Author accepted manuscript.

Published version of record:

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Coats, S. & Laippala, V. (eds.). *Linguistics across Disciplinary Borders: The March of Data,* 213–241. London: Bloomsbury Academic. https://www.bloomsbury.com/uk/linguistics-across-disciplinary-borders-9781350362260/

Exploring self-identification and the functions of the *identify as* construction in *the LGBTQ+ Reddit Corpus*

Abstract

Computer mediated communication has played an important role in shaping the current discourses on gender and sexuality by bringing together often dispersed minorities and providing an anonymous space to explore questions related to identity (e.g., Darwin 2017, Harper et al. 2016). We explore linguistic self-identification practices among sexual and gender minorities on the discussion forum Reddit.

For this purpose, we have compiled *the LGBTQ+ Reddit Corpus* (*c.* 44 million words). This chapter focuses on investigating the textual functions of the construction *identify as*. Our findings suggest that the construction is a productive discursive means for claiming a specific identity and positioning oneself in the discourse (e.g., *I identify as nonbinary*). However, the construction is also used for labelling others, and employed in meta-discussions, for example. In the paper, we describe the corpus and its compilation, and present our qualitatively orientated analysis, discussing and contextualizing online self-identification practices within the broader discourse on gender and sexuality.

KEYWORDS: Corpus studies, LGBTQ+ studies, self-identification, identity discourse, online discourse, Reddit

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

1 Introduction

Self-identification has become a central topic in conceptualizing gender and sexuality. This chapter explores self-identification through the use of a specific construction, which we refer to as the *identify as* construction. While there are many ways to express one's belonging to an identity-based category inclusion (e.g., *I am X, as a(n) X*), the *identify as* construction has become a salient rhetorical resource for indicating self-identification. This is evident for example in the fact that in the 36-billion-word *enTenTen20* corpus (2020), which we can treat as broadly representative of language use on the English-language internet today, *as* is the most frequently occurring preposition after the lemma IDENTIFY. Moreover, the collocates of this construction often specifically relate to gender and sexual identities. This is shown by the Multiword Sketch for *identify as* (Table 1), where prominent collocations include *woman*, *transgender* and *LGBTQ*. [1]

Table 1. Multiword Sketch for *identify as* in *enTenTen20*

Collocate	Frequency	Score (typicality)
priority	3,411	8.7
(risk) factor	3,537	8.5
area	3,101	8.3
cause	2,571	8.2
woman	**2,876**	**8.1**
transgender	**1,794**	**8**
Christian	1,571	7.6
LGBTQ	**1,408**	**7.6**
source	2,756	7.5

Thus, while the verb *identify* often occurs in passive clauses like example (1), linking subjects (*low vitamin D levels*) to predicative complements (*a risk factor*), another major function for this verb

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

clearly relates to gender-based self-identification and other-identification in active clauses, as illustrated in example (2).

(1) *Low vitamin D levels have been **identified as** a risk factor for several major cardiovascular diseases. (enTenTen20)*

(2) *Those issues can be especially significant for students who **identify as** LGBTQ. (enTenTen20)*

The chapter examines the textual functions of the *identify as* construction in contexts like (2), with a secondary aim of investigating which items typically occupy the subject and predicative complement slots. The results demonstrate the many ways in which the construction is employed in self-identification discourse, further illustrating contemporary trends in conceptualizing both gender and sexuality as diverse and fluid constructs.

Our analysis of self-identification focuses on gender and sexual minorities, as many (linguistic) changes related to this discursive practice have been instigated by these communities (e.g., Hanmer 2010, McInroy and Craig 2018, Zimman and Hayworth 2020a,b). We investigate self-identification in online discourse, since by bringing together often dispersed minorities and providing an anonymous space to explore questions related to identity, this mode of communication has played an important role in shaping the current discourses on gender and sexuality (e.g., Darwin 2017, Harper et al. 2016).

Since we are interested in identity-specific uses of the construction (as per example 2 above), rather than utilizing general corpora, we explore *identify as* in a specialized corpus: *The LGBTQ+ Reddit Corpus* (*c.* 44 million words). Representing online discourse by various LGBTQ+

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

communities[2] on the large discussion forum Reddit, the corpus includes discourse about sexual and gender identities. It covers texts from the time period from 2010 to 2021, containing both original submission posts and subsequent comments. This sizeable corpus enables the corpus-based investigation of the *identify as* construction and its functions in discourse, as it provides a substantial number of instantiations in a variety of contexts (*c.* 10,000 occurrences in the corpus). To explore the textual functions of the construction in adequate detail, we investigate a smaller sample of the concordance lines (n=550). While the small sample is necessitated by the qualitative approach, sampling from a larger corpus has the benefit of providing sufficient representation in terms of diversity in the use of the construction.

The data was sourced from Reddit, which, along with Twitter, has become an increasingly rich resource for corpus-linguistic studies (e.g., Kiesling et al. 2018, Dayter and Messerli 2022). Established in 2005, Reddit is one of the largest discussion forums, with over 50 million unique daily users (Reddit Inc 2023). Reddit is an apt resource for linguistic analysis since the data are readily available, for example through the big data and social media analysis website *pushift.io* (Baumgartner et al. 2020). Reddit is also a convenient site specifically for studying questions related to gender and sexual identities, because it contains a large number of subforums, i.e., subreddits, devoted to the discussion of these topics. We introduce the corpus in more detail in section 3.

2 Background

This chapter investigates discourse related to sexual and gender identities. While much more nuanced conceptualizations of identity exist (see, e.g., Stets and Burke 2014), for our purposes we employ a broad definition of identities as "the social positioning of self and others" (Bucholtz and

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

Hall 2010: 18).[3] We further consider identities to be intersubjectively constructed in discourse, both in the sense of their constitution being negotiated collectively, as well as individuals constructing and representing their identities discursively (e.g., Bamberg, De Fina and Schiffrin 2011, Bucholtz and Hall 2010). Certainly, identities can be seen to have an innate component, a person's "self-concept" (e.g., Tajfel 1978: 63), and parts of one's identity may be rooted in "biological determinants" (see Dillon, Worthington and Moradi 2011), but our focus remains on the discursive nature of identity.

How gender and sexuality are understood has varied historically and depending on the socio-cultural context (see e.g., Jackson 2006, Weeks 2017). Furthermore, gender and sexuality are intrinsically connected, particularly in the way gender is tied into the organization of sexual categories (e.g., Jackson 2006). This means that changes in the conceptualization of gender ultimately affect the categories of sexuality. It is via discourse that the perception of both gender and sexuality has broadened from binary understandings to encompass a plurality of possibilities (e.g., Diamond, Pardo and Butterworth 2011, Dillon, Worthington and Moradi 2011). Once considered as fixed and "given", gender and sexual identities are now widely conceptualized as a matter of self-determination to a much greater degree than previously (e.g., Zimman 2019). Certainly, changes at the societal level have also been required: Singh (2019) attributes the public recognition of minority sexual identities in the 21st century to the international framework of human rights both in Western and non-Western contexts.

Early work on sexual behavior already conceptualized binary sexuality as a spectrum, on which individuals may identify to varying degrees as either heterosexual or homosexual (Kinsey, Pomeroy and Martin 1948). Yet, more recent work has established gender and sexual identities as even more fluid and flexible than previously thought, not only changing contextually or culturally,

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

but even intra-individually throughout a person's lifetime (e.g., Cordoba 2020, Diamond 2008, Manley, Diamond and van Anders 2015). What is more, criteria for category placement might vary between individuals. For example, sexual practices do not always align with self-identification in an uncomplicated manner, e.g., some self-identifying heterosexual men may engage in male-to-male relations (e.g., Dillon, Worthington and Moradi 2011). In addition, individuals with non-normative identities are more likely to endorse complex and multiple gender and sexual identities (Galupo, Mitchell and Davis 2015). Thus, considerable variation may exist between individuals identifying within the same category.

Corpus studies provide a useful avenue for studying discourse of both gender and sexuality. Corpora have already been extensively used in this area, and one of the main questions has concerned the discursive construction of identities (for overviews, see Motschenbacher 2018 and Loureiro-Porto and Hiltunen 2020).[4] Previous corpus studies have often focused on media representations of different LGBTQ+ identities in various public discourses, such as gay men and transgender people (e.g., Baker 2005, 2014, Zottola 2021), bisexuals (Wilkinson 2019), and LGBT refugees (Wilkinson 2020),

However, while these and other studies on (media) representation help us understand how identities are collectively constructed in discourse (e.g., Wilkinson 2020), their obvious limitation is that they do not directly provide information about the linguistic practices through which community members themselves construct their identities. To investigate these questions, online discourse produced by the members of the LGBTQ+ communities are potentially more relevant. Such studies include, for example, an ethnographic investigation of a genderqueer Reddit community (Darwin 2017), a case study of transgender vloggers (Jones 2019), and an investigation of trans women's activism on YouTube (de Lima Lopes 2022), as well as several corpus studies.

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

The usefulness of online data for corpus studies was already demonstrated by King (2012, 2015), who explored participant interactions and "sex talk" on *Queer Chatrooms*, and Milani (2013), who showed that the construction of queer male identities in personal profiles of an online community was often intertwined with notions of ethnicity. More recently, Webster (2022) has shown how transgender discourses on Twitter challenge conventional notions of sex and gender, focusing on the intersection of being transgender and lesbian. Particularly relevant to the present chapter are the studies of Webster (2019) and Zimman and Hayworth (2020a, b). The former investigated identity construction in introduction posts by transfeminine individuals, which reveal considerable variation in the way discussants claim identities for themselves in a discussion forum, in contrast to the often-assumed homogeneity of trans identities. On the other hand, Zimman and Hayworth investigated the terms that transgender individuals claimed for themselves in the 2000's on the blogging website *LiveJournal;* they identified an important terminological change as the community moved from using *transgenderist, transgendered* and *transsexual* to the plain form *transgender*. The present chapter continues the investigation of self-identification practices with discussion forum data from the 2010's and early 2020's.

3 The LGBTQ+ Reddit Corpus

The corpus includes texts from chosen LGBTQ+ related subreddits, which typically include discussion about gender, sexuality, and identities. The subreddits were identified in a bottom-up fashion by relying on internet searches and reading the discussions on the forum. The final list contains 51 subreddits, including general subreddits such as *r/lgbt, r/trans*, *r/demisexuality*, location-specific subreddits such as *r/LGBTChicago, r/LGBTireland, NonBinaryUK*, Q&A

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

subreddits such as *r/asktransgender*, as well as a few hobby-related subreddits, e.g., *r/lgbt_superheroe*s. (For a full list of subreddits, see Table 9 in Appendix A.)

An exploration into the subreddits and their chronology further suggested that the number of relevant discussions has constantly increased in the history of Reddit. As many of the relevant subreddits only emerged in the second decade of the 21st century, we decided to sample subreddits covering the 11-year time period from January 2010 to November 2021. For each month during the time span, the corpus includes about 600 submissions with their subsequent comments (ranging between 5000–7000 comments per month). Especially in the early years, there were generally fewer comments per submission. Since not all subreddits are represented in the early 2010's, and some are considerably smaller than others, the corpus is not fully balanced in terms of the number of texts per year nor per subreddit (see Appendix A). The text and word counts of the corpus are shown in Table 2.

Table 2. Statistics of the LGBTQ+ Reddit Corpus

Number of subreddits	51
Number of posts	88,769
Number of comments	782,278
Number of tokens	43,962,874

There were two methods of data collection. With some exceptions, the data for 2017–2021 was streamed from the Pushshift API, using the Pmaw Python package (Podolak 2022). The data for 2010–2016 was sourced from the Pushift data dumps (see Baumgartner et al. 2020) as the API did not enable us to retrieve submission-specific comments. The data were sourced as json-objects, which were converted into Python data frames using the Pandas package. The submissions were

further processed to remove newlines and tabulations, and duplicate posts. All well-formatted hyperlinks were also replaced with a tag [LINK], but links without the *http* remain in the data.

The corpus was stored as a csv-file, which enables the linking of textual content of each submission with relevant metadata, including ID number, subreddit source, type of the submission (post or comment), and time of posting. Titles of posts are separated from the actual text of the submissions, which may occasionally be only a hyperlink, and posts and comments share a common identifier, allowing for contextual analysis and linking posts to different levels of analysis.

While the majority of the data in the corpus represent authentic language use by Reddit users, there are also two edited text types that occur: 1) direct quotes (e.g., from newspapers) and 2) research study invitations. Direct quotes often appear among posts including a link to a newspaper article. Research study invitations are relatively common in the data as well, as researchers sometimes solicit participants on Reddit. Both direct quotes and study invitations are included in the full corpus, but some instances were excluded from the smaller sample selected for analysis (see section 4).

Since Reddit is open to anyone without registration, Reddit data can be considered public without a reasonable expectation of privacy, and thus suitable for research purposes (see European Commission 2021: 13–14). However, as Scott points out, even if there is no restriction to data, the authors may not have "intended the content to be available for general consumption and analysis" (2022: 157). Particularly with vulnerable populations (such as the LGBTQ+ community) and potentially sensitive content, researchers ought to carefully consider matters of anonymity as well as potential negative outcomes for the informants were they to be identified. Since Reddit disallows posting identifying information, usernames and texts are typically anonymous.

Nevertheless, we opted to exclude usernames from the data, and as an additional safeguard, when quoting the data, we consider the identifiability and content of the passages carefully. As a result, quotes have been anonymized by replacing words or paraphrasing to prevent searchability (see Scott 2022: 157).

To gain a first impression of what discourse topics are prominent in the corpus, we ran a keyword analysis. Table 3 shows the top 20 keywords sorted according to effect size (log-ratio), with the 100-million-word British National Corpus used as the reference corpus (BNC 2007). The top keywords include vocabulary related to gender and sexual minorities (e.g., *lgbt, transgender dysphoria*, *demisexual*); names of websites (reddit, facebook, youtube) and features of informal online communication also stand out (e.g., *idk, tbh*). While we do not go further into the topics and their appearance across the corpus, a survey of top keywords clearly indicates that the corpus represents various facets of discourse related to the LGBTQ+ communities — which is our main object of interest— also as far as the actual contents of the texts are concerned.

Table 3. Keyword Analysis of the LGBTQ+ Reddit Corpus

#	Freq	Keyness (LL)	Effect (log ratio)	
1	24387	52347.31	16.5182	lgbt
2	16329	35048.67	15.9395	transgender
3	9637	20683.97	15.1788	reddit
4	15754	33793.86	14.8878	dysphoria
5	6690	14358.51	14.6522	subreddit
6	4715	10119.5	14.1474	nonbinary
7	4714	10117.35	14.1471	mtf
8	4415	9475.61	14.0526	lgbtq
9	4316	9263.13	14.0199	transphobic
10	4230	9078.54	13.9908	agender
11	4154	8915.43	13.9647	facebook
12	3648	7829.41	13.7773	genderqueer
13	2657	5702.47	13.32	tbh
14	2567	5509.3	13.2703	genderfluid
15	2501	5367.65	13.2327	afab
16	2460	5279.66	13.2088	transphobia
17	4803	10290.26	13.1741	idk
18	2362	5069.33	13.1502	website
19	2355	5054.3	13.1459	demisexual
20	2224	4773.15	13.0633	youtube

4 Methods

Building on previous corpus-based work investigating the representation and construction of identities (e.g., Baker 2005, 2014, Milani 2013), we employ a large corpus to investigate a specific construction and its functions in discourse. We focus on the construction *identify as*, illustrated in example (3). This construction roughly corresponds to the simple verb patterns *V as adj* and *V as n* in Francis et al.'s (1996) pattern grammar analysis of English verbs. It is also a particularly relevant one for the present study, as it gives access to the linguistic practice of self-identification through which gender self-determination is realized (Zimman 2019).

(3) *I identify as demisexual.*

This construction is a complex-intransitive construction, which is partially lexically determined: it consists of a subject (*I*), a verb (*identify*), a preposition (*as*) and a complement of the preposition (*demisexual*, which can either be a noun or an adjective), which in our data typically expresses the relevant identity category. Semantically, the predicand is the subject of the clause (here, *I*), and the *as*-phrase is a "marked predicative complement", which expresses a property that is predicated by the person which is picked out by the subject (Huddleston and Pullum 2002: 217, 254). There is also a corresponding transitive construction which is oriented towards the object of the clause (e.g., *we have identified **it** as a problem*, see e.g., Hiltunen 2010), excluded from this study. Francis et al. (1996) identify six meaning groups for their *V as N* pattern (*Work* group, *Function* group, *Begin and end* group, *Rank* group, *Masquerade* group, and Other), but do not, interestingly, include *identify* among the verbs that frequently occur in this pattern. They do, however, note that the pattern in question is productive, and many different verbs can be used in it to indicate what the role of the subject in the predication. Furthermore, based on the analysis of our selected sample (see below), a particular feature of *identify as* seems to be that it is rarely negated (only *c.* 7 percent of occurrences); claiming an identity seems to be more common than rejecting them.

To investigate the *identify as* construction in our material, we wrote a script that exhaustively retrieved the instances of the lemma IDENTIFY followed by *as* in the corpus, including instances of the less-frequent form SELF-IDENTIFY, resulting in 10,833 hits. Instances with intervening words were left out of the analysis, as the time cost involved in searching for them was deemed to outweigh the benefits of improved recall. To allow for a detailed qualitative analysis, we selected a small systematic sample of approximately 5 percent of the concordances. We collected the sample by selecting every n^{th} row with the concordance list being organized by the R2 slot, which

includes the majority of identity labels used as complements in the construction. This approach helped in obtaining a diverse selection of different identity labels in the data, which are one of the foci of our analysis. Table 10 in Appendix A shows which subreddits are represented in our sample (554 concordances).

We excluded from analysis false positives that were not connected to identity discourse (e.g., *AIDS was identified as a disease*) (n=2), as well as direct quotes (n=2). However, cases of imagined speech were included in the analysis, since they still represent the writer's ideas of language use, e.g., *They're certainly not going to take you as seriously if you say "I self-identify as a woman" over "I am a woman".* We also kept survey invitations and reports in the sample, as they may still shape the discourse, i.e., how identities are talked about. Thus, our final sample comprises 550 concordance lines.

For each instance of *identify as*, we identified the subject, the predicative complement and the local textual function, understood as a function specific to groups of lexical items in their textual context (cf. Mahlberg 2007). Each subject was coded on three levels: token (normalizing the spelling),[5] grammatical category, and the reference of the subject (e.g., reference to self, or generic reference). The predicative complements were also coded as tokens, further specifying the type of the token (grammatical category, and whether the token was an identity label), and last, categorizing the complements as references to gender and/or sexuality.

The analysis of the local textual functions was inductive, and the scope was broadened to include the surrounding context (R20–L20). In other words, the functions do not rise directly from the construction *identify as*, but from the context in which it is used. During the first round of close reading, roughly following principles of thematic analysis (Braun and Clarke 2006), the concordances were coded for various first-level functions (such as positioning oneself in the

discourse or engaging in meta-discussions). During subsequent rounds of coding, second-level functions were identified, such as explaining one's identity position or providing definitions.[6] Requiring sufficient familiarity with current scholarly and popular discourses on gender and sexuality, the analysis was primarily carried out by the first author. The approach allowed one researcher to engage thoroughly with the data, particularly beneficial for the identification of different textual functions. Nevertheless, the coding scheme, categorization and unclear cases were discussed among all authors. To ensure reliability, several check-up rounds within categories and across categories were carried out.

5 Patterns of *identify as*

Utilizing the sample of 550 concordances selected for further analysis, in this section we examine the patterns of *identify as*, first exploring the subjects and predicative complements of the construction, and second, delving into the textual functions of the construction. The analysis illustrates that *identify as* is employed to serve various functions in discourse. Overall, the data comprises discussions about identity both at a personal level, referring to the identities of specific people, as well as at a general level, for example, explaining what certain identities are.

5.1. Subjects of *identify as*

Most distinctly, the subjects of *identify as* reveal an emphasis on the self. This is demonstrated by the first-person pronoun *I* being the most frequent subject token (Table 4). Some other personal pronouns appear as well, including *you* and *they*, which direct the reference to other people. In total, personal pronouns appear as the subject in 351 concordances. Additionally, the slot is sometimes occupied by nominal phrases such as *some people*, *most agender people* or *my*

girlfriend. Perhaps reflecting the anonymous nature of the discussion forum, names were rarely used (n=2) although specific references to others were not uncommon (Table 5).

Table 4. Subject Tokens, spelling normalized

Subject Token	*freq ≥ 2*
I	212
you	75
people	33
they	32
she	11
he	10
someone	10
anyone	6
we	6
friend	3
some people	3
no one	3
pansexuals	2
person	2
my girlfriend	2
most people	2
trans people	2

The sample also includes 40 cases where *identify as* appeared either without a subject, including nonfinite constructions (examples 1 and 2) and imperatives (3), or with different types of clausal subjects (4). In many cases, an agent was still present (1), and thus the specificity of the reference could be determined (Table 5).

1. *First started identifying as [gay] when I was [thirteen]* [7]

2. *I went into high school identifying as straight*

3. *Identify as whatever you want*

4. *That makes it easier to identify as* [*a woman*]

Table 5. Subject Reference

Subject Reference	*freq*
Specific total	**341**
self	236
other-person	91
other-group	14
Nonspecific total	**192**
group	102
generic	90
Other	**17**
Total	550

Considering the specificity of the reference (Table 5) further confirms that much of the discourse focuses on specific individuals or groups (62 percent) rather than on an abstract 'other'. Self-references were most frequent, but specific other-references were present as well, for example, addressee-specific *you*s, and nominal phrases which typically referenced friends or partners (5–6).

5. *[My friend] identifies as a genderqueer female*

6. *My [partner] identifies as bi*

Nonspecific references include generic references and references to unspecific yet named groups, including group-level descriptions (7) and, for example, survey invitations (8). Generic references refer to 'everyone' or 'anyone', typically expressed with personal pronouns or indefinite pronouns (9). Some concordances are ambiguous between generic and specific use readings (10), especially with *you;* these are included in the counts for 'other' (Table 5).

7. *because lots of "pansexuals" identify as "bisexual" for ease*

8. *You can participate in this study if you* [...] *identify as a ciswoman*

9. *Even if there isn't that evidence, someone identifying as a woman is enough*
10. *the most important thing is to identify as you see fit*

5.2. Complements of *identify as*

The intricacy of distinctions related to gender and sexuality categories is highlighted in the complements of *identify as.* As shown in Table 6, a variety of different identities are evoked in the discussions. Most of the tokens in the complement position are identity labels (about 80 percent), and more specifically gender or sexuality labels, as expected with this corpus. However, there were also a few political, ideological and ethnic labels (e.g., *feminist, black*), as well as a handful of various descriptions and labels that more-so describe behavior than an identity, e.g., *a man in a dress, a carrot, a chaser, a trap* (Table 7). Acronyms such as FTM or LGBTQ+[8] only rarely appeared as complements (10 occurrences). Other types of complements were uncommon but included, for example, pronouns and clausal constructions; *identify as* also appeared 25 times without a complement (Table 7).

Table 6. Complement Labels, spelling normalized

Complement Labels	*freq ≥ 3*
bi/bisexual	55
lesbian	23
female	22
male	19
straight	18
gay	17
woman	17
nonbinary	17
trans	16
genderqueer	11
queer	10
asexual	10
man	9
agender	8
pansexual	8
transgender	7
girl	6
ace	6
genderfluid	5
pan	5
male or female	4
LGBTQ+	4
panromantic asexual	3
androgynous	3
demisexual	3
women	3

Table 7. Label Category

Label Category	*freq*
GENDER total	**250**
transgender (marked)	127
binary (+transgender context)	46
binary (unmarked)	45
binary + cis	7
general gender	25
SEXUALITY total	**197**
LGBQAP+	172
heterosexuality	19
both LGBQAP+ and heterosexuality	2
general sexuality	4
GENDER AND SEXUALITY total	**15**
gender and sexuality combination	8
LGBT acronym	5
ambiguous	2
OTHER total	**88**
no complement	25
political, 'ism', religious, ethnicity	13
behavior-based label	5
other	45
Total	**550**

As expected, the complements mostly represent minority genders and sexualities (about 64 percent, Table 7). With gender, the most frequent category comprises explicitly marked transgender identities, including unspecific labels such as *trans*, marked binary transgender labels (such as *trans woman*), as well as different nonbinary identities (such as *genderfluid* and *agender*, included under the nonbinary umbrella, e.g., Matsuno and Budge 2017). The marked transgender category mostly includes identity labels, but also descriptions indicating transgender experience

in other ways, such as identifying "without gender", or identifying as "something else" than female or male.

Distinguished from the marked transgender category, sometimes transness was not marked explicitly on the complement but could still be deduced from the surrounding context (11). However, a few cases where the complement itself was simply "gender" or "assigned gender" but the context pointed towards transness, are regarded as general references to gender (12). Moreover, *female* and *male* identities were often claimed without further specification (unmarked; *man, woman*), and only rarely specified as cis (e.g., *cis man*).

11. *I'm just happy to date anyone who looks and identifies as a woman* […] *some MtFs are ridiculously good looking*

12. *Trans to me means I identify as a gender* [*other than what I was*] *assigned at birth*

Most sexuality labels represented lesbian, gay, bi, queer, asexual and pansexual identities (LGBQAP), but included are also labels such as *demisexual* and *sapiosexual.* Heterosexual identities were mostly described as *straight*, with only a few descriptions of *heterosexual*.

Examining the label categories further illustrates that typically discussions about gender and sexuality concern specific identity categories. For example, complements referring to multiple potential identities (13) or complements without specific identity labels (14), coded as *general gender* and *general sexuality* respectively, are notably infrequent.

13. *We identify as male, female, neither*

14. *there's nothing wrong with identifying as one thing now and another thing later*

Similarly, while identities are known to be intersectional, it was rare for discussants to bring up different aspects of their identity as complements of *identify as,* as in example 15 (only 5 occurrences in total). Even more broadly, considering the complement categories, Table 7 shows that complements typically referred to only gender, or only sexuality, and only rarely were both present in the same complement.[9] In a related fashion, discussants typically claimed a single, unmodified label; only in 17 instances did discussants describe either their gender or sexuality by combining different labels (16). There were also a few dozen cases of listing different identities, mostly in generic contexts (example 13). Identity labels were also sometimes modified in different ways, to either highlight the degree of identification, e.g., *100% straight, a very lesbian-leaning bisexual, gay(ish),* to provide additional details, e.g., *a male presenting trans woman*, or to indicate uncertainty, e.g., *probably nonbinary male.*

15. [*People often ask me*] *about identifying as a gay trans man*
16. [*I identify as*] *asexual grey-heteroromantic*

5.3. Functions of *identify as*

In order to assess which purposes the construction serves in the discourse, the concordances were further categorized based on their local textual functions (Table 8). At the first level, five different textual functions were discerned (a sixth category in Table 8 gathers various other instances, further discussed below):

1) positioning oneself in discourse
2) seeking and providing advice
3) engaging in meta level discussions

4) engaging in debates

5) assigning identities onto others.

Table 8. Textual Functions

Functions	*freq*
Positioning total	**226**
explaining	108
pondering	31
discovering	24
questioning	20
mentioning	14
introducing oneself	9
defending	7
testing	4
coming out	2
closeted	2
other	5
Advice total	**71**
giving advice	51
asking for advice	6
question	14
Meta total	**35**
defining	24
pondering	6
linguistic	5
Debate total	**108**
argument	62
explanation	46
Assigning total	**56**
General/other total	**54**
study invitation/report	12
attraction	11
other	31
Total	550

Positioning oneself in the discourse: explaining, discovering and defending identities

The most frequent function for *identify as* was **positioning oneself in discourse**. Most commonly, the discussants were *explaining* how and why they identify with a particular identity category (17–18). There were also many instances, in which the discussants were *pondering* and reflecting on what their identity means to them or wondering how to conceptualize their identity (19–20). In a related fashion, some discussants more directly *questioned* their current form of self-identification, as another form of self-reflection (21). Taking a step further, a few discussants even shared accounts of *testing out* different identities (22–23).

17. *I will likely only ever date women, so now I identify as* [*a lesbian*]
18. *I felt like I didn't identify as anything so I came out as* [*agender*]
19. *I could also be considered nonbinary* [*as I identify as*] *genderfluid*
20. *I could identify as* [*a man but I*] *have very female expressions*
21. [*lately I have started*] *questioning my gender again*
22. *I'm currently (tentatively) identifying as* [*nonbinary*]
23. *I finally began to officially identify as genderfluid* [*after I tried out*] *identifying as a guy for a day*

Discussants also shared narratives related to how they *discovered* their identities, sometimes as a result of personal realization (24), and sometimes after learning about a new identity category (25). The latter occurred especially among bisexual discussants, some of whom embraced pansexuality after discovering the term and/or as a reaction to broadening their understanding of gender. In contrast to accounts of discovering one's identity, there were

surprisingly few *coming-out* narratives (26), and only a few accounts where the discussants explicitly described hiding their identities, i.e., being *closeted* (27).

24. [*When I realized that*], *I started identifying as* [*a lesbian*]

25. *I used to identify as bisexual,* [*until I*] *learned about pansexuality*

26. [*Recently*] *I came out to* [*my partner*], *and with him I identify as male*

27. *I identify as [gay] but I'm not out publicly for fear of the responses I would get*

Serving somewhat more simplistic functions, sometimes discussants merely *mentioned* how they identify, or provided the information as a part of an *introduction* (28). In contrast to simply claiming an identity, sometimes the discussants needed to *defend* their identity position (29–30). A handful of concordances where the discussant was positioning themselves escaped categorization in more detail (*other*, example 31).

28. *About me– I identify as* [*bisexual*]

29. *I [can] identify as whatever I want*

30. *That* [*really offends*] *me,* [*since I identify as a man*], *and* [*I don't feel like having sex with men*] *emasculates me* [*at all*]

31. *I found myself more in synch with [women] who identified as [gay]*

Advice as a component of intersubjective identity construction

Remaining on a personal level of identity discussions, **providing and seeking advice** was a discernable function among the concordances (32–34); included under this function are also questions about how others identify, and some general questions about identities.[10] That there were much fewer instances of asking for advice than there were for providing advice might simply

reflect the nature of the medium; one person asking for advice is likely to receive multiple responses. Advice-seeking and receiving seems to operate as a particularly interactive component of identity building. The discussants often directly asked for advice on how to identify, or how to conceptualize their identity (33–34); one discussant even revealed that a previous discussion with someone on the forum had led to changes in their self-identification (35).

32. *Just because you identify as it doesn't mean you need to take action immediately*

33. [description of same-sex desire] *Can I identify as straight but* [*slightly on the Kinsey scale*?]

34. [*Physically I am bisexual, but emotionally, I have no idea.*] *I feel gay and I identify as gay, even though I like girls.* [*Help, input or advice?*]

35. [*Nice to see you again*] […] [*since we talked last time*] *I started identifying as* [*pansexual*]*, you helped me grow a lot!*

Negotiating identity boundaries through debates and meta-discussions

While the majority of the concordances related to identities at a personal level, there were also more general discussions about identities, including argumentative acts classified as **debates**, and contemplative accounts categorized as **meta-discussions**. With the latter, the discussants were most commonly providing *definitions* for specific identities, an act of negotiating identity boundaries. This category includes instances in which a description was simply provided (36), but also instances in which the act of defining was explicitly stated, as in examples (37) and (38); sometimes the definitions were formatted to resemble dictionary definitions (39). Although negation was rare, occasionally identity boundaries were also negotiated by determining who cannot be included in the description (40).

36. *The term transgender can include people who identify as 'nonbinary'*

37. *The definition of a demiboy is someone who identifies as a boy only in part*

38. *Cisgender means [identifying] as the gender you were assigned at birth*

39. *1. Transwoman - born male identify as female ; Transman - born female identify as male*

40. *an AFAB person cannot identify as transfeminine*

There were also a handful of instances in which the discussant *pondered* about different possibilities in general (41), in contrast to relating it to their own experience (as with the homonymous category under *positioning*). At times, discussants also engaged in *meta-linguistic* discussions, for example on linguistic nuances such as whether it is more credible to claim identities with *identify as* or with *to be* (42–43).

41. *If we all grew up in a genderneutral environment wouldn't we all just identify as human?*

42. *They're certainly not going to take you as seriously if you* [*say*]*"I self-identify as a woman" over "I am a woman"*

43. *I think people generally interpret it more seriously if I say "I currently identify as a boy" than if I say "I might be a boy"*

Acts related to debating can also be seen as negotiating identity boundaries, but the distinction to the meta-category is that *debates* are more argumentative (44). At times discussants for example employed various rhetoric tactics such as hyperbole (45) or rhetoric questions (46) to 'prove a point'. Often, discussants sought to prove their point simply by *explaining* their point of view; such explanatory accounts (47–48) are included as *debate* when they functioned at a general

level (and not as positioning oneself), and when they seemed to serve a function of a counterargument instead of providing definitions (included under *meta*). However, some overlap occurs between these categories.

44. *You can love* [*dolls*] *and dressing up while still identifying as male*

45. *Thing is no one really identifies as* [*a tree*]

46. *I identify as smart. Does that make me smart?*

47. *People that identify as bisexual* [*date transgender and nonbinary people*] *too*

48. *While many people identify as nonbinary, that doesn't mean everyone feels the same way*

Assigning identities and other functions

When identities were **assigned onto others**, the function was considered to be simply referential, not classified in more detail (Table 8). As assigning identities onto specific individuals typically requires some information about the person, it is not surprising that the discussants often indicate that there is a close relationship to the referents (49–50). Nonspecific other-references were mostly generic or referred to groups (51–52).

49. *My ex, who identifies as a lesbian, is now dating a man*

50. *I have a friend who identifies as a lesbian who started dating a trans guy*

51. *I haven't really met anyone who* [*would self-identify*] *as a chaser*

52. *people who identify as bisexual*

Not fitting any of the five first-level functions, two additional categories gather concordances related to *attraction* and *study invitations*. Interestingly, sometimes identities were

discussed from the point of view of *attraction*, both at a personal (53) and at a general level (54). Study invitations and reports were also included in this last category (55–56), as were also instances like example 57 remaining uncategorized as 'other' (Table 8).

53. *it doesn't matter to me what they* […] *identify as, I might think they're hot*

54. *there are many men who identify as straight, yet have had same sex fantasies and contact*

55. *You can participate in this study if you: are 18 or older* […] *identify as a ciswoman*

56. *1 in 4 homeless people in the UK identify as LGBTIQ+*

57. *If you told me that you identified as* [*a guy*] *it would not cause conflict in* [*me]*

Overall, the analysis revealed that in addition to its most direct function of claiming an identity for oneself, the construction *identify as* serves several distinct textual functions. Constructing identities at a personal level, *identify as* was most frequently utilized to position oneself in the discourse, while advice-seeking and receiving highlights the intersubjective nature of identity construction. Yet, identity boundaries were also negotiated at a general level, including various meta-level acts, such as providing definitions for different identity categories, and engaging in debates about what different identities comprise.

6 Discussion

Through our primarily qualitative analysis, we have determined which social actors typically occupy the subject and predicative complement slots in the *identify as* construction, as well as discovered and described the main textual functions of the construction in the sample. Based on the analysis, it is clear that the construction is often employed to position oneself in the discourse,

yet, utilizing the construction in meta-discussions serves a salient function as well, especially in acts of defining identities. Furthermore, demonstrating the heterogeneity and plurality of LGBTQ+ communities, the data include considerable variation in the identity terms discussants claim for themselves (see also Zimman and Hayworth 2020a, Webster 2019). In particular, the analysis highlights the fluidity and flexibility of identities in many ways: this is evident in the numerous descriptions of one's identity changing over time, sometimes as a result of changes in how one feels about oneself, and sometimes as a result of a learning process, for example, after discovering a new identity category.

What is also apparent from the data is that identities may vary based on context. Fluctuation of identity is particularly evident with gender identities such as *bigender* or *genderfluid,* which describe identification with multiple identities at different times (e.g., Matsuno and Budge 2017: 117, see also Corwin 2017), but other types of context-dependency were described in the data as well. For example, off-line and online identities may be different (58) (cf. Marciano 2014), identities may be revealed to only certain people but not to others (59), or they might depend on how one's partner identifies (60–61). Furthermore, while identities are predominantly seen as a matter of self-identification, one discussant also brought up the possibility of modifying one's sexual self-identification based on the desires of one's partner (62).

58. [*in real life*] *I identify as* [*a man*] *because it's my biological sex, and online I identify as* [*a woman*] *because* [*it feels better*]

59. *I identify as* [*genderfluid*] *but* [*I'm only out*] *among close friends*

60. [*Although biologically male, my partner is genderqueer and identifies as my lesbian girlfriend*]

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

61. *Just because my ex came out* […] *as trans, do I necessarily have to stop identifying as straight?*

62. *I am willing to try almost anything to make* [*him happy*]*, including identifying as bisexual*

Most commonly, *identify as* was employed in the first person, preceded by *I*, highlighting the agency of the self in determining one's identity (see Zimman 2019). Similar findings on transgender discussion forum data have previously been reported by Webster (2019), who argued that the high frequency of *I* is explained by its salient function to indicate "personal belonging and locating the self within the local discourse" (2019: 135). The consequent low frequency of plural subjects (such as *we*) in online identity discourse further highlights the emphasis on the individual, instead of group identities, aligning with Webster's results (2019: 136). The emphasis on the self is rather expected, considering the theme of identity discourse and the online discussion forum context, in which actors typically represent themselves. Yet, despite the prevalence of the first person, our data demonstrate that identities are also intersubjectively constructed in online discourse (cf. Bucholtz and Hall 2010). For one, identity boundaries are negotiated discursively, as demonstrated by acts of providing definitions and debating how to frame different identities. In these cases, identities were discussed at a general level, yet there were also instances in which identities were intersubjectively constructed at the individual level, for example, in advice-seeking and receiving.

We can establish further points of contact between our findings and previous scholarship. For example, Zimman and Hayworth (2020a) had identified a variety of different labels transgender individuals claimed in online discourse in the 2000's, including *transgendered,*

transgenderist, and *transsexual,* of which the former two did not appear in the sample of the present study, and *transsexual* appeared once. However, these terms do appear in the full *LGBTQ+ Reddit Corpus*, albeit rarely as complements to *identify as*. These differences might be due to how *identify as* is utilized, but they might also reflect changes over time, as *The LGBTQ+ Reddit Corpus* represents the 2010's and early 2020's; Zimman and Hayworth show a decrease in the use of these terms already in the late 2000s (2020a,b, see also Webster 2019).

Largely missing from the sampled data seem to be coming-out narratives, which is somewhat surprising considering the salience of coming-out in the lives of the LGBTQ+ (see e.g., Zimman 2009, Bamberg, De Fina and Schiffrin 2011).[11] Although our sampling aimed at producing a representative sample, it is possible that our results do not completely reflect the real proportional frequencies of the local textual functions in the text population, given the small sample and the fact that the corpus does not include all texts from the chosen subreddits. However, the absence of coming-out narratives might also be due to the focus on *identify as*; other constructions might be more productive in coming-out narratives. Indeed, self-identification can naturally be realized in many other ways, and further studies would do well to consider other manifestations of self-identification, including copular verb + X (*I am/feel gay*). Similarly, it would be worthwhile to analyze phrases where *as*-constructions are used as adjuncts without the verb *identify,* as in (63), as these seem to function as important rhetorical resources for presenting an argument in these forums.

63. *As a woman, I find your views anti-feminist*

Nevertheless, *identify as* seems to have become a key construction in self-identification discourse, as also indicated by the high frequency of the construction in the corpus. While a

qualitative close reading was only feasible with a relatively small sample of concordances, sampling these instances from a large corpus provided sufficient representation of the different functions, allowing us to make some generalizations and establish links between discursive practices and their lexicogrammatical realizations. Indeed, the qualitative analysis produced a nuanced understanding of the different textual functions of the construction, which might have been overlooked in more quantitatively orientated approaches.

Last, the data considered here represents identity discourse produced primarily by members of the LGBTQ+ communities. The type of interactive identity construction through self-reflection, sharing and communal debates that we have highlighted above might be characteristic of these communities, as claiming a non-normative identity position often requires contemplating one's identity in ways that might not be necessary for individuals securely residing in the heterosexual and cisgender end of the spectrum (cf. Hekanaho, forthcoming). While past research has often focused on either cisgender/heterosexual *or* transgender/homosexual experiences, future research might benefit from comparative approaches, examining potential differences in identity construction. For example, do individuals occupying different identity positions conceptualize their identities similarly or differently?

References

Andler, M. (2022), 'Queer and Straight', in B. Earp, C. Chambers and L. Watson (eds), *The Routledge Handbook of Philosophy of Sex and Sexuality*, 117–130, New York: Routledge.

Baker, P. (2005), *Public Discourses of Gay Men*, London and New York: Routledge.

Baker, P. (2014), 'Bad Wigs and Screaming Mimis: Using Corpus-assisted Techniques to Carry Out Critical Discourse Analysis of the Representation of Trans People in the British Press',

in C. Hart and P. Cap (eds), *Contemporary Critical Discourse Studies*, 211–235, London and New York: Bloomsbury Academic.

Bamberg, M., De Fina, A. and Schiffrin, D. (2011), 'Discourse and Identity Construction', in S. Schwartz, K. Luyckx and V. Vignoles (eds), *Handbook of Identity Theory and Research*, 177–199, New York: Springer.

Braun, V. and Clarke, V. (2006), 'Using Thematic Analysis in Psychology', *Qualitative Research in Psychology*, 3 (2): 77–101. doi:10.1191/1478088706qp063oa.

BNC Consortium (2007). *The British National Corpus*, Oxford Text Archive.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. and Blackburn, J. (2020), The Pushshift Reddit Dataset. Available online: https://arxiv.org/abs/2001.08435 (accessed 03 October 2022).

Bucholtz, M. and Hall, K. (2010), 'Locating Identity in Language', in C. Llamas and D. J. L. Watt (eds), *Language and Identities*, 18–28, Edinburgh: Edinburgh University Press.

Cordoba, S. (2020), *Exploring Non-binary Genders: Language and Identity*. Doctoral dissertation, De Montfort University.

Corwin, A. (2017), 'Emerging Genders: Semiotic Agency and the Performance of Gender among Genderqueer Individuals', *Gender and Language*, 11 (2): 255–277. doi:10.1558/genl.27552.

Darwin, H. (2017), 'Doing Gender Beyond the Binary: A Virtual Ethnography', *Symbolic Interaction*, 40 (3): 317–334.

Dayter, D. and Messerli T. (2022), 'Persuasive Language and Features of Formality on the r/ChangeMyView Subreddit', *Internet Pragmatics*, 5 (1): 165–195. https://doi.org/10.1075/ip.00072.day (accessed 03 October 2022).

de Lima Lopes, R. E. (2022), 'Beyond the Binary: Trans Women's Video Activism on YouTube', *Digital Scholarship in the Humanities*, 37 (1): 67–80. https://doi.org/10.1093/llc/fqab057 (accessed 12 January 2023).

Diamond, L. M. (2008), 'Female Bisexuality from Adolescence to Adulthood: Results from a 10-year Longitudinal Study', *Developmental Psychology*, 44 (1): 5–14. https://doi.org/10.1037/0012-1649.44.1.5 (accessed 03 October 2022).

Diamond, L., Pardo, S. and Butterworth, M. (2011), 'Transgender Experience and Identity', in S. Schwartz, K. Luyckx and V. Vignoles (eds), *Handbook of Identity Theory and Research*, 629–647, New York: Springer.

Dillon, F., Worthington, R. and Moradi, B. (2011), 'Sexual Identity as a Universal Process', in S. Schwartz, K. Luyckx and V. Vignoles (eds), *Handbook of Identity Theory and Research*, 649–670, New York: Springer.

enTenTen20: Corpus of the English Web (2020). Available online: https://www.sketchengine.eu/ententen-english-corpus/ (accessed 13 January 2023).

European Commission (2021), Ethics and Data Protection. Available online: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-and-data-protection_he_en.pdf (accessed 03 October 2022).

Galupo, M. P., Mitchell, R. C. and Davis, K. S. (2015), 'Sexual Minority Self-identification: Multiple Identities and Complexity', *Psychology of Sexual Orientation and Gender Diversity*, 2 (4): 355–364.

Hanmer, R. (2010), 'Internet Fandom, Queer Discourse, and Identities', in C. Pullen and M. Cooper (eds), *LGBT Identity and Online New Media*, 147–158, New York and London: Routledge.

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders*, 213–241. London: Bloomsbury Academic.

Harper, G., Serrano, P., Bruce, D. and Bauermeister, J. (2015), 'The Internet's Multiple Roles in Facilitating the Sexual Orientation Identity Development of Gay and Bisexual Male Adolescents', *American Journal of Men's Health,* 10 (5): 359–376.

Hekanaho, L. (forthcoming). The Communicative Functions of 3rd person singular Pronouns: Cis- and transgender perspectives, in M. Nevala and M. Palander-Collin (eds), *Self- and Other Reference in Social Contexts: From Global to Local Discourses.* John Benjamins Publishing Company.

Hiltunen, T. (2010), *Grammar and Disciplinary Culture*. Doctoral dissertation, University of Helsinki. https://helda.helsinki.fi/handle/10138/19278 (accessed 03 October 2022).

Huddleston, R. D. and Pullum G. K. (2002), *The Cambridge Grammar of the English Language*, Cambridge: Cambridge University Press.

Jackson, S. (2006), 'Gender, Sexuality and Heterosexuality: The Complexity (and Limits) of Heteronormativity', *Feminist Theory*, 7 (1): 105–121.

Jones, L. (2019), 'Discourses of Transnormativity in Vloggers' Identity Construction', *International Journal of the Sociology of Language*, 256: 85–101. https://doi.org/10.1515/ijsl-2018-2013 (accessed 03 October 2022).

Kiesling, S., Pavalanathan, U., Fitzpatrick, J., Han, X. and Eisenstein, J. (2018), 'Interactional Stancetaking in Online Forums', *Computational Linguistics*, 44 (4): 683–718. https://doi.org/10.1162/coli_a_00334 (accessed 03 October 2022).

King, B. (2012), 'Building and Analysing Corpora of Computer-Mediated Communication', in P. Baker (ed), *Contemporary Corpus Linguistics*, 301–320, London and New York: Continuum.

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

King, B. (2015), 'Investigating Digital Sex Talk Practices: A Reflection on Corpus-assisted Discourse Analysis', in R. H. Jones, A. Chik and C.A. Hafner (eds), *Discourse and Digital Practices: Doing Discourse Analysis in the Digital Age*, 130–143, London: Routledge.

Kinsey, A. C., Pomeroy, W. B. and Martin, C. E. (1948), *Sexual Behavior in the Human Male*, Philadelphia: W. B. Saunders.

Li, G., Sham, W. and Wong, W. (2022), 'Are Romantic Orientation and Sexual Orientation Different? Comparisons Using Explicit and Implicit Measurements', *Current Psychology*. https://doi.org/10.1007/s12144-022-03380-9 (accessed 03 October 2022).

Loureiro-Porto, L. and Hiltunen, T. (2020), 'Democratization and Gender-neutrality in English(es)', *Journal of English Linguistics*, 48 (3): 215–232. https://doi.org/10.1177/0075424220935967 (accessed 03 October 2022).

Manley, M. H., Diamond, L. M. and van Anders, S. M. (2015), 'Polyamory, Monoamory, and Sexual Fluidity: A Longitudinal Study of Identity and Sexual Trajectories', *Psychology of Sexual Orientation and Gender Diversity*, 2 2): 168–180.

Mahlberg, M. (2007), 'Clusters, Key Clusters and Local Textual Functions in Dickens', *Corpora*, 2 (1), 1–31.

Marciano, A. (2014), 'Living the VirtuReal: Negotiating Transgender Identity in Cyberspace', *Journal of Computer Mediated Communication*, 19: 824–838. https://doi.org/10.1111/jcc4.12081 (accessed 03 October 2022).

Matsuno, E. and Budge, S. (2017), 'Non-binary/Genderqueer Identities: a Critical Review of the Literature', *Current Sexual Health Reports*, 9: 116–120. doi:0.1007/s11930-017-0111-8.

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders*, 213–241. London: Bloomsbury Academic.

McInroy, L. B. and Craig, S. L. (2018), 'Online Fandom, Identity Milestones, and Self-identification of Sexual/Gender Minority Youth', *Journal of LGBT Youth*, 15 (3): 179–196. https://doi.org/10.1080/19361653.2018.1459220 (accessed 03 October 2022).

Milani, T. (2013), 'Are "Queers" Really "Queer"? Language, Identity and Same-sex Desire in a South African Online Community', *Discourse & Society*, 24 (5): 615–633. doi:10.1177/0957926513486168.

Motschenbacher, H., ed. (2018), 'Corpus Linguistics in Language and Sexuality Studies: Developments and Prospects', *Journal of Language and Sexuality* special issue, 7 (2), Amsterdam: John Benjamins.

Mullany, L. (2010), 'Gender and Interpersonal Pragmatics', in M. Locher and S. Graham (eds), *Interpersonal Pragmatics* (Handbooks of Pragmatics), 225–252, Berlin and New York: De Gruyter Mouton.

Podolak, M. (2022), PMAW: Pushshift Multithread API Wrapper 2.1.3. Available online: https://pypi.org/project/pmaw/ (accessed 03 October 2022).

Reddit Inc (2023), Statistics about Reddit. Available online: https://www.redditinc.com/ (accessed 05 January 2023).

Scott, K. (2022), *Pragmatics Online*, New York: Routledge.

Singh, P. (2019), 'Research on Diversity in Sexual Identities: Beyond Binaries', in M. J. Bosia, S. M. McEvoy and M. Rahman (eds), *The Oxford Handbook of Global LGBT and Sexual Diversity Politics*, 381–396, Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190673741.013.28 (accessed 03 October 2022).

Stets, J. E. and Burke, P. (2014), 'The Development of Identity Theory', in S. R. Thye and E.J. Lawler (eds), *Advances in Group Processes* 31, 57–97.

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

Weeks, J. (2017), *Sex, Politics and Society. The Regulation of Sexuality Since 1800*, London: Routledge.

Webster, L. (2019), '"I am I": Self-constructed Transgender Identities in Internet-mediated Forum Communication', *International Journal of the Sociology of Language*, 256: 129–146. https://doi.org/10.1515/ijsl-2018-2015 (accessed 03 October 2022).

Webster, L. (2022), '"Erase/Rewind": How Transgender Twitter Discourses Challenge and (Re)politicize Lesbian Identities, *Journal of Lesbian Studies*, 26 (2): 174–191. https://doi.org/10.1080/10894160.2021.1978369 (accessed 03 October 2022).

Wilkinson, M. (2019), '"Bisexual Oysters": A Diachronic Corpus- based Critical Discourse Analysis of Bisexual Representation in *The Times* between 1957 and 2017', *Discourse & Communication*, 13 (2): 249– 267.

Wilkinson, M. (2020), 'Discourse Analysis of LGBT Identities', in E. Friginal and J. Hardy (eds), *The Routledge Handbook of Corpus Approaches to Discourse Analysis*, 554–570, London: Routledge.

Zimman, L. (2009), '"The Other Kind of Coming Out": Transgender People and the Coming Out Narrative Genre', *Gender and Language*, 3 (1): 53–80.

Zimman, L. (2019), 'Trans Self-identification and the Language of Neoliberal Selfhood: Agency, Power, and the Limits of Monologic Discourse', *International Journal of the Sociology of Language*, 256: 147–175.

Zimman, L. and Hayworth, W. (2020a), 'Lexical Change as Sociopolitical Change in Trans and Cis Identity Labels: New Methods for the Corpus Analysis of Internet Data', *University of Pennsylvania Working Papers in Linguistics*, 25 (2): Article 17. https://repository.upenn.edu/pwpl/vol25/iss2/17 (accessed 03 October 2022).

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders*, 213–241. London: Bloomsbury Academic.

Zimman, L. and Hayworth, W. (2020b), 'How We Got Here: Short-scale Change in Identity Labels for Trans, Cis, and Non-binary People in the 2000s', *Proceedings of the LSA,* 5 (1). https://doi.org/10.3765/plsa.v5i1.4728 (accessed 03 October 2022).

Zottola, A. (2021), *Transgender Identities in the Press: A Corpus-Based Discourse Analysis,* London: Bloomsbury.

Hekanaho, L., Hiltunen, T., Palander-Collin, M. & Hotti, H. (2024). Exploring self-identification and the functions of the identify as construction in the LGBTQ+ Reddit Corpus. In Laippala, V. & Coats, S. (eds.). *The March of Data: Linguistics across Disciplinary Borders,* 213–241. London: Bloomsbury Academic.

Appendix A

Table 9. Subreddits included in the *LGBTQ+ Reddit Corpus*

	Subreddit	texts		Subreddit	texts
1	r/lgbt	186943	27	r/androgyny	4585
2	r/asktransgender	140711	28	r/questioning	4354
3	r/actuallesbians	111953	29	r/LGBTindia	3907
4	r/ftm	51819	30	r/NonBinaryTalk	3473
5	r/ainbow	51500	31	r/transprogrammer	3022
6	r/bisexual	44039	32	r/queer	2974
7	r/LGBTeens	33424	33	r/sapiosexual	2092
8	r/transgender	28811	34	r/Nonbinaryteens	2077
9	r/gay	28162	35	r/lgbt_superheroes	2003
10	r/asexuality	18986	36	r/transgenderteens	1997
11	r/transtimelines	18297	37	r/bigender	1464
12	r/genderqueer	11514	38	r/LGBTAustralia	1440
13	r/NonBinary	10293	39	r/DualGender	1060
14	r/MtF	9928	40	r/GenderFluxx	885
15	r/trans	9581	41	r/LGBTireland	744
16	r/demisexuality	9522	42	r/UKLGBT	633
17	r/gaysian	9382	43	r/nonbinaryUK	535
18	r/Asexual	7753	44	r/LGBTQ_AnimalCrossing	525
19	r/TransSpace	7413	45	r/demigender	376
20	pansexual	7269	46	r/queerottawa	200
21	r/agender	6857	47	r/lgbtstudies	115
22	r/ennnnnnnnnnnbbbbbby	6428	48	r/lgbtnyc	109
23	r/BisexualTeens	5746	49	r/bisexuality	37
24	r/genderfluid	5618	50	r/GLBTChicago	11
25	r/AskLGBT	5197	51	r/greygender	9
26	r/comingout	4975			

Table 10. Subreddits included in the Analyzed Sample (n=554)

	Subreddit	texts		Subreddit	texts
1	r/lgbt	94	19	r/DualGender	6
2	r/asktransgender	80	20	r/androgyny	5
3	r/actuallesbians	52	21	r/genderfluid	5
4	r/bisexual	42	22	r/queer	5
5	r/genderqueer	35	23	r/trans	5
6	r/agender	27	24	r/TransSpace	5
7	r/ainbow	24	25	r/demigender	4
8	r/asexuality	21	26	r/sapiosexual	4
9	r/AskLGBT	18	27	r/comingout	3
10	r/pansexual	18	28	r/gay	3
11	r/ftm	16	29	r/GenderFluxx	3
12	r/NonBinary	16	30	r/BisexualTeens	2
13	r/questioning	12	31	r/bigender	1
14	r/demisexuality	10	32	r/lgbtnyc	1
15	r/LGBTeens	9	33	r/MtF	1
16	r/NonBinaryTalk	9	34	r/Nonbinaryteens	1
17	r/transgender	9	35	r/UKLGBT	1
18	r/Asexual	7			

[1] Word Sketches are essentially lists of collocations categorized by grammatical relations and are available through Sketch Engine. For more information, see https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/.
[2] We employ the acronym LGBTQ+ in the broadest sense to refer to all gender and sexual minorities.
[3] For the purposes of this study, we do not differentiate between sexual identity and orientation (see e.g., Andler 2022 and Dillon, Worthington and Moradi 2011), nor between romantic and sexual orientation (see e.g., Li, Sham and Wong 2022), but instead focus on social identities that may encompass aspects of all of these.
[4] There has been some debate over the disciplinary status of *Language and Sexuality*. Here we follow Mullany and others in treating this area of research as tightly linked with *Language and Gender* research while recognizing that this relationship is inherently complex (2010: 237); this is evident in our analysis of the discussion forum data.
[5] Spelling was normalized for tokens, e.g., 'non-binary' and 'nonbinary' were both coded as 'nonbinary'.
[6] Some concordances could have more than one function; double-coding was pursued at first, however, since this resulted in only a dozen double-codes, only the most salient category was retained.
[7] Brackets indicate the quote has been altered (during the anonymization process).
[8] There were some variations of the acronym in the data (e.g., LGBT, LGBTIQ), which we have normalized as LGBT. The acronyms FTM (female-to-male) and MTF (male-to-female) appeared rarely.
[9] In two cases, the label *demi* was used and it could not be specified from context whether the reference was to *demisexual*, *demi-girl* or *demi-boy*, hence coded as *ambiguous*.

[10] The corpus includes data from two large advice-subreddits (r/asktransgender and r/asklgbt, see Appendix A), explaining the prevalence of this category.
[11] Although the subreddit r/comingout was included in data collection, it only provided three concordance lines in the analyzed sample (n=550).