Analyzing Linguistic Variation Using Discursive Worlds

Analyzing Linguistic Variation Using Discursive Worlds

Abstract: Researchers in variationist sociolinguistics have long sought to develop social measures that are more sophisticated than demographic categories such as age, gender and social class, while still being useful for quantitative analysis. This paper presents one such new measure: discursive worlds. For each speaker in a corpus, their discursive world is operationalized through compiling a list of specific referents cited in their interview. These lists are then used to construct similarity spaces locating the speakers along dimensions that are discursively relevant in the corpus. Using common clustering algorithms, the corpus speakers are then partitioned into categories, and this partition can be used in statistical analysis. We show how this method can be used to analyze two lexical variables in the *Cartographie linguistique des féminismes* (CaFé) corpus, a corpus of francophone interviews with feminist and queer activists, for which, we argue, quantitative analysis using classic demographic categories is inappropriate.

1. Introduction

This paper presents a new method for analyzing patterns of linguistic variation from a variationist perspective. Ever since the pioneering work of Labov and colleagues in the 1960s and 1970s, researchers have shown that the use of one sociolinguistic variant over another in a large corpus of vernacular speech is often conditioned by social properties such as the gender, social class, age or race of the speakers in the corpus. This being said, from the very beginning of variationist sociolinguistics, many have expressed doubt with respect to how enlightening these large demographic categories actually are for understanding how people use language. For example, Sankoff & Laberge (1978) say, “our experience with the analysis of the Montreal French corpus leads to the realization that directly correlating linguistically variable behavior with social class

membership, whether defined stratificationally or dialectically, is not a well motivated procedure" (Sankoff & Laberge, 1978: 239). They say this because, as they observe, definitions of social class made without language in mind miss crucial distinctions that go beyond broad class categories, for example "ignor[ing] established facts such as that teachers, actors and receptionists tend to speak a more standard variety than other people of similar social or economic position" (ibid, 239). Instead, Sankoff & Laberge develop a measure based on operationalizing Bourdieu & Boltanski (1975)'s *marché linguistique* (linguistic marketplace), which is a sophisticated theory modeling the relationship between language and the social world. In a similar vein, Milroy & Milroy (1985, 1992) argue that studying individuals' social networks provides a better understanding of how they use and change language than abstract sociological concepts like social class. Other arguments about the need to rethink the use of demographic properties, particularly social class, came from studies applying the variationist method to speech communities other than large North American cities. For example, both Rickford (1986) and Eckert (1989) find that a social class measure could not even be meaningfully applied in a variationist study of an East-Indian sugar estate community or an American high school, and that the optimal analysis of linguistic variation in these speech communities relies on more local categories (*estate class vs non-estate class* or *jocks vs burnouts* respectively).

Although these reflections started with class, since the 1990s, there have been moves to likewise reconsider other demographic categories in linguistic variation such as gender (Eckert 1989, Cheshire 2002, Levon 2015, Eckert & Podesva 2021, Becker, Khan & Zimman 2022, among many others), age (Eckert 2017), and race (Rosa & Flores 2017, Charity, Mallinson & Bucholtz 2020, among many others). Although most variationist work continues to operationalize social categories in an uncritical way (see Meyerhoff & Ehrlich (2019) for discussion), one line of

quantitative research attempts to refine traditional categories to ones incorporating intersections between gender, trans-ness, sexuality, race, place and profession (see Podesva & Hofwegen 2016, Hazenberg 2017, King 2021, Becker et al. 2022 for some examples). However, the close proximity of scholars working on language and gender and/or race to critical studies have caused many to question whether employing static gender or racial categories in linguistic analyses are appropriate in the first place. Both constructivist and deconstructivist – or in the european ideological frame materialist and post-structuralist – approaches (Delphy 1977, Butler 1990, Wittig 1992, Guillaumin 1992) define gender or race as operations of classification, of categorization producing social relationships and their operating categories (or norms). These categories come into existence, among others, through discursive practices situated in ideological frames. Such categories do not drive linguistic variation but, rather, are themselves produced through language. These ideas are taken up, for example, in discursive approaches to sociolinguistics, which study how we do gender (West & Zimmerman 1987), race (Alim 2016) or sexuality (Cameron & Kulick 2003). A variationist investigation aiming to build on a more complex and ideologically refined approach to social categories could therefore benefit from the insights of constructivist, non-essentialist approaches in which the categories used in quantitative analysis of a sociolinguistic variable emerged from the discourses in the speech communities using that variable.

In this paper, we present a new method for obtaining categories for quantitative variationist analysis that does not assume a fixed content, meaning or homogeneity of social categories, and instead allows them to emerge from ideologically situated interactions, in this case, linguistic ones. Instead of analyzing patterns of variation using categories like social class, gender, race or age, we will use (what we call) *discursive worlds*. Loosely inspired by work on *worlds* by the French social theorists Boltanski & Thévenot (1991) as well as the notion of *discursive formations* (Pêcheux and

Fuchs 1975), we define a person's discursive world as their ideological structure, including figures, topics, values and objects that are salient and valued for them. In this framework, worlds are tantamount to orders of worth that are based on moral values distributing the worth of persons, objects and actions. This notion allows us to understand what critical and moral stance guides someone's actions, according to a specific axiology. Similar to how Sankoff & Laberge (1978) operationalized Bourdieu & Boltanski's linguistic marketplace for corpus research, we will present a method for operationalizing a corpus speaker's discursive world: we will take a speaker's discursive world to be characterized as the set of specific referents, i.e. proper names for people, places and things, that they utter in their interview. We hypothesize that the information obtained through looking essentially at what speakers talk about in an interview encodes as much, indeed in many cases even more, of the social information about them that is relevant for modeling linguistic variation than knowing their age, gender, social class or racial background, when this information is even available. We then show how the set of discursive worlds of a corpus can be transformed into n-dimensional spatial structures that encode similarity relations between the speakers by means of multi-dimensional scaling techniques (see Borg & Groenen 2005). Similar techniques have been previously used in variationist sociolinguistics to group corpus speakers according to their linguistic features (see Horvath & Sankoff 1987), and have been used to group linguistic features together in register studies (see, for example, Biber 1995); however, we show that these structures allow us to identify the relevant discursive dimensions that are found in the corpus in an empirically based "bottom up" way. We argue that these methods allow for a new picture of the relevant social distinctions in the corpus, not just the linguistic ones. Once we have obtained the similarity spaces from discursive worlds, we show how to use clustering algorithms (such as k-means clustering (MacQueen 1967)) to identify meaningful subgroups of speakers, and

these clusterings can then be used as factors in multivariate statistical analyses. These clustering techniques partition the discursive space into discrete groups (i.e. categories), but the partitions are not fixed. The flexibility of this method, we argue, allows for a treatment of the social categories used in quantitative analysis as dynamic and emerging from the discourses in the corpus itself. What we present thus represents a continuation of research programs in both classic "first wave" variationist sociolinguistics (see Eckert 2012 for a discussion), such as Sankoff & Laberge, and critical approaches to social categories, as seen in, for example, Bowker & Star (2000), Butler (1990, 1997) and Guillaumin (1992).

The second part of our paper illustrates how this method can be applied to study a series of lexical variables in Parisian French in the *Cartographie linguistique des féminismes (CaFé)* 'Linguistic cartography of feminisms' corpus (AUTHORS, 2024). CaFé is a corpus of 102 sociolinguistic interviews with feminist and queer activists in Paris (France), Marseille (France) and Montréal (Québec, Canada) that we collected in 2020 and 2021 order to study the contemporary discursive formations of feminism and queer. We argue that discursive worlds allow us to study variation in this corpus in a quantitative manner, all while providing new information about the social structure of feminist and queer activism in the northern francophone world.

The paper is set out as follows: in section 2, we present the CaFé corpus and discuss how both traditional and more nuanced approaches to gender, race and class are not optimal for studying the variation found within it. In section 3, we introduce *discursive worlds* and show how we construct discursive spaces based on them, following the procedure in Bendifallah, Abbou, Douven & Burnett (2023). We then show how to construct socio-ideological categories based on the spaces which can be used in quantitative analyses of linguistic variation. In section 4, we present two short quantitative studies of lexical variation in the Parisian subcorpus of CaFé:

violences faites aux femmes vs *violences sexistes et sexuelles* `violence against women vs sexual and sexist violence', and *prostitution vs travail du sexe* 'prostitution vs sex work'. We show that the discursive world measure conditions these variables, unlike many other social factors available for the corpus. Finally, section 5 concludes with a discussion of the usefulness of discursive worlds for the analysis of linguistic variation in more traditional variationist corpora.

2. The Cartographie linguistique des féminismes (CaFé) corpus

The CaFé corpus consists of semi-directed francophone interviews on themes related to feminist and queer activism and engagement. It was constructed within the context of the *REMOVED FOR REVIEW* project (ERC StG no XXXXX, PI: AUTHOR). It is composed of 102 90 minute interviews of people who are engaged in what we described as "feminism, women's issues and/or activism for queer and sexual rights" (*le féminisme, la cause des femmes, le queer, ou la lutte pour les sexualités*) in Paris (42 interviews), Montréal (40 interviews) and Marseille (20 interviews). In these interviews, we collected their positions on issues related to gender and sexuality, and their link with language. Although feminist activism is our primary focus, the entanglement of gender and sexualities both in the socio-history of social movements and in the scientific literature, made us group together feminism and queer activists as a way to get a more complete picture of the fight for issues around gender and sexuality, particularly in France. Because of this focus, interviews in the CaFé corpus were based around a questionnaire including questions about participants' biographical journeys, their discovery of feminism and their ideological positions related to a wide range of issues relevant to feminism, including the state, violence, (anti)racism, expertise, and age, generations, feminist linguistic practices, among others. For more information on the content of the interviews, see AUTHORS (2024).

In constructing the corpus, we considered it more desirable to take into account the structure of the communities that we are studying, rather than trying to balance demographic properties to help the statistical analysis. For example, it would be absurd to attempt to include an equal number of cis men as cis women in a corpus of feminist activists, given that cis male feminist activists are quite rare compared to, for example, cis female ones. We also considered balancing the corpus according to more locally relevant categories: as described by Bereni (2012) or Pavard, Francourt & Zancarini-Fournel (2020), francophone feminist activist communities have been structured by conflicts between groups identifying using labels corresponding to different theoretical or political orientations. These theoretical or political orientations are known as *courants* in French: *différentialiste* vs *non-essentialiste; matérialiste vs queer; universaliste vs intersectionnelle*, and many more. However, historical and sociological research (Bereni (2012), Pavard et al. (2020), Bendifallah et al. (2023)) has shown the labels used in feminist identification – and even the label "feminist" itself – have changed throughout time and are currently in flux. Therefore, instead of demographic properties or self-declared labels, we structured the corpus in terms of feminist/queer *practices*, which are varied and directly observable, with the hypothesis that different types of practice situate the speakers in different discursive worlds. We therefore recruited speakers based on their type of engagement with feminist/queer activism:

- **Academics:** Professional engagement for training, creation and production of feminist and queer knowledge. In this category, we find primarily gender studies scholars, but also people who promote a feminist reading of science (scientists, doctors) or promoting women in STEM.
- **Professionals:** Professional activity in the construction, negotiation and application of laws, rules and procedures related to feminism and women's and sexual rights. This

category is composed of diversity practitioners, salaried community activists, lawyers and instructors.

- **Associative:** Volunteer activities in communities or in associations or organizations related to feminism/queer activism. This group is made up of activists, spokespeople and others from associations, collectives, political parties or unions.
- **Media:** Engagement in the diffusion of feminist/queer ideas and knowledge in the media, the publishing world, and the paper and online press. This group includes authors, editors, librarians, translators and journalists.
- **Collective:** Engagement in networks of solidarity or protest, volunteer activities that take place in activist spaces with low organization (online activism, grassroots collectives, informal networks etc.).

In the Parisian subcorpus, which is the focus of this paper, the categories are almost equal, with 9 academics, 9 professionals, 10 activists in associations, 7 media, and 6 collective members.

As mentioned in the introduction, our interest in constructing social factors based on discursive world is not only in that it provides social categories that emerge from the corpus itself, but also we hypothesized that such categories might actually provide more accurate models for patterns of linguistic variation than those based on demographic properties. We will therefore investigate how a breakdown of the CaFé corpus based on discursive worlds compares to breakdowns based on other kinds of social categories. We start with age: in the Paris corpus, we have speakers of a wide variety of ages, ranging from 19 to 72 years old. The speakers are somewhat evenly distributed across three age categories: speakers 60 and over (12/41), the generation that were children/teenagers during the second wave feminist movements in 1970s; speakers under 35 (19/41), who were in their 20s during the #metoo movement in 2017; and

speakers from 35 to 59, who fall in the middle of these two movements (10/41). While CaFé is relatively balanced with respect to age, this is not the case for social class/education. The Parisian subcorpus in particular is highly skewed towards the extremely well educated and the French demographic agency (INSEE)'s *cadres et professions intellectuelles supérieures* 'executives and higher intellectual professions' category. All 41 speakers in the corpus have at least an undergraduate degree, and 13 even have a PhD. Unsurprisingly, there is a positive correlation between having a PhD in our corpus and being older (Spearman's rho coefficient for rank-biserial correlations between dichotomous and ordinal data: r_{rb}= 0.45, p<0.01). There is also a positive correlation between having a PhD and participating in the most formally organized kinds of engagement (r_{rb} = 0.50, p<0.001), with all the PhD holders except for one participating in associations, being professionals or academics. Because of its homogeneity, social class/education is not the most relevant social factor for studying linguistic variation in CaFé.

While social class/education is too homogeneous for interesting quantitative linguistic analyses, other demographic factors, such as gender and race, are too heterogeneous. For example, we let speakers communicate their gender and sexual identities to us in their interviews, and in this way we obtained an extensive list of gender identities, many of which are not theoretically easy or desirable to group together. In CaFé, some speakers identify as *femme*/*homme cis* 'cis woman/man', others identify as *femme/homme trans*. Some identify as *femme* only (no *cis/trans*) and don't mention a transition; whereas, others identify only as *femme* (no *cis/trans*) and mention their transition in the interview. Others still cite other terms, such as *meuf* 'chick', *fille* 'girl', *non-binaire* 'non-binary', *personne* 'person' and even *alien* 'alien' (albeit jokingly). Taken together, the result is a gender factor that is not suitable for quantitative analysis. A similar point applies to race: the existence of racialized people implies the existence of racializing actors and acts

(Pfefferkorn 2011, Kergoat 2011). In order to avoid the interview being a racializing place, it seemed relevant to us to not identify people as already "racialized". We therefore created space in the interview to discuss racialization issues as a way to not have a color-blind practice while not assigning fixed racialized identities. We did this through asking participants 1) whether the notion of *race* is acceptable to them, and if so, 2) whether/how they thought their racial background affects their feminism. Proceeding in this manner, however, results in a race/ethnicity category that cannot be reduced to two or three categories, as is required for quantitative analysis of a corpus composed of only 41 speakers. As is common in France (see Beaman & Petts 2020), not every participant thought that race is an acceptable notion, and many of those who did expressed complex identities, such as ``white passing" or "white with experience of racism". In addition, the French people of Asian descent in our corpus may have a different experience of racism than the French people of North African descent, who may have a different experience from the French people of Central African descent in our corpus, and these differences may turn out to be relevant for how speakers use language. With CaFé, we are therefore faced with a difficult situation: we have a corpus that is very diverse from both a gender and racial perspective, and we know from previous research that the social relations of gender and race often play a role in structuring quantitative patterns of linguistic variation. However, recognizing the complexities of the way our speakers experience their gender and racial identities make it impossible to construct a category that both respects these complexities and is suitable for quantitative analysis. In the rest of the paper, we argue that categories responsible for linguistic variation can be constructed not through asking participants how they self-identify, but through looking at what they say in the interview.

Moving away from demographic categories to more ideological ones, we can now look at the breakdown of the corpus in terms of feminist *courant* labels. Participants were asked in the

interview whether their feminism had a label or an adjective, and the results were varied. Many speakers answered at least a single label, the most common being *intersectionnel* (12/41), *matérialiste* (9/41), and *queer* (3/41). As in Bendifallah et al. (2023), a couple speakers proposed combinations of these terms: *matérialiste queer* (one person), *matérialiste intersectionnel* (one person). We also had one *écoféministe*, one *féministe lutte des classes* 'class war feminist', and one *anarcha-feministe*. There were also some speakers who start their answer to the question of how they qualify their feminism with *I don't like labels*, but then continue to provide the label that they feel closest to (4/41). Finally, there were a number of speakers who didn't provide a label, either because they simply identified as *féministe* or because they didn't actually identify as *féministe* (see AUTHORS, 2023), and some did not even seem to understand what our question was about (8/41). We observe a relationship between age and label, shown in Table 1: all speakers under 35 understood the question about their feminism's adjective, and the answers were about evenly split between *intersectionnel* and *matérialiste*, with a couple of people identifying as *queer* feminists. Feminists over 60, however, either did not have a label for their feminism, or identified with an orientation other than the most common three: *anarcha-feminism, féminisme lutte des classes* or *écoféminisme.* Table 1 groups together feminists with no label and those with unique labels (anarcha-feminism, lutte des classes, écoféminisme).

Table 1: Number of speakers is the Parisian CaFé corpus, by age and self-identification

	Intersectionnel	Matérialiste	Queer	No label/Other
Under 35	9	8	2	0
35-59	5	1	1	3
60+	1	1	1	9

Interpreting the results in Table 1 is hard. While these labels were all provided by the participants themselves, understanding their meaning is difficult in isolation. As discussed above, the terms *intersectionnel, matérialiste* and *queer* are all plurivocal, having different meanings for different people. In the next section, we will argue that discursive worlds can be helpful for understanding what lies behind participants' self-identification with these labels.

3. Building discursive worlds in the CaFé corpus

The theory behind our *discursive world* measure starts from the observation that, for many sociolinguistic variables, analyses framed in terms of age, gender, social class or race are, at the end of the day, proxies for aspects of the social worlds that individuals inhabit and the kinds of interactional situations they find themselves in. Why do we find generalizations of the form *women are more likely to use variant X than men* ? Because more women than men in the study found themselves in social situations in which variant X is more useful to them in their interactional goals and/or persona construction in the interview. However, as Eckert (1989) shows, the minority of women whose interactional goals and persona construction are different from the mainstream female ones do not satisfy this generalization, and instead they use the variants that are more useful to construct a non-mainstream persona. Thus, generalizations about language use that invoke gender should be understood as only distantly related to the gender categories, and this is independent of whether there are two, three or nine of them (see also Silverstein 1985 and Ochs 1993 for more discussion). Likewise, why do we find generalizations of the form *working class people are more likely to use variant Y than upper middle class people*? Because variant Y is preferred by more working class people in the study because of its familiarity or social meanings, which are easier or more useful to make the kinds of interactional moves or construct the kinds of

personae that more working class people want to do in the interview than upper middle class people. Again however, as Labov (1963) showed, if two people of the same social class have different goals or want to come across as different 'kinds' of people in the sociolinguistic interview, their patterns of linguistic variation will be different.

Since gender, social class or other social relations constrain the interpretability and shape the personae and interactional goals of speakers, more accurate generalizations about the distribution of linguistic variants will ultimately come from a better understanding of the personae and goals found in the corpus. And clearly the way to determine this is to look carefully at the content of speakers' interviews: what do the participants talk about in response to the very general questions asked by the interviewer? Qualitative discourse analysis is one way of extracting the desired information, but it is not without its problems. Discourse analysis is extremely time consuming for large corpora and relies on researchers' intuitions that, with large datasets, can quickly become unreliable (see Mautner, 2016). Our idea, then, to develop a more tractable and empirically based method is to identify the specific referents that speakers cite in their interviews and to construct categories for quantitative analysis based on how similar speakers are in their citations. Feminist authors and figures are regularly mentioned in feminist discourses in order to establish common knowledge, create connivance or conflict, or ideologically situate oneself. These figures therefore work as a source of epistemic authority in the discourse of the speakers and play an important role in designing their ideological landscape. For this reason, we chose to focus on these elements.

More specifically, in the transcription of each interview, we tagged each occurrence of a proper name. We then extracted all the proper names from the interview transcripts, compiling a list of specific referents that each participant cited. We then developed a measure of similarity

between participants based on the similarity between their discursive worlds; that is, based on how many proper names overlap in their respective lists. As an example, consider the lists of four speakers: speaker 12, speaker 13, speaker 24 and speaker 42, shown in Table 2.

Table 2: Lists describing discursive worlds of speaker 12, speaker 13, speaker 24, speaker 42

Speaker 12	Speaker 13	Speaker 24	Speaker 42
Adèle Haenel Angela Davis Alice Coffin Assemblée Nationale Fédération des aveugles de France Elizabeth Badinter Benoite Groult Caroline de Haas Catherine Coutelle CESE CFDT CGT CGPME Charlotte Bienaimé Dominique Joseph Édouard Philippe Elle FAGE Françoise Milewski Françoise Vergès HCE INED INSEE JOC La Poudre Lauren Bastide Les Couilles sur la Table Les Gouines Rouges LOSC Marlène Schiappa OCDE Emmanuel Macron La manif pour tous MEDEF Le Ministère des Finances François Mitterrand Niky de Saint Phalle Osez-le-féminisme Le Planning Familial Raphaëlle Rémy Leuleu Pierre de Ronsard Léopold Sédar Senghor	Assemblée Nationale Audre Lorde Causette Christine Delphy CESE Ernestine Ronai Facebook Fédération des aveugles de France FNSEA France Culture Gérard Darmanin GEPS Gwenaëlle Perrier Gwenola Ricordeau HCE IEP Instagram Joan Scott Les Couilles sur la Table Libération Marine Le Pen Marlène Schiappa MEDEF Ministère de la Culture Osez-le-féminisme Observatoire contre les violences faites aux femmes Le Planning Familial Sara Ahmed Sciences Po U. Toulouse -Le Mirail U. Paris 8 U. Paris 1 (Sorbonne) UNEF Victoire Tuaillon	Ma grand-mère n'était pas une féministe U. Paris 2 La Cité des Chances Dilnur Reyhan Femen Humans for Women Imazi Reine Kiffe ta Race Lallab Les Ours à Plumes La maison des femmes du 93 Marie da Silva Marlène Schiappa Olympe de Gouges Philippe Juvin Raphaël Glucksmann Simone de Beauvoir Lycée Stanislas Twitter Éric Zemmour l'École Alsacienne	Amina Wadud Aminata Dramane Traoré Aoua Keita bell hooks Jean Michel Blanquer Carmen Diop Caroline Fourest Elle Femen Françoise Vergès Instagram Kimberlé Crenshaw Lallab Maboula Soumahoro Mariama Ba Marie Rose Moro Marlène Schiappa Maryse Condé Médiapart MLF Nargesse Bibimoune Parti socialiste Rokaya Diallo Ségou SOS racisme Tobie Nathan Women Sense Tour

Sophie Binet Christiane Taubira UNEF Najat Vallaud-Belkacem Véronique Sehier Victoire Tuaillon Wa-thiong'o			

Speaker 12 and speaker 13 are colleagues (professionals) in a large governmental agency related to social planning and justice. This relationship can be seen in the fact that they have a very high number of referents in common: 11. Furthermore, these referents are often governmental organizations (*l'Assemblée Nationale, HCE, CESE*), unions and professional networks or large mainstream feminist associations (*Osez-le-féminisme, le planning familial*). On the other hand, aspects of their discursive worlds are very different. Speaker 12 talks much more about governmental agencies and unions, even citing the names of highly placed individuals in these organizations (*Dominique Joseph* (CESE), *Véronique Sehier* (Planning Familial), *Sophie Binet* (CGT)). In contrast, speaker 13 cites social media platforms (*Instagram, Facebook*) and higher education institutions. Given these differences, it is not surprising to find out that speaker 12 is older (in her 60s) and occupies a government position with more responsibility than speaker 13, who is in her 20s. Thus we see that the discursive world (as operationalized as lists of specific referents) can encode aspects related to age and profession.

Comparing with speaker 24, we see that this speaker has only one referent in common with speaker 12 and/or speaker 13: the controversial politician, and former minister of equality between women and men (2017-2020), *Marlène Schiappa*. Otherwise, the people, places and things that speaker 24 talks about are very different: the associations (*La Cité des Chances*, *La maison des femmes du 93, Imazi Reine, Lallab)* are smaller and more focused on decolonialism and racial and religious justice. In this way, the discursive world (measured in this way) can also encode aspects of participants' racialized identities: speaker 24 describes herself in the following way (speaker

24, line 1006): “alors quand bien même ma peau est claire je peux pas être perçue comme blanche aux yeux des gens et ça c'est juste un fait” `while my skin is light I can’t be perceived as white in people’s eyes, and that’s just a fact’; whereas, both speaker 12 and speaker 13 describe themselves as white. While the latter two are clearly interested in engaging intellectually with intersectional and decolonial ideas, as witnessed by their citations of *Angela Davis*, *Audre Lorde* or *Françoise Vergès*, the organizations and cultural objects that they talk about do not center racial questions in the way that those mentioned by speaker 24 do. Finally, we can compare the discursive worlds of these three speakers with that of speaker 42, who identifies as a black muslim woman. Like speaker 24, speaker 42 also cites the muslim feminist association *Lallab*, the radical anti-religious collective *Femen*. Like most of the speakers in our corpus, *Marlène Schiappa* comes up at least once in the interview, and not in a positive light.

Note that in the construction of discursive worlds, we take into account only whether a speaker mentions a particular person, place or thing, not how they feel about them. While Schiappa is universally disliked in our corpus, many other figures are more controversial. This is the case of *Rokhaya Diallo*, who is listed as an important feminist by speaker 42 (1a), but disliked by speaker 7 (1b), who disagrees with her stance on natural Afropean hair. Despite their conflicting opinions on Diallo, the fact that both speaker 42 and speaker 7 cite her will create a link between the two speakers in the subsequent analysis.

(1) a. il y a des femmes comme Maboula Soumahoro mais qui ont juste quelques années de plus que moi, Rokhaya Diallo qui ont juste quelques petites années [Interviewer: ouais mais qui des fois peuvent être quand même de la génération euh ouais] voilà (Speaker 42)

'There are women like Maboula Soumahoro but who are just a few years older than me Rokhaya Diallo who are just a few little years [Interviewer: yeah but who sometimes may be still from the generation hum yes] that's it

b. voilà je pense que c'était Rokhaya Diallo que tout le monde connaît bah non je suis pas d'accord avec elle là au moment quand elle commence à dire "on fait pas des trucs pour les cheveux" enfin bon euh (Speaker 7)

'That's it I think it was Rokhaya Diallo that everybody knows bah no I don't agree with her when she starts saying "we don't do anything for hair" well huh'

To sum up, we can compare the lists of referents produced by speakers in the corpus, and this comparison will allow us to construct a measure of similarity between them[1]. In particular, **the more items in the discursive world two speakers have in common, the more similar we will consider them to be with respect to this measure.** In this way, speaker 12 and speaker 13 will be considered more similar to each other than either of them will be to speaker 24. Speaker 42 will be less similar to speaker 13 than she will be to speaker 12 or speaker 24. We encode this information in a matrix which, for all 41 speakers in the corpus, represents the number of referents they have in common.

3.1 Constructing similarity spaces

As a basis for our discursive world measure(s), and in order to visualize the similarity relations between speakers in the corpus, we will construct a similarity space based on the similarity matrix. Similarity spaces are one-dimensional or multidimensional structures with a metric defined on them. The dimensions represent basic features that objects can have, and the metric measures

similarity between the representations of objects in the space: the greater the distance between the representations of two objects, the more dissimilar these objects are in the respect corresponding to the space; conversely, the closer the representations are, the more similar the objects are in that respect (see Gärdenfors 2014). In this paper, we will follow the procedure used in Bendifallah et al. 2023, who have previously constructed similarity and conceptual spaces from sets of French feminists. Following these authors, we first transform the discursive world matrix into a distance matrix by computing the Pearson correlation of the matrix and then transforming the correlation matrix into a distance matrix using the formula underlying the *cor2dist* function from Revelle's psych package for R (Revelle 2023). Then, using the SMACOF algorithm, implemented in the *smacof* package in R (Mair, Groenen & de Leeuw 2022), we performed multidimensional scaling on this distance matrix[2]. We considered the results of multidimensional scaling for up to six dimensions. In choosing the number of dimensions, there is always a trade off between favoring a structure with low dimensionality, that is easier to visualize and can often be easier to interpret, and a structure with high dimensionality, which is almost always a better fit to the data. The goodness of fit of an MDS structure is commonly measured in terms of the *stress value*, which measures how closely the distances between objects in the configuration resulting from the MDS procedure match the similarities between the items underlying the distance matrix. The lower the stress value, the better the fit. Figure 1 shows the stress values for MDS structures of 1-6 dimensions, and compares them to the stress values of 250 MDS structures generated by random distance matrices. This panel shows that the similarity space generated from our distance matrix has a consistently better fit than spaces generated from random data, and that, although there is no sharp elbow, the angle gets slightly less sharp going from 3 to 4 dimensions. We therefore present a three dimensional structure.

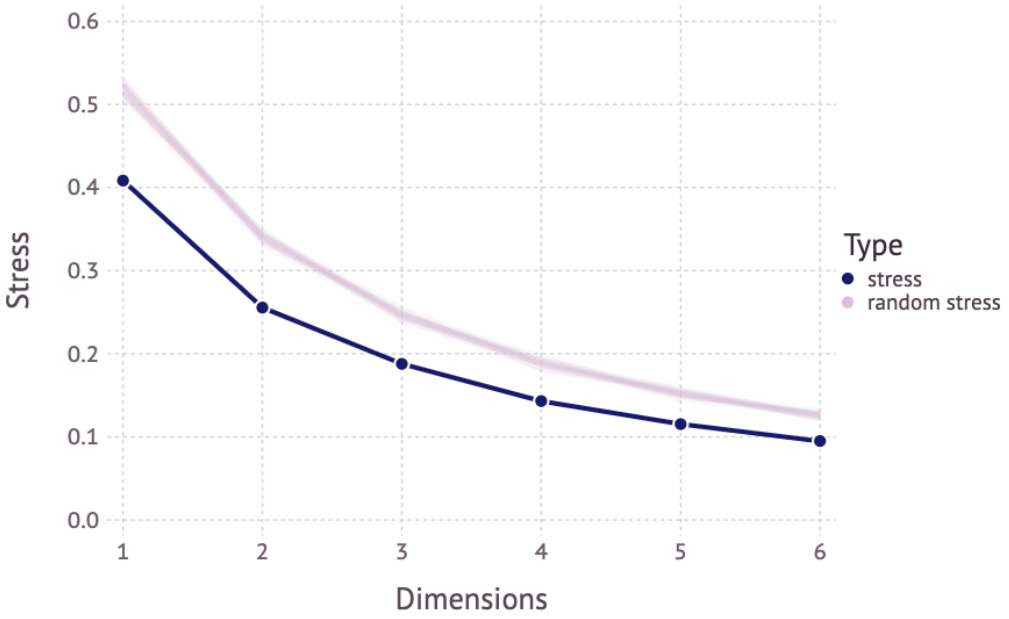

Figure 1: Stress values for MDS structures for 1-6 dimensions based on our distance matrix, compared to 250 MDS structures based on random distance matrices.

The first two dimensions of the similarity space are shown in Figure 2.

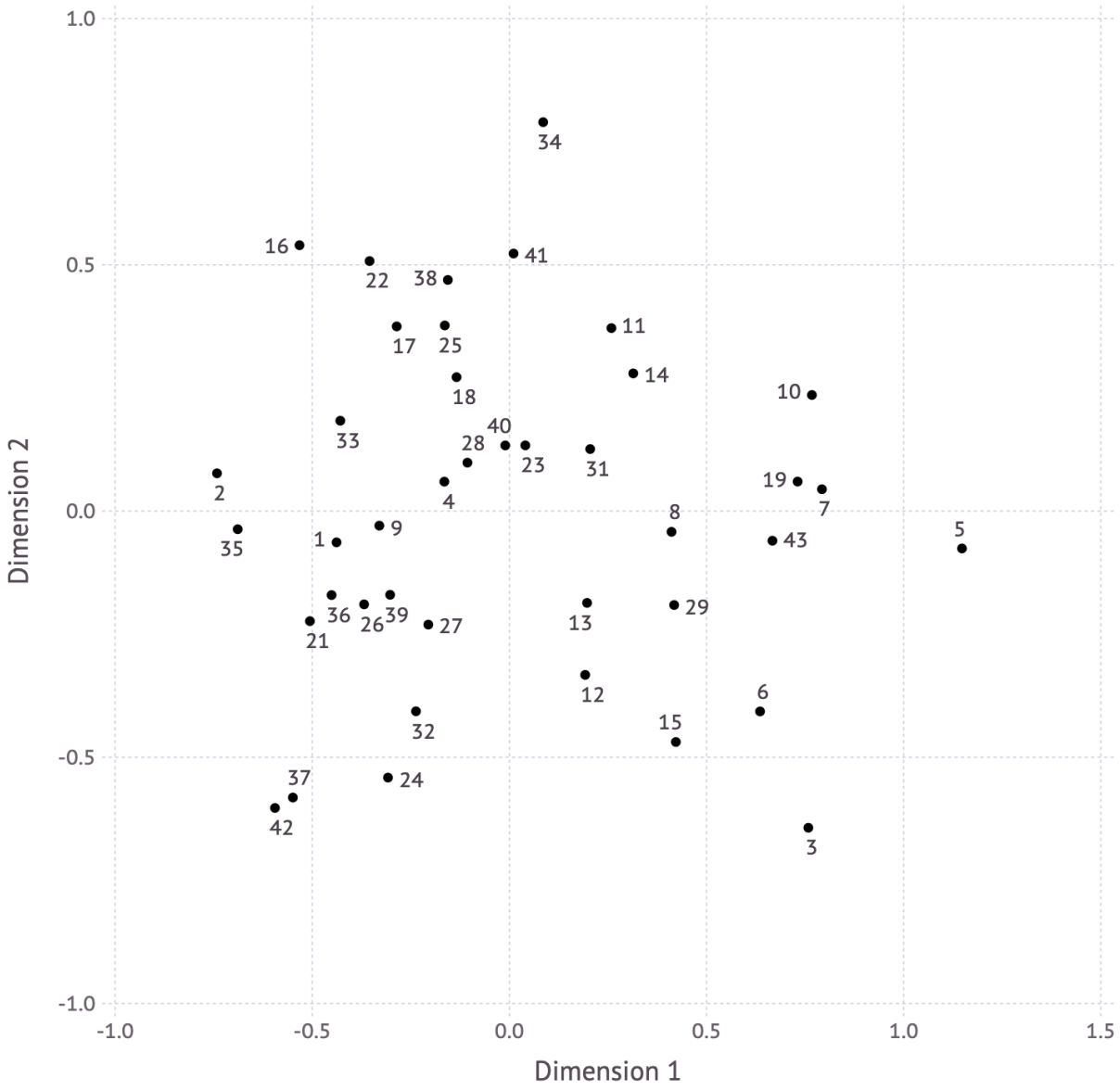

Figure 2: Dimensions 1 and 2 of similarity space for Parisian CaFé corpus

Dimension 1 starts with speakers 2, 35, 42 and 37 on one end, and 5, 3, 6, 7 and 10 on the other. Although no interpretation is definitive since these structures are produced based on the distance matrix and non-deterministic multidimensional scaling algorithms, one possible interpretation is that dimension 1 distinguishes speakers based on whether they are more likely to cite authors or works of academic or popular culture versus institutions or organizations. Speaker 5 is a doctor

who is involved in a medical organization centered around women, and she almost exclusively cites governmental bodies (*le sénat, l'académie française, l'assemblée nationale*), medical associations, or politicians. Likewise, speaker 3, a diversity professional in a large Parisian university, cites governmental and university associations and figures, and only one feminist thinker (*Michelle Perrot*). The right side of the figure is people with feminists whose engagements are in national-level associations (speakers 6, 8, 10) or political parties (speaker 19).

On the left side of the figure, we have speaker 42 (see Table 2), whose discursive world is primarily composed of feminist authors, activists, artists and (social) media. Speaker 2 is an extreme case, where her discursive world contains only 12 items, none of which are governmental organizations and only two of which are feminist associations (*Nous toutes, Féminicités*). Almost all the rest are social media platforms. Speaker 35 cites her alma mater, *Sciences Po*, but otherwise, her discursive world is a mix of feminist authors and artists and social media platforms. For this reason, we propose that dimension 1 encodes a distinction ranging from the **theoretical and cultural** aspects of feminism to its **institutional** aspects.

Dimension 2 appears to primarily differentiate the speakers on the *theoretical/cultural* side, there being empty space in the top right corner of Figure 1. On one end of dimension 2 lie speakers 34, 16, 41 and 22, and on the other lie speakers 42, 37, 24. As mentioned above, both speakers 42 and 24 cite authors and organizations with a focus on racial justice, and this is also the case for speaker 37, who cites organizations like *Amina, Gazelle* and l'*Association des femmes maliennes*, and authors like *bell hooks, Amandine Gay,* and *Angela Davis.* On the other end of dimension 2, we have speaker 16 who cites no authors, artists or organizations focused on racial justice, but rather cites figures focusing on sexuality and/or the economic justice/far left: *Andrea Dworkin, Monique Wittig, Karl Marx, Leon Trotsky, Vladimir Lenin, Garces, Nuit Debout, Aides*. Note that

speaker 16 describes being classified as *meuf arabe* 'arab chick' and *meuf racisée* 'racialized chick', so here we see that, while there may be some relation, one's discursive world is not uniquely determined by properties such as being racialized. Likewise, speaker 34 cites no organizations or authors focusing on racial questions, and the vast majority of the people, places and things that she talks about are related to gender and/or sexuality: *Sam Bourcier, Paul Preciado, Eliane Viennot, Eliot Page, la Mutinerie, Mots-Clés, Raphael Haddad*. Even looking on the "institutional" side of dimension 1, we see a contrast on dimension 2 between speaker 3, whose discursive world is filled with academic and governmental institutions or organizations, and speaker 10, who, although she cites some governmental institutions, also cites organizations focusing on sexuality like *SOS-homophobie, Association des parents et futurs parents gays et lesbiens, queer code*.

Dimensions 2 and 3 are depicted in Figure 3.

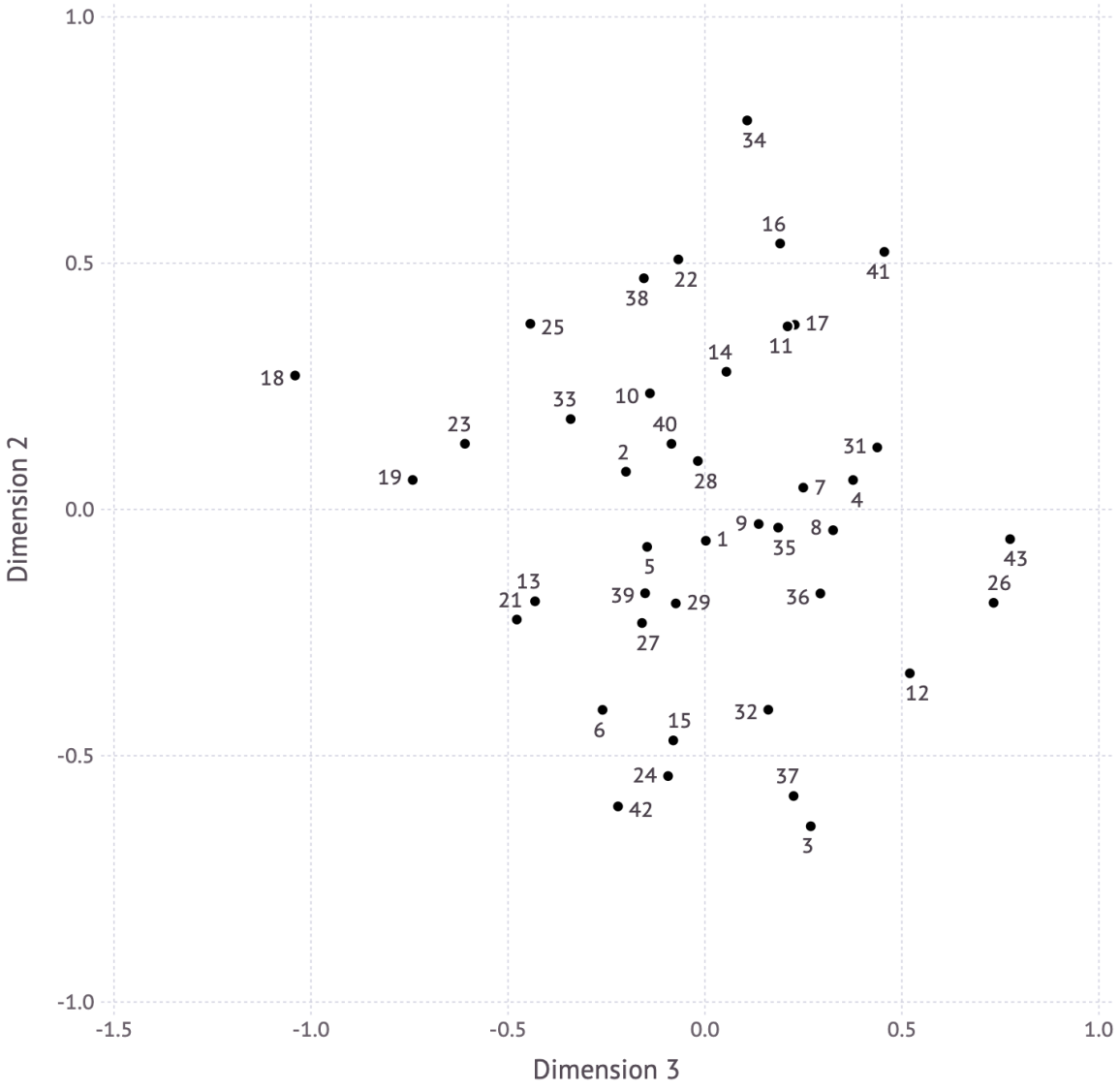

Figure 3: Dimensions 2 and 3 of similarity space for Parisian CaFé corpus

Dimension 3 is a little harder to interpret: on one end, we have speaker 43 and speaker 26. These two speakers are the only ones in the corpus who talk about ecofeminism. Speaker 26 cites *Starhawk,* the book *Sorcières, sages femmes et infirmières* and *Françoise d'Eaubonne,* while speaker 43, a self-described "witch", also cites *Starhawk,* along with *Le Bûcher* and *Carlos Castaneda*. On the other end, we have speaker 18, speaker 19 and speaker 23. Speaker 18 is from the media group, citing mostly sexuality focused authors and artists (so on the theoretical/cultural

side). Speaker 19 is a member of a political party and almost only cites institutions, organizations and political figures. Speaker 23 is somewhat in the middle, citing institutions, organizations and people devoted to gender equality. These three speakers do have a history of far left activism in common, but this is not straightforwardly reflected in their citations. We therefore tentatively hypothesize that dimension 3 can be interpreted as an interest for nature (including ecofeminism) vs sociocultural issues, but because this is very uncertain, dimension 3 will not play a large role in the analyses below.

3.2 Cutting up the space into categories

Again following Bendifallah et al. (2023), we use clustering methods to group the speakers into categories based on their place in the similarity space derived from their discursive worlds. In this paper, however, we are interested in partitioning speakers into classes for use in statistical models of patterns of linguistic variation. It is because of the many ways in which discursive worlds allow us to do this that, we argue, our approach captures some of the dynamicity and context-sensitivity that constructivist approaches to categories stress as crucial to understanding them.

The three dimensions that our similarity space provides can, in principle, be used in statistical analysis, each as a factor that could be potentially relevant for some linguistic variable. In other words, we believe that the discursive world method could be used to study any kind of sociolinguistic variable: phonetic/phonological, morpho-syntactic, lexical or discourse/pragmatic; however, it is possible that the way in which we cut up the similarity space may be different for different kinds of variables. The dimensions that we extract are continuous and it is possible that they are most useful for analyzing continuous sociolinguistic variables, or at least variables whose occurrences are sufficiently frequent as to make the ultra fine-grained distinctions between, for example, speaker 23 and speaker 40 or speakers 27 and speaker 39 relevant. However, for less

frequent variables or variables for which we wish to have discrete categories, there are also empirically-based ways of constructing such categories. Again, not only do we have multiple options for partitioning space into categories, there is actually no single static "right" way of doing so[3]. How many categories we make will depend on considerations related to the properties of our dataset and, more importantly, our theoretical hypotheses driving our investigation. Since this paper is exploratory in nature, the illustrations that we will give will be guided primarily by how much data we will have. Since we will be looking at lexical variables in the next section which are not particularly frequent, we will mostly be interested in partitioning the set of speakers into two or three groups in an optimal way.

The particular clustering algorithm that we employ in this paper is k-means clustering (MacQueen 1967), implemented in R (using the Stats package). K-means clustering is commonly used in machine learning to partition *n* observations or objects into *k* clusters that, very broadly speaking, minimize the distance between members of the same cluster, and maximize the distance between members of different clusters. We use k-means clustering in this paper because it is well-understood and has been shown to be useful across a very wide range of tasks in a wide range of domains. Using k-means clustering for k=2, we arrive at the partition in Figure 4. Roughly, the two-way partition appears to make a distinction between the speakers on the cultural side of feminism and those on its institutional side.

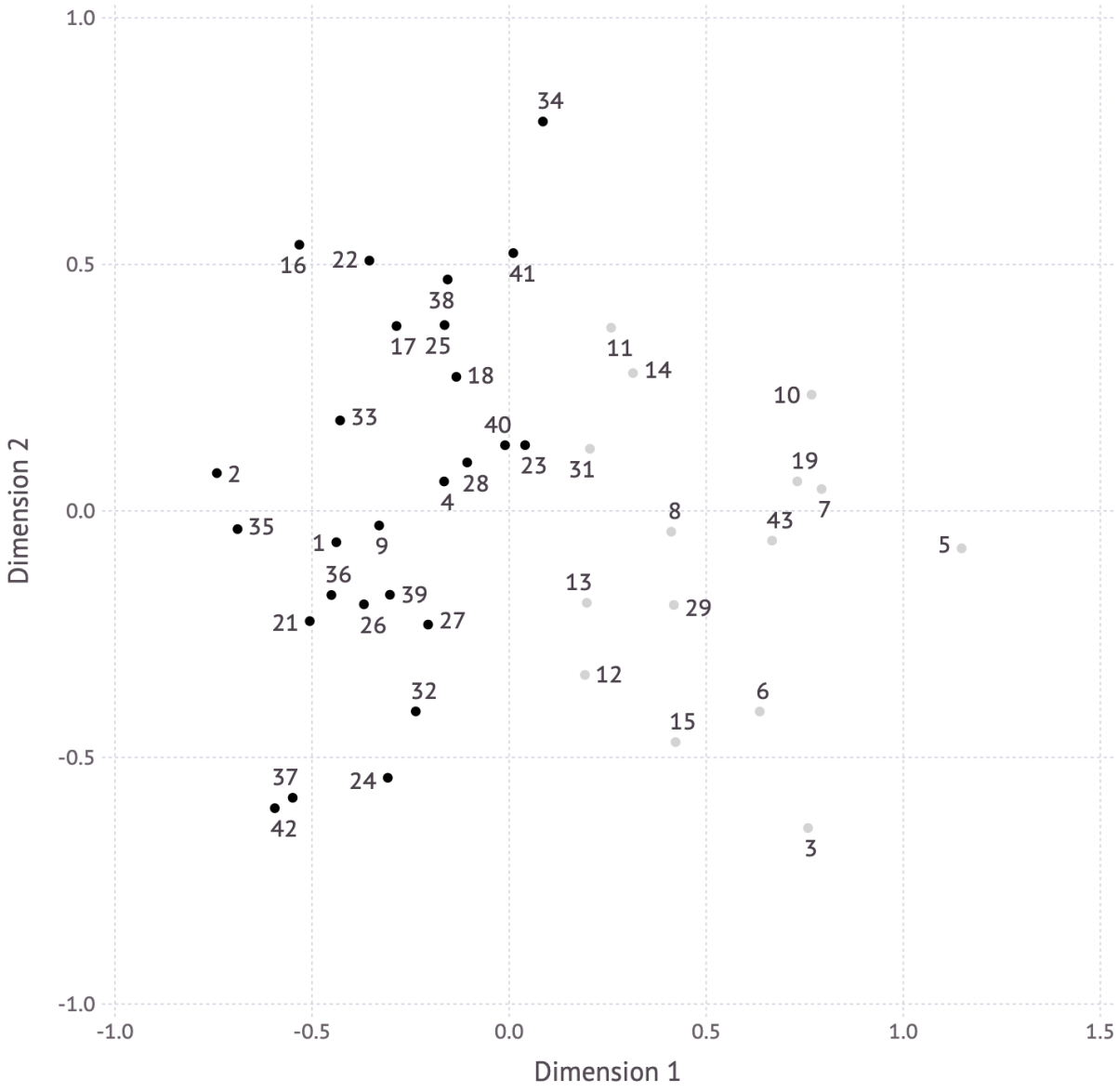

Figure 4: K-means clustering analysis for k=2

With this method, we are not limited to partitioning the set of speakers into two categories. Looking at k-means clustering for k=3, we find the partition shown in Figure 5. Here, the algorithm predominantly makes a distinction on the "theoretical/cultural side", separating culturally oriented feminists primarily focused on sexuality from those primarily focused on race. As we suggested above, the sexuality-race distinction is not particularly relevant for speakers on the "institutional side": speaker 10, speaker 19, speaker 3 and speaker 6 are all part of the same category.

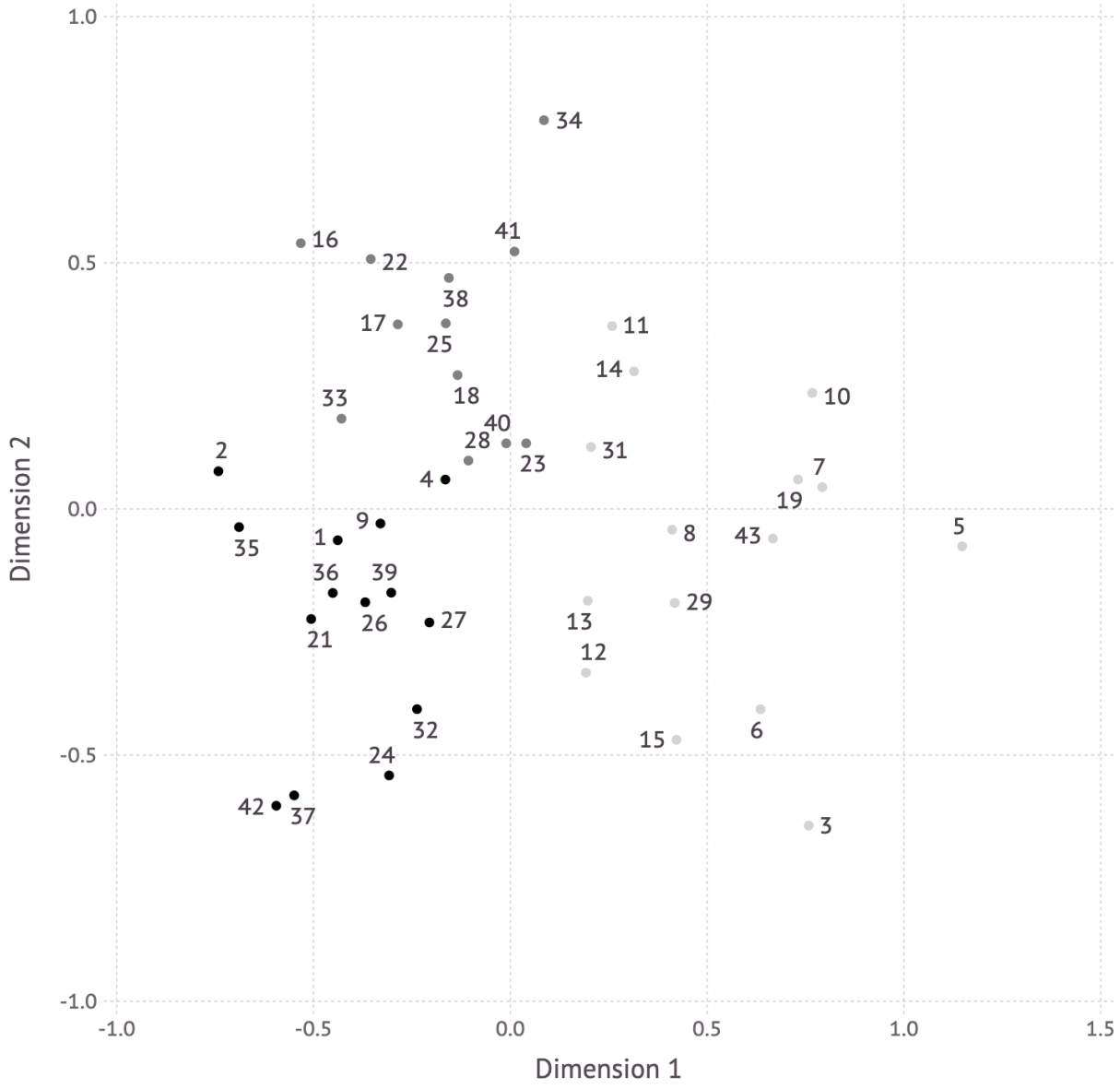

Figure 5: Kmeans clustering analysis for k=3

We can now investigate whether being in the institutional or cultural cluster, or being in one of the two cultural clusters, correlates with the other social factors we have for the corpus. Table 3 shows the breakdown of the CaFé Parisian corpus according to age and discursive world (3 clusters). We see that there are no speakers over 60 in the Cultural A group (more focused on race), which is the group with the most under 35s. In contrast, the Institutional group has the highest number of speakers over 60 and the lowest number of speakers under 35. Cultural B (more

focused on sexuality) is interesting because its speakers are evenly spread out across the age groups.

Table 3: Speakers according to age and discursive world (3 cluster)

	Under 35	35 - 59	60+
Theoretical/cultural A (focus on race)	10	2	0
Theoretical/cultural B (sexuality)	6	5	5
Institutional	3	3	7

There is also a significant but weak correlation between discursive world (2 clusters) and engagement (Phi coefficient for the correlation between two binary variables: $Phi = 0.34$, $p<0.05$). Table 4 shows this correlation: it is essentially driven by the fact that members of informal collectives and media are much more likely to be in the cultural discursive world than in the institutional one. Activists in associations, professionals and academics are roughly equally spread out across the two worlds.

Table 4: Speakers according to engagement type and discursive world (2 cluster)

	Collective	Media	Association	Professional	Academic
Theoretical/Cultural	5	7	6	3	5
Institutional	1	0	4	6	4

We get some more clarification as to how the two theoretical/cultural discursive groups are different through looking at how they break down according to feminist label. As shown in Table 5, two thirds of the speakers in cultural group A (focus on race) identify as *intersectionnel*, and none have no label. cultural group B (focus on sexuality) has the largest number of queer and materialist feminists, but also contains some intersectional feminists and some non-labelers.

Finally, the institutional group shows the opposite pattern from cultural group A: two thirds either identify as not having a label or with some more esoteric one.

Table 5: Speakers according to feminist label and discursive world (3 cluster)

	Intersectionnel	**Matérialiste/Queer**	**Nothing/Other**
Theoretical/cultural A (focus on race)	8	4	0
Theoretical/cultural B (focus on sexuality)	3	8	5
Institutional	4	1	8

The discursive world (3 cluster) measure thus enlightens us as to what lies behind our participants' use of the labels *intersectionnel, matérialiste* and *queer*: speakers who identify as *intersectionnel(le)* are more likely to talk about people and organizations devoted to racial questions than those who identify as *matérialiste* or *queer*, who are more focused on sexuality. We therefore argue that one way in which discursive worlds can be helpful to variationist sociolinguistics is through helping us better understand the local group labels that participants use themselves. In the next section, we will argue that discursive worlds can also be useful for analyzing quantitative patterns of linguistic variation.

4. Analyzing linguistic variation using discursive worlds

In this section we present two quantitative studies focusing on sociolinguistic variables related to feminism (the focus of CaFé). We have chosen lexical-discursive variables because such variables are not subject to linguistic (Labov's internal) factors in the same way that phonological or morpho-syntactic variables are. This makes testing the usefulness of the discursive world measure much easier; however, it means that our datasets are smaller than for sociophonetic or morpho-

syntactic variables. Consequently, our statistical analyses will use Bayesian logistic regression models rather than frequentist regression models because Bayesian modeling is better adapted to smaller datasets (see Sorensen, Hohenstein & Vasishth, 2016 for arguments that Bayesian modeling is to be preferred for studies in linguistics, psychology and cognitive science).

4.1 Violences faites aux femmes vs violences sexistes et sexuelles

The first variable we will study is the alternation between *violences faites aux femmes* 'violence against women' vs *violences sexistes et sexuelles* 'sexist and sexual violence'. Both of these expressions are used in institutional discourse, with *violences faites aux femmes* being the older variant found in texts from the 1990s, such as the United Nations' *Déclaration sur l'élimination de la violence à l'égard des femmes* 'Declaration on the elimination of violence against women' in 1993. According to Lochon (2021)'s study of terms for violence against women in the French press from 1989-2019, *violences sexistes et sexuelles* started to be used, little by little, in the beginning of the 2000s, and it has now become the official term used by the French state (see https://arretonslesviolences.gouv.fr/).

In the CaFé corpus, we find that the two terms are in variation when our speakers talk about the violence that they are fighting against. A case of intra-speaker variation is shown in (2), from speaker 19, and the distribution of the two variants are shown in Table 6, broken down according to discursive world (2 cluster).

(2) a. le reste enfin il se passe rien quoi, enfin je veux dire, que ce soit les **violences faites aux femmes**, que ce soit sur les discriminations LGBTI-phobes (Speaker 19)

'The rest, well, nothing happens, I mean, whether it's violence against women, whether it's LGBT-phobic discrimination'

b. enfin c'est consubstantiel de la domination masculine quoi, les **violences sexistes et sexuelles** (Speaker 19)

'At the end of the day, male domination and sexist and sexual violence are consubstantial'

Table 6: *Violences faites aux femmes* vs *violences sexistes et sexuelles* (2 cluster)

	Violences faites aux femmes	*Violences sexistes et sexuelles*
Theoretical/cultural	10	23 (68%)
Institutional	17 (63%)	11
Total	27	34

We ran Bayesian logistic regression analyses in R (R Core Team, 2023; Posit Team, 2023) testing to see whether there is evidence for a relationship between our social factors (discursive world (2 clusters, 3 clusters), age, education (PhD), engagement, feminist label) and the use of *violences faites aux femmes* and *violences sexistes et sexuelles.* The full detailed results of these analyses, which include speaker as a random effect, are given in the supplementary material at https://osf.io/hu9fy/?view_only=043fcd19e2be49d2830c9c543d65ba43. To summarize, we find that the only factor for which there is robust evidence (P(est.<0)=.97) for an effect is discursive world with 2 clusters[4]: as suggested by Table 6, speakers in the institutional world use less *violences sexistes et sexuelles* than speakers in the theoretical/cultural world. One possible way of understanding the pattern in Table 6 is as a stage in the ongoing "bureaucratic appropriation" of *violences sexistes et sexuelles*, a term from the feminist theoretical/cultural world. Feminists play a role trying to influence the terminology of the State and administration. For example, as described by Orellana and Kunert (2014) (see also Fassin 2008), at the beginning of the 2000s, while large international organizations like the United Nations and the European Union started introducing English terms like *gender-equality* and *gender-based violence* to replace terms

referring specifically to women in official texts, both the French State and some feminists with a mediatic audience were very resistant to using the term *genre* 'gender', either because it threatened the universalist project or it hid women as privileged subjects of feminism. Consequently the French context resisted more than other countries in appropriating *genre* into its official discourses on equality, violence and discrimination. In this dynamic, less institutional spaces try to make available feminist concepts or phrasing to those who drive gender policies. This has been observed for a number of terms like *féminicide* (Nugara 2013), *viol conjugal* 'marital rape' (Brown et al. 2017) or *intersectionalité* (Raus 2018). *Violences sexistes et sexuelles* is a good example of such circulations, as it denotes a certain expertise in the field of feminism while being recognised and used in official documentation. In other words, the pattern in Table 6 suggests that *violences sexistes et sexuelles* is an instance of activists or academics "passing" concepts from the theoretical/cultural to the institutional world. We therefore see that discursive worlds can track patterns of variation that other available social factors miss and, in doing so, allow us formulate new hypotheses about changing sociolinguistic variables.

4.2 Prostitution vs travail du sexe

Our second variable is the alternation between *prostitution* and *travail du sexe* 'sex work'. *Travail du sexe* is a calque from the English *sex work*, a term that became popularized in the United States in the 1970s, as a way of resisting the "degradation" and "objectification" of sex workers around the feminist rhetoric about prostitution at the time. According to this view (articulated most notably by Leigh (2011)), the English word *prostitute* was yet another euphemism trying to hide the shame that prostitution abolitionists held towards people who sell sexual services. The words *sex work* and *sex worker* were introduced to highlight the agency and subjecthood of people engaged in

these practices, and as a way to approach the sex industry as a question of "work" not "moral" (Butler & Rubin 1994). After being successfully adopted in anglophone North America, *sex work* was translated into French (*travail du sexe* or *travail sexuel*) by activists in Québec and France in the early 1990s, particularly in response to the increased stigmatization of people selling sexual services during the 1980s in the context of the HIV/Aids pandemic (Simonin 2016). The two expressions, *prostitution* and *travail du sexe*, are in variation in our corpus, with a single speaker often using both (3).

(3) le deuxième axe c'est que **le travail du sexe** est une violence faite aux femmes on prend euh forcément ce ce en france en tout cas euh ce ce cet axe là **la prostitution le travail du sexe** est une violence faite aux femmes (Speaker 9)

'The second axis is that sex work is a violence against women we take uh necessarily this this in France at least uh this this axis prostitution sex work is a violence against women'

We extracted all occurrences of expressions built on the root *prostitution* (*prostitution, prostitué(e), (se) prostituer* etc) and expressions built on the root *travail du sexe* (*travail du sexe, travailleuse, travailleur du sexe, TDS*). The distribution of the variants *prostitution* vs *travail du sexe* is shown in Table 7. This table shows that there is a large difference between speakers in the institutional world and those in the theoretical/cultural world: more institutionally oriented speakers used *travail du sexe* only one time; whereas, this newer variant was used more by more culturally oriented speakers.

Table 7: *Prostitution vs travail du sexe*, according to discursive world (3 cluster)

	Travail du sexe	*Prostitution*
Theoretical/cultural A (focus on race)	127 (79%)	34

Theoretical/cultural B (focus on sexuality)	29 (85%)	5
Institutional	1	17 (94%)
Total	133	80

We ran Bayesian logistic regression analyses testing to see whether there is evidence for a relationship between our social factors (discursive world (2 clusters, 3 clusters), age, education (PhD), engagement, feminist label) and the use of *travail du sexe* and *prostitution.* The full details of these analyses are shown in the supplementary material on OSF; like the *violences* variable studied above, there is robust evidence for an effect of discursive world (2 cluster) (P(est.<0)=.99) with speakers in the institutional world strongly dispreferring *travail du sexe*. For discursive world (3 cluster), there is also robust evidence (P(est.<0)=.98) for a difference between speakers in the institutional world and speakers in theoretical/cultural world A (focused on race), with speakers in the second theoretical/cultural cluster having a stronger preference for *travail du sexe*. However, contrary to *violences,* we find robust evidence for effects of other social factors: age (older speakers use less *travail du sexe*), education (PhD holders use less *travail du sexe*), label (intersectional feminists use more *travail du sexe* than materialist/queer feminists, who use more *travail du sexe* than non-labelers) and engagement (speakers in informal collectives, media and associations use more *travail du sexe* than professional and academics). In other words, this variable appears to break down along the main correlations described in the previous section. In cases such as this, where multiple kinds of factors condition a variable, we argue that discursive worlds can be helpful for interpreting the effects of other social factors. For example, Simonin (2016) highlights the anti-institutional stances taken by advocates of *travail du sexe,* saying,

> Le mouvement attribue donc une responsabilité causale de la stigmatisation aux pouvoirs publics qui appliquent des politiques spécifiques et aux mouvements sociaux qui défendent

> le projet d'abolition, et revendique une responsabilité politique des « travailleur·se·s sexuel·le·s », affirmant ainsi leur implication comme nécessaire à la résolution du problème." (Simonin 2016: 27).
>
> 'The movement thus attributes a causal responsibility for the stigmatization to public powers which apply specific policies and to social movements who defend the project of abolition, and claim a political responsibility from "sex workers", affirming their implication as necessary to resolve the problem.'

The anti-institutional social meaning of *travail du sexe* can be seen more transparently through how it partitions participants based on discursive world, compared to age or education factors (with which discursive world is correlated). Feminists who view their activism as lying within the institutional world would avoid *travail du sexe*, since they do not necessarily ascribe to the view that institutional policies are making the lives of people who sell sexual services more difficult. Discursive worlds also enlighten us as to why engagement has a significant statistical effect: the least institutionalized forms of engagement (informal collectives and media) are the ones who are more likely to adopt an anti-institutional stance, and therefore whole-heartedly adopt *travail du sexe*.

5. Conclusion

In this paper, we showed how to construct what we call *discursive worlds* from sociolinguistic corpora. Discursive worlds were operationalized through building lists of proper names used by speakers in the corpus, and, by virtue of this fact, we argue that they tap more directly into the aspects of speakers' social worlds that are relevant for predicting linguistic variation than categories like gender, race or social class. We illustrated this proposition with a study of two

lexical-discursive variables in the CaFé corpus. We argued that finding categories for quantitative analysis was necessary for the CaFé corpus because, by virtue of it representing speech communities composed of feminist and queer activists, the traditional demographic categories could not be applied. We also argued that discursive worlds can be useful for helping understand the labels that corpus speakers apply to themselves and to others.

More generally, we suggest that discursive worlds may be relevant for studying variation in other sociolinguistic corpora, including those originally constructed with balancing age, gender, social class and race in mind. Discursive worlds can be applied to all variationist corpora: everyone who is interviewed talks, and when they talk, they talk about people, places and things. If these referents can be extracted from more traditional Labovian corpora, it is trivial to construct the discursive world measures, which can then be used in statistical analyses of patterns of variation. This being said, we suspect that which variables end up being conditioned by discursive world will depend on the topics discussed in the interviews. In CaFé, the topics are standardized: every participant was asked the same questions, and the questions were about feminist activism. It is not therefore surprising that a measure derived from the content of the interviews was successful at predicting variation in expressions relevant to feminism; indeed, the corpus was constructed specifically to investigate these kinds of variables (see AUTHORS 2024). It is an open question whether the discursive world measures, derived from more traditional Labovian corpora, will do a good job at predicting variables that have been shown to be traditionally conditioned by age, gender, social class etc. in previous variationist studies. We tend to think that the discursive world method could work well, since without a standardized questionnaire, the list of referents cited by speakers in a classic sociolinguistic interview will be even more differentiated. On the other hand, it is possible that, without the questionnaire, speakers' lists of referents cited may have too little

overlap to make the construction of a similarity space enlightening. This is a topic for future research. What also remains to be seen is how this method treats sociolinguistic variables that are conditioned both by speakers' social properties and by interactional aspects of the discourse context. For example, the omission of the French negative particle *ne* (eg. *Je (ne) l'aime pas* 'I don't like it') is conditioned not only by properties like age and education (with more educated speakers omitting less *ne*), but also formality (*ne* is omitted less in formal contexts), see Flesch et al. (2024) for a recent review. We would expect that a discursive worlds analysis of *ne* omission in CaFé would be able to capture only the contribution of age or education factors, but this should be empirically tested. Whatever the outcome of these future investigations may be, we hope to have demonstrated that discursive world measures, which are flexible and constructed from speakers' discourse in the interviews, are another step towards a variationist sociolinguistics that incorporates insights of constructivist theories of social categories into quantitative corpus studies.

Endnotes

1 The goal of this paper is to provide a kind of "proof of concept" argument that looking at the referents speakers cite in their interviews, constructing a measure of similarity based on overlaps, and using similarity spaces and (optionally) clustering algorithms to construct factors for statistical analysis has the potential to help quantitative sociolinguists bridge part of the theoretical gap that currently exists between variationists and other sociolinguists working in constructivist approaches. In the first study using this method, we decided to adopt the simplest (reasonable)

measure of similarity that we could think of: comparing overlap in lists of referents. As we will show later in the paper, this measure of similarity seems to work reasonably well for the linguistic variables we chose to analyze; however, we can certainly wonder whether this is the optimal measure of similarity. For one thing, since we simply count the referents, and speakers' interviews differ in length, loquacious speakers who cite many figures (and have longer lists) will tend to be considered more similar to a variety of speakers (because there will probably be more overlap) than those who are more timid, i.e. whose lists are shorter, despite their underlying ideological structures being very similar. This is, arguably, not a great result. Likewise, we make no distinctions between speakers who mention a referent once and those who mention them multiple times, which could also possibly indicate some ideological difference. In the future, we would like to explore alternative measures of similarity, based on "richer" representations that would take into account the length of referent lists and how often names are repeated. We think it is entirely possible that such alternatives could yield more useful categories for quantitative linguistic analysis; however, we would need to test this in future work. We thank an anonymous reviewer for very helpful discussions of these points and for some suggestions of alternative ways of measuring similarities between speakers.

2 More detailed explanations of this methodology and the algorithms used are given in Bendifallah et al. (2023).

3 Of course, once the analyst makes the decision to treat a particular factor in a certain way in statistical analysis (as a continuous dimension, as a two-way or three-way categorical partition, or whatever), the categories involved become static. This is necessary for statistical analysis (at least using the standard tools).

4 There is also weak evidence ((P(est.>0)=.86) for an effect of engagement: informal collective and media members use more *violences sexistes et sexuelles* than *violences faites aux femmes*. See the supplementary material.

References

AUTHORS. (2024). Devenir féministe à Paris et Montréal: Récits de vie dans le corpus CaFé.

Alim, S. (2016). *Raciolinguistics: How Language Shapes Our Ideas About Race.* Oxford University Press.

Beaman, Jean, & Amy Petts. (2020). Towards a global theory of colorblindness: Comparing colorblind racial ideology in France and the United States. *Sociology Compass*, *14*(4), e12774.

Becker, Kara., Sameer Khan & Zimman, Lal. (2022). Beyond binary gender: creaky voice, gender, and the variationist enterprise. *Language Variation and Change*, *34*(2), 215-238.

Bendifallah, Lina, Julie Abbou, Heather Burnett & Igor Douven. (2023). Conceptual Spaces for Conceptual Engineering? Feminism as a case study. In revision for *Journal of Philosophy and Psychology.*

Bereni, Laure. (2012). Penser la transversalité des mobilisations féministes: l'espace de la cause des femmes. In Christine Bard (ed). Les féministes de la 2ème vague, Presses universitaires de Rennes, pp.27-41

Biber, D. (1995). Dimensions of register variation: A cross-linguistic comparison. Cambridge University Press.

Boltanski, L. & L. Thévenot. (1991). *De la justification. Les économies de la grandeur.* Paris: Gallimard.

Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Bourdieu, Pierre., & Boltanski, Luc. (1975). Le fétichisme de la langue. *Actes de la recherche en sciences sociales*, *1*(4), 2-32.

Bowker, G. C., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT press.

Cameron, Deborah & Don Kulick. (2003). *Language and sexuality*. Cambridge University Press.

Charity Hudley, Anne, Christine Mallinson & Mary Bucholtz. (2020). Toward racial justice in linguistics. *Language*, *96*(4), e200-e235.

Brown, G., Delessert, T. & Roca i Escoda, M. (2017). Du devoir marital au viol conjugal. Étude sur l'évolution du droit pénal suisse. *Droit et société,* 97, 595-614.

Butler, Judith. (1997). *Excitable Speech: A Politics of the Performative*. Routledge.

Butler, Judith. (1990). *Gender Trouble*. *Feminism and the Subversion of Identity*. Routledge.

Cheshire, Jenny. (2002). Sex and gender in variationist research. In J.K.Chambers, P.Trudgill & N.Schilling-Estes (eds.), *The handbook of variation and change.* Oxford: Blackwell. 423–43.

Delphy, Christine. (1977). *The main enemy: A materialist analysis of women's oppression (Explorations in feminism)*. Women's Research and Resources Centre Publications

Eckert, Penelope & Podesva, Robert. J. (2021). Non-binary approaches to gender and sexuality. *The Routledge handbook of language, gender, and sexuality*, 25-36.

Eckert, Penelope. (2017). Age as a sociolinguistic variable. *The handbook of sociolinguistics*, 151-167.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, *41*, 87-100.

Eckert, Penelope. (1989). *Jocks and burnouts: Social categories and identity in the high school*. Teachers college press.

Fassin, Éric. (2008). L'empire du genre: L'histoire politique ambiguë d'un outil conceptuel. *L'homme*, (3), 375-392.

Foucault, Michel. (1976). *Histoire de la sexualité.* Vol 1. Paris: Gallimard.

Gardenfors, Peter. (2014). *The geometry of meaning.* MIT press.

Guillaumin, Colette. (1992). *Sexe, Race et Pratique du pouvoir. L'idée de Nature*. Paris, Côté-femmes.

Hazenberg, E. (2017). Liminality as a lens on social meaning: A cross-variable analysis of gender in New Zealand English. Doctoral dissertation, Te Herenga Waka-Victoria University of Wellington.

Horvath, B., & Sankoff, D. (1987). Delimiting the Sydney speech community. *Language in Society*, 16(2), 179-204.

Kergoat, D. (2011). Comprendre les rapports sociaux. *Raison présente*, *178*(1), 11-21.

King, S. (2021). Rethinking race and place: The role of persona in sound change reversal. *Journal of Sociolinguistics*, 25(2), 159-178.

Kulick Don & Cameron Deborah. 2003. *Language and sexuality.* Cambridge: Cambridge University Press.

Labov, William. (1963). The social motivation of a sound change. *Word*, *19*(3), 273-309.

Leigh, Carol. (2011). Inventing sex work. In *Whores and other feminists* (pp. 225-231). Routledge.

Levon, Erez. (2015). Integrating intersectionality in language, gender, and sexuality research. *Language and Linguistics Compass*, *9*(7), 295-308.

Lochon, A. (2021). Trente ans de médiatisation des violences sexistes et sexuelles: L'exemple de deux journaux français. *Emulations-Revue de sciences sociales*.

Mair, P., Groenen, P. J. F., & de Leeuw, J. (2022). More on multidimensional scaling in R: smacof version 2. *Journal of Statistical Software*, 102(10), 1–47

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297.

Mautner, G. (2016). Checks and balances: How corpus linguistics can contribute to CDA. *Methods of critical discourse studies*, 154-179.

Meyerhoff, M., & Ehrlich, S. (2019). Language, gender, and sexuality. *Annual Review of Linguistics,* 5, 455-475.

Nugara Silvia. 2013. "Contacts linguistiques et nomination des violences faites aux femmes dans les documents internationaux. Féminicide/fémicide". *Travaux du CLAIX*, 24

Ochs, Elinor. (1993). « Indexing Gender ». In *Sex and Gender Hierarchies*, Barbara D. Miller, 335-58. Cambridge: Cambridge University Press.

Orellana, M. H., & Kunert, S. (2014). Du genre dans les discours institutionnels de lutte contre les violences faites aux femmes. *Synergies Italie*, (10).

Pavard, Bibia, Rochefort, Florence & Zancarini-Fournel, Michelle. (2020). *Ne nous libérez pas, on s'en charge*. La Découverte.

Pêcheux M., Fuchs C. (1975). Mises au point et perspectives à propos de l'analyse automatique du discours. *Langage* 37, 7-80.

Pfefferkorn, Roland. (2011). Rapports de racisation, de classe, de sexe... *Migrations, racismes, résistances*, (1), 193-208.

Podesva, R. J., Van Hofwegen, J. (2016). s/exuality in smalltown California: Gender normativity and the acoustic realization of/s. In Levon, E., & Mendes, R. B. (eds.). Language, sexuality, and power: Studies in intersectional linguistics, 16-88.

Posit team (2023). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. http://www.posit.co/.

Raus, Rachele. 2018. "Circulation et traduction française des termes à l'international : le cas d'« intersectionnalité »" *GLAD!* 5. Online.

Revelle, W. (2023). psych: Procedures for psychological, psychometric, and personality research [R package version 2.3.3]. Northwestern University. Evanston IL. https://CRAN.R-project.org/package=psych

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rickford, John. R. (1986). The need for new approaches to social class analysis in sociolinguistics. *Language and communication*, *6*(3), 215-221.

Rosa, Johnathan., & Flores, Nelson. (2017). Unsettling race and language: Toward a raciolinguistic perspective. *Language in society*, *46*(5), 621-647.

Rubin, G., & Butler, J. (1994). Sexual traffic. *A Journal of Feminist Cultural Studies*, *6*(2), 63-99.

Sankoff, David, & Laberge, Suzanne. (1978). The linguistic market and the statistical explanation of variability. *Linguistic variation: Models and methods*, *239*, 250.

Silverstein, Michael. 1985. « Language and the culture of gender: at the intersection of structure, usage and ideology ». In *Semiotic Mediation*, E. Mertz, R.J. Parmentier, 219-59. Orlando: Academic Press.

Simonin, D. (2016). *Le «travail du sexe». Genèses et usages d'une catégorie politique* (Doctoral dissertation, Université de Lyon).

Sorensen, Tanner, Sven Hohenstein & Shravan Vasishth (2016). « Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists », The Quantitative Methods for Psychology, vol.12: 175–200.

West, C., & Zimmerman, D. H. (1987). Doing gender. *Gender & society*, *1*(2), 125-151.

Wittig, M. (1992). *The Straight Mind*. Boston : Beacon Press.