6 Language Use and Identity

6.1 Introduction

The topic of identity has long been of interest to researchers in linguistics, including in the context of health communication. The term *identity* emerged in social science literature during the 1950s, with studies on the topic tending to fall into one of two broad categories; the first involves viewing identity as 'intrapsychic' (i.e., as an internal and fixed quality that reflects 'who we really are'), while the second views identity as a socially constructed role that is acquired or even imposed (Gleason, 1983). Between these categories is the notion of 'ego identity' (Habermas, 1975), which represents a socialised sense of individuality (Baker, 2010: 10). Research in the social sciences had tended to view identity in the second sense identified by Gleason (1983). Indeed, Preece (2016: 3) describes a shift in (applied) linguistic research 'from viewing identity as a set of fixed characteristics that are learned or biologically based to seeing identity as a social construct'.

Language use, or 'discourse', is one of the ways in which identity can be socially constructed. As Burr (2004: 106) puts it, identities are 'constructed out of the discourses culturally available to us, and which we draw upon in our communications with other people'. Importantly, this process is never static or 'complete' but, rather, is active, ongoing and dynamic (Benwell and Stokoe, 2006: 4). For linguists working in fields such as discourse analysis and socio-linguistics, an important consideration is how we might go about actually identifying the kinds of language use or discourses that people use to construct identities. As we will see in Chapter 10, we can approach the construction of identity from a perspective of representation, for example, by studying the language used to represent particular social actors or groups. Another way we can study identity, though, is to examine the language used by a given social actor or group of interest, to get a sense of how their identities are reflected in, and indeed constructed through, the language they use. This latter perspective is the focus of this chapter.

Corpus studies of language and identity often use annotation as a way of marking up the language in the corpus according to the relevant demographic

characteristics of the language users represented within it (Baker, 2010). Annotation at the text level is likely to indicate the genre or mode of the texts contained within the corpus. Annotation at the linguistic level, meanwhile, usually indicates a word's grammatical class or the semantic field to which it belongs (Leech, 1997). In this chapter we focus on the use of annotation to study identity: the annotation of a corpus using demographic metadata relating to the characteristics of the language users represented within it, for instance relating to their age or sex identity (Baker, 2010). For example, when investigating sex, language users might be grouped into categories such as male, female, non-binary, and so forth (Baker, 2014; Baker and Brookes, 2021). Meanwhile, for age the groupings may reflect different age groups such as adolescents, people in their twenties, people in their thirties and so on. The annotations reflecting these categories can be used to divide the corpus into a series of sub-corpora which can then form the basis of an analysis, allowing us to focus on particular groups (e.g., men) or cross-sections of groups (e.g., men in their thirties), or to compare patterns across different groups (e.g., to compare people in their thirties against people in their forties). Thus, annotation has become particularly important to corpus linguistics because it can facilitate more complex and sophisticated analyses.

However, not all researchers will need to, or even *want* to, use annotation in their research. For many studies, it is possible to carry out analyses, including of identity, without the help of annotation. Meanwhile, some researchers regard annotation rather negatively, and it has been argued (e.g., by Sinclair, 1992) that annotation imposes upon the theoretical purity of the corpus and that the particular tags have the potential to be applied inconsistently or inaccurately. Once a suitable taxonomy has been established, we can employ corpus tools to automatically apply tags according to our instructions and can supplement this automated process by manually checking and correcting erroneous tags assigned to a corpus by a computational tagger. However, we should also bear in mind that the larger the corpus we are dealing with, the more laborious and less practical this process of manual checking becomes. Yet, even if we do want to perform annotation on our corpus data because we believe that it will support our analysis, it might not be possible for us to do so if we do not possess the information or metadata necessary to reliably tag the corpus in a way that would be relevant to our research. This obstacle is particularly relevant when using annotation to study identity, as the application of demographic tags depends on the annotator having access to reliable demographic metadata about the language users featured in the corpus. However, such information is not always available, especially when studying anonymous (particularly online) communicative contexts.

In this chapter we present two case studies which demonstrate how we might go about incorporating the consideration of language users' identities into

a corpus analysis. First, we present an analysis of identity in a corpus of health communication data for which demographic metadata about the language users in the corpus *was* available. Second, we present an analysis of identity in a corpus containing texts belonging to a similar genre of health communication, but for which reliable demographic metadata was not available to the researchers (see Section 2.4 for background discussion relating to these studies). Taken together, the case studies presented over the coming pages respectively demonstrate, on the one hand, the affordances of demographic annotation for studying identity and, on the other hand, how as analysts we might work around a lack of such annotations in order to study identity in a corpus of health communication data.

6.2 Using Demographic Metadata

The first case study we present in this chapter, as noted, involves the use of demographic metadata annotation to study identity in a corpus of health communication data. Specifically, the study in question examines language use and sex identity in a corpus of patient evaluations of cancer care services in England (Baker and Brookes, 2022). It is worth first briefly discussing the notion of patient feedback and the context of the patients' evaluations. Since the 1980s, patient feedback exercises have been undertaken by an increasing number of healthcare providers across the globe in order to monitor the quality of the services they provide and stimulate improvements where these might be needed (Vingerhoets et al., 2001). While the reliability of patient feedback as an indicator of the technical quality of care can be debated (Coulter, 2006), it has nevertheless become a staple way of measuring and regulating healthcare standards (Graham and Woods, 2013) while also ensuring public involvement in the design and improvement of healthcare provision (Coulter, 2013).

6.2.1 The Corpus and Its Context: Patient Feedback on Cancer Services in England

Patient feedback can be obtained in a variety of ways, and the data analysed in this study was a specialised corpus of written feedback on cancer care services provided by patients responding to England's Cancer Patient Experience Survey (CPES). Responses were provided both through online and so-called pen-and-paper forms, with the latter subsequently being digitised to render them amenable to computational corpus analysis. The CPES form allows patients to provide both quantitative and qualitative feedback. The quantitative part of the form asks, 'Overall, how would you rate your care?', to which patients can respond by providing a score between 0 and 10, where a score of 0 indicates that they definitely would not recommend a service and a score of 10

indicates that they definitely would recommend it. Respondents could then describe their experiences and explain the score they gave by providing qualitative feedback across three free-text boxes preceded by the following questions: 'Was there anything particularly good about your NHS cancer care?'; 'Was there anything that could have been improved?'; and 'Any other comments?'. For the purposes of this study, all the comments each patient provided were combined to form a single 'text' (i.e., one text per piece of feedback).

The resulting corpus comprised 214,340 comments (14,403,694 words), relating to cancer services provided in England between 2015 and 2018. This data was made available to the researchers by NHS England's Insight and Feedback team in a spreadsheet format in which each comment was accompanied by details about each patient's care and certain socio-demographic characteristics. Some of these details were obtained by the organisation from patient records, while others were provided by the patients themselves, as requested on the feedback form. The researchers then used this metadata to annotate each comment with details about the patient who provided it. These included details relating to the patients' treatment, including the duration of treatment and the site on which they received it, as well as about their identity, including age, ethnicity, first language, sex, and sexuality. Here we focus on the analysis based on sex.

6.2.2 *Analysis of Language and Sex Identity Based on Demographic Metadata*

There were three options for patient sex in the CPES form: 'Male', 'Female' and 'Prefer not to say'. The vast majority of patients did respond to this question on the form, with approximately 54 per cent of the respondents identifying as female and approximately 46 per cent identifying as male. The researchers mounted the corpus on CQPweb (Hardie, 2012) and used the sex-related tags to divide the corpus into two sub-corpora: one containing comments provided by female patients and the other containing comments provided by male patients.

To compare the language used by these groups, the researchers then compared these sub-corpora against each other using the keywords technique (introduced in Chapter 1). Specifically, they generated two sets of keywords: one by using the male comments sub-corpus as the target and comparing it against the female comments sub-corpus as the reference, and the second by swapping these around and using the female comments sub-corpus as the target and the male comments sub-corpus as the reference. The resulting sets of keywords represent language use that was characteristic, respectively, of male and female patients' comments, when compared with each other. The following list shows the top-30 keywords

for male patient comments, compared to female patients' comments (ranked by frequency (in brackets))[1]. These keywords were obtained using log likelihood (Dunning, 1993) with a log ratio cut-off (Hardie, 2014).

class (4,969), treatment (46,228), hospital (39,990), good (25,776), no (23,674), by (20,851), (18,719), GP (15,544), first (13,534), NHS (12,360), yes (7,509), months (6,853), blood (6,019), test (4,094), problem (3,892), condition (3,479), general (3,452), thanks (2,870), bladder (2,611), attention (2,518), carried (2,382), bowel (2,348), quality (1,520), professionalism (1,396), period (1,336), myeloma (1,306), removal (1,298), kidney (1,258), successful (854), endoscopy (846)

In addition to using statistical measures, a further step the researchers implemented was to remove keywords that occurred less than 50 times per million words (PMW) in both corpora. This step helped filter out those keywords that denoted proper nouns and sex-specific types of cancer, as well as the treatments associated with these. It should also be noted that CQPweb counts punctuation marks as tokens (and thus as potential keywords), which is why a bracket was key for the male patients' comments.

The researchers then closely analysed the keywords in order to identify their main functions in the feedback and whether and how these might relate to the patients' sex identities. To do this, they used the concordance view to access the uses of each keyword within its wider textual contexts and usually accessed entire comments when interpreting the keywords' functions. For keywords that occurred more than 100 times, the researchers analysed 100 randomly sampled cases.

Analysing these keywords, the researchers were able to identify a series of differences in the language used by male and female patients. The keyness of the generic illness-referring nouns *condition* and *problem* in the male patients' comments indicated a pronounced focus in these texts on the particular health issues that caused male patients to have to visit a provider. By analysing these words within their contexts of use, the researchers found that the male patients used them to characterise their care in terms of processes, of which they, their bodies and their health problems were the objects. This also helped to account for the keywords *carried*, *blood*, *endoscopy*, *removal* and *test*. Here is an example:

> My **condition** was discussed after a series of **blood tests** for diabetes, prostate **problems** and erectile dysfunction.

[1] Note that these keywords ranked as the top 30 when ordered according to the log likelihood scores assigned to them. For statistical information about the keywords, see Baker and Brookes (2022: 18–19).

Medical staff were often presented as performing the procedures that the male patients underwent, and could be indexed through uses of the keywords *general*, *hospital*, and *NHS,* which constituted a kind of metonymic reference whereby the male patients presented the performance of a few individual staff members in terms of representing an entire hospital or even the health system as a whole.

> Very happy with all aspects of treatment I have received from the **NHS**. Very impressed.

Rhetorically, this kind of reference has the potential to present the patient's experiences as applying not just to them or the individual staff members who treated them, but to staff comprising an entire hospital, the wider healthcare system and potentially other patients.

The male patients' comments also exhibited a marked use of words quantifying time (e.g., *months*, *period*), which tended to be used to quantify the amounts of time that the male patients had to wait for (typically) diagnoses and appointments.

> The first thing to do on entering the department was always to look at the screen to whether there were any delays. These often changed mainly to extend the waiting **period** up to 1 hour.

This attention to detail also helps account for the male patients' marked use of brackets, which they use to specify details such as the type of cancer they received treatment for, the type of treatment or procedures they underwent, and the ward or unit they received treatment on.

> Too many hospitals were involved ([name of hospital and ward]**)** in my opinion this resulted in months of inactivity and delay.

A somewhat curious set of keywords, also reflective of style, are the words *no*, *thanks* and *yes*. Rather than reflecting how the male patients linguistically performed evaluation, these words arose as key because of the way these patients interacted with the feedback form. In particular, the male patients interacted almost dialogically with the voice of the feedback form, answering the prompt questions literally (*no*, *yes*) and performing the speech act of thanking service providers, who were assumed to be represented by the voice of the form.

> **Yes**. It was dealt with in a timely fashion and consideration was given by the consultant for the eventual cosmetic appearance of the site following removal of the lesion.

> Just **thanks** for keeping me alive!

To account for this trend, the researchers considered another factor: age. They looked at the frequencies of the keywords *no*, *thanks* and *yes* by male and female patients at different ages. They found that the use of these terms

increased with age, and their qualitative analysis of these words by patients at different age groups confirmed that this more dialogic style and literal interpretation of the prompt questions framing the free text boxes was indeed a feature more typical of older patients, particularly those over the age of 65 (see Baker and Brookes, 2022: 24). It could be the case that younger patients were more accustomed to the (synthetic) personalisation of such public discourse, rather than regarding it as an attempt by the organisation to establish personal dialogue. Importantly, by looking at the wider set of demographic metadata available to them, the researchers also found that the male patients featured were, on average, older than the female patients represented in the corpus. For example, male patients aged 75–84 contributed 22.42 per cent of all the words in the male patients' comments overall (compared to just 8.06 per cent contributed by this age group in the female patients' comments), while male patients aged 85+ contributed 3.9 per cent of all the words in the male patients' comments (compared to 1.45 per cent contributed by this age group in the female patients' comments). Thus, the researchers concluded that the keyness of these items was a product of the socio-demographic make-up of the corpus, with the sample of male patients represented within it being older, on average, than the sample of female patients.

Keywords for the female patients' comments were then obtained by comparing these against the male patients' comments. The resulting keywords, which are shown in the following list, were obtained in the same way as the male patients' keywords seen earlier.[2]

I (302,785), had (73,119), me (63,423), they (39,214), when (31,967), you (31,028), n't (29,681), so (28,576), chemotherapy (24,365), did (21,234), nurse (20,478), caring (18,756), feel (16,105), felt (13,101), radiotherapy (12,972), unit (11,739), kind (11,074), she (10,403), wait (10,315), everyone (9,969), oncologist (7,901), wonderful (7,106), her (6,637), amazing (6,436), chemo (6,056), supportive (4,998), busy (4,188), husband (3,164), lovely (2,961), lump (2,183)

While the male patients' comments focused characteristically on procedural and transactional aspects of service, the female patients tended to discursively situate themselves within their comments (as reflected, e.g., in the keywords *I* and *me*). This more personalised focus also gave rise to a discussion of the emotional impacts that the female patients' experiences had on them (as indicated in uses of the keywords *felt* and *feel*). Other keywords in the female patients' comments provide further evidence for this focus on interpersonal aspects of care, as staff members are evaluated using keywords such as *kind*, *lovely*, *supportive*, and *caring*.

[2] For statistical information about the keywords, see Baker and Brookes (2022: 18–19).

The pronounced focus on interpersonal skills in the female patients' comments also gave rise to a stronger focus on individuals (*she*, *oncologist*, *her*, *nurse*), including the roles and experiences of relatives (*husband*), and units and smaller teams of staff (*they*, *unit*). Meanwhile, the keywords *chemotherapy*, *radiotherapy* and *chemo*, while referring ostensibly to types of treatment, were also used to refer to teams of staff (including metonymically through references to specific wards).

> The nurses in the **chemo** ward were absolutely fantastic at [name] Unit.

Therefore, while the male patients tended to present their evaluations as relevant to entire hospitals and the wider healthcare system, the female patients tended to focus on individuals or, at the most, small teams.

The keyword *everyone* was found to be frequently used in reference to other patients who attend a particular provider. The researchers interpreted this as a rhetorical strategy whereby the female patients rendered their experiences as more generalisable. A similar strategy these patients used to this end was to use the general *you*.

> **Everyone** is looked after in the same wonderful way.

> All staff in the cancer care unit are friendly, caring and helpful. They all welcome **you** and take care of **you** as if **you** are a part of the family.

Like the male patients, the female patients also focused on the theme of waits, as indicated in the keyword *wait*. However, the female patients were found to describe and evaluate waits in less-precise terms than the male patients (specifying the duration of their waits in just 15 per cent of cases). This was why the words *months* and *period* were key for the male patients' comments compared to the female patients', even if both groups focussed on waits in their feedback.

> Sometimes I have a long **wait** on surgery day but I don't think this can be helped.

In summary, by using socio-demographic metadata to annotate their corpus, then, Baker and Brookes (2022) were able to use the keywords technique to compare comments from the male and female patients represented in their data. This comparison revealed some differences in terms of the thematic content of the comments and what male and female patients focussed on or foregrounded within their feedback, finding, for example, that where male patients focussed on transactional aspects of care (e.g., operations), female patients tended to focus more on the people involved in their care, as well as the interpersonal relationships they established with staff. Moreover, analysis of the keywords also indicated how even shared areas of focus (e.g., waits) could be described and evaluated using characteristically different types of language, which involved providing differing levels of detail around those waits. Baker and

Brookes's (2022) study also demonstrated the value in having a wider set of demographic metadata tags to draw on for the purposes of interpreting patterns. In particular, they observed how male patients tended to engage in a more dialogic way with the feedback form than female patients did, and that this resulted in a markedly frequent use of words such as 'yes', 'no', and 'thanks', which functioned to answer the rhetorical prompt questions that framed the free text boxes. By drawing on metadata relating to patient age, the authors were able to conclude that this was likely to be an age effect, with older patients exhibiting this dialogic style more than younger ones. Thus, the authors argued that the (on average) older sample of male patients represented in their corpus was likely to be the reason for this difference. Without this wider demographic metadata, the authors might have concluded that this pattern was related to sex identity, rather than being down to a (likely) intersection of sex and age.

6.3 Using Mentions of Identity in the Data

The availability of reliable demographic metadata can, as the previously provided case study demonstrates, prove useful for organising corpus data for the purposes of studying identity. However, such metadata is not always available to us as analysts, particularly when we are studying language produced in anonymised contexts, such as many forms of online communication and social media. In such cases, if we are interested in studying the relationship between language use and identity, we must find another way of 'identifying the identities' of the language users represented in our data. One way we can do this is by using the language in the data itself, and looking for cases where language users openly disclose the particular aspects of their identity that we are interested in studying. Baker and co-authors (2019) employed this same approach. This study, which was a predecessor to that described in the previous section, involved a comprehensive analysis of the language of online patient feedback about a wide range of areas of healthcare provision in England (i.e., not just cancer services, as was the focus of Baker and Brookes's (2022) study). Their analysis was based on 228,113 online patient comments (28,971,412 words), posted to the NHS Choices website between 2013 and 2015.

6.3.1 Identifying Mentions of Sex Identity in a Corpus

As in the later study by Baker and Brookes (2022), Baker and co-authors (2019) undertook analyses that were designed to answer a series of questions that were put to them by their stakeholder partners in the NHS (see Chapter 2). As in the other study, here the stakeholder partners were also interested in learning about how patients' identities, including their sex identities, might influence the kind of feedback they gave and the language they used for their feedback. However, unlike

in the other study, the set of patients' comments available to Baker and co-authors (2019) did not contain demographic metadata, since the online form through which patients posted their comments did not provide the facility through which such information could be provided. As such, to analyse the influence of patients' sex on the comments, Baker and co-authors (2019) searched for instances where patients mentioned their sex identity. In this way, they explored how sex identity categories 'cropped up', were 'oriented to' or otherwise 'noticed' by patients in their comments. This approach was broadly inspired by the approaches to 'membership categories' (Sacks, 1995) and 'person reference forms' (Schegloff, 1996) in conversation analysis, and offered the practical advantage of allowing the analysts to study how issues such as patients' sex identity figure in the comments, despite the absence of demographic metadata.

To find instances in which patients mentioned their sex identity, the analysts searched the corpus of comments for stretches of text in which either of the words *be*, *is*, *are*, *was* or *were* was followed by *man* or *woman* within the next five words (see also Chapter 2). This search yielded 518 cases for *man* and 332 for *woman*. Not all results actually involved cases where patients identified their own sex identity, as sometimes they could refer to that of another person (e.g., 'My GP is a good man'). Once such cases were removed, the analysts then took 200 cases (at random) for male self-identifiers and another 200 cases for female self-identifiers.

6.3.2 *Analysis of Language and Sex Identity Based on the Mention of Identity*

Baker and co-authors (2019) carried out various kinds of analyses on the resulting four hundred comments, including generating sets of keywords by comparing the comments from the male self-identifiers against the female ones, and vice versa. Tables 6.3 and 6.4 show, respectively, the keywords for the male and female patients' comments when these are compared against each other. Note that rather than take the top keywords from each list, by using relatively small samples of comments for this portion of their analysis, Baker and co-authors (2019) were able analyse all keywords produced through these comparisons (they considered all keywords that occurred at least 10 times and received a log likelihood score of at least 15.13, $p < 0.0001$.).

Keywords for male patients' comments compared to female patients' comments, ranked by frequency (in brackets)

the (1,688), of (543), have (484), this (343), you (262), been (189), appointment (157), practice (133), years (104), always (79), good (70), dentist (61), old (60), year (58), male (57), NHS (57), times (50), problem (48), given (47), many (45), helpful (43), minutes (34), condition (33), advice (29), men (28), poor (25), working (25), dental (25), wife (23), difficult (22), three (22), number

(20), seems (18), surgeries (16), results (16), write (15), consultation (15), referral (15), non (13), following (13), bipolar (13), allowed (12), money (12), visits (12), recent (11), pains (11), five (10), doubt (10), surely (10)

Keywords for female patients' comments compared to male patients' comments, ranked by frequency (in brackets)

i (1,994), and (1,379), was (734), my (697), me (478), they (461), on (288), am (249), woman (153), after (139), hospital (110), said (89), never (83), did (75), life (63), ward (58), female (53), experience (53), first (53), didn't (50), her (45), baby (43), room (40), came (40), husband (39), lady (31), birth (31), pregnant (31), couldn't (30), women (29), around (29), lovely (29), wonderful (26), pregnancy (25), midwife (24), labour (21), grateful (20), midwives (18), amazing (17), impressed (17), nice (17), breast (15), crying (14), elderly (14), antibiotics (14), notes (13), booked (13), broken (11), drs (10), appt (10)

Unsurprisingly, the keywords show that female patients are more likely to use female-marked terms (usually when talking about themselves) and pregnancy, while male patients are more likely to use male-marked terms. Perhaps more interestingly, Baker and co-authors (2019) point out the differences in pronoun use. Female patients have a wider range of personal pronoun keywords and are more likely to use first-person forms (e.g., *me*, *my*, *I*). These tended to be used to describe personal experiences, often within narratives.

> Since there was no available appointments for **me**, **I** had to take a phone call consultation which **I** didn't mind. **They** told **me** to bring in a urine sample and **they** would leave the antibiotics with the receptionist.

The only pronoun keyword for the male patients was *you*. This tended to be used by the male patients to present their own experiences, but in a more impersonal way. This keyword was used regularly with the conditional *if* as a way of both addressing the presumed reader of the comment and making the attested experience appear more generalisable.

> If **you** call when the switchboard opens more often than not the line is engaged.

This use of the generalisable *you* by the male patients supports Charteris-Black and Seale's (2010: 64–5) observation that male patients tend to use more second-person pronouns than female patients, which they hypothesise is intended to 'make their accounts seem less personal and more objective and generalizable'. The patterns in personal pronoun usage suggested by the keywords listed indeed suggest different stereotypical strategies in terms of

the ways that male and female patients compose their feedback, with some female patients being more likely to describe their experience in personal terms and some male patients attempting instead to impersonalise and accordingly generalise their narratives.

Another key difference identified by Baker and co-authors (2019) concerned the use of adjectives, especially those used in the explicitly evaluative aspects of feedback. Female patients' key evaluative adjectives included *lovely*, *nice*, *wonderful*, and *amazing*. Adjectives such as *lovely* and *nice* were argued to give a general impression of positive evaluation, but the authors also point out that they can be used in so many contexts that they eventually lose their meaning (becoming semantically 'empty'), especially if they are used frequently.

> I love this practice and can't sing its praises enough, the doctors, nurses and reception staff are all **nice** the practice management are **lovely** and always speaks or acknowledges.

An interesting keyword for the male patients is *pains*. Baker and co-authors (2019) point out that the word 'pain' can either be a count or a non-count noun, where saying we 'feel pain' is a non-count use but saying we 'feel a pain' is a count use, as it makes the pain a countable entity. The authors argue that when male patients discuss pain they are perhaps more likely to use the count, plural form as a way of emphasising the severity of their pain in order to legitimise their complaint without violating the gendered assumption that men should not feel pain, or at least the expectation that they should not usually complain about it (see also Jaworska and Ryan, 2018).

> I recently have called many times to make an appointment about my arm **pains**

Another notable difference indicated in the keywords is that the male patients are more likely to use words denoting time and quantification (i.e., *times*, *always*, *years*, *year*, *minutes*, *recent*, *many*, *three*, *five*, *number*). In some cases, a single comment could involve a rather intense use of such quantification, as in the following example:

> They **always** take between 20 to 40 **minutes** to speak with someone. Compare [Anonymised] to my previous surgeries this is at least 5 times faster. You should **always** remember a surgery has thousands of people registered and only a couple of administrators looking after it. If 10 people call at the same time you will have to wait for the 2 people working to get through them, which will take about 5 **minutes** a call, so 25 **minutes** if you were number 10!

Baker and co-authors (2019) interpret comments like this as representing a pronounced use of quantification rhetoric (see also Potter et al., 1991). They suggest that it constitutes another strategy and seems to represent yet another way in which some male patients sought to lend legitimacy and credibility to their complaints, while at the same time strengthening the impact

of their feedback. For instance, in this example, the use of quantification rhetoric renders the feedback more specific and accurate (and thus more compelling) and helps emphasise the amount of inconvenience this patient experienced. Charteris-Black and Seale (2010) similarly observed that men tend to quantify and use numbers more often than female patients when talking about health and illness, a feature which they trace to the adoption of a traditionally masculine discourse style, arguing that 'men following traditional masculine styles have a discursive orientation to measuring and counting entities and processes rather than talking directly about what is happening to their own bodies' (2010: 71).

6.4 Conclusion

In this chapter we have considered two broad approaches to studying identity in a corpus of health communication data: one in which we can rely on demographic metadata tags and another in which we instead rely on cases where language users mention aspects of their identities in the language they produce. We focussed on two case studies which compared the language used by male and female patients in feedback given on healthcare services, with each study representing each of these approaches. Both approaches were helpful, in the sense that they allowed the analysts to organise and subsequently analyse their data in a manner that revealed interesting differences which the authors linked to differences in the performance of gender identities at the discursive level.

Patients' identities and the ways they construct these in their comments were found to have ramifications for how they evaluated healthcare services, including the types of linguistic strategies they used to frame this evaluation, contextualise their perspectives and legitimise their arguments. Interestingly, although both studies used quite different approaches, we can nevertheless observe some similarities in their findings. For example, among other things, both studies found that female patients tended to employ a more personalised style of feedback, while male patients tended to draw more on numbers and quantification. This suggests that some of the patterns reported across both studies are likely to be features of feedback in general, while also reflecting broader distinctions in the discourses associated with performances of particular gender identities. At the same time, the similarities in findings could be viewed as a kind of methodological triangulation; since these findings, which could apply to patient feedback as a genre in general, were arrived at through separate methodological routes, we can perhaps be more confident about their validity.

This point notwithstanding, either approach was also found to present certain advantages but also disadvantages, relative to the other. The annotation-based approach, although resource-intensive due to the need to annotate a large corpus with socio-demographic metadata, offers clear advantages in providing

comprehensive insights into the socio-demographic balance of the corpus. Moreover, annotating a corpus with reliable demographic metadata permits the subsequent analysis of differences (and similarities) in language use among people from different identity-based backgrounds, at scale. An obvious advantage of this is that findings can be supported with statistical evidence, thus rendering them more robust and potentially more generalisable to the populations or healthcare contexts under study. However, a challenge of the approach is that it can create the trap of oversimplifying identity by relying on broad socio-demographic categories, potentially leading to us overlooking nuanced types of identity relations. Another notable limitation is the potential misinterpretation of statistically significant correlations as *causal* relationships, which might lead to us obscuring from our analytical gaze the possible influence of other aspects of language users' identity, in favour of focusing only on statistically significant trends.

The alternative approach, which involves relying on mentions of identity by language users in the language they produce, brings the advantage that we, as analysts, can perhaps be more confident that the aspects of identity we are focusing on to explain a linguistic pattern are indeed relevant to that pattern (since the language users invoke this themselves in their discourse). In other words, we can be more confident that an aspect of identity is relevant, as the language user has made it relevant by invoking it in their discourse. However, this approach also has several limitations, perhaps foremost being the relatively small amounts of data it is likely to give us (indeed, Baker et al. (2019) were only able to compare 200 texts each for male and female patients, from a corpus of more than 228,000 texts). The likely small sizes of samples obtained this way can restrict our analytical optionality (e.g., in terms of statistical measures available to us). Moreover, it can pose issues regarding data representativeness, both in terms of how far we can generalise on the basis of such small datasets, as well as what texts sampled in this way actually represent. In other words, we should ask: do texts in which people mention their identity represent a particular sort of language use, and can this be generalised to the wider population under study? The answer to this question will likely depend on the kind of language use being studied but should be considered critically if we are to take this kind of approach to sampling texts.

In view of the kinds of issues discussed previously, Baker and Brookes (2022) suggest that the most robust approach to studying identity in a corpus might be a mixed one which involves combining both of the approaches explored in this chapter. They suggest that this could involve assembling a large, demographically annotated corpus (where possible), but using the kind of approach based on cases where patients mention their identity (utilised by Baker et al. (2019)) to identify aspects of identity that are relevant to the kind of discourse and context under study, which can then be subjected to larger

scale, quantitative analysis. Importantly, they stress that such an approach could help not only in identifying which aspects of identity are deemed relevant to the language users themselves but also support the interpretation of observed patterns and help account for the role of intersectional aspects of identity in a more data-driven way (i.e., by looking at the interaction of different aspects of identity because language users have evoked these themselves, rather than merely creating intersectional categories just because we have the demographic metadata available to us to do so).

References

Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press.

(2014). *Using Corpora to Analyze Gender*. Bloomsbury.

Baker, P. and Brookes, G. (2021). Lovely Nurses, Rude Receptionists, and Patronising Doctors: Determining the Impact of Gender Stereotyping on Patient Feedback. In J. Angouri and J. Baxter (eds.), *The Routledge Handbook of Language, Gender, and Sexuality* (pp. 559–71). Routledge.

(2022). *Analysing Language, Sex and Age in a Corpus of Patient Feedback: A Comparison of Approaches*. Cambridge University Press.

Baker, P., Brookes, G. and Evans, C. (2019). *The Language of Patient Feedback: A Corpus Linguistic Study of Online Health Communication*. Routledge.

Benwell, B. and Stokoe, E. (2006). *Discourse and Identity*. Edinburgh University Press.

Burr, V. (2004). *Social Constructionism*, 2nd ed. Routledge.

Charteris-Black, J. and Seale, C. (2010). *Gender and the Language of Illness*. Palgrave Macmillan.

Coulter, A. (2006). Can Patients Assess the Quality of Health Care? *British Medical Journal*, *333*, 1–2. https://doi.org/10.1136/bmj.333.7557.1.

(2013). Understanding the Experience of Illness and Treatment. In S. Ziebland, A. Coulter, J. D. Calabrese and L. Locock (eds.), *Understanding and Using Health Experiences: Improving Patient Care* (pp. 6–15). Oxford University Press.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, *19*(1), 61–74. Available at https://aclanthology.org/J93-1003.

Gleason, P. (1983). Identifying Identity: A Semantic History. *Journal of American History*, *69*(4), 910–31. https://doi.org/10.2307/1901196.

Graham, C. and Woods, P. (2013). Patient Experience Surveys. In S. Ziebland, A. Coulter, J. D. Calabrese and L. Locock (eds.), *Understanding and Using Health Experiences: Improving Patient Care* (pp. 81–93). Oxford University Press.

Habermas, J. (1975). Moral Development and Ego Identity. *Telos, 1975*(24), 41–55. https://doi.org/10.3817/0675024041.

Hardie, A. (2012). CQPweb: Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics*, *17*(3), 380–409. https://doi.org/10.1075/ijcl.17.3.04har.

(2014). *Log Ratio: An Informal Introduction*. Online. http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/. Accessed 16 May 2023.

Jaworska, S. and Ryan, K. (2018). Gender and the Language of Pain in Chronic and Terminal Illness: A Corpus-Based Discourse Analysis of Patients' Narratives. *Social Science & Medicine*, *215*, 107–14. https://doi.org/10.1016/j.socscimed.2018.09.002.

Leech, G. (1997). Introducing Corpus Annotation. In R. Garside, G. Leech and T. McEnery (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 1–18). Routledge.

Potter, J., Wetherell, M. and Chitty, A. (1991). Quantification Rhetoric – Cancer on Television. *Discourse and Society*, *2*(3), 333–65. https://doi.org/10.1177/0957926591002003005.

Preece, S. (2016). Introduction: Language and Identity in Applied Linguistics. In S. Preece (ed.), *The Routledge Handbook of Language and Identity* (pp. 1–16). Routledge.

Sacks, H. (1995). *Lectures on Conversation: Volumes 1 and 2*. Blackwell.

Schegloff, E. A. (1996). Some Practices for Referring to Persons in Talk-in-Interaction: A Partial Sketch of a Systematics. In A. Fox (ed.), *Studies in Anaphora* (pp. 437–85). John Benjamins.

Sinclair, J. M. (1992). The Automatic Analysis of Corpora. In J. Svartvik (ed.), *Directions in Corpus Linguistics: Proceedings of the Nobel Symposium 82* (pp. 379–97). Mouton De Gruyter.

Vingerhoets, E., Wensing, M. and Grol R. (2001). Feedback of Patients' Evaluations of General Practice Care: A Randomised Trial. *BMJ Quality & Safety*, *10*, 224–8. https://doi.org/10.1136/qhc.0100224.