Behavior-induced identity inferences, politico-ideological polarization and the emergence of sociolinguistically marked terms

Gerhard Schaden

Received: December 2023/Accepted: March 2024
Published on line: July 2024

Behavior-induced identity inferences, politico-ideological polarization and the emergence of sociolinguistically marked terms

Inférences d'identité, polarisation politico-idéologique et émergence de termes sociolinguistiquement marqués

Gerhard Schaden
Université de Lille & CNRS UMR 7110 LLF & UMR 8163 STL
gerhard.schaden@univ-lille.fr

Abstract

This paper investigates behavior-induced identity inferences (as exemplified by phenomena such as *virtue signaling*), and their implications for politico-ideological polarization and the emergence of linguistic forms associated with specific politico-ideological positions (such as *great replacement,* used nearly exclusively by members of the extreme right). Through three simulations, it will be shown that behavior-induced identity inferences consistently increase politico-ideological polarization. However, the emergence of expressions linked to particular politico-ideological stances requires the additional process of *schismogenesis*, that is, a differentiation process between the behavioral profiles of agents belonging to different groups.

Keywords: behavior-induced identity inferences, political polarization, sociolinguistically marked expressions, schismogenesis, computational sociolinguistics.

Résumé

Cet article examine les inférences d'identité engendrées par le comportement d'un agent (comme la *vertu ostentatoire*), et leurs conséquences à la fois sur la polarisation politico-idéologique et sur l'émergence de formes linguistiques associées à des positions politico-idéologiques spécifiques (par exemple, *grand remplacement,* utilisé presque uniquement par des membres de l'extrême droite). Dans trois simulations, il sera montré que les inférences d'identité exacerbent la polarisation politico-idéologique. Cependant, l'association entre une forme linguistique et une position politico-idéologique n'émerge qu'avec le processus

supplémentaire de *schismogenèse*, impliquant la différenciation des profils de comportement d'agents appartenant à des groupes différents.

Mots-clefs : inférences d'identité induites par le comportement, polarisation politique, expressions sociolinguistiquement marquées, schismogenèse, sociolinguistique computationnelle.

1. Introduction

Many words and phrases are strongly associated with particular politico-ideological stances. For instance, consider the expression *great replacement*. According to Wikipedia, this expression refers to the idea that the indigenous European population is being demographically and culturally replaced by non-European immigrants, with the encouragement or complicity of political elites.[1] Moderate rightists, even those who think that immigration should be tightly controlled, have generally avoided using this term, because it is clearly associated with an extreme right-wing political stance, and because it has become an identifier for the extreme right. Valérie Pécresse, candidate for the 2022 French presidential election on behalf of the center-right *Les Républicains* (the party of former presidents Jacques Chirac and Nicolas Sarkozy) was severely criticized for using the phrase in a political meeting as follows:

(1) After 10 years of inertia and poor choices, our destiny is in our hands again. I am convinced that we are condemned neither to the great demotion, nor to the great replacement. Yes, I believe in France.[2]

By uttering (1), V. Pécresse did not directly endorse the great replacement theory, nor did she state that she believed in it. Indeed, (1) has no textual entailment to the effect that the speaker has to believe that the great replacement would be ongoing or true.[3] However,

[1] See https://en.wikipedia.org/wiki/Great_Replacement

[2] The original quote is the following: « Après dix ans d'immobilisme et de mauvais choix, il ne tient qu'à nous de reprendre notre destin en main. J'ai la conviction que nous ne sommes condamnés ni au grand déclassement, ni au grand remplacement. Oui je crois dans la France. » https://www.huffingtonpost.fr/politique/video/valerie-pecresse-dit-grand-remplacement-a-son-meeting-et-ce-n-est-pas-la-premiere-fois_192263.html

[3] The fact that there is no textual entailment can be established by the felicity of the following: *I am convinced that we are condemned neither to the great demotion, nor to the great replacement, because neither of those exist.* Being condemned to is not a factitive predicate, neither when it is used positively, nor when it is used negatively like in (1). Contrast this with a case of clear textual entailment: *I have eaten an ice cream, #because/but ice cream doesn't exist.* Here, the continuation denying the existence of the direct object is clearly infelicitous.

using the term was enough to get her into hot water, and it was generally (and arguably correctly) reported as making an attempt to appeal to the voters of far-right polemicist Eric Zemmour.[4]

Such loaded expressions are not restricted to the European extreme right; similarly marked labels are *big pharma*, *welfare queen*, *female-read person*, *love jihad*, or *Judea and Samaria*, to name just a few. The use of such expressions identifies a speaker as belonging to a particular politico-ideological group, and members of opposed groups will typically avoid these expressions.

In linguistics and philosophy, there is an important and growing literature on such terms, usually under the headings of *expressive meaning*, *dog whistles* or *slurs* (see e.g., Nunberg, 2018; Burnett, 2020; Bolinger, 2020; Gutzmann, 2019; Davis & McCready, 2020).[5] One interesting aspect of such loaded words is that the associations they seem to contain "project out" from many contexts and end up being attributed to the speaker. They may project out even from direct quotations, and therefore, the mere act of quoting may not be permitted for specific categories of speakers (think, e.g., of the *N-word* in a North American context, where admissible use has become more and more restricted over time, and is approaching a total taboo against quoting by Caucasians).

The consensus view on such expressions seems to be that typically, there exists another, more unmarked way of referring to the same entity, and that the marked expression and the unmarked expression are denotationally equivalent.[6] There is, however, not so much consensus on how the marked meaning arises, and what it consists of. One hypothesis is that the negative meaning of something like the N-word is not part of the sign itself, but rather comes from the association of that particular word with a specific kind of user of the word (for the N-word, this would be white supremacists). This, in the terms of Peirce (1894), makes it an *index*.

In this paper, I investigate how and under which circumstances words (or linguistic elements or constructions more generally) can acquire such a politico-ideological association, and thus, become "code words" (see Khoo, 2017) for certain politico-ideological orientations. Studying such expressions will also involve an investigation of their social impact, and the social conditions required for their emergence. An important question is whether linguistic practices can play a role in the formation and maintenance

[4] Zemmour outperformed Pécresse in the first round of the presidential election, but neither qualified for the second round.

[5] These terms do not necessarily refer to the same phenomena. However, they all are (at least partially) interpreted as indicating group membership, which is why they are of interest here.

[6] This majority position is not universally shared; for an opposing view, see Croom (2013).

of social groups, and in the construction of social identity of an agent within their own ingroup and in opposition to competing groups.

The paper is structured as follows: in Section 2, I will introduce the notions of *identity* and *identity inference*, and their crucial importance in the domain of argumentation (that is, contexts where speaker and hearer interests are *a priori* not aligned). Section 3 lays out the assumptions made in the simulations which form the core of this paper, and how the simulations implement notions seen in Section 2. Sections 4-6 present and discuss one type of simulation each. Section 7 contains a more general comparison and discussion of the results of the simulations, both concerning politico-ideological polarization and the linguistic consequences of identity-relevant behavior, and also the evolutionary rationale of sociolinguistically marked words. Section 8 concludes the paper.

2. Preliminaries: identity, identity inferences, and argumentation

In this section, I will address what identity is, how it can be (linguistically) expressed to others by means of identity-relevant (linguistic) behavior, how and why identity (inferences) seem(s) to be especially salient in the context of (linguistic) suasion, and also, why simulations provide answers to difficulties raised by the mechanisms of performed identity.

2.1. Notions of identity

I will start by presenting several concepts of (social) identity, and their importance for an agent living in society. The issue of identity and identity construction has received much attention in many domains of the humanities and social sciences, and this paper cannot do justice to the literature in this domain. A common theme is however that identity functions as a bridge between the individual and society.

In one recent treatment, Moeller and D'Ambrosio (2021) distinguish three different forms of (social) identity: *sincerity*, *authenticity*, and *profilicity* – the former two based on Trilling (1971). They define *sincerity* (p. 10) as prevalent in more traditional societies, where identity was achieved via the sincere enactment of a social role – where the social role was assigned based on the situation (social class, gender, etc.) one was born into. In sincerity-based identity, individuals have no agency in creating their identity; they are expected to conform to some preexisting social role. With modernity and more dynamic societies, Moeller and D'Ambrosio (2021, p. 12f.) argue that another way of conceiving identity was born, namely *authenticity*. They take authenticity to be based on the idea that the "real" identity of an individual resides somewhere beneath the social roles. The third mode of having an identity is called *profilicity*. It refers to a type of identity that has to be curated – the profile. This type of identity is directly destined for social consumption, and it seems to be particularly salient in the times of social media. The main point where the

profile is distinct from older modes of identity is that it is created with a view towards second-order observation: our profile tells others how we would like to be seen *as being seen* (see Moeller & D'Ambrosio, 2021, p. 15). In the terms of this paper: Profilicity corresponds to identity created in order to induce an identity inference.[7]

According to *Identity Process Theory* (see Jaspal, 2014), self-identity is structured along at least two lines: first, its *content*, and second, a *value-affect dimension* (Jaspal, 2014, pp. 4-5). This separation is useful especially when examining components of a person's identity that do not necessarily fit together well, and the way an individual reacts to such inconsistencies. Jaspal (2014, p. 4) gives the example of men coming from socially conservative religious backgrounds and realizing that they are gay. In terms of the content of their identity, they thus add *homosexual* to their previous identity content, which also contains the property of being *religious*. Initially, religiosity is positively valued, whereas homosexuality is negatively valued. Updates on both these dimensions (content and affect) may lead to a more positive evaluation of homosexuality, and possibly, a downplaying or even elimination of the religious component (see Loewenthal, 2014, p. 327). While identity can change thus over time, Loewenthal stresses that a search for coherence is often involved in the evolution of personal identity.

This conception of self-identity has been exploited in the literature on political polarization (see, e.g., Mason, 2018a, 2018b). Notice that the political and ideological positioning or ideology of an individual is also part of their identity. Mason (2018a, pp. 868-869) distinguishes two types of ideology: an *issue-based* ideology (which is based upon policy attitudes; and thus: *content*), as opposed to what she calls an *identity-based* (or *symbolic*) ideology – which corresponds to the *value-affect dimension*. Her observation is that someone may identify strongly with (in the context of the contemporary USA) either liberals or conservatives, without, however, necessarily strongly favoring the policies associated with the corresponding group. She stresses the idea that partisanship seems to be based more on criteria of affect and social attachment than on adhesion to any particular issue of policy: partisans increasingly dislike each other, but there is often no direct connection to issue-based disagreements or object-level opinions (see Mason, 2018a, pp. 870, 885).

At the most fundamental level, identity describes who I am with respect to socially relevant categories. However, identity is at least potentially semiotically inert: I can very well secretly be a communist without anybody suspecting my true political leanings. In some circumstances, I would not want this type of information to become public, and I

[7] Notice that at least sincerity and profilicity contain elements of *public behavior*: Sincerity requires enactment, and the profile by definition has to be communicated. We will come back to this in Section 2.2.

may actively (try to) suppress signs of this aspect of my identity. But often (and increasingly so under the regiment of profilicity), I am not satisfied by simply *being X*; I want *you to know that I am X*. As already stated by Goffman (1969, p. 81), identity is therefore often related to behavioral markers (and thus, *performed*), and this is what we will turn to in Section 2.2.

2.2. Identity inferences and identity-relevant behavior

The performance and recognition of aspects of identity is conceptually complex. In what follows, I try to outline the most important cases we will need to worry about. As the title of the paper announced, this paper deals with *behavior-induced identity inferences* generally, that is, inferences by an agent A – based on aspects of the behavioral profile of another agent B – which categorize B into socially relevant classes. There are thus two elements at stake: the inferences by agent A, and the behavior by agent B. In this paper, I will thus only be interested in identity inferences that rely on observed (or at least: observable) behavior of agent B. This is not the only possible source of identity inferences: upon learning some property of B, e.g., that they are a citizen of Pakistan, A may draw (consciously or not) identity inferences about B, even without any interaction with or any other knowledge about B. Such inferences by preconception (or, if you will, by Bayesian priors) will not be studied here. Neither will I dwell for the purposes of this paper about the well-foundedness of the produced identity inferences.

Let us now turn to B's behavior. While most aspects of a person's behavior can be exploited for identity inferences, this is probably not the case for every single one of them.[8] Following Klein, Spears & Reicher (2007, p. 29), I will thus distinguish *identity-relevant behavior* from other types of behavior. Within identity-relevant behaviors, there are also several subclasses to be distinguished, most importantly, *strategic* and *non-strategic* identity-relevant behaviors.

Strategic identity-relevant behaviors are behaviors that have been deliberately crafted by an agent in order to trigger a determined identity inference in another agent. I will follow Davies (2021, p. 6) and call these in what follows *identity displays*.[9] In other scholarly

[8] For instance, assume that we observe some person sleeping on their side, rather than on their back. At the time of writing, I fail to see what this could possibly tell us about what that person is with respect to any relevant social category (beyond maybe them being members of the *ad hoc* category of "side sleepers"). That is not to say that sleeping on one's side could not be exploited for identity inferences, nor that it never will be (or never has been).

[9] The notion of *identity display* itself is attributed by Davies to Hample and Irions (2015). The term "display" ultimately goes back to the disciplines of ethology and zoology, but is used there in a way that does not correspond to Davies' use (or Goffman's use of *performance*). For instance, Nelson and Jackson (2007, p. 1659) define *display behavior* as "the use of *signals* that have been *evolutionarily*

traditions, this phenomenon is also referred to as the *impression management* of an agent (see Goffman, 1969), or their *identity performance* (e.g., Reed et al., 2012). The wearing of ceremonial garb (for instance, donning black formal clothing for Western funerals) would be an instance of identity display; Valérie Pécresse's use of "great replacement" in (1) probably would also qualify.

Concerning non-strategic identity-relevant behavior, there are two cases to consider. The behavior may either have not been performed deliberately, or the performer of the behavior may be unaware that the behavior can support some identity inference. For the former category, I will use the term of *behavioral identity leak*. Such leaks may come in several flavors. For instance, an agent may be aware that some behavior can give rise to identity inferences, but may not be able to control or suppress it. A (foreign language or regional) accent would be an instance of this: most speakers – even though they may be aware of it – cannot fully eliminate their accent, and classification based on accent is often rather accurate. A similar example would involve a tremor in the voice – indicating nervousness – when speaking in public. On the other hand, an agent may simply lack awareness of some of their identity-relevant behaviors: consider for instance the unintentional and automatic use of *man* or *he* for gender-neutral reference – regardless of whether an observer notices this or not.[10]

Another type of non-strategic identity-relevant behavior are cases where the behavior itself is performed deliberately, but where the performer is not aware of the likely identity inferences that behavior will trigger in the given context. Assume for instance that I travel in China with my family, and that I brought along a Winnie-the-Pooh t-shirt, which was gifted to me by my daughter. While visiting Beijing, I choose to wear this particular t-shirt. My choice is deliberate, in an attempt to please my daughter – and as such, this would qualify as an identity display (e.g., as a caring father). However, it so happens that I am entirely unaware of the fact that Winnie-the-Pooh is censored in China, and that my wearing this t-shirt will likely be interpreted by locals (and more importantly, the police) as criticizing Xi Jinping.[11] Here, while the behavior itself is known by the performer to be identity-relevant and has been performed deliberately, the provoked inference has not been

modified in a manner that enhances their capacity to convey information" [my emphasis]; see also the discussion in footnote 12 below.

[10] The use of singular *they* in such contexts – under the assumption that it is deliberate, and not automatic – would constitute an identity display. Arguably, it is the availability of the alternative in the larger speaker community which confers upon gender-neutral *he* its status as an identity-relevant behavior.

[11] See https://en.wikipedia.org/wiki/Censorship_of_Winnie-the-Pooh_in_China

targeted (and is in fact an effect of cultural ignorance), and fails on this account to be an identity display.

Let me stress at this point again that the identity inferences themselves (be they based on strategic or non-strategic behavior) may be accurate and/or justified, or inaccurate and/or unjustified. An emitter may try to anticipate and control an observer's identity inferences, but there is no effective way of constraining them. For instance, there is no way an emitter could control the identity inferences an observer draws upon viewing a visible tattoo – based on cultural and personal assumptions, some observer may very well classify the emitter as a criminal, and be completely wrong in their assessment.

This rather elaborate classification is necessary because the present article addresses identity-relevant behavior in general, rather than just focusing on identity displays – particularly attempting to neutralize the performer's intention. This is important because of two factors: First, the (inference of) an identity display is often used polemically in argumentative communication, as we will see in Section 2.3; and second, the assumption of whether some behavior has been produced with the intention to be recognized as an identity display may affect its interpretation, as has been pointed out by Goffman (1969) and Davies (2021).

Davies describes the issue as follows: most "standard" linguistic communication simply relies on the addressee recognizing the speaker's intention that led to the production of an utterance. For instance, if I were to greet you, you would recognize my intention to greet you, and as such, my greeting you would be successful. However, in many cases of identity-relevant behavior, the fact (or the mere suspicion) that it has been performed intentionally may cause it to fail to achieve its aim.[12]

Davies (2021, pp. 6-7) discusses the following example: assume that there is a new secretary who frequently – right after coming back from the bathroom – is in a much-

[12] The problem of the (non-)intentionality of identity-relevant behavior overlaps to some degree with the distinction between *signals* and *cues* made in evolutionary biology (see, e.g., Maynard Smith & Harper 2003, p. 3). A *signal* is defined as "any act or structure which alters the behaviour of other organisms, which evolved because of that effect, and which is effective because the receiver's response has also evolved" (Maynard Smith & Harper, 2003, p. 3). A *cue*, in contrast, is simply "any feature of the world, animate or inanimate, that can be used by an animal as a guide to future action" (Maynard Smith & Harper 2003, p. 3). At least behavioral identity leaks are not signals in this classification, but rather cues: A cocaine user's sniffing and nose rubbing are not behaviors that have evolved in order to induce the belief that the emitter is using cocaine (which would be necessary for them if they were signals); these are merely pharmacological consequences of the dominant mode of consumption of the drug, and thus cues which may guide the observer in their future behavior with respect to the emitter. Therefore, talk about identity (or similarly, *virtue*) "signaling" should be avoided.

improved mood, repeatedly scratches his nose, and also repeatedly sniffs. This pattern of behavior (which should be a behavioral identity leak) may lead me to believe that the secretary consumes cocaine in the bathroom, and that therefore, he is a cocaine user – in the same way the presence of spots on someone's skin may make me believe that this person has measles. Now assume that I know for a fact (or that I have formed the belief) that the secretary's behavior is performed with the intention of making me believe that he is a cocaine user (that is, that it is an identity display) – just as my greeting you was performed in the intention of making you believe that I greeted you. As Davies argues, I would (and should) not interpret this as indicating that he is a cocaine user: I should ask myself why the secretary would want me to think that he uses cocaine.[13]

The fact that – in order to be effective – identity displays often must not be seen as having been performed intentionally has an interesting consequence: since they sometimes depend on the observer's inability (or lack of will) to correctly identify the underlying motive, they are in these circumstances a kind of deceit. Trivers (2011) argued that in order to successfully deceive observers, there would be evolutionary pressure for self-deception. For instance, even if the intention to present myself as rich and successful were the true motive for my buying and wearing a Rolex, according to Trivers, my performance would be more successfully if that reason was not consciously accessible to me. I might even come up with and end up believing in rationalizations why I do not wear a Rolex because I want to appear rich and successful, but for some totally unrelated, and perfectly justifiable reason (such as "actually, I'm really interested in the history of watches and the craftsmanship involved, and Rolex contributed in such an amazing way to that history, that I just had to have one of those…"). On the other hand, Trivers notes that this should lead to an arms race with observers, who should be able to detect ever more elusive cues for motives underlying identity display. This pattern should play out particularly in contexts where the interests of emitter and observer are not perfectly aligned, such as contexts of suasion, and more specifically, argumentation.[14] This is what we will examine now.

[13] This example might suggest that identity displays should generally be treated with suspicion, whereas behavioral identity leaks are always reliable. Yet, many identity displays are unproblematic (think again about donning black formal clothing in mourning), even though uncontrolled (and uncontrollable) behavior should generally be more reliable than controlled behavior. The issue in this particular example seems to be that there generally should be no reason why someone should disclose a potentially fireable offense in such a convoluted manner.

[14] In contexts with completely aligned interests, it may be in the observer's best interest to conspire in the self-deception with respect to the ultimate underlying motives.

2.3. Identity inferences in (linguistic) suasion

The observation that identity inferences appear in argumentation is not a recent one. However, the analysis of more specifically recent communicative developments, notably in social media, have caused phenomena and notions that are linked to identity-relevant behaviors, as for instance the term of *virtue signaling*, to increasingly become part of public consciousness (see, e.g., Tosi & Warmke, 2020).

In traditional rhetoric, dating back to Aristotle, identity-relevant behaviors are (more or less) contained under the label of *ethos* (whereas the more issue-based argumentation is subsumed under the label of *logos*). It is important to notice, though, that there are important differences between the concepts of *virtue signaling* and Aristotle's *ethos* as a means of persuasion, and that they can even be seen as being functionally opposed. On the subject of *ethos*, Rapp (2010) states that a speaker has to display practical intelligence, a virtuous character, and good will, and given these properties, the character of the speaker may in itself be an element that convinces the audience. Yet, most often, the notion of *virtue signaling* is used as a means of *resisting* being persuaded. This is clearly the case in (2), where "mere virtue signaling" is opposed to "considered discourse".[15]

(2)

While it may be inevitable in communication that a speaker projects some vision of their own identity (see, e.g., Schulz von Thun, 1981, for a defense of this position), identity-relevant behaviors in argumentation are often remarked upon in political opponents, and, maybe as a consequence, strongly devalued. For instance, the accusation of "mere" virtue signaling in (2) implies that the point is not deeply felt or held by the arguer, or that the point argued for does not have any consequences for them. On the other hand, identity-relevant behaviors in politically (or ideologically) like-minded people are often overlooked – even though there is no reason to assume that this aspect is lacking. As we have seen in Section 2.2, there are good evolutionary reasons why one's own identity displays should

[15] https://twitter.com/Njdoc/status/1628497476750610435

remain unconscious or be misclassified by oneself, whereas (ideologically opposed) other's identity-relevant behaviors are more clearly discernible.

Given the polemical nature of accusations of identity display in discourse, it is not obvious that we will be able to discern clear and unambiguous instances of identity displays (such as *virtue signaling*) in any type of corpus data. Opponents to some cause or proposal are bound to find instances of identity displays where proponents do not perceive them. Both may be intimately convinced that their assertion or denial of virtue signaling is accurate, and they may both be deluded in doing so.[16] Self-serving biases in arguers will conspire to make their own identity-relevant behavior invisible to them, whereas the same biases will over-detect identity displays in opponents.

The difficulty of reliably and objectively detecting identity displays in real-world persons in real-world situations, as well as the tendency to both over-detect such displays in opponents, and to under-detect it in like-minded people, are probably not merely practical problems that could be overcome by clever experimental or corpus design. The distinction between identity display and arguing for a cause may very well be theoretically unsound. Therefore, I suggest that a more indirect approach to the study of identity-relevant behavior is at least advantageous, if not required.

Once it is suggested that neither emitters nor observers can be trusted with respect to identity-relevant behavior, a way of avoiding these issues is to use artificial agents, whose internal states are completely transparent to outside observers, and can be investigated objectively. In other words, this means resorting to simulations.

3. Simulating identity inferences and their linguistic consequences

In order to simulate behavior-induced identity inferences of an agent, and their consequences on a group level, we minimally need to operationalize the notions of *identity* and what it means to perform an *identity-relevant behavior*. For the purposes of the simulation, I will assume that the underlying identity is an agent's politico-ideological position in 2-dimensional conceptual space.[17] Contrary to the more sophisticated theories of identity we have seen in Section 2.1, I do not distinguish *content* from *value-affect* dimensions, in order to keep the simulation simple. The interpretable behavior at the base for identity inferences must be some observable pattern of (linguistic) behavior.

[16] This seems to me to be the main reason why the pioneering effort of Hample and Irions (2015), which relies on the self-assessment of arguers and of their conscious motives, may not be the best way of studying identity-relevant behavior in argumentation.

[17] This is obviously not a full representation of a person's identity, but many traits of (at least contemporary) identity are correlated with politico-ideological positions.

I will describe in Sections 3.1 (for identity) and 3.2 (for identity-relevant behavior) how these are implemented.

3.1. Identity as position in conceptual space

I model politico-ideological space as a 2-dimensional space, similar to the diagrams familiar from the *Political Compass*, as illustrated for some historical political leader and state ideologies in Figure 1.[18] Nothing particular hinges on the assumption of this being a 2D space; it should be seen as a convenient way of incorporating, but not reducing to, the familiar left-right opposition.[19] One could very well use higher-dimensional spaces (3D or beyond), but this will complicate the simulation, without necessarily giving an obvious conceptual advantage in our case, which involves exploring the interaction between identity and identity inferences.

[18] Figure 1 has been recreated after a crowd chart at the *Political Compass*. This picture is for illustrative purposes only. As the Political Compass notes: "We take no responsibility for any specific scores presented on this page".

[19] Similarly, the adoption of the axis-labels "left vs. right" and "authoritarian vs. libertarian" should not be taken as an endorsement of the compass's analyses, nor as an assumption of its universality across time and space. These labels are purely for convenience of presentation, and the simulations merely require the existence of an ideological space that can be modeled in (at least) 2 dimensions. They do not presuppose any particular content for these dimensions (that is, these dimensions might just as well be "monotheism vs. polytheism" on the x-axis, and "free will vs. predestination" on the y-axis).

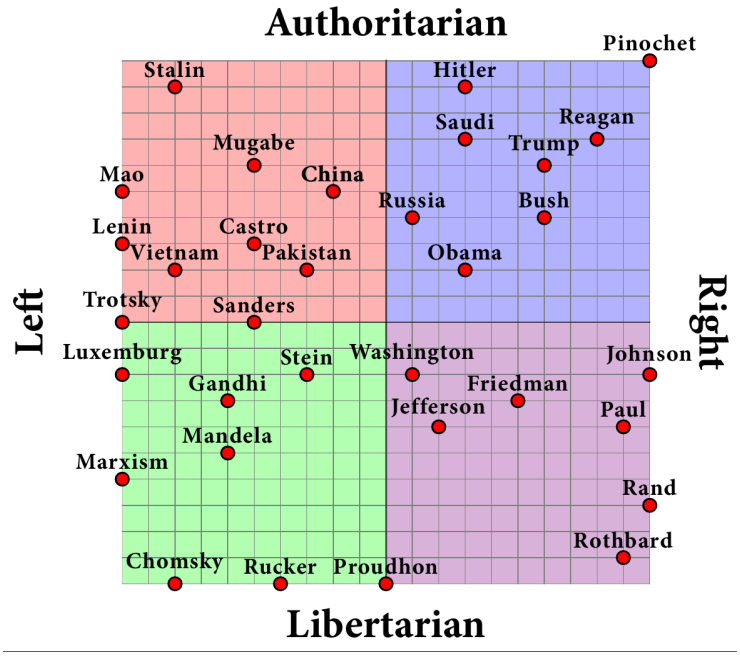

Figure 1. Political Compass crowd chart

At the beginning of a simulation, agents are placed at random in this politico-ideological space, using a beta distribution, which assigns more weight to the center than to the extremes.[20] Therefore, agents are initially placed clustered around the center of the grid, and extreme positions are relatively rare. An agent's identity is their position on the grid, and can be given with two coordinates: the perfect centrist resides thus at (0,0), whereas the position of Pinochet in Figure 1 would be (1,1), and Proudhon would be located at (0,-1).

3.2. Identity-relevant behavior as linguistic behavior in situations with options

At the beginning of the simulation, agents are assigned at random a frequency distribution of the use of different *sociolinguistic variables*.[21] In sociolinguistics, sociolinguistic variables are linguistic variants that can in principle be substituted one for the other, but which are correlated with some social or intentional community. In principle, a sociolinguistic

[20] The shape parameters α and β of the beta distribution used in the simulations were $\alpha = \beta = 3$. Since a beta distribution distributes values in the interval [0,1], and I projected the 2D grid into the space [-1,1], the outcome of the random draw n was scaled by $2n - 1$.

[21] As far as I am aware, nothing in this simulation requires the variation to be necessarily interpreted as linguistic variation, as opposed to any other type of variation (as in clothing, food, hairstyle, etc. – or any combination thereof). Variant A vs. B in Situation 1 could in principle concern wearing skirts vs. trousers, and they might be about having short vs. long hair in Situation 2, and eating vegan vs. omnivorously in Situation 3, etc. – given a suitable definition of what is to be understood as a *situation*. That being said, I will only discuss the linguistic side of variation in this paper.

variable could correspond to any linguistic entity (different pronunciation of a phoneme, different forms of grammatical items, etc.), but for ease of exposition, I will generally and tacitly assume that relevant sociolinguistic variables in this paper are different lexical items, such as words or phrases.

Such a frequency distribution can be represented as follows:

	% variant A	% variant B
Situation 1	0.50	0.50
Situation 2	0.90	0.10
Situation 3	0.20	0.80
Situation 4	0.33	0.67
Situation 5	0.30	0.70

Table 1. Sample frequency distribution for sociolinguistic variables

For each situation in Table 1, there are two linguistic variants that could be used in that situation. I assume that all agents understand in principle both variants and could use them in a given situation. The variants have no intrinsic value at the beginning of the simulation since they are assigned at random.

An agent with the profile as shown in Table 1 – if confronted with Situation 1 – will either use variant A or B with 50% probability. When confronted with Situation 2, such an agent will use 90% of the time variant A, and only 10% of the time variant B. Notice again that I assume that all variants for all situations are different, so we would have 10 different words here, where two words are respectively synonyms, and both can be used with equal efficiency in a given situation.

Any one of these 10 variants could end up being associated with different politico-ideological formations.[22] For an example of this, let us reconsider the N-word in a North American context. The situation would be one where the speaker needs to refer to an Afro-American. Here, the N-word would be variant A, and *Afro-American* variant B. Denotationally, these forms are equivalent (they refer to the same set of persons), but they do not express or reveal the same attitude of the speaker with respect to their referents.
I will refer to the full frequency distribution of an agent as depicted in Table 1 as their *Linguistic Usage Profile*. The linguistic usage profile is an observable behavior, and forms thus the agent's identity-relevant behavior.

Summing up: The initial placement in the politico-ideological grid is random, and so is the linguistic usage profile. In this way, there is no intrinsic social meaning associated with any word, and thereis also no intrinsic bias towards one area of the politico-

[22] A *politico-ideological formation* refers to any cluster of agents at some place of the politico-ideological grid. I use the term *formation* to avoid conveying that these clusters correspond to any organized or long-lived groups.

ideological space – apart from the initial distribution of the agents on the board, which is biased towards more central positions by means of a beta distribution.

3.3. The progression of a simulation

In the simulation, we want to observe the change in both an agent's identity, and in their interpretable behavior. Every agent will do two things: 1) they will adapt their linguistic usage profile in order to make it more similar to other agents' linguistic usage profiles in their vicinity; and 2) they will adapt their identity (that is, their position on the grid) in order to move towards a neighborhood where other agents have a linguistic usage profile that is more similar to their own. (1) describes thus a modification of their identity-relevant behavior, whereas (2) describes a modification of their identity.[23]

How exactly is this done? The way an agent's position is changed remains stable throughout the simulations: within some distance d from their current position, the agent compares other agents' linguistic profile to their own linguistic profile. Within this window of distance d in the 2D space from the agent's position, other agents will influence them, whereas agents further removed from the agent's position will not influence them.[24] If on average, the speakers on the left (within distance d) are closer to the agent's linguistic profile than the speakers on the right (within distance d), then the agent will move to the left by some distance δ (and vice versa); if on average, the speakers that are more libertarian are closer to the agent's linguistic profile than the speakers that are more authoritarian, then the agent will move by some δ towards the libertarian side (and vice versa).Thus, the two ideological dimensions of the space are treated as being independent.

The modification of the linguistic profile differs from simulation to simulation. There is, however, one element that will be stable across simulations, and this is *local conformity*. The agent will observe other agents' linguistic profile within the window of distance d, calculate the average of other agents' linguistic profiles, and adapt their own linguistic profile accordingly, once again by moving towards the local mean for some γ.

[23] This modification of identity is a factor contributing to *homophily* (see Jackson 2008, pp. 68-69), that is, to the fact that agents with similar identity-relevant behavior will end up in similar regions of the 2D board.

[24] This window of distance d may be thought of as something akin to the *Overton-window*, that is, the window of publicly acceptable statements (see https://en.wikipedia.org/wiki/Overton_window).

	% variant A	% variant B
Situation 1	0.50	0.50
Situation 2	0.90	0.10
Situation 3	0.20	0.80
Situation 4	0.33	0.67
Situation 5	0.30	0.70

	% variant A	% variant B
Situation 1	0.50	0.50
Situation 2	0.50	0.50
Situation 3	0.40	0.60
Situation 4	0.67	0.33
Situation 5	0.50	0.50

Table 2. Sample distribution (left) and sample average of other agents (right)

Assume that the left entry in Table 2 represents some agent's linguistic profile, and similarly, that the right entry corresponds to the average linguistic profile of all other agents within the first agent's relevant neighborhood. For Situation 1, the agent's profile is already aligned with the local mean, so nothing will change here. However, for Situations 2–5, the agent is not aligned with the local mean. Therefore, local conformity has the effect that the agent will move towards the local mean by some γ, which will lead to a new profile of the agent as given in Table 3 (assuming a γ of 0.05).

	% variant A	% variant B
Situation 1	0.50	0.50
Situation 2	0.85	0.15
Situation 3	0.25	0.75
Situation 4	0.38	0.62
Situation 5	0.35	0.65

Table 3. Updated linguistic usage profile of the agent in Table 2

Every simulation will contain 100 agents assigned a politico-ideological position and a lexical profile at random. Every individual simulation will imply 1000 rounds of updating the agents' linguistic profiles and their position in the politico-ideological grid, in order to give the agents the time to settle into a pattern, and thus, to achieve *sorting*, in Mason's (2018a, 2018b) terms. Furthermore, in order to minimize the effect of chance, every simulation of 1000 rounds will be run 100 times.

3.4. Aims of the simulation

I will be especially interested whether the parameters of the simulation allow to derive a linguistic situation where some variant is strongly associated with some politico-ideological formation or the other. In the real world, such a state of affairs seems to prevail, and therefore, one issue is to determine under which circumstances this obtains. A further question is whether the paradigm could lead to a *language schism*, that is, a situation where speakers start to develop differing coding systems based on their politico-ideological

formation. This would be the case if not only *some* linguistic variants came to be ideologically marked, but *all* of them, and in a categorical and consistent way. Third, I will investigate whether identity inferences in themselves lead to politico-ideological polarization.

Before moving on to the simulations, let me address again why and how this setup corresponds to identity-relevant behavior. In the simulations, the agents only have a position in a grid (i.e., their *identity*), and a linguistic usage profile (i.e., their *identity-relevant behavior*). They do not try to convince other agents of anything; they only adapt their way of speaking to their neighbors, and they adapt their position on the politico-ideological grid in order to move to an area more in line with their way of speaking. The way of speaking and the politico-ideological position are both connected to the identity of the agents, and this is the only information available in the simulations. Therefore, we have completely eliminated any other component from the equation, and any change we will see will be causally related to agents' identity-relevant behavior (and two other parameters in the simulation introduced below).

In the remainder of the paper, I will perform three different types of simulation: in Section 4, I will investigate a configuration with pure local conformity (i.e., the mechanism outlined above). Section 5 shows a configuration where agents belong to preexisting politico-ideological groups, and where there is a tendency to diverge from the identity-relevant behavior of agents in outgroups. Section 6 investigates the consequences of the process of divergence, but where groups are emergent, and do not exist prior to the simulation.

4. Simulation 1: Pure local conformity

4.1. Presentation

In the first simulation, we will observe what happens when linguistic adaptation is limited to local conformity, and we will investigate the consequences on linguistic diversity and politico-ideological polarization. The algorithm used here can be described informally as follows: given a position in the 2D grid, and a linguistic profile, the agent checks who their neighbors are within some distance to their own position. Then, based on this, the agent adjusts their position in the grid, in order to move to a region where other agents are more similar in their linguistic profiles. At the same time, the agent also adjusts their linguistic profile in order to be closer to the local mean. This procedure of updating the politico-ideological position and the linguistic profile is repeated 1000 times. This forms one single run of the simulation. Since the process is strongly dependent on random factors, 100 simulation runs have been performed, and their outcome compared.

4.2. Results of the simulation

In this simulation, the initially very moderate average politico-ideological polarization increased to extreme degrees. Politico-ideological polarization of a population is defined to be the mean distance of the population from the center. The minimal political polarization one might obtain would be 0 if all agents were located at the center (that is, at coordinates (0,0)). The maximal polarization would be 1, and such a situation would occur if all agents are located at one of the four corners (that is, at coordinates (-1,-1), (-1,1), (1,-1), and (1,1)).[25]

The initial placement on the grid is done by random, using a beta distribution; as can be seen in Figure 2, this results in a mean polarization of around 0.3 for the initial placement. After 1000 rounds of adjustment, polarization has gone up to an average of 0.8 after the final round of adjustments.

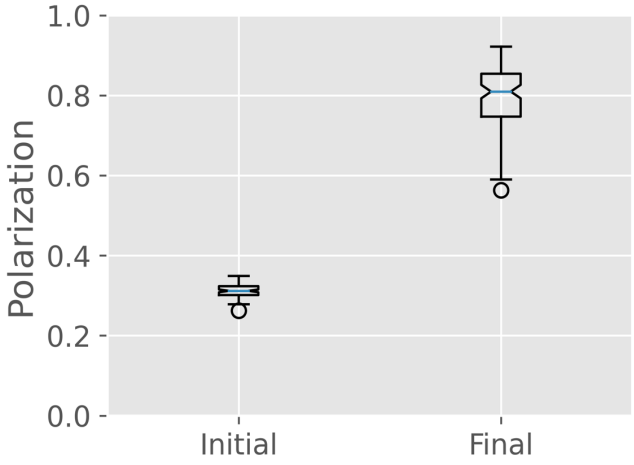

Figure 2. Identity inferences increase polarization

It is interesting to observe that this strong increase in political polarization is not necessarily in itself correlated with an increase in the diversity of agents' linguistic usage profiles. Linguistic profiles are assigned uniformly at random at the beginning, making sure

[25] The mean polarization is calculated as follows: given a population of n agents with positions of (x_n, y_n) respectively, sum the absolute values of x_n and y_n, and divide the result by $2n$. This measure calculates divergence with respect to the center (0,0) of the grid. It would be possible to use another measure for polarization, which would be based on the agents' average position on the board. These two measures would diverge strongly if the entire population had converged toward an extreme position (say, (1,1)). In this case, the measure used here would indicate a polarization of 1, whereas a measure with respect to the population mean would indicate a polarization of 0. For the simulations performed in this paper, there does not seem to be any major difference between these two ways of measuring polarization, so only the first one is reported.

that each agent will initially assign a probability mass of at least 0.02 to any form.[26] Linguistic diversity is calculated as the divergence of an agent from the population mean. If linguistic diversity equals 0, everybody in the population has exactly the same linguistic profile.[27]

As is shown in Figure 3, the result of this simulation is that linguistic diversity ends up being eliminated. As a consequence, there will be no word that would be marked for any politico-ideological formation, since all agents, whatever may be their position within the 2D grid, speak exactly in the same way at the end of an individual simulation run.

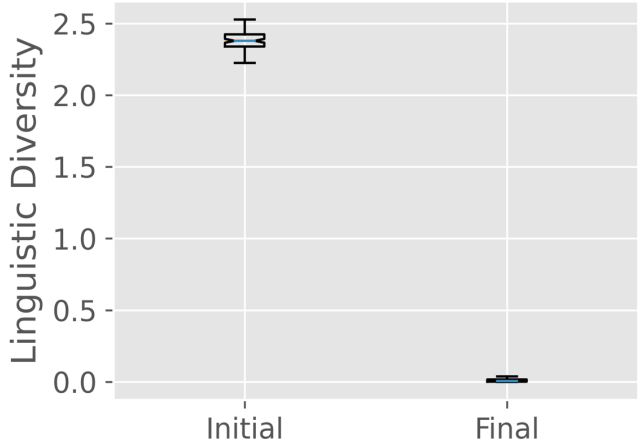

Figure 3. Linguistic diversity is eliminated with pure local conformity

4.3. Discussion

In this first simulation, we end up with great politico-ideological polarization, while there is no or only minimal linguistic differentiation remaining between agents. The absence of linguistic differentiation is to be expected, in the sense that there is no mechanism of differentiation between agents, and their only available adjustment strategy is convergence

[26] Thus, the proportion for variant A is drawn uniformly at random in the interval [0.02,0.98], and the proportion for variant B is calculated as *1 – proportion of variant A*.

[27] More technically, the linguistic diversity of a population is calculated as follows: Given a population of *n* agents with their linguistic usage profiles, we calculate for each agent the distance of their linguistic usage profile from the average linguistic usage profile in the population by taking for each cell in the table the population average, from which we subtract the agent's corresponding value, take the absolute value of this, and finally sum these values across all cells. For instance, consider the agent on the left in Table 2, and the population mean on the right of Table 2. The distance for the cell in Situation 2, variant A is 0.4, and the global distance of the agent from the population mean is 2.24. The linguistic diversity in the population is then defined as the sum of these individual distances to the population mean, divided by the number of agents *n*.

towards the local mean. Therefore, at best, linguistic diversity could be maintained at the initial level.

However, the amount of polarization is unexpected, especially given that there is hardly any diversity this polarization could feed on. This may be due to the fact that the computational agents have full access to other agents' profiles, and that they can detect differences that a human would not be sensible to, if having to infer the frequencies from actual linguistic data. Yet, it is an interesting result of this simulation that high levels of politico-ideological polarization can obtain even in configurations where there is basically no linguistic differentiation between agents, and where initially no pattern of speaking was indicative of any politico-ideological formation.

Yet, since our aim is to find a way of generating a situation where certain linguistic choices are correlated with certain politico-ideological formations, this simulation fell short. The way the simulation will have to be changed needs to involve some way of introducing divergence between agents' linguistic profiles, and not only of convergence towards some local mean. This will be explored in Section 5.

5. Simulation 2: Schismogenesis with fixed politico-ideological formations

In order to achieve our aim of provoking linguistic diversification, I will introduce a new concept, namely *schismogenesis*. This term comes ultimately from the American anthropologist Gregory Bateson (1935), but I will use the notion as found in Graeber and Wengrow (2021, pp. 56-57), in order to refer to "people's tendency to define themselves against one another, and thus, to become more extreme in opposition, as the process continues".[28]

[28] It is not clear to me that the use of the term in Graeber and Wengrow (2021) really conforms to the original intent of Bateson (1935). Bateson defines schismogenesis in terms of *behaviors*, which have to show a strict differentiation between ingroup-only behavior, and outgroup-only behavior.

In any case, the idea in Graeber and Wengrow (2021, p. 57) is specifically meant to apply this process to relations between societies:

> Bateson was interested in psychological processes within societies, but there's every reason to believe something similar happens *between* societies as well. People come to define themselves against their neighbors. Urbanites thus become more urbane, as barbarians become more barbarous. If 'national character' can really be said to exist, it can only be as a result of such schismogenetic processes: English people trying to become as little as possible like French, French people as little like Germans, and so on. If nothing else, they will all definitely exaggerate their differences in arguing with one another. [emphasis in the original]

In practical terms for the simulations, *schismogenesis* can be seen as the result of a negative reaction to identity displays from members of opposing groups. The basic intuition is the following: Assume that there is some person X from group Z that I strongly dislike, and who has a highly salient (to me) linguistic behavior for Situation 2: instead of using variant A 70% of the time and variant B 30% of the time (like I do), X uses A 30% of the time, and B 70% of the time. In order not to be mistaken for a member of Z in Situation 2, I modify my linguistic profile such as to further increase the difference between me and X: I increase the use of variant A to a probability of 0.75 and decrease the use of variant B to 0.25. If X dislikes the group Y and me to which I belong, X may similarly decrease the use of A to 0.25 and increase the use of B to 0.75. Therefore, we have both become more extreme in our linguistic behavior, and this may also have repercussions for our respective identities.

Since schismogenesis necessarily involves some distinction between the ingroup and (possibly several) outgroup(s) of an agent, we need to provide such groups, and assign each agent with an ingroup and outgroups.

5.1 Presentation

In the present simulation, schismogenesis with respect to politico-ideological formations is implemented as follows: each agent is attributed one of five ingroups, based on their position on the grid – centrist, right-authoritarian, left-authoritarian, left-libertarian or right-libertarian. An agent may change their group if they drift out of their old group by updating their position. However, the groups themselves are preexisting and immutable: no attempt is made to identify clusters based on agents' position on the grid, and the positions of the groups do not change through the simulation (even if the number of agents per group may fluctuate, and a group can become void of members). The basic layout is depicted in Figure 4.

If we substitute *societies* by *politico-ideological groups* in the quote above, we get the notion of *schismogenesis* as used in this paper.

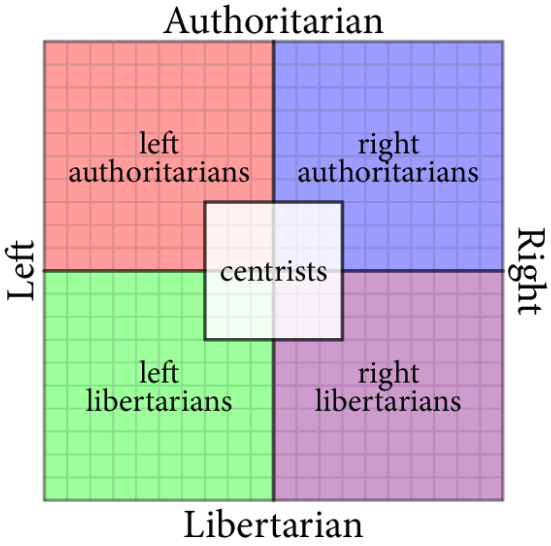

Figure 4. Fixed politico-ideological formations

As before, at each turn, any given agent will conform locally, which is defined as adjusting their linguistic profile by moving towards the mean of their ingroup, and they will also adjust their position on the grid in order to arrive at some position where other agents' linguistic behavior is more similar to the agent's linguistic profile.

However, additionally, each agent will be assigned uniformly at random one agent of each outgroup at the beginning of each turn of the simulation, and they will try to diverge in their linguistic profile from these members of the outgroups. This means that the member of an outgroup the agent will diverge from is not stable from round to round – assuming the agent is centrist, they might get assigned in the first round agent x from the right-authoritarian group, and in the second round agent y. Agent y may not have been a member of the right-authoritarians in the first round, but ended up there by drift. The reason agents from outgroups are assigned at random is the following: If everybody were assigned the same agent to diverge from in each round, this would give that agent some power over the first agent. If the outgroup agent to diverge from would be the same for all agents of (all) other groups, this would mean that this specific agent would have some platform that makes their opinion accessible to (all) other groups. This might be appropriate if we want to model the influence of journalists or other opinion makers (think of cases like Tucker Carlson or Trevor Noah), but, since we are preoccupied with the impact of identity inference itself, rather than signal boosting in the media, I kept as much of an egalitarian outlook as possible – and thus the assignment at random.

The divergence process itself is implemented as follows: the agent will determine for every selected agent from the outgroups the linguistic element they are maximally different from. Then, they will attempt to further increase the distance by adding some distance δ. Should there be a case where an agent is maximally distant with respect to the same

linguistic element to two different outgroups, but in opposite directions, then the agent will move towards the mean point of the two distributions, in order to maximize distance to both individual elements.

In order to see how this works, consider the agent x on the left in Table 4, and two selected agents from two different outgroups (e.g., assuming that x belongs to the centrists, the agent in the center might be a left-authoritarian, and the agent on the right a right-libertarian).

	% var. A	% var. B
Sit. 1	0.52	0.48
Sit. 2	0.64	0.36
Sit. 3	0.81	0.19
Sit. 4	0.82	0.18
Sit. 5	0.66	0.34

	% var. A	% var. B
Sit. 1	0.99	0.01
Sit. 2	0.66	0.44
Sit. 3	0.97	0.03
Sit. 4	0.94	0.06
Sit. 5	0.89	0.11

	% var. A	% var. B
Sit. 1	0.01	0.99
Sit. 2	0.54	0.46
Sit. 3	0.79	0.21
Sit. 4	0.81	0.19
Sit. 5	0.67	0.33

Table 4. The focal agent (left) and two outgroup agents (center and right)

Comparing their profile to the profile of the agent in the center, agent x will determine that they are maximally different with respect to Situation 1 – the distance is of 0.47 with respect to the variables in Situation 1, but only 0.02 with respect to Situation 2. Therefore, agent x will try to maximize the distance with respect to the agent in the center by increasing that distance by some δ (say 0.05). However, assume that x has to diverge from both the agent in the center, and the agent on the right of Table 1. In both cases, the distance is maximal with respect to Situation 1. In this case, x will try to maximize distance to both agents by moving by some δ towards the point that is equidistant from these agents (which would be here 0.5).

5.2. Results of the simulation

As was the case in the first simulation, we can observe that politico-ideological polarization increases considerably between the beginning and the end of simulation runs. This is illustrated in Figure 5, which reports the mean politico-ideological radicalization at the beginning and the end of the 100 simulation runs. Notice that the average degree of polarization is lower than in the case of the first simulation (where the mean was around 0.8).

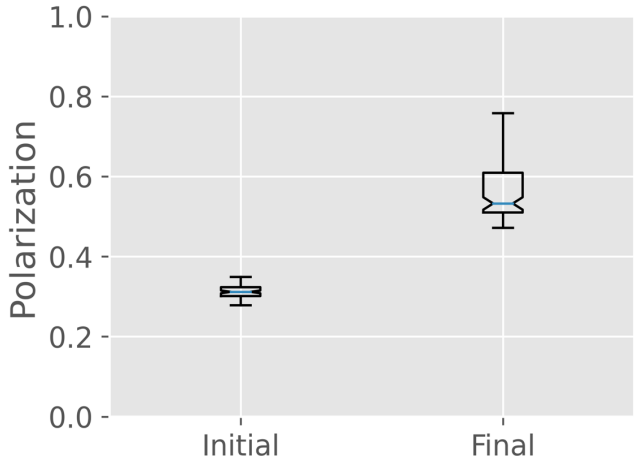

Figure 5. Politico-ideological polarization with fixed groups

Contrary to the preceding setup, the initially occurring linguistic diversity remains stable, or is slightly increased, and will not be erased by ongoing updates, as can be seen in Figure 6. Notice that the spread we obtain at the end does correspond, however, to a different kind of linguistic diversity than the initial random assignment. At the end of the simulation, the population will have converged to very similar usage profiles for 2–4 linguistic elements. However, the remainder will be associated with some politico-ideological group or the other. This pattern will be examined more in detail in Section 7.2. Therefore, we have obtained in this simulation a feature we looked for at the beginning, namely the development of linguistic forms that are strongly indexically associated with some politico-ideological formation, and which outsiders seek to avoid. Remember that a real-world example of such a linguistic phenomenon presented in the introduction was *great replacement* – which is confined within radical fringes of the right wing, and which more moderate opponents to immigration generally tend to avoid.

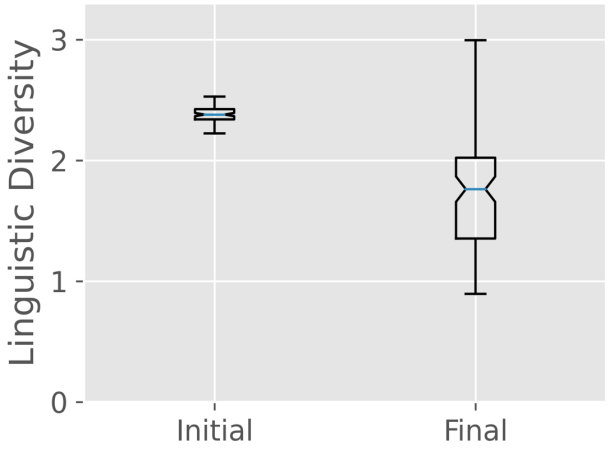

Figure 6. Linguistic diversity is maintained with schismogenesis and fixed groups

5.3. Discussion

Once again, we observe in this simulation a consistent tendency towards politico-ideological polarization in the population of agents. This tendency will remain a constant feature in all simulations presented in this paper.

Since we observed no tendency of linguistically correlating a given form to some politico-ideological formation in Simulation 1 whatsoever, but such correlations did form in the present simulation, we can conclude that these associations are dependent on the process of schismogenesis. That is, without a mechanism to diverge behaviorally in their identity-relevant behavior from agents perceived as members of some outgroup(s), all agents in the population converge to the same linguistic profile.

6. Simulation 3: Schismogenesis with agents as centers of their own politico-ideological formations

In the last simulation, we abandon the idea of preexisting and immutable politico-ideological groups and set up each agent as the center of their own politico-ideological formation – the right one, as the agent would certainly see it, if they were a person of flesh and blood.

This simulation is meant to explore the influence of the perspective dependence of the politico-ideological landscape, where any viewpoint is judged with respect to the proximity or distance with respect to an agent's place in the grid. Additionally, we abandon the extremely artificial constraint of imposing a fixed number of preexisting groups in the simulation, also in order to see what impact the existence of fixed groups has.

6.1. Presentation

The processes of local conformity and schismogenesis towards outgroups are once again both available. The ingroup is calculated around a window of 0.3 distance of the agent: anything that has a distance of less than 0.3 in both authoritarianism and left-right will be in the ingroup; the outgroups are defined by a distance of more than 0.5 on either direction in the 2D space. Such a layout has been depicted in Figure 7 for a (very) moderate right-libertarian. Notice that there is a gap between members of the ingroup, and members of ideological outgroups. Agents that are neither in an agent's ingroup, nor in one of their outgroups, will have no influence whatsoever on that agent's linguistic or politico-ideological profile.

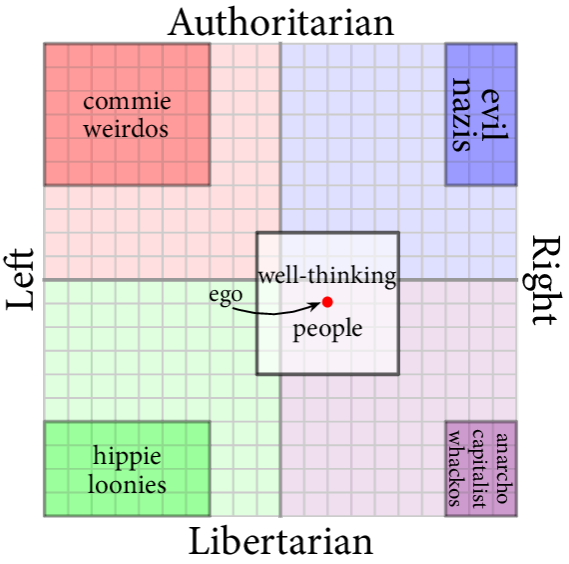

Figure 7. The Agent as the center of their own politico-ideological space

The ingroup of 'well-thinking people' in Figure 7 is centered on the ideological position of the agent (depicted by the red dot). Notice that – because the agent is located on the right – there is more opposition space on the left than on the right; and since the agent is located on the libertarian side, there is more opposition space on the authoritarian side. Notice that in this setup, nothing requires all opposition spaces to be filled. A more extreme agent may not find any opposable terrain to one or two sides of their space, whereas some particular opposition space may seem inflated beyond what a more centrist agent would conceive of. Such a state of affairs is illustrated in Figure 8. The extreme positioning at the upper right corner has two notable consequences: first, the ingroup space is compressed because there is not enough space left to the right and the authoritarian side in order to

get the full window size.[29] Second, the absence of space towards these two sides also conditions that there can be only one opposition space, here based on the left-libertarian spectrum. Notice that the space attributed to left-libertarians by the agent contains territory in the right-authoritarian quadrant that goes beyond the territory of centrists in Figure 4 in the simulation with fixed politico-ideological groups. Thus, positions that other agents may see as being (moderately) right-authoritarian will be qualified by the agent in Figure 8 as left-libertarian.

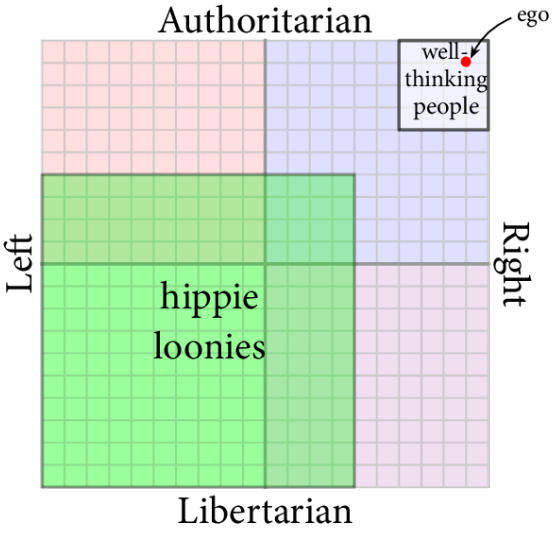

Figure 8. A distorted view of politico-ideological space

Furthermore, the agent whose position is illustrated in Figure 8 will probably not see themself as a radical and extreme ideological outlier, but rather see the majority of the population in the thralls of an extreme (and wrong) politico-ideological memeplex.

6.2. Results of the simulation

In this simulation, we once again observe a strong increase in politico-ideological polarization, as is illustrated in Figure 9. With respect to the preceding simulation, the mean polarization is higher than what we obtained with fixed and preexisting groups, but with considerably less spread. However, once again, the average level of polarization remains lower than in the first simulation, where there was no schismogenesis.

[29] The simulation thus assumes that there are absolute edges of politico-ideological space, and does not assume that space is a torus, which would correspond to the *horse-shoe theory* of politics: extreme right and left (or libertarians and authoritarians) would be assumed to be adjacent in position in such a case.

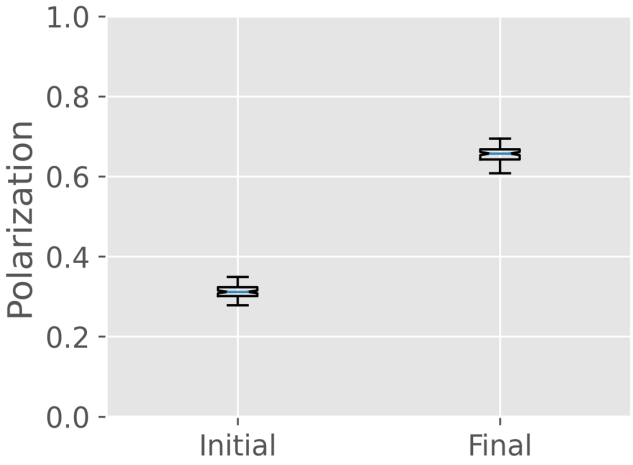

Figure 9. Politico-ideological polarization without fixed groups

When we look at linguistic diversity (as depicted in Figure 10), we observe a higher degree of diversity in this particular simulation. The reason is probably that on the population level, there is no clear definition of how many groups need to be distinguished from one another, and so, in many cases, all five linguistic variables used in the simulation will be recruited in order to achieve a politico-ideological effect.

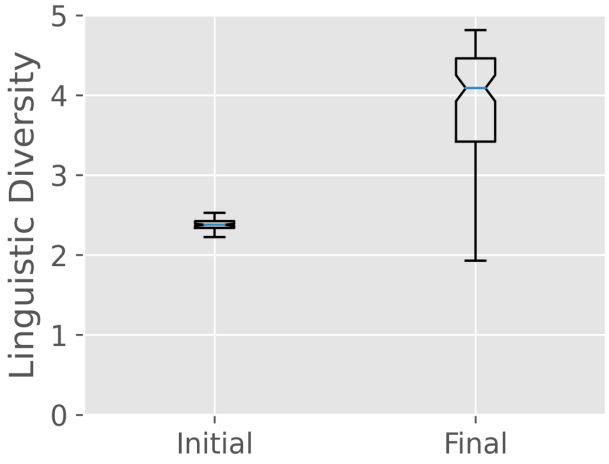

Figure 10. Linguistic diversity is maintained with schismogenesis and without fixed groups

7. General discussion

Now that we have seen the three different simulations, and the impact of minor variations on the outcome, let us take a step back, and consider what they have in common, where they differ, and whether more general lessons can be deduced from them.

7.1. Identity inferences increase politico-ideological polarization

A major result of the simulations is that the occurrence of politico-ideological polarization is independent of the process of schismogenesis, that is, the process making (parts of) the

linguistic profiles of agents more dissimilar. If anything, the occurrence of schismogenesis seems to reduce the degree of polarization, as is illustrated in Figure 11: We obtain by far the highest degree of polarization in the first simulation lacking schismogenesis (noted 'No SG'), whereas the lowest degree of polarization occurs in the second simulation (noted 'SG + FG'), combining schismogenesis and fixed ideological groups. We obtain an intermediate average, with, however, very little spread, if each individual is taken to be the center of their own politico-ideological world (noted 'SG - FG').

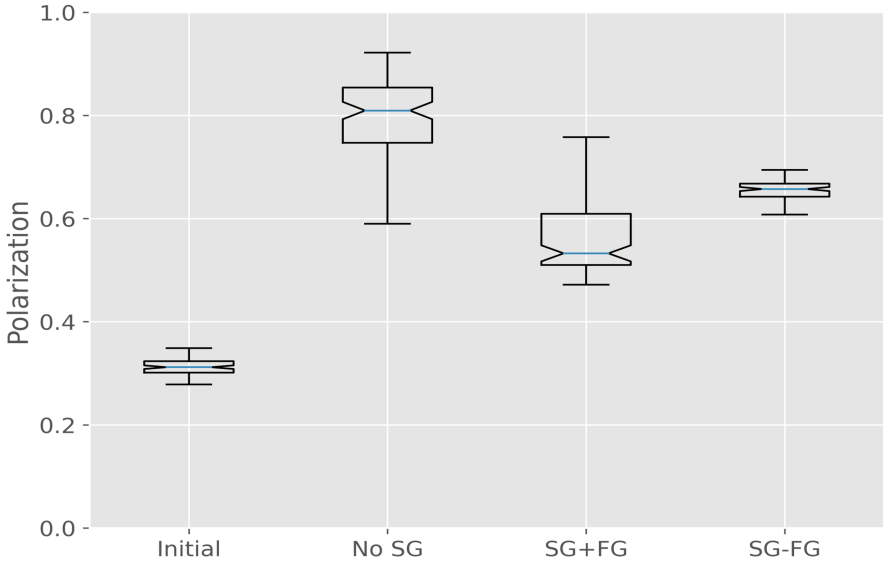

Figure 11. A comparison of the degree of polarization across the 3 simulations

Extrapolating from these results and their difference to the preceding type of simulation to the real world is not obvious. While there are arguably no politico-ideological formations completely independent from persons, there are politico-ideological formations that are relatively independent from persons, in the sense that they preexist their members and have some sort of ideological continuity across time: political parties and civic institutions. Strong parties and civic institutions may thus be able to limit political polarization somewhat – which would confirm the observations and the argument made by Putnam (2000). However, if Putnam is to be believed, these institutions have been in decline in recent decades, and so, their capability to impose culturally recognized politico-ideological groups may already have been jeopardized, which would mean that our current world corresponds most closely to the third simulation, where each agent is the center of their own politico-ideological world.

7.2. The evolution of sociolinguistic markedness

Let us now consider the linguistic results of our simulations. We have seen that we obtain linguistic differentiation between agents placed on different spaces in the politico-ideological grid only through schismogenesis. A real-life motivation for schismogenesis as

implemented in the simulations would be the desire not to speak like a member of an outgroup, in order not to be mistakenly identified as such a person. Notice, however, that schismogenesis here applied only on the level of *linguistic behavior*, and not on the level of the politico-ideological placement: there is no mechanism in the simulation that could correspond to a desire of an agent to become politically more or less radical in itself.

We have seen that in principle and under the assumptions made in the model, even extreme politico-ideological polarization does not depend on maintaining some linguistic diversity. Remember that in the simulations, five contexts were used where linguistic elements could in principle be distinctive for some politico-ideological formation. The limitation to five contexts was done for reasons of convenience: in principle, every couple of contextual variants in a given language might be used to express some politico-ideological formation.

In Figure 12, I have presented these linguistic differences in a slightly different way than what we have seen in Sections 4–6, which were based on the divergence per agent from the population mean, rather than on the divergence per word. Assume for instance that the population mean for Situation 1 is to use variant A with probability 0.7, and variant B with probability 0.3. Assume that Agent 62 uses variant A and variant B for Situation 1 at precisely the population mean: the score for divergence will be 0. However, if the population mean for some Situation 2 would be 0.5 for variant A and 0.5 for variant B, and the agent used variant A at 1.0 and variant B at 0, the divergence in this particular case would be 0.5. Figure 12 shows for every linguistic item at the end of every simulation run how divergent the use of a given linguistic choice is with respect to the population mean.

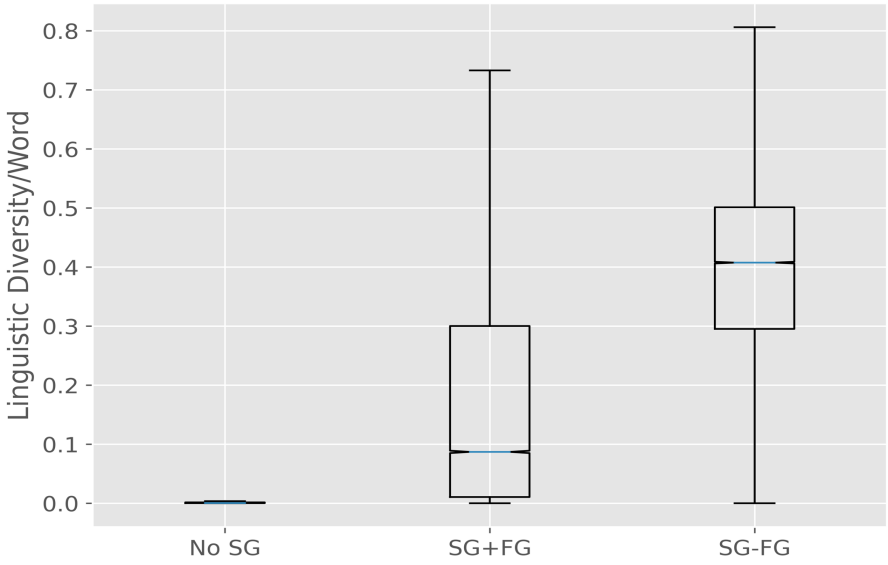

Figure 12. A comparison of the linguistic distance per linguistic item with respect to the population mean at the end of the simulation

If we compare the outcomes of the simulations as depicted in Figure 12, we see that without schismogenesis (noted 'No SG'), the entire population will converge to some common language use, where there is virtually no variation left whatsoever. We observe some variation per linguistic item in both the simulation with fixed groups (noted 'SG + FG'), and also in the simulation without fixed groups (noted 'SG-FG'). However, there is considerably more variation on a per-linguistic item basis in the latter than in theformer configuration. I already stated that this has to do with the number of distinctions that have to be made on the population level: a rather low number in the case of the fixed groups, and a potentially much larger number in the case without fixed groups.

The important issue here is, however, not variation *per se*, but variation as indicative of some politico-ideological positioning of the agents. This can be illustrated with the aggregate distribution of the linguistic distance per linguistic item, comparing the beginning and the end of Simulation 3 (i.e., schismogenesis without stable groups), as depicted in Figure 13. The initial distribution of the linguistic variation on a word basis is done by random assignment, and is basically flat for distances up to 0.45, trailing off for higher values. The distribution at the end of the simulation looks very different: here, the distribution is bimodal – there is one peak at very low values, and another, much higher one, at high values of divergence from the population mean with respect to that word. Thus, there is a tendency for agents of either having very low divergence with respect to some word usage from the population mean, or a very high divergence to the population mean. In the former case, there is a population-wide agreement on how to use these words; in the latter, there are subgroups in the population with widely diverging uses with respect to those linguistic choices. In other words, variation has become to an important degree an

indexical (in the sense of Peirce), and the choice of the use of one word vs. another in some context by some agent is now pointing towards some political or ideological viewpoint.

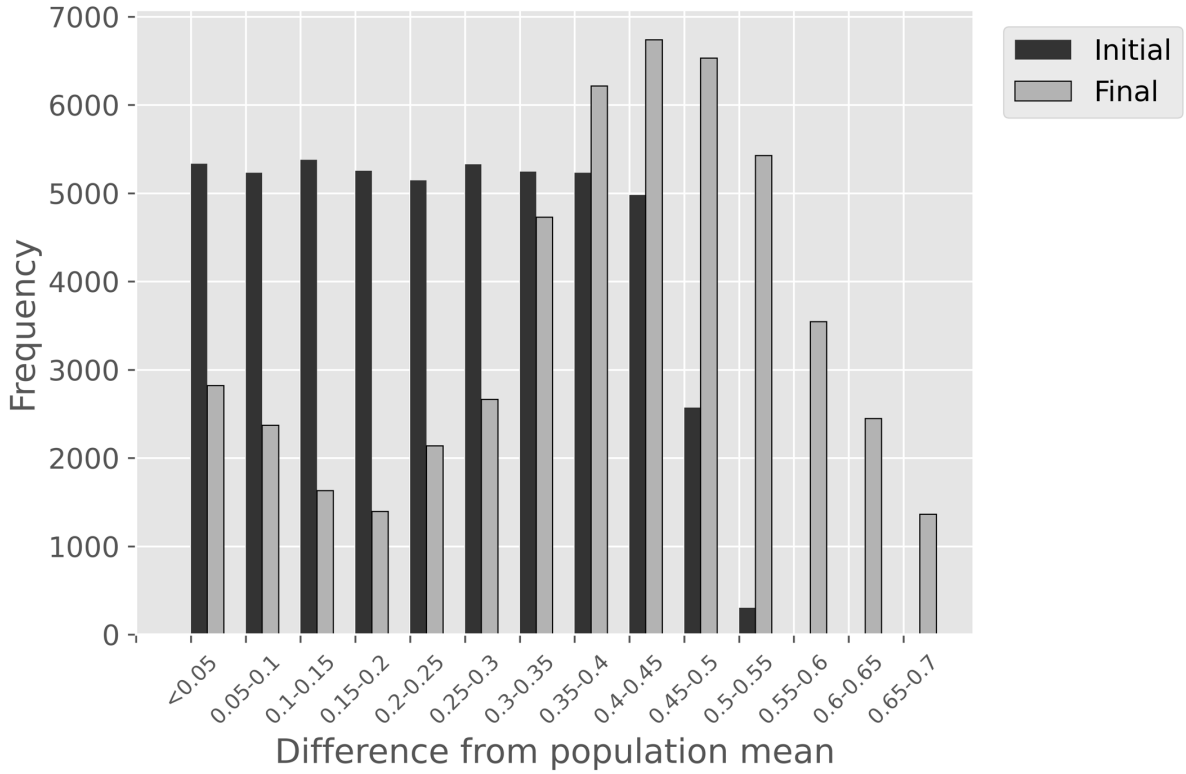

Figure 13. A comparison of the distance per linguistic item with respect to the population mean at the beginning vs. the end of Simulation 3

On a more individual level, let us zoom in on the effect of this pattern in one run from the second type of simulation (with schismogenesis and fixed groups). Let us begin with the following two agents, representative of two clusters of opposing politico-ideological formations: the agent at the left of Table 5 at the extreme right-authoritarian end of things, and the agent on the right the extreme left, and very mildly anti-authoritarian pole – which makes this agent a representative of the left libertarians in this simulation setup.[30]

[30] For reasons of presentation, all numbers in Tables 5 and 6 have been rounded to 2 decimals. Remember that a centrist is someone who is at a distance of less than 0.3 from (0,0) in both dimensions of left-right and libertarianism-authoritarianism.

Position: (0.95, 1)

	% variant A	% variant B
Situation 1	0.00	1.00
Situation 2	0.85	0.15
Situation 3	0.84	0.16
Situation 4	0.82	0.18
Situation 5	0.70	0.30

Position: (-1, -0.05)

	% variant A	% variant B
Situation 1	1.00	0.00
Situation 2	0.00	1.00
Situation 3	0.65	0.34
Situation 4	0.80	0.20
Situation 5	0.56	0.44

Table 5. On the left, the linguistic profile of a sample, right-authoritarian agent; on the right, the profile of a sample left-libertarian agent

Notice the categorical opposition with respect to Situation 1: the right-authoritarian always uses the second variant here, whereas the left libertarian always uses the first variant. There is also a big difference with respect to Situation 2: the right-authoritarians have a very robust preference towards the first variant, whereas the left-libertarians will only use the second. However, these instances of categorical or strongly opposing preferences are not generally observed: Preferences go into the same direction for Situation 3–5, and are nearly identical for Situation 4.

Therefore, the relevant elements for diverging identity inference inducing behaviors in this example and between these two agents are Situation 1 and Situation 2. Even with minimal interaction, an agent belonging to one of these two groups would be able to determine whether their interlocutor is a member of the ingroup, or a member of an outgroup.

Now, let us compare these agents with the three agents in Table 6 – on the left, a left-authoritarian, in the center – a centrist, and on the right – a right libertarian.

Position: (-0.2, 0.4)

	% var. A	% var. B
Sit. 1	0.00	1.00
Sit. 2	0.00	1.00
Sit. 3	0.81	0.19
Sit. 4	0.82	0.18
Sit. 5	0.66	0.34

Position: (0.2, -0.295)

	% var. A	% var. B
Sit. 1	1.00	0.00
Sit. 2	1.00	0.00
Sit. 3	0.97	0.03
Sit. 4	0.94	0.06
Sit. 5	0.89	0.11

Position: (0.05, -1)

	% var. A	% var. B
Sit. 1	0.00	1.00
Sit. 2	1.00	0.00
Sit. 3	0.79	0.21
Sit. 4	0.81	0.19
Sit. 5	0.67	0.33

Table 6. The linguistic profiles of three representative agents: a left-authoritarian, a centrist, and a right-libertarian (from left to right)

The most important distinction, and the only one that is categorical, occurs with respect to Situation 1 and Situation 2 – but since we have 5 groups, this is not sufficient. Notice, however, that we do get the 4 possible choices for categorical variation: the centrist has

$\left\langle \begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix} \right\rangle$, the left authoritarian has $\left\langle \begin{smallmatrix} 0 & 1 \\ 0 & 1 \end{smallmatrix} \right\rangle$, the left libertarian has $\left\langle \begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix} \right\rangle$, and in the right libertarian, we obtain $\left\langle \begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix} \right\rangle$. These elements are, however, not fully compositional: while for Situation 2, we can see that the pattern ⟨0,1⟩ is associated with the left, there is no association of authoritarianism or libertarianism for any configuration in Situation 1. Therefore, it is the full pattern in Situation 1 and Situation 2 that gives us holistically the (most) information with respect to group membership, but it cannot (here) be broken down into component parts meaning *left* and *libertarian*. There is thus no pressure towards some sort of compositional expression of indexical values in the simulations.

But it still is the case that in simulations with schismogenesis, indexical correlations of the use of linguistic forms and politico-ideological positioning have emerged, whereas such correlations were entirely absent from the simulation without schismogenesis. Therefore, schismogenesis is the causal factor underlying the development of such correlations in the simulations, and at least a possible causal factor for such correlations in actual language use.

Now that we have identified how identity inferences influence group formation and linguistic behavior, the question remains of what the social function of such patterns may be. This will be addressed in Section 7.3.

7.3. Sociolinguistic markedness and tag-based communication

For social animals, it is very often important to identify which individual belongs to one's own group, and is thus due cooperative behavior, and which individual belongs to a competing group, and should thus not be cooperated with. Group affiliation is in principle not easily perceivable, but there may be evolutionary pressure to develop some perceivable trait in order to indicate cooperative potential, exemplified in a thought experiment by a "green beard" in Dawkins (2006 [1976], p. 89). A green beard would be an example of a biologically inherited system of marking of cooperative potential. However, similar markers could also be cultural in nature (think, for instance, of tattoos or scarifications indicating lineage, or of social groups using a specific way of dressing or hairstyle to demarcate themselves from outgroups). In order to be reliable, such indicators need to be difficult to fake, and preferably, should not impose big costs on authentic cooperators.

Cohen (2012) makes the case that such cultural markers (which she calls "tags") could also be present in natural languages, and she argues more specifically that accent in natural languages could function as an indicator of the group affiliation of some individual (and thus, of their cooperative potential). For most individuals, it is difficult to display a native accent in a foreign language, and the cost of the acquisition of one's native accent is nonexistent. Therefore, accent satisfies the requirement of being difficult to fake, and has the additional benefit of being essentially cost free for group members. In a response to

Cohen (2012), Dediu and Dingemanse (2012, p. 606) point out that there need not exist one single system of tags in a natural language, but that there could be a much richer set of possibly interlocking tags, having differing associated costs. However, while they propose alternative cultural tags, they do not consider alternative *linguistic* tags in their response; nor do they take into account the possibility of shifting allegiances, which add further complications.

While the issue of the (linguistic) marking of group affiliation is too complicated to be dealt with in any satisfactory way in this paper, I want to point out that group affiliation is multifaceted, and that different affiliations may be operating on differing time scales. This was probably always the case to some degree, but is more salient in the contemporary world. That is, at some moment *t*, an agent may belong to different professional, socio-cultural and religious groups whose interests may be more or less divergent. Similarly, these group affiliations may have changed (or change) through time (through immigration, conversion, etc.).

And while my accent may reliably identify me as a native speaker of language X, thus indicating my regional or ethnic origins and allegiances, it is unable to differentiate any group loyalties I may have acquired later in life. However, these group loyalties may influence who I tend to cooperate with and who I oppose. Therefore, tags for tracking such later group affiliations might be useful. Indeed, there are linguistic phenomena precisely suited for this purpose. Consider, for instance, *jargon*. Acquiring the jargon of a professional field or leisure activity, and using it correctly, requires time, dedication, and is difficult to fake. Once acquired, however, the maintenance cost is low. Such a tag is appropriate for groups one enters and is unlikely to oppose once joined.

However, it is not suitable for tracking politico-ideological affiliations in the contemporary world, which may shift dramatically in relatively short periods of time, and where many individuals can be expected to change their politico-ideological affiliation several times within their lifetime. Once their affiliation has shifted, individuals may strongly oppose the members and beliefs of their previous politico-ideological allegiance.[31] In such circumstances, the ideal tag would be moderately easy to acquire, but would also require at least some moderate cost of maintenance. A constantly shifting set of sociolinguistically marked words could provide such a tag: while it may be easy enough to fake one specific linguistic marker of a group one does not belong to, it is much more difficult to do so for a multitude of linguistic markers. Additionally, sociolinguistically marked words are often very emotionally charged, which adds to their reliability.

[31] Think, for instance, of the French *Nouveaux Philosophes*, who started their careers mostly as Maoists, but turned then to liberal or even conservative positions.

Summing up, sociolinguistically marked words – especially if the set of sociolinguistically targeted words is constantly changing – offer a means of indicating group membership in a context where membership in the group, but possibly also the politico-ideological ideology of the group itself, is in constant flux. A question arising from this idea, but which has to be left to future research is whether more traditional societies (even highly complex ones) with stable politico-ideological formations have used or still use the sociolinguistic marking of words in a manner similar to contemporary Western societies. If the assumptions made in this section are correct, this should not be the case.

8. Conclusions and perspectives

In this paper, I have investigated in three different simulations the circumstances in which a politico-ideological marking of some linguistic form can emerge (as for instance, *pro-life* is associated with conservatism, whereas *pro-choice* is associated with liberalism). I have found that such an association emerges only in simulations where there is some means of inducing differentiation between the linguistic profiles of agents. Such a tendency towards a behavioral differentiation between ingroup and outgroup was identified as *schismogenesis,* and depends on a negative reaction to other agents' identity-relevant behavior. Beyond the question of *how* sociolinguistically marked words can emerge in a simulation, I also speculated on the reason *why* such words could be useful in natural languages as spoken today: Changing constellations of sociolinguistically marked words provide a useful tag for tracking group membership in relatively unstable groups.

With respect to the existing literature on the issue of politico-ideological polarization (see e.g., Mason, 2018a, 2018b; Mercier, 2020; Dorst, 2022), the paper had the modest aim of investigating whether polarization could be the result of feedback loops between changes in identity, caused by agents' identity-relevant behavior and identity inferences; and changes in identity-relevant behavior caused by an agent's identity. I showed that this can be the case under some circumstances within an artificial society. The main result with respect to polarization is that behavior-induced identity inferences are sufficient to lead to (an important degree of) politico-ideological polarization in all experimental setups considered, irrespective of whether there is a tendency of agents to differentiate themselves in their identity-relevant behavior.

The results of the present paper call for three types of further research: The first issue is to investigate how general the results actually are. It would be preferable to establish mathematical proofs showing which parameters determine polarization and the emergence of sociolinguistically marked words, and how realistic it is to assume these parameters to be active in real life. Second, since it is probably inevitable that some kind of identity-relevant behavior should occur in any type of conversation (at least, this is the view taken in theories of communication like Schulz von Thun, 1981), and since we do not (always)

observe the extreme levels of polarization in our societies, it stands to reason that there have to be counteracting tendencies and forces which lead to decreasing levels of polarization, but whose precise nature is unknown at this time. It seems in any case likely that factors mitigating polarization are weaker in online social media-based communication (with their current incentive structure using likes and dislikes) than in offline face-to-face communication. Third, the simulations were concerned with conceptual space, but it could just as easily be interpreted as concerning physical space. Applied to dialects, the data suggests that the emergence and maintenance of dialectal variation depend on the will of speakers to differentiate themselves from their neighbors. It remains to be seen whether this is indeed the case.

Acknowledgments

All simulations and data analysis have been performed with Python; the diagrams have been drawn with Matplotlib (see Hunter, 2007). The code and the raw data should be available with the article; they are also available from the author on demand.

I would like to express my sincere gratitude to Bart de Boer for his feedback on an early version of this paper. My thanks also go to the three anonymous reviewers, whose critiques and suggestions helped refine the quality and clarity of this work. Finally, I wish to thank the editors of *Lexique* (Kristel van Goethem, Véronique Lagae, Dany Amiot and Delphine Tribout) for their guidance and encouragement throughout the publication process. All remaining errors and omissions are mine alone.

References

Bateson, G. (1935). Culture contact and schismogenesis. *Man*, *35,* 178–183.

Bolinger, R. J. (2020). Contested slurs: Delimiting the linguistic community. *Grazer Philosophische Studien*, *97*(1), 11–30. https://doi.org/10.1163/18756736-09701003

Burnett, H. (2020). A persona-based semantics for slurs. *Grazer Philosophische Studien, 97*(1), 31–62. https://doi.org/10.1163/18756735-09701004

Cohen, E. (2012). The evolution of tag-based cooperation in Humans: The case for accent. *Current Anthropology 53*(5), 588-616. https://doi.org/10.1086/667654

Croom, A. M. (2013). How to do things with slurs: Studies in the way of derogatory words. *Language and Communication*, *33*(3), 177–204. https://doi.org/10.1016/j.langcom.2013.03.008

Davies, A. (2021). Identity display: Another motive for metalinguistic disagreement. *Inquiry, 64*(8)*,* 861-882. https://doi.org/10.1080/0020174X.2020.1712229

Davis, C., & McCready, E. (2020). The instability of slurs. *Grazer Philosophische Studien, 97*(1), 63–85. https://doi.org/10.1163/1875635-09701005

Dawkins, R. (2006 [1976]). *The selfish gene. 30th anniversary edition*. Oxford University Press.

Dediu, D., & Dingemanse, M. (2012). More than accent: Linguistic and cultural cues in the emergence of tag-based cooperation. *Current Anthropology, 53*(5), 606-607.

Dorst, K. (2022). Rational Polarization. *The Philosophical Review*. http://dx.doi.org/10.2139/ssrn.3918498

Graeber, D., & Wengrow, D. (2021). *The dawn of everything: A new history of Humanity*. Allen Lane.

Goffman, E. (1969). *The presentation of self in everyday life*. Penguin Books.

Gutzmann, D. (2019). *The grammar of expressivity*. Oxford University Press.

Hample, D., & Irions, A. (2015). Arguing to display identity. *Argumentation, 29*(4), 389-416. https://doi.org/10.1007/s10503-015-9351-9

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90-95. https://doi.org/10.1109/MCSE.2007.55

Jackson, M. O. (2008). *Social and economic networks*. Princeton University Press.

Jaspal, R. (2014). Social psychological debates about identity. In R. Jaspal & G.M. Breakwell (Eds). *Identity Process Theory: Identity, Social Action and Social Change* (pp. 3-19). Cambridge University Press.

Khoo, J. (2017). Code words in political discourse. *Philosophical Topics, 45*(2), 33-64. https://doi.org/10.5840/philtopics201745213

Klein, O., Spears, R., & Reicher, S. (2007). Social Identity Performance: Extending the strategic side of SIDE. *Personality and Social Psychology Review, 11*(1), 28-45. https://doi.org/10.1177/1088868306294588

Loewenthal, K. M. (2014). Religion, identity and mental health. In R. Jaspal & G.M. Breakwell (Eds). *Identity Process Theory: Identity, Social Action and Social Change* (pp. 316-334). Cambridge University Press.

Mason, L. (2018a). Ideologues without issues: The polarizing consequences of ideological identities. *Public Opinion Quarterly, 82*: 866-887. https://doi.org/10.1093/poq/nfy005

Mason, L. (2018b). *Uncivil agreement: How politics became our identity*. University of Chicago Press.

Maynard, S. J., & Harper, D. (2003). *Animal Signals*. Oxford University Press.

Mercier, H. (2020). *Not born yesterday: The science of who we trust and what we believe.* Princeton University Press.

Moeller, H.-G., & D'Ambrosio, P. J. (2021). *You and your profile: Identity after authenticity*. Columbia University Press.

Nelson, X. J., & Jackson, R. R. (2007). Complex display behaviour during the intraspecific interactions of myrmecomorphic jumping spiders (Araneae, Salticidae). *Journal of Natural History*, *41,* 1659-1678. https://doi.org/10.1080/00222930701450504

Nunberg, G. (2018). The social life of slurs. In D. Fogal, D. W. Harris & M. Moss (Eds.), *New work on speech acts* (pp. 237–295). Oxford University Press.

Peirce, C. S. (1894). What is a sign?

https://peirce.sitehost.iu.edu/ep/ep2/ep2book/ch02/ch02.htm

Putnam, R. D. (2000). *Bowling alone: America's declining social capital*. Simon & Schuster.

Rapp, C. (2010). Aristotle's rhetoric. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2010). Metaphysics Research Lab, Stanford University.

Reed, A., Forehand, M.R., Puntoni, S., & Warlop, L. (2012). Identity-Based Consumer Behavior. *International Journal of Research in Marketing, 29*(4), 310–321. https://doi.org/10.1016/j.ijresmar.2012.08.002.

Schulz von Thun, F. (1981). *Miteinander reden: 1. Störungen und Klärungen. Allgemeine Psychologie der Kommunikation*. Rowohlt.

Tosi, J., & Warmke, B. (2020). *Grandstanding: The use and abuse of moral talk*. Oxford University Press.

Trivers, R. (2011). *The Folly of Fools: The Logic of Deceit and Self-Deception*. Basic Books.

Trilling, L. (1971). *Sincerity and authenticity*. Harvard University Press.