Croatian Journal of Philosophy

Vol. XXV · No. 75 · 2025

Croatian Journal of Philosophy Vol. XXV · No. 75 · 2025

Institute of Philosophy

ISSN 1333-1108

Croatian Journal of Philosophy

1333-1108 (Print)
1847-6139 (Online)

Published by
Institute of Philosophy
Ulica grada Vukovara 54/IV, 10000 Zagreb, Croatia
www.ifzg.hr

Available online at https://cjp.ifzg.hr
and https://hrcak.srce.hr/en/cjp

Croatian Journal of Philosophy is published three times a year. It publishes original scientific papers in the field of philosophy.

Croatian Journal of Philosophy is indexed in *The Philosopher's Index, PhilPapers, Scopus, ERIH PLUS* and in *Arts & Humanities Citation Index* (*Web of Science*).

Croatian Journal of Philosophy is published with the support of the Ministry of Science, Education and Youth of the Republic of Croatia.

Instructions for Contributors

All submissions should be made through our online submission system: https://ojs.srce.hr/index.php/cjp/about/submissions. Submissions must be in English and formatted to be double-spaced with suitably wide margins, an A4 page size, and automatic page numbering.

Articles are normally no longer than 8,000 words of main text including bibliography. The Journal will consider longer papers, but, once these limits are exceeded, authors should bear in mind the editorial policy that the acceptance bar raises with increasing length.

Manuscripts should be compiled in the following order: cover page; title; abstract (not exceeding 200 words); keywords (3 to 6); main text; appendices (as appropriate); references.

All the authors of a paper should include their full names, affiliations, postal addresses, telephone and fax numbers and email addresses on the cover page of the manuscript. If a paper is co-written, one author should be identified as the Corresponding Author. The cover page must be submitted as a separate document. All submitted manuscripts must be prepared for blind review, with revealing acknowledgements and self-identifying references removed.

Sources are cited in the text by the author's last name, the publication date of the work cited, and a page number if needed, e.g. (Barber 2007: 324). Full details appear in the reference list in which the year of publication appears immediately after the author's name:

Barber, A. 2007. "Linguistic Structure and the Brain." *Croatian Journal of Philosophy* 21 (7): 317–341.

Williamson, T. 2013. *Identity and Discrimination*. Oxford: Wiley-Blackwell.

The publication of a manuscript in the Croatian Journal of Philosophy is expected to follow standards of ethical behavior for all parties involved in the publishing process: authors, editors, and reviewers. The journal follows the principles of the Committee on Publication Ethics (https://publicationethics.org/resources/flowcharts).

CROATIAN
JOURNAL
OF PHILOSOPHY

Vol. XXV · No. 75 · 2025

Articles

Introduction

This is indeed a very special issue of Croatian Journal of Philosophy. *I have the (sad) honor to be the guest editor of this issue that serves both as a proceedings issue to the Philosophy of Language and Linguistics conference and as an In memoriam issue to our beloved Dunja Jutronić who sadly passed away this summer. There are three* in memoriam *pieces for Dunja—by Boran Berčić, Barry Smith, and Mirela Fus-Holmedal—each offering a personal account of who she was to them and reflecting the powerful friendships formed over the years. We thank them for their heartful celebrations of Dunja.*

It is only fitting that I also dedicate a few words to Dunja. To begin, I want to say that it was a tremendous honor when Dunja invited me to serve as guest editor for this issue of Croatian Journal of Philosophy *during the 2024 Philosophy of Language and Linguistics conference. That honor now carries a note of sorrow, as Dunja is no longer with us.*

When I think about Dunja, there are three words that come to my mind: force of nature. And she truly was; taking care of her three little daughters during war times in Zadar while simultaneously building a flourishing career in both linguistics and philosophy, and, on top of that, being a sports enthusiast, running, and swimming in marathons at the age of 80. But, to me, she was more than that. First, she was my professor who taught me Old English and Chaucer. Later, she became a colleague—reintroduced to me by my mentor, Nenad Miščević—and the three of us used to spend time together at conferences. Finally, she was my friend. I will always remember her as being very encouraging, helpful and insightful. I am sorry that our time was cut short just when we started to spend more time together. But, I am grateful for the time we did have and that I had the chance to know her and to learn from her, academically and in life.

She truly will remain to be an inspiration.

To continue with Dunja's legacy, I will present the papers in this issue.

The first part of the issue is devoted to the discussions of Una Stojnić and Ernie Lepore's book Inflammatory Language *(OUP 2025), held at the 2024 Philosophy of Language and Linguistics conference. We didn't know it at the time, but it would be the last one Dunja attended. Stojnić and Lepore give a valuable precis of their book where they investigate slurs – words that derogate individuals solely on the basis of their group*

membership (race, gender, religion, etc.). The authors challenge common accounts of slurs and offer their own: they argue the offensive sting slurs carry arises primarily from associations triggered by the word's articulatory form (its sound or spelling). These open-ended associations carry socially entrenched histories of bigotry and exclusion, which are reactivated whenever the slur is uttered, even in pure resemblance cases. Their account, they claim, explains a whole arena of slurs' puzzles.

Robin Jeshion challenges Stojić and Lepore's account by claiming that Inflammatory Language *overlooks Multiple Mechanism theories, which are more than capable to explain hyperprojectivity. The other half of her paper addresses the Specificity Problem and claims that her own previously developed theory, Identity Expressivism, is equipped to tackle this issue. Christopher Hom's paper defends a content-based view against common criticisms of slur theories, showing it can adequately respond to each challenge. It examines Stojnić and Lepore's Articulation Account, arguing that the theory is both under-specified and overly ambitious, bringing forth certain dilemmas and challenges for their account. Louise Antony defends a lexical account of slurs, showing it better explains why slurs offend, their acquisition, mishearing, evolution, and reclamation, compared with the Articulation Account. Matthew Stone critically examines Stojnić and Lepore's claim that articulations are central to slur analysis, acknowledging their role in linguistic intuitions and social effects, but argues that slurs are best understood through a combination of prohibitions, word-level associations, and conceptual connections.*

The second part of this issue consists of papers presented at the 2023 Philosophy of Language and Linguistics conference. Adam Michael Sennet and Tyrus Fisher challenge von Fintel's semantics of subjunctive conditionals, showing that a Lewisian approach with pragmatic considerations better accounts for reverse Sobel sequences and NPI licensing. Indrek Reiland's paper was presented as part of the Special session on Elmar Unnsteinsson's book, Talking about: An Intentionalist Theory of Reference *(OUP 2022). In the paper, Reiland critically examines Unnsteinsson's Collapse Argument, showing that Easy views of saying or expressing do not collapse into Gricean views, because the intentions required for rationalizing an act are distinct from those that constitute saying or expressing. Alex Radulescu examines Unnsteinsson's claim that both combinatory and separatory confusion impair reference, arguing that separatory confusion—thinking one person is two—does not necessarily corrupt our ability to refer. Mirela Fus-Holmedal investigates how linguistic plausible deniability facilitates politically manipulative speech through dogwhistles, racial figleaves, and generic stereotypes. The paper shows that plausible deniability both shields such speech from criticism and allows it to spread more efficiently, increasing its impact. It further highlights the ethical and political significance of language, arguing that we should both combat pernicious manipulation*

and consider ways to harness plausible deniability for positive purposes, thereby bringing normative concerns into the philosophy of language.

Dunja, you will be missed!

JULIJA PERHAT
University of Rijeka, Rijeka, Croatia

Dunja Jutronić (1943–2025)

In memoriam
Dunja Jutronić (1943–2025)[*]

Thank you for giving me the opportunity to reflect on the life and work of Dunja Jutronić.

I know what Dunja would say – that I had written far more than necessary.

I first met Dunja in the second half of the 1980s. On several occasions she gave talks at informal gatherings of philosophers in Rijeka, the so-called sjedeljke. *Since she lived in Zadar, she was not a frequent participant—perhaps only when passing through on her way to Zagreb. At that time, she was working on nativism, and I remember her lecture at Nenad Smokrović's apartment on the city market. The question she addressed is one of the perennial problems in the philosophy of language: Is language innate or learned? Nativism holds that language is innate, while empiricism holds that it is acquired. The nativist thesis was revived by Chomsky with his claim that transformational grammar is innate. Despite Chomsky's enormous authority, Dunja remained skeptical of the innateness of language. She defended the more plausible view that we are born with a general disposition to acquire language, but not with language itself. Her style was always clear and straightforward. She wrote and spoke with focus, avoiding unnecessary terminology and elaborate formulations. She claimed only what she could substantiate, she was cautious in making assertions. Her work on nativism from the 1980s culminated in the book* Linguistics and Philosophy, *published in the Filozofska istraživanja series in 1991. A decade later, when new arguments in favor of nativism had crystallized, she critically examined them in her article "Arguments Against Nativism," published in* Metodički ogledi *(2003). That essay is available online, and may well serve as a model of how a scholarly article should be written.*

In Zadar, Dunja was active in an informal philosophy circle that met regularly, composed mainly of members of the Department of Philosophy. Among its participants were Nenad Miščević, Vanda Božičević, Arne Markusović, Darko Polšek, Dragana Sekulić, Boran Berić, and Slavko Brkić. Later, Elvio Baccarini and I joined, once we were employed in Zadar. It was through this group that her interest in philoso-

[*] *This is the speech that I gave on August 21, 2025 on the Krasica cemetery, near Rijeka.*

phy deepened—particularly in the philosophy of language—though her formal background was in linguistics. Meetings were held weekly, often in her apartment on the Old Town Peninsula. A downstairs neighbor, his family name was Trupac, found these gatherings suspicious and reported them to the police, worried about "subversive intellectuals" (perhaps in the climate of 1971 Croatian Spring in which intellectual meetings were mistrusted by the authorities). While the neighbor annoyed everyone, Dunja herself did not condemn him. Though she was naturally combative, she bore no resentment. The man had survived a Nazi concentration camp during WW2, and that experience marked his behavior. Dunja had understanding for this and responded with empathy rather than judgment.

The war years in Zadar were extremely difficult, especially for a mother of three. The Faculty building, with its deep basements, served as a shelter. In the early days, there were up to one hundred detonations per minute—I counted them, though few believe me even now. It was impossible to step outside; there was no water, no food, ATMs did not function. Later we would risk venturing out for coffee at Čulina's café near the Faculty, despite the ever-present threat of shelling. Life under constant danger continued in Zadar for more than three years. Worse still was the purge of "undesirables" carried out by the Dean of the Faculty with the support of certain people from the Ministry of Science. Within a matter of days, dozens—perhaps over seventy people—lost their positions. In such an atmosphere, taking a position in Maribor and relocating to Rijeka was both logical and lifesaving, and the salary was better. Yet the move weighed heavily on her daughters, Jelena, Katina, and Gordana, who had already endured multiple displacements—from Zadar to Split, from Split to Zagreb, and then to Rijeka.

In Maribor, her knowledge and expertise were fully recognized. She was appointed full professor in the newly established department and twice served as chair of the Department for Anglistics and Americanistics. She supervised nearly eighty bachelor theses and taught a broad range of courses, including History of the English Language. She once recited Middle English texts for us on a drive back from Maribor to Rijeka; the language, reminiscent of modern Dutch, fascinated us, since she rarely spoke of her Anglicist work—we usually discussed philosophy or organizational matters. Dunja drove extensively: first a white VW Golf 1, then a Seat Cordoba, later a Seat Leon. For twenty years, every week, she traveled from Rijeka to Maribor, always accompanied by Nenad Miščević, who neither drove nor owned a car. They shared the costs of fuel and tolls. At Nenad's funeral, Dunja opened her eulogy with the words: "Life is a journey." The metaphor of travel was the leitmotif of her address.

Sport was an essential part of Dunja's life. She once set a national record in the athletics triathlon, I think it was set in Belgrade in 1963—a record destined never to be broken, since neither the discipline nor the

state in which it was achieved still exists. After moving to Rijeka, first to Srdoči and then to Krnjevo, she immediately joined a recreational running group. Every Sunday morning in Kastav we ran routes of seven (to Zvirić) or twelve kilometers (a full circle), and did so for more than a decade. Initially she struggled for breath—unsurprising after the war years in Zadar spent raising three children—but regular training soon restored her fitness. She competed in Winter League races in Kastav and Kostrena. She always wore white leather Reebok Classic sneakers. After she developed problems with her foot, she turned to swimming, her other great passion, training regularly at 6 a.m. at the Kantrida pool. In both running and swimming she won numerous medals, particularly in her age group.

The Inter-University Centre in Dubrovnik is an exceptional academic institution in Croatia, enabling direct contact with leading international scholars and keeping us abreast of the latest developments. Dunja was deeply engaged there, serving as course director for two important annual conferences. The September conference Philosophy of Language and Linguistics regularly hosted Michael Devitt, Barry Smith, and Michael Glanzberg; it will now continue under the direction of Mirela Fus-Holmedal and Julija Perhat. She also led the April Philosophy of Science, whose participants included James Robert Brown, David Davies, James McAllister, Joseph Berkovitz, and Zvonimir Šikić. We expect younger people from our Department to continue this work: Zdenka Brzović and Vito Balorda. This event, the very first and longest-running IUC conference, marked its fiftieth anniversary this year. It was the only conference that continued even during the war, and its significance for us cannot be overstated. In those pre-internet years, direct international contact was more vital than nowadays.

In her linguistic research, Dunja studied Croatian emigrants in the United States, especially those from Dalmatia. This was the subject of her doctoral work at Penn State. Emigrant communities are of particular interest to linguists because they preserve the language as it was spoken at the time of departure—sometimes a century earlier. Their speech remains "frozen," offering a living record of Dalmatian dialects as they were spoken three generations ago. Analogous cases include the Rhaeto-Romance or Vlach dialect once spoken in some twenty villages on Učka mountain; today the largest Rhaeto-Romance community is found not in Istra but in New York, as a consequence of emigration.

Dunja worked a lot on naturalism in the philosophy of language, a contemporary and compelling approach that treats language as a natural phenomenon, to be studied as one would any other. She worked closely with Michael Devitt, her longtime colleague and friend, one of the leading figures in the field. Together they explored the causal theory of reference and meaning, grappling in particular with the so-called qua *problem: to explain the meaning it is not enough simply to point to an object; one must also specify the relevant aspect—its color, mate-*

rial, shape, edibility, and so on. This challenge appeared to force causal theorists to concede something to their rivals, the descriptivists. Their discussions resulted in the substantial edited volume The Maribor Papers in Naturalized Semantics *(University of Maribor, 1997).*

As a native of Split, Dunja felt a duty to preserve the city's dialect. She accomplished this in her Rječnik splitskog govora – A Dictionary of Split Dialect *(Durieux, 2006), published both in Split–Croatian and Split–Croatian–English versions. There one finds, for example, that katriga means "chair" and ponistra means "window." As language changes and fades, younger generations risk forgetting it altogether. Such dictionaries are therefore invaluable, and for this contribution we are indebted to Dunja.*

Above all, what remains with us is Dunja's spirit and character. She was, in the best sense of the phrase, a "get-things-done" person. We were all happy when she was organizing conferences, knowing everything would run smoothly: reimbursements processed, travel expenses paid, and always on time. A few years ago, we stayed a couple of days in her house in Sutivan, she gave us her house key and instructions for watering the garden: not only the plants in the middle, but also those at the back, behind the house. "Here's the hose—it's long enough!" she said. She wanted things done properly and seen through to completion.

Dunja enriched our lives with her presence. Her spirit and character will remain with all of us who were fortunate enough to know her and to share time with her. Dunja, we love you!

BORAN BERČIĆ
University of Rijeka, Rijeka, Croatia

- *Professors at the University of Novi Sad: Life Stories*, recorded by Svenka Savić, January 2011, University of Novi Sad jubilee volume (1960–2015), pp. 191–205.
- *Od jezika k filozofiji in nazaj: Festschrift on the 75th Birthday of Dunja Jutronić*, eds. Bojan Borstner, Tomaž Onič, University of Maribor, 2019.

Eulogy for Dunja Jutronić (1943–2025)

Arriving in Dubrovnik for the 2025 Philosophy of Language and Linguistics conference, it was hard to believe that Dunja wouldn't be amongst us welcoming everyone on Sunday night to Troubadour in the Old Town for pre-conference drinks. We had a sense that the conference would be marked by her absence.

For twenty years, Dunja was the lynchpin and key organiser of this series, bringing together a remarkable cast of characters each September to focus on language and linguistics. She started the existing series with Michael Devitt in 2005, although the conference was not always called, Philosophy of Language and Linguistics. Originally, it's title was Mental Phenomena. That was because it was the continuation of a series of meetings at the IUC with an important Dubrovnik pedigree. The Mental Phenomena Workshop was the brain-child of Mike Martin and Tim Crane, from UCL at the time, whose aim was to rebuild the philosophy connections between the UK and the former Yugoslavia following the end of the war in 1994. The workshops started in 1997 and were funded for this purpose by the British Academy for three years. These British Academy workshops (and grant award) celebrated the outstanding, and commendably brave, work of Kathy Wilks in championing analytic philosophy in pre- and post-communist Europe, and her support for Croatian philosophers throughout the war. Her courage and commitment to the IUC and the Croatian people during the siege of Dubrovnik are commemorated in a plaque at the Pile Gate.

The work of Kathy, together with her colleagues, David Charles, Justin Broakes, and the philosopher of science, Bill Newton Smith, established a strong connection between Oxford and philosophers from the former Yugoslavia gathering at Dubrovnik. Now it was London's turn and the aim was to provide a space where philosophers from these now separate countries could once again come together and pursue topics in the philosophy of mind. Matthew Nudds took over the running of Mental Phenomena, which is when I started regularly attending conferences in the IUC. By 2004, there was a desire to pass the baton, and this is where Dunja and Michael Devitt stepped in. They suggested a conference on philosophy of language and linguistics, and Michael was just about to produce his Ignorance of Language book. There were enough

philosophers in the generative grammar tradition who sought to relieve him of that ignorance to make for a lively set of meetings. And so, the first conference in our series was born in 2005. Many of us didn't mind the title, Mental Phenomena, since as Chomskians we took language to be a mental phenomenon anyway. However, with the IUC's permission the title was amended in the following years to The Philosophy of Language and Linguistics and it has been supported ever since by the University of London's Institute of Philosophy through its Centre for Logic and Language (CeLL), with Dunja and Michael Devitt steering its course.

To begin with the meetings were titanic struggles between the leading supporters of experimental generative grammar and the common sense, thought experimental, philosophers of language. In session after session, Dunja held the ring between Michael Devitt, Geoff Pullum and Paul Pietroski, Stephen Neale, John Collins. Frankie Egan, Una Stojnic and Jeff King took us to loftier places, and at times we indulged in the semantics-pragmatics Holy Wars. We have always been extremely grateful to the Croation Journal of Philosophy for publishing so many papers documenting the elaborate skirmishes from these meetings.

Dunja was constantly aware that we mustn't just turn to the big guys as speakers at these conferences. We needed to encourage younger researchers and especially philosophers from the region. She was tireless in consulting and sending out messages to younger people and encouraging them to come. The late Nenad Miščević, who even his best friends wouldn't call organised, did his bit in encouraging a string of young philosophers to attend.

However, it was Dunja who was the mainstay of these events. Very gently and persuasively, she led from the back, asking people's opinion, trying out suggestions, and tirelessly sending email reminders to the errant co-organisers who were never chided for their tardy replies in order to keep things moving and make sure the conferences happened. She encouraged Frankie Egan, Michael Glanzberg, Una Stojnić and me to be part of the organising committee, and more recently, Mirela Fus-Holmedal. Nenad was mostly an honorary part of the organising committee although Dunja knew it was best to include him and defer to him while quietly getting on with things behind the scenes. She knew his larger-than-life presence could be tricky but she was fond of him and gave him his due. Characteristically modest and other-regarding, she never claimed any of the success for herself, but I think she was rightly proud of what she had achieved at the IUC: The Philosophy of Language and Linguistics conferences.

We won't forget her knowing irony and gentle admonishments. She had a quiet wisdom and the measure of all of us. The occasional and kindly given roll of her eyes, her constant concern for fairness and equity, including people, especially the young researchers, and her equally attentive obsession with the air conditioning, which she considered to

need constant fine-tuning through those long mornings and afternoons in IUC classrooms. Most of all, we will remember Dunja taking soundings of where we should all go to swim at lunchtimes: Dančа or the Cove? And there she would set off at an even pace and swim steadily till she was far out of sight. Once we had all swum, talked and sorted out whatever philosophical topic we were chewing over and decided where we would go for lunch we would scan the horizon and spot the steady stroke that we knew was her. I like to think of her as out there still. Thank you for everything, Dunja. You were always much appreciated and will always be much missed.

BARRY C. SMITH
University of London, London, UK

Swimming into Memory and Beyond: Farewell to Dunja[*]

This year, just under a month ago, the Philosophy of Language and Linguistics community lost one of its own, one of its founding pillars. We lost our Dunja. Let me share a few anecdotes about Dunja that I personally witnessed.

How I Met Dunja—Twice

I first met Dunja when I was just a bachelor student. Back then, she was simply "a professor" to my mind, a linguist whom my mentor at the time, Professor Nenad Miščević, brought to our reading group. It didn't take long before I realized that I had stepped into Dunja and Nenad's world, their personal and academic stories from before, during, and after the war in the former Yugoslavia.

I listened to accounts of their commutes together to work in different Croatian cities: Zagreb, Zadar, Rijeka, Split, as well as abroad, to Maribor and Central European University (now in Vienna, but at that time—in their time—located in Budapest).

Only recently did I truly grasp the impact their stories had on my life. Their way of living inspired mine: from applying, to studying, to working abroad, and even seeing commuting to another city or another country for work as normal. Though we have shared these ways of living, my professional life has turned out perhaps a bit less dramatic, less innovative, and, most importantly, not as deeply shaped by war and politics—at least so far.

Dunja came into my life twice, or rather, as I've recently realized, I came into hers. Just over three years ago, Dunja reached out to me again, completely out of the blue, inviting me to become one of the co-directors of this conference. A special, invisible bond we had built long ago suddenly revealed itself. When I saw Dunja again in Dubrovnik in 2023, it had been about eight years since our last meeting. The first thing that struck me was how nothing had changed. The spirit of the conference hadn't changed. Dunja hadn't changed. The same presence, the same grounding force she always carried with her!

[*] Read at *In Memoriam* for Dunja Jutronić, Inter-University Centre Dubrovnik, Croatia, September 8, 2025 (Revised in Oslo and Lørenskog, Norway, October–November 2025).

Dunja Being Nenad's—and Everyone Else's—Sidekick

It's hard for me to talk about Dunja without mentioning Nenad Miščević. Many of you may have known him, if not directly, then through Dunja. She was his grounding force, and I witnessed that both personally and through Nenad. When Nenad was overly idealistic or on the verge of making what he himself sensed might be the wrong choice, she was often the one to bring him back to solid ground, as he once admitted to me.

Dunja was truly supportive, loyal, and empowering, both academically and personally. She had a gift for finding the right words. I always felt she treated us with respect, regardless of our academic or social position. I remember presenting at this very conference for the first time as a student. My talk was much shorter than the usual sixty minutes most people take—I nervously sprinted through my notes. Afterward, she said something like, "short and sweet," but in a way that made me feel I had succeeded. She had a Socratic, midwife-like approach—trusting, guiding, and teaching by showing.

Witnessing Dunja's Two Legacies: Philosophy of Language and Linguistics Conference, and Croatian Journal of Philosophy

Dunja leaves behind many legacies. I want to highlight two that are most closely tied to my relationship with her and to the work that connected us.

She was deeply loyal to this conference, which she helped establish in 2005, twenty-one years ago. I still regret not emailing her the final version of this year's program before she passed, although I did send it afterward. My impression was that it mattered deeply to her that the conference would continue.

Dunja was also a long-time editor of the special issue of the Croatian Journal of Philosophy*, dedicated to topics in philosophy of language and linguistics, often featuring work presented at this conference. After I took over her role just a few months before she learned she was ill, I recall her saying she felt deeply relieved to have both things sorted out. I realized then that it was never about her—it was about the survival of something she cared about profoundly.*

Witnessing Dunja's Sporting Spirit—Never Giving Up

Dunja's determination was not limited to academics; it extended to sports. She was once a junior triathlon champion of Yugoslavia. It's been said that her record has never been surpassed—and likely never will be—since Yugoslavia no longer exists, nor does that particular version of the triathlon. Yet this secure win did not stop her. In the 21st century, she actively took part in swimming marathons, often as the oldest competitor. The most recent public record I found shows she participated in one in 2022. And many of you will remember her swims, or

perhaps even swam with her, during lunch breaks at the Philosophy of Language and Linguistics Conference over the years.

Speaking of never giving up, I suspected—but didn't want to believe—that she was very ill. Nor did she, at least not in front of me. She remained positive. Even in her last email to me, she was complaining that Michael Devitt simply did not understand that she would need time to recover—and that him coming to Rijeka to see her this September, instead of attending this very conference in Dubrovnik, was too early in her recovery process!

I want to share part of Dunja's email that, I believe, captures her spirit:

'Danas mi je rodjendan, 81. koji uzas!!! (...) Uzas ne znaci nista vise nego te silne godine…' */ 'Today is my birthday, the 81st. What a horror!!! (…) Horror means nothing more than all those countless years…' (Dunja Jutronić, personal correspondence, November 17, 2024).*

For those who did not know her, this may sound unusual; for those of us who did, it is a vivid picture of how Dunja could be at once blunt, realistic, and full of determination to live and carry on.

Celebrating Dunja

This year's conference program honors Dunja in three ways: with an In memoriam for Dunja Jutronić*, with the* Dunja Jutronić Memorial Lecture *by Barry C. Smith, titled* "LLMs: Tools or Colleagues?"*, and with the* Dunja Jutronić Memorial Swim.

Let me leave you—and bid farewell to Dunja—with a verse from a poem titled Kada iz velike posude *(*When from a Large Vessel*) by her favorite Croatian poet, Danijel Dragojević:*

Za tu vodu imam razumijevanja /
For that water I have understanding /
kao za sebe u najboljim časovima /
as for myself in my best hours /
kao za riječ koja je izgubila rečenicu /
as for a word that lost its sentence /
na putu do nas. /
on the way to us. /

MIRELA FUS-HOLMEDAL
Norwegian University of Science and Technology, Trondheim, Norway

Croatian Journal of Philosophy
Vol. XXV, No. 75, 2025
https://doi.org/10.52685/cjp.25.75.1
Received: November 1, 2025
Accepted: November 14, 2025

The Evolving Understanding of Slurs: An Inquiry into Meaning and Effect of Slurs

UNA STOJNIĆ
Princeton University, Princeton, USA

ERNIE LEPORE
Rutgers University, New Brunswick, USA

In this précis, we summarize the key themes and arguments of the Inflammatory Language: Its Linguistics and Philosophy.

Keywords: Slurs; pejoratives; content; taboo; pejorative tone; articulations.

Slurs are powerful linguistic weapons. These expressions denigrate those they target purely on the basis of group membership (e.g., on the basis of race, ethnicity, origin, religion, gender, sexual orientation, or ideology). They are highly inflammatory; slurring someone involves a transgression more serious than a mere insult, so much so that even mere tokenings of slurs often have a full-on taboo status, and are subject to media censorship, and even legislation. They have a "viscerally palpable effect," a characteristic pejorative sting that makes them prone to offend and harm but also imbue them with rhetorical potency that mere neutral descriptors for the group they target lack. This pejorative effect underwrites the intuition many theorists voice that tokenings of slurs as opposed to the corresponding neutral descriptors might be necessary in certain context in order to achieve a particular goal—political, pedagogical, artistic, and so on.[1] But characteristically, even tokenings that aren't intended to offend and aim instead to create a rhetorical effect that furthers a political, pedagogical, or artistic goal remain inherently risky. We thus commonly see discussions of whether particular tokenings of slurs in literature, poetry, film, theatre, or in

[1] See, for instance, Camp 2013.

political activist discourse, or within a pedagogical setting, were warranted or apt, or gratuitous or flippant. Indeed, even mere quotations and displays of slurs retain their pejorative sting. Scholarly publications on slurs frequently open with a disclosure that slurs will be mentioned (not used) in the text, followed by an apology for the offense this is prone to cause. Indeed, mere *displays* of slurs are often subject to outright media censorship and even legislation. In *Inflammatory Language*, our question is what makes slurs so inflammatory? What is the source and nature of their characteristic pejorative sting?

In searching the philosophical literature on pejorative language, and in particular, on slurs, prior to this century, there is little to be found. The few extant discussions are mostly about the inferentialism. For example, there are published discussions by Robert Brandom and by Michael Dummett that exploit distributional intuitions most speakers share about slurs (Dummett 1981; Brandom 2002). The reason why philosophers, by and large, avoided theorizing about pejorative language is because everyone—philosopher or not—held it, implicitly or explicitly, that slurs derogate and are prone to offend because of what they mean.

In contrast to the relative neglect of the past centuries, the last few decades have seen an explosion in the philosophical and linguistic discussion of slurs, with a wide range of proposals on offer as to what makes them so inflammatory and gives them the pejorative sting. But this natural pre-theoretical stance, and a *prima facie* plausible theoretical starting point—that the key to understanding the inflammatory nature of slurs lies in their meaning—is retained throughout much of that literature. The rough idea is that slurs have some offensive, pejorative meaning, and so that tokenings of slurs tend to be offensive because they express or otherwise evoke that pejorative meaning. Much of the discussion and disagreement in the literature then lies in characterizing this meaning—whether as descriptive, expressive, or some other kind *sui generis* content—and the ways in which it is conveyed—semantically, as a matter of truth-conditions, or presupposition, or conventional implicature, or pragmatically, as a matter of conversational implicature.

In *Inflammatory Language,* we argue that this unanimity about the meaning approach to the pejorative sting of slurs is radically mistaken. In the book, we critically examine various proposals in the literature that posit pejorative content. Such proposals are numerous—indeed, we are not exaggerating much when we say that for any of the familiar ways of conveying or expressing content—be it as *at issue* truth-conditional content, or presupposed or conventionally implicated content, or perhaps pragmatically implicated content—there is a proposal out there in the literature that argues that that is the way in which the pejorative content of slurs is conveyed. Similarly, there is a whole array of proposals as to what the *nature* of this pejorative content is: some argue it is descriptive, representational content, others that it

is purely affective or expressive, and yet others that it is *sui generis* type of perspectival content that combines both affective and descriptive components.

But regardless of the near consensus that the pejorative sting of slurs is a matter of meaning, and the many content proposals on offer as to the nature of that meaning, we argue that no content account can ultimately be successful. We raise various challenges for different accounts specifically, but we also argue there is a host of challenges that in principle tell against *any* content account one might offer: in short, there is no special pejorative meaning either semantically encoded in, or pragmatically conveyed by, slur terms. Here we briefly summarize just a few of the main challenges we raise in the book.

A peculiar thing that one immediately encounters when inquiring into the meaning of slurs is a lack of non-schematic proposals as to what particular slurs actually mean. This is already reflected in the practice most contemporary dictionaries adopt when listing the meaning of slurs. The O.E.D., for example, offers pretty much the same definition for each slur term: it states that it denotes individuals that belong to a certain group, and that the term is offensive or derogatory (with occasional remarks about the term's etymology, frequency or common usage). This schematic approach is present in most contemporary accounts of slurs that posit special "pejorative" content—either pragmatically implied or semantically encoded. So, to mention just a few concrete proposals, we get accounts that tell us, e.g., that a slur S targeting a group G predicates of an individual x, that x ought to be a target of negative moral evaluation because of being a member of G (Hom 2008; Hom and May 2013; see also Neufeld (2019) for a related proposal); or that it presupposes that (the speaker believes that) the members of G are despicable on account of being members of G (Schlenker 2008; Stojanovic and Cepollaro 2015; Cepollaro 2015); or that it conventionally implicates that the speaker has (occurrent or standing) negative affect towards the members of G (on account of being members of G) or having negative traits stereotypically associated with G on account of being members of G (McCready 2013; Williamson 2009; Potts 2007). In each case we are given a meaning schema, the instances of which lay out the meanings of particular slurs.

The schematic accounts of this sort—much like the dictionary practice—essentially posit the same meaning for all slurs terms, modulo the group membership. This, however, is intuitively inadequate. There is a great deal of variation in the severity and character of the offensive effect of different slurs—including different slurs that target the same group of people. (Consider, for instance, distinct slurs targeting women, or gay men, or African Americans.) Yet the schematic accounts above would tell us that any two slurs targeting, say, women, have the same meaning: e.g., that they predicate that the subject ought to be the target of negative evaluation because of being a woman, or that they presuppose that (the speaker believes that) women are despicable

on account of being women, or that they conventionally implicate a negative affect towards women, and so forth. This seems woefully inadequate.

This lack of specificity, we argue, is not a coincidence. One might be tempted to overcome it by positing different specific contents along the lines of those specified by the schema for different slurs. So, one might be tempted, say, to argue that different slurs targeting the same group encode or express different negative attitudes or different affects, and that those different attitudes or affects are of different severity and strength. But this reaction is on the wrong track. For one, it's hard to see how any such semantically encoded content could be what underlies the intuitions about the severity and differences in the pejorative sting of different slurs targeting one group; for while it's undeniable that there are such variations in the strength and severity, competent speakers routinely disagree over which slurs for a particular group are "worse" or more severe, and why.[2] But these disagreements do not display linguistic ignorance or incompetence with *meaning*. And there can be changes in the nature and severity of a slur's offensive sting over time—as well as differences across dialects; but such changes don't suggest a change in *meaning*.

Moreover, for any bit of specific semantic content one might posit, one can consistently deny that content while predicating a slur. This is already evident even with the bare-bones schematic proposals above; viz.:[3]

1.
   a. Hermione is a mudblood, but she ought not be subject to negative moral evaluation on account of being muggle-born.
   b. It's just false that mudbloods are inferior/despicable on account of being muggleborn. I have nothing but respect for them!

[2] Consider, for instance, this discussion over the comparative severity of distinct slurs for woman engaging in casual sex: https://www.thestudentroom.co.uk/showthread.php?t=2171660. One can easily find an abundance of further examples of debates of this kind on the web.

[3] 'Mudblood' is an extremely offensive slur in the fictional world of Harry Potter, targeting wizards born to non-wizard parents. 'Muggle-born' is its neutral counterpart. We stick to presenting examples involving fictionalized slurs, rather than quoting actual slurs; this is not because we advocate silentism or blanket censorship of such tokenings. We emphatically do not. But as we'll see, even mere quotations and displays of slurs (or more precisely, their articulations), carry the pejorative sting. That doesn't mean all such tokenings constitute transgressions: we argue the sting can be not only weaponized to harm, but also used to achieve various potentially worthy rhetorical, artistic, pedagogical, political or other goals. However, we prefer not to judge whether any such tokening on our part would be apt or warranted. We trust our readers can test their intuitions against non-fictionalized counterparts of our examples featuring slurs they are familiar with. For a more thorough discussion of these issues, see Ch. 1 of *The Inflammatory Language*.

c. Hermione is a mudblood, and I *love* mudbloods! They are my favorite wizards!
d. Mudbloods have no negative features due to being muggle-born.

By contrast, denying the predicated, presuppositional, or conventionally implicated content, be it descriptive or expressive, *does* lead to linguistic incoherence:

2.
a. # The period between October 18th and November 1st is two weeks, but not a fortnight.
b. # John stopped smoking, but he never used to smoke.
c. # John is tall but fast, and there's no contrast between height and speed.
d. # Joe, who is a spy, is in hiding; and he's not a spy.
e. # [Angrily:] That goddamned dog is barking again! [Cheerily:] I'm pleased with the dog.

Making the schematic content more specific does nothing to alleviate this problem: take whatever description, attitude, or affect you think a slur encodes—we submit you can consistently predicate the slur while denying that description, or affect, or attitude.

One might at this point be tempted by an account according to which the pejorative sting of a slur is a matter of conversationally implicated or otherwise pragmatically implied content, for such content is cancellable, and thus deniable.[4] But this temptation should likewise be resisted. First, even a proponent of a pragmatic account has to give a plausible account of what the specific content implicated or otherwise pragmatically conveyed by tokenings of slurs is. Such account must respect the variation in the strength and severity of the offensive potential of different slurs (including different slurs for the same group), and at the same time posit content that is plausibly grasped and tracked by ordinary speakers. We are pessimistic about the prospects of meeting this challenge. But more importantly, while predications of slurs are consistent with denials of any particular bit of descriptive or affective content, as the data above show, the pejorative sting of slurs is precisely *not* cancellable. Indeed, that is already plain to see in examples (2a)–(2d). While the examples illustrate that one can consistently predicate a slur while denying the supposed pejorative content, at the same time, the tokenings of slurs in (2a)–(2d) still retain their offensive sting: (2a)–(2d) are still prone to cause offense, and have radically different effect than their counterparts with the slur replaced with the corresponding neutral descriptor. (We invite our readers to test their judgements with non-fictional slurs they are familiar with.) In short: any bit of (descriptive or affective) content is cancellable, but the pejorative sting is not—hence, the pejorative sting is not a matter of content.

[4] See, e.g., Nunberg 2018; Jorgensen Bolinger 2017.

Indeed, one of the key challenges for content accounts is what we call the "hyperprojectivity" of the pejorative effect: not only does it remain even under negation (3a), but it persists in—i.e., projects out of—a wide-variety of embeddings, including presupposition filters and plugs ((3b)–(3c)):

3.
   a. Hermione is not a mudblood.
   b. If muggleborns are inferior on account of being muggleborns, then Hermione is a mudblood.
   c. Draco said that Hermione is a mudblood, but I don't think muggleborn wizards are despicable on account of being muggleborn.

By contrast, negated predications don't predicate anything of the subject (or anyone else), while presupposed content doesn't project out of filters when the presupposition is locally satisfied, nor does it project under plugs. For example, (4a) doesn't predicate being Italian of Mary, while (4b) and (4c) don't presuppose that John used to smoke:

4.
   a. Mary is not Italian.
   b. If John used to smoke, then he stopped smoking.
   c. Mary said that John stopped smoking, but he never smoked.

Indeed, the pejorative sting notoriously projects even out of environments that normally render any type of meaning inert—the purely quotative uses, and indeed even mere displays of slurs. In purely quotative environments, an expression is merely mentioned, not used to express its meaning; so, for instance, semantic content is rendered inert in meaning attributions (both on the right-, and left-hand side) (Anderson and Lepore 2013):[5]

5.
   a. "'David is Jewish' means David is Jewish" predicates nothing of David.
   b. "'John is tall but handsome'" means John is tall but handsome draws no contrast between height and handsomeness.
   c. "'John, who is my friend, is happy'" means that John, who is my friend, is happy" doesn't convey anyone has friends.
   d. "'Joe stopped smoking' means Joe stopped smoking" doesn't presuppose Joe once smoked.
   e. "'ouch' means ouch" does not express a state of mind (e.g., pain).
   f. "'vous' means vous" does not commit the speaker to addressing anyone in a formal/ polite manner nor does it signal politeness.

[5] Examples (5a)–(5d) are from Anderson and Lepore (2013); (5e)–(5f) are our addition.

Moreover, mere quotation generally renders meaning inert:

6.
   a. 'Italian' has 7 letters.
   b. 'Italian' is an English word.
   c. 'Italian'

By contrast, the pejorative sting of a slur persists even in such environments (as suggested by well-documented real-world incidents, empirical studies, as well as editorial practices):[6]

7.
   a. 'mudblood' means mudblood.
   b. 'mudblood' has 8 letters.
   c. 'mudblood' is a slur term.
   d. 'mudblood'

On the whole, then, we are pessimistic about the prospects of pejorative content accounts. The data above are damning for any such account.

While content-based accounts are by far the most prevalent in the literature, there is an alternative minority position: explain the pejorative sting of a slur not as a matter of its meaning, but through some other feature of the word. So, for instance, Prohibitionists argue that slurs are prohibited, taboo words, and it is the violations of the prohibition that their tokenings constitute that generate their pejorative sting (Anderson and Lepore 2013). The Pejorative Tone Account, in turn, maintains that slurs carry "pejorative tone," understood in a broadly Fregean sense: they give rise to a wide-range of open-ended, pernicious associations due to a variety of socio-historical, cultural, psychological, and other factors, and it is because of these associations that slurs have pejorative sting and taboo status (Lepore and Stone 2018; c.f. Frege 1897/1979).

While these accounts might fare better insofar as they don't attempt to explain the pejorative sting in terms of meaning and thus can offer an explanation as to why even mere quotes and displays of slurs retain the pejorative sting, they still fall short. In the book, we raise a range of problems and challenges for both Prohibitionism and the Pejorative Tone Account. In the interest of space, here we shall only point to the most pressing issues that no account that ties the explanation of the pejorative sting of slurs either to their meaning or the words themselves can adequately address.

To see the main issue, we begin with the phenomenon we've dubbed "inheritance": the pejorative sting of slurs is not only hyperprojective, but it is infectious in that it carries over even to expressions accidentally matching slurs in articulation: in their orthographic and phonological form. The phenomenon can be illustrated by the debate and contro-

[6] See, e.g., Fasoli, et al. 2015; Carnaghi & Mass 2008; Cepollaro, Sulpizio, Bianchi 2019. For a more in-depth discussion of the relevant empirical studies, real-world cases, and editorial practices, see Ch. 6 of *The Inflammatory Language.*

versy over the offensiveness of an English adverb which bears phonetic (and orthographic) similarity to the N-word, but is otherwise etymologically, and semantically unrelated (Kennedy 2002: 94–95). Kennedy reports fiercely divisive opinions surrounding the infamous incident, where a white speaker, addressing a largely African-American audience chose this (rarely used) adverb instead of any of the many (much more frequently used) synonyms (e.g., 'ungenerous', 'stingy'), sparking outrage. Even if one thought that the original outrage was somehow a result of an etymological or semantic confusion, the controversy continues alongside a widespread recognition that the similarity is a matter of mere orthographic accident (to borrow Quine's apt phrase (1953: 67)), and the adverb remains tainted; as O'Hehir puts it, the adverb now "carries a permanent taint: The only person who would conceivably use it now would be a snickering, anti-p.c. asshole trying to make an obnoxious point" (O'Hehir 2020).

Indeed, the phenomenon of inheritance arises even in cases where there is *no* temptation to post an etymological confusion. Just one illustration is a relatively recent case where outrage and offense were sparked in a class setting by an instructor mentioning (not using!) a Mandarin demonstrative term ('那个'), which acoustically resembles the N-word. In this case, even at the time of the incident, there wasn't any confusion about which word was being tokened. The tokening occurred in a context of discussing filler-words, as a part of an explanation that, in Mandarin, this word—which the speaker explained, prior to tokening it, to be a correlate of the English demonstrative 'that'—is often used as a filler-word. The speaker preceded the tokening with an explicit explanation of which word of which language will be tokened.

Note that we are not interested in the question of whether any of these tokenings constituted a wrongdoing, or whether their speakers were blameworthy; nor are we adjudicating whether any offense they caused was sincere, appropriate or warranted. But it would hardly be necessary to debate these issues, issue apologies, or write an op-ed, let alone numerous ones, on whether these reactions were warranted or appropriate, if it were *obvious* that there was no pejorative sting triggered in these instances.

And it's worth emphasizing that the phenomenon isn't isolated or exclusive to a single slur. Indeed, it is quite common. To illustrate with just one more example, consider, for instance, the incident involving a Krispy Kreme doughnuts advertisement that aired in Australia, where the word 'congratulations' was rendered with two doughnuts replacing "o" so that a part of the word accidentally matched a highly charged racial slur.[7] This sparked outrage, the add was removed, and the company issued an apology. Note, here, the orthographic accident is quite

[7] https://www.mirror.co.uk/news/world-news/krispy-kreme-apologises-after-racist-30761080

apparent: no slur was tokened, and the orthographic string (stylized by way of doughnuts) that accidentally matched the articulation of a slur was a proper part of the articulation of the word actually tokened.

Indeed, the phenomenon is a general one that we commonly see with offensive gestures, symbols, and imagery. Consider, for instance, the Nazi Hakenkreuz: its violent and offensive symbolism not only gives rise to taboo and prohibitions— often legally codified— against its displays but its pejorative sting easily transfers to certain closely resembling symbols that carry positive connotations in various Eurasian religions, including Hinduism, Jainism, and Buddhism. This is apparent both in debates over the possibility of "reclaiming" these symbols and "restoring" the original positive connotations, but also in the common attitudes toward them in the West.[8]

We illustrate with an incident involving toy pandas sporting the manji symbol— a good luck symbol that is a mirror image of the Swastika—that ended up in Christmas crackers in Canada, causing widespread outrage, and a subsequent apology and recall from the manufacturer.[9] This is exactly the same phenomenon, we believe, we find with the inheritance of the pejorative sting of slurs in articulatorily resembling expressions.

The phenomenon of inheritance is difficult to explain on the content-based accounts and the word-based accounts alike: for in tokening the now infamous adverb, or the Mandarin demonstrative, or 'congratulations,' no slurs were tokened: and so no slur was even mentioned, let alone used, and thus its (putative) pejorative meaning expressed, or a slur-related taboo violated, or tone triggered.

We instead argue that the pejorative sting of slurs isn't a matter of their meaning, nor even slur *words* themselves. Rather, we argue that it is (certain of the) articulations of slurs words that harbor the pejorative sting. The sting, in turn, is constituted by the open-ended cluster of negative associations rooted in various socio-historical, cultural, and psychological factors. These associations are open-ended and not content-like: they vary across speakers, time and dialects, and for a given speaker across different times and circumstances. However, while variable, they share a common thread, insofar as they are rooted in the same set of socio-historical and cultural facts. Finally, and importantly, these associations do not attach to words, but rather to their articulations: (certain of the) slurs' orthographic and phonological forms.

Our account—the Articulation Account—is uniquely well positioned to capture the full range of data concerning the slurs' pejorative sting. First, it captures that a denial of any specific pejorative content alongside the predication of a slur doesn't lead to *linguistic* incoher-

[8] See, for instance, Heller 2000.

[9] https://www.cbc.ca/news/canada/toy-pandas-bearing-swastikas-a-cultural-mix-up-1.343550; https://www.theglobeandmail.com/news/national/swastikas-still-hurt-toy-error-shows/article4143067/.

ence: slurs do not semantically encode, or pragmatically convey, any such content, and so, there's no inconsistency or incoherence in predicating a slur, while denying such content. The account can also explain why there can be variation in the presence and severity of the offensive sting across time and dialects, as well as across different speakers, and even for a single speaker at different times, without positing a change in meaning, or linguistic confusion. The associations that tie to articulations change dynamically over time, and across communities, as a result of changes in various socio-historical and cultural factors; and they are open-ended, in that no particular bit of content or affect is semantically encoded, nor it must be evoked by any given agent in order for them to be linguistically competent.

The Articulation Account also explains why the pejorative sting is hyperprojective, indeed, why it persists even in purely quotative environments and mere displays of slurs: for it is the articulation of the slur, which is present in such environments, that carries the negative associations that constitute the pejorative sting.

Similarly, we can explain why the tokening of a slur is neither necessary nor sufficient for triggering the pejorative sting. On the one hand, the articulation itself can trigger the pejorative sting without the slur being tokened. This is already reflected in the phenomenon of inheritance: in tokening the Mandarin demonstrative, for instance, the speaker didn't thereby token a slur. It is the articulation of the slur that's tokened, due to the orthographic—or rather, phonological—accident, which in turn carries the pejorative sting. That no slur is tokened in this case should be evident as soon as one reflects on the fact that the written counterpart, '那个', triggers no pejorative sting whatsoever. To insist otherwise would be to confuse words and their articulations.[10]

On the other hand, absent a slurring articulation a slur loses its pejorative sting: if one attempts to token a slur, but misarticulates it so badly that it's difficult to even recognize, the tokening of the slur loses its pejorative sting. If one knows the speaker misarticulated a slur, one might then wonder whether the speaker had malicious intentions or not, whether they are blameworthy or not, and so forth, but one won't thereby experience the same viscerally palpable effect as if one had been confronted by the articulation of the slur directly. This is just as if someone tries to punch you, but misses, or barely grazes you. If you know what they tried to do, you might wonder whether they had any malicious intentions, whether they are blameworthy and so forth; however, you won't thereby experience the pain of the would-be punch.

One might be tempted by the thought that we've conflated metaphysics and epistemology: isn't it simply the case that absent (standard) slurring articulations the pejorative sting is absent because we

[10] For further discussion of the important distinction between words and articulations, see, e.g., Stojnić 2001; Hawthorne & Lepore 2011; Kaplan 1990; 2011.

fail to recognize that the word tokened is a slur? And conversely, isn't it simply because we confuse standard articulations of slurs for the tokenings of slurs, even when no slur is tokened, that such articulations retain the pejorative sting? If so, this is consistent with the idea that it's the slurs that carry the pejorative sting, while the articulations merely serve as evidence as to whether a slur was tokened or not.

This reaction should be resisted. Note that in none of the inheritance examples above there's any confusion about which word was tokened. For instance, the speaker in the Mandarin demonstrative incident announces ahead of time which word of which language will be tokened. So, these examples cannot be chalked up to mere confusion. Nor should we think that the absence of the offensive sting in the absence of (certain of) the standard articulations of slurs is merely a byproduct of epistemic confusion, a failure to recognize that a slur has been tokened.

To see this, it is useful to reflect on another important phenomenon that has been largely neglected in the literature—the variance in the presence and severity of the offensive effect across different articulations of one and the same slur term. A vivid example of this is the phenomenon of graphic pejoratives in logographic languages, like Mandarin. Graphic pejoratives allow that only certain written, but not spoken, articulations of a particular term can carry the offensive sting. For example, historically, many exonyms were considered as pejorative only when written with one of several possible, phonologically indistinguishable phono-semantic compounds (characters). Matisoff reports that names of people deemed lesser would be deemed offensive when they featured the 'beast radical' (犭) (e.g., '猺,' i.e., 'Yao'), a rendering typically avoided in formal correspondences (Matisoff, 1986: 6). Thus, we have a case where only one of several possible standard articulations of a single term, a particular spelling, carries the pejorative sting. If the sting were a matter of a word or its meaning, this would be baffling, for a word retains its identity, and meaning, regardless of the choice of whatever acceptable spelling. Nor can it be that articulations of slurs trigger offense merely derivatively, because we recognize them as articulations of slurs, and that when slurs are tokened absent slurring articulations, the sting is absent due to a failure to recognize that a slur is tokened. With graphic pejoratives, standard articulations of the same word vary in whether they trigger the offensive sting. Yet, the word is what it is, and means what it means, regardless of how you spell or pronounce it.

The phenomenon of variance in the offensive sting across different articulations of one and the same term is likewise not isolated, nor exclusive to logographic languages. In *Inflammatory Language* we discuss a wide range of such cases, including the use of minced oaths, variations in spelling or pronunciation designed to blunt the offensive sting, as well as similar uses of the censoring asterisk, as in:

8.
   a. Hermione is a mudbl**d.
   b. The word 'mudbl**d' is censored in this sentence.

Indeed, variation in the associations attached to different articulations of a single term can explain the effects of variation in articulation that sometimes accompanies reclamation—the reclaimed in-group uses of a slur as a term of pride, a form of solidarity, camaraderie, activism, or simply a neutral moniker of self-designation. Consider, for instance, the in-group uses of the N-word in its non-rhotic pronunciation (and the corresponding spelling variation). While out-group tokenings of both non-rhotic and rhotic articulations are unequivocally offensive, there is a debate both over the availability for in-group reclamation, and comparative offensiveness, of these different articulations of the slur. Whatever one makes of such debates, they wouldn't even *prima facie* make sense if it were the term itself that already triggered the pejorative sting. On the Articulation Account, by contrast, the phenomenon is entirely unremarkable: different articulations of one and the same term can vary in associations and thus in the presence and severity of the offensive sting.

In sum, we've argued that the pejorative effect of slurs is constituted by a cluster of associations that tie to (certain of) the *articulations* of slurs, due to a variety of socio-historical, cultural, psychological, and other factors. While it might be *prima facie* unpalatable to accept that the pejorative sting of slurs is not a matter of meaning, or even language, we've argued that there's overwhelming evidence for this view. The Articulation Account is uniquely well-positioned to explain the full range of puzzling data concerning the behavior of slurs and their pejorative sting.

We emphasize that while associations that constitute the pejorative sting can be weaponized to denigrate, harm or offend, they can also be triggered accidentally or negligently, or exploited for a variety of other effects: rhetorical, pedagogical, artistic and so on. But such tokenings do remain inherently risky. Whether any particular tokening is apt and necessary, or problematic, flippant or gratuitous, whether it constitutes a transgression, let alone one with special moral timbre, and whether any offense it causes is warranted, these are all questions parasitic on the fact that the tokenings of slurring articulations trigger the pejorative sting. But ultimately, we argue, the sting is not a matter of semantics or pragmatics, or even language: it is not slurring *words*, but *articulations* that carry the offensive sting.

References

Anderson, L., and Lepore, E. 2013. "Slurring Words. " *Noûs* 47 (1): 25–48.

Brandom, R. 2002. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge: Harvard University Press.

Camp, E. 2013. "Slurring Perspectives." *Analytic Philosophy* 54 (3): 330–349.

Carnaghi, A., and Mass, A. 2008. "Derogatory language in intergroup context: Are 'gay' and 'fag' synonymous?" In Y. Kashima, K. Fiedler, & P. Freytag (ed.). *Language-based approaches to the formation, maintenance, and transformation of stereotypes,* 117–134.

Cepollaro, B. 2015. "In Defense of a Presuppositional Account of Slurs." *Language Sciences* 52: 36–45.

Cepollaro, B., and Stojanovic, I. 2016. "Hybrid evaluatives." *Grazer Philosophische Studien* 93 (3): 458–488.

Cepollaro, B., Sulpizio, S., and Bianchi, C. 2019. "How bad is it to report a slur? An empirical investigation." *Journal of Pragmatics* 146: 32–42.

Dummett, M. 1981. *Frege: Philosophy of Language.* Cambridge: Harvard University Press.

Fasoli, F., Paladino, M. P., Carnaghi, A., Jetten, J., Brock, B., and Bain, P. G. 2015. "Not 'just words': Exposure to homophobic epithets leads to dehumanizing and physical distancing from gay men." *European Journal of Social Psychology* 46 (2): 237–248.

Frege, G. 1897/1979. "Logic." In H. Hermes, F. Kambartel, & K. F. (ed.). *Posthumous Writings: Gottlob Frege*. Blackwell, 126–151.

Hawthorne, J., and Lepore, E. 2011. "On Words." *Journal of Philosophy* 108 (9): 447–485.

Heller, S. 2000. *The Swastika: Symbol Beyond Redemption?*. New York: Allworth Press.

Hom, C. 2008. "The semantics of racial epithets." *Journal of Philosophy* 105 (8): 416–440.

Hom, C., and May, R. 2013. "Moral and semantic innocence." *Analytic Philosophy* 54 (3): 293–313.

Jeshion, R. 2013. "Expressivism and the offensiveness of slurs." *Philosophical Perspectives* 27 (1): 231–259.

Jorgensen Bolinger, R. 2017. "Pragmatics of Slurs." *Noûs* 51 (3): 439–462.

Kaplan, D. 1990. "Words." *Aristotelian Society Supplementary Volume* 64 (1): 93–119.

Kaplan, D. 2011. "Words on Words." *Journal of Philosophy* 108 (9): 504–529.

Kennedy, R. 2002. *Nigger: The Strange Career of a Troublesome Word*. New York: Pantheon Books.

Lepore, E., and Stone, M. 2018. "The Pejorative Tone." In D. Sosa (ed.). *Bad Words: Philosophical Perspectives on Slurs*. Oxford: Oxford University Press, 132–154.

Matisoff, J. 1986. "The Languages and Dialects of Tibeto-Burman: An Alphabetic/Genetic Listing, with Some Prefatory Remarks on Ethnonymic and Glossonymic Complications." In J. McCoy, Light, & T (ed.). *Contribution to Sino-Tibetan Studies*. Brill, 3–57.

McCready, E. 2013. "Varieties of conventional implicature." *Semantics and Pragmatics* 3: 1–5.

Neufeld, E. 2019. "An essentialist theory of the meaning of slurs." *Philosophers' Imprint* 19: 1–29.

Nunberg, G. 2018. "The Social Life of Slurs." In N. W. Acts, Godal, Daniel; Harris, Daniel; Moss, Matt (ed.). *New Work on Speech Acts*. Oxford: Oxford, 237–293.

O'Hehir, A. 2020. "So much for youth apathy: Student radicalism escapes the '60s at last." Retrieved from *Salon*: https://www.salon.com/2015/11/17/so_much_for_youth_apathy_student_radicalism_escapes_the_60s_at_last/

Potts, C. 2007. "The expressive dimension." *Theoretical Linguistics* 33 (2): 165–198.

Quine, W. 1953. *From a Logical Point of View*. Cambridge: Harvard University Press.

Schlenker, P. 2007. "Expressive presuppositions." *Theoretical Linguistics* 33 (2): 237–245.

Stojnić, U. 2022. "Just Words: Intentions, Tolerance, and Lexical Selection." *Philosophy and Phenomenological Research* 105 (1): 3–17.

Stojnić, U., & Lepore, E. 2025. *Inflammatory Language: Its Linguistics and Philosophy*. Oxford: Oxford University Press.

Williamson, T. 2009. "Reference, inference, and the semantics of pejoratives." In L. P. Almog Joseph (ed.). *The Philosophy of David Kaplan*. Oxford: Oxford University, 137–158.

Croatian Journal of Philosophy
Vol. XXV, No. 75, 2025
https://doi.org/10.52685/cjp.25.75.2
Received: November 1, 2025
Accepted: December 1, 2025

Slurs, Inflammatory Language*, and the Specificity Problem*

ROBIN JESHION
University of Southern California, Los Angeles, USA

In Inflammatory Language*, Una Stojnić and Ernie Lepore argue that no extant theory of slurs can explain slurs' hyperprojectivity, emphasizing their difficulties in accounting for acoustic and phonological resemblance cases in which a word merely sounds like a slur. Further, all content theories confront the Specificity Problem, the charge that the content view's content, whatever it is, is too specific to encompass the full range of competent weapon uses of slurs. One half of this paper concerns hyperprojectivity. I argue that there is a gap in Inflammatory Language's overarching dialectic that results from excluding a range of theories. Some theories of slurs are what I call single mechanism views: they aim to explain all the phenomena with a single explanatory mechanism. Multiple mechanism views exploit more than one. Within* Inflammatory Language*, multiple mechanism theories are bypassed. Yet multiple mechanism theories possess resources to explain slurs' hyperprojectivity. The other half of this paper addresses the Specificity Problem. I argue that a view I have developed in previous writings, Identity Expressivism, does not succumb to the problem. I craft a version of the Specificity Problem tailor-made for the theory and rooted in Stojnić and Lepore case against other expressivist theories. Identity Expressivism is, I argue, uncompromised by the Specificity Problem.*

Keywords: Slurs; semantics; pejoratives; epithets; expressives; hate speech.

1 The overarching dialectic of Inflammatory Language

Within their rich investigation of slurs in *Inflammatory Language*, Una Stojnić and Ernie Lepore advance both a negative and a positive thesis. Their negative thesis is that no extant theory of slurs can explain slurs' hyperprojectivity. Their positive thesis is that the Articulation Account

they develop within *Inflammatory Language* is capable of doing so.[1] This paper dominantly concerns their negative thesis. Here, I argue that their negative thesis does not go through.

I focus on Stojnić and Lepore's arguments for the negative thesis as applied to content theories. They advance two main arguments. The first is that all content theories are unable to explain slurs' hyperprojectivity, emphasizing inheritance cases on which a word sounds like a slur. The second is that all content theories confront the *Specificity Problem*, the charge that the content view's content, whatever it is, is too *specific* to encompass the full range of competent weapon uses of slurs.

The first half of this paper addresses the claim that content views are bereft of resources to explain hyperprojectivity. I argue that there is a gap in *Inflammatory Language*'s overarching dialectic that results from excluding a range of theories. Some theories of slurs are what I call single mechanism views: they aim to explain all the phenomena with a single explanatory mechanism. Multiple mechanism views exploit more than one. Within *Inflammatory Language*, multiple mechanism theories are bypassed. The reason why appears to be that Stojnić and Lepore either assume that all theories must be single mechanism or illegitimately rule out of hand multiple mechanism approaches. Yet multiple mechanism theories possess resources to explain slurs' hyperprojectivity.

The second half of this paper concerns Stojnić and Lepore's Specificity Problem. My aim is to demonstrate that the view I have developed in previous writings, Identity Expressivism, does not succumb to the problem. While Stojnić and Lepore don't directly address Identity Expressivism, I craft a version of the Specificity Problem tailor-made for the theory and rooted in Stojnić and Lepore case against other expressivist theories. Identity Expressivism is, I argue, uncompromised by the Specificity Problem.

2.1 Slurs' hyperprojectivity and mere orthographic and acoustic resemblance cases

Stojnić and Lepore frame their investigation in *Inflammatory Language* as one that primarily aims to explain what they call slurs' 'pejorative sting'. They ask: "What is the nature and source of [slurs'] pejorative sting, that makes slurs such powerful linguistic weapons?" (Stojnić and Lepore 2025: 2). In keeping with much of the slur's literature, they aim to account for how slurs function as tools of derogation, their capacity to cause offense when used literally as weapons, their derogatory and

[1] In Jeshion (2025), I discuss the positive thesis. There I explicate why the Articulation Account cannot explain how slurs function as tools of derogation, cannot account for instances of incomplete understanding of slurs, and is hard-pressed to analyze slurs' linguistic standing as pejorative expressions. I also raise concerns about some of their data.

offense potential when used within attitude attributions, in direct and indirect quotation, and even when the slur is mentioned. As noted, they advance two theses, one negative, one positive. The negative thesis is that no extant view in the contemporary literature is adequate to the task of explaining the full range of phenomena of how slurs sting. They break these down into two types, what they call content views and non-content views. Content views construe slurs as encoding or conveying a pejorative meaning or message. Different content views assign a different variety of content. Semantic content theories invoke a description, expressive, or perspective as a slur's meaning or within its use conditions. Pragmatic content theories typically convey descriptive, expressive, or ideological affiliation, yet do so via pragmatic mechanisms, like presuppositions or generalized conversational implicatures.[2] All content theories, say Stojnić and Lepore, are defective because "there is no pejorative content that can capture the behavior of a slur's offensive potential" (Stojnić and Lepore 2025: 2).

Non-content views make no appeal to meaning or pragmatically conveying a content. They appeal, rather, to sociolinguistic features of slurs or to the psychological impact of hearing a slur to account for their 'offensive potential'. For instance, Prohibitionism locates slurs' sting in their taboo status, in the offensiveness of breaking prohibitions on using or saying slurs (Anderson and Lepore 2013). Word Associationism locates that sting in associations triggered upon seeing or hearing a slur (Lepore and Stone 2014). For both non-content views, *slurring words themselves* are the primary source of slurs' sting. For Prohibitionism, speakers offend by uttering prohibited words. For Word Associationism, speakers trigger associations in their audience by uttering slurs. To Stojnić and Lepore, therein lies a problem common to all extant non-content theories, the fact that they ultimately root slurs' pejorative sting in slurring word themselves.

Why is there no pejorative content that can explain slurs' offense potential? And what is the error in rooting slurs' sting in slurring words themselves? According to Stojnić and Lepore, content and non-content views confront cases of pejorative sting via the *mere orthographic or acoustic resemblance to a slur*. These are their inheritance cases. Here is one: in 1999, a white aide to the mayor of Washington DC used the expression *niggardly* to describe a budget. Although the word shares orthographic and acoustic resemblance with the racial slur, it is unrelated etymologically and semantically. Nevertheless, upon hearing the aide's utterance, one of his black colleagues took offense, lodging a complaint. A similar case occurred in 2020, at my own university, USC. While conducting a lecture about cross-cultural dialogue, a communications professor noted how different languages employ different filler

[2] Some content views: Semantic: Bach (2018), Camp (2013), (2018), Davis and McCready (2020), Hom (2008), Hom and May (2013), (2015), Jeshion (2013b), (2018), Marquez and Garcia-Carpintero (2020), Potts (2007), Richard (2008). Pragmatic: Jorgensen Bolinger (2017), Nunberg (2018).

expressions. In English, we use 'um' and 'er'. Speakers of Mandarin, he noted, use a filler term, '那个', an expression that is pronounced like the n-word. In the lecture, the professor enunciated the word. He did so with no intention to provoke – indeed, apparently entirely oblivious to its possible effects. A group of students in the course expressed outrage over the incident, saying it caused them pain and upset. Call these *mere orthographic/acoustic resemblance cases.* Call the problem to explain all the 'offense potential', where this includes explaining derogation of weapon uses, derogatory variation, offense of belief attributions, mere mentions, and so on *the problem of slurs' hyperprojectivity*.

Of mere orthographic/acoustic resemblance cases, Stojnić and Lepore write:

> Such cases are particularly puzzling for content-based accounts: the offensive potential is inherited even though no slur is either used or mentioned. Indeed,…this type of data presents an issue for virtually all extant theories – content or non-content. The problem is that the expressions that inherit the pejorative effect are distinct from slurs and do not share either content or etymology with them. So, any account that ties the pejorative effect to an expression will leave unexplained why merely resembling an articulation of a slur leads to a pejorative effect being inherited by tokenings of a resembling expression…. (Stojnić and Lepore 2025: 75)

They add that for content views specifically, orthographic/acoustic resemblance cases present an additional issue: "the explanation of inheritance cannot be gotten by way of meaning – for the expressions that inherit the pejorative effect don't share any aspect of the slur's meaning" (Stojnić and Lepore 2025: 75).

Stojnić and Lepore's other thesis is positive: that the novel view they craft and dub the Articulation Account can explain all the phenomena. According to the Articulation Account, the offensive potential of slurs resides in negative association triggered by

> certain of [a slur's] articulations – its phonological and/or orthographic forms. These associations attach to particular articulations of slurs through multifarious factors – causal, historical, cultural, and psychological. Token articulations trigger them even when they do not accompany a tokening of a slur, while tokenings of slurs, absent particular standard articulations that harbor these associations, fail to trigger them. In short, the offensive potential has nothing to do with slurring words; it is a result of associations triggered by articulations of slurs. (Stojnić and Lepore 2025: 2)

For them, articulations and not slurring words are the ultimate source of slurs' sting. The theory is, then, expressly designed to account for mere orthographic/acoustic resemblance cases. Since, in the case of the Mandarin filler term it is, they say, the sheer acoustic resemblance that provoked the outrage, they invoke the sound itself as what accounts for slurs' 'offense potential'. They maintain further, that by appealing *exclusively* to articulations, the Articulation Account successfully accounts for all the phenomena, including all aspects of slurs' hyperprojectivity. This is their *Comprehensiveness Thesis*. They also maintain that their positive account is "uniquely well-positioned to account for

the full range of data and do so in a fully uniform manner" (Stojnić and Lepore 2025: 6). This is their *Uniformity Thesis*.

2.2 Single mechanism and multiple mechanism theories of slurs

Implicit in Stojnić and Lepore's *Comprehensiveness* and *Uniformity* Theses is a commitment to what I call a single mechanism approach to theorizing about slurs. As its name suggests, single mechanism accounts posit a single mechanism to explain all the phenomena regarding the full range of slurs' hyperprojectivity. The single mechanism that Stojnić and Lepore embrace to explain slurs' pejorative potential is articulations. Other theorists also appear to plumb for a single mechanism approach. For instance, Anderson and Lepore emphasize but one mechanism – prohibition-breaking – to explain hyperprojectivity (Anderson and Lepore 2013).

By contrast, multiple mechanism approaches posit various mechanisms to explain the full range of phenomena regarding the derogatory and pejorative potential of slurs. In earlier papers, I advanced a multiple mechanism theory of slurs (Jeshion 2013b, 2018). The theory advocates an Identity Expressivist semantics of slurs whose primary role is to capture why slurs are, by convention, pejorative expressions, and how, in weapon uses speakers derogate their targets, a matter fully independent of whether anyone takes offense to the act or any downstream causal harmful and painful effects due to the utterance (Jeshion 2013b). The account couples the expressivist mechanisms in the semantics with additional pragmatic and sociolinguistic features of slurs, for there are a "plethora of reasons why a use of a slurring term can be offensive, and we must tease apart the sources of offensiveness of slurring terms due to their semantic properties and those that are due to pragmatic phenomena and sociolinguistic properties"(Jeshion 2013b: 234).

I advanced several mechanisms contributing to the offense profile of slur-utterances that supplement those that derive exclusively from the semantics. One was stereotype-activation. While, I argued, slurs do not semantically encode stereotypes, they have the power to activate, to trigger, thoughts of stereotypes in hearers, and cause harm and offense due to that activation (Jeshion 2013a, 2013b). Another mechanism concerned how utterances of slurs interact with historical institutional bigotry and oppression of targets: in general, utterances of slurs will compound the effects of oppressive ideologies and will cause far more psychological and social damage to those in historically oppressed groups. Still, because the mechanism is causal and associationist, the exact impact will vary radically from utterance to utterance, and from slur to slur. In addition to stereotype activation and interaction with ideologies and historical oppression, other psychological and social factors can contribute to intensifying or diminishing the impact of a slur.

The main message: as always with perlocutionary effects, downstream discourse effects are wide-ranging, manifold, complicated. And finally, I argued that Lepore and Anderson were right to identify prohibition-breaking as important. While it is just one piece of the story and cannot explain derogation, breaking the taboo of uttering a slur does cause offense to those invested in the prohibition's integrity. Thus, another sociolinguistic mechanism, prohibition-breaking, contributes to the offense profile of slur utterances (Jeshion 2013b). Several other theorists, including Diaz-Legaspe, Liu, and Stainton, Rappaport, and Rinner and Hiecke have also advanced their own multiple mechanism theories or at least advocated handling some of the hyperprojective phenomena with multiple mechanisms.[3]

2.3 Exclusion of multiple mechanism accounts

Return now to *Inflammatory Language*'s negative thesis that no extant theory can explain slurs' hyperprojectivity. The most important thing to recognize about Stojnić and Lepore's dialectic is that all of arguments are directed at single mechanism views. They curiously ignore multiple mechanism theories. Moreover, throughout the book, they repeatedly deploy a mode of argumentation that appears to rule out of hand multiple mechanism approaches – perhaps for an underlying rationale that they do not specify. Alternatively, they may simply implicitly assume that all theories must be single mechanism. In fact, in their critical discussion of many different views, they never allow for a theory to tack on an additional mechanism to explain aspects of hyperprojectivity, including mere orthographic/acoustic resemblance cases. This is precisely what Rappaport does to account for the toxicity in mere mentions, positing that slurs undergo distinct neurolinguistic processing, routed through their phonological forms. He could easily deploy it as well to explain mere acoustic resemblance cases. The move seems in no way implausible, and certainly not illegitimate *for* simply adding an additional explanatory mechanism.

Let's capture Stojnić and Lepore's implicit assumption thus:

> *Single Mechanism Necessity Assumption*: if a theory is unable to explain hyperprojectivity with a single mechanism, that theory cannot explain hyperprojectivity.

[3] Diaz-Legaspe, Liu, and Stainton (2019) offer a register-based account of the conventional rules governing slurs, which would count as a content view. It is multiple mechanism because they acknowledge that nonlinguistic performance effects are necessary to fully account for hyperprojectivity. Rappaport (2019) advances a theory that couples together a Camp and Nunberg style group-allegiance or perspective signaling plus a subsidiary neurological processing mechanism to explain the toxicity of slurs in indirect quotation and slur mentions. Rinner and Hiecke (2021) do not themselves advance a particular semantic or pragmatic analysis of slurs. Nevertheless, they argue that content views can handle derogation, while another mechanism handles slur mentions, namely the psychological efficacy (via associations, presumably) of being reminded of the content.

The Single Mechanism Necessity Assumption is false. After all, there are several multiple mechanism theories of slurs that, in addition to positing a content, advance other mechanisms – hearers' associations, neurolinguistic factors, stereotype activation, historical facts, and sociolinguistic properties of words – to explain hyperprojectivity. They are not, *for that*, unable to explain hyperprojectivity. And they are not, *for that*, inadequate theories.

So, the Single Mechanism Necessity Assumption is false. For this reason, one might naturally be skeptical of the claim that numerous arguments in *Inflammatory Language* depend upon it. To back this claim, then, consider the following three arguments.

First, addressing Potts' expressivist conventional implicature theory,[4] Stojnić and Lepore note that different slurs for the same group vary in their pejorative effects and impact. They write:

> This is particularly difficult to capture on attitudinal or fully non-descriptive approaches. Such accounts *have to argue that distinct slurs for the same target group signal different negative attitudes of differing strength*, in way that underscores speaker intuitions about the associated pejorative effect. We find it highly dubious that fine-grained distinctions in attitudes or affects …underwrite such intuitions. We are thus pessimistic about the success of CI accounts. (Stojnić and Lepore 2025: 51, my emphasis)

This argument presumes that Potts has no resources available to him other than his semantics to explain derogatory variation of different slurring words for the same group.

Second, addressing Camp's perspectivalist theory,[5] Stojnić and Lepore note that slurs often "evoke facts that go well beyond" that which is encoded in perspectives, in particular they may invoke "socio-historical facts about who used or coined the term, including socio-cultural facts concerning power structures and structures of oppression; or various phonological/orthographic features"; or they may evoke imagery. The only way for Camp to rectify her theory, they maintain, is to widen perspectives to incorporate "all open-ended associations that can affect pejorative potential." This, they say, only exacerbates problems by bloating the semantics, making disagreements in pejorative effect linguist disagreements (Stojnić and Lepore 2025: 59–61).

Lastly, in summarizing their case against content views, Stojnić and Lepore claim that all such views are faced with an insurmountable problem of slur-meaning insulation, namely, that in quotation and in mere displays of slurs, the slur's meaning is inert. They write:

> we have seen that the projective behavior of the pejorative potential is way more robust than any of those accounts – semantic or pragmatic – predict. Indeed, the effect remains even in environments that render meaning inert – meaning attributions, quotative environments, and even mere displays of slurs. In as much as these environments render meaning inert, that they

[4] Potts (2005), (2007).

[5] Camp (2013), (2018).

> still preserve the offensive potential of slurs suggests that the latter *cannot be cashed out in terms of meaning*. (Stojnić and Lepore 2025: 129)[6]

Their concluding sentence, that the persistent sting of slurs within meaning-insulated constructions *cannot* be explained in terms of meaning is presented here as a nail in the coffin of all content theories.

In each of the three arguments, Stojnić and Lepore *assume* that the only mechanism available to explain derogatory variation, offense of a use or mention, or same-sounding articulation is the semantic mechanism the theorist embraces: expressivist conventional implicature, for Potts; perspectives for Camp; and for all other content theories, whatever mechanism is exploited to capture content. From this, they conclude – rightly, I maintain – that content, however it is cashed out, is insufficient to account for *all* the phenomena. Yet they then conclude from *that*, that the theory itself is thereby defective. This last inference depends essentially on the false *Single Mechanism Necessity Assumption*.

Now, what general lessons we can learn from this dialectical gap and assumption within *Inflammatory Language*? I offer three. First, *assuming* that the goal of our communal project of understanding slurs is to capture all the effects, intended and unintended, of every communicative act and every sound people make that involves the assertion or mention of a slur and of all intentional and unintentional acts involving humans making noises that sound like a slur, then every theory can – and I think should – reject a single mechanism approach. No content theorist is *restricted* from supplementing their account by positing pragmatic and sociolinguistic mechanisms to explain this extraordinarily vast range of discourse effects. Further, it should not be assumed that the mechanisms used to explain how slurs derogate and cause offense be exactly the same as those used to explain why homonyms can trigger offense.

Second, this raises an interesting question about the proper methodological principles for multiple mechanism theories. Obviously, appealing to one mechanism is always acceptable. Equally obviously, adding mechanisms *willy-nilly* is unacceptable. Advancing and justifying guiding principles and rationales is a large task, not one I can take up fully here, though some of §3.1 speaks to this matter.

Last, we must take stock and ask whether our communal project ought to be this vast in its scope. Is it in fact incumbent upon a theory of slurs to account for all the intended and unintended discourse effects of utterances and mentions of slurs and all articulations that may sound like slurs? While this is an important methodological question, here is not the place to take a deep dive into what, precisely, our com-

[6] This style of argument also occurs in numerous other places throughout the book. Cf., chapter three, where the authors maintain that the only way for a presuppositional theorist to account for the different discourse effects of different slurs is to "posit different pejorative presuppositional contents for different slurs."(42) Cf., also their footnote 75, (61–62).

munal project should be. But I will register my skepticism that it is incumbent upon us to grapple with every effect of a speech act that looks or sounds like a slur and, in particular, to treat it as on a par with intentional weapon uses of such terms. What is puzzling is that mere orthographic/acoustic resemblance cases are the prime variety of datum inspiring Stojnić and Lepore's Articulation Account. In my view, this is misguided. As I noted elsewhere, outrage in response to slur homophones that are not the n-word are exceptionally rare.[7] Further, it seems to me a banal fact about communicative acts, and intentional acts more generally, that those that appear to listeners to have the same properties of those that systematically cause offense will also sometimes trigger that same effect. Such phenomena shouldn't induce novel theory.

In sum: Stojnić and Lepore's negative thesis is that all extant theories are unable to explain slurs' hyperprojectivity. Content views cannot appeal to content to explain why utterances on non-slurring words that sound or look like slurs trigger offense. And since both content and non-content views maintain slurs themselves – the words, not their articulations – are the ultimate source of their pejorative sting, both cannot explain the offense triggered in mere orthographic/acoustic resemblance cases. Thus far, I have taken a preliminary step toward undermining this negative thesis by arguing that Stojnić and Lepore suppose that all theories must deploy but a single explanatory mechanism. This assumption is unargued and false. Any particular multiple mechanism theory of slurs may well be wrong. But they are not wrong *for* positing multiple mechanisms. This leaves a path open for the viability of multiple mechanism theories.

3 The Specificity Problem: A defense for Identity Expressivism

In this second section, I extend the case against the negative thesis by demonstrating the power of my favored multiple mechanism theory, a view I call Identity Expressivism.[8] On Identity Expressivism, slurs are tools of derogation. In weapon uses of slurs, speakers regard their targets with contempt, and with the slur express their contempt. It counts squarely as a content view. Stojnić and Lepore's most powerful arguments against all content views is given by their *Specificity Problem*. This is the charge that the content in any given content view is too *specific* to encompass the full range of competent weapon uses of slurs. I demonstrate that key features of Identity Expressivism disable the Specificity Problem. I begin by briefly laying out four methodological assumptions I adopt. Next, I give a bare-bones overview of Identity Expressivism, and then turn to the Specificity Problem.

[7] Jeshion (2025).

[8] The view is developed in Jeshion (2013b), (2016), (2017), (2018).

3.1 Methodological assumptions of Identity Expressivism

Different methodologies, starting points, and underlying assumptions often strongly impact an analysis of slurs. Although I cannot detail the whole range or offer full justifications for them, here I lay out four of my own, all of which I believe are or ought to be shared by Stojnić and Lepore. Each plays a critical role in the dialectic regarding the Specificity Problem.

One: Slurs are pejorative lexical items.
From a linguistic perspective, there are two basic starting points for examining slurs. One is to view slurs, first and foremost, as *types of lexical items*. Another is to take as most basic a type of speech act, the act of slurring, where such acts can occur in the absence of a word standardly branded as a slur. I adopt the first perspective. This approach affords insight into how slurs are related to other lexical items, including other pejoratives and dysphemisms, and also kind terms, names, and the many types of social deictics, including honorifics, pronouns indicating (in)formality, nicknames, diminutives, and other hypocorisms. Critically, this starting point allows us to understand slurs' lexical dynamics: how they can change their linguistic properties over time and become polysemous.[9]

Two: Basic uses of slurs are weapon uses, applied to the target group.
Slurs are used in various ways. Notoriously, slurs are used to derogate and dehumanize. Yet they are also used, typically by target group members themselves, to neutrally reference the group or group-member or to convey positive valences of solidarity, endearment, friendship, or empowerment. Slurs are also used to express bigotry that is exception-making and to widen the domain of application beyond the target group.

Ideally, a full theory of slurs ought to have the resources to explain *all* of the phenomena. However, no linguistic theory that isn't hyper contextualist will be able to satisfy all the phenomena without appealing, at some point, to slurs' polysemy. I adopt the following approach: in basic uses, slurs are used to pejoratively reference an important social group, defined by race, ethnicity, nationality, religion, gender, sexual orientation, and so on. Often such social groups have a corresponding non-derogating term, a neutral counterpart, as a preferred term for the group – but not always.

Basic uses have two properties. They function to derogate. They are thus aptly deemed *weapon uses*. Basic uses also aim to refer to all and only those in the group that the slur targets. I call such uses *group-*

[9] Stojnić and Lepore treat slurring terms as first and foremost lexical items, more primitive than slurring speech acts. They announce that 'slurs are pejorative by design'(2). However, it is unclear how, precisely, they explain slurs' standing as *pejorative* lexical items. Cf., Jeshion (2025).

referencing (G-referencing) uses of slurs. Identity Expressivism's semantics is designed to apply to these basic uses.

While I construe group-referencing weapon uses as basic, Identity Expressivism is also designed to account for *non-basic uses*, like those with positive and neutral valence and with different extensions (what I call G-contracting and G-extending uses). I explain non-basic uses as securing what they convey via novel extensions from basic uses. The theory uses familiar tools of semantic change and polysemy to explain this multiplicity of meanings.

Note: our methodological assumption that selects a class as basic does *not* entail that weapon group-referencing uses are the only uses whose meanings are conventionalized. The n-word has a conventionalized meaning as a social deictic to convey friendship. Similarly, our methodological assumption does not entail that for every slurring word, its weapon group-referencing use is currently or at some other time the most common use of that expression. For instance, the slurring word *queer* has a basic use, a weapon group referencing use, whereby it functions as a linguistic tool to derogate LGBTQ+ persons. Through most of the twentieth century, in most western communities, the basic use was the dominant use. With the reclamation of *queer* beginning in the late 1980s, the word became polysemous: the neutral use designating a group with a novel identity – *being queer* – proliferated, becoming conventionalized. Currently, at least in many western locales one quarter way through the twenty first century, the basic use is, happily, no longer the dominant use. This is all compatible with Identity Expressivism.

Three: Sharply distinguish derogation, offensiveness, and the moral assessment of uses of a slur.

I sharply distinguish between *derogation* and *offense*.[10] I use *derogate* and cognates to label what speakers do – what they perform – when deploying slurs as weapons. Such speakers derogate their targets and they do so regardless of how anyone responds to the act. Derogation is, therefore, not the same as *offense taken*, understood as an actual response to an utterance. It also differs from *offense potential*, understood as a possible response to an utterance or as a measure of possible responses. Offense taken and offense potential both concern *post-act* (actual and potential) *responses* to a slur content, utterance, mention, attribution, or even to non-slur articulations, utterances that just sound or look like a slur.

[10] Officially, Stojnić and Lepore would appear to agree. Slurs, they say, are "epithets that *derogate* purely on the basis of group membership, e.g., on the basis of race, ethnicity, origin, religion, gender, sexual orientation, or ideology."(2, emphasis mine) and are "used to derogate, disparage, offend, insult, or cause harm."(5) Nevertheless, they standardly conflate the distinction between derogation and offense with a covering term 'pejorative potential' or 'offensive potential'. This creates problems for their Articulation Account's ability to explain cases (below) like *Homophobia*, a point argued in Jeshion (2025).

I sharply separate the *moral assessment* of an act of using a slur from the non-normative account of what speakers do with slurs. That is, to say that a speaker derogates or offends with a slur is not *thereby* to say that their act is worthy of moral criticism or condemnation. That is a further matter.

Four: A linguistic theory of slurs must be capable of explaining both derogation and offense.
All agree that slurs are tools of derogation. In the vast range of basic uses of slurs, speakers derogate *and* cause offense. How can we separate the two? The best way is to consider cases where a speaker derogates with a slur while the actual and potential offense are made null. Here is one:

> *Homophobia*: A and C are friends, both homophobic. Their homophobic attitudes are mutual knowledge between them. They frequently use *f****t* amongst themselves to designate people they deem gay. Both A and C believe that B is gay. Both know that B does not understand the slur – for, they know, B barely understands English. In isolation of all other people except C, A says to B, 'You are a f****t'. Since B has never heard the slur before, he has no associations whatsoever with it. Going forward, B will never be in contact with anyone who does have any associations. A and C know this fact.

Given the set-up, there is neither any actual offense taken nor any offense potential. Yet with the slur, A has derogated B. This is palpable and should be uncontroversial. That there is derogation here is rooted straightaway in intuition. It employs no theory-laden assumptions about the mechanisms to account for the derogation. Theories of slurs ought to be able to explain the derogation in cases like *Homophobia*.

3.2 Identity Expressivism

Identity Expressivism advances a semantics for what I call canonical slurring terms, those attacking race, nationality, religion, sexual orientation, ability, political affiliation, and possibly other important social categories.[11] In their basic uses as weapons, slurs have three components to their semantics:

> *Group Designating Component:* A slurring term designates the group G or set of individuals that is designated by its neutral counterpart. Slurs are designationally-equivalent with their neutral counterparts. The group-designating component contributes to the

[11] I contrast canonical slurs from descriptive slurs like *beaner* for Hispanic and *slanty-eyes* for Asians. Like canonical slurs, descriptive slur refer to a target group, yet they also contain an additional stereotype that is truth-conditionally inert. *Beaner* can be 'aptly' applied to Hispanic persons that do not eat beans. I also contrast canonical slurs with gendered slurs like *bitch*, *c**t, *slut*, for girls and women, and *sissy* and *pussy* for boys and men.

truth-value of propositions in the same way that neutral counterparts do.

Expressive Component: Slurring terms are vehicles for expressing affective stances and attitudes: with a slurring term, speakers express contemptuous regard for the members of a group G on account of their being in G or having a group-defining property g. They express that the target is lesser, on a moral dimension, on account of being in the group. The expressive component makes no contribution to truth-conditions.

Identifying Component: Slurring terms are tools for specifying, what, by the speaker's lights, the targets of the slur *are*. Speakers use slurs to map a negative characteristic-defining *social identity* onto the members of the group G. The identifying component makes no contribution to truth-conditions.

For example, [1] and [2] are designationally and truth-conditionally equivalent,

[1] Jake is a Kike

[2] Jake is a Jew

Yet *kike* differs from *Jew* insofar as it is used to express contempt for Jews on account of being Jews. In expressing contemptuous regard, the speaker of [1] does not *say* or *assert* the truth-evaluable proposition

[3] Jake is a person deserving of contempt on account of being Jewish.

And neither does the speaker *say* or *assert*

[4] Jake has the fundamental negative character-defining feature of being Jewish.

It is rather that *by* expressing an affective stance of contempt for Jews on account of being Jews, the speaker displays that what the speaker is, *qua* person, is a Jew and thereby lesser.[12] In this way, the semantics of [1] is essentially equivalent to that of [5] and [6],

[5] Jake is a goddamn Jew

[6] Jake is a Jew^{C}

where *goddamn* in [5] functions as a pejorative expressive modifier, and the superscript '*C*' in [6] indicates that *Jew* is spoken with contemptuous intonation.

Lastly, in giving this expressivist semantics of slurs, we have yet to say how slurs derogate. After all, words – particular lexical items – are not by themselves derogations. *Derogations are acts,* communicative acts, performed by people. Identity expressivism accounts for how, with a basic use of a slur, speakers derogate their targets. In sincerely using a slur with full understanding, a speaker expresses contempt for the

[12] These characterizations leave off many fine-points. Cf., Jeshion (2013b) and (2018) for far more depth.

target, and consequently both *communicates* and *treats* the target as a lesser person, as less deserving of full respect, on account of being in the group.[13]

3.3 The Specificity Problem

Finally, we return to the Specificity Problem. Recall that Stojnić and Lepore advance it as the challenge that the content view's content, whatever it is, is too *specific* to encompass the vast range of competent uses of slurs. Importantly, unlike their point about slurs' hyperprojectivity, the Specificity Problem is directed toward basic weapon uses of slurs: not reclaimed uses, not belief attributions, not mentions, not words that sound like slurs. To them, the Specificity Problem undermines all content views, whether the content is descriptive, a speaker's attitude, an emotion, or an ideology encoded as a perspective. Since Stojnić and Lepore don't themselves address Identity Expresivism, I detail how they wield the Specificity Problem against semantic accounts that also posit a close connection between slurs and the emotions: Schlenker's attitudinal presuppositional account of slurs and versions of expressivism due to McCready and Potts. From this, I craft the strongest version of the Specificity Problem for Identity Expressivism, and demonstrate why it fails to undermine the theory.

In *Inflammatory Language*, Stojnić and Lepore present a version of the Specificity Problem to a view that bears *some* similarities to Identity Expressivism, Schlenker's attitudinal presuppositional account of slurs.[14] On Schlenker's view, a slur is truth conditionally equivalent with its neutral counterpart. Here it aligns with Identity Expressivism. Schlenker's analysis differs from Identity Expressivism along several dimensions. For him, the slur is governed by a presupposition that the speaker believes that members of the group the slur targets are despicable. Thus, while Identity Expressivism offers a rule of use to characterize the semantics, Schlenker posits a presupposition as the mechanism to convey the speaker's attitude. More important with regard to the Specificity Problem, Schlenker's view requires the speaker has a standing belief that members of the target group are despicable. Identity Expressivism makes no such assumption. It trades on speakers having affective stances on the group, not any particular belief about them. Finally, Schlenker's belief content is indeed quite specific: that the target group and its members are despicable.

In pressing the problem to Schlenker's account, Stojnić and Lepore claim that it is coherent and perfectly felicitous to use the slur while issuing a denial of the alleged presupposition (Stojnić and Lepore 39-43). Here are two of their key examples against Schlenker: [15]

[13] For a much fuller treatment of the moral psychology and how dehumanization if performed, see Jeshion (2018).

[14] Schlenker (2007).

[15] Here and throughout the rest of this section, my examples are extracted directly from *Inflammatory Language*, yet are altered by shifting the slur, and the

[7] Jake is a kike, but I don't believe that kikes are despicable.

[8] There are so many f****ts in that bar. But I don't believe that f****ts are despicable.

Though my reasons for finding [7] and [8] coherent may well be different from theirs, I agree with Stojnić and Lepore. In each utterance, the speaker can be competently using the slur as a weapon while denying they believe that Jewish or gay people are despicable. *Believing* that a whole group is despicable is one thing. Having an *affective stance* of contempt toward a group is another. Contemptuous regard may exist in the absence of standing beliefs that a group is despicable – or, in fact, even that the group is contemptable. And thus, using a slur is compatible with outright sincere denials that the target group is despicable. In my view, this version of the Specificity Problem appropriately pinpoints my own prime concern with Schlenker's account, namely that competent use of a slur does not require that speakers have any specific negative *beliefs* about the group – and further, that such beliefs are not, in the first instance, what is conveyed or otherwise communicated with slurs.

Stojnić and Lepore also wield the Specificity Problem towards views considerably closer to Identity Expressivism. One is McCready's expressivist view, whereby speakers using slurs express negative attitudes toward the slur's target. The problem is that "predications of slurs in the absence of such attitudes are not linguistically infelicitous".[16] To demonstrate competent use of a slur in the absence of negative attitudes, they offer examples like these:

[9] I don't think C***ks are despicable. [49]

[10] John is a n****r. N****rs are very respectable people. [49]

[11] F****ts are good! We should respect them. [49]

Of [9]-[11], Stojnić and Lepore say they are neither linguistically incoherent, linguistically deviant, nor linguistically infelicitous.

The other view they address is inspired by Potts' treatment of expressives (Potts 2005, 2007) like *damn* when used as intensifiers to modify nouns, as in

[12] That damn dog is barking again!

According to Potts, the speaker must be "undergoing a heightened, negative affect towards the dog". On such a view of slurs, in

[1] Jake is a kike

the speaker expresses a heightened negative affect toward Jews. To such a view, Stojnić and Lepore write

person. They are also trimmed to excise redundancy. I choose to insert slurs in the example sentences, replacing Stojnić and Lepore's Harry Potter 'mudblood', for several reasons. Chief among them is that I had to myself reread the sentences with real slurs in order to test them for felicity. Mistakes in felicity judgments can far more easily arise from considering fictional rather than real-life slurs.

[16] Cf., McCready (2010), Stojnić and Lepore (49).

> one needn't have a negative affect toward the target group – not even a standing one, let alone occurrent – in order to felicitously utter the term. We can imagine a casual conversation among bigots, who invariably drop slur terms to reference the target group, regardless of whether they are currently experiencing any negative affect….Such utterances are perfectly felicitous when conjoined with denials of negative attitudes and affects. This, however, is not so for ordinary expressives, such as 'damn', which are infelicitous when conjoined with any such denial. (Stojnić and Lepore 2025: 50).

Here are the contrast cases Stojnić and Lepore advance to support their objection:

> [13] Leo is a Kike. But, to be clear, I have no ill feelings toward kikes.
>
> [13'] #Leo is a goddamn Jew. But to be clear, I have no ill feelings toward Jews.
>
> [14] You are a f****t. I love f****ts, they are wonderful people!
>
> [14'] #You are a goddamn homosexual. I love homosexuals, they are wonderful people! (Stojnić and Lepore 2025: 50, 76).

According to Stojnić and Lepore, the predications with the expressive intensifier *goddamn* modifying *Jew* and *homosexual* become linguistically infelicitous with avowals of respect and love, as in [13'] and [14']. Here they agree with Potts. However, they maintain, the slur predications in [13] and [14] do not. They are linguistically aright, entirely felicitous.

Now, will these objections directly transfer to Identity Expressivism? Not straightaway. While the McCready and Potts accounts are both expressivist, they are in some important respects different from my own. The most fundamental is that both require that with a slur, speakers express some *negatively valanced attitude* or *affect*, not, in particular, contemptuous regard.[17] With this in mind, we can construct a version of the Specificity Problem tailor-made for Identity Expressivism. It goes as follows:

Identity Expressivism's expressivist component is too confining because:

[17] There is another important difference with the Potts analysis. Whereas Potts holds felicitous uses of an expressive modifier like *damn* and *fuck'n* requires the speaker be in a 'heightened emotional state right this minute', (Potts 2007: 171), that is a claim I explicitly reject – and for *both* expressives modifying nouns for social groups and for slurs. When combined with nouns as in *goddamn Jew* and *fuck'n homosexual*, the combined expression functions, on my view, exactly like slurs do. As we saw above, on Identity Expressivism, *Jake is a kike* is *synonymous* with *Jake is a goddamn Jew*. On this score, Stojnić and Lepore differential assessments of [8] and [8'] and [9] and [9'] seem to me exceptionally surprising and counterintuitive. We should register the same assessments about saying *You are a f****t!* and *You are a goddamn homosexual*! Following up either one with *I love faggots/homosexuals*! is certainly linguistically deviant. In what follows, I argue the full case for the slur.

I. Competent users of slurs need not be occurrently experiencing feelings of contempt toward the group.
II. Competent users of slurs need not have any standing feelings of contempt toward the group.
III. Competent users of slurs may entirely lack all contempt or contemptuous attitudes toward the group.

In support of I-III, they can use the linguistic examples above. The rationale then becomes:

IV. [9], [10], [11], [13], [14] manifest competent use. They are in no way linguistically incoherent, deviant, or infelicitous.

This is our construction of Stojnić and Lepore's *Specificity Problem for Identity Expressivism*. It is an important, radical, and powerful challenge.

The key to why this objection does not compromise Identity Expressivism is to understand the nature of contempt and the way it can be expressed in language. In previous work, I advanced a detailed analysis of contempt. A main impetus of the analysis was to address earlier, less radical versions of the Specificity Problem by Camp and Anderson and Lepore.[18] There I drew heavily on two pools of research: one, empirical research by emotions theorists like Paul Ekman, Klaus Scherer, Ira Roseman, and June Tagneny; the other, philosophical analyses by moral psychologists like Michelle Mason, Macalaster Bell, and Steve Darwall.[19] The following summarizes main points pertinent for combatting Stojnić and Lepore's Specificity Problem for Identity Expressivism. All of [i]-[vii] concern the *nature of contempt* itself.

[i] Contempt is a complex, though basic emotion, typically experienced as a *form of regard*. It is an *affective stance*.

[ii] Contempt is a hierarchizing affective stance. In harboring contempt for a person, one looks down on them, regarding them as low, as lesser.

[iii] Contemptuous regard must be sharply distinguished from contempt's *behavioral manifestations*. These can vary enormously,

[18] Cf., Jeshion (2018). Camp's objection is that not all weapon-uses of slurs are associated with contempt. Although some are, she acknowledged, others are associated with "disgust, fear, dismissiveness" Camp (2013: 10). Further, one need not outwardly manifest contemptuous regard like hateful feeling. To underscore the point, she offered examples likes this felicitous weapon-use of slurs: *I'm glad we have so many Sp*cs at our school; they always bring the best food.* Anderson and Lepore offered examples that register compliments toward the group: *Ch*nks are so much smarter than the rest of us.* I argued that Camp and Anderson and Lepore's challenges rest on false assumptions about the nature of contemptuous regard and its expression in language.

[19] Cf., Ekman and Friesan (1986), Ekman and Heider (1988), Ekman (1994), Mason (2003), (2016), Scherer (2003), (2013), Tangney, J. Stuewig, J. & Mashek, D. (2007), Bell (2013), Roseman (2018).

ranging from manifestations of hate, demonstrations of disgust, mocking laughter, sneering, condescending pity, amused dismissiveness, simmering resentment, and simple indifference.

[iv] Because it is a standing form of regard, having contempt toward X at time t is compatible with the absence of behavioral manifestations of contempt toward X at time t.

[v] Contemptuous regard doesn't have a specific phenomenology. Unlike emotions like anger and fear, there is no specific way that it feels like to regard another with contempt. There is no well-defined 'feeling of contempt'.

[vi] Having contemptuous regard toward X is compatible with appreciating X's positive features.

[vii] Having contemptuous regard toward X is compatible with being self-blind to one's own contempt. People harbor contempt toward others without knowing they do.

From these features of the nature of contempt, we can straightforwardly derive three points about the *expression of contempt in language*.

[viii] One can felicitously use a term that expresses contemptuous regard without accompanying it with any behavioral manifestations of contempt.

[ix] One can felicitously use a term that expresses contemptuous regard without experiencing or communicating any alleged 'feelings of contempt'.

[x] Expressing contemptuous regard for X is compatible with sincerely communicating appreciation of X's positive features.

For deeper understanding of and support for each point, I refer the reader to the full analysis and especially the empirical and philosophical accounts that underwrite it. Even still, absent that, it is easy to recognize the truth of many of them. For instance, if you regard a certain president with contempt, you can do so *sans* behavioral manifestations, and without any specific phenomenology. Further, with respect to [vi] and [x], there is simply no problem with your expressing your contempt thus

[15] *Goddamn Trump* came into office in 2025 fantastically organized and focused,

while acknowledging (resentfully) positive properties.

These features of contempt straightaway undercut claims I and II of Stojnić and Lepore's Specificity Problem as a case against Identity Expressivism. Bigots can indeed competently use slurs without occurrently experiencing 'feelings' of contempt or manifesting any behaviors that signal contempt. Thus, Identity Expressivism is in fact compatible with I and II.

What about III, Stojnić and Lepore's claim that no negative attitude or affect or stance is necessary for competent use?[20] This is different. Identity Expressivism *requires* that competent speakers have contemptuous regard, a stance obviously incompatible with the total absence of standing negative attitudes, affects, feelings, and opinions about the target group. In support of III, Stojnić and Lepore offer a series of examples, consolidated here.

[9] I don't think C***ks are despicable.

[10] John is a n****r. N****rs are very respectable people.

[11] F****ts are good! We should respect them.

[13] Leo is a Kike. But, to be clear, I have no ill feelings toward kikes.

[14] You are a f****t. I love f****ts, they are wonderful people!

These uses of slurs are coupled with: denials of thinking the group despicable [9], avowals of respect [10] and directives for all to respect the group [11], denials of having any negative feelings [13], and expressions of reverence and love toward the group [14]. I take each in turn.

Unlike the rest, there is no problem with [9]: sincere, knowledgeable denial that one thinks the group despicable can be felicitously used with a slur, as discussed above in the Schlenker analysis. Furthermore, contemptuous regard only requires looking down on the group, seeing them as lesser; it does not require thinking they are *despicable*. Thus, [9] offers no support to III, as it does not show absence of all negative feeling. And by itself, [9] is fully compatible with Identity Expressivism.

By contrast, the avowals of respect in [10] and directives for all to respect the group in [11] strike one – strike me – as highly marked, at least if the respect entails regarding the group as *in no way* lesser. To appreciate this, remember that in testing for felicitousness, we must assume the speaker's use of the slur is a basic use, a weapon use. We must also assume that the speaker possesses full linguistic understanding of the slur. In the absence of these conditions, [10] and [11] might sound acceptable. Otherwise, they land as highly marked, smacking of a speaker ludicrously squirming toward some curious form of plausible deniability.

The same should be said for the wholesale denial of all negative feelings in [13] and avowals of love and reverence for the group in [14]. They immediately land as deviant, infelicitous. One wants to say: why are you derogating the group you (profess to) love? Why choose the slur, if you truly harbor no ill feelings? To even make sense of such a weapon use with full understanding, we lean toward ascribing some measure of irrationality, positing either a blindness to their own affects and feelings coupled with a smarmy form of insincerity.

[20] Cf,. also Anderson and Lepore's claim that slur use is compatible with harboring "no negative opinions towards" the target group (2013).

A comparison with racially coded dogwhistles sheds light on this. Dogwhistles are communicative devices that enable speakers to send cloaked messages. Racially-coded dogwhistles are expressions like *inner city* that can be used to function to communicate racist attitudes. Unlike racial slurs which are universally recognized as blatant tool of derogation, words commonly used to dogwhistle have basic uses that are completely neutral, as in *The inner city kids need state-supported lunches*. Relatedly, they don't wear their racism on their sleeves. When used as dogwhistles, the racist message is somehow hidden – available only to those in the know, or at least thinly veiled. Take, for instance, the racist dogwhistle *inner city* in this selection from Donald Trump's *Time to Get Tough: Making American #1 Again*:

> If we keep on this path, if we reelect Barack Obama, the American we leave our kids and grandkids won't look like the America we were blessed to grow up in. The American Dream will be in hock. The shining city on the hill will start to look like an inner-city wreck.

On influential analyses by Saul and Khoo, dogwhistles do not have their hidden meaning as part of their semantic content.[21] *Inner city* is synonymous with *densely populated urban area*; its meaning is devoid of anything about race. Trump uses *inner city* to *signal* race because he knows something about his hearers beliefs: that the inner city "brings to mind poor, crime-ridden, African-American neighborhoods" (Khoo 2017: 34). Thus, while saying only, 'innocently', that a second Obama term will result in an America that looks like a densely populated urban area, he messages that it will result in "an America dominated by poor, lazy, and criminal African Americans" (Khoo 2017: 34).

Dogwhistlers often accompany their utterances with what Saul has aptly called *figleaves*, addendums to the utterance that function as a way to plead innocent. When someone calls Trump's remarks out as racist, he might say *I was talking about the troubles in cities. Anyway, I love Black people!* Read: no racism here. Literal figleaves, as on Michaelangelo's David, provide cover for artists whose artworks would otherwise be viewed as overly sexual. In similar fashion, racist figleaves provide some modicum of cover for dogwhistlers' racism (Saul 2024: chapter 1). Notably, the figleaf of avowing love for Black people isn't incompatible with use of *inner city*.

Now return to [14]. After saying *You are a f****t*, the bigot follows up with *I love f****ts*! One deeply derogates with the slur and in the next moment professes love. The latter statement wreaks of both linguistic incoherence and insincerity, as do each of the follow-up remarks in [10], [11], [13]. They make the most sense construed as speakers desperately, yet ludicrously, grabbing for figleaves. Of course, the follow ups cannot provide cover, but the practice of grabbing a figleaf has become so normalized, it is easy to imagine someone thinking it could. Unlike *Anyway, I love Black people!* as an addendum to the dogwhistle,

[21] Cf., Saul (2018) and Khoo (2017).

an utterance neither linguistically incoherent nor infelicitous, professing love for the group while using the slur is blatantly both. All in all, it's deeply misguided to pass these utterances off as linguistically nondeviant. Therefore, none provide support for III. Our dialectic reveals that this Specificity Problem leaves Identity Expressivism uncompromised.

4 Conclusion: Looking forward

In *Inflammatory Language*, Stojnić and Lepore boldly widen the purview of phenomena for theories of slurs to explain. The widening, together with their commitment to a single mechanism view, impels them to advance the remarkable view that slurring words themselves are not the ultimate source of slurs' pejorative sting. This piece has been dominantly defensive, aimed at showing that even if we grant the centrality and importance of mere orthographic and acoustic resemblance cases, their negative thesis does not stand. The first half demonstrated that their Single Mechanism Necessity Assumption rules out multiple mechanism theories. The second half demonstrated that their Specificity Problem does not undermine Identity Expressivism. However, Stojnić and Lepore's widening does an important service in bringing to the forefront deep and critical questions: What is the proper scope of a theory of slurs – what phenomena must it explain? What phenomena, if any, should take priority? What starting assumptions shall we make? What are the proper methodological principles guiding our investigations? Currently, the literature on slurs is disunified. We need a reckoning on the full and diverse range of issues before us and a systematic meta-analysis of the most pressing linguistic, psychological, social, and moral problems comprehensive theories of slurs must address.

References

Anderson, L. and Lepore, E. 2013. "Slurring Words." *Noûs* 47: 25–48.

Bach, K. 2018 "Loaded Words: On the Semantics and Pragmatics of Slurs." In Sosa (ed.). *Bad Words*. Oxford: Oxford University Press, 60–76.

Camp, E. 2013. "Slurring Perspectives" *Analytic Philosophy* 54 (3): 330–349.

Camp, E. 2018. "A Dual Act Analysis of Slurs." In Sosa (ed.). *Bad Words*. Oxford: Oxford University Press, 29–59.

Davis, C. and McCready, E. 2020. "The Instability of Slurs." *Grazer Philosophische Studien* 93 (3): 63–85.

Diaz-Legaspe, Liu, and Stainton. 2019. "Slurs and Register: A Case Study in Meaning Pluralism." *Mind and Language* 35 (2): 156–182.

Ekman, P. and Friesan, W. 1986. "A New Pan-Cultural Facial Expression of Emotion." *Motivation and Emotion* 10 (2): 159–168.

Ekman, P. and Heider, K. 1988. "The Universality of Contempt Expression: A Replication" *Motivation and Emotion* 12: 303–308.

Ekman, P. 1994. "Strong Evidence for Universals in Facial Expression: A Reply to Russell's Mistaken Critique." *Psychological Bulletin* 115: 268–287.

Fasoli, F. A. Maass, and A. Carnaghi. 2015. "Labeling and Discrimination: Do Homophobic Epithets Undermine Fair Distribution of Resources?" *British Journal of Social Psychology* 54 (2): 383–393.

Fischer, A. & Roseman, I. 2007. "Beat Them or Ban Them: The Characteristics and Social Functions of Anger and Contempt." *Journal of Personality and Social Psychology* 93: 103–115.

Hom, C. 2008. "The Semantics of Racial Epithets" *Journal of Philosophy* 105 (8): 416–440.

Hom, C. and May, R. 2013. "Moral and Semantic Innocence." *Analytic Philosophy* 54 (3): 293–313.

Hom, C. and May, R. 2015. "Pejoratives as Fiction." In Sosa (ed.). *Bad Words*. Oxford: Oxford University Press, 108–131.

Jeshion, R. 2013a. "Slurs and Stereotypes." *Analytic Philosophy* 54 (3): 314–329.

Jeshion, R. 2013b. "Expressivism and the Offensiveness of Slurs." *Philosophical Perspectives* 27: 231–259.

Jeshion, R. 2016. "Slur Creation, Bigotry Formation: The Power of Expressivism" *Phenomenology and Mind* 1: 130–139.

Jeshion, R. 2017. "Loaded Words, Expressive Words: Assessing Two Semantic Frameworks for Slurs" *Croatian Journal of Philosophy* 50: 111–130.

Jeshion, R. 2018. "Slurs, Dehumanization, and the Expression of Contempt." In Sosa (ed.). *Bad Words*. Oxford: Oxford University Press, 77–107.

Jeshion, R. 2021. "Varieties of Pejoratives." In Khoo and Sterkin (eds.). *The Routledge Handbook of Social and Political Philosophy of Language*. Routledge Press, 211–231.

Jeshion, R. 2025. "Slurs, Articulations, and Inflammatory Language" forthcoming. In Sosa and Lepore (eds.). *Oxford Studies in Philosophy of Language*, volume 4.

Jorgensen Bolinger, R. 2017. "The Pragmatics of Slurs." *Noûs* 51 (3): 439–462.

Khoo, J. 2017. "Code Words in Political Discourse." *Philosophical Topics* 45 (2): 33–64.

Lepore, E. and Stone, M. 2014. *Imagination and Convention*. Oxford: Oxford University Press.

Marques, T. and Garcia-Carpintero, M. 2020. "Really Expressive Presuppositions and How to Block Them." *Grazer Philosophische Studien* 97 (1): 138–158.

Mason, M. 2003. "Contempt as a Moral Attitude." *Ethics* 113: 234–272.

Mason, M. 2016. "Contempt at the Limits of Reactivity." In Mason, ed., *The Moral Psychology of Contempt*. London: Rowman & Littlefield Publishers, 173–192.

Nunberg, G. 2018. "The Social Life of Slurs." In Fogel, Harris, & Moss (eds.). *New Work on Speech Acts*. Oxford University Press, 237–295.

Potts, C. 2005. *The Logic of Conventional Implicatures*. Oxford: Oxford University Press.

Potts, C. 2007. "The Expressive Dimension." *Theoretical Linguistics* 33 (2): 165–197.

Rappaport, J. 2019. "Communicating With Slurs." *The Philosophical Quarterly* 69: 795–816.

Richard, M. 2008. *When Truth Gives Out*. Oxford: Oxford University Press.

Rinner, S. and Hieke, A. 2021. "Slurs Under Quotation." *Philosophical Studies* 1483–1494.

Roseman, I. 2018. "Rejecting the Unworthy: The Causes, Components, and Consequences of Contempt." In Mason (ed.). *The Moral Psychology of Contempt*, 107–130.

Saul, J. 2018. "Dogwhistles, Political Manipulation, and Philosophy of Language." In Fogal, D., Harris, D., Moss, M. (eds.). *New Work on Speech Acts*. Oxford University Press, 360–383.

Saul, J. 2024. *Dogwhistles and Figleaves: How Manipulative Language Spreads Racism and Falsehood*. Oxford: Oxford University Press.

Scherer, K. 2003. "Vocal Communication of Emotion: A Review of Research Paradigms." *Speech Communication* 40: 227–256.

Scherer, K. 2013. "Emotion in Action, Interaction, Music, and Speech." In Arbib (ed.). *Language, Music, and the Brain: A Mysterious Relationship*. MIT Press, 107–139.

Schlenker, P. 2007. "Expressive Presuppositions." *Theoretical Linguistics* 33 (2): 237–245.

Stillman, R. 2021. "Slurs as Ballistic Speech." *Synthese* 199: 6827–6843.

Stojnić, U. and Lepore, U. 2025. *Inflammatory Language*, final manuscript version.

Tangney, J. Stuewig, J. and Mashek, D. 2007. "Moral Emotions and Moral Behavior." *Annual Review of Psychology* 58 (10): 345–372.

Croatian Journal of Philosophy
Vol. XXV, No. 75, 2025
https://doi.org/10.52685/cjp.25.75.3
Received: September 4, 2025
Accepted: November 14, 2025

Inflammatory Content: Reply to Stojnić and Lepore's Inflammatory Language[1]

CHRISTOPHER HOM*
Texas Tech University, Lubbock, USA

In their Inflammatory Language*, Stojnić and Lepore present four major criticisms of content-based views of pejorative language: 1) the Projection Argument; 2) the Hyperprojection Argument; 3) the Specificity Argument; and 4) the Reclamation Argument. This paper argues that a content-based view can adequately respond to each of these criticisms. The paper goes on to consider their positive view, the Articulation Account, and argues that it suffers from being both under-specified and overly ambitious. Even when the view is plausibly precisified as a functional role theory of articulations, a serious dilemma arises: focusing on the sound or shape of the articulation is problematic when considering counterexamples like the racist use of 'Monday' as code for the N-word, and focusing on the functional role of the articulation is problematic when considering how little functional similarity there is between the Mandarin demonstrative term ('那个') and the N-word. The paper also presents the applications of two external criticisms deriving from the Identity Thesis and the Framework Fallacy and concludes that the Articulation Account falls short of being a leading contender in the analytic space of views for pejorative language.*

Keywords: Slurs; pejoratives; hate speech; philosophy of language.

[1] This paper indirectly mentions pejorative terms like the N-word with either 'the N-word' or 'n*' depending on which is most grammatically appropriate.

* Thanks to David Boylan, Ray Buchanan, Robert May, Gary Ostertag, and Jeremy Schwartz for their helpful discussion, and to the organizers of the Philosophy of Linguistics and Language Conference at IUC Dubrovnik for giving me the opportunity to present an earlier version of this paper. I am also indebted to Una Stojnić and Ernest Lepore for the precision and care of their reconstruction and criticism of my views, and I hope the following reciprocates in that generous spirit of inquiry.

What white people have to do is try to find out in their own hearts why it was necessary to have a n* in the first place. Because I am not a n*, I am a man! But if you think I'm a n*, it means you need him. And the question the white population of this country has got to ask itself … If I am not the n* here, and you the white people invented him, then you've got to find out why. And the future of the country depends on that, whether or not it's able to ask [itself] that question.

-James Baldwin, interview with Kenneth Clark, 1963

Introduction

In their *Inflammatory Language*, Stojnić and Lepore present four major arguments against content-based views of pejorative language which are presented in Section 1. In Section 2, I argue that a content-based view can adequately respond to each of these criticisms. Section 3 goes on to consider their positive view, the *Articulation Account*, and argues that it suffers from being both under-specified and overly ambitious.

1. Arguments against content-based views

Stojnić and Lepore (2025) offer four major arguments against content-based views of pejoratives:[2]

1. The Projection Argument
2. The Hyperprojection Argument
3. The Specificity Argument
4. The Reclamation Argument

Understanding the Projection Argument requires understanding the phenomenon of projectivity where the offensiveness of a pejorative term appears to scope out from not just logical and intensional operators like negation and attitude report verbs but also from questions, event quantifiers, tense operators, and fictional contexts.[3] Consider the following examples:

(1) Negation: Baldwin is not an n*.
(2) Attitude report: Mary believes that Baldwin is an n*.
(3) Question: Is Baldwin an n*?
(4) Event Quantification: Every time the firm hires an n* Mary complains.

[2] Content-based theories explain pejorative words like racial slurs primarily through their semantic contents. Prominent examples include Hom (2008), Hom and May (2013, 2018), and Neufeld (2019).

[3] Ordinary predicates do not typically scope out of embedded environments. Consider sentences like: "Baldwin is not a Ukrainian", "Mary believes that Baldwin is a Ukrainian", and "Is Baldwin a Ukrainian?". When sincerely uttered, the speaker takes no positive stand on whether Baldwin is a Ukrainian. In the negation case, the speaker even explicitly rejects it. So there is no sense in which an attitude toward Baldwin being a Ukrainian shines through the embedded context. In each case, the particular linguistic operator (negation, attitude report verb, and interrogative) completely seals off the semantic content of the predicate, leaving nothing further that the speaker is committed to.

(5) Tense Operators: That n* Baldwin was late for work yesterday.

(6) Fiction: "Tell me this one thing. How much is a n* supposed to take?" (Morrison 1987: 289)[4]

The Projection Argument holds that since offensive potential appears to scope out of all of these embedded environments, and semantic content does not, the offensive potential of a pejorative term like the N-word is not identical to its semantic content.[5] For example, in (2), the offensiveness of the N-word is not completely sealed in the attitude report and scopes out to indicate something about the racist attitude of the speaker. Contrast that with "Mary believes that Baldwin is a Ukrainian" where the embedded clause is completely sealed and does not allow any scoping out of attitudes.

Hyperprojection is the observation that projection of offensive potential extends to quotation and phonological variants (what Stojnić and Lepore call *inheritance cases*):[6]

(7) Quotation: "You can really only be destroyed by believing that you really are what the white world calls a "n*"."[7]

(8) Phonological variation: the Mandarin demonstrative term ('那个'), which acoustically resembles the N-word, and corresponds to the English filler word 'uh'.[8]

The Hyperprojection Argument is simply the Projection Argument augmented with cases of quotation and phonological variation.[9] Since offensive potential appears to scope out of quotation and through mere phonological similarity, and semantic content does not, the offensive potential of a pejorative term like the N-word is not identical to its semantic content.

The third major argument is the Specificity Argument which takes the form of an apparent dilemma facing content theories of pejoratives that Stojnić and Lepore attribute directly to Hornsby (2001):

> Is it possible, for every derogatory word, to spell out the faulty consequences to which anyone who uses it is committed? If a coarse articulation of the attitudes of those who endorse uses of a derogatory word is attempted, then

[4] The N-word fully articulated in the original text.

[5] The argument has been well-documented: Cruse (1986), Kaplan (1999), Potts (2005, 2007), Hom (2008, 2010, 2012, 2020), Richard (2008), Anderson and Lepore (2013), Camp (2013) Hom and May (2013, 2025), Cepollaro (2015), and Langton (2018).

[6] See Stojnić and Lepore (2025: 74). The expansion of the projection argument to include cases of quotation and inheritance are also well-documented in the literature; see Hom (2008, 2020) and Hom and May (2025).

[7] Baldwin (1963), N-word fully articulated in original text. Note that this is a case of double quotation.

[8] See Fadel (2020).

[9] As Stojnić and Lepore (2025: 75) notes, inheritance cases are "particularly puzzling for content-based accounts: the offensive potential is inherited even though no slur is either used or mentioned".

> differences between different derogatory words will be ignored. But if a fine-grained articulation is attempted, it will not be credible that what we spell out are regularly consequences accruing to the commitments of speakers who apply the word.

Under the first horn of the dilemma is a coarse-grained specification of the negative content of a pejorative term. The problem is that there are distinct slurs that target the same group that seem to have different potential to offend. The assumption seems to be that what the content theorist must say about the N-word holds for all other racial pejoratives for African Americans, yet there are some pejoratives for African Americans that do not reach the maximal offensive potential of the N-word.

Under the second horn of the dilemma is a fine-grained specification of the negative content. The problem is that a speaker can disavow any particular element of the proposed, bigoted content and yet competently speak about targeted members of the slur. Consider the example of the reformed bigot from Stojnić and Lepore:

(9) I have nothing but respect and admiration for S*'s.; S*'s don't deserve negative moral evaluation because of their group membership.

Such a speaker "would not be manifesting linguistic incompetence, even if the choice of words is insensitive (and retains its pejorative punch)" (Stojnić and Lepore 2025: 20). My interpretation of (9) is that such a speaker simply *does* manifest linguistic incompetence but let us consider a weaker version of the problem. Consider an *un*reformed bigoted speaker who utters:

(10) I hate S*'s but not because they have the common stereotypical property, P_1; S*'s are despicable for other reasons.

The unreformed bigoted speaker does not seem to manifest linguistic incompetence even when they explicitly reject a property that the content-theorist attributes to the semantic content of the pejorative, S*. The assumption seems to be that linguistic competence with a term requires the speaker to know each and every component of the term's semantic content.

The problem under the second horn of the dilemma is a variant of Moore's Open Question Argument in metaethics. The unreformed bigot seems to be able to felicitously and competently ask questions like:

(11) I know this g is not P_1, but are they still an s*?

If this is a felicitous question for a competent, unreformed bigot (where members of group g are targeted by a slur s* that encodes at least the stereotypical property P_1), then it does not seem like s* encodes P_1 as part of its semantic content. The open question generalizes to every property component that a content theorist might plausibly assign to any pejorative slur term, and so the content view is supposedly undermined.

The fourth major argument that Stojnić and Lepore present against content theories of pejoratives is the Reclamation Argument. Reclamation is the phenomenon whereby in-group members repurpose a pejorative term previously used by out-group members to target them. The repurposing is typically political in nature, e.g. signaling mutual allegiance, building camaraderie, defusing the pejorative potential of the term by taking control away from oppressive out-group members, etc. Their negative argument considers two specific content theories: a semantic, ambiguity account and a pragmatic, echoic account. Both theories "offer no explanation as to why it is virtually impossible for out-group speakers to eliminate ambiguity concerning their attitudes regardless of how favorable the context might be" (Stojnić and Lepore 2025: 78). Assuming that there are no other alternative explanations for reclamation for content-based accounts, Stojnić and Lepore conclude that "[t]ogether with the problems raised by hyper-projectivity, inheritance, and the specificity challenge, they motivate a search for a non-content based account" (2025: 79). In the next section, I argue that such a conclusion is certainly premature, if not ill-guided.

2. Responses to Stojnić and Lepore

With regard to Stojnić and Lepore's criticisms of older responses to the phenomenon of hyperprojection, their points are well-taken and were previously addressed in the literature as the *Predicative Response* (Hom 2020). According to this response, intuitions of projection and hyperprojection occur because "when speakers predicate with the derogatory content of [s*] (in almost any context), they do two things that are often offensive: (1) they force hearers to entertain a degenerate way of classifying the world, and (2) they signal that they themselves approve of this classificatory scheme as normatively appropriate" (Hom 2020: 299) I will not rehearse the entirety of Hom (2020) but let me remind the audience of the two core tenets of *Forced Imagery Triggering* (FIT) and the *Conversational Implicature of Predication* (CIP) that compose the Predicative Response:

> The response rests on properly distinguishing between assertion and predication. While assertion is (roughly) the speech act of putting forward a proposition and endorsing it as true, usually with a sentence that has assertoric mood, predication can be conceived as the cognitive application of a property to an object. The application of a property to an object requires the proper sorting of the object relative to the rule that is given by the property. For example, to apply the predicate "is green" to an object *o* is to sort *o* relative to green things and nongreen things. This is to take seriously the idea that propositional content is the object of the attitudes and that propositional content stands for what we cognitively entertain in understanding a sentence. So if the conceptual content associated with a slur is deeply vile or unjust, it makes sense that hearers are disturbed when they predicate with its content. To sort the world according to a deeply racist concept, for example, is to think of the world as the racist does. For nonracists, that way

> of looking at the world is flawed and disturbing. *Even when a slurring sentence is negated, conditionalized, and so on, predication itself is not undone.* Such a predication forces on the hearer a particular way of thinking that is offensive, and this is one part of the explanation that the derogatory content of a slur appears to scope out when embedded. Call this the *Forced Imagery Triggering* (FIT) account. (Hom 2020: 298, emphasis added)
>
> In predicating with the detailed and disturbing content of [s*], hearers who don't share the ideology that supports [s*] are typically offended in having to view the world this way. In addition, the speaker is also pragmatically implicating her general endorsement that [s*] encodes an appropriate way of dividing the world. This higher order, general endorsement is also offensive to hearers who do not share in the ideology that supports this conceptual scheme. Call this the *Conversational Implicature of Predication* (CIP) account. (Hom 2020: 299)

The Predicative Response to the Hyperprojectivity Argument says that predicating the semantic content of pejoratives can be doubly offensive to hearers: a) because the content is so disagreeable to anti-bigots; and b) because predicating with this content is to endorse this way sorting the world. This offense can be generated *whenever* the pejorative is predicated, and so this includes all of the embedding cases.

Let me discuss an additional aspect that contributes to the hyperprojection worry. Because racial pejorative terms like the N-word are socially taboo, and directly asserting them can incur steep social costs, this motivates the recognition of a new kind of speech act that I call *linguistic ventriloquism*. This occurs when a speaker uses language in a way that is technically defeasible but with the intention of triggering the offensive content of the language. Triggering offensive content is sometimes legitimate (e.g., a pedagogical context where the instructor deems it necessary for understanding the nature of bigotry, or a case brief where a judge is writing about the complexities of hate speech in their particular ruling) but in a deeply oppressive society, the act is more often for the purpose of indirectly denigrating the target members of the audience with plausible deniability. In this context, an out-group speaker may also be seeking to vicariously experience uttering a racial pejorative (either as reclaimed or non-reclaimed). For some, there can be a perverse kind of linguistic thrill in uttering the N-word and pretending to publicly violate social taboos, either as a racist or as an out-group member using the reclaimed pejorative. Because bigotry so thoroughly inhabits contemporary society, it is difficult to escape the potential, nearby interpretation that any embedded expression of the N-word by a non-Black speaker is either straightforwardly racist or a form of linguistic ventriloquism. Examples include using the synonym for 'stingy' that sounds like the N-word at a budget meeting in front of a mostly African American audience[10], vocalizing the Mandarin demonstrative term ('那个'), which acoustically resembles the N-word, singing the fully articulated N-word on stage at a concert[11], and using the word

[10] See Hom (2008: footnote 17).

[11] This occurred at a Kendrick Lamar concert in 2018. See BBC News (2018).

'Monday' as code for the N-word.[12] Some of these examples may be benign though the last example is clearly *racist* linguistic ventriloquism. In some cases, there are potential off-setting contextual factors, e.g. the use/mention distinction, the pedagogical goals involved in such utterances, etc. But there are also very nearby racist interpretations that too easily suggest that such a speaker, especially one who vocalizes their point with more than sufficient frequency and intensity, is performing the act of linguistic ventriloquism to denigrate African Americans through an incredibly thin linguistic veil. This is what makes the classroom examples difficult to assess, and why the pedagogical decision procedures are so delicate.[13] Together with the Predicative Response, there are good reasons to think that a pragmatic explanation for hyperprojectivity is successful for content-based views.

Turning to the Specificity Argument, recall it was presented in the form of a dilemma. On the first horn, if the content for a racial pejorative term is too course-grained (e.g. *despicable for being a g*), then it will be applicable for different pejoratives for the same group where those pejoratives vary in their offensive potential. On the second horn, if the content for a racial pejorative is too fine-grained, then it faces a form of the Open Question argument. The content theorist has the tools to saw off both horns. For the first horn, it must be recognized that a certain level of theoretical abstraction is not only helpful but required. If we look too closely at the facts on the ground, then what we find are all kinds of lunatic ideologies that motivate all kinds of lunatic linguistic practices. I suppose that it might be an interesting and perhaps even helpful socio-linguistic project to provide a complete taxonomy of the various slurs for African Americans along with their underlying nuances, histories, and ideological influences. However, this kind of empirical work only minimally engages with higher-order conceptual and philosophical questions of meaning. For the second horn, the Open Question argument fails to acknowledge that the externalist nature of the content of a pejorative divorces linguistic competence from conceptual knowledge. Just as speakers could competently use the word 'water' without knowing almost anything about the chemical makeup of its referent, speakers can competently use a pejorative like the N-word without knowing almost anything about the social or normative composition of its content.[14] On either horn of the dilemma, content-based views have powerful responses and so the Specificity Argument is defused.

[12] See Zimmer (2012).

[13] Collaborative work on this issue is forthcoming with Robert May and Brian Soucek.

[14] There are obviously significant differences between natural kind terms and racial pejorative terms, and so the extension of semantic externalism from the former to the latter requires a significant argument. Such an argument would extend beyond the scope of the current project but see Hom (2008) for at least initial reasons that favor this extension. Thanks to Ray Buchanan for helpful discussion on this point.

The Reclamation Argument against content-based views is puzzlingly weak as it relies on the mistaken assumption that the only semantic content view of reclamation relies on ambiguity, and the only pragmatic content view of reclamation is an echoic one.[15] Hom (2020) acknowledges a fundamental worry with ambiguity accounts of reclamation; namely that they risk making the act of reclamation vacuous. Instead, a different kind of pragmatic view of reclamation is offered; one that is made available by the semantic content expressed by pejoratives:[16]

> Appropriation should be seen as an act of semantic protest to the wide endorsement of the ideology that supports the appropriated slur. A speaker S1 demonstrates that she stands against the ideology by predicating the slur [s]* to another discourse participant S2 where S1 and S2 both belong to group g. The act of appropriation puts oneself and one's allies up as targets of the slur. Appropriately understood as not self-hating or a poor attempt at humor, the speaker of the appropriated slur demonstrates that she stands against the ideological scheme that provides the rule for sorting according to the slur. This is one way of cancelling wide endorsement. Initially, self-predication is only coherent as either an act of self-hatred, false consciousness, or bad humor. So hearers will search for alternative explanations for a cooperative speaker self-inflicting such disturbing content—for example, for some political purpose. (Hom 2020: 302)

The primary thrust of Stojnić and Lepore's criticism rests on the lack of explanation for why reclamation for the N-word is closed off to out-group members. On the pragmatic account presented in Hom (2020), the answer is that for out-group speakers, the affirmative alternative explanations for their pejorative utterances are quite distant because of the prevalence of bigotry (and the fact that pejoratives encode contents that express that bigotry), and because the speech act of linguistic ventriloquism is so correspondingly close by. Note the simplicity of the pragmatic explanation given the content semantically encoded by the pejorative. It also strikes me that the strength of the barrier for out-group members to make use of reclaimed pejoratives is somewhat exaggerated. For the word 'gay', the gay community intentionally dismantled barriers for its use so that non-gay speakers could use the reclaimed word. It has also been reported to me that certain out-group members (members of the Latino community) do felicitously make use of the reclaimed N-word among African American discourse participants.[17] Because members of these particular Latino communities commonly experience the bigotry and the injustice associated with the unreclaimed N-word, the pragmatic explanation is more closely available, and the possibility of linguistic ventriloquism is more distant. With

[15] See Stojnić and Lepore (2025: 78).

[16] The pragmatic move is actually consistent with semantic ambiguity, but I will not argue for this here.

[17] Thanks to Angelina Alvarez-Carrera, Luis Arbelaez, and Jordan Botello for their helpful discussion on this topic.

this plausible pragmatic explanation, content-based views of pejorative terms are shielded from the Reclamation Argument.

3. Assessing the articulation account

With content-based views defended from the Stojnić and Lepore criticisms, we turn to evaluate their positive theory called the *Articulation Account*. An articulation is a sound or shape that is tokened as a physical event. Often, there is overlap between the tokening of an articulation and the tokening of a linguistic expression. When I utter the word, 'dog' (or write it down on a piece of paper), I token a particular articulation that is also a tokening of a word. Someone else can also token that articulation but in a different volume or pitch (or in cursive rather than block print). Articulations can also be produced by accident without any linguistic intention. For example, if the wind blows through a spinning windmill above, the soundwaves can combine in such a way that create the auditory articulation of 'dog' for hearers below. The view holds that articulations that are standardly associated with pejorative words are the primary vehicles of analysis, not the words themselves, and these articulations automatically trigger negative associations. These associations explain inheritance and reclamation, as well as the other data surrounding pejoratives. In their own words: "Our thesis is … that certain articulations automatically trigger negative associations, which come to be associated with these articulations through a web of complex causal, socio-historical, cultural, and psychological factors" (Stojnić and Lepore 2025: 117).

As initially stated, the thesis of the Articulation Account faces two initial problems. The first problem is *valence variation*, or the fact that articulations can have associations that are positive, negative, or neutral. Consider the racist articulation of slurs in conversations with exclusively racist participants. Their associations with this language are not negative (or at least not negative to the extent that they are for anti-racists). Racists think that such words are appropriate and fitting for their targets. It is part of their natural worldview (perhaps even one ordained by God) so some if not all of their associations will be non-negative – the racist is happy with racist associations.[18] Another case of valence variation occurs in contexts of reclamation. Such a context exclusively involves members of the in-group who are targeted by the racist use of the term, but the articulation has *positive* associations of solidarity, alliance, collective resistance, mutual support, etc.[19] Finally, there are articulations that are similar to racial pejoratives and yet invoke no negative or pejorative association. For example, consider the

[18] The example suggests something like an assessment-sensitive analysis given by MacFarlane (2014). Thanks to Ray Buchanan for pointing out this connection.

[19] Oddly, these cases are explicitly acknowledged. See Stojnić and Lepore (2025: 78 and 120).

C-word that is a slur targeting Chinese people. The words 'chin', 'sink', 'ching', 'ink', etc. display articulatory similarity and yet have no negative association whatsoever.[20] The case can also be reversed with words that display no articulatory similarity and yet have the same negative association, e.g., the example of racists who use the word 'Monday' as code for the N-word. Notice crucially that 'Monday' does not plausibly encode the sound or shape of the N-word but the N-word itself. This undercuts the Stojnić and Lepore position that articulations are the primary vehicles of analysis over words.

The second initial problem for the thesis of the Articulation Account as initially stated is *automaticity*. Not all the associations with pejorative articulations are automatic. Some can be conceptual and hence inferential. It even appears as if there is an internal contradiction in Stojnić and Lepore's statement of their view. Contrast the statement of their thesis above with what they say here: "*This is not to say, of course, that tokening an articulation automatically causes offense*. What's automatically triggered is the pejorative effect—an open-ended cluster of cognitive/affective associations". (Stojnić and Lepore 2025: 111, emphasis added)

The initial problems raise the question of what exactly is being automatically triggered by an articulation? The full thesis refers to a complex web of 'socio-historical, cultural, and psychological factors'. Their most definitive elaboration of this point occurs as a continuation of the previous quote:

> Triggering these associations can, in certain circumstances, for certain individuals, cause offense, and an agent can intentionally token the slur in order to weaponize this effect to cause offense. But, we've seen, it can equally be exploited for other purposes—to create a particular kind of an effect, pedagogical, artistic, or other; and it can, in turn, cause a range of other emotions and reactions in the audience, besides offense, depending on a variety of factors, including the circumstances of tokening, the speaker's intentions and audiences' recognition of those intentions, and the audiences' familiarity with, and situatedness within, the socio-cultural and historical context that grounds the relevant associations, and their individual sensibilities and proclivities. In all these cases, it is because the articulations give rise to these associations that they can be used and abused to create these various effects; and it is these articulations, not slurs, that trigger the associations. (Stojnić and Lepore 2025: 111–112)

Taking a step back, note that aside from distinguishing articulations from words, there is most certainly widespread agreement that complex, socio-historical, cultural, and psychological facts determine the negative effects of pejorative language. But until we are told what the relevant facts are and what the relevant associations are, it is not clear what we have actually gained. To assess the theory, we have to be told a lot more about what the historical contexts are and how they ground these associations.

[20] See Jeshion (2025) and Ostertag (forthcoming) for many more examples in the same vein.

With the backdrop of these concerns, I interpret their theory as holding that articulations play certain *functional roles*. Articulations are associated with complex functions that pair input domains with output results where input domains include a wide array of socio-historical, cultural, and psychological factors including the initial beliefs, dispositions and values of the conversational participants, and where output results include updated beliefs, dispositions, and emotional reactions which can be negative, positive, or neutral.

To help clarify, let us refine their thesis to avoid both of the initial worries of valence variation and automaticity, and make explicit the functional role aspect of the view. Following their lead, let an *articulation* (an event) be any sound or shape (commonly associated with the tokening of words but not essentially so). Let an *association* (a relation) be a function that takes a complex domain as input (including a speaker's beliefs, values, history, culture, and various facts about the context) and gives a complex domain as output (including a hearer's beliefs, attitudes, and behavioral dispositions). Consider the following refinement on the Articulation Account:

> *Functional Role Articulationism (FRA)*: Articulations play certain functional roles in instantiating complex association functions whose overall domains include relevant causal, socio-historical, cultural, and psychological factors. In this way, articulations generate certain pejorative effects.

Consider the kinds of detailed associations that must be part of these functions. For an articulation of the straightforwardly racist N-word, the view must specify something like the following association function:

Input	Output
Speaker: member of the out-group	Hearer's Emotional State: offense, anger, fear
Hearer: member of the in-group	Hearer's Belief Update: speaker is a bigot
Speaker beliefs: biased against members of the in-group	Hearer's Disposition: terminate the conversation, explicitly sanction the speaker, punch the speaker
Personal relationship between speaker/hearer: none	
Time: 21st century	
Location: United States	
Non-fictional	

Table 1: Partial Specification of the Functional Profile for the N-word (racist)

Next consider the specification of the reclamatory aspect of the association for the N-word:

Input	Output
Speaker: member of the in-group	Hearer's Emotional State: camaraderie, support, alliance
Hearer: member of the in-group	Hearer's Belief Update: speaker is an ally
Speaker beliefs: non-biased against members of the in-group	Hearer's Disposition: acknowledge the speaker as an ally
Personal relationship between speaker/hearer: friendly	
Time: 21st century	
Location: United States	
Non-fictional	

Table 2: Partial Specification of the Functional Profile for the N-word (reclamatory)

It is these kinds of complex, 'open-ended', functional pairings that make up the individual associations that in turn compose the entire functional role of an articulation. But notice that this initial attempt at the partial specification of the functional role for the N-word is still a massive simplification. The association function must take into consideration individual histories and psychologies.[21] For example, in the straightforwardly racist portion of the function, there is a tremendous amount of variation in the hearer's potential dispositions for action. Consider the different ways such an articulation could play out depending on just the different social roles and power dynamics of the conversation. For example, contrast how a Black employee being slurred with the N-word by their non-Black boss would (or even could) respond with how a wealthy and powerful Black superstar like LeBron James might respond to being slurred with the N-word by a non-Black fan on the court during a game. Individual roles and power dynamics in a context must be incorporated into the function.[22]

3.1 Internal criticisms of the articulation account

Upon careful reflection, the details of the functional specification for pejorative articulations seem impossibly fine-grained. There are many other specific factors that must be accommodated for a full specification

[21] As Stojnić and Lepore note, "members of a target group might carry different, more nuanced associations, reflecting direct experience with a history of discrimination and oppression" (2025: 121).

[22] See Popa-Wyatt and Wyatt (2018) for a detailed consideration of these kinds of conversational kinematics.

of the functional role of an articulation of the N-word. Such factors are made salient by questions like:

- What kind of facial expression does the speaker have?
- What kind of vocal tone did the speaker use?
- Is the context a physically dangerous one for the hearer?
- What are the relative genders of the speaker and hearer?
- What are the relative physical sizes of the speaker and hearer?
- Have the speaker and hearer had previous conversations?
- Is the hearer more/less sensitive to racial slurs?
- Is the hearer more/less inclined to take racial antagonism personally?
- Does the hearer believe that males are generally more racist than females?
- Does the hearer believe that blonde-haired speakers are generally more racist than brunette-haired speakers?
- Has the hearer been having a good/bad day?[23]
- Has the hearer skipped lunch that day making them more/less cranky?
- Does the speaker remind the hearer of a racist colleague?

A difference in any of the above input factors can change the output results of an articulation of the N-word. Furthermore, the list of questions specifying potentially impactful contextual factors seems endless. To offer a complete theory, the function must specify the exhaustive range of causes and effects associated with hearing the sounds or seeing the shapes of the word. Such a detailed specification is very much nontrivial and potentially quite disparate. Associations can be extremely extensive, varied, and haphazard. Their explanations may reside at a very specific psychological level for individual agents. Every quirk, bias, misunderstanding, misconception, background assumption, heuristic, and cognitive habit must be reflected in a complete specification of the association function.[24]

The view is supposed to incorporate every functional variable involving associations having to do with articulations of pejoratives. This is an incredibly wide-ranging metatheoretic assumption. If true, then it becomes unclear what is at stake for such a view. Consider that any dynamic system can be represented functionally, e.g. chemical systems, meteorological systems, digestive systems, raindrops on a window, etc. The mere existential claim that there is a functional specification for a particular set of articulations is trivial, so much so that such a view is not even falsifiable because no meaningful prediction is produced. Perhaps Stojnić and Lepore's claim is not merely existential but also one

[23] See Doris (2002) for evidence from moral psychology that normative behavior is influenced by trivial contextual factors.

[24] Even if such an exhaustive specification is given, it is not at all clear that human behaviors and responses will be accurately predicted. Thanks to Robert May for pointing this out.

that grounds the functional specification in particular facts about society, history, culture, and psychology. To the extent that the grounding is actually specified, the view escapes the triviality of simply making the existential claim. But from the considerations above, such specificity is flagrantly absent.

The immense power attributed to FRA has another negative unintended consequence. Humans make many associations as a result of their basic cognitive architecture. The explanatory domain of FRA seems to include *anything* that might signal human significance, e.g. words, symbols, gestures, vocal intonations, facial expressions, clothing, skin tone, colors, jewelry, religious artifacts, artistic imagery, metaphors, motorcycles, tattoos, facial hair, food, etc. There seems to be no limit to the domain of associations, and hence no limit to the applicability of FRA as a *semiotic* theory of human culture. Perhaps Stojnić and Lepore would welcome such a powerful consequence of their theory but there are reasons to be suspicious.

Focusing on language, there are many articulations of words that have negative associations. Consider 'bald' which has negative associations for men perhaps involving aging and loss of sexual virility and negative associations for women perhaps involving aging and loss of femininity. There is surely a complex, functional role played by articulations of 'bald' that coordinate the complex socio-historical, cultural and psychological facts which ground these gendered and ageist associations. But what distinguishes words like 'bald' from pejoratives? There seems to be no principled difference between the articulation of terms like 'bald' and the articulations of terms like the N-word. Perhaps the negative associations for the N-word are simply far greater so this is a distinction in degree and not kind. But it is easy to construct an expression with far greater negative associations that at least rival the N-word but which itself is not intuitively a pejorative; e.g. 'narcissist Nazi pedophile'.

Other distinctions under the precisification of the Articulation Account also get the wrong intuitive results. The functional complexity of FRA actually *undermines* the ability to explain inheritance cases. Consider again the case of the N-word and the Mandarin demonstrative term ('那个'). The explanation of the offensive potential of the Mandarin term was supposed to be explained by its articulatory similarity to that of the N-word. But consider the overall scope of the functional roles of these two articulations. One has very specific associations relative to English and its speakers, the other has very specific associations relative to Mandarin and its speakers. For the most part, they are linguistically, geographically, historically, culturally, and psychologically distinct. Their articulations play very different functional roles, and so their similarity is actually extremely marginal. There is an incredibly small amount of overlap between these articulations when considering their overall functional roles. In fact, their overlap might occur in just

a handful of contexts—namely the one where the teacher reportedly made the observation that the Mandarin demonstrative sounds similar to the N-word. To put the point another way, of the billions of articulations of the demonstrative generated by Mandarin speakers and of the billions of articulations of the N-word generated by American-English speakers, there is almost no functional overlap whatsoever. Mandarin speakers will have many kinds of associations with the Mandarin demonstrative that they do not have with the N-word, and English speakers will have many kinds of associations with the N-word that they do not have with the Mandarin demonstrative.[25] With such tremendous functional dissimilarity, the articulations of the N-word and the Mandarin demonstrative are actually not very similar at all, and this undermines the explanation of the very data that Stojnić and Lepore prioritize for their view. It is worth noting here the damaging significance of this point for the Articulation Account.[26]

These problems form a dilemma for the Articulation Account. If the focus is on the articulation itself (sound/shape) as an explanation for pejorative potential, then the view faces difficulty from cases where articulatory resemblance is absent but pejorative potential is not (e.g. 'Mondays'), and also from cases where pejorative potential is absent but articulatory resemblance is not (e.g. 'chin', 'sink', the reclaimed N-word). If the focus is on the articulation as satisfying a complex functional role, then the view faces difficulty in explaining inheritance cases (e.g. the functional dissimilarity between the N-word and the Mandarin demonstrative). What this highlights is the overall vagueness of Stojnić and Lepore's presentation of the Articulation Account, and the worry that there is a vicious ambiguity with which the theory is being originally deployed.[27]

[25] It is even worse when you consider the billions of natural, orthographic accidents over time that have produced either kind of articulations (e.g. awkward sneezes, wind blowing through the trees, random sound waves colliding, certain positionings of rocks, stars, or grains of sand, etc.).

[26] A similar argument is made in Ostertag (forthcoming).

[27] Lepore's affinity for the philosophy of Donald Davidson suggests an alternative metatheoretic interpretation of their book project. Recall Davidson's position on metaphor; roughly that there is no metaphoric content, and that metaphors are just like bumps to the head of the hearer in the attempt to cause them to think differently about something (Davidson 1978). Similarly, one might hold that beyond articulations having offensive associations, there is simply nothing further to say about pejorative language. Call the position that any theory of pejorative language is false, *Defeatism*. If this was part of their overall proposal, then the negative chapters (perhaps even together with the failure of FRA) form a kind of inductive argument against any theory of pejorative words. Note that this would make the above defense of content-based views all the more significant. But if defeatism is actually one of their goals, why keep it hidden? They would owe it to readers to explicitly own up to it and to give explicit arguments for it. Without an argument for defeatism, there is little to motivate this potential hidden assumption. Thanks to Ray Buchanan for helpful discussion on this point.

3.2 External criticisms of the articulation account

While the previous section addresses criticisms that are internal to the Articulation Account, there are significant criticisms that are theoretically external to the Articulation Account. The first is the *Identity Thesis* and the Racist Frege Puzzle that results from it. The second is the *Framework Fallacy*. Both problems are developed in the literature but let me present brief versions of each and apply them directly to the Articulation Account.

The Identity Thesis holds that any pejorative term, s*, and its neutral correlate term, g, are semantically equivalent.[28] The result is that the two terms are co-extensive, i.e. they refer to the same set of people. The critical question arises for any theory of pejoratives: does the theory subscribe to the Identity Thesis? Though they offer no explicit confirmation, there is good reason to think that Stojnić and Lepore are committed to it.[29]

There are significant negative implications of the Identity Thesis. For example, it entails that sentences like "all African Americans are n*'s" is literally true. As racist and cringeworthy as this might sound, things are even worse. Since claims like 'all African Americans are African American' are not only true but *necessarily* true, so too is the racist claim, 'all African Americans are n*'s'. The result is that such a racist claim has the same alethic status as mathematical or logical truths. One might have thought that the racist had radically *false* beliefs about the world, but not according to the Identity Thesis. Under this thesis, the racist holds deeper insight into metaphysical reality than the anti-racist who would 'wrongly' deny the racist claim as false.

The Identity Thesis suggests a racist version of Frege's Puzzle. The puzzle asks what explains the cognitive difference between the following sentences as the first is trivial and knowable *a priori* while the second is non-trivial and knowable *a posteriori*:

(12) African Americans = African Americans
(13) African Americans = N*'s'

The Identity Thesis blocks the ability to explain this difference in terms of a semantic difference between the terms flanking the identity sign in (13). Like a direct reference theorist in the original Frege Puzzle, the proponent of the Identity Theory might be tempted to make use of modes of presentation (or ways of thinking) to solve the puzzle.[30] For Stojnić and Lepore, the corresponding move is potentially made by saying that the trivial case has no negative associations (i.e. no negative

[28] See Hom (2008) and Hom and May (2013, 2018, 2025).

[29] Stojnić and Lepore's commitment to their wager argument (2025: 23) strongly suggests this, and Robert May reports in p.c. that Stojnić explicitly confirmed their commitment to the Identity Thesis at her colloquium at U.C. Berkeley in the spring of 2025.

[30] See Sennet and Copp (2017).

modes of presentation) but the non-trivial case does, and this explains their difference in cognitive significance.[31]

The content theorist's response is simply to invert the racist Frege puzzle so that the pejorative occurs in both identity statements:

(12') N*'s = N*'s

(13) African Americans = N*'s

Since articulations of both identity claims would have negative associations, the negative associations can no longer distinguish their cognitive significance. If Stojnić and Lepore attempt to retreat to standard modes of presentation (perhaps generated through associations with the articulations) to distinguish the identity claims, there are already extensive arguments in the literature for why such a move does not look promising.[32]

To hold the Identity Thesis precludes the elegant, simple response to the Racist Frege Puzzle available to the content-based theorist who distinguishes the semantic content of a pejorative from that of its neutral counterpart: *there simply is no puzzle*. For content views, the terms s* and g are *not* semantically identical and hence non-coreferential. So believing that s*'s are s*'s is like believing that unicorns are unicorns (trivially true), while believing that g's are s*s is like believing that white horses are unicorns (false). Because the embedded identity claims express distinct propositions, you should rationally believe the first and disbelieve the second, allowing for an accurate, consistent, non-racist, representation of the world.

The final worry related to the Identity Thesis is what I call the *Baldwin Point* which directly references the epigraph at the start of this paper.[33] James Baldwin was a leading 20th century civil rights activist and acclaimed author on issues of race in America. He was quoted as uttering the following sentence:

(14) I am not a n*. [I am a man]

Holding the Identity Thesis implies that Baldwin, a Black man, spoke *falsely* in making this utterance. He *is* African American, and according to the Identity Thesis, he *is* an n*. Under the Identity Thesis, Baldwin is like the person who does not realize that beech trees are elm trees, or that aluminum is molybdenum. Though Baldwin was a leading, Black

[31] A lot of conceptual work would need to be filled in here, but as shown in the next paragraph, the task is moot.

[32] See Hom and May (2025: 644–648, and fn 22): the guiding principle against such a move is "what Schiffer calls *Frege's Constraint*: "There are distinct modes of presentation m and m' such that x believes y to be such-and-such under m and disbelieves it under m' only if x fails to realize that m and m' are modes of one and the same thing." (1992: 502). The relevant point here is that you cannot rationally believe and disbelieve something at the same time under modes you realize are modes of the same thing."

[33] This point is echoed in Hom and May (2025: 648).

civil rights activist, he was supposedly conceptually confused about his own racial identity.

The Identity Theorist is not without recourse and could appeal to *metalinguistic negation*.[34] An utterance like the following:

(15) I am not *cold*. [I am freezing.]

is not technically false because it has a metalinguistic interpretation whereby the speaker is denying the appropriateness of using the term, not the literal predication of the term's content:

(15_M) 'Cold' does not appropriately apply to me. [I am freezing.]

The follow-up in brackets shows that *being freezing* is consistent with *being cold*, but not consistent with appropriately being called merely 'cold', and so the parallel supposedly holds in (14):

(14_M) The N-word does not appropriately apply to me. [I am a man.]

The denial is of the linguistic application of the N-word to Baldwin and not of the claim that he is African American. *Being a man* is actually consistent with *being an n** (i.e. because according to the Identity Thesis proponent, *being a man* is consistent with *being African American*).[35]

The metalinguistic strategy is not limited to cases of negation. The metalinguistic move is required for any sentence where there is apparently acceptable embedding of a pejorative term:

(16) If you think I'm a n*, then it means you need him.
(17) There are no n*'s.
(18) Institutions that treat Chinese as c*'s are racist.
(19) Every time the department interviews an asshole, George rolls his eyes.

The examples demonstrate that the metalinguistic strategy requires expansion to metalinguistic conditionalization, metalinguistic existential generalization, metalinguistic attitude reporting, metalinguistic quantification over events, etc. Unlike in the application of metalinguistic negation which was motivated in alleviating the apparent contradiction, such motivation is absent in the expanded set of cases, and so the metalinguistic strategy seems ad hoc. A further and more serious problem is that it is not even clear that metalinguistic reinterpretation uniformly works:

(16') If you think the N-word appropriately applies to me, then it means you need the N-word (to appropriately apply to me). (?)
(17') There is no one appropriately called by the N-word.
(18') Institutions that call Chinese people by the C-word are racist (?).[36]

[34] See Horn (1989), Hom (2020), and Stojnić and Lepore (2025: fn 30).

[35] The continuation of the quote in (14) suggests that Baldwin did not agree with the consistency claim between *being an n** and *being human*.

[36] The complaint against such racist institutions does not seem to center around the *linguistic* practices of the institution but rather its *material* practices.

(19') Every time the department interviews someone appropriately called by the term 'asshole', George rolls his eyes. (?)[37]

Let the *Baldwin Point* state that when Baldwin utters (14), he speaks truly. As a Black civil rights leader who was one of the most prominent literary figures in the 20th century, it is completely implausible that Baldwin was either conceptually or linguistically confused about the N-word. It is difficult to overemphasize that this is unassailable, primary data. Baldwin is one of the foremost conceptual and linguistic experts on racial pejoratives, and content-based theorists are on concrete footing in validating his judgements. If the Baldwin Point is true, then the Identity Thesis is false.

Putting the Baldwin Point another way, we can ask what belief gets expressed when Baldwin utters (14)? Does he believe that he is not African American? Does he have a metalinguistic belief about the applicability of the N-word? The positive answers sound completely implausible. On the other hand, there is a simple, elegant, negative answer available to the content theorist: the belief Baldwin expresses with (14) is the denial of a deeply personal, degrading, belief that reflects his experiences as a victim of racial oppression. He expresses that by using the N-word to encode and reflect the racist ideology of his time. The content of the N-word is emblematic of the negative stereotypes and prescriptions that anti-Black racism stands for. The N-word is a *bad* word because it encodes viciously *bad* content, the content of racism that persists in American society from Baldwin's time to this day. As a civil rights leader, Baldwin rejects this attempted targeting by racists with this ideological term. To cling to the Identity Thesis in the face of data like the Baldwin Point is inexplicable.

Let me turn to the last theoretically external concern for the Articulation Account. Start with the basic observation that many things have the capacity to cause offense, e.g. calling someone 'bald', interrupting someone in conversation, using the wrong silverware utensil, refraining from taking off one's shoes in a traditional Japanese household, chewing with your mouth open, insufficiently tipping, facial hair, body odor, etc. There might be a causal theory that explains all of this by appealing to the socio-historical and psychological facts that ground these offensive associations. The worry of theoretical overreach has already been suggested, and I do not believe that the Articulation Account has successfully met their own explanatory challenge. But let me address the status of the challenge for content-based accounts. It is important to emphasize that *content-based accounts do not aim to provide a comprehensive account of the offensive potential of pejorative terms, nor should they*. Why should a semantic account of language

[37] This seems like an implausibly strained reading for (19). Change 'dick' for 'asshole' and suppose that one of the unpleasant candidates is named 'Richard' and then this candidate *would* be appropriately called by the term 'Dick'. Particularly problematic for Stojnić and Lepore is the fact that the articulation of 'dick' is identical in both pejorative and nickname contexts.

provide a complete explanation of how socio-historical, cultural, and psychological factors affect the offense either given or taken by utterances of language? The distinction was explicit between offense and derogation, where derogation is the application or predication of the prescriptive semantic content expressed by pejoratives, and where offense, while overlapping with derogation is obviously distinct. Simply observe that some utterances can be offensive and non-derogatory (e.g. 'bald') while others can be non-offensive but derogatory (e.g. reclaimed utterances of the N-word between Black conversants). To misattribute the task for content theories of explaining offensiveness commits the *Framework Fallacy*: to "commit such a fallacy is to assume one's own view from the start in order to criticize a rival. It is to saddle a rival view with commitments to which it does not actually subscribe."[38] The Articulation Account commits the Framework Fallacy in assuming that offensive associations must be explained by semantic content views of pejoratives. This is precisely what the semantic content view rejects.[39] It does not however reject the possibility that other kinds of theories could help to explain offensive associations with pejoratives (i.e. theories in pragmatics, psychology, sociology, etc.). Because a semantic content view of pejoratives does not explain all of the potential associations with pejorative terms does not mean that the view has failed. It only means that its theoretical domain is limited to concepts and meaning, and does not include the entirety of the perlocutionary effects for such language.

Conclusion

The Articulation Account, intending to add to the diversity of the analytic space of theories for pejoratives, is undermotivated and underspecified. In the attempt to give the most plausible specification for the Articulation Account, insurmountable problems become immediately apparent. At the same time the negative arguments against content-based views can be dispensed with, and the motivations for content-based views are also immediately apparent. In light of this, the Articulation Account falls short of its aspiration to become the leading contender in the analytic space of pejoratives, and given the above considerations, it is unclear whether it is a contender at all.

References

Anderson, L. and Lepore, E. 2011. "Slurring Words." *Noûs* 47 (1): 25–48.

Baldwin, J. 1963. *The Fire Next Time*. New York: Dial Press.

BBC News. 2018. "Kendrick Lamar Stops White Fan Using N-Word on Stage at Concert." *BBC News Service*. May 22. https://www.bbc.com/news/newsbeat-44209141.

[38] See Hom and May (2025: 641).

[39] See Hom (2010: footnote 5; 2012: 397) and Hom and May (2018: 116ff).

Camp, E. 2013. "Slurring Perspectives." *Analytic Philosophy* 54 (3): 330–349.

Cepollaro, B. 2015. "In Defense of a Presuppositional Account of Slurs." *Language Sciences* 52: 36–45.

Cruse, D. A. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.

Davidson, D. 1978. "What Metaphors Mean." *Critical Inquiry* 5 (1): 31–47.

Doris, J. 2002. *Lack of Character: Personality and Moral Behavior*. Cambridge: Cambridge University Press.

Fadel, L. 2020. "Professor Is at Center of Controversy over Chinese Word That Sounded Like Racial Slur." *NPR: All Things Considered*. September 16. https://www.npr.org/2020/ 09/16/913693813/professor-is-at-center-of-controversy-over-chinese-word-that-sounded- like-racial.

Hom, C. 2008. "The Semantics of Racial Epithets." *Journal of Philosophy* 105 (8): 416–440.

Hom, C. 2010. "Pejoratives." *Philosophy Compass* 5 (2): 164–185.

Hom, C. 2012. "A Puzzle About Pejoratives." *Philosophical Studies* 159 (3): 383–405.

Hom, C. 2020. "Slurs, Assertion, and Predication." In S. Goldberg (ed.). *The Oxford Handbook of Assertion*. Oxford: Oxford University Press, 287–305.

Hom, C. and May, R. 2013. "Moral and Semantic Innocence." *Analytic Philosophy* 54 (3): 293–313.

Hom, C. and May, R. 2018. "Pejoratives as Fiction." In D. Sosa (ed.). *Bad Words: Philosophical Perspectives on Slurs*. Oxford: Oxford University Press, 108–131.

Hom, C. and May, R. 2025. "The Metatheoretic Foundation for Racial Epithets." In E. Lepore and U. Stojnić (eds.). *Oxford Handbook of Contemporary Philosophy of Language*. Oxford: Oxford University Press, 636–652.

Horn, L. 1989. *A Natural History of Negation*. Chicago: University of Chicago Press.

Hornsby, J. 2001. "Meaning and uselessness: How to think about derogatory words." In P. French and H. Wettstein (eds.). *Figurative Language, Volume XXV (Midwest Studies in Philosophy)*. Oxford: Wiley-Blackwell, 128–141.

Jeshion, R. 2025. "Slurs, Articulations, and Inflammatory Language." In E. Sosa and E. Lepore (eds.). *Oxford Studies in Philosophy of Language, Vol. 4*. Oxford: Oxford University Press, forthcoming.

Kaplan, D. 1999. "What is Meaning?—Explorations in the Theory of Meaning as Use." Brief Version—Draft no. 1. Unpublished manuscript. UCLA.

Langton, R. 2018. "The Authority of Hate Speech." In J. Gardner, L. Green, and B. Leiter (eds.). *Oxford Studies in Philosophy of Law, Vol. 3*. Oxford: Oxford University Press, 123–152.

MacFarlane, J. 2014. *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Oxford University Press.

Morrison, T. 2007. *Beloved*. London: Vintage Classics.

Neufeld, E. 2019. "An Essentialist Theory of the Meaning of Slurs." *Philosophers' Imprint* 19 (35).

Ostertag, G. Forthcoming. "Critical Notice: *Oh Slur, Where is Thy Sting*?" *Analysis Reviews*.

Popa-Wyatt, M. and Wyatt, J. 2018. "Slurs, Roles and Power." *Philosophical Studies* 175 (11): 2879–2906.

Potts, C. 2005. *The Logic of Conventional Implicatures*. Oxford: Oxford University Press.

Potts, C. 2007. "The Expressive Dimension." *Theoretical Linguistics* 33 (2): 165–198.

Richard, M. 2008. *When Truth Gives Out*. Oxford: Oxford University Press.

Sennet, A. and Copp, D. 2017. "Pejoratives and Ways of Thinking." *Analytic Philosophy* 58 (3): 248–271.

Stojnić, U. and Lepore, E., 2025. *Inflammatory Language: Its Linguistics and Philosophy*. Oxford: Oxford University Press.

Zimmer, B. 2012. "How did 'Monday' become a racist slur?" *Boston Globe*, July 29. https://www.bostonglobe.com/ideas/2012/07/28/how-did-monday-become-racist-slur-how-did-monday-become-racist-slur/Mf4fQEVcXabGKHFaDMZ4NO/story.html.

Croatian Journal of Philosophy
Vol. XXV, No. 75, 2025
https://doi.org/10.52685/cjp.25.75.4
Received: July 28, 2025
Accepted: September 24, 2025

A Defense of Lexical Accounts of Slurs: Comments on Stojnić and Lepore's Inflammatory Language

LOUISE ANTONY
University of Massachusetts, Amherst, USA and Rutgers University, New Brunswick, USA

I defend a lexical account of slurs against criticisms mounted by Stojnić and Lepore, and present positive reasons for preferring a lexical account over the articulation account they put forward. A lexical account, I argue, explains why articulations of a slur give offense: viz., they are recognized as articulations of a particular word. A lexical account also does better than the articulation account in explaining the naive acquisition and use of slurs, facts about mishearing and mispronouncing, the evolution of certain words into or away from status as slurs, and the fact that "reclaimed" slurring words still give offense when used as slurs.

Keywords: Norms; normativity; functions; social roles; social construction; social kinds; convention.

I admire this book enormously; I cannot claim to have fully digested it. I completely agree with Stojnić's and Lepore's criticisms of content views of slurs, partly just because of my background bias toward externalism in semantics. So I'm prepared to treat slurs as simply meaning their denotations, just as neutral counterpart terms do (or at least, to the same extent as they do), and to look to non-semantic factors to explain the phenomena. And I also agree with the critiques they offer of expressivist and other views that appeal to speaker intentions. (I do have sympathy with the idea (Bolinger, Nunberg, Kirk-Gianni) that a speaker's *choice* (or presumed choice) to use a slurring expression carries information about the speaker's intentions, attitudes or affiliations, particularly if there is a neutral counterpart term available but unused, but I think that word choice in general can carry such information, so there is nothing here that is specific to slurs.)

Another reason I have for rejecting both semantic and expressivist views of slurs – and this reason is something not fully discussed by Stojnić and Lepore – is the fact that slurs can be acquired and used *naively*. That is, the terms can be picked up and used by someone who does not know that the term is a slur. A person can acquire a slurring term without knowing the stereotype associated with the term (if there is one), and without knowing even that the term is, can be used to, or is taken to express contempt. Here's an example: a good while back, when my husband and I lived in Raleigh, N. C., the two of us were chatting with our neighbor, a native Southerner. He had just started a career as a realtor, and we asked him how things were going. "Great," he said. "I just made a terrific deal – the guy wanted $n, but I managed to Jew him down." My husband is Jewish; my neighbor knew that. But it was clear from the immediate change in his expression as he spoke that he had not until that moment connected the expression he was using with Jewish people. He apologized profusely and insisted that that was "just what we called it." My husband and I believed him. After all, he and I had both grown up using the expressions "gyp" (which I would have spelled "J-I-P") and "Paddy wagon" without the least awareness that these were slurs against (respectively) the Romani and the Irish peoples.

Here's a slightly different case, but one which also illustrates naïve acquisition: one summer while I was in graduate school, I got a job as a counselor at a day camp in Dorchester, MA, a city that in the 1970's and 80's had seen a great deal of racial strife. The camp was multi-racial and the administrators were trying valiantly to foster inter-racial friendships. My charges were 8-yr.-olds, but they had already picked up some nasty terms. When some of the Black kids got mad at me, they would, with semantic appropriateness, call me a "honky." When some of the white kids got mad at me, however, they would hurl the N-word at me. What I think was going on was that the white kids (and maybe the Black kids, too) had naively mistaken a racially specific slurring term for a general pejorative.

One more – possible – example: Mark Twain's novel *Huckleberry Finn* is written in the voice of its main character, a 13 (or so) year-old boy in antebellum Missouri. The novel is now notorious because of Twain's use of the N-word in rendering the thoughts and speech of its characters. Twain tells his reader, in an explanatory forenote, that:

> a number of dialects are used, to wit: the Missouri negro dialect; the extremest form of the backwoods Southwestern dialect; the ordinary "Pike County" dialect; and four modified varieties of this last. The shadings have not been done in a haphazard fashion, or by guesswork; but painstakingly, and with the trustworthy guidance and support of personal familiarity with these several forms of speech. (Twain 1884)

(Twain's purpose in adding this note is not, apparently, to apologize for his use of the N-word, but to explain the differences in his various characters' patterns of speech to the "many readers" who, he feared, "would

[without the note] suppose that all these characters were trying to talk alike and not succeeding.") If Twain's research can be trusted, the term he used and put into his characters' voices was *the* term for African-Americans at the time and place, and among the populations, in which the novel is set. If that's the case, then Huck would presumably have acquired the term simply as the common name for Black people, the way liberals and "well-bred" bigots of my parents' generation picked up the terms "Negro" and "Colored" as names for Black people.

Indeed, the move from "Negro" to "Black" as the preferred term for Black people in the U.S. supports, I think, the lexical view. Brian Palmer of *Slate* explains how the terms "Negro" and "colored" moved from being acceptable and even preferred, to being strongly disfavored. Palmer marks the turning point to be the historic speech by Stokely Carmichael in Mississippi in 1966, in which he coined the expression "black power." According to Palmer, Carmichael, in his book *Black Power: The Politics of Liberation in America*, "persuasively argued that the term ["Negro"] implied black inferiority."

Palmer continues:

> Prominent black publications like *Ebony* switched from *Negro* to *black* at the end of the decade, and the masses soon followed. According to a 1968 *Newsweek* poll, more than two-thirds of black Americans still preferred *Negro*, but *black* had become the majority preference by 1974. Both the Associated Press and the *New York Times* abandoned *Negro* in the 1970s, and by the mid-1980s, even the most hidebound institutions, like the U.S. Supreme Court, had largely stopped using *Negro*. (Palmer 2010)

This progression is difficult to understand on any semantic or expressivist view. And neither, I think, can the Articulation Account make sense of it, especially since the N-word was, in the dialects of at least many Americans, *an articulation* of the "respectful" word "Negro."

To be clear, the point I am making here is *linguistic*, not moral. I am not saying that the fact that a slur has been acquired naively makes a person who deploys it is blameless. As is recognized by everyone in this debate, a slur can cause pain whatever the intentions of its speaker, and at a certain point in social evolution, persons become responsible for knowing such matters as that certain expressions are slurs. Rather, the point is that even if the meaning of a slur is something derogatory, one *can* use the slur without knowing that. This is a consequence of externalism, and what Putnam called the "social division of linguistic labor:" we pick up terms from the others in our milieu without needing to know what the others know about them. We certainly do not need, in general, to know the etiology of an expression in order to use it; neither do we need to know how and why a slur counts as a slur. Even had I known that "gyp" derived from "gypsy," that wouldn't have told me that the term was a slur, because I didn't know that the term was a derogatory term for the Roma, nor that these people were derogated as cheaters and thieves.

But to return to Stojnić and Lepore's own discussion. Although I agree with their criticisms of semantic and expressivist views, I am satisfied with an approach to slurs that they reject, namely the *lexical* account. What makes a term a slurring term, I submit, is the fact that the term has a certain kind of history. I'm not sure exactly how to characterize this kind of history precisely, but typically, it has these features: it is, first of all, a name for a group that the members of that group did not give themselves, and secondly, the name was given in the context of oppression, discrimination, or domination. It is also typically the case that the term was and is used by individuals who endorse the oppressive relationship or who advocate (or at least would favor) restoring the oppressive relationship. Slurring terms are offensive to those to whom the terms "apply" because the terms are recognized as having this sort of history. (Uses of the terms by those to whom they apply – ingroup slang or "reclaimed" uses – have to be dealt with separately, as everyone acknowledges.) And the offense occurs because of the history of the term, and often regardless of the user's specific intentions or attitudes. The *potency* of a slur has much to do with the extent to which the oppressive condition still exists, or the extent to which there are people who resent or want to undo social progress.

I was convinced of this much by Luvell Anderson and (an earlier person-stage of) Ernest Lepore (Anderson and Lepore 2011: 25–48). Anderson and Lepore went on, however, to say that slurring words were words that had been *prohibited* – rendered taboo – by some set of authorized authorities. Stojnić and Lepore take this element of the view to be central (as indeed, Anderson and Lepore do) and accordingly criticize "Prohibitionism," and raise several important issues for it. But what about the part of the Anderson-Lepore view that simply says that to be a slur is to be a *word* with a certain history and sociology? One can endorse this part without taking the prohibitionism on board. So, apart from the prohibitionism, what's wrong with a lexical account?

Stojnić and Lepore argue that it is not *words per se* that give offense – it is only *articulations* of words that give offense. They write: "an offense is provoked by specific vehicles for presenting slurs, *not* by occurrences of slurs themselves" (Stojnić and Lepore 2021: 747). I am very puzzled by this distinction. I grant that *no* word has any causal power in itself – words are presumably abstract objects of some sort, and abstract objects have to get themselves realized or materially instantiated in some way or other in order to have any causal effect. Stojnić and Lepore deny that words are abstract objects; they say that they are "shapes." I am very puzzled by this, since I think of shapes as abstract objects, too. But I don't think it's important to settle the metaphysics of words. As long as words are distinct from their individual articulations, I want to go with the words. I contend that the *effects* of the articulations of slurs that S &L make central to their account are themselves

effects of the *recognition* that a given articulation is *an* articulation *of a certain word*. (More on this model in a minute.)

I fully agree with Stojnić and Lepore that it is centrally characteristic of slurs that their articulations evoke negative and often quite painful reactions on the part of auditors of the slur, particularly if the auditor is a member of the group targeted by the slur. But the Articulation Account does not explain *why* articulations of slurs trigger negative reactions – why should the articulation of *these* words have these effects? The explanation certainly seems to be – somehow -- a matter of the history and sociology of the word. Stojnić and Lepore seem to think, however, that somehow the articulations themselves, independent of the words that are expressed, do all the work. "[C]ertain articulations automatically trigger negative associations, which come to be associated with these articulations through a web of complex causal, socio-historical, cultural, and psychological factors." My point, though, is that the Articulation Account offers no explanation for why *articulations* of a word should trigger negative associations, if not because they are articulations of offensive *words*. It's not like the sound or the orthography are inherently offensive, or have, independently of our identifying them as articulations of certain words, negative valence.

Stojnić and Lepore, consider, but do not accept, Mandelbaum and Young's suggestion that there is something inherently negative or distasteful in certain "phonesthetic" features (specifically, "velar plosives") and that the over-representation of such features in slurs may be significant. But I actually found their argument persuasive: if velar plosives turn out to be a significant factor in explaining why certain words (not necessarily slurs) are "disliked" more than others, then it makes sense that such sounds would show up more frequently in insulting or hurtful expressions. I take it, in other words, that Mandelbaum and Young's hypothesis is friendly to the lexical view.

Stojnić and Lepore reject the picture that I'm painting – they say it's a confusion to think that articulations only play a role in triggering offensive effects "insofar as they help us recognize which word is uttered" (Stojnić and Lepore 2021: 752) because such a picture involves the assumption that there's a multi-stage process – first hear the sounds, then figure out what word was articulated, then take offense. But (a) I don't deny that associations can bypass rational or cognitive processing; and (b) I take it for granted, as Stojnić and Lepore seem not to, that very rapid word recognition might *seem* not to involve cognitive processing even when it does; whether this is so is not a question that reflection on the phenomenology can solve.

In making this criticism, Stojnić and Lepore seem to be focused on cognitively *unmediated* association; this focus is misplaced (and, as we'll see in a minute, has to be shifted in order to account for *reclaiming* of slurs.) We know from critiques of behaviorism that the existence

of an association between stimulus and response does *not* mean that there is no cognitive mediation.[1] The explanation for one-trial extinction can be – generally is – that a necessary cognitive link has been broken – that's why the human being who is shown that the shock machine has been unplugged will have no or reduced galvanic skin response when they sit down at the apparatus.

Moreover, I think that Stojnić and Lepore's account seems to deny that the *cognitive state* of simply learning that someone used a certain slur – even if the slur itself is not articulated – can produce the offensive effects. But I think that, as a matter of fact, this happens. Suppose I say to Hermione: "Draco called you a you-know-what" – Hermione might well be– I say probably would be – triggered in a similar way to the way she would have been if I had articulated the slur. (Can we evaluate such counterfactuals? Why not?)

Let's turn then, to facts about the "reclamation" of slurs. While Stojnić and Lepore earlier seem to presuppose a behavioristic view of associations, their discussion here seems to make the pertinent associations highly cognitive – Stojnić and Lepore say that while negative associations cannot be erased, they can be "exploited or subverted in order to signal camaraderie or solidarity." But while one can simply choose to proudly declare oneself a "slut" or a "bitch" or a "queer," if the terms' potency to offend was really fundamentally a matter of ingrained associations, it's hard to see how reclamation could happen quickly, or voluntarily. I am not, obviously, saying that an association *couldn't* be broken by the adoption of some new propositional attitude (since that would conflict with what I've said about (e.g.) one-trial extinction). But it doesn't follow that any given association can always be broken in this way.

Relatedly, Stojnić and Lepore fail to explain why, even after reclamation by members of the in-group, *articulations* by members of *outgroups* still have the power to trigger offensive effects in members of the in-group. Surely the difference is entirely cognitive – the sound of the articulations of the slurring words could be identical between an ingrouper's pronunciation and an out-grouper's; the difference in effect would lie in the *knowledge* of what the out-grouper was doing.

Stojnić and Lepore say that their account does a much better job than its competitors in explaining certain central phenomena involving slurs. One important phenomenon is the fact that slurs *project*, and do so indiscriminately. It doesn't matter if I embed a slurring expression inside a context of indirect quotation – it still retains (or may do) its

[1] See the masterful review of "anomalies" in the behaviorist literature in Brewer (1974). Brewer distinguishes between the existence of an S-R or operant conditioning regularity, and the behaviorist's *explanation* for the regularity. He argues, compellingly, that even apparently very simple S-R regularities are best explained on a cognitivist model, as is shown by the sort of case of one-trial extinction described above.

stinging effect. But I think that the lexical account does just as well in explaining this, once one grants that words always have their effects *via* their articulation. The mere pronouncing of the slurring word can trigger the same painful reactions in an auditor even if the auditor recognizes the slur is being mentioned rather than used by the speaker.

(I want to just float an alternative – or perhaps additional – explanation, compatible with just about any account of slurs, for why slurs fail to project out of indirect discourse contexts. The fact is that we lack, in colloquial and (especially oral) speech, good mechanisms for indicating *direct* as opposed to indirect quotation. In written speech, we can use quotation marks, and in oral speech, we can say explicitly things like "Draco said, and I quote, ….." But these explicit indicators that words are being mentioned and not used are not obligatory. (So-called "air quotes," in oral speech, are ambiguous between indicating that the words air-quoted are merely being mentioned, and indicating that the speaker is using them in a "so-called" way.) In ordinary contexts, saying or writing "Draco said that Hermione is a mudblood" is simply ambiguous between a reading in which I'm *paraphrasing* Draco and one in which I'm *quoting* Draco. (And then of course there's the possibility that I'm essentially doing both – that is, Draco used the word "mudblood" and I also use that derogatory word to refer to muggles.) I wonder if there would be less evidence of projection if we were always perfectly explicit when we are mentioning the offensive word rather than using it. And even lesser if we, say, offer an explanation and apology in advance: suppose that Draco has beaten up Hermione, and the issue is whether the degree of offense should be raised because it was a hate crime (as can happen in the U.S.). The prosecutor might say: "I've very sorry to have to use the language I'm about to use, because I know that it is offensive to many people, and it is not language that I use myself, but I need you to know exactly what Draco said before he attacked Hermione: 'You are a rotten mudblood.'" I am not suggesting that the occurrence of the slur "mudblood" would not trigger difficult feelings for the persons targeted by the slur, but such feelings would, I submit, be different in nature from the ones triggered by a hateful *use* of the term.)

In further defense of a lexical account, I want to point out some important questions that I think the Articulation Account fails to answer, illustrated by the following anecdote, recounted in the British paper *The Independent*. The matter involved the philosopher Sidney Morgenbesser – renowned for his quick wit as well as his philosophical depth:

> [An] unfortunate encounter with the police occurred when he lit up his pipe on the way out of a [New York City] subway station. Morgenbesser protested to the officer who tried to stop him that the rules covered smoking in the station, not outside. The cop conceded he had a point, but said: "If I let you get away with it, I'd have to let everyone get away with it." To which Morgenbesser, in a famously misunderstood line, retorted: "Who do you think

> you are, Kant?" Hauled off to the precinct lock-up, Morgenbesser only won his freedom after a colleague showed up and explained the Categorical Imperative to the nonplussed boys in blue. (Gumbel 2004)

One important question this anecdote raises is: what counts as an *articulation of* a slur? Did Morgenbesser call the cop a c*nt, or did he not? Another question: when one is *misunderstood* as having uttered a slur, is that fact enough for the person *to have in fact* uttered a slur? (I'm not asking about the culpability of such a person – I'm only wondering if they've articulated a slur or not.) Note that S & L contend that it is a count against lexical accounts that some non-slurring words have articulations similar enough to common articulations of a slurring term to trigger the same articulations. But I don't see that the lexical account has any problem with this. The fact is that almost any word can be *misheard*, and the fact that it's misheard *as being* an articulation of word X doesn't tell us anything, other than that the articulations of some words are identical or similar to each other.

So I agree that the explanation for why pejorative effects project so indiscriminately has to do with associative effects – although I'll say a little more about the limits of this in a minute – but it's not *just* a matter of association; the effects of the association can be erased in at least some cases where one finds out more. If cop in the Morgenbesser story was capable of realizing that Morgenbesser was invoking the name of the philosopher Immanuel Kant, and wasn't calling him the C-word, he might have just chuckled and said, "Oh, I see – OK." (He might still have been insulted or made angry, but on a different basis.) That is, it might have mattered to the cop *what word* Morgenbesser was articulating. But it would be the same if Morgenbesser had been *misspeaking* – if he actually was trying to call the cop a c*nt, but mispronounced it as "kant" – if the cop understood Morgenbesser as invoking a philosopher, he would probably become insulted if he came to realize what Morgenbesser had been *trying* to say.

Finally, the Articulation account also has – or appears to have – an implausible consequence, namely, that certain speakers might be *unable* to slur. It's a general fact about words that it is through their articulations that we come to know *which* words have been articulated. Stojnić and Lepore cite as a virtue of the Articulation Account that "absent certain articulations, slurs do not offend." Well, in general, if a slur is misarticulated, the auditor may well fail to appreciate that a slur has been uttered. And no one can take offense without at least believing that an offense has been committed – or at least attempted. But someone with an uncommon speech pattern (maybe our hypothetical Morgenbesser trying to insult the cop) or someone who is not a native speaker of the language in which a certain word is a slur, may be unable to articulate the slur in a way that a target of the slur can understand. Do Stojnić and Lepore want to say that a person *cannot commit* the slur that they are trying to commit? Consider the following quote:

> Absent a particular articulation that carries negative associations, the tokening loses its sting; and the presence of articulation—the specific acoustic or orthographic shape that matches (or closely resembles) the canonical articulation of a slur on its own can trigger the pejorative potential, even when no word—and so no slur—has been tokened. (Stojnić and Lepore 2025: 108)

But I'm looking at the opposite case: Suppose that someone, say Sam, who Ray intended to insult with the slur, figures out – say, from hearing Ray express certain opinions and intentions (not involving the slurring word), or maybe by being told that that was Ray's intention, that Ray was trying to articulate the slur – I expect that this knowledge would be sufficient for Sam to take offense in exactly the same way and to the same degree as they would have had Ray successfully articulated the slur.

Stojnić and Lepore address this worry – they say: "But here, the speaker *is* still tokening a term: one can badly misarticulate a term while still articulating that term" (Stojnić and Lepore 2025: 110) but I don't see what, on their account warrants this conclusion. I say, it's the fact that they are trying to articulate a slurring *word* that makes them guilty of committing a slur. The usual negative associations are not triggered, yet there was still a slur committed.

They say:

> An expression that starts off as neutral can, over time, acquire negative associations, becoming a slur; unless we think of all such changes as ones in *meaning*, we cannot conclude that articulations trigger pejorative effects only derivatively, by articulating terms with offensive meanings. (Stojnić and Lepore 2025: 112)

I quite agree that there's no reason to view such processes in terms of changes in *meaning*, but that doesn't count against the view that it's the word – because of its history and sociology – that can, over time, become a slur.

In conclusion, let me reiterate that I have enormous respect for this book, and that I learned a great deal about slurs from reading it. Because I have here focused on my disagreements with Stojnić and Lepore, I have not been able to do justice to the richly detailed, original and systematic arguments they make against views with which the three of us disagree, systematizing and illuminating a number of difficult issues about the interaction among semantics, pragmatics, and social psychology. Anyone interested in slurs – or in language in general – needs to read this book.

References

Anderson, L. and Lepore, E. 2011. "Slurring Words." *Noûs* 47 (1): 25–48.

Brewer, W. F. 1974. "There is No Convincing Evidence for Operant or Classical Conditioning in Adult Humans." In W. B. Weimer and D. S. Palermo (eds.). *Cognition and the Symbolic Processes*. Hillsdale: Erlbaum, 1–42.

Gumbel, A. 2004. "Professor Sidney Morgenbesser." *The Independent*. August 5, 2004.

Cited at: https://en.wikipedia.org/wiki/Sidney_Morgenbesser [accessed August 23, 2024]

Palmer, B. 2010. "When Did the Word Negro Become Taboo?" *Slate* January 11, 2010. https://slate.com/news-and-politics/2010/01/how-old-was-harry-reid-when-the-word-negro-became-taboo.html Accessed July 16, 2025.

Stojnić, U. and Lepore, E. 2021. "Inescapable articulations: Vessels of lexical effects." *Noûs* 56 (3): 742–760.

Stojnić, U. and Lepore, E. 2025. *Inflammatory Language: Its Linguistics and Philosophy*. Oxford: Oxford University Press.

Twain, M. 1884. "Explanatory." *The Adventures of Huckleberry Finn*. Glassbook Classic, https://contentserver.adobe.com/store/books/HuckFinn.pdf

Croatian Journal of Philosophy
Vol. XXV, No. 75, 2025
https://doi.org/10.52685/cjp.25.75.5
Received: August 8, 2025
Accepted: October 31, 2025

Articulations and associations: Comments on Stojnić and Lepore's Inflammatory Language

MATTHEW STONE*
Rutgers University, New Brunswick, USA

I critically examine Stojnić and Lepore's key claim from their book Inflammatory Language that articulations are particularly important to the analysis of slurs. I agree with Stojnić and Lepore that we should endorse the significance of articulations for our linguistic intuitions, especially when it comes to the causal and social powers of language; I argue that for locating this approach to the articulations of slurs in a broader context of fiction and taboos. However, while I accept Stojnić and Lepore's evidence for theorizing some slurs through associations at the level of articulations, I argue for continuing to account for slurs in part through prohibitions and through associations at the level of words and concepts.

Keywords: Slurs; associations; norms; words; concepts.

1. Introduction

In recent years, philosophers of language have begun to discuss slurs intensively. In part because people have such strong intuitions about slurs, slurs help us address new questions about the nature of meaning in language, the role of language in culture, the politics of language and interaction, and the psychology of language processing. Una Stojnić and Ernie Lepore's book *Inflammatory Language* (2025) continues this trend with several exciting and provocative suggestions. As I explain in this paper, their work brings an important new set of issues to the

* A preliminary version of this paper was presented at the Workshop on Linguistics and Philosophy of Language in September 2024; this presentation benefits from discussions with the participants of the workshop. Supported by NSF awards 2021628, 2119265, and 2427646.

table, encompassing a wide range of novel arguments and considerations, informed by important and previously unexplored categories of data. The excitement of these contributions lies in part in the ability of philosophers to engage with them critically. I think it's a testament to the success of their book that the counterarguments and clarifications that I'll be presenting in this paper could only have been formulated against the distinctions and claims that they put forward in the book.

A variety of proposals aim to characterize what makes slurs offensive. An important direction has been to link the offensive character of slurs to the meaning that's conventionally and arbitrarily associated with them. Generally, the offensive content is taken to be some kind of claim that disparages or demeans a target group. Theories differ about how that disparaging content is associated with the slur.

One idea is that it is part of the propositional content normally conveyed by the slur, what linguists tend to call its at-issue content. This is a proposal that's most associated with Christopher Hom (2008). Another idea is that the disparaging content is a background assumption that the speaker takes to be uncontroversially shared with the hearer: in other words, that it is a presupposition. This proposal is often associated with Philippe Schlenker (2007). A different way of associating meanings with slur terms is to assume that the content is not assumed to be shared with the hearer, but is also not assumed to be put forward, as part of the main claim being advanced by the speaker. Many terms have such arbitrary incidental contributions to meaning; they generally are known as conventional implicatures, following Christopher Potts (2005). Timothy Williamson (2009) highlights the ability of this conventional implicature analysis to capture the disparaging content of slurs.

Yet another approach aims to assimilate the content of slurs, not to an explicit disparaging meaning, but rather to an implicit attitude or inference that the speaker and perhaps the audience are expected to take regarding the use of the slur. One example of this kind of approach is to account for slurs in terms of expressive meaning (Potts 2007)—a contribution that indicates that the speaker and perhaps the audience should take a negative attitude towards the target group without thereby making a specific claim about the properties of the group that merit this negative attitude. A different approach is taken by Elisabeth Camp (2013), who accounts for slurs in terms of disparaging perspectives—that is, patterns of inference and prominence that emphasize negative stereotypes about the target group without thereby encoding a particular propositional contribution. Her analogy is the perspective from Wittgenstein's Duck–Rabbit, of seeing the lines as either forming duck or forming a rabbit.

As you can see, there's an incredible variety of directions that can be used to conceptualize a disparaging meaning as encoded into the slur itself. At the same time, many researchers take a very different kind of

approach, and this is the approach that Ernie Lepore has developed in a range of different work over the last decade and more (Anderson and Lepore 2013a, Anderson and Lepore 2013b, Lepore and Stone 2018). For Lepore, the idea of accounting for the meaning the offensive character of slurs in terms of meaning is misguided. There is no unique signature of offensive meaning in slurs that differentiates slurs from neutral counterparts. Instead, as Lepore and Stone (2018) put it, slurs differ from their neutral counterparts in heterogeneous, open-ended and often non propositional ways, including words whose loaded associations evoke and perpetuate offensive imagery. The important point for Lepore and Stone is that these loaded associations aren't conventionally associated with words but instead arise organically from the words' history and the interpretive practices of audiences that look in expansive and open-ended ways in understanding what's said.

In the face of this heterogeneity, Anderson and Lepore (2013a, 2013b) claim that ultimately, the only thing that slurs share is that it's prohibited to use them. This prohibitionist approach offers a very different kind of explanation for the offensive character of slurs, one that doesn't point at meaning and the direct relationship of speaker and audience that meaning helps to mediate, but instead points more broadly at the social and cultural context for language use, in which the regimentation of language takes on political objectives and gives political meanings over and above the conventional meanings for the social significance of the use of a word with a particular political status.

This is the context in which Stojnić and Lepore's work unfolds. Following Lepore's previous suggestions, Stojnić and Lepore reject semantic views. They also criticize and reject prohibitionism as giving an inadequate story about the interpretation of slurs. And to tie these arguments together, they characterize the associations of slurs in a new way, one that they argue directly can account both for when we find slurs and related items offensive and what offense those slurs seem to carry. It is in this positive proposal—and specifically in the idea that articulations are particularly important to the analysis of slurs—that a range of new data and new arguments come into play.

Articulations are systematic ways of realizing linguistic terms, and they were first brought to attention by Hawthorne and Lepore (2011) in their work on the metaphysics of words. As Hawthorne and Lepore observe, "[o]ne and the same word can be written on a pad with a pen, typed on a sheet of paper, projected on a screen, spoken out loud, signed with a gesture, or Brailled on a plaque" (2011: 13). Articulations are these distinctive ways of presenting the word through a particular medium; the word is the underlying abstraction that's realized across all the different articulations. This metaphysics is not just a philosophical curiosity. It plays an important role in helping to get clear on what's behind ordinary language users' talk about linguistic forms. For example, it's intuitive to say that 'color' and 'colour' are two spellings of the same

word. You can make sense of this claim by thinking of the spellings in quotes as exhibiting articulations, which can, in fact, be different articulations of the same abstract word.

What Stojnić and Lepore argue is that what is problematic about slurs is their articulations. It's in the articulations, not in the words, that we should look for offensiveness. To bring home this point, they offer a telling example from Chinese (reported by Matisoff 1986) where the writing system offers two different characters to represent a particular word, the name "Yao." The two characters differ in the radicals that make them up. One, 傜, features the person radical, suggesting that the character denotes a person. That's the unoffensive articulation. Historically, there's also been a character 猺 that combines the same phonological cue for the name "Yao" with a radical for beast.

Obviously, it can be upsetting to be denoted by a character that implicitly identifies the semantic category that one belongs to as an animal. But of course, it's not the name, not the word, that's offensive, only the articulation with the character 猺. There's nothing offensive at all about the alternate articulation 傜 with the person radical.

Thinking more broadly, Stojnić and Lepore suggest that we should always look for the offensiveness of slur terms in the associations of their articulations, and that the offensiveness that we find there is sufficient to explain what's bad about slurs without necessarily appealing to any kind of prohibition on the use of slurs. These claims take Stojnić and Lepore directly into dialogue with a paper that Lepore and I wrote on pejorative tone.

What I'll do in this paper is offer a few lines of defense or counter argument around the points that Lepore and Stone (2018) make in that work. I'll suggest that even if prohibitions might not carry the load that Anderson and Lepore say they do, it's still clear that slurs (and many other kinds of offensive words) are prohibited. This prohibition does sometimes shape the social implications of their use. Finally, while it's clear that articulations do have associations, I will try to present a range of arguments that show that words also have associations. It may well be the associations of slurs as words, as opposed to the associations of their articulations, that underpin the offensiveness of tokens of slurs. Thus, while the attention to articulations and associations in Stojnić and Lepore's book is, in my opinion, an important advance, I would not draw the radical conclusions they draw from their reflections.

2. The intuitive significance of articulations

One way to join Stojnić and Lepore in their appreciation of articulations is to observe that articulations are given a special place in folk theories of the causal effects of language. That's particularly evident in the principles that govern fantasy worlds, where readers readily accept

magical workings of language that are mediated by the properties of articulations.

In Tolkien's epic, *The Fellowship of the Ring* (1954), we discover that the gates of Moria, the famous underground Dwarven city, are inscribed with a message that says, "speak friend and enter." The runes apparently invite the audience, addressed as a friend, to speak some unstated password. In fact, though, the solution to this puzzle is simply that "friend" is the password.

For our purposes, what's notable about the inscription in the story is that that speaking the word is what opens the door. Presenting the door with the word in some other way—with an inscription for example—would not satisfy the conditions of the spell.

There's the opposite kind of magic in the manga *Death Note* (Ohba and Obata 2003-2008), where writing someone's name in the titular magical notebook engenders the person's death. Of course, talking to the book, saying someone's name, is harmless.

The contrast between the two cases shows the sensitivity of the effects of language to an appropriate articulation. The magic attaches whatever articulation it does, sometimes the spoken word, sometimes the written word, and perhaps, in other cases, to some other articulation.

Now, there is no magic in our world (as best we can tell). What these literary possibilities—these intuitive thought experiments—tell us, then, is not how language works, but rather how people are prepared to conceptualize it. If language users thought that articulations were inert, that it was the words that did the work, these fantasy logics would not be compelling. The resonance of the stories points to our everyday readiness to accept a privileged role for articulations in the workings of language.

Sensitivity to articulation seems to matter particularly for taboos around the use of language. It's very often aspects of the articulation that are taboo, not just the word or its meaning. In such examples, the properties of articulations that Stojnić and Lepore call attention to are very much in evidence.

By Jewish custom, the Hebrew name for God cannot be pronounced; it can, however, be written (although with special considerations for the disposition of these inscriptions, see Toy and Blau 1906). A distinctive prohibition against its use applies to the spoken form.

A further example comes from the taboo on the dead that many indigenous cultures maintain, particularly the indigenous cultures of Australia (Dixon 2002). These taboos prohibit speaking the name of someone who has died during a subsequent mourning period. In some communities, even similar sounding words are to be avoided—a clear constraint on articulations (reminiscent of the infamous "niggardly" controversy; see O'Hehir 2015). In other communities, the use of the name is prohibited, even to refer to other people (or nonhuman entities)

whose name involves the same form. If you think (following Kaplan 1990) that these different kinds of semantic interpretations amount to differences in vocabulary, then even this taboo is a constraint on articulations rather than a constraint on the use of words.

In sum, we should endorse the significance of articulations for our linguistic intuitions, especially when it comes to the causal and social powers of language (as manifest in fiction and taboos). Stojnić and Lepore have done a great service by connecting the discussion of slurs to this broader context.

3. Language and social norms

We can agree that articulations are important, however, while still acknowledging that articulations are just one among many levels of linguistic representation with special political and social significance.

After the 2022 invasion of Ukraine, Russian authorities insisted on describing events as a "special military operation." To call the hostilities a "war" was a politically charged description, one that could be understood (and punished) as an expression of resistance to Putin's propaganda machine and perhaps even as opposition to the regime itself (Al Jazeera 2022). Obviously, it's not merely articulating word "war" that is significant in this case. Russians continued with officially endorsed commemorations of the "Great Patriotic War," for example. Rather, the problem is the frank categorization that the scale of time, treasure, and blood at play in Ukraine merits a different kind of categorization than a mere "operation." The political repercussions of this case seem to align with observations often made about the contestation and litigation of meaning in recent work in the philosophy of language. Communities can police conventional meanings they do not wish to see used, just as they police articulations they regard as problematic.

Communities can also police speakers for expressive actions in ways that aren't easily tied down to either form or meaning. The Thai laws of Lèse-Majesté are a case in point (Connors 2002) They criminalize any expressive behavior that carries a (generally understood) message of disrespect toward the royal family. Enforcement is not limited to language; gestures or other behaviors with plain social meanings are also prohibited. In one extreme example, wearing black on the King's birthday was understood as an expression of this kind of disrespect, and led to charges under the law (Prachatal English 2014). Here too, it seems like we cannot adequately theorize the norms at play merely at the level of articulations.

It seems that some social and political rules around languages do involve articulations, but others seem to involve words, references, meanings and more. Regardless, whenever rules exist, it can mean something to break them. Speakers can use a violation of the rule to make a point—for example, to indicate the urgency and passion of their feelings, their contempt for the powers that be, or simply their inabil-

ity to formulate their ideas within the bounds of propriety. For that reason, Lepore and Stone argued that the interpretation of slurs as offensive is often colored by the fact that speakers who use them are self-consciously failing to conform to an established norm.

This is the reason why I think that just establishing the associations of articulations, and their often-problematic status, doesn't necessarily obviate the need to invoke prohibitionism in giving a full account of the offensiveness of slurs.

4. Words, articulations, and associations

My assessment of associations matches my assessment of prohibitions. Already in this paper, I have argued that it is crucial to recognize taboo articulations, but that it is also necessary to situate such taboos among the many kinds of prohibitions that exist in language and culture. In a similar way, I claim that it is often crucial to theorize associations at the level of articulations, but that it is also necessary to situate such associations among the many different levels of analysis where language carries associations. All those different kinds of associations may well be relevant to the treatment of inflammatory language.

In fact, in the cognitive science of language, the most influential appeals to association have been found at the level of words or concepts. A famous example is Collins and Quillian's (1969) spreading activation theory of semantic memory and priming. They proposed that the activation of a concept in memory facilitates the processing of associated concepts. Classic examples of such facilitation—an occurrence of "dog" speeding recognition of the word "cat" or the word "bark", and occurrence of "come" speeding recognition of the word "go"—clearly indicate the semantic or conceptual basis for moving from one word to another. Similarity of sound or form isn't what drives semantic priming—as the name itself of course makes clear.

Cognitive science has a rich tradition of work that explains problematic attitudes and judgments on the part of speakers in terms of associations. That tradition does not straightforwardly support Stojnić and Lepore's interpretation of associations, however. Instead, it builds on theories of semantically based facilitation, like Collins and Quillian's. One influential paradigm comes from implicit association tests (Greenwald et al 1998) that purport to reveal the problematic stereotypes that influence subjects subconsciously as they perform language processing tasks (and other tasks with stimuli associated with target groups). The received explanation is that experimental effects—typically, faster or more accurate responses for stereotype-aligned judgments—reflect the facilitation engendered by subjects' biased and stereotyped conceptual networks. For example, if subjects have a concept of Blackness that is associated with concepts of poverty or crime, then they'll be more disposed and faster to draw inferences about crime in contexts where the concept of Blackness has been activated. Again, those are associations

at the level of concepts, not at the level of words or articulations, at least the way the theories are generally presented.

For Stojnić and Lepore, however, it's the articulations and their associations that are primary, that are really driving the show. Can they respond to evidence in favor of semantic associations? I think they can, to a surprising degree, because of the surprising richness of textual associations as a proxy for semantic reasoning—a point brought home over the last decade by computational models of "distributional semantics."

Already in the 1950s, the linguist Firth famously linked studies of semantics to studies of textual associations with his slogan "you shall know a word by the company it keeps" (1957:11). In computational linguistics and cognitive computational modeling, researchers have operationalized this suggestion to demonstrate how much inference can be guided just by distributional similarity across linguistic articulations.

The word2vec model (Mikolov et al 2013), for example, induces vector representations—that is, numerical coordinates in a high-dimensional space—to capture trends in the co-occurrences of words in gigantic internet-scale corpora. These representations just summarize, as Firth would put it, the company a word tends to keep. But you can nevertheless analyze these vectors to discover indicators for morphological relations like the difference between singular and plural, for semantic relations like the marking of gender in the words "king" and "queen," for general knowledge like the relations between countries and their capitals, and for harmful stereotypes such as the problematically gendered textual associations of "sewing" as feminine on the one hand and "carpentry" as masculine on the other (Bolukbasi et al 2016).

Often these observations or findings are taken to imply that textual associations can be a perfect proxy for semantic associations, meaning that, from the point of view of Stojnić and Lepore's argument, the associations of articulations capture all semantic associations and more. But this expansive interpretation of models like word2vec is tendentious. Finley et al (2017), for example, in their analysis of semantic inferences and distributional textual representations, find that only a small number of semantic relationships are well modeled in the vector embedding space, and those tend to be relationships that have extremely clear signatures in the distribution of frequent, correlated items.

This is a fast-moving research area—but the very pace of progress is a telling indication that questions about the difference between textual associations and semantic associations are questions for science more broadly, not just questions about philosophical intuitions. In fact, there are some spectacular examples in the literature that show that our intuitions about forms are misleading: speakers tend to overattribute to articulations properties that ultimately might better be characterized as semantic. (There is perhaps an echo of the magical understanding of the articulations of language I opened with in Section 2.)

The word "moist", as studied by Thibodeau (2016), is a good example. Lots of people hate the word "moist". When you ask why, they tend to speculate that it's the phonological properties of the word that are the cause of their displeasure. As one participant put it, "it just has an ugly sound that makes whatever you're talking about sound gross." However, Thibodeau and colleagues found instead that a better explanation for people's unpleasant experience with the word was its semantic associations with disgusting bodily functions. With the case of "moist" in mind, one should be very skeptical of one's intuitions that it's somehow the form of a word that is making it objectionable, rather than something about what property or referent the word evokes.

How might we go about distinguishing between associations at the level of meaning and associations at the level of form? One of the exciting contributions of Stojnić and Lepore's book is to press us to ask this question. I'm intrigued at the direction of research that it leads to, which I'll briefly explain in closing,

A good way to distinguish between words and articulations is provided by verbs. In many languages, verbs have productive morphology, which means that you can encounter a verb as a word through an articulation—through a specific form—that you may never have encountered before at all, and that you certainly have not encountered frequently enough for you to link that form with specific associations.

But verbs can be inflammatory, just as nouns and adjectives can. Adam Sennet and David Copp (2020) call attention to pejorative verbs—verbs for behaviors that are derived from slur terms and incorporate denigrating stereotypes. In English, verb morphology is comparatively simple, so Sennet and Copp can't give us English examples of the productive slurring character of pejorative verbs across a generative class of articulations. But we might hypothesize that there could be slurring verbs in languages with productive morphology. For such verbs, you might point to a rare or unattested morphological form of this slurring verb as a slur whose offensiveness must depend on the word that it's made of, not of the form, since the form has never been previously articulated.

As a morphologically impoverished English speaker, I don't have much to add about the cross-linguistic productivity of slurring verbs, but it seems like an important topic, and finding examples of this kind would be quite interesting. Nevertheless, rather than leaving the phenomenon as a mere speculative possibility, I'd like to point to a possibly related case to suggests that productive slurring verbs might not be so hard to find.

What I'll be using is the Japanese verb "yagaru" (Protonstorm 2021), which is a suffix that fits into the productive morphology of the Japanese verb. It shows contempt for whoever is doing the action in a way that in certain contexts and registers can be inflammatory. As an example, consider this example:

何を かんがえやがったんだ
nani-o kangae-yagat-ta-n-da
What-acc think-yagaru-past-compl-be
What is it that you were fucking thinking?

It is, as the saying goes, "a nasty question."

The challenging suggestion here would be that "yagaru" inflection productively adds contemptuous tone to a verb but does so by creating forms and articulations that you may never have heard before. If that's right, then it seems necessary to associate the tone with the morpheme, not the articulation. It's the morpheme that's doing the work. An analogous slur would show the limitations of thinking about associations purely at the level of articulations. It would show that we cannot dispense with the other associations for words, concepts, and objects in the world that psychologists regularly postulate and theorize.

5. Conclusion

In sum, Stojnić and Lepore's book has convinced me that articulations are an important part of the story of the offensiveness of inflammatory language. But I'm not ready to abandon prohibitions as part of the explanation, nor am I ready to abandon associations at the levels of words or concepts as an important explanatory tool to describe the effects of slurs and pinpoint their offensive nature.

References

Al Jazeera (no byline) 2022. "Do not call Ukraine invasion a 'war', Russia tells media, schools". 2 March 2022. https://www.aljazeera.com/news/2022/3/2/do-not-call-ukraine-invasion-a-war-russia-tells-media-schools

Anderson, L. and Lepore, E. 2013a. "What Did You Call Me? Slurs as Prohibited Words." *Analytic Philosophy* 54 (3): 350–363.

Anderson, L. and Lepore, E. 2013b. "Slurring Words." *Noûs* 47 (1): 25–48.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. 2016. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." In *Proceedings of the 30th International Conference on Neural Information Processing Systems. (NIPS'16),* 4356–4364.

Camp, E. 2013. "Slurring Perspectives." *Analytic Philosophy* 54 (3): 330–349.

Collins, A. M., and M. Ross Quillian. 1969. "Retrieval time from semantic memory." *Journal of Verbal Learning & Verbal Behavior* 8 (2): 240–247.

Connors, M. K. 2002. *Democracy and National Identity in Thailand*. London: Routledge.

Dixon, R. M. W. 2002. *Australian Languages: Their Nature and Developments*. Cambridge: Cambridge University Press.

Finley, G., Farmer, S., and Pakhomov, S. 2017. "What analogies reveal about word vectors and their compositionality." In N. Ide, A. Herbelot, and L. Màrquez (eds.). *Proceedings of the 6th joint conference on lexical and computational semantics*, 1–11.

Firth, J. R. 1957. *Studies in Linguistic Analysis*. London: Blackwell.

Greenwald, A. G., D. E. McGhee, and J. L. K. Schwartz. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* 74 (6): 1464–1480.

Hawthorne, J. and Lepore, E. 2011. "On Words." *The Journal of Philosophy* 108 (9): 447–485.

Hom, C. 2008. "The Semantics of Racial Epithets." *Journal of Philosophy* 105 (8): 416–40.

Kaplan, D. 1990. "Words." *Proceedings of the Aristotelian Society*, Supplementary Volumes, lxiv. 93–119.

Lepore, E. and Stone, M. 2018. "Pejorative Tone." In D. Sosa (ed.). *Bad Words: Philosophical Perspectives on Slurs*. Oxford: Oxford University Press, 132–154.

Matisoff, J. 1986. "The languages and dialects of Tibeto-Burman: An alphabetic/genetic listing, with some prefatory remarks on ethnonymic and glossonymic complications." In: J. McCoy and T. Light (ed.). *Contributions to Sino-Tibetan Studies*. Leiden: E. J. Brill, 3–57.

Mikolov T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. 2013. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems* 26.

Potts, C. 2005. *The Logic of Conventional Implicatures*. Oxford: Oxford University Press.

Potts, C. 2007. "The Expressive Dimension," *Theoretical Linguistics* 33: 165–198.

Prachatai English (no byline) 2014. "3 lèse majesté complaints filed against women wearing black around King's birthday" 10 December 2014. https://prachataienglish.com/node/4584

Protonstorm 2021. "Talk Like a Japanese Pirate with 'Yagaru.'" *Medium.com,* Jun 4 2021. https://medium.com/anitay-official/talk-like-a-japanese-pirate-with-yagaru-8b79fba2062f Retrieved July 2025.

Ohba, T. and Obata, T. 2003–2006. *Death Note*. Serialized In *Weekly Shonen Jump,* December 2003-May 2006, Shueisha.

O'Hehir, A. 2015. "So Much for Youth Apathy: Student Radicalism Escapes the '60s at Last," Salon.com, November 17, 2015. Retrieved February 17, 2018, from http://www.salon.com/2015/11/17/so_much_for_youth_apathy_student_radicalism_escapes_the_60s_at_last/

Schlenker, P. 2007. "Expressive Presuppositions." *Theoretical Linguistics* 33: 237–245.

Sennet, A. and Copp, D. 2020. "Pejorative Verbs and the Prospects for a Unified Theory of Slurs." *Analytic Philosophy* 61: 130–151.

Stojnić, U. and Lepore, E. 2025. *Inflammatory Language: Its Linguistics and Philosophy*. Oxford: Oxford University Press.

Thibodeau, P. H. 2016. "A Moist Crevice for Word Aversion: In Semantics Not Sounds." PLoS ONE 11 (4).

Tolkien, J. R. R. 1954. *The Fellowship of the Ring*. London: Allen and Unwin.

Toy, C. H. and Blau, L. 1906. "Tetragrammaton". In Cyrus Adler and Isodore Signer, (eds.). *The Jewish encyclopedia; a descriptive record of the history, religion, literature, and customs of the Jewish people from the earliest times to the present day*. Funk and Wagnalls, vol 12: 118–120.

Williamson, T. 2009. "Reference, Inference, and the Semantics of Pejoratives." In J. Almog and P. Leonardi (eds.). *Festschrift for David Kaplan*. Oxford: Oxford University Press.

Croatian Journal of Philosophy
Vol. XXV, No. 75, 2025
https://doi.org/10.52685/cjp.25.75.6
Received: September 11, 2024
Accepted: November 8, 2024

Sobel-esque Sequences and Felicity Judgments in Philosophy of Language

ADAM MICHAEL SENNET
University of California, Davis, USA

TYRUS FISHER
Independent scholar

This paper considers reverse Sobel sequences, NPI licensing, and related speaker judgments as they bear on von Fintel-style dynamic approaches to the semantics of subjunctive conditionals. We argue that neither reverse Sobel sequences nor von Fintel's explanation of NPI licensing speak in favor of von Fintel's semantics and that the alternative Lewisian approach, augmented by Moss-style pragmatic considerations, can accommodate and predict the relevant data at least as well as, and in some cases better than, von Fintel's view. Our arguments include counterexamples to von Fintel's semantics with respect to both its treatment of reverse Sobel sequences and NPI licensing.

Keywords: Subjunctive Conditional; Sobel sequence; pragmatics; dynamic semantics; counterfactual

1. Introduction

Work in philosophical semantics typically proceeds without a clear-cut answer to the question of how infelicity data should guide semantic analyses. It would be great if we could reliably infer constraints on positing semantic values from utterance or sentence infelicity. Sadly, we should know better.

A case in point concerns Sobel Sequences (hereafter SSs) and their reverse counterparts, Reverse Sobel Sequence (RSSs). These provide a nice example of unclarity about how felicity judgments should guide our theorizing. For, if the semantic import of the data here were clear, much of the controversy surrounding RSSs would evaporate. It's agreed

by all that the following sequence is felicitous given normal assumptions about parades and the presence of Pedro Martinez at the relevant parade:

(1) a. If Sophie had gone to the parade she would have seen Pedro.
b. But if Sophie had gone to the parade and been stuck behind a tall person, she wouldn't have seen Pedro.

Even Al Hájek (ms.), who argues that counterfactuals such as (1a) and (1b) are both false (regardless of order) will agree that the above sentences uttered in this sequence are felicitous. On the other hand, the reverse counterparts of such sequences are often clearly infelicitous:

(2) a. If Sophie had gone to the parade and been stuck behind a tall person, she wouldn't have seen Pedro.
b. # But If Sophie had gone to the parade she would have seen Pedro.

No one denies the infelicity of (2b) when uttered after (2a), and it is tempting to explain this infelicity via a story involving the claim that (2b) is false when uttered after (2a). If we could be sure this was the correct way to interpret the infelicity data here, we would be sure of an interesting and substantive constraint on the semantics of subjunctive conditionals. Indeed, the infelicity of (2b) here has been taken as motivation and even evidence for a range of theories that predict its falsity. Kai von Fintel (2001) provides what we take to be the paradigm case of an account that predicts the truth of (1a) and (1b) but the falsity of (2b) when uttered in the order given by (2) (Cf. von Fintel 2001: 132, 146, endnote 8).[1]

On the other hand, there is work that explains infelicity in terms of some type of conversational incoherence. On such a view, the infelicity of (2b) is to be explained by appeal to an assertability failure, while preserving the possibility that (2b) is true in some contexts even when uttered after (2a). We will take Sarah Moss' (2012) augmentation of a Lewis-style semantics with her pragmatic principle (EI) as our paradigm case here.

In any case, a first point to appreciate is that the infelicity data doesn't give an obvious advantage to either view, but it does make questions about the right theoretical treatment here pressing and interestingly hard to settle.

In this paper, we defend the following claims:

(a) *Against If to Might* Von Fintel touts a proposed entailment of sentences like (2a) as evidence for his theory's explanation of the falsity of sentences like (2b) (von Fintel 2001, Gillies 2007, von Fintel & Gillies 2012). But there are counterexamples to the alleged entailment.

[1] Von Fintel's 2001 and Gillies' 2007, and Moss' 2012 are the best known treatments of RSSs. In this paper we do not consider Gillies' treatment. This is primarily because Gillies' machinery is more complex. Nevertheless, our points in this paper apply for the most part to Gillies' treatment.

(b) *NPIs and SDE* Von Fintel touts his explanation of negative-polarity-item licensing in conditional antecedents as evidence for his view. But von Fintel's explanation overgenerates by predicting NPI licensing in environments that don't license NPIs, so it doesn't seem like his explanation works.

(c) *Need Pragmatics* Von Fintel's theory utilizes only a limited range of tools to explain the predicted falsity of (2b). To substantially improve the empirical coverage of theories like his, one would need pragmatic elements of the same sort that Moss already appeals to directly in her treatment.

(d) *True RSSs* Some RSSs are comprised of sentences that are jointly true regardless of their order of utterance. But von Fintel's theory entails that the second sentence of any RSS is inconsistent with the first.

(a), (b), (c), and (d) are considerations that count against accepting von Fintel's view. Interestingly, each of (a) through (d) concern features of von Fintel's theory that others have thought count in favor of von Fintel's view over the Moss-Lewis view, but we show they do not count as such.

Section 2 describes von Fintel's theory and the Moss-Lewis view. Section 3 presents and defends (a) and (b), and Section 4 does the same for (c) and (d).

All in all, we think that RSSs and their surrounding phenomena end up supporting views like Moss' (2012) better than views like von Fintel's. This is surprising given how RSS's and the related felicity-judgment data have been used to motivate and support von Fintel's theory over Lewisian alternatives.[2]

2. Background: von Fintel's theory and the Moss-Lewis view

Crediting J. Howard Sobel for calling his attention to such sequences, David Lewis (1973b: 10) observed that the sentences of SSs are quite often jointly assertable. But the joint truth of such a pair would constitute a counterexample to the validity of *Antecedent Strengthening* (AS):

(AS) $\phi > \psi \; F \; (\phi \wedge \chi) > \psi$[3]

Since (AS) is predicted to fail on a Lewis-Stalnaker theory, the felicity of Sobel sequences was thought by Lewis and others to speak in favour of such theories. But von Fintel (2001) and Gillies (2007),

[2] Cory Nichols (2017) has recently presented argumentation in a similar vein to some of our argumentation in this paper. We note that the present paper differs from Nichols' in that we consider von Fintel's argumentation concerning NPI licensing, and we consider a number of cases not considered by Nichols (and he considers some cases we do not).

[3] We use '*F*' here to denote natural-language entailment. We will let context disambiguate when we use this symbol to denote the entailment relation of some formal theory.

following an observation of Irene Heim's, present a problem for Lewis-Stalnaker theories: reversing the order of utterance for the sentences of a felicitous Sobel sequence makes the last sentence unassertable.[4]

2.1 Von Fintel's theory

Von Fintel (2001) takes the infelicity of RSSs as a data point and a mark of the falsity of their second coordinate. The basic idea of von Fintel's theory and his explanation of the infelicity of RSSs is that a context carries a *modal horizon* with it. This modal horizon comprises the possibilities that are available for interpreting subjunctive conditionals. If we consider a discourse-initial subjunctive conditional, then on von Fintel's view a counterfactual is to be evaluated just as in Lewis' theory of counterfactuals (assuming Limit):[5] a counterfactual $\phi > \psi$ is true just in case all the nearest ϕ worlds are ψ worlds.

But, if a more distant possibility, χ, is mentioned in the antecedent of a subjunctive conditional, then the modal horizon expands just enough to let in all the nearest χ worlds plus all the worlds at least as close to the actual world as the nearest χ worlds. Von Fintel motivates this expansion by positing a presupposition carried by subjunctive conditionals. The presupposition is that the modal horizon contains a world satisfying the antecedent. Thus, if the modal horizon doesn't contain such a world, it expands to include one. In the formal semantics, the modal horizon is modeled as a set of worlds and any expansion comprises a superset, as above, of the worlds previously in the horizon.

For simplicity of presentation, we consider only bare conditionals (conditionals with no embedded conditionals or other modals). We describe the theory more precisely for this special case with a few definitions:

> *The nearness relation* $\leq$ is a function from worlds to orderings of worlds corresponding to some metric of similarity assumed to be given by context.[6]

[4] As noted by Moss (2012 footnote 5), von Fintel credits Heim with noticing the infelicity of RSSs (von Fintel 2001: 130).

[5] Limit being the thesis that for any proposition ϕ and any world w, there is a closest ϕ world to w, provided there are some ϕ worlds at all. Assuming Limit typically allows one to simplify one's preferred statement of truth conditions for counterfactuals. The Limit assumption makes a difference to which counterfactuals come out true at a world (Cf. Lewis 1973b: 20). The affect of Limit on the logic generated by a semantics is more subtle. For a system of spheres (SOS) semantics, Limit has no characteristic axiom (Cf. Lewis 1973a: 121). However, Nute shows that adding Limit may change the theory's deduction principles. In particular, every SOS semantics satisfying Limit validates the deduction principle GCP (Nute 1980: 72).

For more on discoveries about the content and consequences of Limit Assumption(s) in formal frameworks, see William Starr's (2019) SEP entry, "Counterfactuals" (in particular, the supplement, "Formal Constraints on Similarity") as well as Stefan Kaufmann's (2017) "The Limit Assumption."

[6] We make no assumptions here about the properties of these orderings. However, following Lewis, it is typical to hold that the orderings should be total, reflexive,

Accessibility functions An accessibility function, f^n, is a function given by context from worlds to sets of worlds (modal horizons). We may think of each accessibility function as mapping a world of utterance to the contextually-live counterfactual possibilities. Letting f^0 denote a discourse-initial accessibility function, we assume for simplicity that $f^0(w) = \{w\}$. We assume a given discourse and stipulate that the counting numbers indexing the accessibility functions track the number of counterfactuals uttered in this discourse.

Then, on von Fintel's theory we can define the *context change potential* (CCP) of a counterfactual $\phi > \psi$ as a function from accessibility functions to accessibility functions such that:

CCP $f^n|\phi > \psi|^{\leq} = \lambda w.\ f^n(w) \cap \{w^l : \forall w^{ll}[w^{ll} \in [\![\phi]\!] \supset w^l \leq_w w^{ll}]\} = \lambda w.\ f^{n+1}(w)$. And the truth conditions:

vF Truth conditions $[\![\phi > \psi]\!]^{f^{n\leq}}(w) = 1$ iff $\forall w^l \in f^n|\phi > \psi|^{\leq}(w)$: $w^l \in [\![\phi]\!] \supset w^l \in [\![\psi]\!]$,

equivalently:

iff $\forall w^l \in f^{n+1}(w)$ $[w^l \in [\![\phi]\!] \supset w^l \in [\![\psi]\!]]$.[7]

In effect, we have an analysis of the counterfactual as a strict conditional $\Box(\phi \supset \psi)$, where the necessity operator is interpreted by the quantified expression: $\forall w^l \in f^{n+1}(w)$, and the uttered counterfactual determines the operative set of accessible worlds, $f^{n+1}(w)$ (the modal horizon), as a function of the set $f^n(w)$ and the antecedent ϕ. Hence, the range of quantification continually expands so as to track the contextually-live subjunctive possibilities. As von Fintel (2001) highlights at the start of his paper, on this analysis the meaning of a counterfactual has two aspects: "it alters the initial context c [corresponding above to $f^n(w)$] to a new context c^l [corresponding above to $f^{n+1}(w)$]. . . and maps c^l to the proposition p [the proposition expressed by the counterfactual] in a systematic. . . way" (von Fintel 2001: 123; our brackets). Thus we

and transitive (i.e., they are total preorders). But see Pollock (1976: 43–44, 1981), Hiddleston (2005) and Briggs (2012) for presentation and/or motivation of semantic theories that violate totality.

[7] Given the definitions above, we can present Lewisian truth conditions (assuming Limit) as follows:

Weird Lewis Truth Conditions

$[\![\phi > \psi]\!]^{f^n}(w) = 1$ iff $\forall w \in f^0(w) \cap \{w^l : \forall w^{ll}[w^{ll} \in [\![\phi]\!] \supset w^l \leq w, w^{ll}]\}$: $w^l \in [\![\phi]\!] \supset w^l \in [\![\psi]\!]$

The above is equivalent to more typical presentations of Lewisian truth conditions to the effect that a counter- factual $\phi > \psi$ is true just in case all the nearest ϕ worlds are ψ worlds. Of course, referencing the indexed accessibility functions we used to characterize von Fintel's theory is unmotivated beyond the purpose of comparison with the von Fintel semantics. This is because, on a Lewisian view, the relevant accessibility function does not change across utterances of counterfactuals (at least not as a matter of semantical rules). That is, $\phi > \psi$ is true just in case all the ϕ worlds nearest to the world of utterance according to $\leq w$ are ψ worlds.

have the two-step procedure of semantic interpretation displayed by von Fintel (2001) at the outset of his paper.

(1) $c\,|\,\alpha\,|\, = c^{l}$
$[\![\alpha]\!]^{c}_{l} = p$

Applying von Fintel's account to Sobel and reverse Sobel sequences is straightforward. Suppose (1a) is uttered at a context in which what it expresses comes out true. Then, given standard assumptions, all the nearest worlds where Sophie went to the parade are worlds where she saw Pedro. None of those worlds are worlds where she got stuck behind a tall person. However, on von Fintel's theory, an utterance of (1b) forces the modal horizon to change in such a way as to include all the nearest worlds where Sophie is stuck behind a tall person because the initial set contained no such worlds. And, presumably, at all of those worlds, Sophie's view of Pedro is obscured. So both (1a) and (1b) are true.

On the other hand, if the order of utterance is reversed, so that (2a/1b) is uttered first, the modal horizon is updated to include some worlds where Sophie went to the parade and got stuck behind a tall person. The resulting modal horizon contains parade worlds where Sophie didn't see Pedro. Consequently, when (2b) is uttered, there are worlds in the modal horizon that satisfy its antecedent while falsifying its consequent.

Von Fintel's proposed semantics has the following two interesting features, both of which he explicitly considers advantages. First, his semantics validates the inference from $(\phi \wedge \psi) > \chi$ to $\phi > \psi$.[8] This is easy to see: if $(\phi \wedge \psi) > \chi$ is uttered truly, this will have the effect of expanding the modal horizon so that it includes a world that satisfies $\phi \wedge \psi$. And that is just what it takes to satisfy $\phi > \psi$.[9] Second, von Fintel's semantics, in virtue of treating subjunctive conditionals as strict, renders the antecedents of subjunctive conditionals as downward entailing contexts (DECs). This, von Fintel claims, helps explain why subjunctive conditionals license Negative Polarity Items (NPIs) such as 'any' and 'even'. DECs have long been thought to hold the key to understanding NPI licensing. We discuss this apparent advantage in section 3.2 below.

2.2 Moss' pragmatic account

As noted, RSSs look problematic for Lewisian theories of counterfactuals. These theories predict that the coordinates of an RSS will be true

[8] Here $\phi > \psi$ is the dual of $\phi > \psi$ in the typical (if controversial) sense that the latter is equivalent to $\neg(\phi > \neg\psi)$. In pseudo-English: the inference from 'If it were that ϕ and ψ, then it would be that χ' to 'If it were that ϕ, then might ψ'.

[9] That is, unless there are no $\phi \wedge \psi$ worlds, but then the context can't satisfy the utterance due to presupposition failure. We discuss this point below.

together just in case the coordinates of its corresponding SS are true together. But that is just what Heim's RSSs put pressure on.[10]

Here is a standard statement of Lewisian truth conditions (with Limit):

Lewis truth conditions $[\![\phi > \psi]\!]\ (w) = 1$ iff $\{w^l : w^l \in [\![>\phi]\!] \wedge \forall w^{ll} \in [\phi],[w^l \leq_w w^{ll}]\} \subseteq [\![\psi]\!]$

Moss (2012) defends the adequacy of Lewisian truth conditions from the challenge posed by RSSs by augmenting them with a pragmatic principle, EI. Moss' formulation of EI is:

(EI) It is epistemically irresponsible to utter sentence S in context C if there is some proposition ϕ and possibility μ such that when the speaker utters S:
(i) S expresses ϕ in C
(ii) ϕ is incompatible with μ
(iii) μ is a salient possibility
(iv) The speaker of S cannot rule out μ (Moss 2012: 568).[11]

(EI) is independently plausible. Further, anyone who likes a broadly Gricean story about a cooperative principle—in particular the maxim of quality—or is sympathetic to knowledge as a norm of assertion should find it plausible that (EI) explains the infelicity of utterances that appear to violate it. Moreover, (EI)'s power to make plausible in-

[10] We have been writing throughout of "Lewisian semantics" rather than "Stalnaker-Lewis semantics" because we are restricting our attention to theories involving accessibility functions from worlds to *sets* of worlds, rather than from worlds to worlds as in Stalnaker (1968). Though of course, for (nearly?) all purposes, special cases of Lewisian semantic theories involving functions from worlds to singleton sets are equivalent to the Stalnakerian machinery.

[11] Notice that (EI) concerns the speaker's epistemic condition and, correspondingly, the felicity of the utterance.

However, Moss claims the following in her endnote 12:

> For simplicity, I talk as if infelicity is a property of utterances. Strictly speaking, infelicity is audience- relative: an utterance sounds infelicitous to an agent insofar as she takes the resulting assertion to be epistemically irresponsible (Moss 2012: 585).

A few points worth taking seriously: First, Moss conflates infelicity with judgments of infelicity. This puts into interesting relief the question of what is our target of explanation—is it infelicity judgments or utterance infelicity? These are not obviously the same.

Second, and relatedly, this raises a meta-puzzle about RSSs: if the infelicity depends on the epistemic status of the speaker, why is it that when we encounter cases like (2) we detect any infelicity at all? Notice that (2a) and (2b) aren't being asserted. (Moss isn't really telling us about Sophie and some parade.) As such, there is no irresponsibility to judge. It's tempting to respond by saying that in considering (2a) and (2b) we imagine someone asserting them and then make the judgment. This may be right, but why do we take our imaginary speaker to have an impoverished epistemic state? It's strange that we aren't more charitable towards our own imaginary conversants. It's worth noting that von Fintel's view has an answer to our question: the dynamics of the system predict the falsity of the sentence at the context, irrespective of the speaker's epistemic state.

felicity predictions applies to utterances in general rather than to the particular case of conditionals. Consider the following example.

(3) a. We're going to have a good time skiing chair 10 at Kirkwood today.
b. But Chair 10 has been closed a lot this week because of high winds.

On the other hand, the following sounds bad

(4) a. Chair 10 has been closed a lot this week because of high winds.
b. # But we're going to have a good time skiing chair 10 at Kirkwood today.

EI facilitates an explanation of what is going on here: When one speaker asserts that Chair 10 has been closing a lot due to high winds, an effect of this is that the possibility that Chair 10 will be closed today becomes salient. We are not in an epistemic position to rule out that Chair 10 is on wind hold today and this possibility is incompatible with having fun skiing lines off of Chair 10. Hence, the infelicity is predicted.

Once we have EI as a plausible constraint on epistemically-responsible assertion, we can utilize it to explain the infelicity of RSSs. Here is the initial RSS again:

(2) (a) If Sophie had gone to the parade and been stuck behind a tall person, she wouldn't have seen Pedro.
(b) # But If Sophie had gone to the parade she would have seen Pedro.

The idea is that asserting (2a) raises to salience the possibility that if Sophie had gone to the parade she might have been stuck behind a tall person. And on Lewis' semantics, this possibility is inconsistent with 2b.[12] Hence, assuming the newly salient possibility can't be ruled out, EI predicts the infelicity. Notice, however, that infelicity is not falsity and Lewis' semantics tells us that (2a) and (2b) are perfectly consistent.

2.3 Scorekeeping

One might think that as far as RSSs go, the score looks tied for von Fintel vs Moss-Lewis. However, von Fintel's view retains some apparent advantages. First, on Moss' account it is somewhat mysterious why uttering "If Sophie had gone to the parade and been stuck behind a tall person, she wouldn't have seen Pedro" should make "If Sophie had gone to the parade, she might have been stuck behind a tall person" a salient possibility. Certainly, EI tells us that when such a possibility is raised to salience and we are not in an epistemic position to rule it out, it is irresponsible to utter something incompatible with this possibility. But

[12] On Lewis' semantics, '$P > R$' is true iff there is an R world among the nearest P worlds. But if all R worlds are $\neg Q$ worlds, '$P > Q$ is false. (Where, again, $\phi > \psi$ is the dual of $\phi > \psi$ in the typical (if controversial) sense that the latter is equivalent to $\neg(\phi > \neg\psi)$.) Moss provides derivations illustrating this for a system corresponding to Lewis' favored semantics (Moss 2012: 569–572).

it doesn't tell us anything about how or under what conditions such a possibility is raised to salience.[13]

On the other hand, von Fintel's semantics gives us a story about how the possibility at issue is relevant that is as explicit as one could want—on von Fintel's theory, a subjunctive conditional of the form $(\phi \wedge \psi) > \chi$ entails $\phi > \psi$.

The second apparent advantage for von Fintel's view is that, on account of treating subjunctive conditionals as strict conditionals, the view can explain the fact that the antecedents of subjunctive conditionals licence negative polarity items. Given that being downward monotone gives rise to the licensing of NPIs, his dynamic approach explains the relevant NPI licensing, since strict conditionals are downward monotone in their antecedents.

The first supposed advantage is actually a disadvantage. Below we offer counterexamples to the entailment at issue. The second advantage is merely apparent as well: Building on work by Anastasia Giannakidou (2006, 2011), we show that being a downward entailing environment is neither necessary nor sufficient for licensing NPIs.

3. Against if to might, and NPIs and SDE

3.1 Against if to might

On von Fintel's semantics $(\phi \wedge \psi) > \chi$ Strawson entails $\phi > \psi$, where *ϕ Strawson entails ψ* iff for every context such that ϕ and ψ are both defined—which requires having their presuppositions satisfied—the truth of ϕ guarantees the truth of ψ.[14] (We'll say that an inference is *Strawson valid* iff its premises Strawson entail its conclusion.) The inference from $(\phi \wedge \psi) > \chi$ to $\phi > \psi$ is Strawson valid because an utterance of $(\phi \wedge \psi) > \chi$ will have the effect of expanding the modal horizon so that it includes a world that satisfies $\phi \wedge \psi$. And that is just what it takes to satisfy $\phi > \psi$. This observation might be thought to weigh heavily in favor of von Fintel's theory and against the Moss-Lewis view.

Unfortunately for the von Fintel-style theorist, there are counterexamples to the entailment claim at issue. Consider the following example:

(5) Context: We are discussing (the fictional) Gandalf the White's awesome power in *The Lord of the Rings*.

[13] This was first brought to our attention by I-Sen Chen. Von Fintel made this point as well in a 2012 talk at the conference *What If* at the University of Konstanz (von Fintel and Gillies 2012).

[14] Strawson gets the honour on account of his application of the concept to sentences such as 'Every F is G'. Strawson argued that while 'Every *F* is *G*' doesn't entail 'Some F is G', it does when the presupposition that there is at least one F is satisfied (Strawson 1950: 343–344). Von Fintel (other paper) defends the view that a Strawson Entailing context licenses NPIs even if it is weaker than the requirement of downward entailment.

a. If Sophie had gone to the parade and had her eyes cursed by Gandalf, then she wouldn't have seen Pedro.
b. # So if Sophie had gone to the parade, she might have had her eyes cursed by Gandalf.[15]

Surely (5a) is true. But is a subsequent assertion of (5b) true or felicitous? We feel certain that (5b) is false since Gandalf doesn't exist. It seems wild to accept (5b) on the basis of an acceptance of (5a).

Of course, there are some philosophical moves available to von Fintel but the quick fixes don't seem to offer the desired succor. An obvious idea is to deny worlds that satisfy the antecedent of (5a) a place within the modal horizon. The modal horizon is there to track possibilities and Gandalf showing up is not a possibility to take seriously. This, in effect, is to mimic part of Moss' explanation though implemented in the semantics. But this seems like a bad move—if the modal horizon comprises possibilities being tracked, then pretty clearly it should sometimes admit worlds where Gandalf curses eyeballs. This can be illustrated by considering cases involving modal subordination. The following sounds fine and seems to be a case of elaborating on possibilities:

(6) a. If Sophie had gone to the parade and had her eyes cursed by Gandalf, then she wouldn't have seen Pedro.
b. Yeah, she would have wandered around terrified and bumping into people. It would have been a disaster.

To make matters worse, there are infelicitous RSSs involving similar possibilities:

(7) a. If Sophie had her eyes cursed by Gandalf but then drank the magical antidote, she would have seen Pedro.
b. # But If Sophie had her eyes cursed by Gandalf, she wouldn't have seen Pedro.

[15] Regarding the Gandalf case, one commentator reports that since the initial context of (5) is underspecified, he wonders if in fact (5b) might sound felicitous when uttered after (5a) in any adequately specified context such that (5a) is felicitous. Some have also worried that contexts involving talk of fictional characters are atypical and might muddy our intuitions. Here is a similar case involving real people and a more detailed context:

(i) Context: Tim lives in India, and that's where he is today. Tim is a funny guy so whenever the authors of this paper see him he makes jokes that we laugh at. There was a time when Tim and the authors of this paper would often meet for beers at a local brewery. The authors of this paper almost met for a beer at that brewery today but decided to stay home and work on this paper via Zoom instead. The following conditionals were uttered in the Zoom meeting.
a. If we had grabbed a beer at the brewery today and Tim had joined us, we would have laughed.
b. # So if we had grabbed a beer at the brewery today, Tim might have joined us.
(ib) sounds bad to us, since we know Tim is in India and couldn't have visited Fieldwork today regardless of whether we had gone there.

(7b) is pretty clearly infelicitous, and, as such, assuming the infelicity stems from the same source as the original RSS, the von Fintelian can't deny Gandalf-worlds places within modal horizons. Since a place within the modal horizon brings along the requisite entailment, a view that explains the infelicity of RSSs in this way inherits the problems of over-generation. There are less silly counterexamples to the entailment claim at issue. Some concern descriptions of empirical hypotheses, their test conditions, and what inferences are warranted.

Consider:

(8) Context: Boyle meant to put a candle in his air pump on Monday but didn't get around to it. On Tuesday he uttered the following:

a. If a vacuum had been created around the candle and it had continued to burn, then my hypothesis would have been confirmed.[16]

b. # So if a vacuum had been created around the candle, it might have continued to burn.

On the relevant (non-epistemic) reading of 'might' (as the dual of 'would'), the argument has a true premise and a false conclusion.[17]

But if von Fintel is correct, the argument is sound. Notice that Moss's EI principle is suggestive of an explanation of what's going on in (5). Take, for example, an utterance of (5b). The hearer of this assertion is in a position to rule out that Gandalf will arrive at any parades attended by Sophie. So, it is natural to expect the infelicity judgment since it is often epistemically irresponsible to make assertions that are inconsistent with a salient possibility. Gandalf's absence from the parade regardless of Sophie's presence is not merely possible, but known. (However, it remains true that *if* Gandalf were to attend a parade and curse someone's eyes, that person wouldn't see Pedro.) Keep in mind, of course, that EI is meant to articulate necessary and not sufficient conditions for infelicity, so while the explanation sits well with the principle, it is no entailment of it.

But what of (8)? Notice that Boyle and his contemporaries had not yet ruled out that the candle would keep burning. Nevertheless the relevant notion of ontic modality expressed by 'might' in (8) seems to make that sentence false as uttered. If we consider that what is important to the truth value of (8) is the set of actual physical and chemical

[16] We gather that Boyle endorsed the phlogiston theory of combustion, and in his time it was an open question whether a flame would continue to burn in a vacuum. At the time some theorists, including Boyle apparently, suspected that in the absence of gas surrounding the candle, Phlogiston particles would be able to escape more easily thus aiding combustion (Cf, e.g., Myers 2003: 20).

[17] Whether a non-epistemic reading of 'might' is the most natural reading or not, it is available. In any case, the 'might' that von Fintel's theory concerns for subjunctive might conditionals is non-epistemic.

laws or regularities, then perhaps Karen Lewis' (2018) account (which we discuss below in 4.1.1) explains this case—simply take what we just called "importance" to be relevance. However, we aren't certain this is a case involving Lewis' intended notion of relevance.[18]

3.2 NPIs and SDE

Von Fintel's view offers an explanation for NPI licensing in the antecedents of counterfactuals. 'any' is a negative polarity item:

(9) #I see any coins.
(10) I don't see any coins.

And 'any' is licensed in the antecedents of subjunctive conditionals:

(11) If I were to see any coins, I would take them. These observations call for explanation.

Bill Ladusaw (1980) famously argued that the key to NPI licensing is downward entailment. Specifically, NPIs are licensed exactly in syntactic positions where subset inferences are valid. Call the scope of such an expression an *environment*. In (10), the second argument position of 'see' is a downward entailing environment. Any substitution into that position with a term whose extension is a subset of *[*coin*]* will preserve truth. However, there are cases of NPI licensing environments that are not downward entailing (see von Fintel 1999 and Giannakidou 2006). For example, 'only' seems to license NPIs but is not downward entailing:

(12) Only Jim ate any beans.
(13) Only Jim ate any beans with mold on them.

The truth of (12) doesn't suffice for the truth of (13) since there may not be any moldy beans.[19] Similar points involve 'regret'. 'regret' licenses NPIs—such as 'ever' and 'any'—but fails to be downward entailing:

(14) Bernard regrets that he ever bought a guitar.
(15) Bernard regrets that he ever bought a polka-dotted guitar.

(14) doesn't entail (15) since Bernard bought a guitar and regretted it, but the guitar has no polka dots.[20]

[18] In any case, our observations have been independent of Karen Lewis' work and our discussion here differs from hers in that we have been considering what the impact of allowing resetting and non-expansion of the modal horizon could be on a von Fintel style semantics. In particular, we have considered how trying to ameliorate the issue on a view like von Fintel's seems to require invoking claims at a pragmatic level that have the unfortunate effect of making modal discourse utilizing counterfactuals too unlike discourse involving other modal vocabulary used to track possibilities.

[19] The claim that 'Only Jim ate some moldy beans' entails that Jim ate some beans is not without its critics. See, McCawley (1993: 312) for a similar example and discussion.

[20] See Giannakidou (2006: 577) for a similar example.

As von Fintel notices, the problem looks to be related to factivity presuppositions. Consider again the 'only' and the 'regret' cases: if the factivity presuppositions are satisfied, then the inferences look good. Accordingly, von Fintel proposed a modification of the Ladusaw approach: NPIs are licensed in environments that are downward entailing when all the con- textual and linguistic presuppositions of all the sentences at issue are satisfied. He (1999) pursues this idea to great effect by formulating and offering up the notion of Strawson entailment.

Why is the foregoing relevant? The antecedents of subjunctive conditionals license (some) NPIs. So his theory nicely predicts NPI licensing in the antecedents of conditionals: von Fintel's semantics renders the antecedents of subjunctive conditionals SDE environments.
Thus, while

(16) If John were taller, he could dunk. doesn't seem to entail:

(17) If John were taller but couldn't jump, he could dunk.

(16) does Strawson downward entail (17).

To see that the inference is Strawson valid we may reason as follows. If (16) is true and its presuppositions are satisfied, then there is either at least one world within the modal horizon at which John is taller and can't jump, or there is no such world. If there is such a world, then John can dunk at it. If there is no such world, then (17)'s presuppositions aren't satisfied so it can't constitute a counterexample to the SDE claim at issue. In either case, the inference is Strawson valid. Thus, von Fintel secures an answer for an otherwise puzzling question: why are NPIs licensed in the antecedents of subjunctive conditionals, given that the antecedents of conditionals aren't downward entailing?

By contrast, Lewis' semantics treats subjunctive conditionals as variably strict, and the antecedents of such conditionals don't validate subset inferences. As such, the Lewisian is left with no obvious explanation of why the antecedents of subjunctive conditionals license negative polarity items.[21]

3.2.1 Does SDE explain NPI Licensing?

In any case, all is not well for von Fintel's own explanation of NPI licensing. The antecedents of conditionals don't license all NPIs. For example, 'yet' appears to be an NPI.

[21] We hasten to add that strict conditionals are downward entailing and Lewis' account is one of subjunctive conditionals as variably *strict*. We think there is room for the Lewisian to offer an explanation of NPI licensing as a function of downward entailment should such a thing be desired. One might begin with the idea that NPI's are licensed in sentential contexts that are "no-variation downward monotone", where a sentential context is no- variation downward monotone just in case if the premise and conclusion of the inference involve quantification over the same set of worlds, then the sentential context is downward monotone. We won't explore this idea further here.

(18) I haven't seen her yet.
(19) #I have seen her yet.

But 'yet' is not licensed in the antecedents of conditionals:

(20) # If I had seen her yet, I would have remembered it.

If SDE contexts license NPIs and conditional antecedents are SDE, it is not clear why 'yet' isn't licensed.

Second, and more importantly, von Fintel's treatment of the antecedents as SDE environments is only as compelling as the explanation of NPI licensing. But the claim that SDE environments are NPI licensing environments is highly controversial. The trouble starts with overgeneration. Recall that a sentence φ Strawson entails ψ just in case ψ is entailed by φ given that all the presuppositions of both φ and ψ are satisfied. Giannakidou (2006) shows that it's not hard to find counterexamples to the claim that SDE environments license NPIs. Consider the following cases:[22]

(21) John is unique in eating a vegetable.
(22) Uniquely, John ate some vegetables.

It is a presupposition of (23) and (24) that someone ate kale (as evidenced, e.g., by the negation test for presupposition).

(23) John is unique in eating kale.
(24) Uniquely, John ate some kale.

Recall that for Strawson entailment the contexts that matter are all and only those in which the presuppositions of *all* the involved sentences are satisfied. Hence, the environment: '*s* was unique in having eaten *x*' is Strawson downward monotone and (21)/(22) Strawson down- ward entail (23)/(24) (respectively). However these environments do not license NPIs such as 'any':

(25) #John is unique in eating any vegetable(s).
(26) #Uniquely, John ate any vegetables.

Similar points attend focus construction:

(27) # JOHN ate any vegetable.

Similar points also attend 'ever':

(28) #It was John who ever ate a vegetable.
(29) #It was John who ever ate a vegetable.
(30) #JOHN ever ate a vegetable.
(31) #Uniquely, John ever eating a vegetable.

Problematically, these sentences are similar to one's involving 'only' which clearly does license NPIs:

(32) Only John ate any vegetables.

The problem, then, is that the explanation of NPI licensing by environments that are Strawson-downward entailing is subject to apparent counterexamples (see Giannakidou 2006).

[22] These cases are ours but inspired by some of Giannakidou's cases involving 'only' and SDE (Giannakidou 2006: 579).

Giannakidou also provides counterexamples involving modals that license some NPIs while not generating SDE environments:

(33) John may order any meal.
(34) Any visitors must register at the front desk.
(35) The search committee can hire anyone they like.
(36) The search committee would interview anyone.

Of course, the environments after the NPIs above do support subset inferences:

(37) John may order any meal. ⇒ John may order any meal on the menu.

But, they aren't SDE:

(38) John may order a meal. ⇏ John may order an expensive meal.

Von Fintel's view is that NPIs are licensed just by environments that are SDE and this forms a key part of his argument that subjunctive conditionals are strict conditionals.

Interestingly, despite licensing occurrences of 'any', one can see that many other NPIs are *not* licensed by modals:

(39) ?John may/must ever go to Paris. (John may not ever go to Paris)
(40) ?John can be all that smart. (John can't be all that smart)
(41) ?The search committee would have much time to consider it. (The search committee wouldn't have much time to consider it.)

By contrast, the antecedents of subjunctive conditionals do license these NPIs:

(42) If John had ever gone to Paris. . .
(43) If John were/had been all that smart. . .

This seems right, but it doesn't remove the mystery: what is the connection between (some) NPI licensing and modality? The claim that the connection is grounded in SDE clearly overgenerates.

Von Fintel claimed that his semantics has the virtue of preserving a powerful explanation of NPI licensing. Namely, that NPI licensing occurs in SDE environments. But if modals are generally licensers of NPIs but not generally producers of SDE environments his semantics lacks the claimed virtue. Moreover, given the widely held view that subjunctive conditionals are modals, one can't defend von Fintel's semantics by suggesting that modals are a special case of NPI licensers that don't generate SDE environments.[23]

[23] It may be that presupposition satisfaction is part of the correct story of NPI licensing, but there is reason to be suspicious that the story involves a mere appeal to Strawson downward entailment. A promising line of explanation treats non-veridical contexts as the key to NPI licensing. A function f is veridical just in case $f(p)$ entails or presupposes p. Negation is thus non-veridical and, most relevant for our purposes, so are the antecedents of subjunctive conditionals. We refer the reader to Giannakidou (2006).

4. Need Pragmatics and True RSSs

4.1 Need Pragmatics

We have argued that the supposed advantages of a strict account of conditionals like von Fintel's are not clearly advantages at all and in fact are problematic. Here we argue that von Fintel's theory utilizes an overly limited range of tools to explain the predicted falsity of (2b). And to substantially improve the empirical coverage of theories like his, one would need pragmatic elements of the sort that Moss already appeals to directly in her treatment.

Consider the following RSS-like sequence:

(44) a. If Sophie had gone to the parade and hadn't been stuck behind a tall-person, she would have seen Pedro.
b. #But, if Sophie had gone to the parade, she would have seen Pedro.

Clearly, (44b) is infelicitous. Also clear is that a strict application of the von Fintel semantics does nothing to predict the infelicity of (44b). After all, the minimal expansion of the modal horizon needed to include a world in which Sophie doesn't get stuck behind a person is no expansion at all. The infelicity remains but the explanation of the infelicity doesn't apply. As such, it looks like the von Fintel explanation (and ones that utilize similar mechanisms) of RSSs is no help in explaining similar sequences—sequences that seem importantly similar to RSSs.

On the other hand, Moss's view handles the case pretty easily: the antecedent raises to salience both the worlds at which Sophie doesn't get stuck behind a tall person as well as ones where she does. How does the antecedent make this happen? Unclear. Perhaps it is just a feature of our psychology that when we hear the possibility $\neg\phi$ mentioned and we can rule neither it or its negation out, both become contextually salient. In any case, whatever the correct story is about the mechanism at work, something like this is what happens.

What are the range of responses for the proponent of a von Fintel-type view? We will consider a few.

First, one might claim that von Fintel was trying to explain Sobel sequences and (44) doesn't comprise such a sequence. So why should von Fintel be on the hook to explain the infelicity of (44)? Our answer can probably be anticipated: A theory that explains likes alike should be preferred to one that does not. Sequences like (44) are clearly a lot like RSSs.

A second, more subtle, response: The purpose of the modal horizon is ultimately to allow participants to keep track of possibilities under discussion. An utterance of (44a) followed by (44b) typically indicates that the speaker wants to consider the contrast between worlds where the parade-going Sophie is not stuck behind a tall person with ones in which she is. Since the speaker indicates a desire to consider such worlds, cooperative partners will take their indirect salience to be rea-

son to include them in the modal horizon. As such, we will have worlds in the modal horizon in which Sophie goes to the parade and fails to see Pedro and as such, (44b) is predicted to be false on von Fintel's view after all.

Notice that the expansion of the modal horizon according to this explanation is not generated by the semantics. But it invokes the very resources that Moss calls upon to explain the infelicity—an appeal to pragmatics. Since these RSS-like sequences seem so similar to RSSs and SSs, this is highly suggestive that something like salience is at the heart of a good explanation of all Sobel-esque sequences.

Third, one might complain that our case is not like an RSS on account of conversational dynamics involving coherence. Plausibly, SSs and our RSSs depend on the discourse relation of contrast–typically marked by an occurrence of 'but' at the start of the second sentence.[24] Moreover, it is reasonable to complain that adding the word 'but' at the start of the second sentence in our RSS-like sequence generates infelicity by trying to mark a contrast that doesn't exist. (Both sentences mention a possibility in which Sophie sees Pedro.)

Fortunately, other cases seem to include the contrast characteristic of RSSs (and SSs) but still allow us to make the point we want to make:

(45) Context: Pedro participated in two parades today, one in the morning and one in the afternoon. Sophie, so far as we know, went to neither. However, we do know she couldn't have gotten out of work long enough to attend both parades but she could have gotten away to attend one of them.
   a. Heidi: If Sophie had gone to the morning parade and been stuck behind a tall person, she wouldn't have seen Pedro.
   b. Lana: # But if Sophie had gone to the afternoon parade, she would have seen Pedro.[25]

[24] This was pointed out to us by I-sen Chen.

[25] Moss offers a similar case in which the information that someone other than Sophie got stuck at the parade behind someone tall is asserted rather than mentioned in the antecedent of a conditional. Moss suggests:

> Consider the following sequence:
>
> (11a) Do you remember when Kate got stuck behind a tall person and missed seeing Pedro in her first baseball parade?
>
> (11b) # But if Sophie had gone to the New York Mets parade, she would have seen Pedro.
>
> (11a) is not a counterfactual. But it nevertheless raises the possibility that if Sophie had gone to the parade, she might have been stuck behind a tall person. My analysis predicts that (11b) is therefore infelicitous. Gillies and von Fintel do not predict this. Since (11a) is not a counterfactual, or even a modal sentence, it does not prompt any expansion of the domain over which counterfactuals quantify (Moss 2012: 578; her numbering).

As Moss points out, von Fintel and other strict conditional theorists may not be too worried about her sort of case, because such cases aren't sequences of counterfactuals. But with our sort of case, which does involve counterfactuals, this move is not available.

The only way we can hear (45b) as felicitous is if we understand its utterer to be communicating, in part, that if Sophie had gone to the second parade she wouldn't have been stuck behind a tall person. The important point for us is that on von Fintel's account the infelicity of the second sentence is not predicted. This is because on his account the antecedent of the first conditional broadens the modal horizon so as to let in worlds at which Sophie goes to the morning parade and is stuck behind a tall person. However, his story does not tell us that the modal horizon is expanded in such a way to include worlds where Sophie gets stuck behind a tall person at the evening parade. One might think that what the antecedent of the first conditional makes salient is some generic possibility in which Sophie gets stuck behind a tall person at some parade, so the modal horizon gets expanded in this way. But that is not predicted by von Fintel's semantics.

We can foresee a response on von Fintel's behalf: One might worry that the semantics von Fintel offers requires expanding the modal horizon to include all the worlds at least as close to ours as the closest antecedent worlds. Some of those worlds may not satisfy the antecedent.[26] So, for example, in the case of (45), perhaps the right thing to say is that the afternoon parade worlds where Sophie is stuck behind someone tall are as close as the morning parade worlds, and so an utterance of (45a) enriches the modal horizon with those worlds as well. In that case, we could explain the infelicity of (45b) by reference to worlds already in the modal horizon that are inconsistent with the truth of (45b). See Arregui (2009). This reply is on to something but it would require us to treat similarity very heavy handedly. Against such heavy handedness, notice that we can clearly setup the context in ways that make Sophie's going to the afternoon parade much more dissimilar to the world of utterance than the morning parade. Moreover, we can change the case to include cases that seem rather dissimilar yet the relevant judgments still obtain:

(46) Context: Sophie and Pedro love parades, but both fear subways and go out of their ways to avoid riding them.
   a. Heidi: If Sophie had gone to the morning parade and been stuck behind a tall person, she wouldn't have seen Pedro.
   b. Lana: # But if Sophie and Pedro had stood near each other on the subway, she would have seen Pedro.

The mere mention of having one's view obstructed by a tall person seems to destroy the felicity of a whole range of subjunctive conditionals somewhat independently of the nearness of the possibilities implicated in the antecedents of those conditionals.

4.1.1 Karen Lewis on the Moss-Lewis View

Recently, Karen Lewis (2017) has offered some criticism of the Moss-Lewis view that we consider below. K. Lewis (2017) offers a hybrid view

[26] Thanks to an anonymous reviewer for emphasizing this to us.

of sorts, according to which RSSs are inconsistent, but this has nothing to do with any modal horizon because her semantics for counterfactuals involves an ordering on possible worlds as a function of two distinct proper- ties of worlds: similarity and relevance. As with D. Lewis' view, the semantics she offers is one according to which $\phi > \psi$ is true iff every ϕ world nearest to the actual world is a ψ world. But when a world is relevant enough to conversational purposes this may have the consequence that the world is nearer to the actual world than some more similar but less relevant worlds. For discussion of K. Lewis' notion of relevance and how it interacts with salience and similarity, we refer the reader to K. Lewis' 2016 and 2017.

We think K. Lewis' account is fine as far as it goes, though we worry about how relevance may interact with other semantic phenomena, such as modal subordination, where salience, not relevance seems to hold the key to which worlds are tracked. (A possibility may be made salient yet simply be so irrelevant to conversational purposes that no world that realizes the possibility is ranked among the nearest worlds.) In any case, given that salience doesn't suffice for relevance, the entailment doing the explanatory work for von Fintel is absent on K. Lewis' view. We take this to be a distinguishing feature between von Fintel–style semantic views and that class of views we group with the Moss-Lewis approach because they share a reliance on features of pragmatics to explain the infelicities at issue.

We feel obliged to offer some response to some of Karen Lewis' recent (Lewis 2017) criticisms of the Moss-Lewis view. K. Lewis' criticisms are fourfold:

1 Moss cannot give a theory-neutral account of what gets raised to salience (Lewis 2017: 23).
2 Moss cannot account for infelicitous Heim sequences involving possibilities not among the most similar worlds satisfying the antecedent (Lewis 2017: 12, 23).
3 In predicting what and how many possibilities are raised to salience in various cases, Moss' view may become too complicated (Lewis 2017: 15, 23).
4 Unlike K. Lewis' view and the Fintelian view, Moss' view has difficulty handling retraction sequences (Lewis 2017: 15, 23).

We find problems 1., 2., and 4 to be the most pressing and will consider them in turn.

Regarding the first problem, recall that Moss' view is that for an RSS having the form

(n) a. $(P \wedge R) > \neg Q$
b. $P > Q$

b. is infelicitous if a. raises to salience a possibility that cannot be ruled out and that is incompatible with b. The problem is that which possibilities are inconsistent with a. is sensitive to which semantic theory of subjunctive conditionals is correct. For example, positing that $P >$

R is raised to salience works for D. Lewis but not Stalnaker, and vice versa for $P > R$. And positing that both are raised to salience works for Stalnaker but not for D. Lewis (Lewis 2017: 13).

Accordingly, K. Lewis suggests that the best option for the proponent of Moss' view is to hold that what gets raised to salience is that there is a $P \wedge R$ world among the nearest P-worlds (which works whether Lewis' or Stalnaker's semantics is correct). But this move, she argues, saddles Moss with the second problem.

Regarding the second problem, K. Lewis offers the following case:

(47) Context (quoting K. Lewis 2017: 14): Suppose Logan and Claudia are in the same group of friends, and so often end up at the same social events. But Logan and Claudia notoriously do not get along, and when they are together, it is absolutely no fun to be around them (though each on his or her own is a very fun person). The conversational participants are discussing a recent party which neither Logan nor Claudia attended because they both had work to do. While Logan is very attentive—he is unlikely to miss a deadline if he can help it—Claudia is easily distracted from her work and almost went to the party, only to be swayed at the last second not to go by her guilty conscience.
   a. If Logan and Claudia had come to the party, it would have been no fun at all.
   b. # But of course, if Logan had come to the party, it would have been fun.
   c. # But of course, if Claudia had come to the party, it would have been fun.

K. Lewis' analysis is that b. and c. are infelicitous but that the Moss view cannot explain the infelicity of c. This is because we know it was always unlikely that Logan would have attended. So we know that if Claudia had attended, Logan would not have. Hence, EI cannot explain the infelicity here because we can rule out that Logan and Claudia would have both been at the party.

We believe the case does not demonstrate K. Lewis' conclusion. The problem is that the context gives us by stipulation that it is unlikely that Logan would attend. But unlikeliness (in any of the many ways of precisifying the notion) does not make for greater distance from the actual world. Suppose a die will be rolled and that you have taken a bet that it will come up six. Observe (the fairly well-known point) that just because it is unlikely you will win, this does not mean that worlds where you win are farther away than worlds where you don't. Rather, the correct judgment seems to be that exactly one-sixth of the most similar worlds are worlds where you win. The unlikeliness of an event does not on its own make worlds where that event occurs more distant.

Notice further that if we change the context to read as follows, then 47c sounds fine:

(48) Context (compare K. Lewis 2017: 14): Suppose Logan and Claudia are in the same group of friends, and so often end up at the same social events. But Logan and Claudia notoriously do not get along, and when they are together, it is absolutely no fun to be around them (though each on his or her own is a very fun person). The conversational participants are discussing a recent party which neither Logan nor Claudia attended because they both had work to do. Logan is very attentive to work deadlines and the possibility that he would attend is ruled out. On the other hand, Claudia is easily distracted from her work and almost went to the party, only to be swayed at the last second not to go by her guilty conscience.

a. If Logan and Claudia had come to the party, it would have been no fun at all.

b. # But of course, if Logan had come to the party, it would have been fun.

c. But of course, if Claudia had come to the party, it would have been fun.

Keep in mind that if K. Lewis' claim is correct that we and the conversational participants could rule out Logan's attendance all along, then the language in our modified context should make no difference to the relevant felicity judgments. But it does.

Turning to the fourth problem now, K. Lewis offers the following example of a *retraction sequence*:

(49) a. A: If Sophie had gone to the parade, she would have seen Pedro dance.

b. B: But of course, if Sophie had gone to the parade and been stuck behind someone tall, she wouldn't have seen Pedro dance.

c. A: Alright, I guess then, if Sophie had gone to the parade, she might not have seen Pedro dance.

The sequence is felicitous, and as K. Lewis puts it, "At first, this seems like just the sort of sequence that (EI), or at least something in its spirit, is poised to account for. According to the account, [the above sequence] raises to salience the counterfactual *If Sophie had gone to the parade, she might have been stuck behind someone tall*. Since this makes [(49a)] unassertable in this context, given that it conflicts with a possibility raised to salience, it would make sense that a speaker might go back and amend or retract her initial assertion, even though it was perfectly good in the original context" (Lewis 2017: 16 [brackets added]). The problem that K. Lewis brings up is that on D. Lewis' analysis 49c has the logical form $P > \neg Q$, and on his analysis this is logically equivalent to $\neg(P > Q)$. But then 49a is logically inconsistent with 49c.

Since 49a and 49c are both true by hypothesis, K. Lewis' example seems to have the consequence that either the D. Lewis analysis of

might counterfactuals is false or the Moss-Lewis view has the unpalatable consequence that there are very many felicitous but false might counterfactuals like 49c.

Whether or not the theoretical cost of positing lots of felicitous and false sentences is too high, we see little reason why a proponent of the Moss-Lewis view cannot simply accept that utterances of might conditionals in natural language are very often syntactically ambiguous between D. Lewis' favored analysis and the Stalnaker analysis according to which *If P then might not Q* is equivalent to saying *It is epistemically possible that if P then not Q*. And we think that the context surrounding K. Lewis' retraction sequence heavily suggests an epistemic reading of the occurrence of 'might'.

In sum, we quite like K. Lewis' account of counterfactuals and what it tells us about the effect of relevance on their truth values. However, we think that while the criticisms she offers of the Moss-Lews view are interesting and worth thinking through carefully, we do not think they ultimately show the Moss-Lewis view to be in trouble.

4.2 True RSSs

The previous section aimed to show that a proper explanation of RSS-like phenomena in general involve a pragmatic dimension. One might worry that this leaves open whether von Fintel has the correct story about RSSs in particular. In this section, we consider RSSs directly and argue, pace von Fintel's semantics, that their coordinates are not semantically inconsistent.

We start from the following case involving an RSS.

(50) Context: Ours is a world at which Sophie did in fact go to the parade this morning, and she saw Pedro there. However, neither Heidi or Lana know this. Part of their conversation proceeds as follows.
   a. Heidi: If Sophie had gone to the parade this morning and had her view (of Pedro) blocked by a tall person, she wouldn't have seen Pedro.
   b. Lana: #But if Sophie had gone to the parade this morning, she would have seen Pedro.

Lana's utterance is infelicitous in context. However, even a counterfactual skeptic like Al Hájek will accept the truth of both of these subjunctive conditionals (since (50a) has an antecedent that entails its consequent and (50b) has a true antecedent and consequent—i.e., it is a true-true subjunctive conditional. Many philosophers accept Strong Centering, and that principle guarantees the truth of true-true subjunctive conditionals.[27]

But Strong Centering is ultimately inessential to the issue here—all that would be needed to show that the Fintelian semantics is incorrect

[27] By *Strong Centering* we mean the inference principle: *(SC)* $\phi \wedge \psi \Rightarrow \phi > \psi$.

would be a demonstration of the existence of some true-true subjunctive conditionals like (50b) that are true and for reasons not defeated by utterances of conditionals like (50a). But on von Fintel's view, an utterance of (50b) is inconsistent, so *cannot* be true as uttered. But we think it is pretty clear that if Heidi and Lana were to go on to disagree about the truth of Heidi's utterance, we could easily settle the issue in favor of Heidi's side by pointing out that we were at the parade with Sophie and we know she saw Pedro.

Now, it still it remains open to the Fintelian to simply hold on to the claim that no RSS can have simultaneously true coordinates. But here are two more, we think very unpalatable consequences of this insistence:

First, if von Fintel's semantics is correct, we get a surprising disanalogy between subjunctive conditionals and other modals. Uttering a subjunctive conditional when it is known that the antecedent is true typically generates infelicity. [28]

People very rarely utter subjunctive conditionals when they know that both antecedent and consequent are true and such utterances typically generate infelicity. But notice how similar this looks to what happens to utterances of other modal expressions in analogous circumstances:

(51) Context: We are at a party and have just both observed that Jordan is in attendance. Then I say one of the following:
  a. Jordan might attend the party tonight.
  b. Jordan should at least make an appearance at the party tonight.

It is certainly odd for a speaker to say that someone *might* or *should* attend a party when participants in the conversation know the person has attended, and the oddness here is similar to the oddness of uttering a conditional in the subjunctive mood when conversational participants know that the antecedent is true. However, despite the oddness of (51a) and (51b), the natural judgment, we think, is that it is perfectly true that Jordan might attend the party tonight given that he is in fact in attendance tonight, and it may well be true that Jordan should attend regardless of whether I say anything about that. The point is that the Fintelian must claim that unlike in cases involving other modals, analogous utterances of subjunctive conditionals cannot be true when uttered in an RSS.

Further, if we avail ourselves of rigidifying operator we can construct sequences of conditionals that come out true on von Fintel's semantics, but which appear to generate infelicity in exactly the way that RSSs do: Let '*Actually*' be an operator that shifts the circumstance of evaluation to the actual world of utterance (in a manner something like that introduced in Kaplan (1989)). Now consider the following variant of (50):

[28] But see Iatridou (2000) for discussion.

(52) Context: Ours is a world at which Sophie did in fact go to the parade this morning, and she saw Pedro there. However, neither Heidi or Lana know this. Part of their conversation proceeds as follows.
   a. Heidi: If *Actually* (Sophie had gone to the parade this morning) and Sophie had been stuck behind a tall person, She wouldn't have seen Pedro.
   b. Lana: #But if *Actually* (Sophie had gone to the parade this morning), *Actually* (she would have seen Pedro).

More colloquially:

   a. Heidi: If Sophie had actually gone to the parade this morning and she had gotten stuck behind a tall person, she wouldn't have seen Pedro.
   b. Lana: #But if Sophie had actually gone to the parade this morning, she actually would have seen Pedro.

Notice that the first conditional is non-vacuously true because its first conjunct is necessarily true and every world in the modal horizon that satisfies the second conjunct also satisfies the consequent. And the second coordinate is true because both antecedent and consequent are necessarily true given that Sophie actually went to the parade and saw Pedro. However, despite the fact that the Fintelian semantics tells us both sentences of 52b are true, the sequence looks like it generates infelicity *in exactly the way* that "real" RSSs do.

Of course it is open to the Fintelian to deny that such an operator could be added to English (not to mention impossible that such an operator already is already present in our language). But that should look implausible to anyone who isn't already committed to the Fintelian picture. Again, our point is not that the Fintelian is unable to hold on to their view in the face of the above considerations, but rather that for anyone with an open mind, the above considerations show some serious drawbacks of the view and should count heavily against it.

We also wish to point out that our (50) is somewhat different than the following case presented by Moss:

(53) Context (quoting Moss 2012: 574): Suppose John and Mary are our mutual friends. John was going to ask Mary to marry him, but chickened out at the last minute. I know Mary much better than you do, and you ask me whether Mary might have said yes if John had proposed. I tell you that I swore to Mary that I would never actually tell anyone that information, which means that strictly speaking, I cannot answer your question. But I say that I will go so far as to tell you two facts:
   a. If John had proposed to Mary and she had said yes, he would have been really happy.
   b. But if John had proposed, he would have been really unhappy.

We think that a proponent of von Fintel's view has a reasonable explanation of what is going on in (53). Let's look more closely, then, at von Fintel's claim that the second sentence of an RSS will be false and why he thinks that the truth conditions of such sentences are sensitive to context in this way.

Von Fintel writes:

> What I mean by "no longer true" is not that the objective facts have changed. It is the parameters of the discourse that have changed so that the proposition expressed by the first counterfactual in the initial context can no longer be ex- pressed by the same linguistic expression in the new context. Compare the fact that the claim that France is hexagonal may be true in a context where it is preceded by Italy has the shape of a boot, but may cease to be true in a later context where the standards of precision have been sharpened (von Fintel 2001: 146, note 8).

The Fintelian can say that the utterance of (53b) after (53a) *is* inconsistent and, hence, surprising.[29] This effect of surprise is capitalized upon in order trigger the inference on the part of the listener. Assuming the speaker is taken to be reliable and in this instance committed to the activity of cooperative communication, the hearer must perform a pragmatic repair on the modal horizon so as to make the second utterance come out true, and the only way to do this consistent with (53a) is to kick all the worlds out of the horizon where John asks and Mary accepts. This is a pragmatic explanation to be sure, but it is precisely because the utterance of (53b) was initially inconsistent, as predicted by von Fintel, that the oddity of the utterance was striking enough to spur the pragmatic repair and the intended inference. This looks okay for von Fintel's view.

Our case involving Lana and Heidi is different. In our case, the second coordinate remains infelicitous by the lights of Heidi and any other listener unaware that Sophie did in fact go see Pedro. So the proponent of von Fintel's theory can't say that it is some accommodation effect making the second RSS coordinate ultimately come out true. Rather, Sophie went to the parade and saw Pedro so, contra von Fintel, both (50a) and (50b) are true together despite the generated infelicity.

This still leaves the questions of why Lana's utterance is infelicitous in context. However, the Moss-Lewis view supplies a story that we think should be counted among the best candidate explanations: Lana's utterance is infelicitous because it runs afoul of the EI principle: The (nonactual) possibility of being stuck behind a tall person should Sophie have attended the parade was brought to salience by Heidi's utterance of (50a), and neither Heidi or Lana were in an epistemically appropriate position to rule that possibility out. And the Moss-Lewis style explanation of the infelicity generated in 52b is the same, which is exactly what a theorist should desire.

[29] Moss describes such an explanation (Moss 2012: 578–579).

5. Conclusion

In this paper we considered the case for von Fintel-style semantic theories of subjunctive conditionals. A great deal hangs on this as von Fintel's semantics (and ones like it) recommend a fairly non-conservative approach to the semantics of subjunctive conditionals. Given the generally agreed upon links between subjunctive conditionals and modals, the departure is most likely not limited to subjunctive conditionals (as discussed in Section 3.1).

The motivation for von Fintel-style theories comes in large part from intuitions regarding the truth values of conditionals in a sequence. There are long standing questions in the philosophy of language regarding our methodology: when and how far can we trust our ability to separate truth value judgments from judgments of felicity? After all, trained philosophers know there is a difference between falsehoods and infelicitous truths. Given that we are good at this, why can't we settle the dispute more directly by intuiting whether or not the second sentence of an RSS is true? If it's false then it is hard to see how a view like von Fintel's doesn't win the day. If it's infelicitous but true, it looks like views like von Fintel's are basically non-starters. It's downright odd that we have to try to settle these disputes indirectly. But that's where we are. And to that extent, we think the case for von-Fintel semantics is pretty weak and the case against it pretty strong.

References

Arregui, A. 2009. "On similarity in counterfactuals." *Linguistics and Philosophy* 32 (3): 245–278.

Briggs, R. 2012. "Interventionist counterfactuals." *Philosophical Studies* 160: 139–166.

von Fintel, K. 1999. "NPI licensing, Strawson entailment, and context dependency." *Journal of Semantics* 16: 97–148.

von Fintel, K. 2001. Counterfactuals in a dynamic context." In M. Kenstowicz (ed.). *Kan Hale: A Life in Language*. Cambridge: MIT Press, 123–152.

von Fintel, K. 2012. "Subjunctive conditionals." In D. Graff Fara and G. Russel (eds.). *The Routledge Companion to the Philosophy of Language*. New York: Routledge, 466–477.

von Fintel, K. and T. Gillies. 2012. "Should von Fintel and Gillies be mothballed?" Invited talk at conference "What if? On the Meaning Epistemology, and Scientific Relevance of Counterfactual Claims and Thought Experiments", University of Konstanz. Konstanz Germany. October 27, 2012. Slides retrieved from http://web.mit.edu/fintel/fintel-gillies-2012-mothball-konstanz.pdf on 4/22/2018.

Giannakidou, A. 2006. "Only, emotive factive verbs, and the dual nature of polarity dependency." *Language* 82 (3): 575–603.

Giannakidou, A. 2019. "3. Negative and positive polarity items". In P. Portner, C. Maienborn and K. von Heusinger (ed.). *Semantics - Sentence and Information Structure*. Berlin, Boston: De Gruyter Mouton, 69–134.

Gillies, T. 2007. "Counterfactual scorekeeping." *Linguistics and Philosophy* 30: 329–360.

Hiddleston, E. 2005. "A causal theory of counterfactuals." *Noûs* 39 (4): 632–657.

Iatridou, S. 2000. "The grammatical ingredients of counterfactuality." *Linguistic Inquiry* 31 (2): 231–270.

Kaplan, D. 1989. "Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals." In *Themes from Kaplan*. Oxford: Oxford University Press.

Kaufmann, S. 2017. "The limit assumption." *Semantics and Pragmatics* 10 (18): 1–29.

Ladusaw, W. 1980. *Polarity Sensitivity as Inherent Scope Relations*. New York: Garland.

Lewis, D. 1973a. "Causation." *Journal of Philosophy* 70: 556–67.

Lewis, D. 1973b. *Counterfactuals*. London: Blackwell Publishing.

Lewis, K. 2016. "Elusive counterfactuals." *Noûs* 50 (2): 286–313.

Lewis, K. 2018. "Counterfactual discourse in context." *Noûs* 52 (3): 481–507.

McCawley, J. 1993. *Everything That Linguists Have Always Wanted to Know about Logic. . . But Were Ashamed to Ask* (2nd ed.). Chicago: University of Chicago.

Moss, S. 2012. "On the pragmatics of counterfactuals." *Noûs* 46 (3): 561–586.

Myers, R. 2003. *The Basics of Chemistry*. Santa Barbara: Greenwood Publishing Group.

Nichols, C. 2017. "Strict conditional accounts of counterfactuals." *Linguistics and Philosophy* 40 (6): 621–645.

Nute, D. 1980. *Topics in Conditional Logic, Volume 20*. Dordrecht: D. Reidel Publishing.

Pollock, J. 1976. *Subjunctive Reasoning*. Dordrecht: Reidel.

Stalnaker, R. C. 1968. A theory of conditionals. In R. Stalnaker (ed.). *Studies in Logical Theory*. London: Blackwell Publishing, 98–112.

Starr, W. 2019. "Counterfactuals." In E. N. Zalta (ed.). *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition). https://plato.stanford.edu/archives/spr2019/entries/counterfactuals

Strawson, P. F. 1950. "On referring." *Mind* 59 (235): 320–344.

Croatian Journal of Philosophy
Vol. XXV, No. 75, 2025
https://doi.org/10.52685/cjp.25.75.7
Received: September 11, 2024
Accepted: November 8, 2024

Easy Does It: Unnsteinsson on Saying and Gricean Intentions

INDREK REILAND*
University of Vienna, Vienna, Austria

This paper critically examines Unnsteinsson's Collapse Argument, which contends that "Easy" views of saying something or expressing a proposition collapse into the Gricean view (Unnsteinsson 2022: Ch. 4). Easy views maintain that saying/expressing is simply a matter of uttering a sentence with its meaning, without requiring Gricean communicative intentions. Unnsteinsson argues that Easy views must appeal to such intentions to explain what makes saying/expression intentional and rational and that this collapses them into the Gricean view. I show that this argument fails for several reasons. First, the intentions that the Easy views must posit to explain what makes saying/expressing rational are not equivalent to the Gricean communicative intentions. Second, the constitutive question of what makes an act into a saying/expressing and the rationalizing question of what makes it rational are distinct. Thus, even if Easy theorists would have to appeal to something like Gricean communicative intentions in answering the latter question, this wouldn't cause their answer to the former question to collapse into the Gricean answer.

Keywords: Language; meaning; saying; intentions; Grice; Unnsteinsson.

1. Introduction

What is it, in uttering a sentence, to say something or express a proposition? Elmar Unnsteinsson's recent book *Talking About: An Intentionalist Theory of Reference* involves an extended argument against views which take this to be, in his words, "easy" (Unnsteinsson 2022: Ch. 4).

* I want to thank Alex Radulescu, Eliot Michaelson, Elmar Unnsteinsson, Rory Harder, and the audience at the Philosophy of Language and Linguistics conference at the Interuniversity Center in Dubrovnik in September 2023 for helpful comments and discussion. This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/PAT1233624.

On such *Easy* views, saying/expressing is a matter of uttering a sentence that has a meaning in a language while participating in a convention, being subject to a rule, or expressing a thought. It doesn't involve Gricean communicative intentions, intentions to produce effects in a particular addressee. In contrast, Unnsteinsson aims to argue that saying/expressing something isn't easy and must involve Gricean communicative intentions. The reason is that he thinks that any *Easy* view will have to posit something like Gricean communicative intentions to explain what makes saying/expressing intentional and rational, and therefore collapses into the Gricean view.

My aim in this paper is to discuss Unnsteinsson's *Collapse Argument* and show that it fails for several reasons. I will first demonstrate that it depends on two unargued assumptions that some *Easy* theorists don't grant, namely that saying/expressing is a necessarily intentional action and one that must furthermore be intended to have effects on a particular hearer. I will then show that even if we go along with these assumptions, the *Easy* views still do not collapse into the Gricean view. First, the intention that the *Easy* theorists would have to posit in explaining what makes saying/expressing intentional and rational is one of intending for the hearer to come to believe that one has *said* that p/*expressed* the proposition that p. Such intentions, while structurally similar to Gricean communicative intentions, are semantically infused in a way that the latter aren't. Second, and fundamentally, Unnsteinsson's argument runs together the *constitutive* question what it is to say/express with the *rationalizing* question of what makes saying/expressing intentional and rational. But, as I will demonstrate at length, these are distinct questions. Even if the *Easy* theorists would have to posit something like a Gricean intention to explain what makes saying intentional and rational, this still wouldn't cause their view of saying/expressing to collapse into the Gricean view. Thus, the *Collapse Argument* argument leaves the *Easy* views of what is it to say/express completely untouched.

I will proceed as follows. First, I'll reframe the issue in a way that should be acceptable to all parties in the dispute (Section 1). Next, I'll discuss Unnsteinsson's *Collapse Argument* and show that the intention that the *Easy* theorists would have to posit to explain what makes saying/expressing intentional and rational is not equivalent to a Gricean communicative intention (Section 2). Finally, I'll discuss the fundamental problem with the argument, namely that it conflates constitutive and rationalizing questions about language use and demonstrate why this shouldn't be done (Section 3).

2. Reframing the issue

To start, we have to make sure that we frame the issue in a manner acceptable to all parties to the dispute. Unnsteinsson frames it in terms of the speaker's *meaning* something:

> Cicero was an orator. Normally in making such an utterance part of the speaker's or writer's goal will be to *mean* something, in this instance to mean something that is true if and only if Cicero really was an orator and false otherwise. Speaking truly (falsely) about Cicero involves *meaning* something by one's utterance such that its truth (falsity) depends on the state of a worldly object, namely Cicero. (Unnsteinsson 2022: 75).

He then introduces the *Easy* views which he aims to argue against:

> There is a long and powerful tradition of theorists who argue that meaning something by what one utters is easy. It is easy in the sense that the speaker is not required to intend the utterance to produce an effect in some addressee. Rather, by uttering something, speakers can simply voice their thoughts or, even, undergo episodes of thinking those thoughts. ... I will call this tradition 'expressionism'. … Arguably, the rallying cry of easy meanings is routine in other quarters as well. Conventionalists would say that all the speaker needs to perform a meaningful speech act is to participate in a convention of a specific type (e.g., Lepore and Stone 2015). Normativists would say that the speaker only needs to be subject to a norm (e.g., Brandom 1994). (Unnsteinsson 2022: 75–76).[1]

In contrast, Gricean intentionalists think that meaning something isn't as easy as simply voicing their thoughts, participating in a convention, or being subject to a norm, but rather requires Gricean intentions to have effects on a particular addressee.

Framing the issue in terms of the speaker's *meaning* something is problematic. It's clear that what is fundamentally of interest to Unnsteinsson is speaking truly or falsely at all (Unnsteinsson 2022: 75). He calls this 'meaning something', but this departs both from the pre-theoretical use of 'to mean' and the Gricean notion of 'speaker meaning' both of which are a matter of attempted communication. This is why those who hold *Easy* views typically deny that speaking truly or falsely is the same as the speaker's meaning something, either in a pre-theoretic sense or in the Gricean sense, where the latter is instead thought to be a matter of communication that goes beyond language use (Bach and Harnish 1979, Bach 2001). In other words, if the issue is framed in terms of 'meaning something', those in the *Easy* camp could simply insist that speaking truly or falsely is easy while granting that meaning something requires Gricean intentions. But this wouldn't satisfy Unnsteinsson, who is interested in the former.

It is therefore better to frame the issue directly in terms of speaking truly or falsely, that is, in terms of a notion like saying that p or expressing the proposition that p.[2] Thus, what we'll take to be at issue

[1] Note that these versions of the *Easy* view aren't necessarily in competition with each other. There are *Easy* views which are simultaneously expressionist, conventionalist, and normativist. For example, one could think that to say/express is to use a sentence with its meaning, which is to be subject to a conventionally accepted rule, as a result of which one counts as expressing a thought (Reiland 2023b, 2025a, Schroeder 2008). For recent explicit defenses of *Easy* views see Fisher 2024, Michaelson 2022, and Reiland 2023a, 2023b, 2025a, 2025b.

[2] The choice between these two ways of putting it depends on one's view of mood. It is natural to think that in uttering 'p' with its meaning one *says* that p (Austin

is the question what is it, in uttering 'p', to say that p or express the proposition that p. On *Easy* views this doesn't involve Gricean communicative intentions, which, if at all, are taken to be relevant only to the further act of speaker meaning or attempted communication. In contrast, Unnsteinsson thinks that saying something or expressing a proposition must involve such intentions.

3. The Collapse Argument

Unnsteinsson's main argument against *Easy* views is that they collapse into the Gricean view. He directly targets only what he calls "expressionism" but thinks that his argument can be extended to other versions of the view (Unnsteinsson 2022: 76). Here's his construal of his target:

Expressionism: *S* says that *p* /expresses the proposition that *p* by uttering *U* if and only if *U* expresses *S*'s thought that *p* (Unnsteinsson 2022: 81).[3]

This schematic view is compatible with different ideas about what it is, in uttering an expression, to express a thought. For example, one could take expression to be a causal relation (Rosenthal 1986), an intentional relation (Davis 2003, Green 2007) or a conventional or normative relation (Alston 2000, Kaplan MS, Reiland 2023b, 2025a, Schroeder 2008). Unnsteinsson's argument doesn't depend on any particular theory of expression so we can abstract away from this, for the time being.

Here is his preliminary presentation of the *Collapse Argument*:

> In a nutshell, the argument against expressionism is that expressing a thought—the expressionist's basic theoretical posit—must be construed as an intentional action on all fours, requiring explanation in terms of the speaker's beliefs, desires, and intentions. If so, I argue, the act of intentionally expressing a thought collapses into the act of uttering something with the intention to mean something to someone. Thus, the expressionist fails to carve out an intentional act different from the Gricean act of speaker meaning. (Unnsteinsson 2022: 76)

I will argue that this argument fails for several reasons. However, to see this, we need to work through it slowly.

The first step of the *Collapse Argument* is to claim that the act of expression must itself be an intentional action and thus must be explained in terms of an intention to have some effect on the actual world:

1962: VIII). However, saying seems forceful in a minimal, representational sense: to say that p is not just to present the proposition as an object, but to present it as true (Reiland 2024). Those who take mood to encode just content and no force at all therefore typically prefer putting their view by saying that in uttering 'p' with its meaning one *expresses* the proposition that p which is just to present the proposition as an object. Since this difference won't matter for us here, we will leave this open and put things in terms of say/express. For discussion see Reiland 2024.

[3] I've replaced 'means' with 'says'. It would also be more accurate to put the view by saying that it's not *U*, an expression-type, but *S*'s utterance of *U*, an act, that expresses *S*'s thought that p.

> ...if the label 'expressing a *p*-thought' names anything of interest to a theory of meaning, it must name some person's rational, intentional action, which is normally explained in terms of that person's intention to achieve some perceived good or benefit. It follows that this person (*S*) intends, minimally, to have some *p*-thought-involving effect on the actual world. (Unnsteinsson 2022: 84)

Here Unnsteinsson makes an unargued assumption that expression is a necessarily intentional action. Some *Easy* theorists might already balk at this and argue that at least some linguistic expression could be unintentional, but expression nonetheless.[4] But let's grant this step for now.

The second step of the argument is to claim that the intention to have effects on the actual world must be an intention to directly have effects on minded creatures, more specifically, a particular addressee or hearer:

> Let's use the label 'T_p' for *S*'s *p*-thought. So, to explain *S*'s behavior we would normally postulate some T_p-involving intention-state which guides and controls her action. … If a speaker *S* intends to have T_p-related effects on the actual world those effects must be intended as effects on minded creatures. ... *S* will only have intentions to cause T_p-effects in creatures which S takes to have the capacity to be influenced in the T_p-way by her intentional action. Otherwise we have no rational explanation of her action. *S*'s intentional act of expressing T_p by uttering something can be labeled '$E_s(T_p)$'. So, what we are saying is that *S* would not, if *S* is a competent speaker, form the intention to express T_p at all, unless some minded creature —for all we have said, it will be possible that $S = H$ —is supposed by *S* to be influenced by $E_s(T_p)$ in a cognitive way. (Unnsteinsson 2022: 84–85)

Here Unnsteinsson makes an unargued jump from the claim that the relevant intention is an intention to have effects on the actual world to the claim that it must therefore be an intention to have direct effects on a particular addressee or hearer. Many *Easy* theorists would balk at this and argue that the intention could be to have certain effects that do not directly mention a particular addressee or hearer. For example, the intention might be simply to externalize the thought or record information (Chomsky 1975: 55–77). Or the intention might be to have an effect on the public conversational record (Camp 2018, Lepore and Stone 2015). But again, let's grant this step for now.

[4] For example, consider people with Tourette's syndrome, some of whom are subject to vocal tics, consisting of utterances of words and sentences. Or consider automated announcements, utterances made by chatbots, and LLMs etc. It's not clear that all of these utterances amount to meaningful utterances and thus to acts of expression. But if they do, it's at least an open question whether they are intentional actions. Thanks to Alex Radulescu for discussion.

It is also important to realize that even if it turns out that expression is always an intentional action, it doesn't immediately follow that what one wants to express is what one ends up expressing. Many, if not most *Easy* views deny this, claiming that a speaker who is mistaken about the linguistic meaning of an expression might use it meaningfully, wanting to say one thing, but ending up saying something else (Burge 1979, Dummett 1986, Fisher 2024, Reiland 2023a, 2025b).

The final step of the argument is to claim that the relevant intended effect on a hearer that the expressionist must posit is equivalent to the sort of effect on a hearer involved in Gricean communicative intentions. Discerning the exact argument for this is not straightforward, so let's walk through the steps of Unnsteinsson's reasoning slowly:

> So, finally, we can safely conclude that there will be some H which S expects to be capable of recognizing or inferring that the act of $E_s(T_p)$ was performed with the intention to produce T_p-related cognitive effects in H. … From this point on it is easy to derive several interesting consequences about the cognitive state of someone like S. First, in performing an act of expressing a thought, S will expect some H to be able to infer from the act that it was performed with an intention to have a cognitive effect on H. This intention is a propositional attitude of some kind. If S forms the belief that a particular propositional attitude, say $M(p)$, where M stands for 'meaning,' is the easiest and most likely one to be immediately inferred by H, S can only be rational in performing the act if S really expects H to infer that the act was performed by someone in the $M(p)$-state. Second, if $M(p)$ is indeed the most immediate inference by S's own lights and S would have no reason to engage in expressive behavior other than in order to have cognitive effects on a minded creature, it follows that the $M(p)$-effect is the cognitive effect S primarily intends to produce, if she (S) is rational. She may intend to produce many other cognitive effects, for sure, but as we have set things up, $M(p)$ is comprised of a mental attitude toward a proposition with truth conditions, such that it is believed by the speaker to be her most easily and immediately inferable mental state. On most accounts of communication, this is simply what an act of attempted communication consists in. Trying to communicate some message is to produce a signal from which the recipient can grasp the message. An important part of human communicative messages is standardly thought of as a proposition or attitude to a proposition and, so, if a speaker takes it that a given propositional attitude is most likely to come to a hearer's mind on perceiving the signal, producing the attitude will, under normal circumstances, be the primary intended purpose of producing the signal in the first place. Communication succeeds when the hearer grasps the intended propositional attitude or, perhaps, some mental content sufficiently similar to that attitude for all intents and purposes. It follows, like Schutz had surmised, that in every rational and intentionally produced act of expressing a contentful thought, what is expressed is intended as some kind of communication, presupposing a recipient of the message. (Unnsteinsson 2022: 87–88)

Note, first, that what Unnsteinsson is interested in is the intention that makes S's act of saying/expressing intentional and rational. We've gone along, for the time being, with the assumption that this must be an intention to have an effect on a particular hearer H. However, it is still distracting to call it "$M(p)$" or a "meaning-intention". After all, it is just whatever intention it is that makes S's act intentional and rational, and it is an open question whether this is equivalent to a Gricean communicative intention that explains acts of speaker meaning. Unnsteinsson thinks it is. He characterizes the Gricean communicative intention as follows, relying on Sperber and Wilson:

> Basically, the speaker's intention to communicate, or the 'communicative intention, involves at least two more basic intentions. The first is the effective intention which Sperber and Wilson call the 'informative' intention. This is simply an intention to have a cognitive effect on someone; I can intend to produce a belief, but also an intention. This may explain, for example, differences between declaratives and imperatives. More importantly, for current purposes, I can very well intend to induce an attitude in someone without thereby *meaning* anything by my action. A cop could plant evidence on a crime scene to make everyone believe that the butler did it. The cop does not express that belief by the act of planting evidence, however. She could express that belief by ostentatiously displaying the evidence to someone or simply saying that she believes the evidence is there. The crucial difference in the latter case is the cop's *signaling* intention, as I will call it, namely the intention that the audience recognize that she wants the audience to believe something. This is the intention to get the hearer to recognize the speaker's effective intention, as Thom Scott-Phillips (2015) would say, to "signal signalhood". The combination of effective and signaling intentions partly determines the speaker's overall communicative intention, also called the *meaning*-intention. … The point here is to say that the communicative intention is precisely the type of propositional attitude common to all rational acts of expressing a thought that we were looking for, even, I argue, in cases of self-directed speech. (Unnsteinsson 2022: 88–89)

At a relatively general level, this is a standard Gricean view of communicative intentions on which they consist of a pair of audience-directed intentions. Informally, the first intention is to induce a particular *response* (belief, activated belief, second-order belief, intention etc.) in the audience. This is what Unnsteinsson calls the "effective intention". The second intention is to *reveal* the first intention to the audience. This is what Unnsteinsson calls the "signaling intention" (compare Harris 2025).

Unnsteinsson thus thinks that even on the expressionist view and other *Easy* views, the intention that makes *S*'s act of saying/expressing intentional and rational is equivalent to the Gricean communicative intention, characterized as above. Here's his final train of thought to this effect:

> So, finally, we can specify the invariant cognitive effect intended by an act of expressing the thought that *p* as that of making it possible for some hearer to infer, on the basis of the utterance, that the speaker intends the hearer to have a particular propositional attitude. Or, more simply, that *S M*-intends some proposition *p*. Thus we have our *M*(*p*)-effect as postulated before. If this is right, the distinction between performing an intentional act of expressing the thought that *p* and doing something with the *M*-intention that *p* simply collapses. There is no difference between the two.

This is where we're supposed to get the collapse. The intention that makes *S*'s saying/expressing intentional or rational is supposed to be equivalent to the Gricean communicative intention. However, this is not true.

On the assumption that all acts of expression are intentional and have to involve intentions to have effects on a particular hearer, it is

true that the intention that makes *S*'s act of saying/expressing intentional and rational must be an intention for the hearer to produce a response and to reveal this to the audience. The intention is thus structurally similar to the Gricean communicative intention. But it is not equivalent to a Gricean communicative intention! This is because, on *Easy* views, such an intention is naturally taken to be for *H* to come to believe that *S* has *said* that p, *expressed* the proposition that p or *expressed* the thought that p. Such intentions are semantically infused in the sense that they mention notions like saying or expressing. In contrast, the Gricean communicative intention is for *H* to come to believe something about the world etc. Such intentions are semantically innocent in that they don't mention concepts like saying, expressing etc. The former sorts of intentions presuppose linguistic meaning etc. and couldn't therefore be used to explain acts of non-linguistic communication or linguistic meaning. In contrast, Gricean communicative intentions don't presuppose linguistic meaning etc. and could be used to explain acts of non-linguistic communication and linguistic meaning.

To sum up, there is no collapse. Even if we go along with all of Unnsteinsson's assumptions, the *Easy* theorists will still think that the intention that makes *S*'s act of saying intentional and rational is distinct from the Gricean communicative intention.

But there is an even more fundamental problem with the whole argument. The question at issue is what it is to say that p/express the proposition that p. *Easy* theorists give a particular view of this that doesn't mention Gricean communicative intentions. Unnsteinsson's aim is to argue that such views collapse into the Gricean view. However, this is not actually what he's done. What he's argued for is that *Easy* theorists have to posit a particular sort of intention to explain what makes saying/expressing intentional and rational. But this is relevant to the question what it is to say/express only if the explanation of what it is to say/express is the same or is somehow dependent on an explanation of what makes an act of saying/expressing intentional and rational. Unnsteinsson seems to be assuming that it is:

> Let me restate the point briefly. Expressionists explain why an utterance means that *p* by holding that the utterance is an act of expressing a thought, where *p* is the content of the thought. What I have argued, in essence, is that expressive acts of this sort cannot be fully explained unless we assume that speakers intend, by performing such acts, to induce some attitude *A* to the proposition *p* in some addressee. Essentially, this is because of the peculiarly cognitive nature of the act-type; the presence of a competence to perform such acts is best explained by their intended effects on creatures with specific rational dispositions or receptivities. If so, the act of expressing a thought is no different from the act of intending to produce a thought in someone. (Unnsteinsson 2022: 90)

The claim that the act of expressing the thought is no different than the act of intending to produce a response assumes that what makes

it the case that one says etc. is the same that makes one's saying etc. rational. In the next section I will show that this is false. The result is that even if Unnsteinsson were right that the *Easy* theorists would have to posit something like a Gricean intention to explain what makes one's saying/expressing intentional and rational, it still wouldn't cause their view of what it is to say/express to collapse into the Gricean view.

4. Constitutive vs. rationalizing questions about language use

At every level of language use, we can and ought to distinguish between the following two questions:

(*Constitutive*) What is it to perform the relevant sort of act (e. g. use a sentence, say/express, speaker mean etc.)?

(*Rationalizing*) What makes one's act intentional and rational?

I will go on to show that these are separate questions that require distinct answers.

In discussing levels of language use, it is helpful to start with Austin's distinctions between phonetic, phatic, and rhetic acts which together constitute what he called the locutionary act:

> The phonetic act is merely the act of uttering certain noises. The phatic act is the uttering of certain vocables or words, i. e. noises of certain types, belonging to and as belonging to, a certain vocabulary, conforming to and as conforming to a certain grammar. The rhetic act is the performance of an act of using those vocables with a certain more-or-less definite sense and reference. Thus 'He said "The cat is on the mat"', reports a phatic act, whereas 'He said that the cat was on the mat' reports a rhetic act. A similar contrast is illustrated by the pairs:
>
> 'He said "I shall be there"', 'He said he would be there';
>
> 'He said "Get out"', 'He told me to get out';
>
> 'He said "Is it in Oxford or Cambridge?"'; 'He asked whether it was in Oxford or Cambridge' (Austin 1962: 95)

A *phonetic* act is simply an act of making certain noises or marks or gestures. Parrots can perform phonetic acts. In contrast, a *phatic* act is an act of making certain noises etc. and therein using (uttering etc.) expressions of a particular language. Non-linguistic creatures can't perform phatic acts. As Austin puts it: "If a monkey makes a noise indistinguishable from 'go' it is still not a phatic act" (Austin 1962: 96). A *rhetic* act is an act of performing a phatic act with *its* meaning (or with one of its meanings if it has several) in the language while fixing the reference of the expressions that need their reference fixed. If one performs a rhetic act with a full sentence then one also performs a *locutionary* act of saying, asking or telling-to which is just an abstraction from the rhetic act which disregards the particular sentence used (Austin 1962: 97–98, for discussion see Reiland 2024).

Let's reserve the term 'uses of expressions' for Austin's phatic acts.[5] Uses of expressions are acts like utterings, inscribings and gesturings etc. What turns a phonetic act, a mere making of noise, into a use of an expression or a phatic act? Many appeal to *articulatory intentions*: intentions, in making a noise, to articulate a particular expression (Cappelen 1999, Hawthorne and Lepore 2011, Neale 2016: 265).[6] Let's assume this view for a moment, since it allows us to make the point that what makes a phonetic act into a use is quite distinct from what makes the use itself intentional and rational. What makes a phonetic act into a use is an articulatory intention. That is what explains why one made the noise or performed the phonetic act and what makes the act of making the noise intentional and rational. But it isn't what makes the use itself intentional and rational. Typically, what makes the use intentional and rational is the speaker's desire to, in using the expression, to perform a locutionary act: to say/express something. But it might also be the speaker's desire to practice pronunciation etc..

Let's reserve the term 'meaningful use of an expression' for Austin's rhetic acts. What turns a use of an expression or a phatic act into a meaningful use or a rhetic act? Many people who hold *Easy* views appeal to *semantic intentions*: intentions, in uttering a sentence, to use it with its meaning in the language (Forguson 1973: 163–165, Evans 1982: 387, Kaplan 1989: 602, Reiland 2025b, Salmon 2004: 257). Depending on the variety of *Easy* view, this might taken to amount to an intention to participate in a convention, be subject to a rule, or to express a thought. Again, let's assume this view for a moment, since it allows us to make the point that what makes a use into a meaningful use is quite distinct from what makes the meaningful use itself intentional and rational. What makes a use into a meaningful use is a semantic intention. That is what explains why one used the sentence and what makes using the sentence intentional and rational. But it isn't what makes the meaningful use or a rhetic act intentional and rational! Typically, what makes the meaningful use intentional and rational is the speaker's desire to, in meaningfully using the expression, to communicate something to an audience. But it might also be to record information for one's private use, to go on the conversational record etc.

Let's reserve the term 'linguistic act' for Austin's locutionary acts. As already hinted at above, the relationships between phonetic, phatic, and rhetic acts and the relationship between rhetic acts and locutionary acts or meaningful uses and linguistic acts are disanalogous. The former three are nested in that the more sophisticated act is constituted by the more basic act + something extra like an articulatory or semantic intention. But the latter two don't stand in this relationship: locutionary acts aren't rhetic acts + something extra. Rather, locution-

[5] Searle calls these utterance acts (Searle 1969: 24) while Alston calls them sentential acts (Alston 2000: 26).

[6] For criticism and alternative views see Munroe 2022, Stojnić 2022.

ary or linguistic acts are just abstractions from meaningful uses or rhetic acts where we disregard the particular sentence used. To report a meaningful use or a rhetic act we say:

(1) Dan used 'Bertrand is British' to say that Bertrand is British/ express the proposition that Bertrand is British.

To report a linguistic or locutionary act we say:

(2) Dan said that Bertrand is British/expressed the proposition that Bertrand is British.

Thus, nothing turns a rhetic act into a locutionary act. Rather a locutionary act is an abstraction from a rhetic act. Still, we can distinguish rhetic or what Heck calls *semantic* descriptions of such acts like in (1) from locutionary or what Heck calls *propositional* descriptions like in (2) (Heck 2006: 30–32). And the point remains. What makes a use of a sentence into a meaningful use and thus into a linguistic act such as saying/expressing, is quite distinct from what makes the linguistic act intentional and rational. Again, what makes a use into a meaningful use and a linguistic act is a semantic intention. That is what explains why one used the sentence and what makes using the sentence intentional and rational. But it isn't what makes the meaningful use or a linguistic act itself intentional and rational! Typically, what makes the meaningful use and the linguistic act of saying/expressing intentional and rational is the speaker's desire to communicate something to the audience etc.

Let's illustrate the distinction between answers to constitutive and rationalizing questions by walking through a concrete example. Take a speaker, Dan who utters 'Bertrand is British' to communicate to Stephen that the point made before in the conversation, say, that descriptions are Russellian, is too obvious to discuss. Here's the structure of nested acts together with pieces of practical reasoning that explains why a particular act was intentional and rational:

1. *Phonetic act*: making the noise /Bertrand is British/

Reasoning: I want to make this noise to utter 'Bertrand is British' (in order to…). Making the noise with an *articulatory intention* to utter 'Bertrand is British' is a way of doing it. Therefore, I will make the noise with the articulatory intention.

2. *Use/Phatic act*: uttering 'Bertrand is British'

Reasoning: I want to use the sentence 'Bertrand is British' in order to say that Bertrand is British (in order to…). Using the sentence 'Bertrand is British' with a *semantic intention* is a way of saying that Bertrand is British. Therefore, I will use the sentence with the semantic intention.

3. *Meaningful use/Rhetic act* & *Linguistic/Locutionary act*: using 'Bertrand is British' with its meaning in English / saying that Bertrand is British.

Reasoning: I want to communicate that the point made before in the conversation, that descriptions are Russellian, is too obvious to discuss. Saying something commonly known, such as that Bertrand is British, is a way to do that due to general pragmatic principles. Therefore, I will say that Bertrand is British.

4. *Speaker meaning/communicating*: that the point made before, that descriptions are Russellian, is too obvious to discuss.

On each step, we can see that what makes the more basic act into the more sophisticated act could be a particular intention (articulatory, semantic, communicative). However, what makes the resulting more sophisticated act itself intentional and rational is a further desire-belief pair that the speaker has.[7]

This should be enough to demonstrate that we shouldn't run together constitutive questions about language use such as what it is to use a sentence, to use it meaningfully or say/express, or even what it is to speaker mean, with questions about what makes these acts intentional and rational.

Coming back to Unnsteinsson's *Collapse Argument*, its conclusion was that *Easy* theorists have to posit something like a Gricean intention to explain what makes one's saying/expressing intentional and rational. But even if this were true, this would only pertain to the rationalizing question and still wouldn't cause their view of what it is to say/express to collapse into the Gricean view.

5. Conclusion

On *Easy* views, saying/expressing something is a matter of uttering a sentence with its meaning in a language which is thought to be a matter of participating in a convention, being subject to a rule, or expressing a thought. Unnsteinsson's aim is to argue that saying/expressing isn't easy and must involve Gricean intentions. His argument is that any *Easy* view must posit something like Gricean intentions to explain what makes saying/expressing intentional and rational. He therefore thinks that the *Easy* views collapse into the Gricean view.

I've shown that this argument fails for several reasons. First, even if we grant Unnsteinsson all of his assumptions, the intentions that the *Easy* theorists would have to posit are still not equivalent to Gricean communicative intentions because they are semantically infused

[7] Note that the same sort of difference is evident on the Gricean view of speaker meaning. On that view what makes one's bare, non-communicative action (gesture, utterance etc.) into a case of speaker meaning or attempted communication is the communicative intention. The communicative intention is thus what makes the *bare action* intentional and rational. It's what explains why one made the gesture, utterance etc. But it isn't what makes the act of *speaker meaning* or attempted communication itself intentional and rational! What makes the act of speaker meaning rational is whatever beliefs and desires the speaker has that make them want to mean or communicate something.

in the way the latter aren't. Second, and fundamentally, Unnsteinsson's argument conflates the *constitutive* question what it is to say/express with the *rationalizing* question what makes saying/expressing intentional and rational. The main lesson of this paper is that this shouldn't be done. These are distinct questions. The result is that even if Unnsteinsson were right that the *Easy* theorists would have to posit something like a Gricean intention to explain what makes one's saying/expressing intentional and rational, it still wouldn't cause their view of saying/expressing to collapse into the Gricean view. Thus, the *Collapse Argument* leaves the *Easy* views of what is it to say/express completely untouched.

References

Alston, W. 2000. *Illocutionary Acts and Sentence Meaning*. Ithaca: Cornell University Press.

Austin, J. L. 1962. *How To Do Things With Words*. Cambridge: Harvard University Press.

Bach, K. 2001. "You don't say." *Synthese* 128: 15-44.

Bach, K. & Harnish, R. 1979. *Linguistic Communication and Speech Acts*. Cambridge: MIT Press.

Burge, T. 1979/2007. "Individualism and the Mental". In *Foundations of Mind*. Oxford: Oxford University Press, 100–150.

Cappelen, H. 1999. "Intentions in Words." *Nous* 33: 92–102.

Chomsky, N. 1975. *Reflections on Language*. New York: Pantheon.

Davis, W. 2003. *Meaning, Expression, and Thought*. Cambridge: Cambridge University Press.

Dummett, M. 1986. "'A Nice Derangement of Epitaphs': Some Comments on Davidson and Hacking." In E. Lepore (ed.). *Truth and Interpretation*. Oxford: Basil Blackwell, 459–477.

Fisher, S. 2024. "That's not what you said! Semantic constraints on literal speech." *Mind & Language*: 1–16.

Green, M. 2007. *Self-Expression*. Oxford: Oxford University Press.

Harris, D. 2025 (forthcoming). "Gricean Communication, Natural Language, and Human Evolution." In B. Geurts, R. Moore (eds.). *Evolutionary Pragmatics*. Oxford: Oxford University Press.

Hawthorne, J. and Lepore, E. 2011. "On words." *The Journal of Philosophy* 108: 447–485.

Heck, R. 2006. "Reason and Language." In C. Macdonald, G. Macdonald (eds.). *McDowell and His Critics*. Oxford: Blackwell, 22–45.

Kaplan, D. 1989. "Afterthoughts." In J. Almog, J. Perry, H. Wettstein (eds.). *Themes from Kaplan*. Oxford: Oxford University Press, 565–614.

Kaplan, D. MS. "The Meaning of Ouch and Oops." http://eecoppock.info/PragmaticsSoSe2012/kaplan.pdf.

Lepore, E. and Stone, M. 2015. *Imagination and Convention*. Oxford: Oxford University Press.

Michaelson, E. 2022. "Speaker's Reference, Semantic Reference, Sneaky Reference." *Mind and Language* 37: 856–875.

Munroe, W. 2022. "What it takes to make a word (token)." *Synthese* 200: 1–30.

Neale, S. 2016. "Silent Reference." In G. Ostertag (ed.). *Meanings and Other Things*. Oxford: Oxford University Press, 229–341.

Reiland, I. 2023a. "Linguistic Mistakes." *Erkenntnis* 88: 2191–2206.

Reiland, I. 2023b. "Rules of Use." *Mind and Language* 38: 566–583.

Reiland, I. 2024. "Austin vs. Searle on Locutionary and Illocutionary Acts." *Inquiry*. Latest Articles. https://doi.org/10.1080/0020174X.2024.2380322.

Reiland, I. 2025a. "Meaningfulness, Conventions, and Rules." *Journal of the American Philosophical Association* 11: 431–446.

Reiland, I. 2025b. "Semantic Intentions." Forthcoming in *Australasian Journal of Philosophy*.

Rosenthal, D. 1986. "Intentionality." *Midwest Studies in Philosophy* 10: 151–184.

Salmon, N. 2004. "The Good, the Bad, and the Ugly." In A. Bezuidenhout, M. Reimer (eds.). *Descriptions and Beyond*. Oxford: Clarendon Press, 230–260.

Schroeder, M. 2008. "Expression for Expressivists". *Philosophy and Phenomenological Research* 76: 86–116.

Searle, J. 1969. *Speech Acts*. Cambridge: Cambridge University Press.

Stojnić, U. 2022. "Just Words: Intentions, Tolerance, and Lexical Selection." *Philosophy and Phenomenological Research* 105: 3–17.

Unnsteinsson, E. 2022. *Talking About: An Intentionalist Theory of Reference*. Oxford: Oxford University Press.

Croatian Journal of Philosophy
Vol. XXV, No. 75, 2025
https://doi.org/10.52685/cjp.25.75.8
Received: September 11, 2024
Accepted: November 8, 2024

Separatory Confusion Does Not Corrupt

ALEXANDRU RADULESCU
University of Missouri, Columbia, USA

If I am confused, and I think two people are one and the same, that may impair my ability to refer to either of them. This is combinatory confusion. What if I am confused, and think that one person is actually two people? This is separatory confusion, and it seems quite different. After all, even in my confusion, my thoughts and my referential devices seem to track back to a single individual. Unnsteinsson has recently argued that both types of confusion corrupt, i.e. they may prevent us from referring the right way. In this paper, I examine the four arguments he offers for this conclusion, and I argue that the intuitive view that separatory confusion does not corrupt can withstand his challenge.

Keywords: Demonstratives; reference; intentionalism; objectivism.

1. Introduction

Confusion can corrupt. That is, when a speaker is confused about something, and they attempt to refer to it, their attempt may well fail because of their confusion. Here is a simple example:

> *Teddy Bears:* my daughter gets a teddy bear called "Bill," and loves it. Afraid that it might get damaged, the next day I buy her an identical teddy bear. I make sure every night to replace one with the other, so that neither gets visibly more used than the other. She often asks for Bill when she wants to play with her dolls, and I give her that day's teddy bear.[1]

My daughter is confused: she thinks there is one single bear, and in fact there are two. Since she plays with them equally, there is no

[1] The case is adapted from (Unnsteinsson 2022: 27); henceforth, all references surrounded by square brackets will be to this book. Similar cases have been widely discussed in the literature, but often in the context of belief reports, or other substitution worries. I will follow Unnsteinsson's lead here, and focus on the relevance of these cases to theories of reference *simpliciter*.

reason to think that "Bill" names just one of them. Suppose that years later, my daughter says "I really loved Bill". Does she succeed in referring to anything? Or, if you prefer, does "Bill" refer to anything? Intuitively, there are two candidates, but neither is more likely than the other. Is she referring to both? That seems wrong too: she loved them both, but she intended to talk about one single bear, the only brown bear she thinks she ever owned. Since there is no such bear, a natural answer is that my daughter did not refer to anything. My daughter, it turns out, had two intentions where she thought she had only one, and since each terminated in a different object, reference fails. Of course, I knew what to do when she asked for Bill: I would hand her one of the two bears. And I now know how to handle the question: I come clean about my well-intentioned deceit. So the name is not useless noise. But it fails to refer, as does my daughter.

One might resist this description of the case in several ways. One might, for instance, insist that "Bill" refers to the first bear to whom the name was attached, whatever happened next. Or one might claim that the speaker's intentions do not determine reference, and might look to other contextual factors to fix it in each context. One might reject the belief model of confusion, and insist that confusion does not reduce to false beliefs.[2] Or one might claim that each time she referred to that day's bear (which does not tell us what she referred to years later). I propose that we put these worries aside, and accept the claim that reference fails here. I want, instead, to see how corruptive confusion is.

Teddy Bears is a case about combinatory confusion: my daughter confused two things as being one. I want to address a different type of confusion: does separatory confusion also corrupt? That is, in cases where the speaker confuses one thing as two, can they refer to that thing, or does their confusion prevent them from doing so, like it prevented my daughter?

At first glance, one might think that there is no danger of corruption here. Let us look at an example:

> *Paderewski:* Paderewski is a famous pianist who was also a famous politician. Peter believes that politicians do not have the time to become accomplished pianists, nor vice versa, so he believes that there are two people called "Paderewski", one a politician, and one a musician. At some political event, he is introduced to Paderewski, whom he takes to be politician-Paderewski. He tells the politician "Paderewski is my favorite pianist; I just wish I could meet him one day".[3]

Did Peter succeed in referring to Paderewski? Well, he got the name right, and we can assume that no other Paderewski was relevant to the conversation. He also intended to refer to the pianist, who does in fact have that name. Sure, at other times he intended to refer to the poli-

[2] See (Goodman 2024), and Unnsteinsson's answer (Unnsteinsson 2024).

[3] This case is adapted from [39], who adapts it from (Kripke 1979). Kripke's focus is belief, not speaker reference, and the differences are important, but they are irrelevant to our present goals.

tician-Paderewski, but both of those intentions terminate in the same person, so whenever he used the name "Paderewski", it looks like he was just referring to Paderewski. There seems to be no problem here.

Recently, Unnsteinsson has argued otherwise: he claims that both kinds of confusion can corrupt, and hence that there was something wrong about Peter's attempt to refer in our scenario. In this paper, I defend separatory confusion from Unnsteinsson's accusations of corruption. We will focus our discussion on proper names, but nothing depends on this choice, other than the ease of finding examples in the literature.

Here is the plan. In §2 I will present Unnsteinsson's claim in some detail, enough to be able to understand the view, especially some key notions introduced or refined by Unnsteinsson. In §3, I present the four arguments that he offers specifically for the conclusion that separatory confusion can corrupt. Along the way, we see that the arguments are usually presented for the broader claim that all confusion can corrupt, but I argue that the arguments work only for combinatory confusion.

2. Spelling out Unnsteinsson's Claim

Before we get to Unnsteinsson's arguments, we need to look at the details of what he claims: that combinatory confusion leads to the corruption of *speaker reference*, with respect to its *proper function*, but only *in cases where the confusion is relevant* and *of the right type*. We need to spell out all four italicized parts of that claim before we can defend separatory confusion.

2.1. Unnsteinsson's notion of speaker reference

The notion of speaker reference belongs to the Gricean tradition of speaker meaning.[4] The claim about the corruptive power of confusion is made solely about speaker reference, thus allowing that it is possible that other types of reference could well go through. The claim is that the kind of reference that Griceans take as basic, which is about certain complex communicative intentions that the speaker has, can get corrupted by confusion.

Unnsteinsson puts his own spin on speaker reference. The differences from Grice and the tradition that grew out of his work are important and interesting, just like the reasons Unnsteinsson offers for those differences, but for the purposes of this paper I will simply take on Unnsteinsson's definition:

> *Speaker Reference:* "Speaker S refers to object o with e in uttering U at time t iff U is an utterance of the linguistic expression type $\Sigma[e]$, and for some H and some propositional attitude A, S utters U intending
> (1) to produce thereby in H an A(p[o])-state;
> (2) H to use the e-part of U as direct evidence that the content of the A-state S intends to produce in H is an o-dependent proposition;

[4] See (Grice 1989; Schiffer 1972; Neale 1992) for *loci classici*.

(3) *H* to recognize *S*'s intention (1);
(4) *U* to recognize that *S* intends *U* to satisfy (1) on the basis of (3)." [131][5]

The novel part of this definition is condition (2), and it is the one that will matter most for our discussion of separatory confusion. The part of the claim that we are interested in is that when a speaker attempts to refer with an expression, they have a communicative intention about a particular object, so that they use that singular expression as evidence for the hearer to figure out what the speaker intended to communicate about that object.

Let us see how this applies to combinatory confusion: when my daughter uses "Bill", that cannot function as evidence that would help me figure out which object she wanted to say something about. And the deeper reason for that is that she really did not have a singular thought about either of the bears that she intended to communicate, so nothing could count as evidence for communicative intentions about such a thought. So (1) fails (there is no intended proposition), and (2) also fails (nothing can count as such evidence), along with all other conditions (since they incorporate (1)). These failures make it that the mechanisms for referring fail to do their job. We will come back to separatory confusion in §2.4.

2.2. Unnsteinsson's notion of proper function

Unnsteinsson's main focus in this book is not reference *simpliciter*. Rather, it is the proper function of speaker reference.[6] Here is the central claim:

> *The Proper Function of Reference* "The mechanism of reference has the proper function of directing the hearer's attention to the referent as a part of the means by which some attitude involving that referent is produced in the hearer." (Unnsteinsson 2022: 141)

As we can see, this is related to part (2) of the definition of speaker reference: since the speaker intends to use the singular expression as evidence about which object they intend to communicate, then it is natural to claim that a device of reference is doing its job when it does provide such evidence.[7] As we saw above, confusion can lead to a failure of speaker reference, and this leads to a constraint on the proper functioning of devices of reference:

> *Edenic Constraint on Speaker Reference:* "If *S* is relevantly confused about object *o* at time *t*, S cannot successfully perform an act of speaker reference to refer to *o* at *t*, i.e., *S*'s act of referring will be constitutively barred from performing its proper function." [151]

[5] "Σ[*e*]" is an expression of type Σ that contains expression *e*, and "*p*[*o*]" is a proposition whose truth conditions depend on object *o*.

[6] Unnsteinsson is using Millikan's notion of a proper function here. See (Millikan 1984; Garson 2013; Neander 2017).

[7] Talk of attention is somewhat surprising here, since attention is not obviously needed in all cases in which condition (2) works out. But again, the focus of this paper is elsewhere, and attention will play no role hereafter.

Since the focus is on the proper function of, say, proper names, this leaves open two options in cases where names fail to fulfill that function: we can either say that reference always fails, or that reference may succeed, just not in the right way. Unnsteinsson is not perfectly clear about this. In one place, he says that "strictly speaking, we assign no referent to Peter's utterance of 'Paderewski'", so it looks like some cases of confusion lead to complete reference failure. In another place, he says that "such acts are best understood, theoretically speaking, in terms of deviations from proper function", and that if reference happens, it's just a matter of luck [157].

Fortunately, we can avoid this issue: I propose that we stay neutral on whether reference happens when the edenic constraint is violated, and that we instead focus on the claim that proper names cannot perform their proper function when the speaker is relevantly confused.

2.3. When confusion is relevant

Suppose that I am confusing Francis Bacon, the Early Modern natural philosopher, with Francis Bacon, the 20th century painter. This fact does not by itself guarantee that all my uses of "Francis Bacon" fail to refer in the right way. Suppose that I am an art critic, who is at best marginally interested in Early Modern goings on, and has merely heard the name "Francis Bacon" as having been a philosopher, and I took that to be about the painter, whose style I am very familiar with. This confusion notwithstanding, so long as I am in a conversation about painters and their styles, when I unfavorably compare the contemporary painter Adrian Ghenie with Bacon, I can do so using both their names, and my confusion does not corrupt.

The reason is that though I am confused, that confusion is not relevant to the communicative project at hand. My intentions are firmly anchored in my painter-beliefs, and there is no part of my plan that relates to my philosopher-beliefs. The addressee, presumably, is also interested in painting styles, and will easily figure out that I am talking about the painter. My belief about philosophy simply does not come into play, and does not serve as bad evidence for the interlocutor.[8]

There is room for clarification here: suppose, as above, that my philosopher-beliefs are indeed irrelevant to my overall plan, but the hearer thinks that they are relevant. There is a sense in which the hearer can be misled by my utterance, due to their beliefs about my beliefs (which I do have, and are merely inactive at this point). This is a case where the two considerations seem to come apart: what is it that matters, the structure of my communicative plan, or what the hearer reasonably takes the plan to be?

[8] See [163]. Unnsteinsson uses a separatory confusion example there, but I want to keep that issue until §3, so I am using a combinatory confusion example.

While this is indeed a worry, I will only rely on cases that work exactly as Unnsteinsson wants them: cases where the speaker's confusion is relevant, and when that issue is a salient one for the hearer as well.

2.4. Which Kinds of Confusion Corrupt

Thus far, we have only given examples of confusion that originate in identity beliefs: confusing one person as two, or two as one. Of course, we ordinarily talk about many more ways of being confused. I may, for instance, mistakenly think that "Big Ben" is the name of a tower, rather than a clock. This involves merely having a false belief about a thing, a false belief that involves no identity.

Unnsteinsson holds that only identity-based confusion can corrupt. The reasons he offers are broadly externalist in nature: one may believe that the Earth is flat, talk about the shape of the Earth, and still succeed in referring to the Earth, because reference does not happen via property satisfaction.[9] Note that the case does not change if the discussion is specifically about the shape of the Earth, so even if the confusion is relevant to the discussion, even its main topic, the name performs its function properly.

This introduces a constraint on Unnsteinsson's project: the arguments for the claim that a particular kind of confusion corrupts must not generalize to property-confusion. To see why, consider the line of thinking above: the speaker fails the Edenic constraint just in case they are not providing the right kind of evidence for the hearer to figure out which object they intend to communicate about. The flat-Earther will say of the Earth that it is flat, and one may come away from that doubting that they intend to say something about the Earth, for instance if the hearer has never come across or even heard of anyone who believes that the Earth is flat. And yet, according to Unnsteinsson, reference succeeds edenically [166]. For now, I just want to flag the issue. We will come back to this tension in the next section.

3. The arguments

As I count them, Unnsteinsson offers four arguments for the claim that separatory confusion corrupts. This count is somewhat arbitrary, because these are not offered explicitly as separate arguments, and, more importantly, because the arguments are related. Still, one could offer versions of these arguments separately, so they each deserve their own discussion. In this section, I show that none of the arguments suffices to show that separatory confusion corrupts.

3.1. The No-Determination Argument

According to Unnsteinsson, a speaker succeeds in referring to an object by using a name just in case the speaker's relevant intentions fix that

[9] See [166] and the citations therein.

object as the referent. So one way for a speaker to fail to refer is for the intentions to determine no referent. I take this to be the most general, and the most fundamental, argument for the claim that confusion corrupts.

This kind of argument applies nicely to combinatory confusion: since my daughter's intentions fail to differentiate between the two teddy bears, when she uses "Bill", her intentions cannot determine either as the referent (and there are no other plausible candidates). Hence, reference failure.

What about separatory confusion? Even if the speaker has two intentions where optimally they would have only one, the problem is quite different: after all, both intentions pick out the same thing, so there seems to be no contradiction. Here is Unnsteinsson, talking about the Paderewski case:

> *No-Determination Argument:* "if reference is metaphysically determined by Peter's referential intention, [we have a problem]. Peter definitely intends to refer to Paderewski and definitely intends not to refer to the man in front of him, who is identical to Paderewski. Hence, there is a problem: as long as the speaker's separatory confusion is sufficiently relevant to the determination of a singular referential intention on a given occasion of utterance, the same problem will arise" [150; see also 162].

In combinatory confusion cases, the failure of edenic reference stems from having intentions that terminate in two different things. In separatory confusion cases, Unnsteinsson finds a different problem: we have the intention to refer to a thing, plus the intention *not* to refer to that thing. This introduces a novel requirement for edenic reference: it must not only be the case that the speaker's intentions terminate in a particular object, it must also be the case that the speaker have no intentions not to refer to that object. Should we grant this constraint?

One problem with the No-Determination argument, as applied to separatory confusion, is that Unnsteinsson does not explain why it does not generalize to other kinds of false beliefs that the speaker might have about the purported referent. Consider the flat-Earther. Unnsteinsson claims that the speaker does manage to refer edenically to the Earth by using its name. Presumably, the reason is that their intention terminates in the same planet as our intentions. But the speaker also very much intends not to refer to a spherical thing. Since the Earth is spherical, can we not describe the speaker's intentions as conflicting in such a way that they fail to terminate in a single object, since there are no flat planets available for them to refer to?

I do grant Unnsteinsson the claim that false beliefs about the properties of the purported referent are irrelevant to the speaker's referring abilities. But why not say the same thing about confusion cases? In combinatory confusion, given all that we granted at the beginning of this paper, there is an answer here: if the speaker's intentions are so messed up that they fail to terminate, they can fix no object as the referent. But in the case of separatory confusion, the intentions to refer

do terminate in a particular object. Sure, there are other intentions around, like the intention not to refer to the person in front of you. But why not treat that like the intention not to refer to a spherical planet? Intentions not to refer to something do not obviously need to play the same kind of role as the intentions to refer.

3.2. The Bad Evidence Argument

The No-Determination argument was about a constitutive feature of speaker reference: the referent must be determined by the speaker's intentions. No determining, no referring. The next argument is about a different feature, namely the claim that the proper function of referring terms is to provide evidence to the hearer, evidence that would help them figure out the object that the speaker intends to communicate about. Here is Unnsteinsson, this time discussing a case where Lois Lane does not know that Superman is Clark Kent:

> *Bad Evidence Argument:* "The hearer's attention is intentionally directed at an object *o* while, simultaneously, it is intentionally not directed at an *x* such that $x = o$. More specifically, there is no object which is Superman while not being Clark and, so, the utterance is evidence for no actual object (or, perhaps, an impossible object) while its proper function is to provide such evidence. And this is precisely because of Lois's false identity belief." [162]

The Bad Evidence argument is, in principle, separable from the No-Determination argument. First, it is obviously possible for the speaker's intentions to determine a particular referent, while the utterance provides poor evidence for the hearer to figure out the purported referent. Unnsteinsson would probably agree that this is possible, for reasons unconnected to identity beliefs. Second, even in cases of false identity beliefs, it is in principle possible to claim that speakers who suffer from combinatory confusion might have intentions that determine a referent, but that their false beliefs provide bad evidence for the hearer, and hence that the overall referential act is not edenic. Unnsteinsson clearly believes that both arguments show that all confusion corrupts, but I will treat the two arguments as separate.

Alas, I believe that the Bad Evidence argument has the same problem as the No-Determination argument: if it works, it generalizes to all false beliefs about the referent, which is a conclusion that Unnsteinsson wants to avoid. The reason is that Lois Lane does not just intend to draw our attention to Clark Kent. She also intends to draw our attention to someone whose eyesight is not perfect. But Clark Kent's glasses are just for disguise. So again, there is something in Lane's thinking that could be misleading, especially to someone who is aware of Lane's beliefs, and their relevance to the discussion. But if this type of false belief does not give evidence bad enough to corrupt, why think that false identity beliefs do? In Ch. 2, Unnsteinsson talks at length, and

persuasively, about the special place that identity beliefs play in our cognitive architecture, and about how they can make it difficult for us to express our beliefs. But as far as I can tell, none of these obviously count as reasons to say that all kinds of confusion lead to the failure of edenic reference. After all, it can always be the case that certain faculties are insulated from the damage wrought by certain types of sub-optimal functioning, no matter how acute that may be.

There is another issue as well: the cases do not seem to me to provide any evidence of bad evidence. Here is why: suppose that we know that Superman is Clark Kent, but we know that Lois Lane does not know, and we do not want to inform her. She says to us: "I find Clark Kent boring". Since we know about her confusion, we know that her telling us that is not evidence about her beliefs about Superman. In fact, we might even know that she does not think Superman boring. All this is rational on her part. Has she provided us with bad evidence? Well, if we simply take into account her intentions not to refer to Superman, that would be bad evidence for us hearers. But we know about her confusion. We would not be misled. Instead, we would not care about her intention not to refer to Superman, and we would work off her intention to refer to Clark Kent.

One might object that this discussion rides on our being informed about Lois Lane's confusion. So let us look at an addressee who is just as ignorant as Lois Lane: Perry White, who is the boss of both Lois Lane and Clark Kent. Suppose that Lane says to White: "I find Clark Kent boring". White takes the use of "Clark Kent" as evidence that Lane wants to say something about Clark Kent. He also knows, let us assume, that neither of them finds Superman boring. So White takes the use of "Clark Kent" not to be intended to pick out Superman. Does Lane thereby provide bad evidence to White? I think not. Lane and White are both thinking about Clark Kent, and the evidence worked well to make that happen. Their coordination is partly based on a shared mistake, but it is not accidental. And, as we just saw, sharing in the mistake is not necessary for what appears to be successful referring.

All this talk of what the addressee knows about the speaker may raise the following worry: is the hearer now merely in the business of repair, of trying to make the best of the attempted communication, and thus aren't they going beyond what the speaker meant and what the speaker referred to? This is a legitimate worry, and it clearly depends on much deeper issues than could be settled in this paper. I have no knock-down way to prove that this response does not work. But then again, there are ways to resist it. Note that we considered both a case in which the addressee knew about the speaker's confusion, and one where they shared the confusion. If it turns out that, no matter what the addressee believes about the hearer, the attempt to refer does its communicative job, this looks like a sign that perhaps the attempt to

refer did exactly what it was supposed to do, in the way that it was supposed to do it. Much more would need to be said here; so ultimately I leave it as a challenge to Unnsteinsson.[10]

3.3. The Non-Starter Argument

The first two arguments were of a general nature: they were designed to show that confusion leads to corruption because it leads to a failure with respect to some central part of the definition of speaker reference. The next two arguments are of a more consequentialist nature: they purport to show that accepting combinatorily confused speakers as edenic referrers would lead to some counterintuitive claim about language use.

The Non-Starter argument is part of Unnsteinsson's discussion of flat Earth believers, but it can be presented separately from it. Here it is in its original context:

> *The Non-Starter Argument:* "A group of three speakers where the majority of each speaker's Earth-beliefs are false and no two of them share the same false Earth-beliefs would still be able to communicate perfectly with each other about the Earth. If we picture the same situation except that the beliefs in question are all false beliefs about the identity or distinctness of the Earth, the practice of using the name wouldn't get off the ground—the name wouldn't function properly in interpersonal communication." [166-167]

The argument is found in the second sentence above: the claim is that the prevalence of combinatory confusion in a population would make it impossible for the name to circulate in that population, because it would not be able to do its job.[11] I find this argument plausible: if a whole population were confusing two things as one, it would be hard to believe that there would be a sustained use of a particular name. Perhaps a new use could arise out of that, for instance if there were a myth involving two gods, but then the later tradition would take there to have been only one god with two personalities. But keeping focused on the original use, its stability would be fairly hard to sustain.

Unnsteinsson claims that this argument shows that identity confusions in general face this problem [167]. But is that plausible for separatory confusion? Consider one of the cases he discusses most often, that of Superman. There clearly is a widespread practice of using two names for the same individual, where only very few people know the truth. In a limit case, I see no reason why it could not happen that nobody knew the truth at any point in time, and still, I see no reason why their practice would not be stable. They would consistently deal with the same person under two names, and so long as Superman keeps being very careful, nobody may ever come to realize the problem. The practices are constant, in their own way: they keep referring to the

[10] I thank Unnsteinsson for pointing out this response.

[11] This passage in the book comes with a reference to (Devitt 1974: 201), where a similar discussion is restricted to combinatory confusion.

same thing. Most things said using the names are true, or as true as statements generally are, except in the cases where issues of identifying Superman become important, when people will generally think false thoughts. But the names seem to live within a stable, continuing practice.

The conclusion here is that the Non-Starter argument shows that combinatory confusion makes it at best difficult for the practice of using a name to be established and to continue in a regular way. But separatory confusion seems to lead to no such problems.

3.4. The Lack of Consent Argument

If I am combinatorily confused, and I think that two people are a single one, I cannot marry either of them by using the name I use indiscriminately when I meet one of them. Nor, let us assume, can I marry both of them (let us assume that I neither want that, nor would the laws allow me to). So combinatory confusion leads to the impossibility of consenting to marriage, at least if a proper name is involved in the right way in the ceremony.[12]

It is not obvious that consent would be altogether impossible even when combinatory confusion is involved. Suppose that the officiant merely asks whether I would marry the person next to me. I assume that there is exactly one person there, and the fact that I believe a lot of false things about them does not seem to take away my power to refer to them, or my power to marry them. I suspect that Unnsteinsson would disagree with this, and claim that my confusion makes it simply impossible for me to consent, no matter how that consent is to be formulated. I am not entirely convinced by this, but the point remains that if the name is an integral part of the ceremony, something has gone so badly that consent cannot properly be said to have been expressed.

What about separatory confusion? Unnsteinsson claims that this confused person is also not in a position to express consent. Here is the argument, using Lois Lane again:

> *The Lack of Consent Argument:* "Lois Lane marries 'Superman,' but 'Clark Kent' had also proposed to her and she refused. 'They' could also constantly switch places during the ceremony and Lois would accordingly change her beliefs as to whether she is with someone to whom she is getting married or not. Let us assume that marriage must be consensual in that one cannot marry x if one intends not to marry x. Thus, since she intended to marry x by saying 'I do,' and she intended to remain unmarried to y while in fact $x = y$, it is hard to say exactly whom she married, if anyone. Especially if y, the person Lois calls 'Clark Kent,' is quite salient in the context—he ran out of the church yelling—it is reasonable to think that the act was not consensual for Lois. Reference is like marriage in presuming uniqueness and consent." [156]

[12] Unnsteinsson uses a more complex example, but the same point is being made. See [155-156].

As described, this case does not feature any utterance of any name. So the question is whether Lois's "I do", said while Superman is next to her, constitutes consent. Unnsteinsson claims that it does not, since it is particularly salient in the context that Lane does not want to marry Kent. So her confusion, claims Unnsteinsson, makes it that she cannot express consent, because she does not consent to marrying Kent, who is, after all, Superman. Again, I am not sure that this argument works. Consent is about the person next to you. Lane does have false beliefs about him, but that seems to be on a par with false beliefs that anyone may have towards their possible future partner. "I thought you were a different person" does not usually mean literally that the speaker had false identity beliefs; but it does express regret to having consented to marriage, because of false beliefs about the person, and thus makes it clear that consent did happen.

Suppose now that the name "Superman" features prominently in the ceremony. Suppose that Lane says "I consent to marrying Superman". Does that present a problem? Again, I think not. Any hearer will correctly take her to be talking about the person next to her. The evidence provided is strong. Yes, her public refusal to Kent's proposal does make it unclear that the marriage will work out, assuming she does find out the truth. But consent is being given, and the fact that she uses the name seems to make no difference.

One could reasonably disagree with this argument. Is the eventual regret a sign that consent was given, but the speaker is unhappy about it, or is it a sign that the speaker was never in a position to give consent, and are now regretting that turn of events? We could turn to the law here, but that would not settle the matter. Law is about being practical, and certain simplifying conventions are necessary in ways that our talk of consent need not agree with. It might be the case that other types of speech act would provide clarity here; I have not found anything that would satisfy all the parties to the dispute, but I am open to the possibility. Still, I take it that the argument, as presented, does not suffice to prove its point.[13]

4. Conclusion

We have looked at four arguments for the claim that separatory confusion makes edenic reference impossible. They were all designed in parallel to similar arguments for combinatory confusion. I have argued that this parallel fails consistently, and for similar reasons: combinatory confused speakers can think about and refer to the things they are confused about, because their false identity beliefs, and their intention not to refer to the referent under some other guise, seem not to infect the work that their positive referential intention does. The remaining conclusion is that the intuition I expressed at the beginning remains

[13] I thank Unnsteinsson for pointing out the complications here.

the more plausible option: combinatory confusion can lead to a failure to refer; separatory confusion cannot.

References

Devitt, M. 1974. "Singular Terms." *Journal of Philosophy* 71 (7): 183–205.

Garson, J. 2013. "The Functional Sense of Mechanism." *Philosophy of Science 80* (3): 317–333.

Goodman, R. 2024. "Confusion and explanation." *Mind & Language 39* (3): 434–444.

Grice, P. 1989. *Studies in the Way of Words*. Cambridge: Harvard University Press.

Kripke, S. 1979. "A Puzzle about Belief." In A. Margalit (ed.). *Meaning and Use*. Dordrecht: Reidel, 239–283.

Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge: MIT Press.

Neale, S. 1992. "Paul Grice and the Philosophy of Language." *Linguistics and Philosophy 15* (5): 509–559.

Neander, K. 2017. *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge: MIT Press.

Schiffer, S. R. 1972. *Meaning*. Oxford: Oxford University Press.

Unnsteinsson, E. 2022. *Talking About: An Intentionalist Theory of Reference*. Oxford: Oxford University Press.

Unnsteinsson, E. 2024. "Inference and identity." *Mind & Language 39* (3): 445–452.

Croatian Journal of Philosophy
Vol. XXV, No. 75, 2025
https://doi.org/10.52685/cjp.25.75.9
Received: July 7, 2025
Accepted: September 15, 2025

Linguistic Plausible Deniability: The Catalyst for Political Manipulation

MIRELA FUS-HOLMEDAL*
Norwegian University of Science and Technology, Trondheim, Norway

Risky politically manipulative speech has unexpectedly been on the rise. This paper investigates the role that the phenomenon of linguistic plausible deniability plays in the increasing prevalence of politically manipulative speech through dogwhistles, racial figleaves, and generic stereotypes. The paper unfolds in three main stages. First, it suggests that these linguistic devices share the phenomenon of plausible deniability, which, by offering cover for their overtness, mitigates (some) risks of such political speech. Second, it argues that the plausible deniability of these linguistic devices makes them powerful tools for politically manipulative speech as it helps it to spread more efficiently and appear more acceptable. Finally, it elevates the ethical and political dimensions of language to a more central position within the philosophy of language by discussing two normative claims stemming from conceptual engineering: (i) we should combat such pernicious political manipulation, and (ii) we should exploit the effects of plausible deniability for beneficial purposes.

Keywords: Linguistic plausible deniability; political manipulation; dogwhistles; racial figleaves; generic stereotypes; conceptual engineering.

* I would like to acknowledge the anonymous reviewers for their valuable input and suggestions, which significantly contributed to enhancing the quality of this paper. For useful feedback on earlier versions of the paper, I would like to thank the members of *Generics Online Group*; as well as the audiences of the *Hate Speech, Fake News and Freedom of Speech Conference*, 2023 July (University of Rijeka, online); *The 11th European Congress of Analytic Philosophy,* ESAP, 2023 August (University of Vienna, Austria); the *Philosophy of Language and Linguistics Conference,* 2023 September (Inter-University Center, Dubrovnik, Croatia); the *Vitenskapsteoretisk forum*, 2023 October (NTNU, Trondheim, Norway); the *Centre for Language Research*, 2024 June (University of Rijeka, Rijeka, Croatia); the *Perceiving Voice and Speaker Project Seminar*, 2024 December (University of Inland Norway, Lillehammer, Norway).

1. Introduction

Linguistic political manipulation can be broadly defined as the use of linguistic devices by politicians, often devised by their strategists, with the aim of gaining political advances. The propagation of -isms and -phobias, such as racism, sexism, Latinophobia, and Islamophobia, through politically manipulative speech is often seen as a calculated means to an end, exploiting racial resentment or implicit biases to achieve political advances. However, engaging in such tactics overtly carries certain risks for politicians and their strategists, as the audience may reject or condemn their speech, ultimately leading to a decrease in political advances.

Yet risky politically manipulative speech has unexpectedly been on the rise. Take, for instance, Trump's 2016 presidential campaign, which showcased the dissemination of both covert and overt racism through linguistic devices like dogwhistles, racial figleaves, and generic stereotypes. However, as Saul (2017a: 97–98) rightly points out, overt racism in political campaigns was previously 'widely thought to be socially unacceptable and death to a nationwide political campaign'. What has changed? What enables the relatively successful introduction of elements of such overtness in recent political discourse? Is there solely a shift in what is socially permissible or acceptable?

This paper takes as a standpoint that recent shifts in political communication norms and media environments may have increased the strategic value of plausible deniability, making it more instrumentally valuable now than before. It then investigates the role that the phenomenon of linguistic plausible deniability[1] plays in the increasing prevalence of politically manipulative speech through dogwhistles, racial figleaves, and generic stereotypes. The paper unfolds in three main stages. First, it suggests that these linguistic devices share plausible deniability, which, by offering cover for their overtness, can mitigate[2] (some) risks of such political speech[3]. Second, it argues that the plausible deniability of these linguistic devices has additional consequences that can be readily utilized for political manipulation. Finally, it discusses some normative considerations connected to the plausible deniability of these linguistic devices and the political manipulation that arises from these consequences.

[1] *Plausible deniability* hereafter refers specifically to linguistic plausible deniability, unless otherwise stated. See section 2 for more on this phenomenon.

[2] While plausible deniability may be significant, it is not the only factor in mitigating risks of these linguistic devices. Success may also vary, as effectiveness depends on factors such as biases, racial resentment, and the social and cultural background of the audience.

[3] Plausible deniability can potentially offer cover for both covert and overt forms of political speech. However, its effectiveness in relation to overt versions further strengthens the main argument of this paper.

The aim of this paper is to bring together and further elucidate the phenomenon of plausible deniability in the context of politically manipulative speech through linguistic devices such as dogwhistles, racial figleaves, and generic stereotypes[4]. However, this paper does not seek to argue for any particular view of plausible deniability nor these linguistic devices, as their existence is widely acknowledged in the literature. Instead, the paper assumes their presence and provides examples commonly recognized in the field. Its objective is to establish a general connection between these three linguistic devices, plausible deniability, and their role in politically manipulative speech. Moreover, it is worth noting that while the logical structure of generic stereotypes is considered to be relatively complex yet stable, figleaves and dogwhistles can be seen as possessing a more flexible logical structure. This flexibility contributes to the fact that plausible deniability associated with the latter two often requires extensive context and historical understanding to be considered credible. Although the topic of the logical structure and context of these linguistic devices warrants closer examination, it falls beyond the scope of this paper and will not be addressed in detail.

The paper proceeds as follows. Section 2 introduces the phenomenon of plausible deniability and provides motivation for its utilization in the context of politically manipulative speech. Sections 3-5 suggest that linguistic devices such as dogwhistles, racial figleaves, and generic stereotypes allow for plausible deniability, and that plausible deniability helps mitigate some potential risks associated with the messages conveyed through these linguistic devices. Section 6 argues that the plausible deniability of these linguistic devices makes them powerful tools for politically manipulative speech as it helps it to spread more efficiently and appear more acceptable. Furthermore, the section discusses two normative claims stemming from conceptual engineering: (i) we should combat such pernicious political manipulation, and (ii) we should exploit the effects of plausible deniability for beneficial purposes.

2. Plausible deniability

Plausible deniability, in broader terms, refers to the ability of individuals, especially those in positions of authority or power, to deny any knowledge, responsibility, or involvement in actions carried out by others within their organization or hierarchy. It is used in various domains including law, military, government, international relations, politics, espionage, intelligence operations, corporate governance, information security, programming, privacy, computer networks, cryptography, re-

[4] It is worth noting that other linguistic and non-linguistic phenomena, such as physical and linguistic micro-aggressions, accents, physical dogwhistles, propaganda, euphemisms, and speaker voice impressions, could also lead to politically manipulative acts that can be plausibly denied.

ligion, as well as informal, everyday or private speech and actions. For instance, a politician might strategically maintain plausible deniability by avoiding any direct involvement or knowledge of the controversial decision or action, despite being in a position to know.

A common thread in both linguistic and non-linguistic forms of plausible deniability is that denial often hinges on a lack of evidence directly linking those in question to the actions in question (linguistic and/or non-linguistic), even if they were personally involved or intentionally chose to remain ignorant about them. They rely on the assumption that their skeptics will be unable to prove otherwise due to the absence of compelling evidence, making their denial seem credible or believable. In some cases, when such (linguistic and/or non-linguistic) actions in question may involve wrongdoing or illegal activities, the absence of concrete evidence makes it difficult or even impossible to take any legal or punitive action against them based solely on accusations.

Within the philosophy of language, the phenomenon of plausible deniability can be broadly defined as a speaker asserting a certain proposition that she could, if challenged, later plausibly deny and claim that the proposition she asserted is, in fact, some other proposition (for some formulations, see Walton 1996; Pinker 2007; Lee and Pinker 2010; Fricker 2012; Stanley 2015; Peet 2015, 2024; Khoo 2017; Camp 2018; Mazzarella et al. 2018; Mazzarella 2021; Dinges and Zakkou 2023; Lemeire ms.).

To illustrate, let's consider a widely discussed example of plausible deniability involving a driver trying to bribe a police officer after being caught for running a red light by using language that allows for plausible deniability (see Pinker 2007: 437). We can contrast two cases to demonstrate how this bribery could occur[5]. In the first case, the driver could ask the police officer if they 'can take care of the ticket on the spot'. In the second case, the driver could explicitly say to the officer: 'I will give you money if you let me go without a ticket'. While the latter case would be considered explicit bribery, the former one would arguably be seen as a (more) implicit form of bribery. In contrast to the second case, what allows the first case to be considered a successful example of implicit bribery is its potential for plausible deniability. Specifically, if challenged by an incorruptible police officer, the driver in question could plausibly deny engaging in bribery. For instance, if challenged, the driver could deny that by uttering 'Can we take care of the ticket on the spot?' she intended to assert proposition p, where p represents something like 'Let me give you the money instead of paying the fine'. As a plausible defense, she could offer an alternative interpretation of her original statement, suggesting that she meant to assert proposition q, where q represents something like 'I would prefer the option of paying the fine on the spot using my credit card and a mobile terminal' (see Mazzarella 2021; Lemeire ms.).

[5] For the contrast between the two cases, see Mazzarella (2021).

To contrast, certain utterances lack plausible deniability due to e.g. their logical structure or the context in which they are uttered, leaving no room for plausible deniability interpretations[6]. Let's take the universally quantified statement 'All Blacks are violent' as an example of implausible deniability. If challenged, one cannot reasonably deny it by claiming to mean that some Blacks are not violent. When considering plausible deniability in terms of context, there is no context in which the statement 'All X are Y' allows for an interpretation that includes 'Some Xs are not Y'. Furthermore, it is also possible to speak of varying degrees of plausible deniability. Some utterances may be deniable, but they are not particularly plausible, even though they are more deniable than statements that are completely undeniable. For instance, in the case of a police officer, one could argue that the statement 'Can we take care of the ticket on the spot?' is perhaps not the most plausible but is still more deniable than the statement 'Let me give you the money instead of paying the fine', which is entirely undeniable (unless, perhaps, one intends it as a joke). There will be many grey areas in between. For the purposes of this paper, it is sufficient to acknowledge that some utterances are plausibly deniable while others are implausibly deniable.

It has also been suggested that an important prerequisite for the speaker to plausibly deny the content of her utterance is that the audience relies on the context of the utterance to be able to recover the content the speaker wants them to recover. Furthermore, if challenged, the speaker can redirect the focus to another plausible context in the vicinity and claim that the audience recovered the content *p* instead of *q* because the audience relied on the context which was not the context of the utterance she made (see Camp 2018; Mazzarella 2021; Lemeire ms.). For instance, in the case of the driver trying to bribe the police officer, the driver may emphasize the context of paying off the officer instead of considering the alternative context where the driver intends to pay using a mobile terminal. The salience of credit card and mobile payment in this alternative context makes the alternative content *q* plausible enough to deny the content *p* (see Mazzarella 2021: 10). Apart from relying on the context in the vicinity, plausible deniability could also be related to the meaning of concepts. For example, in the case of the phrase 'can we settle it on the spot', the speaker may argue that her intended meaning of "settle" is different from the conventional interpretation. However, such alternative interpretations are typically less convincing, since (private) meanings are often disregarded or difficult to change quickly within the constraints of pragmatic norms and conventions. Additionally, the presuppositions, common ground, and other pragmatic parameters at play could also influence the plausibility of deniability.

[6] For the phenomenon of so-called "implausible deniability", see e.g., Lee and Pinker (2010: 793); Camp (2018: 48); Berstler (2019: 27-28); Dinges and Zakkou (2023).

It is, however, important to reiterate that this paper does not aim to develop a specific account of plausible deniability. Thus, instead of offering a well-designed definition of plausible deniability, this paper adopts a heuristic approach to maintain neutrality regarding the specific accounts of plausible deniability associated with the linguistic devices discussed below. By utilizing the notion of a "communicative message", this paper avoids delving into discussions about the precise nature of what is being plausibly denied (e.g., propositions, assertions, semantic or pragmatic content, concepts, context, etc.).

> *Plausible deniability heuristic:* A speaker communicates a message by using a specific linguistic device that enables her to plausibly deny this message when challenged, claiming that the message she communicated is, in fact, a different one.

This heuristic has some inherent limitations, as it assumes that the plausibility of denial is due to the specific linguistic device used. It is, nevertheless, hoped that the heuristic will serve as a guiding principle in supporting the hypothesis that the identified commonalities among the linguistic devices used in political manipulation examined in this paper, namely dogwhistles, racial figleaves, and generic stereotypes (see sections 3-5), are influenced by the phenomenon of plausible deniability.

When it comes to the main motivation behind utilizing plausible deniability, literature emphasizes the intent to avoid the risks associated with openly communicating a certain message, such as asserting content *p*. For instance, in the example of the driver, her motivation for using a linguistic device that involves plausible deniability for her implicit bribery is to mitigate the risk of being punished for bribing a police officer. Furthermore, it is worth noting that the speaker's motivation to utilize plausible deniability of a linguistic device can also arise after the communicative message has already been delivered and the speaker becomes aware of its potential riskiness. The incentive to employ plausible deniability exists regardless of whether the speaker intentionally used the linguistic device with that purpose in mind before being challenged or if she only realized its potential for plausible deniability after being challenged.

Moving on to the domain of political manipulation, one can imagine similar motivations for why certain linguistic devices that allow for plausible deniability would be preferable in political speech, especially if they can mitigate risks that could lead to the politicians losing current or scaring away their potential supporters. Politicians might sometimes use linguistic devices with plausible deniability for the purposes of political manipulation, without being aware that what they are spreading is deeply racist. Instead, they might only care to spread those messages because they believe it will get them certain political advances or help them win the election, without being interested in spreading racism. In other words, the full consequences of their politi-

cal manipulation might go beyond their initial intentions. They might not care about the message being conveyed; their main objective could be to politically manipulate their supporters into accepting something that aligns with their implicit biases, and to evade accountability for if challenged, all for the sake of gaining more political advantage.

Moreover, plausible deniability can extend from speakers to their audience. It can protect not only politicians but also their supporters, an effect that those engaging in political manipulation (or their strategists) might be especially motivated to exploit (often without their supporters being aware of it). For example, being able to plausibly deny an overtly risky racist message could not only mitigate the risks of being considered a racist for the politician who communicated such a message but also for their supporters. This feature of being able to plausibly deny communicating or supporting a racist message comes in handy because often politicians and especially their supporters do not want to be perceived as explicitly racists (though they might be implicitly biased to be such) nor would want to identify themselves as such.

To sum up, certain linguistic devices with plausible deniability allow individuals to mitigate the risks of their communicative message by enabling them to plausibly deny the communication of a risky message when challenged. Using such linguistic devices is particularly advantageous in the realm of political manipulation. In sections 3-5 it will be suggested that linguistic devices such as dogwhistles, racial figleaves, and generic stereotypes used in political manipulation involve plausible deniability, which contributes to their selection as means of political manipulation.

3. Dogwhistles[7]

Some existing accounts of dogwhistles explain them in terms of implicit presuppositions (see Langton 2012), conversational exercitives (see McGowan 2004, 2012), perlocutionary speech acts and effects (see Saul 2018), and not-at-issue content (see Stanley 2015). This paper draws mainly on Saul's (2018) account of dogwhistles. The focus of this paper is not bound to a particular account of dogwhistles, however. This section is about plausible deniability as a more coarse-grained property of dogwhistles. In that sense, the plausible deniability of dogwhistles is seen as a higher-order phenomenon, not tied to either semantics or pragmatics *per se*[8].

According to Saul (2018), dogwhistling as a form of political[9] manipulation can occur through intentional or unintentional dogwhistles,

[7] It is, however, worth noting that Saul's discussion of dogwhistles and the aim of her paper is more sophisticated than discussed here.

[8] Analogous remarks apply to the case of racial figleaves, where Saul's work on racial figleaves is utilized for the same general purposes (see section 4).

[9] For examples of dogwhistles outside the political domain, see Saul (2018) and Witten (ms.).

both in overt and covert ways. While Saul (2018) uses words such as "deniability" and "challenging" that support the phenomenon of plausible deniability, her paper does not explicitly discuss this phenomenon. This section utilizes her account to show how one can tie these different types of dogwhistles to plausible deniability and how plausible deniability plays a role in choosing political manipulation through dogwhistles.

> *Case #1*[10]*. Dogwhistle:* Our inner cities are a disaster. You get shot walking to the store. They have no education, they have no jobs. I will do more for African Americans and Latinos[11] than she [Hillary Clinton] can ever do in ten lifetimes. All she has done is talk to the African Americans and to the Latinos.

Political manipulation through dogwhistle utterances allows politicians (or their strategists) to send one message to the general electorate and another coded message to the target electorate that the general electorate could challenge (see Goodin and Saward 2005; Lopez 2014; Stanley 2015; Saul 2018; Witten ms.). In Case #1, an utterance expressing a dogwhistle "inner cities" carries a coded meaning that associates black and brown communities with negative attributes such as "disastrous", "dangerous", "uneducated", "jobless". If the politician is challenged or accused of racism, one possible defense for the politician (and those who support his political speech) can be to claim that no inherent connection between black and brown communities and these negative attributes has been implicated. Instead, they can argue that their intention was simply to highlight the difficult circumstance faced by these groups and the need for assistance, something they claim their opposition cannot provide.

3.1 Intentional dogwhistles

Saul's (2018) definition of overt intentional dogwhistles, adopted from Kimberly Witten (ms.), is as follows: 'A[n overt intentional] dogwhistle is a speech act designed, with intent, to allow two plausible interpretations, with one interpretation being a private, coded message targeted for a subset of the general audience, and concealed in such a way that this general audience is unaware of the existence of the second, coded interpretation' (Witten ms.: 2). As an example of an overt intentional dogwhistle, consider George W. Bush's use of the phrase 'wonder-working power' to signal to fundamentalist Christians. While the general audience, specifically non-fundamentalists, may perceive this phrase as ordinary political language, it carries a coded message for its target audience, the fundamentalists, referring to 'the power of Christ'. Saul

[10] https://www.vox.com/policy-and-politics/2016/10/19/13336894/third-presidential-debate-live-transcript-clinton-trump

[11] One could remark that Trump is making this dogwhistle more explicit by naming the social groups the dogwhistle inner cities stand for, i.e. he uncoded the coded.

(2018) argues that this dogwhistle conveys two coded messages to its target audience: (i) it aligns with their religious language (idiolect), and (ii) it signifies group membership by speaking their idiolect.

On the other hand, covert intentional dogwhistles are more complex. They are often connected to the Norm of Racial Equality[12] (see Mendelberg 2001), which gained prominence after the 1960s when overt racism became increasingly unacceptable to most supporters[13]. However, what has remained largely unchanged among the target audience are implicit biases and a belief system referred to by psychologists as 'racial resentment' (see Mendelberg 2001) or 'symbolic racism' (see Tesler and Sears 2010). These dogwhistles are often not consciously recognized by both the general and target audience, partly because they appear to be unrelated to race. As a result, the audience does not oppose them in the same way as they would with explicitly racist dogwhistles. Furthermore, these dogwhistles are 'lending deniability if confronted with racism accusations' making them less risky than more overtly racist dogwhistles (see Saul 2018: 365). As an example of covert intentional dogwhistles, consider the Willie Horton advertisement used in George H. W. Bush's campaign against Michael Dukakis. At the time, Dukakis was leading Bush in the opinion polls. The ad, which was part of negative campaigning criticizing the prison furlough program, featured a black man Willie Horton, a furloughed convict who had raped a woman and stabbed a man in their home. Notably, the ad did not explicitly mention Horton's race. However, following the airing of the ad, Dukakis's lead in the opinion polls began to decline significantly. The ad was later labeled as "racist" and sparked extensive discussion. Subsequently, Dukakis's standing in the polls started to recover. Saul (2018: 366) argues that 'as the possibility of racism was raised, the ad ceased to function wholly on an implicit level. Viewers began to consider the possibility that something racial might be going on. And at this point, Dukakis started to rise in the polls again—some indication that the ad had ceased to be effective once race was explicitly under discussion'.

3.2 Unintentional dogwhistles

Another extremely prevalent category of dogwhistles are unintentional dogwhistles. This category encompasses instances where a dogwhistle is inadvertently transmitted, even though those who transmit it are not doing so intentionally and may not be aware of the dogwhistle in question. Nonetheless, these unintentional dogwhistles can produce the

[12] Mendelberg (2001), who introduced the norm, associates it with 'implicit political communication', while Lopez (2014) links it more directly to dogwhistles.

[13] Mendelberg's research indicates that prior to 1930s, American political discourse allowed explicit use of pejoratives for black people, asserting their inferiority to white people, and supporting legal discrimination in the form of enforced segregation or refusal to hire black people. However, The Norm of Racial Inequality began to erode from the 1930s to the 1960s (see Mendelberg 2001: 67).

same effects as the original intentional dogwhistle. According to Saul's (2018: 368) working definition, unintentional dogwhistles refer to the 'unwitting use of words and/or images that, used intentionally, constitute an intentional dogwhistle, where this use has the same effect as an intentional dogwhistle'. These unintentional dogwhistles can be either covert or overt, often contingent on the initial intentional dogwhistle. As an example of spreading an unintentional dogwhistle, consider the case of individuals such as reporters and TV producers who were responsible for re-showing and disseminating[14] the aforementioned Willie Horton ad from Bush campaign, without being aware that it was initially a covert[15] intentional dogwhistle they were propagating. Another example is the creation of a covert intentional racial dogwhistle by the Republican Party in the 1980s, resulting in the association of "government spending" with racial minorities. This association has since been unintentionally perpetuated in everyday discussions about a country's expenditure (see Saul 2018: 367–369).

3.3 Plausible deniability

Regardless of their subtype, both overt and covert dogwhistling involve coded communication that allows for two plausible interpretations. This characteristic provides different pathways for plausible deniability. For example, according to Khoo (2017), the discussion of deniable norm-violations and the ambiguity of code words like "inner city" involves examining their potential racial and non-racial interpretations. Khoo suggests that the social context and implications associated with these code words play a more significant role in their meaning than their semantic definition. The social meaning can introduce genuine ambiguity, particularly with racial connotations. Khoo's analysis revolves around the idea that code words, despite their plausible deniability, can still have racial effects. He proposes a minimal inference-driven account, which emphasizes that the social implications and racial associations of these code words contribute to their impact. This perspective acknowledges the potential for genuine ambiguity while recognizing the predictive power of plausible deniability and the racial effects that can result from the use of such code words.

In the context of political manipulation through dogwhistles, plausible deniability enables politicians and their target audience to deny the presence of a risky coded message, such as one that would label them as racists, when challenged. For instance, if confronted, they can explicitly deny endorsing or accepting the coded and risky message. Instead, they can endorse the second plausible interpretation, a non-risky message intended for the general audience and claim that it

[14] Saul (2018: 368) refers to them as "amplifier dogwhistles".

[15] For an example of an overt unintentional dogwhistle, see Saul's (2018) discussion of the Dred Scott dogwhistle.

was the only intended message. The plausible deniability of both overt and covert dogwhistles can be viewed as an improvement over unambiguously racist statements, as it provides politicians with a defense strategy. This factor may explain why politicians choose to employ dogwhistles. The following paragraphs examine in more detail how this dynamic unfolds in overt and covert dogwhistling.

In the case of an overt intentional dogwhistle[16], both the politicians and the target audience (as opposed to the general audience) are aware of the coded message. However, due to factors like implicit bias or racial resentment, one or both parties may not be aware of the message's riskiness. Nevertheless, regardless of their initial awareness of the riskiness, they can choose to deny the coded message when challenged. Even Saul highlights the deniability aspect of the overt Dred Scott dogwhistle used by Bush: 'Bush intends to have his anti-abortion message recognized, and recognized as intended. At the same time, though, use of a code phrase gives allows [sic] him to avoid placing his contribution on the record—thus achieving deniability' (Saul 2018: 371). It is also worth noting that while Saul had reservations about the effectiveness of an explicit racist dogwhistle because the target audience would likely recognize it as racist and resist it (see Saul 2018: 365), she admits that 'its efficacy would vary from voter to voter, but the deniability it would bring might well allow for a substantial degree of success. When [she] initially drafted [her] paper, [she] thought an explicit racial dogwhistle would fail, but [she is] now (post-Trump) not at all convinced' (Saul 2018: fn. 8).

In the case of a covert intentional dogwhistle, only the politicians (and their strategists) are aware of the coded message, while the target audience is not. Like in the case of overt intentional dogwhistles, both the politicians (and their strategists) and the target audience may not be aware of the riskiness of the coded message due to factors like implicit bias or racial resentment. Saul points out that these dogwhistles 'would appear on its face to be innocuous and unrelated to race—lending deniability if confronted with racism accusation' (Saul 2018: 365). For instance, in the case of the Willie Horton ad, there were 'no overtly racist assertions that are easily pointed to. And politicians can, and did, easily deny that there was racism in the ad or in their intentions' (Saul 2018: 381). However, like with overt intentional dogwhistles, regardless of their initial awareness of the message's riskiness, both the politicians and the target audience can choose to deny the coded message when challenged. Saul explains that 'it will indeed be *conversationally* challenging to make what has been covert explicit. People will reject what challengers say, and deny that it is true. Sanity may be, and often is, called into question. Challengers will be accused of having

[16] Similarly, once the spreaders become aware of the riskiness of their message, unintentional dogwhistles exhibit the same kind of plausible deniability.

a political agenda' (…) 'the intended audience of the speech acts will probably insist that the analysis is wrong and deny the existence of the covert material' (Saul 2018: 381).

4. Racial figleaves

Racial figleaf[17] utterances provide cover for more overt political manipulation than dogwhistles but also represent an improvement over unambiguously racist statements as they offer politicians a defense strategy. This section utilizes Saul's (2017b, 2024) work[18] to demonstrate that the effectiveness of racial figleaf utterances is consistent with plausible deniability. It also discusses their role in overt political manipulation.

> *Case #2. Racial figleaf:* When[19] Mexico sends its people, they're not sending their best — they're not sending you.[20] They're not sending you. They're sending people that have lots of problems and they're bringing those problems with us. They're bringing drugs. They're bringing crime. They're rapists. And some, I assume, are good people.

Political manipulation through racial[21] figleaves involves additional utterances that serve to provide cover for both racist-sounding statements within a political speech and the politicians themselves. In Case #2, a racial figleaf utterance, 'And some, I assume, are good people', is used alongside an explicitly racist-sounding statement, 'They're rapists', to deflect from the publicly unacceptable act of being a racist. If challenged on their racism, politicians (and their supporters) may use the figleaf utterance to argue that they did not mean that *all* of them are rapists, as they explicitly acknowledged that *some* of them are good people.

4.1. Racial figleaves: varieties and complexities

Most notably, Saul introduced the term 'racial figleaf' with the following definition: 'an utterance made in addition to an otherwise overtly racist one, that serves the function of calling into question the racism of the speaker and the utterance' (Saul 2017b: 98). Racial figleaves can

[17] See Saul (2017b: fn. 1) for, non-linguistic, human figleaves, and Saul (2024) for figleaves more generally.

[18] See fn. 7 of this paper.

[19] https://archive.nytimes.com/www.nytimes.com/politics/first-draft/2015/06/16/choice-words-from-donald-trump-presidential-candidate/

[20] "You" can be interpreted as another dogwhistle if one understands it as a coded reference to individuals who harbor resentment or implicit biases against "problematic foreigners".

[21] It's important to note that figleaves exist in other domains beyond race (see Saul 2024).

be categorized as synchronic[22] or diachronic[23], depending on whether the figleaf utterance is provided roughly at the same time or substantially later than the problematic utterance. Some well-known varieties of racial figleaves include the *Denial figleaf* ('I'm not a racist but….'), the *Friendship Assertion figleaf* ('Some of my best friends are … , but . . . [racist utterance]'), and the *Mention figleaf* ('I feel like saying [racist utterance]'). While the first two are often met with skepticism and their acceptance may depend on the audience, the *Mention figleaf* offers more complexity and room for plausible denial to be accepted (see Saul 2017b: 103–107).

Another more complex and arguably more credible racial figleaf is illustrated in Case #2. The typical interpretation of "they" in this speech is that it includes a *generic* statement that sounds racist, rather than a *universal*[24] statement, regarding Mexicans, Mexican (illegal) immigrants, or Mexican (illegal) immigrants sent by the Mexican government[25], which allows for some ambiguity. Unlike universal generalizations, generics have specific truth conditions that do not require all instances to possess the same characteristic for the statement to be considered true. The combination of this feature of generics with the additional figleaf utterance asserting that some of them are assumingly good people enables an interpretation in which both the generic statement and the figleaf utterance can be seen as true (see Saul 2017b: 105). It's important to note the confusion that arises from other instances of the pronoun "they" in the speech, such as 'They're bringing crime', which could also ambiguously refer to one of the aforementioned (sub-)groups of Mexicans or to those who sent them, e.g., the Mexican government. Furthermore, one could argue that the racial figleaf utterance 'And some, I assume, are good people' provides similar cover for

[22] Example of a synchronic racial figleaf: 'I´ve always had a great relationship with the blacks.' https://www.huffpost.com/entry/donald-trump-blacks-lawsuit_n_855553

[23] Example of a diachronic racial figleaf, uttered in an interview that took place after Trump had made several utterances considered to display antiblack racism: 'I have great African-American friendships. I have just amazing relationships, and so many positive things have happened'. https://edition.cnn.com/2015/12/13/politics/donald-trump-antonin-scalia-affirmative-action/

[24] It is worth noting that Tim Kaine, during a debate with Mike Pence, accused Trump of saying that *all* Mexicans are rapists. Pence denied this and mentioned Trump's qualification about some Mexicans being "good people". Fact-checking websites found Kaine's universal quantification to be false. Despite the clarification, some Trump supporters interpreted his statement as slandering all Mexicans, while Pence was able to defend Trump based on the specific wording. For an explanation of this rhetorical move in terms of specific characteristics of generics, see McKeever and Sterken (2021).

[25] For example, Saul shifts and expands the interpretation of "they" between narrower and broader readings, ranging from: 'Mexicans who are sent are rapists' (see Saul 2017b: 104) and 'The Mexicans who come to the United States are rapists' (see Saul 2017b: 105, 111, 112) to 'Mexicans are rapists' (see Saul 2017b: 97, 109, 113).

utterances that label "them" as problematic, criminals, drug users, or rapists.

4.2 Plausible deniability

According to Saul, a racial figleaf effectively counters the inference that the speaker has expressed overt racism, blocking the claim 'The speaker is racist' (see 2017a: 107). For example, the *Denial figleaf* directly denies the claim 'The speaker is racist', but it often fails to convince the audience of its sincerity, especially if the speaker has a history of making racial remarks. The *Friendship Affirmation figleaf*, such as saying 'Some of my best friends are black', attempts to refute the claim 'The speaker is racist' by suggesting that a racist wouldn't have close black friends. The *Mention figleaf* strategically includes the potentially racist utterance within quotation marks, making it more challenging to draw the inference that 'The speaker is racist'. However, it is worth noting that, *pace* Saul (2017b), one may argue that the effectiveness of racial figleaf utterance being consistent with plausible deniability is not because it needs to (always successfully) block the implicature that the statement is racist (or that the speaker harbors racist beliefs) but because it attempts to soften or deny racism by defusing charges of racism, as suggested in section 5[26].

As argued by Saul (2017a), a racial figleaf doesn't necessarily require a direct denial to prevent the inference that 'The speaker is racist'. It is sufficient for the figleaf to create confusion and uncertainty[27]. This understanding is based on a thin version of the Norm of Racial Equality, which aligns with the Ideology of Personalism. According to this ideology, racism is solely a matter of individual beliefs, intentions, and actions[28]. This thin version of the norm allows for the coexistence of racial resentment, which can be measured by agreement with statements like 'Blacks should do the same without any special favors, as Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up[29]'. In Saul's words:

> Due to the Norm of Racial Equality, politicians attempting to exploit racial resentments need to be able to deny that this is what they are doing. Of course, it is far easier to make a convincing denial if you have avoided mentioning race. This is a significant advantage of using an implicit appeal/covert dogwhistle. However, figleaves can be used to provide deniability even when one has been more explicit. Indeed, as we have seen, this deniability

[26] I would like to acknowledge the anonymous reviewer for highlighting this concern.

[27] It is possible that Trump believes both Mexican immigrants are generally rapists and murderers, and that some of them are good people. Such a belief would explain his statements without an intention to hide racism, as he genuinely holds both views (see Saul 2017b: 104).

[28] Hill's (2008) work highlights the individualistic nature of racism.

[29] See Tesler and Sears (2010) who examine racial resentment and its relation to specific beliefs.

> may come in the form of simply denying racism, as in a Denial figleaf. However, the more subtle figleaves offer more possibilities. (Saul 2017b: 109)

On Saul's account, racial figleaf utterances serve as tools for politicians and their target audience to deflect, weaken, or create uncertainty around the potentially risky message that could label them as racists when confronted. This aligns with the idea that racial figleaves are linguistic devices that allow for plausible deniability. However, the effectiveness of different racial figleaves in providing plausible deniability can vary based not only on the type of linguistic device used but also on an individual basis. While none of these figleaves may be 100% effective, they can still be effective with certain groups while potentially less effective with others. For example, they may be less effective with the group targeted by the overt utterance, but their effectiveness does not necessarily need to be decisive[30].

Political figures, including Donald Trump, have effectively utilized racial figleaves to shape public opinion or advance discriminatory policies. By employing these figleaves successfully, they can navigate the boundaries of acceptable speech and, if challenged, maintain a certain level of defense, regardless of their true intentions, racial resentments, or biases. For instance, one could argue that Trump managed, in the 2016 elections, to tap into the prevalent anti-Latino immigrant sentiment in the country and amplify its reach. He did so by employing overt political manipulation using racial figleaves, which potentially garnered him additional votes. Surprisingly, Trump also experienced an increase in support from Latino voters in the 2020 election compared to the 2016 election. This increase in support could be attributed to factors such as his emphasis on economic growth, job creation, as well as his stance against socialist and leftist regimes. However, it is also worth considering that a good faith interpretation of his speeches, which may have included elements of plausible deniability, could have played a role in attracting Latino voters.

5. Generic stereotypes

Generic stereotypes most commonly take the form of a bare plural statement 'Ks are F', such as 'Blacks are violent[31]'. They can appear on their own or in conjunction with other linguistic devices like racial figleaves, which were discussed in the previous section. This section employs the concepts of majority and explanatory generalizations, as well as the ideas of defensive shifting and defensive weakening, to motivate plausible deniability of generic stereotypes. It also highlights their role in overt political manipulation.

Case #3. Generic stereotype: Mexicans are rapists.

[30] Similar can be noted for dogwhistles and generic stereotypes.

[31] For further reference, see Beeghly (2015).

Political manipulation through the use of generic stereotypes exposes the audience to generic beliefs, reinforcing prejudice and implicit biases, and perpetuating harmful or false stereotypes, such as racism or sexism (see Haslanger 2011; Langton, Haslanger and Anderson 2012; Rhodes, Leslie and Tworek 2012; Wodak, Leslie and Rhodes 2015; Leslie 2017; Saul 2017a; Wodak and Leslie 2017; Ritchie 2019; Fus 2021). In Case #3, when a generic stereotype about Mexicans is expressed, it can lead the audience to implicitly adopt the belief that there is a connection between being a rapist and belonging to the group of Mexicans. Generic sentences, unlike universally quantified ones like 'All Mexicans are rapists', allow for exceptions, making them easier to accept and harder to refute (Langton, Haslanger and Anderson 2012). If challenged or accused of racism, politicians (and their supporters) may attempt to claim that the intended meaning was not to imply a non-accidental connection[32] between being a rapist and being a Mexican, or that only some, not all, Mexicans are rapists, as Case #2 explicitly states.

It is worth noting that Case #3 can be viewed as an implicit generalization in Trump's political speech, building upon the previous Case #2. Specifically, as mentioned in section 4.1, his statement 'They are rapists' has been widely perceived as a derogatory and inflammatory generalization, raising questions about whether it pertains to Mexicans in general or specific subgroups within the Mexican population. If one does not agree with this interpretation of Case #2, it is still conceivable that a contemporary politician like Donald Trump could make such a generic stereotype. In case it may still be challenging to imagine it being uttered on its own, without a racial figleaf or a similar device, it is more plausible to consider a politician uttering a generic stereotype such as 'Muslims are terrorists'. For the purposes of this paper, it is sufficient to demonstrate that at least some generic stereotypes could be utilized.

5.1 Majority generalization and explanatory generalization

On the one hand, generic stereotypes can be understood as expressing a majority generalization. For example, they imply that most members of a specific social group, such as Blacks, possess a particular characteristic, like being violent. On the other hand, they can also be interpreted as expressing an explanatory generalization or in-virtue-of-kind-membership. For instance, they may attempt to explain a supposed biological essence, suggesting that Blacks are inherently violent

[32] This is not to claim that, ontologically speaking, racism requires essentialism. Statistical discrimination and statistical stereotyping can still be racist, especially when the alleged statistical connection is groundless. Instead, the denial through defensive shifting or weakening, as argued in 5.1 and 5.2 respectively, is supposed to capture the moves that speakers can employ to defuse charges of racism/sexism/etc., for the purpose of, hopefully, keeping the audience on their side.

in-virtue-of their membership in the social kind 'blacks' (Noyes and Keil 2019; Prasada and Dillingham 2009). However, when attempting to explain a biological essence, speakers engage in what is known as psychological essentializing, which should be distinguished from metaphysical essentializing (Leslie 2013, 2014, 2017). In this context, they treat a certain social group as if there is something inherent in the biological essence of women that makes them less capable of abstract thinking or as if there is a shared cultural or social essence that leads to immigrants being poor (Vasilyeva and Lombrozo 2020; Lemiere ms.).

It is also important to emphasize that generic stereotypes are commonly understood to convey both a majority generalization and an explanatory generalization (Haslanger 2014; Bian and Cimpian 2017; Rosola and Cella 2020; McKeever and Sterken 2021; Lemiere ms.). For instance, Rosola and Cella point out that accepting the statement 'women are bad at abstract thinking' leads one to believe that women, as a group, are generally poor thinkers, attributing this to their presumed 'women are poor thinkers and that this applies to almost every woman, due to their – supposed – underlying nature' (Rosola and Cella 2020: 743).

Given the above, in the context of political manipulation, when a politician (and their supporters) asserts a generic stereotype such as 'Mexicans are rapists', they are asserting both that *most* Mexicans are rapists and that they are inherently dangerous *in-virtue-of* their membership in the social group of Mexicans. If either the politician or their supporters are challenged for falsely claiming that *all* or *most* Mexicans are rapists, they can plausibly deny that they intended to make such a claim. They can argue that there is something *in-virtue-of* being Mexicans that predisposes them to being rapists, even if most of them are not or will never be rapists. It is worth noting that challengers can interpret generic stereotypes either as majority or explanatory generics, and those who are challenged can equally plausibly deny either interpretation. Which interpretation is denied depends on which one is perceived as riskier, and which one is being challenged. The takeaway is that the use of a generic form allows for plausible deniability, unlike more explicit quantification statements such as 'All Mexicans are rapists', 'Some Mexicans are rapists', or 'Most Mexicans are rapists', which do not offer the same level of plausible deniability.

5.2 Defensive shifting and defensive weakening

Langton, Anderson, and Haslanger (2012) employ the concept of defensive shifting to elucidate the two interpretations of generic stereotypes mentioned above. They propose a majority generalization interpretation, which is applicable to generics like 'Cabs are yellow' and 'Barns are red' (Prasada and Dillingham 2009; Prasada, Khemlani, Leslie and Glucksberg 2013). Additionally, they propose a characteristic (or explanatory) generalization interpretation, which suggests that generics

convey that 'those members which do have the property, are disposed to have it by virtue of the fact that they are members of the kind. (…) It does not follow, however, that all or most members of the kind have the property' (Langton, Haslanger and Anderson 2012: 763). Given these two distinct interpretations, Langton, Haslanger, and Anderson (2012) argue that when confronted with accusations of asserting prejudiced generic stereotypes, the speaker can defend themselves by shifting to either of the two interpretations:

> Does [Latinos are lazy] assert a majority generic or a characteristic [i.e. explanatory] generic? Interpret [it] as a majority generic. To combat it, one provides many counterexamples. However, the speaker can then suggest that, although many Latinos aren't lazy, they tend to be—thus embracing the characteristic generic. Instead interpret [it] as a characteristic generic. To combat it one provides evidence that, say, Latinos show no greater tendency towards laziness than any other group. The speaker can then suggest that, although it is not part of the nature or essence of Latinos to be lazy, most are. This slide back and forth between different interpretations of the utterance allows speakers to avoid taking responsibility for the implications of their claims. (Langton, Haslanger and Anderson 2012: 764)

According to them, defensive shifting occurs when, in order to mitigate the risks of their utterance, the speaker who has been challenged for making a generic generalization shifts between the majority and explanatory interpretations, depending on which interpretation they have been challenged on.

Lemeire (ms.), however, argues that Langton, Haslanger and Anderson's (2012) account of defensive shifting overlooks a crucial aspect of the plausible deniability phenomenon. As mentioned above, generic stereotypes are typically interpreted as both majority and explanatory generalizations, such as 'Tigers are striped' or 'Ravens are black', rather than solely as majority generalizations like 'Cabs are yellow' or solely as explanatory generalizations like 'Bulgarians are good weightlifters'. According to Lemeire (ms.), when challenged[33], the speaker (and her supporters) is usually challenged on one interpretation, either the majority generalization or the explanatory generalization. She can then plausibly deny the interpretation she is challenged on (e.g., the majority generalization) and rely on the unchallenged one (e.g., the explanatory generalization), even though she intended to communicate both. Consequently, Lemeire (ms.) and Bowker, Fus-Holmedal, Lemeire, Thakral (ms.), however, argue, such denials represent instances of '*defensively weakening* [of] the content of one's utterance, rather than as defensively shifting between two alternative interpretations'. Consider, for example, a statement made by senior Oxford academic Nick Bostrom: 'Blacks are more stupid than whites[34]'. Despite his sub-

[33] See Lemeire (2021) for two strategies for responding more forcefully to generically formulated stereotypes.

[34] https://thetab.com/uk/oxford/2023/01/12/senior-oxford-uni-academic-argues-blacks-are-more-stupid-than-whites-in-unearthed-emails-29768

sequent apology, he does not appear to renounce his comment regarding relative IQ. Instead, he attributes the disparity to social inequality resulting from unequal access to education, resources, and basic healthcare, rather than a genetic predisposition.

5.3 Plausible deniability

The upshot, then, is that no matter whether generic stereotypes involve defensive shifting[35] or defensive weakening, the accounts discussed above connect them to features of plausible deniability. Specifically, these accounts can be applied to the context of political manipulation to explain how politicians can spread risky (e.g., racially oppressive) messages[36] by incorporating elements of plausible deniability through the use of generic stereotypes in their speeches. This enables politicians and their supporters to mitigate the risks associated with their messages when challenged. Consequently, the presence of plausible deniability in generic stereotypes helps explain why they are favored as more overt forms of political manipulation, shedding light on the shift from covert to more overt strategies.

6. Further consequences and normative considerations

It appears that plausible deniability, as a political phenomenon, often relies on widespread public disagreement about what counts as racist speech—disagreement shaped by a thin interpretation of the Norm of Racial Equality (see section 4.2). Similar dynamics may apply to issues such as sexism, transphobia, and related forms of discrimination, which may be underpinned by similarly thin interpretations of their corresponding equality norms. If the public disagrees about what counts as racism/sexism/transphobia/etc., it is going to be hard for charges of racism/sexism/transphobia/etc. to stick, except in the most obvious cases. There will be political incentives to engage in racist/sexist/transphobic/etc. oppressive speech, especially when addressing audience with simplistic, actively ignorant views[37] about what racism/sexism/transphobia/etc. is and how it manifests in speech/thought/behavior.

This section presents some further arguments for why the plausible deniability of linguistic devices, such as dogwhistles, racial figleaves,

[35] A similar explanation in terms of defensive shifting and weakening could also be applied to dogwhistles and racial figleaves, further reinforcing the idea that they all share plausible deniability.

[36] This point applies regardless of whether the content is conversationally implicated or context-dependent semantic content.

[37] I would like to thank the anonymous reviewer for formulating the importance of this type of active ignorance. While it is important to acknowledge the potential significance of this type of active ignorance in plausible deniability, the limitations of this paper restrict further exploration. At the very least, the paper assumes a connection between active ignorance and the aforementioned thin Norm of Racial Equality.

and generic stereotypes, makes them effective tools for politically manipulative speech. Additionally, it briefly discusses normative considerations from the perspective of conceptual engineering that arise because of these consequences.

6.1 Efficient spread and perceived acceptability

The successful use of linguistic devices with plausible deniability in overt political manipulation can lead to a more efficient spread of the manipulative message. Plausible deniability allows both the transmitters and the recipients of the risky communicative message (e.g., a racist message) to plausibly deny its riskiness. This is especially important in cases where those spreading the message, such as politicians and their supporters, would choose not to transmit or accept it if they were aware of its riskiness. Plausible deniability provides a cover for those intentionally or unintentionally transmitting the risky message, enabling them to deny the risky message when challenged. Moreover, it can incentivize those who are aware of the message's riskiness to deliberately use these linguistic devices, knowing that they can deny it if challenged, thus facilitating the more efficient spread of the risky message.

Furthermore, politically manipulative speech utilizing linguistic devices with plausible deniability can, when successful, appear more acceptable despite its elements of overtness. Plausible deniability protects both the transmitters and the recipients of the risky communicative message by allowing them to (in certain cases) plausibly deny the risky interpretation. If accused of spreading a risky message, such as racism, politicians and their supporters can more easily deflect the accusation by leaning on the non-risky interpretation. Plausible deniability also relies on the principle of charity, which assumes that the speaker's intended message must be the best possible interpretation, in this case, a non-risky one that would not label them as racists. Alternatively, the principle of charity can be extended to one's intentions, so challengers should give the benefit of the doubt to politicians and their supporters once they have pointed out the plausible deniability of their message. Plausible deniability offers a rational alternative interpretation that suffices for charitable interpretation, allowing manipulators to escape more easily. This can further motivate political manipulators to exploit the principle of charity for their purposes.

It is worth noting again that other factors, such as implicit bias and racial resentment, may also facilitate the more efficient spread of the communicative message. These factors can make it harder for politicians and their supporters to perceive that they are spreading and accepting racism, for example. Furthermore, while politicians or their supporters may not initially be aware of the riskiness of their message or its potential for plausible denial, once challenged and made aware of the riskiness, they can still benefit from the plausible deniability of such linguistic devices. They may deny the interpretation that labels

them as racists because they do not want to explicitly commit to racism, even though they may be implicitly biased against different races.

6.2 Some normative considerations

The aforementioned consequences raise important normative considerations that can be addressed through the philosophical method of conceptual engineering. According to this approach, it is recognized that certain representational devices are deficient (descriptive claim) and should be improved (normative claim) (see Cappelen 2018; Burgess, Cappelen and Plunkett 2020).[38] In the context of this paper, one can argue that overt political manipulation through linguistic devices involving plausible deniability is deficient when it leads to socially, morally, or politically harmful effects[39] (descriptive claim), and thus, it should be improved (normative claim).

> *The first normative claim:* We should combat overt political manipulation through linguistic devices with plausible deniability when it results in pernicious effects.

It could be argued that the urgency for improvement is even greater in this case because, as discussed in section 6.1, overt political manipulation through linguistic devices with plausible deniability can appear more acceptable and spread more efficiently. At the same time, plausible deniability makes it challenging for individuals opposing such politically manipulative speech (e.g., anti-racists) to change the minds of politicians, their strategists, or their supporters regarding the dissemination and acceptance of such speech.

Addressing this deficiency requires comprehensive ameliorative projects that extend beyond the scope of this paper. While a systematic approach to ameliorating politically manipulative speech is lacking, recent literature has proposed some solutions for combating harmful effects both within and outside the context of political manipulation. Specifically, for dogwhistles, see Khoo (2017), and Saul (2018); for racial figleaves, see Saul (2017a); and for generics, see Leslie (2017), Saul (2017b), Ritchie (2019), Lemiere (2021), and Fus (2021: ch. 6).

So far, I have suggested that utilizing linguistic devices with plausible deniability for overt political manipulation results in morally, politically, or socially harmful effects, such as the spread of racism or sexism (descriptive claim). However, the plausible deniability of these linguistic devices is not inherently good or bad. On the contrary, some of its consequences, as described in section 6.1, can perhaps be harnessed to achieve morally, politically, or socially beneficial effects.

[38] Given that the deficiency and amelioration in this case do not pertain to concepts, a more suitable term at a higher-order level could be what Fus-Holmedal (2024) refers to as "philosophical engineering".

[39] For objectionable effects of the semantic value, such as socially, morally, or politically objectionable effects, as well as cognitive effects and effects on theorizing, see Cappelen (2018: 33–34), and Fus (2021: ch. 6).

In the context of conceptual engineering, Cappelen (2018) introduces a category of conceptual engineers he calls Exploiters of lexical effects. He argues that: 'There are of course Exploiters with good intentions, but the overall effect of their exploitation is to contribute to and encourage a use of language that undermines what we should treasure the most about it: the continuous exchange of ideas. Exploiters are in effect anti-intellectualist opportunists that contribute to a destruction of genuine communication' (Cappelen 2018:133-134). Similarly, politicians, their strategists, and their supporters can be seen as exploiters of the effects of plausible deniability in linguistic devices such as dogwhistles, racial figleaves, and generic stereotypes. However, one may argue, their aim should be not solely to promote their political agenda and gain power, but also to achieve morally, politically, or socially beneficial effects.

> *The second normative claim:* We should exploit the beneficial effects of overt political manipulation through linguistic devices with plausible deniability.

The underlying assumptions behind this second normative claim are that there can be morally good (political) exploiters or manipulators, and that it is permissible to exploit or manipulate to achieve beneficial goals. Given these assumptions, we should utilize the plausible deniability of linguistic devices such as dogwhistles, racial figleaves, and generics, when they can lead to morally, politically, or socially beneficial effects such as promoting social justice, minority rights, or gender equality (see Cappelen 2018; Fus-Holmedal 2024).

Consider the positive generic stereotype like 'Mexican immigrants are hardworking and have strong family value' that politicians could utter in their political speeches to challenge negative stereotypes and contribute to a more positive perception of a particular group. In some cases, epistemic goals should be overridden[40] by morally, politically, or socially beneficial goals, such as promoting social justice. For instance, introducing a false generic statement like 'girls play football' (see Saul 2017a; Ritchie 2019) could be permissible since, as Saul (2017a: 13) observes, such generics can be 'very important weapons in our anti-prejudice arsenal', or as Ritchie (2019: 38) argues: 'even if it is a false generic generalization, there may be good reasons to assert it in certain political contexts'. Conversely, in other cases, utilizing certain linguistic devices with plausible deniability could serve to express more accurate statements[41]. For example, using descriptively accurate and

[40] For a more in-depth exploration of overriding epistemic goals with morally, politically, or socially beneficial ones, see Fus (2021: ch. 6).

[41] Ritchie (2019) identifies accuracy as one of the primary benefits of social generics. She argues that generics can more accurately describe systematic patterns of violence and discrimination than explicitly quantified claims (Ritchie 2019: 34). This accuracy is one explanation for their social and political effectiveness (Ritchie 2019: 38).

true generic statements like 'women are expected to want children' or 'Blacks face economic, legal, and social discrimination' (see Ritchie 2019: 36) could aid in combating social injustice by providing justification for politicians to introduce policies that could improve the conditions of these marginalized social groups. It is important to notice that such policies may not be justifiable if universal versions of these statements were promoted instead.

One might, however, question the second normative claim—namely, that plausible deniability can be used for good, politically just ends—since examples such as 'Blacks face economic, legal, and social discrimination', 'women are expected to want children', and 'Mexican immigrants are hardworking and have strong family values' are not instances of plausible deniability due to the absence of racist/sexist/transphobic/ implicatures to be denied. In response to this concern, I would like to offer two points. Although these types of statements may not seem, at first glance, to be cases of plausible deniability, as those uttering or accepting them would not be (as frequently) challenged on their racist/sexist/transphobic/ undertones, it does not necessarily mean that they cannot contain such implications. Specifically, there is an interpretation of the above examples suggesting that they stereotype certain social groups, i.e. Blacks, women, Mexicans and can as such be seen to carry racist/sexist/transphobic/ implications. Even if they do not carry them in a strict sense, it does not invalidate the perception of certain audience members, who may interpret them as carrying racist/sexist/transphobic/ implications. The *perceived* racist/sexist/transphobic/ implication is as important in these cases, as political manipulation, whether good or bad, thrives on the audience's interpretation and perception of what can be seen as racist/sexist/transphobic/. For instance, certain members[42] of the audience opposing racism/sexism/transphobia/ might be especially inclined to perceive or emphasize that such statements can be seen as hostile, thereby challenging the ameliorators by focusing on racist/sexist/transphobic/ implications (whether real or perceived), rather than on the non-racist/non-sexist/non-transphobic/ values and goals they aim to promote.

It is also important to recognize that the efficacy of ameliorative approaches is largely an empirical question. This pertains to the broader context of implementing conceptual engineering and, consequently, extends to both the first and second normative claims—whether to counter harmful overt political manipulation through linguistic devices with plausible deniability, or to leverage the beneficial effects of such manipulation. Consider the case of generics, where we are only beginning to unravel the mechanisms dictating the influence of generic

[42] It is also worth noting that not every member of the audience needs to challenge the speaker even though the linguistic device with plausible deniability allows for such challenges. For example, those who already openly adhere to racist/sexist/transphobic/ statements are typically not the ones to challenge the speaker using linguistic devices with plausible deniability.

language on essentialist beliefs, which in turn can contribute to the formation of social stereotypes related to race, gender, transphobia, etc. Recent studies by Foster-Hanson, Leslie and Rhodes (2022) shed light on how generics shape children's concepts. Their findings suggest that simply making a generic claim (e.g., 'girls like pink', 'girls are good at reading'), even if the content is neutral or positive, can lead to the adoption of views that might reinforce stereotypes, regardless of the absence of explicit negative content in the generics themselves. These conclusions prompt discernment when using generics, even for well-intentioned purposes, as positive stereotyping can lead to unintended biases. Furthermore, their studies suggest that ameliorative strategies, which involve responses that maintain 'the generic scope of reference—even if it challenges claims about the referenced features (such as, "no, that's not right about girls" or even "well, boys like dolls too")—are unlikely to limit the spread of essentialist beliefs' (Foster-Hanson, Leslie and Rhodes 2022: 4). Merely negating generic statements about gender would not be sufficient to undermine their influence. Instead, their studies propose two potential solutions to this issue. First, to mitigate the possible negative consequences of the generic, 'one would need to directly challenge the generic scope of the sentence by limiting it to a specific person. For example, when a child hears (or utters themselves) a generic statement about a gender category, a parent might ask which particular person the child is referring to (e.g., "What person do you mean? Yes, Jimmy does like trucks")' (Foster-Hanson, Leslie and Rhodes 2022: 26). Second, to counteract the prescriptive effect of certain generics, the parent might also expand it to 'a superordinate category (e.g., "Lots of kids like trucks")' (Foster-Hanson, Leslie and Rhodes 2022: 26). These findings may suggest that endorsing the first normative claim may not necessarily support the second normative claim, at least in the context of generics.

To summarize, the primary objective of this section was to emphasize the additional consequences of manipulative messaging through linguistic devices with plausible deniability, such as their efficient spread and perceived acceptability. Additionally, the section explored two normative claims that incorporate the phenomenon of plausible deniability within overt political manipulation and suggested potential avenues for exploration within the field of conceptual engineering, without endorsing any specific ameliorative approach.

7. Conclusion

This paper suggested that linguistic devices with plausible deniability have played a significant role in enabling politicians to reintroduce and maintain some elements of overt messaging in the recent era, which was thought to have declined in the 1960s, when the Norm of Racial Equality gained prominence. It has shown how these devices can contribute to the resurgence of certain overt characteristics from the pre-

1960s era but in a more subtle and seemingly plausible manner. As a result, contemporary political speech has become more overt than it was approximately ten years ago while remaining more covert than it was 80–100 years ago.

Furthermore, the paper explored the role of plausible deniability in overt political manipulation, focusing on linguistic devices like dog-whistles, racial figleaves, and generic stereotypes. It discussed the phenomenon of linguistic plausible deniability and demonstrated how these devices can facilitate risky political manipulation. The paper also discussed the consequences that arise from plausible deniability, highlighting the power of linguistic devices with plausible deniability as tools for political manipulation.

Moreover, it contributed to the elevation of ethical and political considerations in the philosophy of language by discussing normative aspects related to plausible deniability and politically manipulative speech from the perspective of conceptual engineering.

References

Baram, M. 2011. *Donald Trump Was Once Sued By Justice Department For Not Renting To Blacks* <https://www.huffpost.com/entry/donald-trump-blacks-lawsuit_n_855553> accessed 24 June 2023.

Beeghly, E. 2015. "What is a stereotype? What is stereotyping." *Hypatia* 30 (4): 675–691.

Berstler, S. 2019. "What's the Good of Language? On the Moral Distinction between Lying and Misleading." *Ethics* 130 (1): 5–31.

Bowker M., Fus-Holmedal M., Lemeire O. and R. Thakral. Manuscript. "Weakening Generic Sterotypes."

Bian, L. and A. Cimpian. 2017. "Are Stereotypes Accurate? A Perspective from the Cognitive Science of Concepts." *Behavioral and Brain Sciences* 40, E3. doi:10.1017/S0140525X15002307.

Burgess, A., Cappelen, H. and D. Plunkett. 2020. *Conceptual Engineering and Conceptual Ethics*. Oxford: Oxford University Press.

Burns, A. 2015. *Choice Words From Donald Trump, Presidential Candidate* <https://archive.nytimes.com/www.nytimes.com/politics/first-draft/2015/06/16/choice-words-from-donald-trump-presidential-candidate/> accessed 24 June 2023.

Camp, E. 2018. "Insinuation, Common Ground, and the Conversational Record." In D. Fogal, D. W. Harris and M. Moss (eds.). *New work on speech acts*. New York: Oxford University Press, 40–65.

Cappelen, H. 2018. *Fixing Language: An Essay on Conceptual Engineering*. Oxford: Oxford University Press.

Dinges, A. and J. Zakkou. 2023. "On Deniability." *Mind* 132/526: 372–401, https://doi.org/10.1093/mind/fzac056.

Foster-Hanson, E., Leslie, S.-J. and M. Rhodes 2022. "Speaking of Kinds: How Correcting Generic Statements can Shape Children's Concept." *Cognitive Science* 46: e13223. https://doi.org/10.1111/cogs.13223.

Fricker, E. 2012. "Stating and Insinuating." *Proceedings of the Aristotelian Society Supplementary Volume* LXXXVI: 61–94.

Fus, M. 2021. *Assert This: 'Philosophers Are Engineers' (A Study of Philosophical Engineering and Generic Judgments*). PhD dissertation. University of St. Andrews and University of Oslo.

Fus-Holmedal, M. 2024. "In Defense of 'Philosophical Engineering': A Novel Terminological Dispute Resolution." In P. Stalmaszczyk (ed.). *Conceptual Engineering: Methodological and Metaphilosophical Issues*. Leiden, The Netherlands: BRILL | mentis, 135–159. https://doi.org/10.30965/9783969753026_008

Golshan, T. 2016. *Full transcript: Hillary Clinton and Donald Trump's final presidential debate* <https://www.vox.com/policy-and-politics/2016/10/19/13336894/third-presidential-debate-live-transcript-clinton-trump> accessed 21April 2023.

Goodin, R. and M. Saward. 2005. "Dogwhistles and Democratic Mandates." *Political Quarterly* 76 (4): 471–476.

Haslanger, S. 2011. "Ideology, Generics, and Common Ground." In C. Witt (ed.). *Feminist Metaphysics*: *Explorations in the Ontology of Sex, Gender and the Self*. Springer Verlag, 179–208.

Haslanger, S. 2014. "The Normal, the Natural and the Good: Generics and Ideology." *Politica & Società* 3: 365–392.

Hill, J. 2008. *The Everyday Language of White Racism*. Chichester: Wiley-Blackwell.

Khoo, J. 2017. "Code Words in Political Discourse." *Philosophical Topics* 45 (2): 33–64.

Langton, R. 2012. "Beyond Belief: Pragmatics in Hate Speech and Pornography." In I. Maitra and M. K. McGowan (eds.). *Speech and Harm: Controversies Over Free Speech*. Oxford: Oxford University Press, 72–93.

Langton, R., Haslanger, S. and L. Anderson. 2012. "Language and Race." In G. Russell and D. Graff Fara (eds.). *The Routledge Companion to Philosophy of Language*. NewYork: Routledge, 753–767.

Lee, J. J. and S. Pinker. 2010. "Rationales for Indirect Speech: The Theory of the Strategic Speaker." *Psychological Review* 117 (3): 785–807.

Lemeire, O. 2021. "Falsifying Generic Stereotypes." *Philosophical Studies* 178 (7): 2293–2312.

Lemiere, O. Manuscript. "The Strong yet Deniable Meaning of Generic Stereotypes."

Leslie S. J. 2013. "Essence and Natural Kinds: When Science Meets Preschooler Intuition." In T. Gendler and J. Hawthorne (eds.). *Oxford Studies in Epistemology*. Oxford: Oxford University Press, vol. 4, 108–165.

Leslie S. J. 2014. "Carving up the Social World with Generics." In J. Knobe, T. Lombrozo and S. Nichols (eds.). *Oxford Studies in Experimental Philosophy*. Oxford: Oxford University Press, vol. 1, 208–231.

Leslie S. J. 2017. "The Original Sin of Cognition: Fear, Prejudice, and Generalization." *Journal of Philosoph,* 8: 393– 421.

Lopez, I. 2014. *Dog Whistle Politics: How Coded Racial Appeals Have Reinvented Racism and Wrecked the Middle Class*. New York: Oxford University Press.

Mazzarella, D. 2021. "'I Didn't Mean to Suggest Anything Like That': Deniability and Context Reconstruction." *Mind & Language* 1–19.

Mazzarella, D., Reinecke, R., Noveck, I. and H. Mercier. 2018. "Saying, Presupposing and Implicating: How Pragmatics Modulates Commitment." *Journal of Pragmatics* 133: 15–27.

McGowan, M. K. 2004. "Conversational Exercitives: Something Else We Do With Our Words." *Linguistics and Philosophy* 27 (1): 93–111.

McGowan, M. K. 2012. "On 'Whites Only' Signs and Racist Hate Speech: Verbal Acts of Racist Discrimination." In I. Maitra and M. K. McGowan (eds.). *Speech and Harm: Controversies Over Free Speech*. Oxford: Oxford University Press, 121–147.

McKeever, M. and R. Sterken. 2021. "Social and Political Aspects of Generic Language and Speech." In Khoo, J. and R.K. Sterken (eds.). *The Routledge Handbook of Social and Political Philosophy of Language*. New York: Routledge, 259–280.

Mendelberg, T. 2001. *The Race Card: Campaign Strategy, Implicit Messages, and the Norm of Equality*. Princeton: Princeton University Press.

Noyes, A. and F.C. Keil. 2019. "Generics Designate Kinds but not Always Essences." *Proceedings of the National Academy of Sciences* 116 (41): 20354–20359.

Peet, A. 2015. "Testimony, Pragmatics, and Plausible Deniability." *Episteme* 12 (1): 29–51.

Peet, A. 2024. "The Puzzle of Plausible Deniability." *Synthese* 203 (156): https://doi.org/10.1007/s11229-024-04600-4.

Pinker, S. 2007. "The Evolutionary Social Psychology of Off-record Indirect Speech Acts." *Intercultural Pragmatics* 4 (4): 437–461.

Prasada, S. and E.M. Dillingham. 2009. "Representation of Principled Connections: A Window onto the Formal Aspect of Common Sense Conception." *Cognitive Science* 33 (3): 401–448.

Prasada, S., Khemlani, S., Leslie, S. J. and S. Glucksberg. 2013. "Conceptual Distinctions amongst Generics." *Cognition* 126 (3): 405–422.

Rhodes, M., Leslie, S.J. and C. Tworek. 2012. "Cultural Transmission of Social Essentialis." *Proceedings of the National Academy of Sciences (PNAS)* 109 (34): 13526–13531.

Ritchie, K. 2019. "Should We Use Racial and Gender Generics?" *Thought: A Journal of Philosophy* 8: 33–41.

Rosola, M. and F. Cella. 2020. "Generics and Epistemic Injustice." *Ethical Theory and Moral Practice* 23 (5): 739–754.

Safire, W. 2008. *Safire's Political Dictionary*. New York: Oxford University Press.

Saul, J. 2017a. "Are Generics Especially Pernicious?" *Inquiry*. doi:10.1080/0020174X.2017.1285995.

Saul, J. 2017b. "Racial Figleaves, the Shifting Boundaries of the Permissible, and the Rise of Donald Trump." *Philosophical Topics* 45 (2): 97–116.

Saul, J. 2018. "Dogwhistles, Political Manipulation, and Philosophy of Language." In F. Daniel, C. Matt and D. Harris (eds.). *Dogwhistles, Political Manipulation, and the Philosophy of Language*. Oxford: Oxford University Press, 360–383.

Saul, J. 2024. *Dogwhistles and Figleaves: How Manipulative Language Spreads Racism and Falsehood*. Oxford: Oxford University Press.

Scott, E. 2015. *Trump Hits Scalia Over Comments on Black Students* <https://edition.cnn.com/2015/12/13/politics/donald-trump-antonin-scalia-affirmative-action/> accessed 24 June 2023.

Stanley, J. 2015. *How Propaganda Works*. Princeton University Press.

Tesler, M. and D. O. Sears. 2010. *Obama's Race: The 2008 Election and the Dream of a Post-Racial America*. Chicago: University of Chicago Press.

Time stuff. 2015. *Here's Donald Trump's Presidential Announcement Speech* <https://time.com/3923128/donald-trump-announcement-speech/> accessed 21 April 2023.

Valentino, N., Hutchings, V. and I. White. 2002. "Cues That Matter: How Political Ads Prime Racial Attitudes During Campaigns." *American Political Science Review* 96 (1): 75–90.

Vasilyeva, N. and T. Lombrozo. 2020. "Structural Thinking about Social Categories: Evidence from Formal Explanations, Generics, and Generalization." *Cognition* 204: 104383.

Walton, D. 1996. "Plausible deniability and evasion of burden of proof." *Argumentation* 10 (1): 47–58.

Witten, K. Manuscript. "Dogwhistle Politics: The New Pitch of an Old Narrative."

Wodak, D., Leslie, S. J. and M. Rhodes. 2015. "What a Loaded Generalization: Generics and Social Cognition." *Philosophy Compass* 10 (9): 625–634.

Wodak, D. and S. J. Leslie. 2017. "The Mark of the Plural: Generic Generalizations and Race." In Taylor, P. C., Alcoff, L. M. and L. Anderson (eds.). *The Routledge Companion to the Philosophy of Race.*, accessed 21 April 2023, Routledge Handbooks Online.

Acknowledgement to Referees

The Editorial Board would like to thank our colleagues listed below for serving as referees of manuscripts submitted to the *Croatian Journal of Philosophy* over the past two years.

Scott Aikin, *Vanderbilt University, Nashville, USA*
Elvio Baccarini, *University of Rijeka, Rijeka, Croatia*
Christopher Bartel, *Appalachian State University, Boone, USA*
Peter Baumann, *Swarthmore College, Swarthmore, USA*
Keith Begley, *Durham University, Durham, UK*
Bálint Békefi, *Central European University, Vienna, Austria*
Nicola Bertoldi, *Catholic University of Louvain, Belgium*
Martina Blečić, *University of Rijeka, Rijeka, Croatia*
Tomislav Bracanović, *Institute of Philosophy, Zagreb, Croatia*
Ian Carter, *University of Pavia, Pavia, Italy*
Daniel Cohnitz, *Utrecht University, Utrecht, Netherlands*
John Collins, *University of East Anglia, Norwich, UK*
Tamara Crnko, *University of Rijeka, Rijeka, Croatia*
Dušan Dožudić, *Institute of Philosophy, Zagreb, Croatia*
Andrius Galisanka, *Wake Forest University, Winston-Salem, USA*
John Gibson, *University of Louisville, Louisville, USA*
Ramiro Glauer, *University of Applied Sciences Potsdam, Potsdam, Germany*
James Gledhill, *George Mason University, Fairfax, USA*
David Grčki, *University of Rijeka, Rijeka, Croatia*
Keith Green, *East Tennessee State University, USA*
Filip Grgić, *Institute of Philosophy, Zagreb, Croatia*
Alison Hall, *De Montfort University, Leicester, UK*
Matthew Hammerton, *Singapore Management University, Singapore*
Ljudevit Hanžek, *University of Split, Split, Croatia*
Pavol Hardoš, *Comenius University, Bratislava, Slovakia*
Clare Hay, *University of Reading, Reading, UK*
Katja Hettich, *The Bauhaus-Universität Weimar, Germany*
Frank Hofmann, *Institute of Philosophy, University of Luxembourg, Luxembourg*

Nurbay Irmak, *Bogazici University, Istanbul, Turkey*
Tomislav Janović, *University of Zagreb, Zagreb, Croatia*
Richard Joyce, *Victoria University of Wellington, New Zealand*
Dunja Jutronić, *University of Split, Split, Croatia*
Seyed Ali Kalantari, *University of Isfahan, Isfahan, Iran*
Eva-Maria Konrad, *Humboldt-Universität zu Berlin, Germany*
Srećko Kovač, *Institute of Philosophy, Zagreb, Croatia*
James Kraft, *University of Edinburgh, Edinburgh, UK*
Changsheng Lai, *Shanghai Jiao Tong University, Shanghai, China*
Gregory Landini, *University of Iowa, Iowa City, USA*
Alexander Linsbichler, *Johannes Kepler University Linz and University of Vienna, Austria*
Jakub Mácha, *Masaryk University, Brno, Czechia*
Ian MacLean-Evans, *York University, Toronto, Canada*
Christian Michel, *Vrije Universiteit Amsterdam, Amsterdam, Netherlands*
Zoltan Miklosi, *Central European University, Vienna, Austria*
Philip Mills, *Goethe University Frankfurt, Germany*
Michael Omoge, *University of Alberta, Camrose, Canada*
Francesco Orsi, *University of Tartu, Tartu, Estonia*
Norbert Paulo, *Ludwig-Maximilian University, Munich, Germany*
Jelena Pavličić Cerović, *University of Belgrade, Belgrade, Serbia*
Davor Pećnjak, *Institute of Philosophy, Zagreb, Croatia*
David Pereplyotchik, *Kent State University, Kent, USA*
Robert Piercey, *Campion College, University of Regina, Canada*
Karol Polcyn, *University of Szczecin, Szczecin, Poland*
Kalle Puolakka, *University of Helsinki, Helsinki, Finland*
Alexandru Radulescu, *University of Missouri, Columbia, USA*
Luke Roelofs, *University of Texas at Arlington, USA*
Lovro Savić, *University of Oxford*, *Oxford, UK*
Karen Simecek, *University of Warwick, Warwick, UK*
Mario Slugan, *Queen Mary University of London, London, UK*
Matej Sušnik, *Institute of Philosophy, Zagreb, Croatia*
Danilo Šuster, *University of Maribor, Maribor, Slovenia*
Guido Tana, *IFILNOVA, Universidade Nova de Lisboa, Lisbon, Portugal / Istituto Universitario di Studi Superiori, Pavia, Italy*
Milica Urban, *University of Rijeka, Rijeka, Croatia*
Iris Vidmar Jovanović, *University of Rijeka, Rijeka, Croatia*
Tzachi Zamir, *Hebrew University of Jerusalem, Jerusalem, Israel*

Table of Contents of Vol. XXV