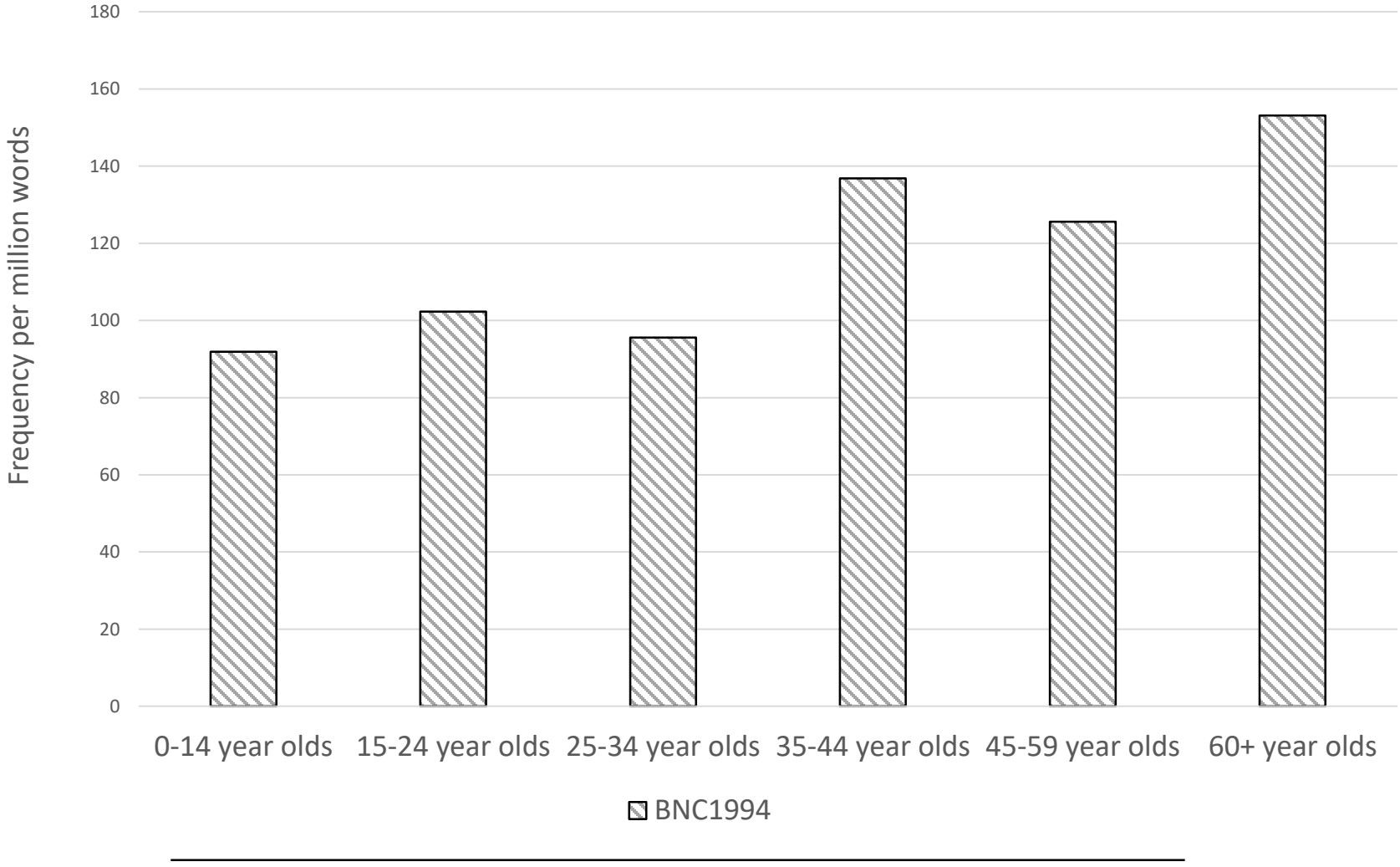
- Age-related effect – how we think/behave at different points across our lives - 20 year olds will differ from 40 year olds e.g. younger people tend to be more liberal and become more conservative as they age.
- Cohort-related effect – how the generation we are born in will affect how we think/behave – e.g. Boomers (1946-1964), Gen-X (1965-1979), Millennials (1980-2000)
- Period-related effect – a change which occurs at a particular point, affecting all age groups and cohorts uniformly (Grusky and Rice 1998)
- Case study: is use of *may* declining over time? (Leech 2002, Millar 2009)

Grusky and Rice (1998) Generation X not so special. www.standford.edu/news/relaged/ 980821genx.html.

Leech, G. (2002) 'Recent grammatical change in English: data, description, theory.' In K. Aijmer and B. Altenberg (eds) *Proceedings of the 2002 ICAME Conference*, Gothenburg, pp. 61-81.

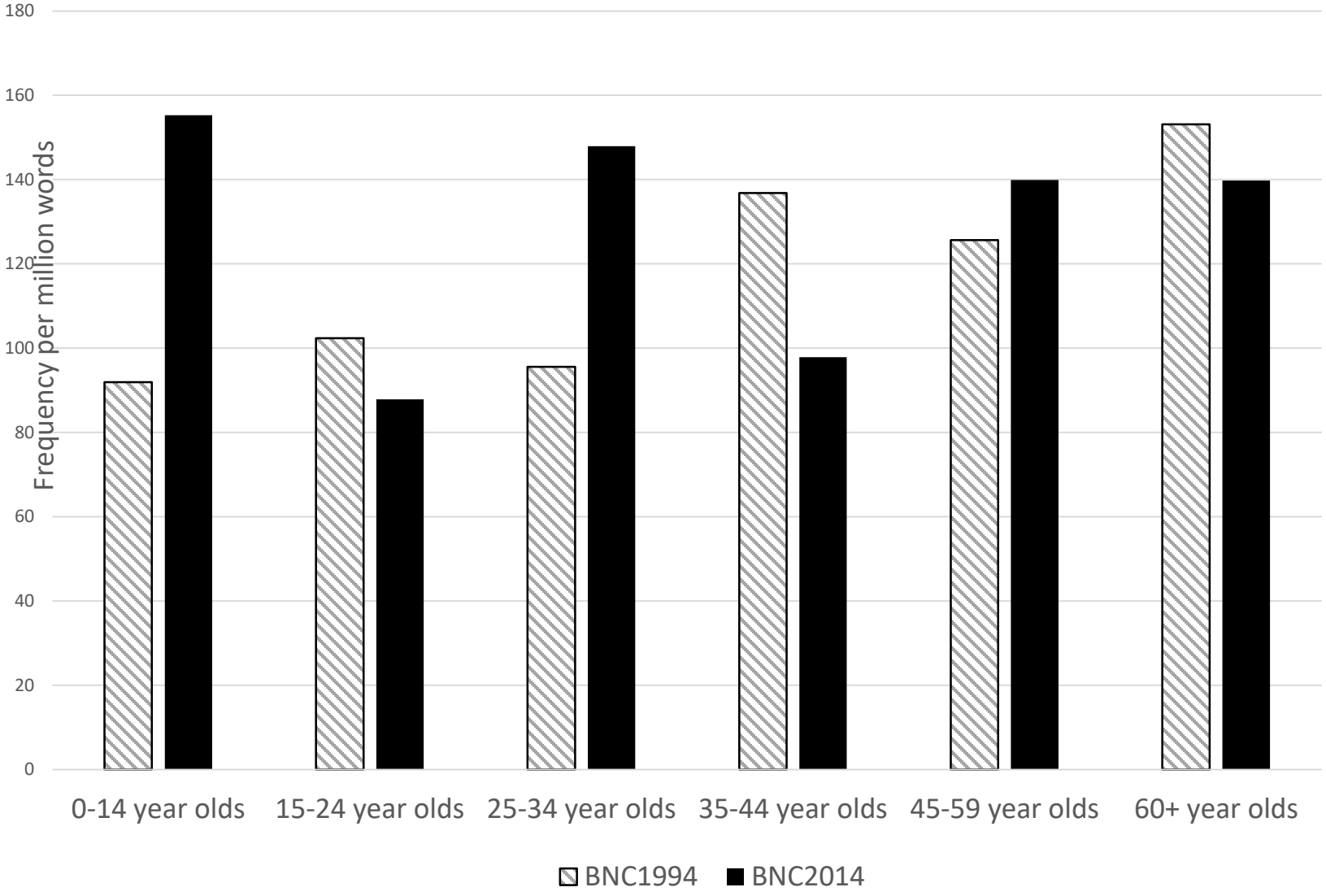
Millar, N. (2009) Modal verbs in TIME. *International Journal of Corpus Linguistics* 14(2): 191-220.

BNC 1994 age groups: *may* as a modal verb

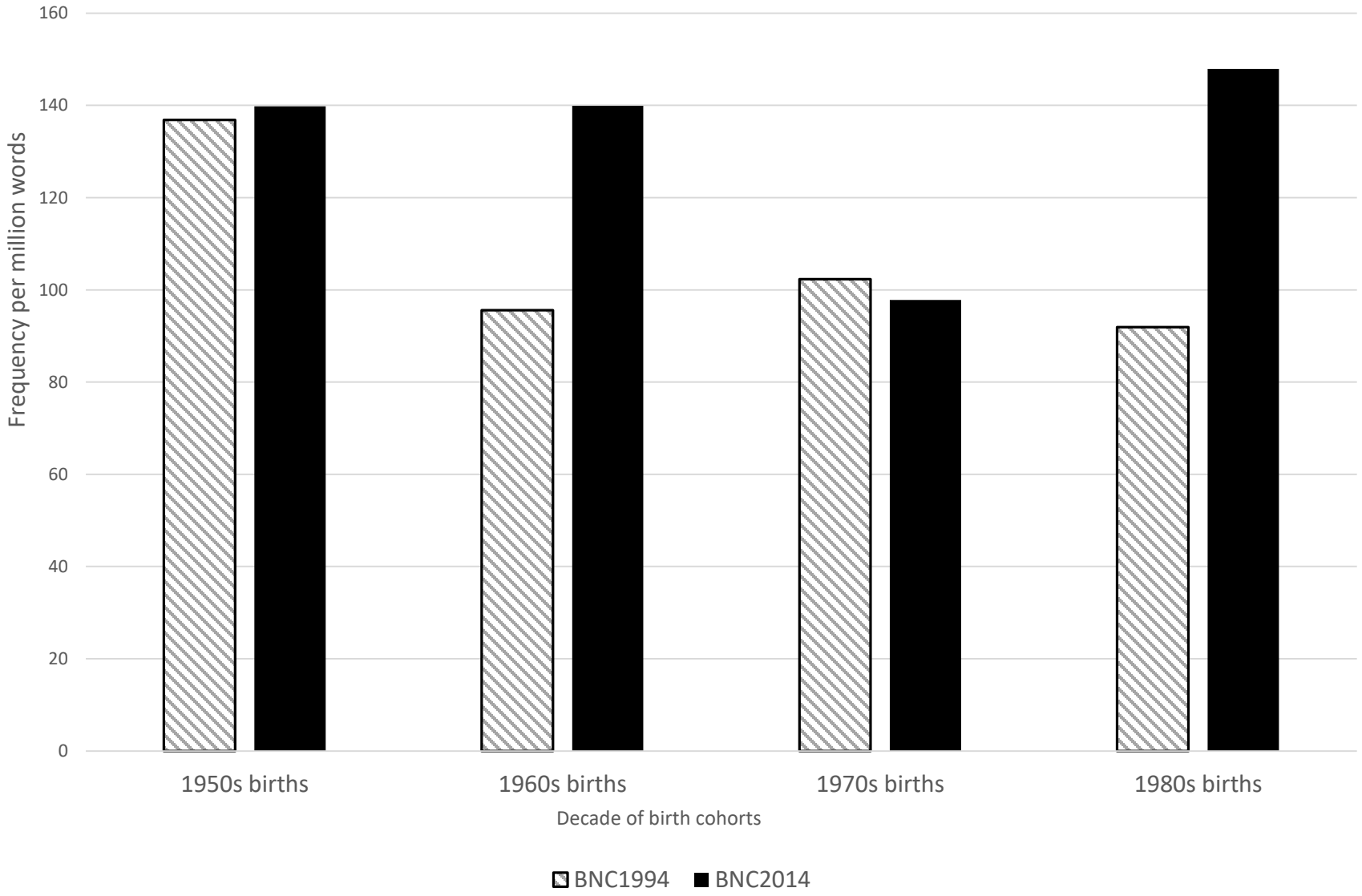


Baker, P. and Heritage, F. (2021) Using corpora in sociolinguistic research. In O'Keefe, A. and McCarthy, M. (eds) *Routledge Handbook of Corpus Linguistics*. 2nd Edition. London: Routledge.

BNC Age groups – comparison of *may* in 1994 and 2014



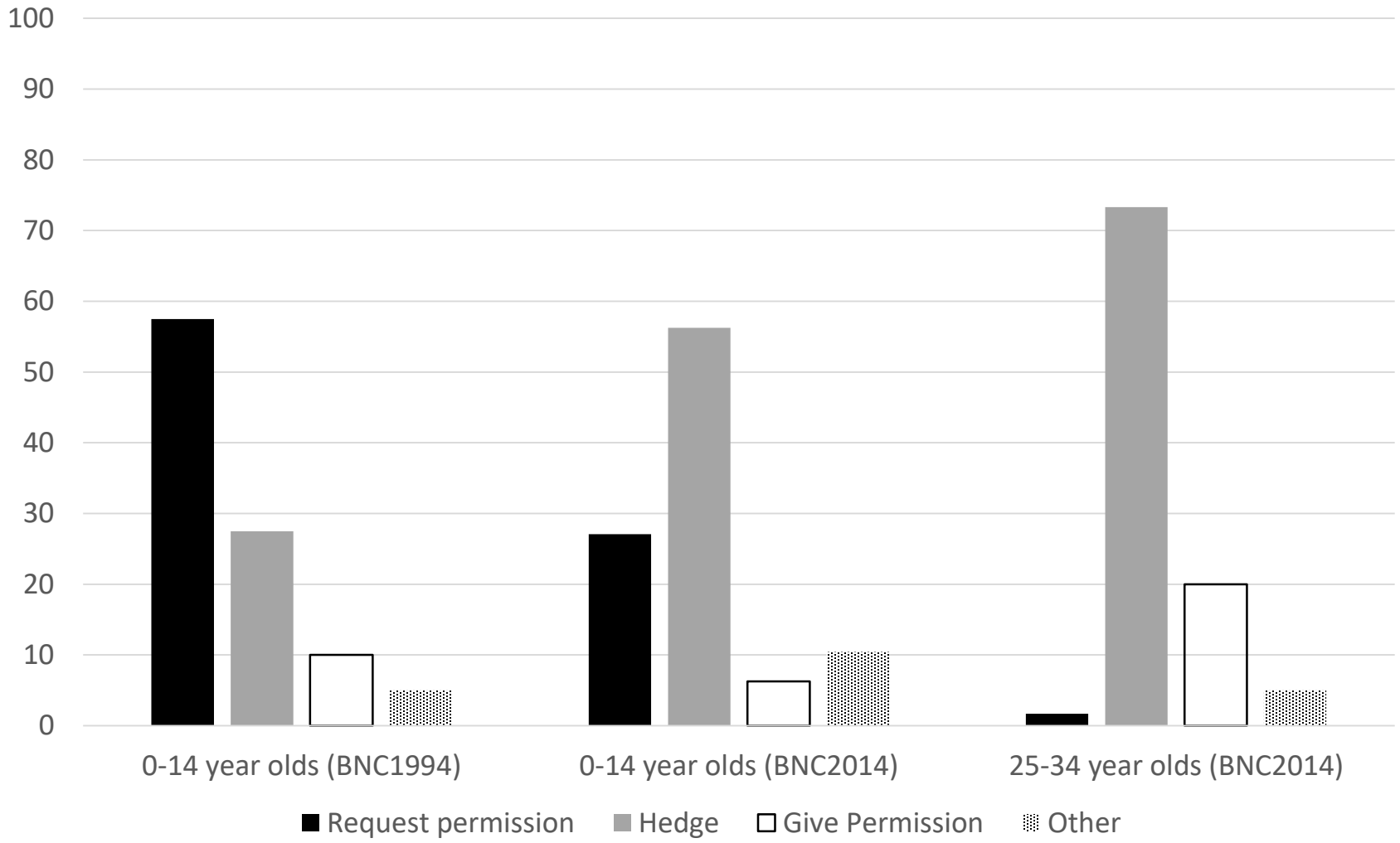
Use of *may* – age cohorts via decade



Contexts of *may*

- 1. Request permission “May I have some milk”
- 2. Give permission “You may do that”
- 3. Hedge a proposition “There may be some milk”
- 4. Others (e.g. not enough context to classify)

Contexts of may



Finding data

1. Build your own corpus, small, specific, lends towards qualitative analyses (Murphy 2010)

2. Use a publicly available large corpus (e.g. BNC Spoken 1994 and 2014 versions) (Brezina et al 2018)

3. pre-categorised data e.g. personal adverts

Brezina, V., Love, R. and Aijmer, K. (eds) (2018) *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*. London: Routledge.

Murphy, B. (2010) *Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk*. Amsterdam: John Benjamins.

Women's adverts

Women seeking men – 22,261 words (189 adverts)

Women seeking women – 21,047 words (208 adverts)

Keywords technique used to compare the sub-corpora directly against one another

---

Baker, P. (2017) Sexuality. In E. Friginal (ed) *Studies in Corpus-Based Sociolinguistics*. London: Routledge, pp. 159-177.

	Women seeking men	Women seeking women
1	man	ladies
2	guy	no
3	me	bi
4	mom	women
5	children	lesbian
6	lonely	w4w
7	gentleman	girls
8	his	female
9	music	pic
10	has	couples
11	live	girl
12	male	stud
13	it	studs
14	the	butch
15	night	femme
16	dick	aa
17	different	feminine
18	marriage	fem
19	swm	men
20	unhappy	woman
21	number	soft
22	sex	hang
23	in	females
24	kind	we

me (key for women seeking men)

- 16 cases (5.5%) used for self-presentation

  *As for **me**, I'm a young woman who is a little lost*

  *About **me**: - 29-Never married/Single – No kids – Latina*

- 242 cases (83%) advertiser in object position (with male as subject)

  *I just want someone who can make **me** happy*

  *I love people who can make **me** laugh, know when it's time to be serious and a gentlemen.*

  *Touch **me**, trust **me**, savour each sensation*

  *Looking for someone to show **me** around town*

  *I am looking for a handsome, professional, well off gentleman who will make **me** feel like a woman.*

we (women seeking women)

- A desire to specify the relationship:

  *Maybe **we** can do Happy Hour or find a good taco Tuesday spot*

  *If you enjoy long walks at the beach, chivalrous behavior and sincerity, perhaps **we** should chat.*

- Implicature of equality, which is grammatically realised

  *Im looking for a female who has a membership already to the gym, so **we** can motivate each other*

  ***We** can make dinner together, watch a good movie at night, and then head to the bedroom*

Culturally speaking, we = the relationship is serious

Machacek, R. (2014) Going plural: making the jump from 'me' to 'we'. Online article at *Swimmingly.com* June 24, 2014

Reynolds, S. and Reynolds, D. (2010) *When I Becomes We… Together as One.* USA: CreateSpace

Finding data

1. Build your own corpus, small, specific, lends towards qualitative analyses (Murphy 2010)

2. Use a publicly available large corpus (e.g. BNC Spoken 1994 and 2014 versions) (Brezina et al 2018)

3. pre-categorised data e.g. personal adverts

4. non-categorised data where categories can be identified by hand e.g. fiction, news articles, transcribed political debates, court transcripts, student essays

Brezina, V., Love, R. and Aijmer, K. (eds) (2018) *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*. London: Routledge.

Murphy, B. (2010) *Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk*. Amsterdam: John Benjamins.

Coupland, N. (2004) Age in Social and Sociolinguistic Theory. In Nussbaum, J. F. and Coupland, J. (eds.) *Handbook of Communication and Aging Research*. Second Ed. London: Routledge, pp. 69–90.

Finding data

1. Build your own corpus, small, specific, lends towards qualitative analyses (Murphy 2010)

2. Use a publicly available large corpus (e.g. BNC Spoken 1994 and 2014 versions) (Brezina et al 2018)

3. pre-categorised data e.g. personal adverts

4. non-categorised data where categories can be identified by hand e.g. fiction, news articles, transcribed political debates, court transcripts, student essays

5. non-categorised data where the speakers/authors make aspects of their identity explicit in the texts e.g. blogs, tweets, forum posts, online reviews

> 'age identity is often nearer to the surface of talk and text than other dimensions of social identification. There are very many social encounters where age is made immediately and obviously salient, and where it becomes a thematic resource for talk.' (Coupland 2004: 84).

Brezina, V., Love, R. and Aijmer, K. (eds) (2018) *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*. London: Routledge.

Murphy, B. (2010) *Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk*. Amsterdam: John Benjamins.

Coupland, N. (2004) Age in Social and Sociolinguistic Theory. In Nussbaum, J. F. and Coupland, J. (eds.) *Handbook of Communication and Aging Research*. Second Ed. London: Routledge, pp. 69–90.

Identifying age and gender in patient feedback (Baker et al 2019)

- Search terms for identifying people in their 20s:
  - 2? year* old
  - twenties
  - 20s

- Search term for identifying women:
  - VB* <<5>> woman

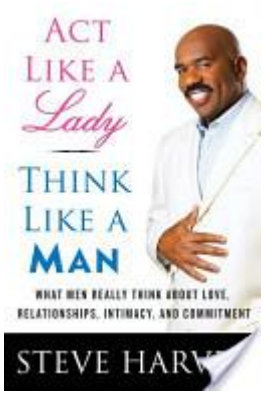
Baker, P., Brookes, G. and Evans, C. (2019) *The Language of Patient Feedback: A corpus linguistic study of online health communication.* London: Routledge.

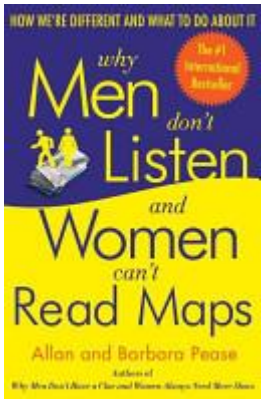
Men's relationship with pain (patient feedback corpus)

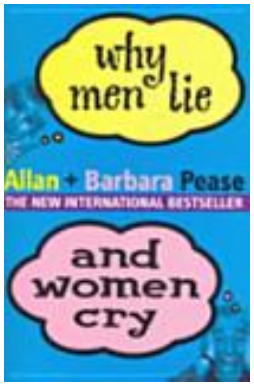
- **Now I am a big 6ft man** who has under gone a lot of dental work and have never felt pain like it. I ended up sahking and crying. I had to wait for an hour afterwards as I couldn't drive my car.
- I was seen x- rayed, treated really well, and then told off for not explaining my pain and told to stop biting my lip, and help them understand my pain, **I explained I was a man,** to which I was told I had damaged my ligaments and shown my x rays
- The dentists are excellent and have always treated me with great care and consideration (**being a man I can't bear pain!**)
- I recently have called many times to make an appointment about my arm **pains**

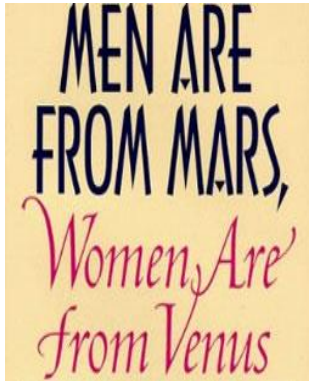
Essentialism

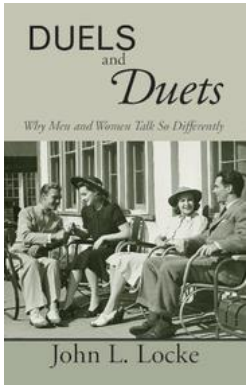
- Over-represent differences between groups e.g. men are x, women are y
- Down-play similarities between groups e.g. both men and women do z
- Ignore differences within groups e.g. some men do x, others do y.
- Hyde (1995) meta-analysis – difference approaching zero
- Eckert and McConnell-Ginet (2003) 'a popular thirst for gender difference'
- Cameron (2008) Mars and Venus is a myth.











Cameron, D. (2008), *The Myth of Mars and Venus*. Oxford: Oxford University Press.
Eckert, P. and McConnell-Ginet, S. (2003), *Language and Gender*. Cambridge: Cambridge University Press.
Hyde, J. (2005), 'The gender similarities hypothesis.' *American Psychologist* 60(6): 581–92.

Findings from corpus linguistics

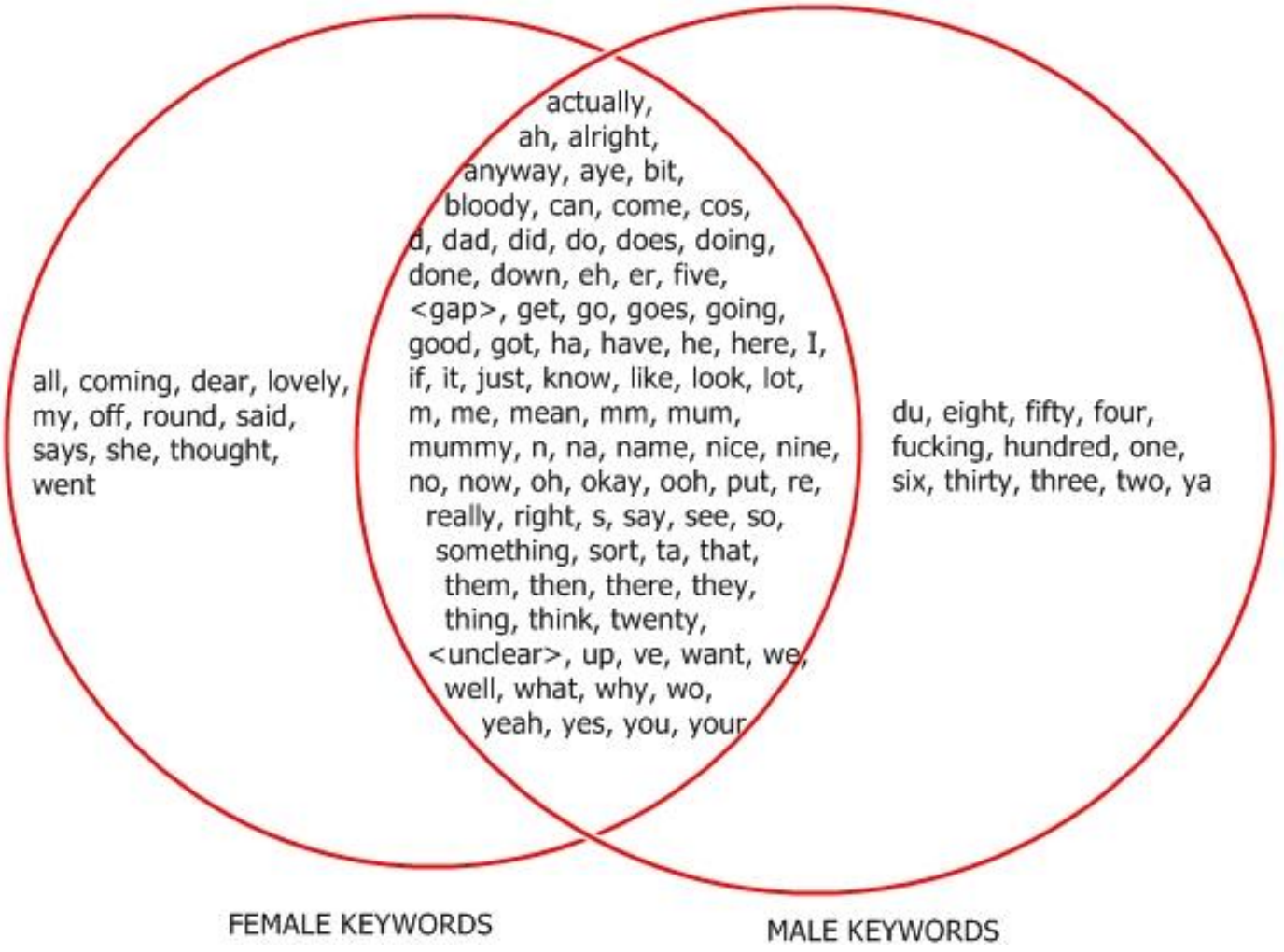
- Rayson et al (1997) females say *I*, *me*, *she*, *her*, *him*, the co-ordinators *because* and *cos* and discourse markers/evaluation *oh*, *lovely*, *nice*, *mm*, *really*. Males used *fuck/fucking* more, as well as other discourse markers: *yeah*, *aye*, *right*, *okay*, *yes, ah* and numbers *two*, *three*, *four*, *hundred*, *number*.
- Koppel et al's (2002) algorithm was able to identify the sex of the writer in BNC texts about 80% of the time. Males used more determiners, numbers, females use more negation, pronouns.
- Schmid (2003) some stereotypes borne out –males use more public affairs and abstract words (*government, tax, option*), females use more clothing, colours and home words (*coat, orange, kitchen*)

Rayson, P., Leech, G., and 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.

Koppel, M., Argamon, S. and Shimoni, A.R. (2002), 'Automatically categorizing written texts by author gender.' *Literary and Linguistic Computing* 17(4): 401-412.

Schmid, H-J. (2003) 'Do men and women really live in different cultures? Evidence from the BNC.' In A.Wilson, R. Rayson and T. McEnery (eds) *Corpus Linguistics by the Lune*. *Lódź Studies in Language 8*. Frankfurt: Peter Lang, pp. 185-221.

Comparing male and female speech against a third (reference) corpus



Lovely and fucking

Lovely

- Used by 318 out of 1,360 female speakers (23%).
- Used by 251 out of 2,448 male speakers (10%).
- Anne uses *lovely* 27 times in 1,974 words
- Kathleen uses *lovely* once in 39,423 words
- 36 out of 1360 women say half of all female cases of *lovely*.

Fucking

- Used by 89 out of 2,448 males (3.63%)
- Used by 57 out of 1,360 females (4.19%)
- Mark said it 274 times in 26,068 words
- 5 male speakers out of 2,448 contributed over half the cases of *fucking*.

Conclusions

- Corpus analysis can show that different social groups use language differently
- Need to do more than just compare frequencies
- Consider differences within a group and similarities between groups
- Consider context of recording e.g. who we're speaking to matters
- Take into account intersectionality
- Go beyond simple word frequencies to consider functions of language