2020. *Proc Ling Soc Amer* 5(1). 499–513. https://doi.org/10.3765/plsa.v5i1.4728.

How we got here: Short-scale change in identity labels for trans, cis, and non-binary people in the 2000s

Lal Zimman (he/him) & Will Hayworth (they/them)*

Abstract. Though understudied in research on language variation and change, the lexicon is a crucial domain for sociopolitical transformations of language. This paper presents a corpus-based sociolinguistic analysis of changes in terms for transgender, cisgender, and non-binary individuals in four online communities on the social media blogging site, LiveJournal.com – one for trans women, one for trans men, one for non-binary people, and another for transgender people in general – that were popular in the 2000s. Using innovative corpus methods that utilize general purpose cloud computing tools, we focus on changes in the popularity of labels for trans, cis, and non-binary people, the factors that impact the variable use of these terms, and what kinds of differences can be observed across the four LiveJournal communities of practice studied. It thereby contributes both to the study of language and identity in trans and queer communities and to the development of methods for studying large datasets of technologically-mediated communication.

Keywords. sociolinguistics; corpus linguistics; language, gender & sexuality; transgender language; lexical change; social media.

1. Introduction: Why is the internet so trans? In a number of ways, the internet is a highly trans modality. This may be articulated explicitly in the form of warnings that online spaces are full of men pretending to be women, or more generally, through the sense that a person's "true" identity cannot be known from digitally-mediated interactions. At other times, the connection between trans people and the internet comes from knowledge that certain digital platforms, such as erstwhile social media giant Tumblr, have provided opportunities for trans discourses to flourish through the establishment of densely populated online trans communities.

There are a number of benefits online spaces offer trans people, including the freedom to take on identities that differ from those occupied in offline contexts, to do so without the cultural baggage attached to our fleshy selves, and to connect trans people who are otherwise socially or geographically isolated. The link between trans people and technologies of interaction is far from a new one: before there was Tumblr, there was LiveJournal (Zimman & Hayworth 2020), and before LiveJournal, there was USENET (Dame 2017). Going back much further, there was trailblazer Virginia Prince's mail-based newsletter for trans people in the mid-to-late 20th century (Stryker 2008). Given the centrality of communicative relationships between trans people in the discovery and articulation of trans identities, online spaces have clearly contributed to the current level and types of awareness of trans people's experiences. Part of recent cultural shifts in how trans people are seen and treated has been the establishment of norms surrounding language use. While we know a bit about trans language in online spaces (Zimman 2014; Dame 2018) and the history of trans communities online (e.g., Giardina 2019), we know little about the history of trans language.

* Thanks to Google for providing credits for this project.
Authors: Lal Zimman, UC Santa Barbara (zimman@ucsb.edu) & Will Hayworth, Google (wsh@google.com).

2. The place of the lexicon in sociolinguistics and in language, gender & sexuality studies. In the study of language variation and change, the lexicon has often been marginalized in relation to other levels of language, particularly phonetics/phonology and morphosyntax. This is in part due to the fact that the lexicon exists above the "level of awareness" (Silverstein 1981) and is therefore more likely to be subject to conscious, intentional intervention rather than reflecting unconscious patterns of stratification and change. By contrast, the lexicon has always been central in the field of language, gender, and sexuality (e.g. Lakoff 1973) precisely because of the way lexical items can be deployed, examined, evaluated, and reconstituted to better suit speakers' political goals. Identity labels have been especially important in queer linguistics and the burgeoning area of trans linguistics (Zimman forthcoming), in which agency, self-definition, social change, resignification, and self-definition have often resulted in a centering of the lexicon (e.g., Chen 1998; McConnell-Ginet 2001; Wong 2005; Hazenberg 2017).

The study presented here considers changes in the distribution of identity terms in trans communities on LiveJournal, a social media platform whose popularity during the 2000s preceded the rise of current social media giants. In so doing, we make two major contributions to the study of the queer/trans lexicon, which previous authors have typically investigated qualitatively using relatively small datasets. First, we augment previous work by using a large corpus of data to investigate quantitative trends over time, offering a broader context for smaller-scale studies. In this respect, we follow corpus sociolinguists such as Paul Baker in using corpus and computational methods to consider queer linguistic questions (e.g., Baker 2003, 2004). Second, our dataset allows us to analyze change over time in digitally mediated interactions by focusing on a social media platform that is no longer popular, at least among English speakers.[1] Where social media data has been used in the analysis of trans and queer identities, the source has typically been synchronic data from currently popular platforms such as YouTube, Twitter, Tumblr, and Reddit, all of which offer APIs to facilitate large-scale data collection. LiveJournal has an API, but its ability to retrieve large numbers of posts, especially older ones, is limited. Maximizing the coverage of the corpus requires crawling the HTML version of the site to discover content and parse it. Because the HTML is formatted for display in a browser rather than for programmatic manipulation, parsing it to extract corpus content and metadata is non-trivial. We therefore discuss our methods for constructing and querying the **trans-livecorpus** analyzed below (see also Zimman & Hayworth 2020).

The goal of this paper is thus to provide historical background for current identity labels for trans, cis, and non-binary people through an analysis of the distribution of such terms in LiveJournal communities for trans people in the 2000s. We expand on the findings of Zimman and Hayworth (2020), in which we report on trends in FTM, the LiveJournal community for trans men and other transmasculine people. The analysis below compares these findings to three other communities as well as exploring additional terms for cis people not discussed by Zimman and Hayworth (2020). Like that piece, this paper is also a demonstration of using cloud computing tools to create and query corpora from semi-structured social media data.

3. Data & methods. The data analyzed in this study come from a corpus constructed by the authors from four communities on LiveJournal. LiveJournal offers a number of benefits for our study. First, it is a highly interactional blogging platform in which users can join an unlimited number of communities that may be defined around particular identities, practices, or interests. Second, it provides some short-scale historical background for the trends around naming that are

[1] LiveJournal is currently owned by a Russian company, Rambler Media Group.

happening online today, as LiveJournal happened to be a popular platform for trans people to build community online in the 2000s, before sites like Facebook and Tumblr existed. Finally, users can select whether posts are visible to anyone with an internet connection or whether they can only be viewed by other community members, which allowed us to avoid collecting sensitive information that users did not want shared beyond the community's membership.

3.1. CORPUS CENSUS. The trans-livecorpus currently consists of four LiveJournal communities, each of which was created for trans people to exchange information and interact with one another: FTM (for trans men and transmasculine people), MTF (for trans women and transfeminine people),[2] TRANSGENDER (for trans people generally), and GENDERQUEER (for people who identify or present outside of the gender binary).[3] It is worth noting, however, that none of these communities was exclusive, and that many individuals joined despite not identifying with the relevant gender category, whether as partners, friends, family members, allies, and the occasional troll.

Community	# of posts	# of comments	# of words	# of tokens
FTM	19,643	207,579	17,034,982 (72%)	63,102 (63%)
TRANSGENDER	5,930	34,498	3,586,914 (15%)	20,888 (21%)
GENDERQUEER	3,167	18,560	1,833,638 (8%)	11,166 (11%)
MTF	1,800	13,310	1,165,400 (5%)	4,635 (5%)
Totals	**30,540**	**273,947**	**23,620,934**	**99,791**

Table 1. Number of posts, comments and words in the corpus and tokens from each community

Table 1 shows the number of posts, comments, and words included in the corpus as well as the number of tokens extracted from each community and included for analysis here. FTM is by far the most popular group, constituting about 72% of the corpus and 63% of the tokens analyzed. TRANSGENDER is next, as 15% of the corpus and 21% of the tokens analyzed. The other two communities had less traffic, at least in the form of public posts, with GENDERQUEER accounting for 8% of the corpus and 11% of the tokens and MTF providing 5% of the corpus and 5% of the tokens. The analysis presented below is based on 99,791 tokens of words that act as labels for trans, cis, and non-binary identities.

3.2. BUILDING & SEARCHING LIVECORPUS: A CLOUD-NATIVE ARCHITECTURE. We built livecorpus after looking for existing LiveJournal parsers and finding nothing complete or usable. LiveJournal's HTML is formatted for display in a browser, not for machine readability, so the raw data needed to be transformed significantly to make analysis tractable. That processing required writing code and finding a place to run it. We chose to do so in the cloud, which allows users to reserve and release computing resources easily, quickly and cheaply. We could have rented our own servers, but starting with the cloud presented an opportunity to go "cloud native" and take

[2] FTM (female-to-male) and MTF (male-to-female) were at one point used as umbrella terms that today would be referred to as transfeminine and transmasculine or assigned male at birth and assigned female at birth.
[3] The names of LiveJournal communities discussed here appear in caps throughout the paper in order to distinguish references to those groups from references to the words *transgender* and *genderqueer*.

advantage of pre-existing managed services. These services also gave us discrete, tested components to build upon.

The livecorpus crawler is written in Python 3 and runs on Google App Engine, which makes it easy to run code without much setup and scales up and down automatically. The crawler app, visualized in Figure 1, fetches HTML pages from LiveJournal's servers and parses them using BeautifulSoup. For pages that contain links to entries, the crawler extracts the links and enqueues them for further processing using Cloud Tasks. The Cloud Tasks push queue dispatches links to the crawler for parsing, ensuring that each entry is fetched and rate limiting our crawl to comply with LiveJournal's bot policy (no more than 5 connections per second). The parsed entries are stored in Cloud Firestore, an auto-scaling document database which doesn't require a schema to be defined before usage, making it easy for us to progressively save more of the crawled data. Cloud Firestore also uses a hierarchical data model (collections have documents which have subcollections), which matches LiveJournal's data structure (community → entries → comments).

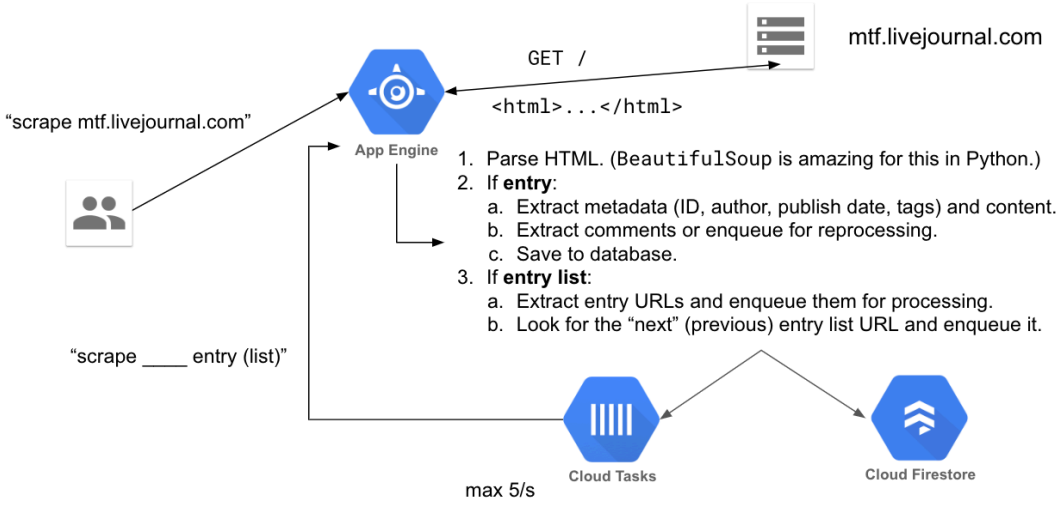

Figure 1. The **livecorpus** crawling pipeline

Analyzing the data efficiently requires a database with different properties. On Google Cloud, that's BigQuery, an analytical data warehouse that can hold petabytes and query them quickly on demand. Cloud Firestore's managed export produces files that BigQuery can load into its native format. We query BigQuery using SQL and get back rows of results that can be converted to CSVs or Google Sheets. To look for particular terms, we used regular expressions, which are well known in corpus linguistics. Figure 2 contains an example that finds all tokens of *trans* with an optional * suffix, separated by word boundaries.

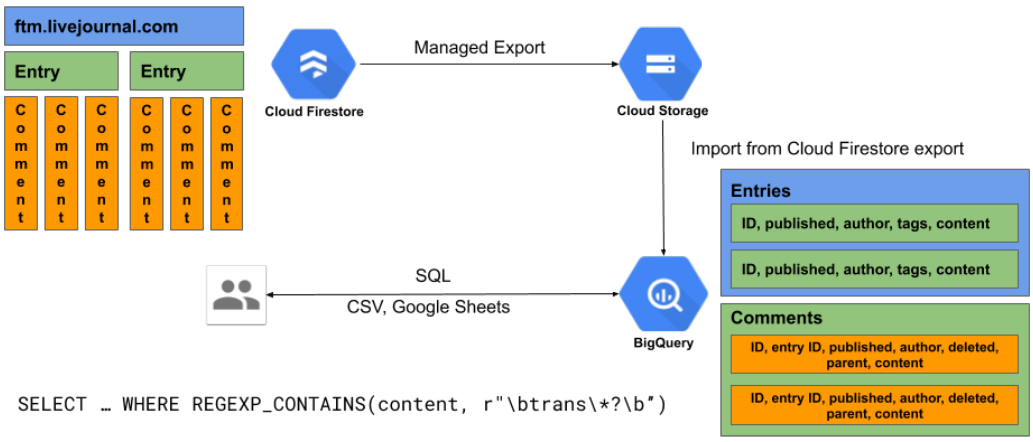

Figure 2. Searching livecorpus

3.3. LIVECORPUS: A "SERVERLESS", "GENERAL" METHOD. This method is unusual because it's "serverless", which describes an emerging style of cloud application architecture. In particular, none of the resources livecorpus uses requires explicit provisioning. We can turn the application on or off, but we don't have to do resource planning or system administration. This high level approach means that Google is responsible for sharing resources on our behalf; we pay only for the small amount of computing resources we actually use. A downside of this implementation is that it's not trivial to run elsewhere; it's optimized for Google's infrastructure. The architecture is portable, however, and could be extended to other cloud platforms.

livecorpus is also built entirely on "general purpose" computing tools. Python is a ubiquitous and powerful language, and the components we built our system on are used across industries and different kinds of applications, from mobile games to scientific research. This broader user base means that it's easier to get help with them. The cloud components scale seamlessly and automatically, so your corpus can grow from megabytes to petabytes without changing much. And because one provider (Google, in this case) is producing these tools to work as complements, they play well together.

3.4. CODING & STATISTICAL ANALYSIS. In order to explore whether the use of identity labels in trans-livecorpus changed over time, the data were coded for a number of factors, many of which were generated by the crawler. Our coding for each token thus far includes 1) whether the token appears in a post or a comment; 2) which post or comment the token occurs in; 3) the username of the author; 4) the date of the post or comment in which the token occurs, binned into months; 5) the lexical category or lexeme; and 6) the gender category of the referent (i.e., as trans, cis, or non-binary). Manual coding of additional factors is ongoing, including whether the author is referring to themself, someone of the same identity, or someone of a different identity; whether they are speaking as themself or voicing another; and the presence of metalinguistic stance.

The statistical analyses we report below are from linear mixed effects regressions with fixed effects of the month in which the token occurred and its interaction with the community from which it was collected. The random effects were the total number of posts that month, to account for changes in community traffic over time, and the author (by username).

4. Analysis #1: Terms for trans people. Our first analysis considers identity terms for trans people. We begin with a summary of our previous findings on the FTM community and then discuss differences from these patterns in the other three communities in trans-livecorpus.

4.1. PATTERNS IN FTM FOR REFERRING TO TRANS PEOPLE. Zimman and Hayworth (2020) report on the use of terms of transgender people in the FTM community, including the popular full forms *transgender*, *transgendered*, and *transsexual*, as well as the short forms *trans* or *trans* + [some class or group of persons] (e.g. *trans woman*, *transperson*, *trans-boy*, etc.) and a few less commonly used terms including *transgenderist*,[4] *transmasculine* and *transfeminine*. In FTM, the most popular options are the short forms: *trans* as a stand-alone descriptor and *trans* + [group] together account for 78% of the tokens of trans identity labels in this community. Zimman and Hayworth also document some significant changes over time, including a decrease in the use of the long forms (*transender[ed]*, *transsexual*) and increase in the short forms, as well as an increasing preference for *transgender* over *transgendered* where the long forms are used. We also find a decrease in the popularity of *transsexual*. Zimman and Hayworth further note the appearance of *transmasculine* in the FTM community, starting in 2006. However, the term remained uncommon at that time.

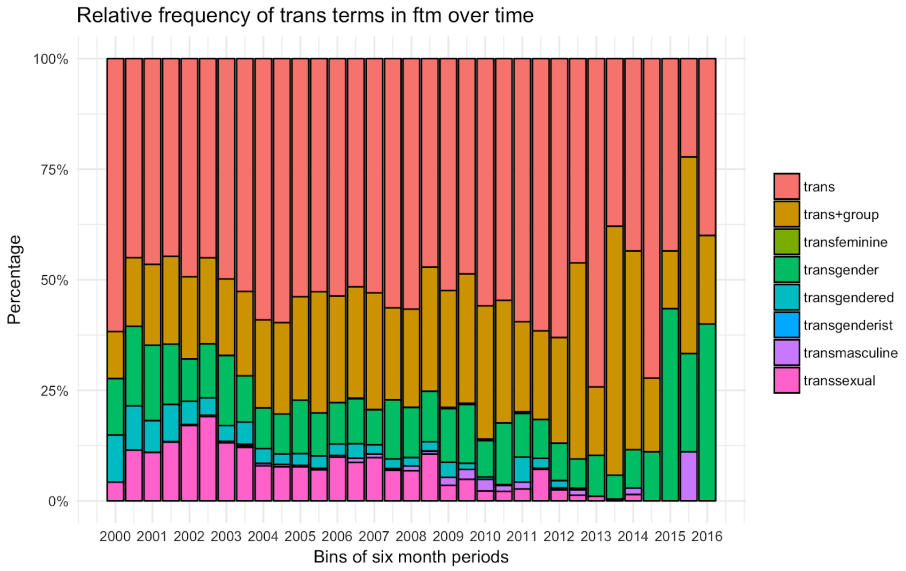

Figure 3. The relative popularity of terms for trans people in the FTM community

Figure 3 shows the relative frequency of each word for trans people we analyzed as a percentage of the total number of tokens in the dataset within the same six month period. This make the patterns of change more visible while accounting for the overall drop-off in traffic on LiveJournal starting in 2007. However, this choice also produces some unusual patterns in the left- and rightmost bars because they are based on a much smaller number of tokens (e.g. the rightmost bars imply that *transgender* enjoyed a resurgence in 2015-16, but this is likely due to the very small number of tokens during from this period.

Overall, Figure 3 shows the popularity of the short forms, *trans* (in red) and *trans + [group]* (in mustard). It also shows how *transgendered* (in light blue) was somewhat well represented in the beginning of the community's existence before decreasing until it is no longer visible after 2012. This mirrors metalinguistic commentary in this and many other trans communities that problematizes the *-ed* ending in *transgendered* (see Zimman & Hayworth 2020).

[4] *Transgenderist* is a term that was popular in the 1990s and before to refer to a transgender person.

Finally, you can see a decrease in the popularity of *transsexual* (in pink), which similarly reflects community associations between this term and pathologizing medical models of trans identity and compulsory gender conformity (Zimman forthcoming). With the situation in FTM in mind, we can now turn to trends in other trans communities on LiveJournal.

4.2. PATTERNS IN OTHER COMMUNITIES FOR REFERRING TO TRANS PEOPLE. The patterns found in FTM are largely the same as those in TRANSGENDER, which is the other high traffic community in the corpus. Figure 4 shows the pattern for TRANSGENDER, where we can again see that the most popular terms are the short forms (in red and mustard), that a decrease in the use of *transgendered* occurred, and that *transsexual* also declined in popularity. Here too, the word *transmasculine* makes an appearance at around the same time of 2008, but *transfeminine* is not common enough to be visible on the plot. This is an interesting asymmetry given that this community was for trans people of all sorts.

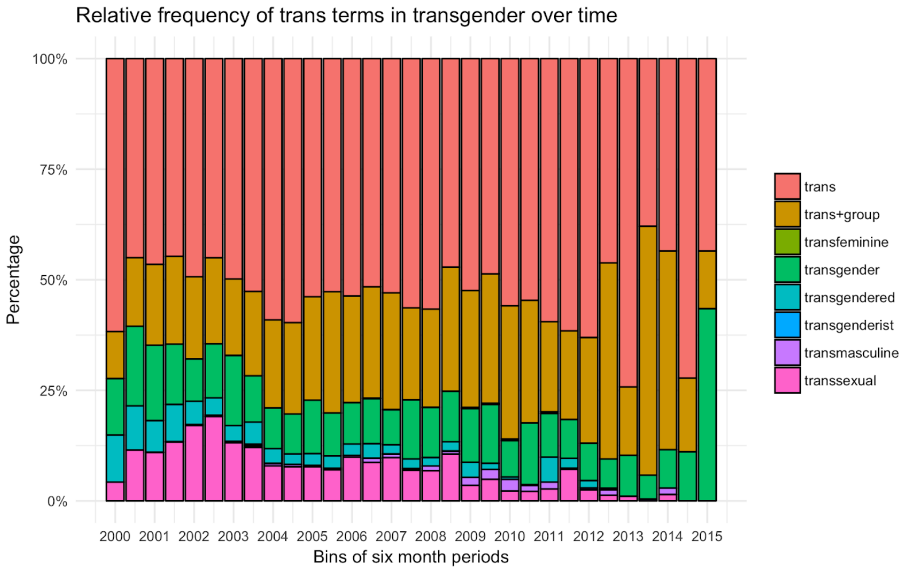

Figure 4. The relative popularity of terms for trans people in the TRANSGENDER community

Figure 5 shows the distribution of terms for trans people in MTF, which is lower in traffic than either FTM or TRANSGENDER. Despite the smaller number tokens, the distribution is not dissimilar to the results for either of the high traffic communities. However, the changes over time in MTF are at times less visually dramatic than that occurring in FTM and TRANSGENDER. Specifically, there is little if any apparent increase in the use of short forms over long forms. Additionally, there is no clear decrease in the use of the long terms problematized in other communities, including *transgendered* and *transsexual*.

Despite having approximately twice as many tokens as MTF, the GENDERQUEER community's usage (Figure 6) is somewhat less uniform and thus harder to interpret than the other groups' due to less consistency across adjacent six month periods. Furthermore, some terms seem to fall out of use in GENDERQUEER – for instance, *transsexual* and perhaps also *transgendered* – only to resurge later on. These patterns highlight the need for more nuanced coding, particularly given that genderqueer people are less likely to be using the word *transsexual* as a term of self-identification, and hence may serve these community members as a category of **dis**identification, which could remain relevant even if binary-identified trans people are using

the term less often. Take note that for GENDERQUEER, like other groups, the last three bars are based on only a few tokens.

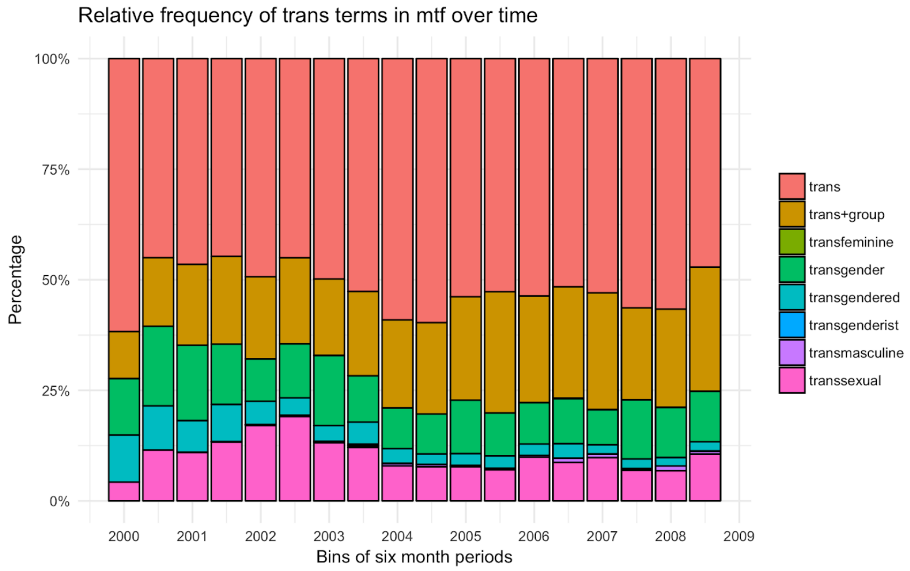

Figure 5. The relative popularity of terms for trans people in the MTF community

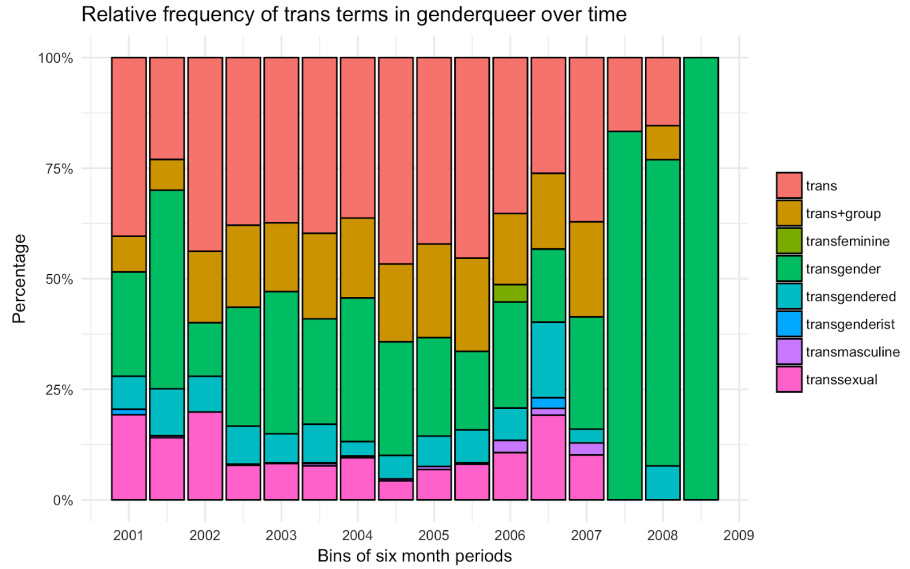

Figure 6. The relative popularity of terms for trans people in the GENDERQUEER community

4.3. THE DECLINE OF THE -ED ENDING. Another specific finding in Zimman and Hayworth (2020) that we revisit here is the status of the *-ed* ending in the word *transgendered* within the FTM community. As Figure 7 shows, in cases where someone used either *transgender* (which we call the plain form) or *transgendered* (the *-ed* form) there is a strong trend away from the *-ed* ending over time. By the end of the life of the FTM community, there were virtually no tokens of *transgendered*.

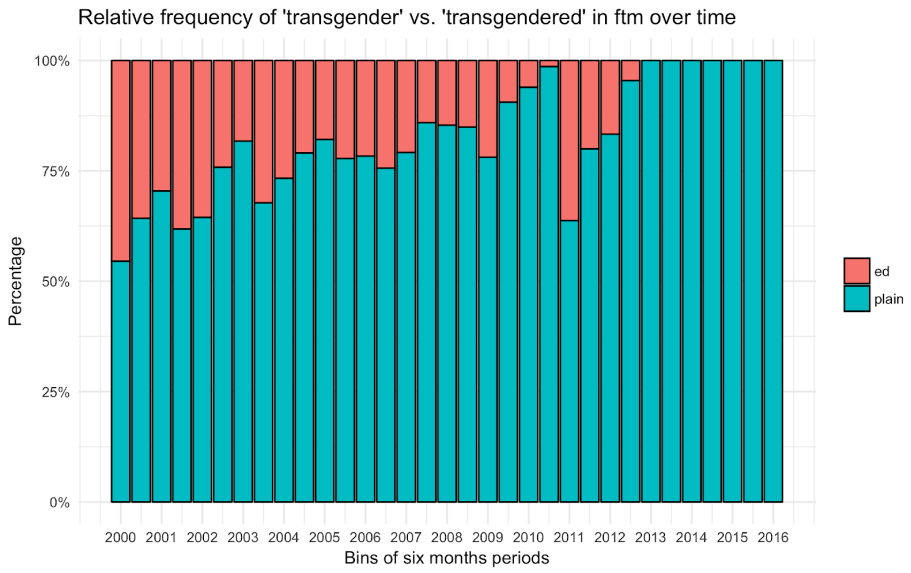

Figure 7. The relative popularity of *transgender* versus *transgendered* in the FTM community

The linear mixed effects regression modeling described above shows an overall decrease in use for the *-ed* form compared to the plain form ($B = -3.65$, $p < 0.001$). However, we also found an interaction between the month of usage and the community, such that the overall downward trend held only for the two most populous communities, FTM (Figure 7) and TRANSGENDER (Figure 8). The statistical results are presented in Table 2.

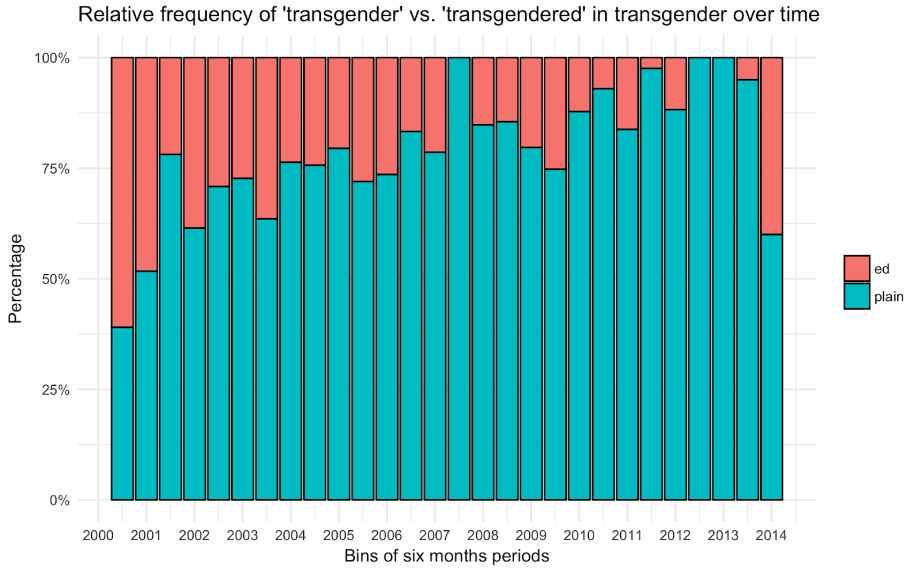

Figure 8. The relative popularity of *transgender* versus *transgendered* in the TRANSGENDER community

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	81.665	1.7969	178	45.445	< 0.001 ***
edendinged	-3.6513	0.4217	15406	-8.659	< 0.001 ***
edendinged:communitygenderqueer	0.8739	0.8531	15346	1.024	0.30563
edendinged:communitymtf	2.1750	1.0151	15070	2.143	< 0.033 *
edendinged:communitytransgender	-1.8588	0.6065	15250	-3.065	< 0.01 **

Table 2. Fixed effects from linear mixed effects regression: The decline of *transgendered*

There was no change in the use of *transgender* versus *transgendered* in the GENDERQUEER group, while the MTF community showed the opposite direction in which the *-ed* ending became more popular over time (B = 2.175, p < 0.05), though the effect size is smaller for this finding than the overall trend away from *-ed* (see Figure 9). Further coding and qualitative analysis will be important to understand why these differences exist.

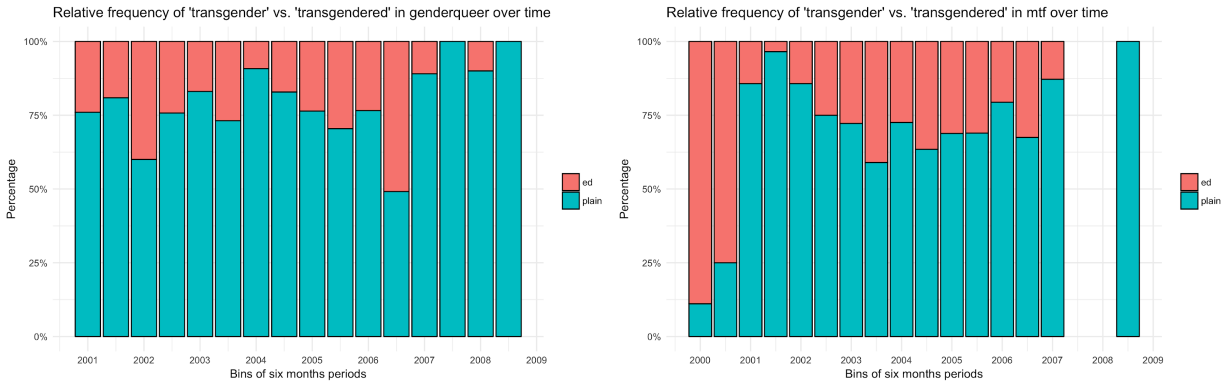

Figure 9. The relative popularity of *transgender* versus *transgendered* in GENDERQUEER (left) and MTF (right)

Although space limitations prevent us from discussing this trend in depth, we reached similar findings in our analysis of the long versus short forms. In the dataset as a whole, the long forms (*transgender[ed]*, *transsexual*) decreased in use in favor of the short forms trans and trans + [group]. In the dataset as a whole, the short forms increased in use (B = 1.471, p <0.001), but interactions showed that the changes were significant only for the FTM and TRANSGENDER communities, whereas no significant change was found for MTF and GENDERQUEER. Again, it is possible that the larger number of tokens in the former communities made for a more robust trend than the smaller number of tokens in the latter communities.

5. Analysis #2: Terms for cis people. Next we want to talk about words for cis (i.e., non-trans) people. Zimman and Hayworth (2020) examined the emergence of *cis(gender)* as a way to mark this unmarked category. Our current analysis expands on this by adding two additional ways of referring to cis people that were once popular in trans communities but is now seen as highly problematic, which are the modifiers *bio(logical)* and *genetic* in phrases like *biological female* or

genetic male (see Zimman 2014 for more on the problematization of these terms). As Zimman and Hayworth report, *cis* did not appear in the FTM community until 2003 and did not become common until a few years later.

Figures 10 and 11 show how *cis* came to replace *bio(logical)* and *genetic* as ways of referring to non-trans people within the FTM and TRANSGENDER communities.

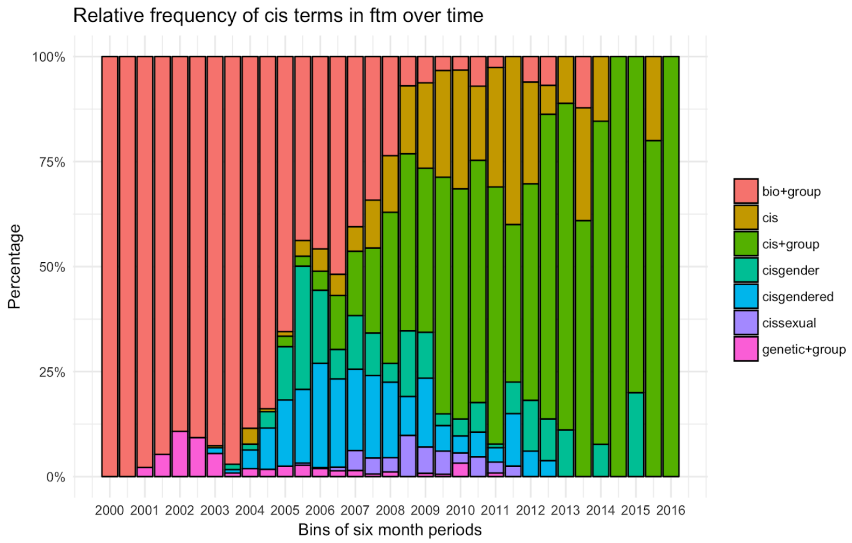

Figure 10. The relative popularity of terms for cis people in the FTM community

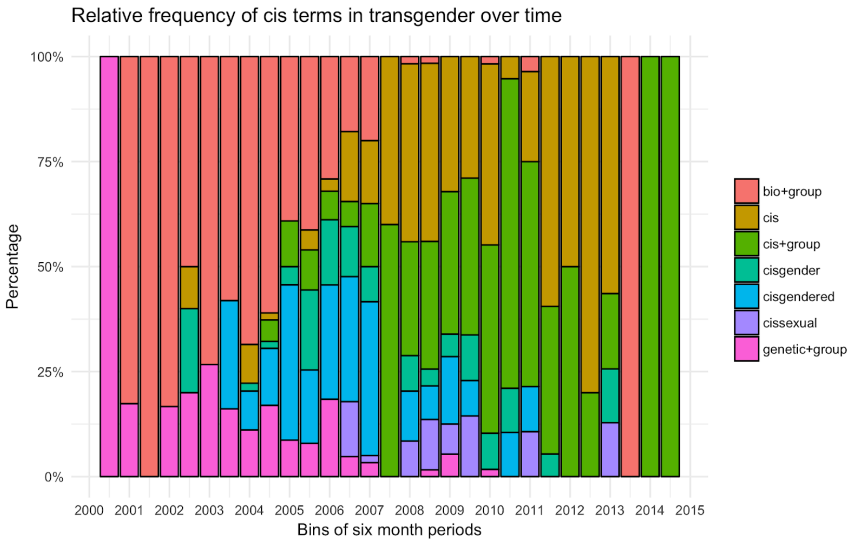

Figure 11. The relative popularity of terms for cis people in the TRANSGENDER community

In both FTM and TRANSGENDER, the earliest terms documented for referring to non-trans people are *bio* and *genetic*. The latter term, however, was found far more often in the TRANSGENDER group than in FTM, which is likely related to the set phrase *genetic girl* (or "GG") which had a long history of use in communities of trans women. However, starting in the mid-2000s, those terms began to shrink in relative usage in both communities as *cis*, *cis* + [group], *cisgender(ed)*, and *cissexual* took their place. Unfortunately, it is more difficult to get a sense of the way *cis* emerged in MTF and GENDERQUEER because of the dramatic decrease in

traffic in those groups right as *cis* was taking hold in the other communities. By 2008, *cis* was the norm and older forms had become regarded as highly problematic within these communities.

6. Analysis #3: Terms for non-binary people. Our final set of questions concerns how *non-binary* came to replace other words such as *genderqueer*, the previously most common umbrella label for individuals who do not identify as either strictly female or strictly male, and the distribution of other terms that refer to those outside the binary, including *agender*, *bigender*, *polygender*, *genderfuck*, *genderfluid*, and *non-binary*.

6.1. PATTERNS IN FTM FOR REFERRING TO NON-BINARY PEOPLE. The distribution of terms for non-binary people is quite different from the plots presented above because the overwhelming majority of tokens (92%) referring to non-binary people are the word *genderqueer*. This demonstrates how firmly established *genderqueer* was as an umbrella label before it was replaced by *non-binary*. Today, *genderqueer* is still a common term of self-identification, but it is now typically used to index a specifically queer gender identity or presentation, and is seen as falling under the non-binary umbrella. Figure 12 shows the terms for non-binary people in FTM with tokens of *genderqueer* removed (n = 6,437) so that other terms (n = 515) can be seen.

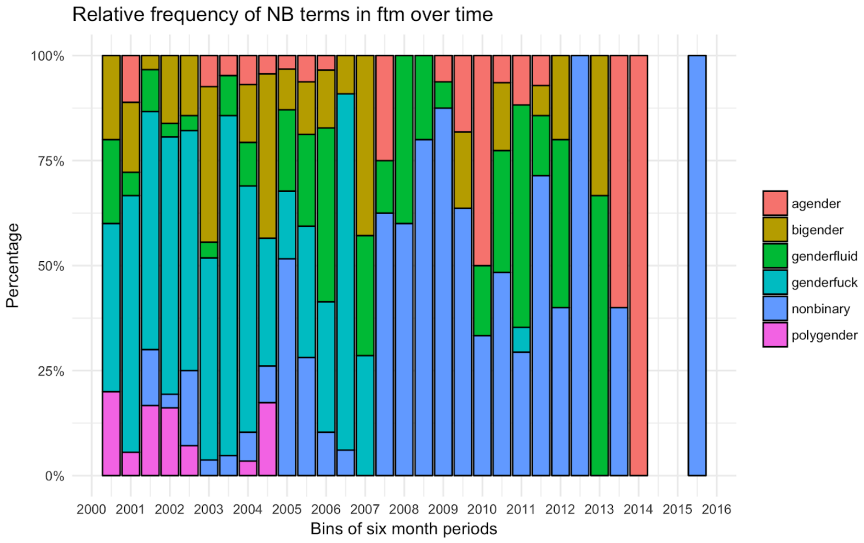

Figure 12: The relative popularity of terms for non-binary people in the FTM community

Figure 12 shows that, in FTM, the most common words after *genderqueer* were *genderfuck* (37.1% of tokens, in light blue), which was most popular prior to around 2007, and *non-binary* (24.5%, in dark blue), which became the most common term after around 2007. *Genderfluid* (14.6%, in green) and *bigender* (13.2%, in mustard) were relatively well represented as well. *Agender* was only 6.8% of tokens (in red), but it seems to be more common in the later years of the community, while *polygender* was the least popular term (3.9%, in pink) and showed up only in the first few years of the community's life.

6.2. PATTERNS IN OTHER COMMUNITIES FOR REFERRING TO NON-BINARY PEOPLE. Turning to the other communities for comparison, we focus on the obvious question of how people in the GENDERQUEER community use these terms (total number of tokens = 3,696).
One major shortcoming of comparing the data from GENDERQUEER to other communities is that the former ends in 2008. However, the data that is available from GENDERQUEER looks a

lot like the pre-2008 data in FTM, as Figure 13 shows. After removing *genderqueer* (87% of tokens), the most common word is *genderfuck* (37% of remaining tokens), followed by *bigender* (22%), *genderfluid* (13%), *polygender* (12%), and *agender* (5%). Because the data ends right as *non-binary* was taking over as a new umbrella term, we cannot see the same sea-change we saw in FTM, but the term does show up as a less common option prior to 2008 (10% of non-*genderqueer* tokens).

The fact that the data in GENDERQUEER drop off much earlier than the other communities is itself worthy of comment. One possibility is that LiveJournal was never as much of a hub for non-binary people than it was for other trans people, and when general usage of LiveJournal began to decline, non-binary users were ready to make use of other platforms. Another possibility is that non-binary people were being pulled away from LiveJournal for other platforms. For instance, Tumblr has often been identified as a platform that was (until recently) saturated with trans discourse and offered opportunities for non-binary visibility in particular. Tumblr was introduced in 2007, right as LiveJournal usage began to drop dramatically.

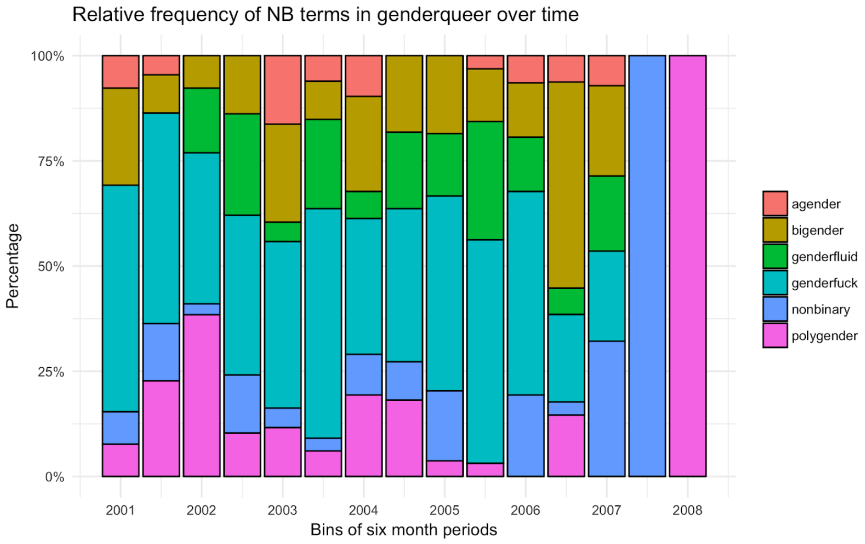

Figure 13. The relative popularity of terms for non-binary people in GENDERQUEER

7. Discussion & conclusions. The trans-livecorpus has allowed us to explore a number of significant shifts in trans people's language use during the 2000s. During that decade, the word *trans* came to predominate, while longer forms like *transgender(ed)* and *transsexual* faded in use. Among the long forms, both the *-ed* ending and the word *transsexual* came to be disfavored, with *transgender* emerging as the long form of choice. We were also able to identify the emergence of *cis* starting as early as 2003, and how it came to fully replace the descriptors *bio(logical)* and *genetic* by 2007. Finally, we were able to identify 2007-2008 as the time that *non-binary* began to take hold as the widely used term for people previously known as *genderqueer*.

Comparing across communities, we found that the patterns in FTM – the largest community in the corpus – did not always hold in other communities. Specifically, where statistical analysis was performed we found that the FTM community patterned with the TRANSGENDER group, while the MTF and GENDERQUEER communities showed different trends (or a lack thereof). Given that the dataset from the latter groups were smaller those from either FTM or TRANSGENDER, one possibility is that the datasets from the less popular groups do not include

enough tokens to observe any changes over time; however, "small" here is relative, as MTF and GENDERQUEER still contributed 1-2 million words each to the corpus. Another potential explanation is that members of FTM are leading at least some of these changes, with other groups following behind. This is borne out by some of the patterns in the data; for instance, the decline of *genetic* as a term for cis people begins in around 2003-2004 in FTM and not until 2006-2007 for MTF (see Figure 8). Additionally, although page limitations prevent us from fully exploring it, the change from the long forms *transgender(ed)* and *transsexual* to the short form *trans* appear to be further in progress in the FTM community than the others. Even if this hypothesis is correct, however, and the members of FTM are leading language change, it is also crucial to recognize the distinctive histories of these communities, as we did with the use of *genetic girl* in past generations of trans women.

If the FTM community is leading in these changes, this opens the door for a problematic comparison to cisgender women, who are often said in variationist sociolinguistic research to lead (certain kinds of) change. But while women's linguistic innovation in studies of sound change has been understood in terms of their subordinated social status (Eckert 1989), theorists of language, gender, and sexuality have documented the ways socially powerful groups exert control over linguistic norms and, specifically, the meaning and acceptable usage of words (e.g. Spender 1980, Braun & Kitzinger 2001, McConnell-Ginet 2001). It is the latter interpretation that is most sensible here, especially given that trans feminine people consistently face more, more intense, and more violent types of transphobia than do transmasculine individuals (e.g. Grant et al. 2012). When thinking about differences between female-assigned and male-assigned trans people, then, we need to remember how differently positioned these groups are with respect to institutional power and to resist easy but ultimately transphobic interpretations based on frameworks that were developed with only cisgender people in mind.

This paper has served as an example of the ways corpus sociolinguistics allows us to weave together concerns with power and normativity with powerful computing tools and large datasets. Many new, widely available computational tools can be extended to quickly create custom corpora out of semi-structured social media data, which was crucial in order to explore patterns that predate the platforms with pre-structured data and easy to use APIs. Such an approach serves as an avenue for representing speakers who are unlikely to appear in existing corpora, whether due to the small size of the community or to lack of access to the contexts from which data are collected. We have taken this approach by building a corpus consisting of interactions within online trans communities, which allowed us to provide an analysis of lexical change that contextualizes trans people's current linguistic. This kind of sociohistorical contextualization is critical if we hope to understand how we got here – and where we need to go next.

References

Baker, Paul. 2003. No effeminates please: A corpus-based analysis of masculinity via personal adverts in Gay News/Times 1973–2000. *The Sociological Review* 51(1). 243–260. https://doi.org/10.1111/j.1467-954X.2003.tb03614.x.

Baker, Paul. 2004. "Unnatural acts": Discourses of homosexuality within the House of Lords debates on gay male law reform. *Journal of Sociolinguistics* 8(1). 88–106. https://doi.org/10.1111/j.1467-9841.2004.00252.x.

Chen, Mel Yuen-Ching. 1998. "I *am* an animal!": Lexical reappropriation, performativity, and queer. In Suzanne Wertheim et al. (eds.), *Engendering communication: Proceedings of the Fifth Berkeley Women and Language Conference*, 129–140. Berkeley, CA: BWLG.

Dame, Avery. 2018. Transgender USENET Archive Project. http://averydame.net/?page_id=506. (20 Feb, 2020).
Eckert, Penelope. 1989. The whole woman: Sex and gender differences in variation. *Language Variation and Change* 1(3). 245–267.
Giardina, Henry. 2019. An oral history of the early trans internet. *Gizmodo*. https://gizmodo.com/an-oral-history-of-the-early-trans-internet-1835702003 (20, Feb, 2020).
Grant, Jaime M., Lisa A. Mottet, Justin Tanis, Jack Harrison & Mara Keisling. 2012. *Injustice at every turn: A report of the National Transgender Discrimination Survey*. Washington, DC: National Center for Transgender Equality & the National Gay and Lesbian Task Force.
Grieve, Jack, Andrea Nini & Diansheng Guo. 2016. Analyzing lexical emergence in Modern American English online. *English Language & Linguistics* 21(1). 99–127. https://doi.org/10.1017/S1360674316000113.
Hazenberg, Evan. 2017. Naming ourselves: Trans self-labelling. In Evan Hazenberg & Miriam Meyerhoff (eds.), *Representing trans: Linguistic, legal, and everyday perspectives*, 204–225. Wellington, New Zealand: Victoria University Press.
Lakoff, Robin. 1973. Language and woman's place. *Language in Society* 2(1). 45–80. https://doi.org/10.1017/S0047404500000051.
McConnell-Ginet, Sally. 2001. "Queering" semantics: Definitional struggles. In Kathryn Campbell-Kibler, Robert J. Podesva et al. (eds.), *Language and sexuality: Contesting meaning in theory and practice*, 137–160. Stanford, CA: CSLI Publications.
Silverstein, Michael. 1981. The limits of awareness. *Working Papers in Sociolinguistics* 84. 1–30.
Spender, Dale. 1980. *Man made language*. New York: Routledge & Kegan Paul Books. Stryker, Susan. 2008. *Transgender history*. Berkeley, CA: Seal Press.
Wong, Andrew D. 2005. The reappropriation of *tongzhi*. *Language in Society* 34(5). 763–793. https://doi.org/10.1017/S0047404505050281.
Zimman, Lal. 2014. The discursive construction of sex: Remaking and reclaiming the gendered body in talk about genitals among trans men. In Lal Zimman, Jenny L. Davis & Joshua Raclaw (eds.), *Queer excursions: Retheorizing binaries in language, gender, and sexuality*. 13–34. Oxford, UK & New York: Oxford University Press.
Zimman, Lal. forthcoming. Transgender language, transgender moment: Toward a trans linguistics. In Rusty Barrett & Kira Hall (eds.), *The Oxford handbook of language and sexuality*. New York: Routledge.
Zimman, Lal & Will Hayworth. 2020. Lexical change as sociopolitical change in trans and cis identity labels: New methods of the corpus analysis of internet data. *Pennsylvania Working Papers in Linguistics* 25(2). https://repository.upenn.edu/pwpl/vol25/iss2/17/.