Hunter Youngquist

Analyzing Binary Relationships of Identity Labels Using Distributional Semantic Models

A Critical Queer Linguistic Analysis of an English Language Subreddit

Abstract

Following the shift towards quantitative, corpus-based analysis in queer linguistics, I examine the usage of identity labels to explore the binary relationships and predicted normative effects in the case of the online community r/lgbt, a subreddit dedicated to minority identity labels and discussion.

I analyze the distribution of the most frequent identity labels of the subreddit in a 2-year period with distributional semantic models, vector-based matrices that capture word distributions as numeric representations, showing evidence for various binaries that co-construct each other within the corpus. Additionally, I utilize concordances and collocations to examine the discourses surrounding gender and sexuality in the comments and submissions subcorpora, showing a more queer-aligned perspective in the former and a label-searching perspective in the latter.

Finally, the results from these techniques demonstrate the overall complex relationships between the many types of labels currently in use and between the subreddit users and their feelings about adopting specific labels to describe their identities.

Keywords: *queer linguistics, critical discourse analysis, Distributional Semantic Models, binaries, identity labels, gender & sexuality*

In recent years, queer linguistics (QL), a relatively new paradigm in studying marginalized queer identities through a critical perspective, has begun to operationalize quantitative techniques from the related field of corpus-based discourse analysis (CBDA) to examine the normative effects present in specific contexts (Santonocito 2020; Motschenbacher and Stegu 2013; Baker 2008). In this paper, I follow this shift toward quantitative analysis with the aim of studying the usage of identity labels in an LGBT subreddit through the results of

distributional semantic models (DSM). These models, adopted from natural language processing, are built on vector-based word distributions and have been applied to critical discourse analysis in the past decade to study lexical distributions in large corpora (LaViolette and Hogan 2019; Bruchansky 2017; Brigadir, Greene, and Cunningham 2015; Peirsman, Heylen, and Geeraerts 2010).

Using this novel technique, I show that the usage of the topmost frequent identity labels within the subreddit shares a close relationship both syntagmatically and paradigmatically with their prototypical binary, providing evidence for mutual co-construct and reification. In contrast, the label 'lesbian' and plurisexual labels, such as 'bi' and 'pan,' do not appear in a close semantic space with their typical binaries, hinting at potentially different relationships and distributions.[1] As a result of these strong binaries and the either/or selection they impose, I also explore the discourses surrounding the comment and the submission subcorpora, where acknowledgement of the inadequacy of labels is marked in the former and the difficulties in adopting constrictive, concrete, and reified binary labels to explain complex internal identities is demonstrated in the latter.

1 Previous Research

1.1 Queer linguistics and critical discourse analysis

Queer linguistics is a critical discipline that seeks to analyze normative structures built through the ways that language interacts with gender, sexuality, race, class, and other inequalities with the ultimate goal of deconstructing the power dynamics caused by binary structures (Leap 2015, 661). The inception of this field was influenced by Judith Butler's work in the late 80s and early 90s in which she emphasized the performative nature of gender and sexuality by considering them as performances rather than essential aspects of identity (1993; 1991; 1990; 1988). This perspective sheds light on the fluid nature of these facets of human identity across individuals, cultures, and history. Crucial to this study, QL rejects the notion of inherent identity labels as a result of this performative perspective and aims to deconstruct how they legitimize a normative ideal and shun those who do not fit within their borders.

The effects of identity labels are discursively constructed and the by-product of a binary system. A strong binary relationship is predicted to cause rigid, normative conceptions of what it means to be a specific identity, as individuals must fit idealized criteria (Motschenbacher 2013). In turn, this may cause conflict as individuals attempt to negotiate their complex identity with the

[1] I will use single quotes around a word to denote a calculated word vector, i.e. 'lesbian' referring to a specific vector that was extracted from a model.

need to assume a label and its pre-established conceptions. In recent years, the quantity of identity labels and the desire for inclusion has led to the usage of large acronym strings, such as LGBTQIA+ which stands for lesbian, gay, bisexual, transgender, queer, intersex, and asexual with the plus sign as a placeholder for other identities. Within this new paradigm, these identity labels are still subject to normative forces as a new default, the LGB portion of the acronym, is established (Oakley 2016, 9).

Due to the critical nature applied to the identification and deconstruction of binaries and the norms that they establish, techniques from corpus-based critical discourse analysis can be integrated with the field of queer linguistics and have been adopted by various studies in the last few decades (Santonocito 2020; Milani 2013; Bachmann 2011; Baker 2006b, 2005). Corpus-based critical discourse analysis is the application of quantitative corpus techniques, such as collocations and concordances, to the study of discourse from a critical perspective. These various and often competing discourses are identified and examined to uncover how they construct and maintain social and power structures. While specific terminology may differ, both fields are concerned with the evaluation and subversion of power structures within society and how they are spread through language and semiotic systems. Additionally, both can benefit from adopting the quantitative techniques of corpus linguistics to challenge common criticisms aimed at critical research such as biases in research and arguments that lack quantifiable evidence (Santonocito 2020, 190).

1.2 Queer identity labels in the online space

The advent of the internet and social media provides a new source to analyze the ways in which queer people express their identity through labels and how they present and discuss them with others. Two studies from different research fields have examined the relationship between online users and gender and sexuality labels, providing evidence for overall positive experiences. From an educational perspective, Lucero utilized an anonymous online survey to gauge the social media experiences of queer youth from various locations (although predominately from Texas) and found that the participants considered social media to be a safe space to perform and explore sexuality and gender (2017, 117).

In another work that assumes a critical discourse perspective, Oakley examined the usage of identity labels in free-form bios and "about me" pages on the social media platform Tumblr, arguing that despite being entrenched in the hegemonic power structures of binaries and their normative effects, the usage of explicit labels in four categories (gender, pronouns, sexual

attraction, and romantic attraction) allowed for nuanced expression of self-identity and ease of communication to both new users and those outside of the community (2016).

While limited in quantity, the results of these previous studies have shown that the online space can be considered a powerful and safe tool for queer people to explore their identities, even when the usage of identity labels is still subject to the normative effects of binary labels.

1.3 Distributional semantic models

Distributional semantic models, or otherwise known as vector space models, are mathematical representations of words based on their statistical co-occurrences. This methodology of constructing lexical word meaning is derived from the distributional hypothesis developed by Harris, which posits that similarity in the context of two words correlates to a similarity in meaning (Lenci 2018; Clark 2015). Due to the corpus-based nature, these models can be adapted to construct lexical meaning in specific discourse contexts, a feature that has been employed in various discourse studies throughout the last decade (LaViolette and Hogan 2019; Bruchansky 2017; Brigadir, Greene, and Cunningham 2015; Peirsman, Heylen, and Geeraerts 2010).

Despite their advantages, DSMs can be built using many combinations of parameters and vector-construction techniques that often alter the type of semantic relations that are extracted. Additionally, the meaning that is derived from DSMs is typically described as semantic similarity and relatedness, two concepts that, while concise and attractive, are quite abstract and vague, encompassing a myriad of other lexical relationships such as hypernymy, hyponymy, synonymy, and antonymy (Lenci 2018; Erk 2012). In efforts to demystify the results of these models, Sahlgren attributed two specific model types (word-by-word and word-by-document) to capturing syntagmatic and paradigmatic relations (2008). Following his work, I constructed both word-by-word and word-by-document models to analyze the results in terms of syntagmatic and paradigmatic relationships between nearest words. All other parameters will be discussed in the methodology section, as they have a greater impact on the performance of the models rather than the semantic relationships that they capture.

2. Materials and Methodologies

2.1 Reddit and r/lgbt subreddit

Reddit is a social media platform that consists of community-driven forum threads based around specific topics, called subreddits. From a discourse analysis perspective, various studies have been performed on target subreddits (Desmarais 2020; LaViolette and Hogan 2019). These subreddits encompass a wide variety of topics, including spaces for minority groups as in the

target of this study, the r/lgbt subreddit. Additionally, subreddits are managed by moderators, who establish and enact the rules for their specific subreddit. For individual users to post, they are required to sign up with an email address, creating an account with a unique username (Desmarais 2020, 37). Another significant aspect of Reddit is the relative anonymity that exists for users; personal information is not required for sign up, and it is normalized to use a customized version of the Reddit mascot, Snoo, for profile pictures.

2.2 Corpus and data collection

The data for this study was obtained from the r/lgbt subreddit during the period between July 1, 2020 and July 1, 2021 using Python and the PushShift Archive, a service that collects and archives reddit posts and comments (Baumgartner et al. 2020). The data was additionally cleaned of posts that contained zero text (i.e., photos or videos), posts that were subsequently deleted by users or moderators resulting in [deleted] or [removed] tags, and non-relevant data columns.

In terms of the ethical nature of this research and the data collection, the interaction between internet research with aspects such as consent, risk, and the public/private distinction is complex, making it difficult to establish strict guidelines (Orton-Johnson 2010). Within the last decade, there have been conflicting views on what constitutes public within the online space. Roberts (2015) argues that there are no ethical issues if online spaces are considered as a public setting. Thus, as the data is readily available through searching and without the need for registration, Reddit can be regarded as a public space (Desmarais 2020).

Despite this, I emphasize the highly sensitive nature of the data from this subreddit, consisting of personal information relating to gender and sexuality. Therefore, all data that was collected went through an anonymization process before being saved to the corpus. As an additional step to this process, I will only analyze the data from an aggregated perspective, avoiding the usage of specific posts and long concordance lines that could allow one to find the post on Reddit.

Corpus	Total Posts	Total Tokens	Average token per post
Corpus	462,424	14,803,992	~32
Submissions	41,433 (~9%)	4,841,387	~117
Comments[2]	420,991 (~91%)	9,962,605	~24

Tab. 1: Corpus and subcorpora statistics

[2] The original data pull from the subreddit included around 1,040,391 comment posts in total. However, due to size limitations in processing and storing the data, 500,000 were randomly removed.

As shown in Tab. 1, the final corpus contains 462,424 posts and 14,803,992 word tokens with an average of 32 words per post. To allow for a more fine-grained analysis, I split the corpus into two subcorpora: submissions and comments. While all posts share similar features on Reddit, they can be categorized as submissions, posts that are submitted to the subreddit and posted on its main feed, and comments, posts that are aimed at a particular submission and are only visible under that specific submission thread. Submissions only make up 9% of the total posts but are almost three times as long, with an average token per post of about 117; thus, submission posts consist of around 1/3 of the linguistic material with 4,841,38 total tokens. In contrast, comments are much shorter at an average of 24 tokens per post but 10 times as numerous at 91% of the total posts. These large discrepancies between the statistics of the two subcorpora can be attributed to their function within the subreddit. Anecdotally, submissions tend to be longer questions, stories, and discussion starters, while comments are shorter, and discussion based.

2.3 Distributional semantic models and corpus tools

The parameters that I adopted to construct the two models can be seen in Tab. 2. Both DSMs were constructed using the R programming language, along with the quanteda and wordspace packages (Benoit et al. 2018; Evert 2014). I additionally preprocessed the models, transforming all words to lowercase and removing symbols, URLs, punctuation, and numbers.

The key parameters of DSMs discussed in the literature can be categorized in two major groups, the ones that determine how to initially construct the word vectors, such as context type, context window, and learning methodology, and those that are applied to transform the constructed model, such as vector dimensionality reduction, weight scores, transformation techniques, and similarity metrics.

Parameters	**Word-by-word Model**	**Word-by-document model**
Context Type	Word Vectors	Document (Posts) Vectors
Context Window	2 words two the left and right	Entire reddit post
Context Learning	Count	Count
Vector Dimensions	138,023 by 138,023	20,764 by 462,424
Weight Score	Simple Log-likelihood	Simple Log-likelihood
Transformation	Log	Log
Similarity Metric	Cosine	Cosine

Tab. 2: A Summary of the hyper parameters of each model

Starting with the parameters used to construct the models, the context type refers to the features that are utilized to define a word's meaning, which, in the case of this paper, include context vectors as other words (word-by-word model) and context vectors as regions (word-by-document model). The context window is the range around a target word in which other words are counted to build the context vector, and for this reason, it is particularly important for the word-by-word model. Finally, the learning methodology is the method used to build the vectors and typically can be considered either count or prediction. For this study, I opted to use the count method which means that each context vector is built from counting co-occurrences.

Despite this method having worse results compared to prediction models when tested in specific semantic tasks, it has been more thoroughly studied and the results are therefore more interpretable (Lenci et al. 2021; Baroni et al. 2014).

Moving to the techniques used to transform the matrices, the first, vector dimensionality reduction, is often applied to models to reduce the amount of feature vectors used to describe each word, revealing latent semantic features and improving model performance (Lenci 2018; Lapesa and Evert 2014). These models are now called implicit due to their abstract and uninterpretable features. For this reason, I have kept the vectors explicit to preserve interpretability at the cost of model performance. The next techniques, feature weighting and transformation, are applied to the individual entries of each feature, which are typically the raw co-occurrence frequencies in count models, adjusting the scores by weighting them against the global statistics of the model matrix. Following the results of Lapesa and Evert, the models of this study have been weighted using simple log-likelihood with a log transformation (2014). The last parameter to discuss is the selection of a method to measure similarity between two-word vectors within the model space. The most common similarity method and the one that I have decided to apply to the model is the distance measure cosine which also has been shown to have good model performance in various semantic tasks (Kiela and Clark 2014; Lapesa and Evert 2014).

In addition to the models, I also employ collocations and concordances to analyze further the context of the model results. These were obtained using the software SketchEngine, which adopts LogDice as the statistical score to measure the correlative strength of collocations. An important feature of logDice is that it ranges from 0-14, with 14 being the strongest possible correlation.

3. Results

3.1 Paradigmatic and syntagmatic models

To analyze the results of the models, I extracted the top five most frequent label embeddings in the subreddit corpus, 'trans' (36,759), 'gay' (36,120), 'bi' (19,613), 'straight' (18,818), and 'lesbian' (12,080). I also included the ten closest word embeddings to each label based on their cosine score for the paradigmatic model and the syntagmatic model, which can be seen in Tab. 3 and Tab. 4 respectively. Before I examine the nearest neighbors, I would like to highlight that the topmost frequent labels to occur within the subreddit are the labels that constitute the letters of the subreddit name, LGBT. This demonstrates the greater visibility of these labels and the effects of the subreddit name and title on the topic of discussion that users engage in. There is an additional methodological motivation for examining only the more frequent identity labels as more frequent words lead to more stable word embeddings; In fact, during the analysis of the embedding for 'lesbian,' which has the lowest frequency of the top 5, the resulting nearest neighbors have high scores and high ranks, showing that the word embedding occurs in a unique model space potentially due to its lower frequency. For this reason, it is apparent that this top five is a sufficient cutting-off point, as labels with lower frequencies may display other peculiar effects.

Examining the results of the paradigmatic model in Tab. 3, the first thing to note is that all the nearest neighbors are other identity labels. Due to the nature of paradigmatic relations, this is unsurprising as it relates to linguistic features that occur in similar slots and that can be effectively substituted. Despite this, examining which labels are closer in vector position shows that three out of the five most frequent identity labels have their binary counterparts in the first or second nearest neighbor with a low rank or high mutual similarity. This includes 'cis' as second nearest to 'trans,' 'straight' as the nearest to 'gay,' and 'gay' as the nearest to 'straight.' One feature of DSMs is that they are unable to distinguish synonyms and antonyms (Lenci, 2017); for example, words such as 'good' and 'bad' may be rated as near as 'good' and 'great,' despite being semantically opposites. This explains why the top two nearest neighbors for 'trans' includes the non-shortened version, 'transgender,' and its normative opposite 'cis.' Additionally, the high cosine and rank of 'queer' as a nearest neighbor to 'gay' may show the increasing popularity of using 'queer' as a less reified synonym for 'gay.' The results of two binary labels that are often defined as opposing and mutually exclusive being the nearest neighbors of each other demonstrates that the paradigmatic usage of these labels within the subreddit follow this opposition, providing evidence for a strong relationship caused by a co-constructed binary.

Trans		Gay		Bi		Straight		Lesbian	
Term	Score	Term	Score	Term	Score	Term	Score	Term	Score
Trans gender	66 (1)	Straight	68.6 (1.5)	Pan	61.8 (1)	Cis	65.8 (1.5)	Bi sexual	68.9 (7.5)
Cis	69 (3.5)	Queer	73.2 (3.5)	Bisexual	63.7 (2)	Gay	68.6 (1.5)	Bi	69.9 (6.5)
Non-binary	69.4 (11)	Lesbian	73.3 (9)	Pan sexual	67.7 (10.5)	Cishet	68.6 (2)	Pan sexual	70.7 (14)
Non binary	69.5 (12)	Trans	73.4 (7)	Non-binary	67.9 (9)	Hetero	69.6 (3)	Pan	71.6 (14.5)
Nb	69.7 (11)	Bi	73.6 (17.5)	Nb	69 (8.5)	Hetero sexual	70.3 (6)	Gender fluid	71.9 (25.5)
Bi	70.5 (9.5)	Bisexual	75.5 (40)	Gender fluid	69.1 (18.5)	Cis gender	70.9 (7.5)	Non-binary	72.2 (25)
Queer	72.1 (5)	Homo sexual	75.9 (12.5)	omni	69.3 (6.5)	Bi	71.6 (12.5)	Asexual	72.3 (23.5)
Straight	72.2 (8)	Cis	76.2 (29.5)	Ace	69.4 (7)	Trans	72.2 (8)	Demi sexual	72.7 (30)
Cis gender	72.7 (12.5)	Hetero	76.3 (13)	Asexual	69.4 (69.4)	Bi sexual	73.5 (25.5)	Ace	72.8 (16)
Gay	73.4 (7)	Hetero sexual	77 (30.5)	Non binary	69.8 (16)	Lesbian	73.9 (15.5)	Sapphic	72.9 (6)

Tab. 3: The nearest neighbors from the word-by-word model (syntagmatic) with cosine score and rank in parenthesis

In contrast to these three labels, the top 10 nearest neighbors of labels 'bi' and 'lesbian' did not include their predicted binary of 'straight.' The label of 'bi' instead is highly similar in both cosine and rank to the label 'pan,' as well as the un-abbreviated form, 'bisexual.' This could be caused by a distinction that is formed between monosexual labels and plurisexual ones, such as bi or pan, which constitute a distribution and semantic space that is less related to that of monosexual identities.[3] Interestingly, the vector for 'lesbian,' a monosexual label, also shares this space, having terms for bisexuality and pansexuality high on its list ('bi,' 'bisexual,' 'pan,' and 'pansexual'). However, this is a non-mutual relationship, as the rank score is relatively high for all of its nearest neighbors. The lack of mutual similarity between 'lesbian' and both

[3] Plurisexual refers to an attraction to multiple genders as opposed to a single gender.

monosexual and plurisexual labels could have been caused by the low overall frequency of the term.

Turning to the word-by-document model in Tab. 4, the results are the nearest neighbors that have a syntagmatic relationship with the label and occur in similar contexts, i.e., similar posts within the subreddit. Unsurprisingly, we can see potential collocates such as the vector for 'people,' 'women,' and 'men' which are used to refer to groups such as "trans people," "straight women," or "gay men." Additionally, there is topically related terminology, including 'dysphoria' for 'trans' and 'attracted' for 'bi,' 'straight,' and 'lesbian.' Beyond these expected vectors, the word-by-document model provides support for the reification of binaries for the vectors 'trans,' 'gay,' and 'straight,' as their prototypical binaries still hold top positions with very low ranks. Due to these binaries appearing in similar posts and contexts, this syntagmatic evidence highlights the co-constructive nature of these labels within the subcorpus.

Trans		**Gay**		**Bi**		**Straight**		**Lesbian**	
Term	Score	Term	Score	Term	Score	Term	Score	Term	Score
People	77.4 (1)	Straight	80 (1)	Pan	71.7 (1)	Gay	80 (1)	Lesbians	80.6 (1)
Cis	77.4 (1.5)	Men	81.5 (4)	Attracted	81.1 (5)	People	80.9 (2)	Women	82.3 (5)
Women	77.9 (3)	People	82.3 (8)	Genders	81.2 (4)	Sexuality	82.8 (3.5)	Bi	83.1 (4.5)
Woman	79.8 (3.5)	Bi	83.5 (7.5)	Attraction	82.4 (9)	Men	83 (8.5)	Men	83.2 (9.5)
Men	80.6 (4)	Lesbian	83.8 (6)	Omni	82.7 (4.5)	Cis	83.1 (6.5)	Attracted	83.4 (14)
Trans phobic	80.8 (3.5)	Man	84 (5.5)	Lesbian	83.1 (4.5)	Bi	83.4 (7.5)	Bisexual	83.5 (7)
Gender	81.86 (14)	Just	84.2 (17)	Men	83.3 (11.5)	Women	83.5 (10.5)	Gay	83.8 (6)
Man	82.5 (6)	Homo phobic	84.3 (7)	People	83.3 (18)	Attracted	83.8 (16.5)	Woman	84.4 (9.5)
Person	82.5 (5)	Like	84.6 (23)	Straight	83.40 (7.5)	Just	84.3 (20)	Girls	84.57 (11.5)
Dysphoria	82.5 (7)	Think	84.8 (16.5)	Women	83.5 (12.5)	Gay	84.3 (7.5)	Label	84.9 (12.5)

Tab. 4: The nearest neighbors from the word-by-document model (paradigmatic) with cosine score and rank in parenthesis

Analyzing the results of outliers from the word-by-word model, 'bi' and 'lesbian,' there are similar results from the word-by-document model, with the predicted binaries either occurring low on the list, as 'straight' in ninth place for 'bi,' or not at all, as in 'lesbian.' Due to the lack of context that DSMs provide, it is now necessary to utilize other corpus techniques, such as collocations and concordances, to further investigate these two outliers and to examine the discourses surrounding identity labels and their binaries.

3.2 Discourses surrounding identity labels

The models have shown a close distributional relationship between typically binary identity labels, both paradigmatically and syntagmatically. However, as the example of the relationship between the labels 'bi' and 'pan' highlights, there are two major issues when analyzing the results of DSMs: their inability to distinguish synonyms and antonyms and their lack of extended context. While this may not be crucial for the interpretation of terms that are conventionally considered binaries (i.e., gay and straight), it raises concerns for less clear relations, such as that of bisexual and pansexual. In order to alleviate these downfalls and further analyze the various discourses surrounding the binaries within the subcorpus, I will employ collocations to examine a key construction that exist in the corpus that supports a binary conception of labels. Additionally, I will examine the comments and submissions subcorpora to further contextualize the various conflicting discourses that underlie the discussion of identity within the subreddit.

3.2.1 "Either/or" discourse

One strong collocation that is present when examining the topmost frequent identity labels occurs in the form of "Identity label or other label." Tab. 5 shows the collocations that follow this structure with the highest logDice scores. This collocation is a predicted side effect of strong binaries, as binary labels stand in mutual opposition, requiring people to select one or the other while discouraging the messy, granular nature of identity.

An interesting addition to this set of collocations is "bi or pan" with the highest logDice score of 12.2. This could mean that a new binary relationship is forming between these plurisexual identities. However, similarities in the definitions of these two labels suggest that confusion in deciding between these two nuanced labels is the source of this strong collocation. According to Hayfield and Karolína, the classic definition of bisexual, "attraction to both genders," may be shifting towards the more inclusive definition, "those of my own gender and other genders," which overlaps more with the definition of pansexual as "romantic and sexual attraction to all

genders, or regardless of genders" (2021). Despite the lack of certainty as to whether these two labels co-construct like traditional binaries, the strong "either/or" collocation further demonstrates the normative requirements of identity labels as users must decide between which of the two best describes their own complex identity.

Collocate	Frequency	LogDice
"bi or pan"	432	12.2
"cis or trans"	283	11.6
"gay or straight"	435	11.2
"lesbian or bisexual"	130	10.5
"trans or non binary"	70	9.7
"straight or bisexual"	52	9.0
"gay or queer"	46	8.2

Tab. 5: Collocates following the "Identity label or other label" pattern

3.2.2 Complex views of labels

Regardless of the evidence for a strong relationship between binary labels and either/or collocations within the subreddit, users of the comment section also acknowledge the natural complexity of gender, sexuality, and identity, and how labels are flawed in portraying this. Strong collocates with the terms "gender" and "sexuality" demonstrate this complex view of personal identity. For example, both terms collocate with "spectrum," resulting in phrases such as "sexuality is a spectrum" (85, 11.7) and "gender is a spectrum" (84, 10.9). Additionally, "sexuality" collocates with "fluid" (171, 12.4) and "gender" with "construct" (156, 11.9). These collocates exemplify that there is an underlying discursive idea in which the users of the subreddit consider the concepts of sexuality and gender to be non-binary spectrums that are socially constructed and subject to change.

In addition to this more queer-aligned conception of gender and sexuality, there is an emphasis on the idea that labels are tools that can be assumed by users to help describe their identity. The term "label" collocates with the verb "fitting" (69, 10.1), emphasizing the users emotional state in selecting an identity label. Additionally, the verb "choose" collocates strongly with "identity," forming the commonly used phrase of "choose to identify as" (80, 10.2) suggesting a choice-based perception of identity labels.

3.2.3 Desire to have a label

The collocations of the previous section illustrate how the comments of the r/lgbt subreddit have a view of sexuality and gender that contrasts with the strong relationship between binaries that resulted from the distributional models, contradicting a more traditional, binary-driven perspective that would be expected to cause such strong either/or effects. Significantly, if there is a general view that sexuality and gender are complex and that labels may be insufficient in describing one's own identity, why are binary labels syntagmatically and paradigmatically close within the semantic space, and why is there a strong either/or collocation with the most frequent labels? Through analyzing further collocations and aggregated concordances, I will now show that this contradiction is driven by the tension caused through the desire of submission users in the subreddit to discover the identity label(s) that most accurately describe themselves and the knowledge that identity is complex and may never be fully explained by labels. Due to the smaller size of the submission subcorpus and the more complex syntactic structures of the target discourse, the results of this section are based on lower frequencies, slightly weaker logDice scores, and qualitative concordance analysis.

First, there are constructions that provide evidence for the difficulties that submission users have in adopting an appropriate identity label to describe their personal identities. Examining the concordance lines for the collocation, "put" + "label" (76, 9.0), the majority relay the difficulties submission posters have in putting labels on themselves. The most common sentiment that I found was an inability to put a label on their identity at 23 of the 76 collocates. Additionally, other common sentiments were the following: those that felt hesitant to put a label on themselves (4), that were asking the forum for label advice (6), that felt they needed to find a label (4), and that found specific labels felt wrong (3). In addition to these more negative or confused sentiments, there were also 8 cases of users who connected well with a label or wanted to find a label to feel better.

Beyond this collocation, various n-grams further depict the difficulties of users in finding the correct label. The lemmatized n-gram "do not fit" occurs 214 times and concordance lines show that this phrase occurs often with users declaring that labels do not fit correctly, with around 56 out of 214 lines having this sentiment.

There is also a more complex structure of the form "I know labels (slot 1), (slot 2)." In this construction, slot 1 is filled with a phrase that acknowledges the shortcomings of labels such as "labels are not important" or "labels are not for everyone." Then, in slot 2, they contrast the previous statement with the desire to have an appropriate label with phrases such as "I would like one" or "I want to be." The primary collocate that I used to locate this construction was

"know" + "label" with a frequency of 61 and a logDice score of 7.3. Of the 61 instances of this collocate, 43 follow this construction. While the logDice is weaker in terms of score, the more complex syntactic structure of this construction naturally occurs less frequently; therefore, in relation to this complexity, along with the size of its raw frequency within the smaller subcorpus and its appearance in 43 distinct posts, this suggests a reoccurring pattern within the submission subcorpus.

4. Closing remarks

The overall picture that is apparent from the various methodologies applied in this study is a complex and often conflicting portrayal of the usage of identity labels within the subreddit. On one hand, the results of the two DSMs have shown that the most frequent identity labels in the subcorpus, 'trans,' 'gay,' and 'straight,' are geometrically similar in both paradigmatic and syntagmatic contexts, providing evidence for the strong co-construction and mutual reification of these labels caused by their usage in the corpus. In contradiction to these strong binaries, the comment section of the subreddit contained various collocations that suggested a more queer-aligned perspective on the role of gender and sexuality labels, highlighting that gender and sexuality are a "spectrum" and labels are just tools that can "fit." With this general acknowledgment of the limitations of labels, the submission posts contained patterns that demonstrated the tension caused by the binary effects of the labels on identity selection for users as they struggled to apply normative and restrictive labels to their complex inner identities.

Despite the strong relationships between binaries found for some of the identity labels, other labels such as 'lesbian' and 'bi' did not align with these predictions. More work done on the usage of these labels in similar and different contexts could shed light on their evolving nature and conception. In addition to these usage-based analyses, testing the inner perception of these labels through surveys and other methodologies would allow for a more holistic explanation on whether the strong relationships between both lesbian and plurisexual labels and 'bi' and 'pan' derive from new binaries or rather from confusion caused due to similarity in definition.

Finally, while the models constructed in this study were built to capture syntagmatic and paradigmatic similarity, other configurations could lead to different results in semantic similarity between terms. Therefore, further work is required to better understand the type of meaning extracted from models based on their parameters and how to best implement them for discourse analysts. Nevertheless, the usage of novel computational techniques, such as distributional semantic models, in the field of queer linguistics will continue to be a powerful

source of new tools to expand the quantitative and qualitative means of analysts in examining and deconstructing the discursive norms surrounding gender and sexuality.

Hunter Youngquist *is a PhD candidate at the University of Verona interested in the integration and application of techniques from corpus linguistics and natural language processing to the study of discourse and pragmatics, with a specific interest in queer linguistics, distributional semantics, and language annotation.*

Works cited

Bachmann, Ingo. "Civil Partnership—'Gay Marriage in All but Name': A Corpus-Driven Analysis of Discourses of Same-Sex Relationships in the UK Parliament." *Corpora* 6 (2011): 77-105.

Baker, Paul. *Public Discourses of Gay Men*. New York: Routledge, 2005.

---. *Sexed Texts: Language, Gender and Sexuality*. City: Equinox Pub., 2008.

---. *Using Corpora in Discourse*. New York: Continuum, 2006.

Baroni, Marco, et al. "Don't Count, Predict! A Systematic Comparison of Context-Counting vs. Context-Predicting Semantic Vectors." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics (2014): 238-247.

Baumgartner, Jason, et al. "The Pushshift Reddit Dataset." *ArXiv:2001.08435* (2020).

Benoit, Kenneth, et al. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3.30 (2018): 774.

Brigadir, Igor, et al. "Analyzing Discourse Communities with Distributional Semantic Models." *Proceedings of the 2015 ACM Web Science Conference*. Association for Computing Machinery (2015): 1-10.

Bruchansky, Christophe. "Political Footprints: Political Discourse Analysis Using Pre-Trained Word Vectors." *ArXiv:1705.06353* (2017). 1-7.

Butler, Judith. *Bodies That Matter: On the Discursive Limits of "Sex."* New York: Routledge, 1993.

---. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge, 1999.

---. "Imitation and Gender Insubordination." *The New Social Theory Reader*, 2nd Edition. London: Routledge, 2008. 13.

---. "Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory." *Theatre Journal* 40.4 (1988): 519-531.

Clark, Stephen. "Vector Space Models of Lexical Meaning." *The Handbook of Contemporary Semantic Theory*. Edited by Shalom Lappin and Chris Fox. Oxford: Wiley-Blackwell, 2015. 493-522.

Desmarais, Angela-Marie. *Men Who Knit: A Social Media Critical Discourse Study (SM-CDS) on the Legitimisation of Men within Reddit's r/Knitting Community*. Auckland: Auckland University of Technology, 2020.

Erk, Katrin. "Vector Space Models of Word Meaning and Phrase Meaning: A Survey." *Language and Linguistics Compass* 6.10 (2012): 635-653.

Evert, Stefan. "Distributional Semantics in R with the Wordspace Package." *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. Dublin City University and Association for Computational Linguistics (2014): 110-14.

Hayfield, Nikki and Karolína Křížová. "It's Like Bisexuality, but It Isn't: Pansexual and Panromantic People's Understandings of Their Identities and Experiences of Becoming Educated about Gender and Sexuality." *Journal of Bisexuality* 21.2 (2021): 167-193.

Kiela, Douwe and Stephen Clark. "A Systematic Study of Semantic Vector Space Model Parameters." *Proceedings of the 2nd Workshop on Continuous Vector Space Models and Their Compositionality (CVSC)*. Association for Computational Linguistics (2014): 21-30.

Lapesa, Gabriella and Stefan Evert. "A Large Scale Evaluation of Distributional Semantic Models: Parameters, Interactions and Model Selection." *Transactions of the Association for Computational Linguistics* 2 (2014): 531-546.

LaViolette, Jack and Bernie Hogan. "Using Platform Signals for Distinguishing Discourses: The Case of Men's Rights and Men's Liberation on Reddit". *Proceedings of the International AAAI Conference on Web and Social Media*. ICWSM (2019): 323-334.

Leap, William. "31 Queer Linguistics as Critical Discourse Analysis." *The Handbook of Discourse Analysis*. Edited by Deborah Tannen, Heidi E. Hamilton and Deborah Schiffrin. New York: John Wiley & Sons, 2015. 661-680.

Lenci, Alessandro, et al. "A Comprehensive Comparative Evaluation and Analysis of Distributional Semantic Models." *ArXiv:2105.09825* (May 2021).

---. "Distributional Models of Word Meaning." *Annual Review of Linguistics* 4.1 (2018): 151-171.

Lucero, Leanna. "Safe Spaces in Online Places: Social Media and LGBTQ Youth." *Multicultural Education Review* 9.2 (2017): 117-128.

Milani, Tommaso M. "Are 'Queers' Really 'Queer'? Language, Identity and Same-Sex Desire in a South African Online Community." *Discourse & Society* 24.5 (2013): 615-633.

Motschenbacher, Heiko and Martin Stegu. "Queer Linguistic Approaches to Discourse." *Discourse & Society* 24.5 (2013): 519-535.

Oakley, Abigail. "Disturbing Hegemonic Discourse: Nonbinary Gender and Sexual Orientation Labeling on Tumblr." *Social Media + Society* 2.3 (2016): 205630511666421.

Orton-Johnson, Kate. "Ethics in Online Research; Evaluating the ESRC Framework for Research Ethics Categorisation of Risk." *Sociological Research Online* 15.4 (2010): 126-130.

Peirsman, Yves, et al. "Applying Word Space Models to Sociolinguistics: Religion Names before and after 9/11." *Advances in Cognitive Sociolinguistics*. Edited by Dirk Geeraerts, Gitte Kristiansen and Yves Peirsman. New York: De Gruyter Mouton, 2010. 111-137.

Roberts, Lynne D. "Ethical Issues in Conducting Qualitative Research in Online Communities." *Qualitative Research in Psychology* 12.3 (2015): 314-325.

Sahlgren, Magnus. "The Distributional Hypothesis." *The Italian Journal of Linguistics* 20.1 (2008): 33-54.

Santonocito, Carmen. "LGBT* People in the Speeches of Italian and British PMs: A Corpus-Assisted Critical Discourse Analysis." *Critical Approaches to Discourse Analysis across Disciplines* 11.2 (2020): 187-212.