Studying language and identity in a corpus of computer-mediated communication with (and without) sociodemographic metadata

Gavin Brookes
Lancaster University
E-mail: g.brookes@lancaster.ac.uk

Abstract

This paper explores the methodological and interpretive implications of analysing language and identity in large corpora of computer-mediated communication (CMC), both with and without sociodemographic metadata. Drawing on a 14-million-word corpus of online patient feedback about UK cancer care, I compare two approaches: one using metadata (e.g. patient-declared sex) and another relying on patients' in-text self-references. The metadata approach enables large-scale, statistically grounded comparisons, revealing broad patterns, such as male patients' focus on procedures and female patients' emphasis on emotional and interpersonal dimensions of care. The self-reference approach, while limited by smaller sample sizes, offers nuanced insights into how patients perceive and mobilise intersecting aspects of identity, including sex and age. The paper highlights the trade-offs between scale and contextual richness, advocating for a combined, bottom-up and top-down approach. It concludes that identity analysis in CMC benefits from attending to both declared demographic categories and emergent, textually embedded identity cues.

Keywords: corpus linguistics, identity, sociolinguistics

1. Introduction

This talk will reflect on the challenge of answering questions relating to language and identity in corpora of computer-mediated communication (CMC) when we, as analysts, do not have access to reliable sociodemographic metadata. The talk reflects on an experiment, reported in Baker and Brookes (2022), which compared the affordances of two approaches to studying identity in CMC: (i.) using sociodemographic metadata; and (ii.) using language users' in-text attestations of their identities. To do this, we performed two sets of analyses, each one adopting either of the approaches noted above, in particular comparing the language used by male and female patients in a corpus of online patient feedback about cancer care services in the UK (14,403,694 tokens).

2. Data and approach(es)

Our methodology, then, comprised two approaches. For the first approach, we used the sociodemographic metadata available to us. Focussing on sex identity, we tagged the corpus and divided it into two sub-corpora, stored and analysed on *CQPweb* (Hardie 2012). One of comments in which patients checked a box to indicate that they identify as male, and another of comments in which patients checked a box to indicate that they identify as female (note that a small number of patients contributing to this corpus identified as 'Other', including non-binary. However, there was not enough data of this kind to facilitate the kind of analysis being undertaken in this study). For the purposes of this experiment, we refer to this approach as the 'metadata approach', as it relied on the sociodemographic metadata that our healthcare provider partners made available to us. Within our corpus, there were 97,774 comments from male patients (5,720,898 tokens) and 116,564 comments from female patients (8,683,079 tokens).

For the second approach, we operated under the artificial assumption that we did *not* have access to any sociodemographic metadata. For this analysis, we adopted an approach resembling one we were forced to adopt in previous work with similar data (Baker et al. 2019), and searched for cases where patients referenced their sex identity within the comments themselves. To exemplify, one patient prefaced their feedback with the phrase, 'As a 52 year-old man…'. On this basis, we determined the patient contributing this comment to identify as male. Again, we grouped the comments into two sub-corpora: one in which patients referred to themselves as male in their comments, and another in which patients referred to themselves as female. And as this approach relied on patients *referring* to their sex identity, we can refer to this approach as the 'self-reference approach'.

3. Findings

3.1. The metadata approach

We then compared the two sets of comments against each other using the keywords technique (statistic: log-likelihood with log ratio). This gave two sets of keywords – one for the male patients' comments compared against the female patients' comments, and one for the female patients' comments compared against the male patients' comments. We focused on the top 30 keywords from each set, ranked by log-likelihood score. This was an arbitrary cut-off but it did give a manageable number of keywords for analysis. These keywords are shown in Table 1 (for full table with statistical information, see Baker and Brookes 2022: 18-19).

class, *bladder*, *treatment*, *good*, *hospital*, *nhs*, *first*, *no*, *by*, *condition*, *test*, *carried*, *blood*, *thanks*, *kidney*, *gp*, *bowel*, *endoscopy*, *)*, *yes*, *quality*, *problem*, *attention*, *period*, *general*, *months*, *removal*, *myeloma*, *professionalism*, *successful*

Table 1: Keywords for male patients' comments versus female patients' comments.

We then analysed these keywords qualitatively with the broad aim of interpreting their uses in terms of recurrent rhetorical patterns and gendered discourses (Sunderland 2004). To do this, we went beyond concordance lines and examined the comments in their entirety.

Male patients were more likely to refer to their cancer and other diseases in their comments, evidenced through the keyness of words such as *bladder*, *bowel*, *kidney* and

myeloma in this data. The keywords also featured more general disease-related terms, such as *problem* and *condition*. Male patients also tended to focus on treatment processes, which evidenced in uses of keywords such as *removal, tests* and *endoscopy.* This tendency also accounts for the keyness of the constituents of the phrasal verb, *carried out*, as well as the word *by*, which tended to be used in passive constructions of medical processes.

Healthcare staff were also indexed by male patients through uses of keywords such as *NHS*, *General* and *Hospital*. These words could function metonymically, being used to denote all staff involved in a patient's care. Through such constructions, male patients could present their feedback as applying not just to a single staff member or team, but to an entire site of care or even the healthcare system as a whole. This could therefore represent a rhetorical strategy used by male patients in particular to generalise and present their complaints as being particularly pressing.

A characteristic theme of the male patients' comments is time, indicated in the keywords *months* and *period*. these tended to be used to quantify the amount of time that male patients had to *wait* for something, typically a diagnosis or an appointment for treatment. While the theme of waiting was frequent in both the male and female patients' comments, the male patients' comments provided more precise quantification of their waits.

The final group of keywords from male patients' comments are the words *no*, *yes* and *thanks*. And these keywords reflected the almost dialogic manner in which these patients in particular interacted with the voice of the feedback form, as in their comments they answered the prompt questions framing the feedback literally – with a *no* or a *yes* – and to express thanks for the quality of the service they received. This feature seems to be an effect of age as well as sex identity. Inspecting the frequencies of these keywords across the age groups, as well as between the sexes, we found that these words were all much more likely to be used by older patients, and by older *male* patients at every age group. Because these words are more common in men at all ages, this feature is likely an effect of the mixture of age and sex as factors.

i, *kind*, *felt*, *n't*, *amazing*, *feel*, *husband*, *she*, *so*, *lovely*, *oncologist*, *chemotherapy*, *me*, *they*, *had*, *radiotherapy*, *her*, *wonderful*, *did*, *you*, *nurse*, *unit*, *when*, *wait*, *supportive*, *lump*, *chemo*, *everyone*, *caring*, *busy*

Table 2: Keywords for female patients' comments versus male patients' comments.

Moving onto the female patients' keywords (Table 2; see also Baker and Brookes 2022: 27-28), and while the male patients' comments focused on procedural and transactional aspects of service, female patients tended to adopt a more personalised style, as reflected in the keyness of the pronouns *I* and *me*. This more gave rise to a more characteristic focus on how female patients' experiences made them *feel*. Staff were also evaluated using keywords such as *kind*, *lovely*, *supportive* and *caring*. They were also evaluated as *amazing* and *wonderful* and using the intensifier *so.* When we analysed 100 uses of each of these latter keywords, we again found that they tended to denote staff interpersonal skills.

Also key for female patients' comments were words indicating a stronger focus on individuals (e.g., *she*, *oncologist*, *her* and *nurse*), as well as words denoting relatives, units and smaller teams of staff. The keywords *chemotherapy*, *radiotherapy* and *chemo*, while ostensibly denoting types of treatment, tended instead to refer to teams of staff. Meanwhile, the keyword *everyone* could refer to staff working in teams or on wards, but at other points referred to other patients. In these cases, the female patients rendered their experiences as more generalisable, and this was also something we saw in uses of the general *you*.

A shared concern for male and female patients is the theme of waiting. When female patients described and evaluated waits, they did so in much less precise terms than male patients did. These patients specified the duration of waits in just 15 per cent of cases, which might be why words such as *months* and *period* are key for male patients' comments compared to female patients' ones.

3.2. The self-reference approach

The first step of this approach was to search for uses of the term 'man' and then the term 'woman'. We then extracted 100 comments in which patients self-identified as male and a hundred comments in which patients identified as female. We manually checked both samples to ensure that patients were indeed referencing their own sex identities, and not someone else's. For this analysis, we were forced to adopt a slightly different approach to obtaining keywords. We began by trying to compare the samples directly against each other, as we did in the metadata approach. However, this yielded a very small number of keywords, and these did not really tell us anything about gender-based patterns. This is likely a result of the small sample sizes that this approach forced us to work with (the maximum number of comments we could have analysed to have balance across male and female patients was 102). As a work-around, we generated keywords by comparing each of our samples against the rest of the comments in our corpus as a whole. And we might regard this reference corpus as a general corpus of cancer patient feedback. So these comparisons gave us two sets of keywords: one for the sample of male patients and one for our sample of female patients (show in in Tables 3 and 4, respectively; see also: Baker and Brookes 2022: 33-34).

man, *old*, *a*, *said*, *i*, *that*, *lucky*, *prostate*, *am*, *life*, *we*, *young*, *now*, *"*, *sick*

Table 3: Keywords for the sample of male patients' comments compared to the rest of the corpus.

age, *old*, *women*, *younger*, *hair*, *wig*, *intelligent*, *children*, *should*, *fertility*, *me*, *said*, *!*, *this*, *ovarian*, *be*, *that*

Table 4: Keywords for the sample of female patients' comments compared to the rest of the corpus.

Because we compared the samples against the same reference corpus, rather than against each other, we had some overlapping keywords, which could be viewed as indicating what is lexically characteristic of feedback in which patients declare their sex identities compared to

feedback more generally. A drawback of this approach is that the differences between the keywords here are not statistically significant between our two samples. However, an advantage of the approach is that it does at least let us look at *similarities* between the two samples, by looking at the overlapping keywords. We then undertook a close analysis of these keywords, proceeding in the same way as we did for the metadata approach.

A striking similarity between both samples is the keyness of the quotative *said*. The fact that this is key suggests that patients in both samples quote others in their comments more often than we might expect in feedback on cancer care in general. This also helps to explain the keyness of the word *that*, which tended to be used to frame quotations. The use of quotations seems to emerge as especially frequent in these samples because the patients' sex identity is often mentioned in the quoted speech. The use of quotes is linked to negative feedback in particular, as patients tended to use quotes when recounting cases in which they were given advice that they viewed as inconsistent or inaccurate, or cases in which they experienced staff rudeness.

Another overlapping feature across both the male and female samples was the use of keywords relating to age. For the male patients, this includes the words *old* and *young*, and for the women's comments we get *age*, *old* and *younger*. Both male and female patients frequently referenced their age in conjunction with their sex for evaluative purposes. For example, both male and female patients referenced their age in order to construct themselves as having particular healthcare requirements. Sometimes these requirements were met and sometimes they were not, and this could determine whether the feedback given was broadly positive or negative. In some cases, the negative evaluation targeted gendered stereotypes that patients attributed to healthcare staff. For example, one male patient complained about being treated like a 'grumpy old man', while a female patient complained about being treated like a 'silly old woman'. Both male and female patients drew on the intersection of age and sex, then, to frame descriptions of experiences in which they felt belittled by staff members.

As well as older age, both the male and female patients in our samples also referenced youth. Some of the male patients used the keyword *young* to construct themselves as socially and sexually active, with these aspects of their identity being linked to both their age and sex. And so this was again about constructing particular healthcare needs, and whether or not these were met could again motivate positive or negative feedback. Where the adjective *young* was key in the male comment sample, the comparative form *younger* was key for the female sample. Female patients tended to use the keyword *younger* to refer either to younger female patients in general, or to hypothetical others. Such comments typically described how particular aspects of service provision would not be suitable for younger female patients, and often made recommendations about how services could be improved for younger women in the future. This pattern, of the female patients issuing recommendations, also helps to account for the keyness of *should* in the female patient sample.

Male patients, on the other hand, frequently produced a discourse of exceptionalism. This was realised, for example, in the keyword *lucky*, which male patients tended to use to describe themselves as being lucky for having been treated by a highly skilled practitioner or team. In other contexts, *lucky* is used by male patients when relaying interactions with staff in which they'd been told that they're *lucky* to be alive. In either case, male commenters imply that their experiences are somehow exceptional or even unique, either in terms of the high standards of care they received, or the severity of their illness. Cases of the latter also help to account for the keyness of *sick* in the male patient sample, as some of the men described how staff informed them that they were 'very *sick* men'.

Another keyword which indicates the male patients' focus on their own experiences is the temporal adverb *now*. While female patients frequently made recommendations as to how services could be improved for others in future, male patients tended to focus instead on the past, in addition to the present. These descriptions of the past took on an almost autobiographical tone, as the male patients often recounted their previous experiences with a provider, and described the different forms of treatment that had brought them to the present – i.e. to the *now*. Thus, male patients used *now* in order to draw comparisons between their current experiences and previous ones. A similar tendancy is observable for uses of the keyword *life*, with male patients either thanking staff for 'saving' or 'improving' their *life*, or evaluating an experience as being the 'worst of [their] *life*'.

4. Conclusions

The metadata had the advantage of allowing us to base our findings on a much larger dataset. This not only meant that we could have greater confidence in the trends we identified, but it also allowed us to perform direct statistical comparisons of our sex-based subsets using the keywords technique. Another advantage of this approach was that we were able to draw on other metadata tags to interpret some of the patterns we found. For instance, our interpretation of the finding that male patients engaged with the feedback form in a more dialogic way, was enriched by our ability to look at age-related metadata too, where we could see that this was a feature of older male patients in particular. This supplementary perspective was only possible because we could draw on this extra sociodemographic information. Without it, we would not have been able to arrive at that interpretation.

Yet the metadata approach also had some shortcomings. While having a vast corpus tagged for sociodemographic information brings lots of clear advantages, *assembling* such a corpus – and with all of this metadata – remains a demanding (and resource-intensive) task. We were helped in this project by our collaboration with NHS England, as our contacts there collected the metadata from patients, in an ethically appropriate manner, at the point at which the feedback was given. They then provided that metadata to us in a format whereby it was relatively straightforward for us to convert it into a series of searchable tags. Without their support, this would have been a much more resource-intensive exercise.

A criticism of sociodemographic annotations is that they often depend on quite broad social categories. In this work, we were forced to work with the categories of 'male' and 'female'. But these broad categories could result in us taking a top-down and overly simplistic view of identity. While these categories might be suitably broad to be operationalizable in a large-scale corpus analysis, they also risk obscuring more nuanced types of identity relations. In other words, what is gained from broad categories in terms of scalability and practicality, might be lost in terms of granularity and contextual nuance.

Relatedly, we should also reflect here on a more general criticism that is often made of studies which correlate social categories with language use. Statistically significant correlations between a social attribute and the use of a linguistic feature are often interpreted as relationships of causation. In other words, if we find that use of a particular word or feature correlates with language users being male, we might be tempted to conclude that this trend occurs *because* those language users are men. However, the marked use of a linguistic feature can be related not just to the particular variable under focus, but to some other aspect of identity, or even a combination of these. It is also important to bear in mind that no set of sociodemographic annotations is ever complete. In our case, there were numerous other aspects of their identities that patients *could have* been asked about but were not. With the kind of data we were working with, then, which was elicited through a survey that we, ourselves, did not design, we were restricted by what the survey creators decided to ask about, either because they thought it was important, or because it was easy to measure and categorise.

Turning to the self-reference approach, an advantage of this was that we could have greater confidence that patients' sex identities were more directly relevant to their comments. We knew this because the patients explicitly oriented to these aspects of their identities in the comments themselves. In this way, this approach gave us something of an interpretive warrant, which meant that we could be more confident that the differences we were observing were indeed related to patients' self-attested sex identities. Another advantage of this approach was that it gave us an arguably more organic route into looking at intersectionality, as patients' orientations to one identity category frequently accompanied, or gave rise to, another. For example, we found that patients who referenced aspects of their sex identity were also particularly likely to reference their age too. These intersectional aspects of their identities were highlighted because patients perceived them as relevant to their experiences, and so to their feedback too.

This approach also had some limitations, too. The first concerns the size of the samples that the approach allowed us to work with. Because patients referred to their identities relatively *in*frequently in their comments, we were forced to work with very small samples. This posed several methodological challenges, and for example meant that we did not have sufficient data to perform a direct keyword comparison between the samples. Our small sample size also reduced the generalisability of our findings. Relatedly, comments in which patients went on-record about their identity in their feedback may not be considered to be representative of all the comments in our corpus as a whole. That is, when patients went on-record about their identities in their comments, they often did so because they perceived these qualities to be central somehow to the type of feedback they were giving, and this was not the norm. As such, this approach might train our focus on certain types of comments which are not necessarily representative of the wider corpus (nor, indeed, the wider context of language that the corpus is intended to represent). If adopting this kind of approach, then, some caution is likely to be needed regarding making generalisations.

Finally, just like sociodemographic metadata can never capture *all* identity variables, we should also be mindful with our self-reference approach that, just because a language user doesn't mention an aspect of their identity explicitly, that doesn't mean that that aspect of their identity is not in fact relevant to their language use in a given context.

A pertinent question we might ask at this point, regards which approach might be best for researchers studying identity in a large corpus of texts. And of course, the answer is likely to depend on the levels of granularity and accuracy required, as well as, on a more practical note, the type of data we are working with (and what its limitations are).

Even if we do have access to reliable sociodemographic metadata, any approach to studying language use and identity (in computer-mediated communication or any other context) will nevertheless stand to benefit from our bringing in qualitative, bottom-up methods of analysis. In this vein, we could combine both of the approaches presented here. Such an analysis could start by looking closely at a sample of texts in the corpus and noting emergent identity categories. We could then use that analysis as means of narrowing our focus to those emergent categories when presented with (a potentially overwhelming range of) sociodemographic tags. This kind of bottom-up approach has the advantage of directing our analytical focus to those aspects of identity that the language users in our corpus *themselves* perceive to be contextually relevant. The reference-based approach that would precede any annotation-based analysis could also help us to account for more subtle or even implied forms of identity self-referencing. And this, in turn, could not only give our analysis focus, but also help us to guarding against an uncritical overreliance on correlational statistics.

5. References

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

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

Hardie, A. (2012). CQPweb: Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380–409.

Sunderland, J. (2004). *Gendered Discourses*. Basingstoke: Palgrave Macmillan.