Charting Sensorial Style: Frontiers in Linguistic Style Analysis

by

Osama Khalid

A thesis submitted in partial fulfillment
of the requirements for the Doctor of Philosophy degree
in Computer Science
in the Graduate College
of The University of Iowa

December 2024

Thesis Committee: Padmini Srinivasan, Thesis Supervisor
Alberto Maria Segre
Rishab Nithyanand
Sanvesh Srivastava
Ethan Kutlu
Bijaya Adhikari

To my mother, Ruby,
Who taught me to love

To my father, Khalid,
Who taught me to think

To my brother Junaid,
Who taught me to code

ACKNOWLEDGMENTS

Imagine an unwieldy child who throws tantrums at odd hours of the day, sometimes keeps you up all night, sometimes for multiple nights. Only ever demanding more attention and nurturing when you think you've given your all. This is what this thesis and the past 5 years have felt like. Except a child eventually grows up and becomes a productive member of society. This thesis on the other hand is as stubbornly resistant to maturity now as it was 5 years ago. The only solace is that like a child, it too was raised by a village.
This thesis would not be what it is today without this village.
This short note is a love letter, a humble attempt at expressing my sincere gratitude to the wonderful people who helped me persevere these past five years.

To My Committee

Padmini: Thank you for being not just an advisor, but a true mentor. Your guidance extended far beyond academic supervision—you knew exactly when to offer praise to build my confidence and when to push me toward greater rigor.

Thank you for teaching me how to be meticulous and thoughtful in my work. Most importantly, thank you for giving me the freedom to follow my intellectual curiosity, even when it led to unexpected places. Though not every experimental path bore fruit, your willingness to let me explore and take

creative risks shaped both this thesis and my growth as a researcher. The innovative approaches presented in this work exist because you created an environment where imagination and creativity could flourish alongside academic discipline.

Alberto: Thank you for teaching me the art of communication through example. I was lucky enough to be your TA during my first year. Your approach to teaching Python to freshmen has been my model for explaining complex ideas to diverse audiences. Being your TA wasn't just about teaching – it was a masterclass in effective communication. Even if I leave academia, your approach to teaching and communicating will always stay with me. Every time I communicate my work to those outside my field, I find myself channeling your ability to make complicated concepts accessible while maintaining engagement.

Rishab: Thank you for your endless patience with my enthusiastic but often unrefined ideas. I appreciate how you listened thoughtfully the first time I burst into your office, convinced I had "theoretically solved" your data void problem—a declaration that proved to be premature, to say the least. Your ability to balance encouragement with gentle guidance helped shape me both as a researcher and as a colleague. I continue to be inspired by how you embrace new ideas with patience and curiosity, even when they come rushing out in a torrent of half-formed excitement.

Sanvesh: I couldn't have finished this thesis on time if you hadn't encouraged me to write and submit Chapter 7 to a conference. Throughout this past year, you've been incredibly generous with your time as I repeatedly showed up to your office hours trying to figure out what should have been simple statistics.

Ethan: Thank you for taking me into your lab when I was scrambling for funding during my last semester. You are not just a mentor, but a true friend – for discussing everything linguistics with me and letting me vent about everything else. What I valued most was knowing I could bring you my most nascent ideas, and you would treat each one with genuine interest and thoughtful consideration. Our weekly meetings, which evolved into breakfast conversations, became a highlight of my week. My only regret is that I got to know you so late in my academic journey. The combination of your academic mentorship and authentic friendship made my final year uniquely meaningful.
Bijaya: Thank you for being extremely accommodating with your time, especially as I was trying to schedule the final defense. This thesis would not have been completed on time if it wasn't for your valuable feedback, especially your comments pertaining to Chapter 7.

To the different labs that adopted me during my PhD.

Starting with

To the members of NLP@UIowa

Jon Rusert: I want to thank you for starting NLP@UIowa. Though our group remained intimate with just three members, it provided exactly the sense of community I needed during those finicky middle years of my PhD. Your influence on my academic development runs deep — as my writing partner, you helped me discover and refine a writing voice that has become an integral part of my academic identity. Thank you for being my foil during our weekly discussions and taking over whenever I would space out in front of

our advisor. Most importantly, the thing I cherish more than our academic collaborations are all the photos of your cat, Belle.

Ingroj: Thank you for always being there for me. For explaining Encoders, Decoders, LLMs. No model was ever too Large or Small that you couldn't explain. This thesis owes a great deal to your explaining how to modify different model architectures.

To the members of the Sparta lab

Manisha: Thank you for making our collaboration on the data voids project such a rewarding experience. Barring the growing pains, watching you take ownership of the project and transform it was truly inspiring.

Hussam: Thank you for teaching me to view my work through a broader lens, showing me how frameworks developed for linguistics could be used in seemingly unrelated domains. And beyond academia, I'm grateful for you joining me for RAGBRAI — the long rides across Iowa gave us plenty of time to discuss everything from abstract theory to life itself.

To the members of the VOICE lab

Danica, Jake, Emerson and Seoyeon: Thank you for being a wonderful set of friends. Thank you for all the dinners, all the coffees, and for helping me when I was packing. And for matching (and often exceeding) my chaotic energy with your own. My only regret is that we didn't get to meet earlier, but the time we did share was invaluable. Your friendships have meant the world to me.

To the Teaching Center

Katherine, Andrew, Hannah and Samantha: Thank you for transforming how I think about education, for helping me discover my passion for teaching, for teaching me how to be an empathetic and compassionate educator, for connecting me to a community of progressive educators who shared the same values and for giving me the language to express what I found wrong with academia. Besides the academics, thank you for recommending a reliable local mechanic.

To the denizens of 14-MLH

Sheryl, Tina and Alli: Thank you for being wonderful pillars of support throughout this entire journey and for helping me navigate the bureaucratic maze of academia. Even when I struggled to keep track of deadlines (this happened quite often), you were always there with patient guidance and timely reminders. Your constant availability, especially during my more frantic moments, helped keep this journey on track.

Matthieu: Thank you for your infectious enthusiasm and for your sense of humor, and for encouraging me to ride and write about RAGBRAI.

To the Linguistics Department

Zuzanna, Jill, Becky, Christine: Thank you for being my introduction to linguistics. For letting me enroll in courses even though I was completely unqualified and lacked any and all prerequisites. Thank you for always checking in, and for making sure that I graduate on time.

Jerzy: Thank you for arranging virtual classes over whatsapp video for me, when I got stuck in Pakistan (This was a recurring theme during my PhD)

Marcin: Thank you for embracing the unconventional aspects of linguistics and giving me the freedom to explore them. Your enthusiasm for my work on emojis showed me that academic rigor doesn't have to mean sacrificing creativity. Your energy, excitement, and pop culture references made every class engaging and memorable.

Onea and Wenqi: Thank you for all the study sessions.

To the people at Google

Dan Liebling: Thank you for being a wonderful manager but more importantly thank you for your infectious excitement and wonderful energy. Thank you especially for encouraging me to focus on making Responsible Tech that has a positive social impact.

Giovanni Motta: Thank you for being so accommodating and allowing me to move between Iowa and California to complete my thesis and defense.

Ghufran Baig: Thank you for helping me with everything! From calculus and C++ in undergrad to Blaze at Google. My life would have been infinitely harder without you.

Asad Ullah Naweed: Thank you for sharing my sense of humor, in all things kitsch, for having faith in me to make it through everything. For hosting me whenever I was in California, and for all our shared meals in both Lahore and in the Bay Area.

To the community around me

Amelia: Thank you for being a wonderful friend and for being extremely generous with your time and energy. Thank you for taking risks with me and encouraging me to explore teaching. For helping me navigate many foreign bureaucracies and airports. For helping me sneak into my first football match, and for introducing me to all the food from the wonderful (chocotorta) to the weird (mondongo). Most importantly, thank you for teaching me that it's okay to not have everything planned and that some problems are best left to "Future Osama and Future Amely".

John Kessler and Kate: Thank you for being there these past 5 years. For the countless dinners and for helping me figure out the transition to adulthood, and for driving all the way to Winterset to pick me up.

To my friends from Pakistan

Abdullah Hasan, Fatima Batool and Fatima Hasan: Thank you for hosting me every time I got stuck in Pakistan, (I kept getting stuck more often than I would've liked) and thank you for introducing me to pop and indie music from the early 2000s.

Rafid and Hamza Malik: Thank you for helping me with my college apps when I was initially applying. And thank you for introducing me to the foodscape of Lahore. Every food-related example in this thesis exists because of our countless explorations of every new restaurant and old *dhaba* in Lahore.

Mohsin Bukhari and Umar Mustafa: Your persistent questioning about my PhD progress, though sometimes exasperating, kept me focused on the

finish line. And true to form, as soon as I defended, you seamlessly moved the goal-post and started asking about marriage.

Joham and Javeria: Thank you for letting me bounce my half-baked ideas for the different chapters. And for offering me a space be it Pakistan or New York.

Fatiqa: Thank you for always being someone I could vent to and vice versa, for the kindred support we shared when we were both looking for jobs.

Saddam and Rana Masood: Thank you for being there throughout my teenage years and twenties. Thank you for teaching me/forcing me to learn how to drive, and literally opening up a new world to me. But most importantly, you showed me how to build deep and meaningful friendships.

To my friends from the Coasts

Saad: From that first day in freshman year when someone told me "Saad knows math but he's also really cool", you've consistently proven both parts true. Every equation in this work passed through your scrutiny - the good math that made it in and the bad math that didn't. Thank you for examining all my half-formed ideas, for visiting me so often, and for joining me on countless adventures both in Pakistan and the US.

Sara: Thank you for being a wonderful friend and an exceptional travel partner. Every comma, every period in this thesis—whether present or absent—owes its existence or nonexistence to your editing. Your writing has inspired me, and your wit has entertained me. Your friendship has made both the academic journey and our travels together infinitely more fun.

To my roommates

Sarmad: Thank you for being both my soundboard and letting me be yours through all the ups and downs of graduate school. Your matching excitement and surpassing curiosity made every conversation memorable. You challenged me to push the boundaries of my productivity, encouraging me to finish this thesis and all the projects in between. The most productive moments of my PhD were spent working on collaborative projects with you. Thank you also for being there to vent with me when things weren't going well, and to celebrate when they were.

Sharaf: Thank you for being a wonderful roommate and an even better friend. From that first freshman anthropology class to my PhD defense, you've been there through every milestone. I especially appreciate how you kept me grounded with your sharp wit and necessary critiques of academia, including (and especially) my own work. Your presence made these past 5-6 years feel like a breeze. Thank you for helping me cultivate a healthy sense of community in Iowa. For indulging me and going on our multi-day road trips even when you obviously didn't want to.

To my family

Junaid: Thank you for being a wonderful companion, mentor, and friend. Thank you for bearing with me as I explored different paths in my 20s, for encouraging me to learn Python, and for learning to accept my inability to make concrete plans.

Talha: Thank you for all the love, for sitting through my defense, and trying

to follow along with all the attention to detail and diligence you could muster. For always helping me scavenge junk-food whenever I was stuck in Lahore. And for your company.

Ruby: Thank you for teaching me everything that mattered – from plumbing to cooking to calculating Euclidean distances. Thank you for teaching me to love others, to be more thoughtful in my actions, and to be more compassionate and kind. Thank you for encouraging me (to the point of coercion) to try new things.

Khalid: Thank you for always saying the right thing and for encouraging me to indulge my passion for the humanities both as an undergrad and as a grad student.

Even as a child, I was resistant to academia. It was only through my parents' exasperated persistence that I managed to make it this far.

To my partner

Sana: A universe of love and thanks for being there with me through it all. For your compassion, your love, and your support through this awfully long journey, through all the مسرت and رنجش. For having all the صبر for all my گلہ. There is no one else I would rather have shared a PhD with.

I have nothing but unbounded love for each and everyone of you!

Final thoughts

This thesis, this entire body of work is فانی and will over time lose relevance (not for you, but definitely for me). But the love I feel for everyone who has helped me write it will never fade.

ABSTRACT

When we communicate, it is not just *'what'* we say, but also *'how'* we say it — our linguistic style — that conveys meaning. The way we use language reveals a lot about us - our background, personality, and even how we perceive the world. This thesis explores an understudied aspect of linguistic style: how people write about sensory experiences like sights, sounds, tastes, and smells, "sensorial style". This thesis bridges a fundamental gap between stylometry (the study of linguistic style) in text collections and sensorial linguistics from psychology by demonstrating that communities have identifiable linguistic styles and that sensory language forms a distinct stylistic dimension.

We formalize the concept of 'sensorial style' using three distinct representations: sensorial synaesthesia, sensorial prevalence, and sensorial diversity. We analyze a range of literary genres including songs, novels, poems, and demonstrate, with sensorial synaesthesia, that sensorial language of individuals exhbit a stable sensorial style.

We extend our analysis to 10 languages and observe that sensorial language use exhibits a consistent hierarchical organization of sensory modalities across languages, with visual terms comprising a vast majority of the sensorial vocabulary. We also find that there is variance in the sensorial diversity of languages. As an example, French and German demonstrate high gustatory diversity compared to Arabic and Urdu, indicating that cultural and linguistic factors significantly influence the richness and variety of sensory descriptions.

We model the relationship between traditional stylometric features and sensorial style. We find that low-dimensional latent representations of traditional stylometric features, like LIWC, effectively capture style information relevant to sensorial language prediction. Building on this, we introduce Stylometrically Lean Interpretable Models (SLIM-LLMs), which combine dimensionality-reduced language models with latent LIWC features.

Our findings not only expand our understanding of language use but also give us a new lens to study how sensory perception is expressed in different cultures and contexts. By developing computational models to represent sensorial style as well as models that integrate traditional stylometric features with sensorial language, we demonstrate the potential for a more comprehensive approach to linguistic analysis. This thesis lays the groundwork for future research exploring the connections between language, cognition, and sensory experience, offering new methods, tools and, perspectives for a wide range of domains.

PUBLIC ABSTRACT

When we communicate, it is not just *'what'* we say, but also *'how'* we say it — our linguistic style — that conveys meaning. The way we use language reveals a lot about us - our background, personality, and even how we perceive the world. This thesis explores an understudied aspect of linguistic style: how people talk about sensory experiences like sights, sounds, tastes, and smells, "sensorial style". This thesis bridges a fundamental gap between stylometry (the study of linguistic style) and sensorial linguistics by demonstrating that communities have identifiable linguistic styles and that sensory language forms a distinct stylistic dimension.

In this thesis, we discovered that sensorial style is a product of meaningful choices in how people communicate. By comparing sensorial language use across cultures, we found both universal tendencies and unique cultural differences in how sensory experiences are described.

We found that people use far more words related to vision than to other senses like smell or touch. This is true regardless of language. However, the balance between senses can shift depending on the context - recipes use more taste-related words, while song lyrics emphasize hearing and emotions. We also found that different cultures vary in how richly they describe each sense. For example, French and German speakers use a wider variety of words to describe tastes and flavors than speakers of other languages especially Arabic and Urdu. Arabic and Urdu speakers tend to use a smaller, more focused set

of words when describing sensory experiences, whether they're talking about sights, sounds, tastes, or smells.

We demonstrate that traditional stylistic features and sensorial style are interconnected. We model this relationship between traditional style and sensorial style using dimensionality-reduced Large Language models that we called SLIM-LLMs. We show that these low-dimensional SLIM-LLMs effectively capture traditional style information relevant to sensorial language prediction.

Our findings not only expand our understanding of language use but also give us a new lens to study how sensory perception is expressed in different cultures and contexts. By developing computational models to represent sensorial style as well as models that integrate traditional stylometric features with sensorial language, we demonstrate the potential for a more comprehensive approach to linguistic analysis. This thesis lays the groundwork for future research exploring the connections between language, cognition, and sensory experience, offering new methods, tools and, perspectives for a wide range of domains.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

It was a forest of cork oaks, and the sun came through the trees in patches, and there were cattle grazing back in the trees.

The Sun Also Rises - Ernest Hemingway

The sun sharpened the walls of the house, and rested like the tip of a fan upon a white blind and made a blue finger-print of shadow under the leaf by the bedroom window.

The Waves - Virginia Woolf

Consider the writing styles of two literary greats of the 20th century — Ernest Hemingway and Virginia Woolf. Even though both authors write about similar phenomena, their approaches to sensory description reveal fundamentally different linguistic styles. Hemingway's concise, unadorned prose emphasizes direct visual imagery through short sentences and sparing use of adjectives. In contrast, Woolf's stream-of-consciousness style weaves together multiple sensory experiences, creating rich tapestries of visual, tactile, and interoceptive sensations.

This variation in how authors encode sensory experiences through language falls within the domain of psychology that studies how language is used to express and communicate sensory perceptions — sensorial linguistics. *Sensorial linguistics* examines patterns in how individuals and groups deploy sensory vocabulary, construct sensory metaphors, and prioritize different sensory modalities in their communication. For instance, some writers may heavily emphasize visual descriptions, while others might focus more on auditory or tactile sensations in their language.

The language we use to communicate is not just a consequence of what we want to communicate but also the particular manner in which we want to say

it. The latter is our *linguistic style*. For instance, one person may use more abstract language, while another might tend towards concrete descriptions. Some people may frequently employ metaphors or similes, while others prefer literal expressions.

Our linguistic style is influenced, among other things, by our individual psychology as well as our social contexts. Linguistic style encompasses a wide range of language features and patterns. It includes features such as sentence structure complexity, vocabulary choice and grammatical constructions, such as passive voice versus active voice, or the frequency of different parts of speech. In this thesis, we are particularly interested in the dimensions of linguistic style that focus on the patterns in sensorial language use. For example, someone might heavily emphasize visual descriptions, while others might focus more on auditory or tactile sensations in their language. Since linguistic style is informed by individual psychological and social pressures, analyzing its features can yield important information about these pressures.

Linguistic style is considered a latent construct of language i.e. we cannot observe linguistic style directly and instead, we rely on quantifiable features that are indicative of linguistic style. As an example, we cannot directly observe or measure "complexity" in a person's use of language, but we can approximate it through various indirect measures such as the average number of words per sentence, the frequency of subordinate clauses, or the use of less common vocabulary. Within the Computer Science tradtion, *Stylometry* focuses on studying linguistic style in text collections, typically using computational or statistical techniques as estimates of this style.

Just as individuals develop unique linguistic styles, groups and communities can develop shared linguistic norms and patterns, i.e. style. These patterns and norms also emerge from repeated interactions within a group. Members may unconsciously adjust their language use to align with the group's norms, a process that contributes to the formation and maintenance of a distinct communal style. As an example, individuals alter their linguistic style when moving between different social contexts or groups in the form of code-switching. A person might use more formal and technical language in a professional setting, then switch to casual slang when interacting with friends. This

indicates that the different contexts have their own linguistic styles. While individual linguistic style has been well-researched in the psycholinguistic literature, there is a gap in our understanding of how shared linguistic styles manifest at the level of abstract aggregates like communities or text genres.

We postulate that such groups also have linguistic style and these can vary across communities and text genres. We expect features may manifest differently in different groups based on their focus and experiences. We might expect certain predictable patterns, like in terms of sensorial language specifically, visual artists using more visual language or chefs employing more gustatory terms. However, in other communities stylistic patterns are not immediately obvious. For instance, even within the same literary genre, authors like Hemingway, Steinbeck and Chandler exhibit markedly different linguistic styles from modernist writers like Woolf, Joyce and Faulkner. Such stylistic variations emerge from the collective experiences and artistic movements these authors belonged to, in addition to their individual inclinations, indicating that systematic analysis of linguistic style — including sensorial language use — can reveal unexpected insights about how literary communities develop and maintain their identities.

Within Computer Science, stylometry has made significant strides in understanding various aspects of linguistic style, such as syntactic patterns, vocabulary richness, and use of function words, however, there remain many unexplored dimensions that could provide valuable insights into language use and human cognition. One such dimension is the stylistic aspect of sensorial language - how sensory experiences are expressed through linguistic choices. This dimension — which we term "sensorial style" — represents how individuals and groups distinctively encode sensory experiences through their linguistic choices. Given the fundamental role of sensory perception in human experience and communication, this gap in our understanding of linguistic style is particularly significant.

Sensorial language use has been studied from the perspective of psychology and the psycholinguistics of individuals. However, relatively little work has been done on how sensorial language is used and how it varies across different communities and text genres. Moreover, much of the existing research in sen-

sorial linguistics [35, 49] has focused on the descriptive analysis of the sensorial language of a limited number of individuals. There has been relatively little work done on analyzing the broader generalizable patterns and trends in the language used within aggregates to encode sensory perception.

Moreover, the scope of most of the research is largely limited to a small number of languages, primarily English, with only a handful of small-scale studies in other languages. This narrow focus presents significant challenges in understanding the full spectrum of sensorial language use across diverse linguistic and cultural contexts. While we know that cultures exhibit fundamentally different patterns in conceptualizing and expressing sensory experiences [51], shaped by their distinct environments, practices, and linguistic frameworks, there is a lack of systematic large-scale text-based studies that could reveal broader patterns and variations in the use of sensory language across cultures. To develop methods and even representations to comprehensively understand and analyze sensorial language, we also need to devise computational approaches and methods that can analyze sensory expressions at scale, enabling us to identify universal patterns and culture-specific variations in how humans encode sensory experiences through language.

In addition to this, even though sensorial style is a form of linguistic style, the relationship between sensorial style and the other forms of linguistic style is still unexplored. For instance, we do not fully understand how the use of sensory language might correlate with overall sentence complexity, or whether individuals who tend to use more abstract language also differ in their use of sensory terms.

Our Contributions

Our work addresses fundamental gaps in stylometric research, focusing particularly on sensorial language use.

Extending Linguistic Style (Chapter 2)

We extend our understanding of linguistic style [36]. We show that linguistic style manifests at the level of abstract aggregates like communities of practice.

We show that communities on social media do indeed have linguistic styles that are both distinctive and representative.

Exploring Sensorial Style (Chapter 3)

We formulate one of the first definitions of sensorial style in literature that uses synaesthesia of the sensory modalities [37]. We show that this formulation of sensorial style is potentially a product of meaningful intent as opposed to random chance. While this definition of sensorial style is very focused, it lays the ground work and opens up the field of sensorial style.

Extrapolating Sensorial Lexicons (Chapter 4)

We develop a systematic method to extrapolate monolingual sensorial lexicons to other languages using BabelNet's multilingual lexical-semantic networks. This provides a foundation for cross-linguistic analysis of sensorial style, across diverse language families and cultures.

Expanding Sensorial Style (Chapter 5)

We broaden our understanding of sensorial style by introducing 2 additional representations of sensorial style — Sensorial Prevalence and Sensorial Diversity. We also develop methods to measure the contribution of individual words to these representations, allowing us to identify which specific words and concepts drive patterns in sensorial style across languages and contexts.

Extending Analysis to Multiple Languages (Chapter 6)

Using the multilingual sensorial lexicons developed in Chapter 4 and the expanded style representations (Prevalence and Diversity) introduced in Chapter 5, we investigate and analyze sensorial style across multiple languages and text types. We examine both parallel Wikipedia articles and culturally-specific texts like recipes and song lyrics to reveal universal patterns and cultural variations in how different languages encode sensory experiences. While

we reveal both universal patterns and cultural variations, our analysis is limited primarily to Indo-European languages and those with significant cultural contact with Indo-European languages.

Explicating Style (Chapter 7)

We explore the interaction between sensorial style and more traditional forms of style. We model and identify low-rank group structures within a form of traditional style. We introduce Stylometrically Lean Interpretable Models (SLIM-LLMs), which provide a more interpretable lens to study the relationship between traditional linguistic style and sensorial style.

Together, these contributions open new avenues for studying human language and perception. By introducing computational methods to analyze and understand sensorial style, we provide researchers with tools to investigate how different cultures and communities express their sensory experiences through language and its relations to lingusitic style.

Chapter 2

EXTENDING LINGUISTIC STYLE

2.1 Introduction

A key attraction in online social media is the capacity to establish specific communities (subreddits, subverses, etc.) for individuals with shared interests. An interesting aspect of these communities is that although their members may come from diverse locations and backgrounds, a shared system of language and communication allows them to engage effectively with each other. This shared communication system evolves naturally, i.e., *in situ*, and is in turn also a part of what defines the community's identity - an identity formed by highlighting the commonalities among group members and differences from other groups [12]. An example of this is the use of shibboleths as linguistic markers of identity [5].

A key component of a community's shared language is its vocabulary [62] - undoubtedly influenced by shared interests. Members interact with each other using familiar content bearing words, phrases, symbols etc. In fact, a shared vocabulary is necessary to effectively engage within the community.

A second key component is the set of para-lingual or stylistic features. These can be explicit like the avoidance of taboo words or more subtle like the use of complex language. While individuals can consciously learn vocabulary to express content, style develops more through subconscious processes [21]. It may be argued that language develops and is used in communities in a similar manner. A dynamic and subtle process of positive and negative feedback perhaps shapes a community's shared style over time.

[0]This work has been published at the Proceedings of the international AAAI conference on web and social media 2020 [36]

Style versus Content Research:

Of the two, content has been studied extensively especially in the field of information retrieval.

Individual writing style has been studied in domains like author attribution research [71]. We know that style can serve as a window into the psychological and sociological state of individuals and provide cues about their gender, occupation and even social class [63]. In fact almost all of the research involving style, or stylometrics, has focused on the individual, leaving open a number of research questions about a community's style. Another limitation in prior research is that researchers rarely maintain a clear distinction between content features and style features and often bundle them all under stylometry [84, 63]. This seriously limits our understanding of the role of style in defining communities.

We see an important opportunity to study the para-language or stylistic features of communities. It is known that groups can have a profound impact on individuals' identities [1]. The study of community-style can give us an additional lens to study the individual in the context of their relation to the community.

Research Goal:

Our goal is to study the linguistic style of 9 communities selected from voat, 4chan and reddit defined around the topics of politics, travel, and television. We study them through a lens made of 262 style-features, taking care to avoid content revealing features. Doing so, we ask the following novel research questions.

RQ1: Do online communities have identifiable linguistic styles? (section 2.5)

If the answer to the above question is yes, then our goal is to identify stylistic features that are distinctive both to a community and globally across communities. (section 2.5)

RQ2: To what extent can we predict community membership based on style alone? Taking care to distinguish style from content, we compare prediction

using style with a baseline strategy where prediction is done more traditionally using content (section 2.5).

Understanding these community-level linguistic styles is cruical not only for analyzing online discourse, but by demonstrating that communities have identifiable linguistic styles, we lay the groundwork for exploring more specific aspects of linguistic style, like patterns that exist in sensorial language use. We also extend the scope of investigation from individuals to more abstract categories such as communities, or even literary genres.

In summary, we find that communities have representative and distinctive style features. Additionally, style predicts community membership with surprisingly excellent results (>95% accuracy and >0.95 F-Score). While this performance is on average statistically equivalent to content-based prediction, we observe that compared to content-based predictions, style-based predictions are more robust to a drastic reduction in training data. In section 2.5 we present a case study comparing the styles of two communities. Section 6 rounds out our study with additional analyses. For example, we find that our style based classifier is excellent at predicting community membership even with thematically similar communities from the same social media platform.

2.2 Background

Distinctiveness of style:

In 2005, Van Halteren et al. introduced the notion of a human 'stylome' analogous to the human genome. This refers to distinctiveness in an individual's linguistic style similar to an individual's distinct genetic makeup. 'Stylometrics' has been the basis of research on identifying author identity [71, 72] and attributes such as gender, race and even social class (e.g., Cheng, Chandramouli, and Subbalakshmi [14]. It has also been used to identify acquired characteristics such as political leanings [65].

Style as an indicator of psychological status:

Pennebaker [62] postulated that, because of the psycholinguistic nature of style, individuals who share similar mental illnesses would have a similar language style. Using Flickr data, Wang et al. [84] employed style similarity to identify individuals susceptible to mental illnesses. Zaman et al. [91] analyzed google search histories with LIWC features to find users with low self-esteem.

Studying style in groups:

Hu et al. (2013) analysed the language style of 3 broadly defined groups: Twitter, email and blogs. These groups, however, do not represent communities of individuals with shared interests, which is our focus [32]. Hu et al. (2016) demonstrated that people with similar occupations share similar language style [31]. However, individuals were selected based on their occupation and not on the basis of networking in 'community' structures.

The study that is closest in spirit to our work in being more community-focused is by Potthast et al. (2017) analyzing the language of three groups of news articles (mainstream, and hyperpartisan right-wing and left-wing articles) [65]. They found similarities in the writing styles of both, extreme right-wing and left-wing articles; the mainstream articles had a different style. But here too, news articles are largely single direction transmitters of information with almost no support for interactivity as in communities. Additionally, they use a limited set of style features.

Limitations in prior work:

(1) There has been no significant focus on examining style in communities; most of the work has focused on individuals and loosely formed groups (email/ news articles). (2) With few exceptions such as Potthast et al. (2017) most papers include content features as part of style analysis [65]. This is not surprising since choice in vocabulary is recognized as a matter of style. But, the downside is that such papers cannot clearly separate signals conveyed through content and those conveyed through style. These limitations motivate our research.

2.3 Methods

Identifying a community's style

A community's style may be understood via its representative stylistic features.

Definition of Representative:

A feature f_x^c is regarded as more representative of a community c than a feature f_y^c, if f_x^c's values show greater consistency across c's posts than f_y^c's values.

More formally we assess the representativeness of feature f_i for a community c by the standard deviation (σ) of its values over c's documents. Since features can have values with different ranges, these values are normalized by scaling them between 0 and 1.

$$\sigma_{f_i^c} = \sqrt{\frac{\sum_{j=1}^{M_c} f_{ij}^c - \overline{f_i^c}}{M_c - 1}} \tag{2.1}$$

Here f_{ij}^c is the normalized feature value in pseudo-document j, M_c is the number of pseudo-documents of community c, $\overline{f_i^c}$ is the mean of the scaled values of the feature. A pseudo-document is the concatenation of a collection of posts and is detailed in the next section.

Identifying distinctive features

There are two perspectives on identifying distinctive features. First, a feature may be distinctive for a specific community and second, a feature may be distinctive across all communities.

Definition of Distinctive for a Community:

A feature f_x^c is more distinctive from a community c's perspective when compared to another feature f_y^c if c's average distance from all the other communities decreases by a larger extent upon exclusion of f_x than upon exclusion of f_y.

Social media	Community	Months	# of comments (Full set)	# of comments (10k-subset)	Time Period
4chan 1%	/pol (/4c/politics)	63	1,697,788	10,265	Nov'13 - Jan'19
	/trv (/4c/travel)	59	6,542	6,542	Feb'14 - Dec'18
	/tv (/4c/television)	74	777,126	9,903	Dec'12 - Jan'19
voat	/v/politics	50	862,501	10,073	Jan'15 - Feb'19
	/v/travel	36	742	742	Jul'15 - Feb'19
	/v/television	50	8,854	8,854	Jan'15 - Feb'19
reddit 10%	/r/politics	77	5,866,346	10,265	Dec'12 - Apr'19
	/r/travel	80	312,862	10,037	Dec'12 - Jul'19
	/r/television	76	830,068	9,931	Jan'13 - Apr'19

Table 2.1: Summary of Datasets: Full set and 10k-subset.

More formally we first define the distance (d) between two communities i and j as follows.

$$d(i,j) = \sqrt{\sum_{k=1}^{N}(\overline{f_k^i} - \overline{f_k^j})^2} \qquad f \in F \tag{2.2}$$

where F is the full set of features with dimension N. We then define the average distance of a community i from all other communities as:

$$\overline{d(i)} = \frac{\sum_{p=1}^{P} d(i,p)}{P} \tag{2.3}$$

where P is the number of communities.

Finally, distinctiveness (ΔC) of a feature α for a community i is defined as follows:

$$\Delta C_i^{\alpha} = \overline{d(i)} - \overline{d^{\alpha}(i)} \tag{2.4}$$

Where $\overline{d^{\alpha}(i)}$ is average distance (Eq. 2.3) computed on the feature space $F - \{\alpha\}$.

Definition of Distinctive Globally:

A feature f_x is more distinctive from a global perspective when compared to another feature f_y if the average distance between all pairs of communities decreases by a larger extent upon exclusion of f_x than upon exclusion of f_y.

Distance between a pair of communities is as in Eq. 2.2. The average is the mean distance across all pairs of communities. We compute the average in two ways, once with all features F and a second time with features in $F - \{\alpha\}$. Their difference represents global distinctiveness and is labelled ΔG^{α} for feature α.

Readability	**Word level features**	**Parts of Speech Frequencies**	**Character Frequencies**
Automated Readability Index (ARI)(D)	Hapax (Dis)Legomena (D)	Verbs(PWS)	Emoji
Coleman-Liau Index (CLI)(D)	Brunét's W Measure(D)	Conjunctions(PWS)	White Space(PWSC)
Dale Chall Readability Index(D)	Yule's K Characteristic(D)	Determiners(PWS)	Digits(PWSC)
Simple Measure of Gobbledygook (SMOG)(D)	Honore's R Measure(D)	Existentials(PWS)	Tabs(PWSC)
Flesch Kincaid Grade Level(D)	Sichel's S Measure(D)	Prepositions(PWS)	Special Characters(PWSC)
Fleash Kincaid Readability Ease(D)	Simpson's Index(D)	Adjectives(PWS)	Uppercase(PWSC)
Gunning Fog Index(D)	Out of Vocabulary Rate(D)	Nouns(PWS)	Linebreaks(PWSC)
	Syllable Frequency(PWS)	Possessives(PWS)	Punctuations(PWSC)
	Short Word Frequency(PWS)	Adverbs(PWS)	Character Count(PWS)
	Elongated Words(PWS)	Possessives(PWS)	
	LIWC-$v3$(PWS)	Interjections(PWS)	

Table 2.2: Categories of Style Features.

Predicting community membership

We approach this as a standard single label, multiple class (9 communities) classification problem. We represent each pseudo-document as a 262 dimensional vector of feature values. These features are described in the next section.

We use a random forest classifier[1] and analyze performance using accuracy, precision, recall and F-score. We use 3 fold cross validation for our prediction experiments ensuring that all comments from the same user (where the user-id is persistent) fall into the same fold.

Baseline:

We use content-based classification as a baseline predictor. We represent each pseudo-document (described later) as a weighted vector of words after stemming and excluding stop words and purely numeric unigrams. The stems are given TF-IDF weights. We use identical 3 fold cross validation and random forest classifier, to calculate the accuracy, precision, recall and F-score.

2.4 Dataset and Features

Dataset

We collected comments posted in 9 communities discussing politics, travel and television from 4chan[2], reddit[3] and voat[4].

reddit:

Reddit is a news aggregation website with over 1 millions subreddits; specific communities with shared interests. Users interact with others by commenting on posts. The Reddit API for data collection has multiple limitations including being rate-limited[5]. Instead, we obtained our data from pushshift.io[6], a reddit archive. We downloaded comments from /r/politics, /r/travel and /r/television subreddits going back to Dec. 2012. We further reduced the large data volume by taking a 10% sample of the data for each community.

[1]We also conducted experiments using a multi-class logistic regression model. However, the results were largely similar to results from the random forest classifiers, therefore we only report the results from the latter.

[2]www.4chan.org

[3]www.reddit.com

[4]www.voat.vo

[5]https://github.com/reddit-archive/reddit/wiki/api

[6]https://github.com/pushshift/api

LIWC-*v*1		**LIWC-*v*2**		**LIWC-*v*3**	
Categories	**Subcategories**	**Categories**	**Subcategories**	**Categories**	**Subcategories**
Linguistic Processes	Function Words Common Verbs Swear Words	Linguistic Processes	Function Words Common Verbs Swear Words	Linguistic Processes	Function Words Common Verbs Swear Words
Psychological Processes	Social Processes Affective Processes Cognitive Processes Perceptual Processes Biological Processes Relativity	Psychological Processes	Social Processes Affective Processes Cognitive Processes Perceptual Processes	Psychological Processes	Cognitive Processes
Spoken Concerns		Spoken Concerns		Spoken Concerns	
Personal Concerns					

Table 2.3: The categories and subcategories of features in each version of LIWC

voat:

Voat is an alt-right variant of reddit [69]. Similar to subreddits, voat has "subverses". However, unlike reddit, voat has neither a well-developed API nor an archive. We manually scraped the comments from the three voat communities /v/politics, /v/travel and /v/television. Our data spanned from Jan. 2015 to Feb. 2019.

4chan:

4chan is an image-board website [8], where users post anonymously. Like subreddits and subverses, most of its 63 boards are thematic in nature. Unlike reddit or voat, 4chan is completely anonymous; pseudonyms are optional but rare. We assume that each comment is posted by a unique user. We downloaded the political discussions (/pol), the travel (/trv) and television(/tv) boards data from an archiving service, 4plebs[7] which goes back to Dec. 2012. We label these as /4c/politics. /4c/travel and /4c/television. Like reddit, 4chan also has an extremely high volume of data. We use a 1% sample of each board. Table 2.1 summarizes our full dataset and a reduced 10k-subset. Unless otherwise specified all results are on the full dataset.

Pseudo-documents:

Our goal is to study the style of a community and not of its specific individuals. Thus, we disregard user identity and create pseudo-documents containing temporal chunks of comments. These pseudo-documents are representative of the community and used to train and test our community-level style and content classifiers. Each pseudo-document holds all comments posted in a single month by a community. As an example, for /4c/politics there are 63 pseudo-documents corresponding to 63 months of data. For some communities, the data can be sparse for a given month, as an example, only 4 comments were posted on the /v/travel forum during Jan. 2016.

[7] http://archive.4plebs.org/

Features

Most prior stylometry projects, such as Feng, Banerjee, and Choi [25], have relied on narrow definitions of style. Since there is no *a priori* theoretical reason for selectivity in style features we use a wide array of features identified from prior literature. Crucially, we exclude features that might also convey information about topic or content. Table 2.2 presents our four style categories.

Readability:

These measure the ease with which one can expect a text to be comprehended. Included are features such as CLI [16], and the Gunning fog index [28]. Prior literature, including Potthast et al. [65], have used a similar set of readability features for stylometric analysis.

Parts of Speech:

These capture syntactic properties of style by calculating the distributions of various parts of speech like nouns and verbs.

Character level features:

These measure orthographical style properties and include features such as the use of white space, punctuation, and emojis.

Word level features:

These assess the diversity and range in vocabulary used (but the specific words appearing in the text are not included as features). We include LIWC in this category as its logic is mostly word dictionary driven, though it also includes measures for properties such as short word counts. LIWC measures the proportion of words which fall into one or more hierarchical categories and their subcategories [80].

Some LIWC categories have the capacity to *leak* information about the text's topical content. We take precautions to avoid including such features

when building style-based classifiers. We do this in order to make a fair comparison between style and content for predicting community membership. For example, if a document scores high on the LIWC category 'Religion' this can indicate that the posts are related to religion. Likewise, the category 'Personal Concerns' has subcategories such as 'money' which includes words related to finance. Again, these potentially indicate post topic.

We create two reduced versions of LIWC which exclude topic leaking categories. In LIWC-$v2$, we remove 9 features belonging to 'Psychological Processes' category (e.g., 'ingestion' which includes words like eat, pizza etc) and 7 from the 'Personal Concern' category (e.g., 'religion') as these almost certainly convey topic. In LIWC-$v3$, we further eliminate 10 features which we suspect *indirectly* leak content - such as sentiment (under 'Affective Processes' category) and 'Perceptual Process' categories (words like touching). Again we take these precautions in order to get a clearer understanding of style versus content in communities. We suggest that this level of caution in avoiding inclusion of topical features in style is one of the strengths of our work. The full LIWC (LIWC-$v1$) has 64 features while LIWC-$v2$ and our most conservative LIWC-$v3$ have 48 and 38 features respectively. Table 2.3 summarizes the features retained in each LIWC version. We use our most conservative LIWC-$v3$ unless otherwise specified.

We measure style features at multiple granularities: word, sentence, post and (pseudo-) document levels as appropriate. For example, the prevalence of conjunctions may be the average of frequencies across posts or sentences or a single proportion of the total words in the document. Some measures are only suitable at the document level. Our final feature set, with LIWC-$v3$, has 262 features in total. Table 2.2 summarizes the features we explore and their granularity.

2.5 Results

The Style of a Community

The style of a community is represented by its dominant stylistic features. Table 2.4 presents the features with the top ten representative scores (Eq. 2.1)

	politics			travel			television		
	voat	**4chan**	**reddit**	**voat**	**4chan**	**reddit**	**voat**	**4chan**	**reddit**
	SYM	swear	RB	simpson	yule	pronoun	auxverb	negate	totChar
	you	inhib	VB	tabs	simpson	IN	FW	adverb	CLI
	nonfl	MD	DT	RBS	VB	TO	VBG	excl	charWord
	digits	we	ipron	filler	upper	adverb	NN	preps	WRB
	WDT	EX	funct	SYM	NNP	VBG	emoji	IN	alphabet
	UH	JJS	VBN	WP$	VBG	preps	upper	VBG	ARI
	assent	ppron	punct.	emoji	swear	past	VBP	short	MD
	VBN	shehe	VBZ	assent	you	VBN	verb	WRB	VBN
	auxverb	you	PRP	shehe	VBP	NNS	NNS	PRP$	CC
	cause	assent	NN	JJR	lines	VBD	VBN	shehe	VBP
Range									
Top 10	(0.085-0.105)	(0.058-0.072)	(0.052-0.059)	(0.096-0.100)	(0.067-0.075)	(0.056-0.062)	(0.081-0.094)	(0.053-0.060)	(0.055-0.058)
Full set	(0.085-0.224)	(0.058-0.241)	(0.052-0.208)	(0.096-0.312)	(0.067-0.281)	(0.056-0.210)	(0.081-0.239)	(0.053-0.246)	(0.055-0.180)

Table 2.4: Top ten representative features for each community.

for each community. Since features calculated at multiple granularities in spirit aim at the same stylistic property, we rank feature types based on the best representativeness score across granularities. The table shows the range of standard deviations as well.

Most (42 of 61, 69%) of the top 10 representative features apply to a single community. 12 features (20%) apply to 2 communities. Verb feature tend to occur commonly, such as 'VBN' (past participle verbs) appearing in five communities and 'VBG' (participle verbs) appearing in 4 communities.

Standard deviations are low:

In general, the standard deviations of the ten most representative features, for all communities, have fairly narrow ranges staying mostly between a low of 0.053 to a high of less than 0.1. Thus these top properties are fairly consistent across the pseudo-documents of a community. Voat communities had the highest range while Reddit communities tended towards the lowest ranges. As expected the standard deviation ranges increase when considering all features as shown in the table.

Representative features within each category:

Here we identify the single feature, within each of the four categories, that has the best rank for representatives on average across the 9 social media. Table 2.5 provides the mean score and standard deviation for these top features. We also include information for LIWC-v3.

Category	Feature	Mean(σ)
Readability	CLI	6.98(0.967)
Words	simpson	0.992(0.004)
POS	VBG/Word	0.025(0.002)
Character	word length	4.26(0.115)
LIWC-*v*3	ppron/words	0.0421(0.006)

Table 2.5: Representative feature in each category.

An interesting feature here is 'CLI' under readability. A score of 6.98 would indicate that the language of social media is simple enough that an individual with almost 7 years of education can understand it. The 'CLI' was the lowest for /4c/television at 5.31 and the highest for /r/politics at 8.81. This would indicate that a Middle School freshman can understand the language on /4c/television, while an individual would need to be almost a high school freshman to understand the language on /r/politics.

Styles are different across communities.

We rank all the style features by the representative score for each community and compare communities to identify style similarities using Spearman's rank-order correlation (Fig. 2.1). 17 of 36 community pairs have weak correlation (less than 0.2) while 15 pairs are weak to moderate correlations (greater than 0.2 and less than 0.4). The three highest correlations, while still at best moderate, all involve /4c/politics. This would indicate that these communities share representative style features - but only to a slight extent.

Takeaway:

Our answer to **RQ1** is that communities do have their own style which is identifiable and that style differs across communities independent of topic and medium.

Distinctive Style Features

Community specific distinctive features:

Table 2.6, shows the top ten features that make each community distinct. Again we consider the best distinctiveness score for a feature type independent of granularity. The table also shows the range of distinctiveness score for the feature (ΔC score). This ranges from 0.075 to 0.178.

Of the 49 distinct features in the top 10 lists, 23 (47%) and 17 (35%) were listed for one or two communities respectively. The use of the present tense was the most common distinctive feature in 5 communities. Brunét's W

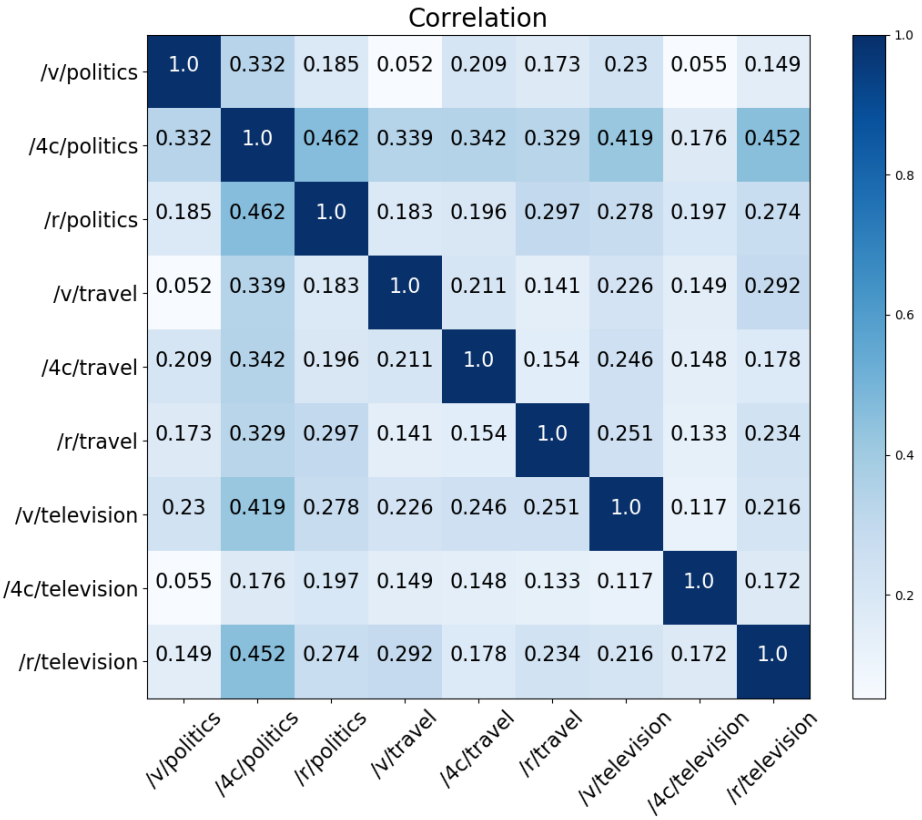


Figure 2.1: Spearman correlation between communities. Each community is represented by a ranking of features by representativeness score.

	Community Specific, ΔC									Global, ΔG	
	politics			travel			television				
	voat	**4chan**	**reddit**	**voat**	**4chan**	**reddit**	**voat**	**4chan**	**reddit**	**Most**	**Least**
	VBP	they	VBP	oov5000	verb	TO	oov5000	CD	RBR	VBP	lines
	brunet	WP	honore	oov1000	present	assent	auxverb	CC	insight	verb	tabs
	you	UH	insight	oov500	honore	preps	VBP	totChar	TO	brunet	emoji
	swear	brunet	VBZ	hapax Leg.	article	VB	verb	EX	we	oov5000	LS
	oov500	verb	they	brunet	VBP	CD	future	present	certain	oov500	FW
	CLI	VBZ	JJ	NNS	adverb	oov500	JJ	auxverb	WP	WP	yule
	shehe	auxverb	preps	future	CLI	inhib	article	verb	adverb	oov1000	simpson
	EX	present	ppron	MD	NN	present	brunet	shehe	upper	present	elongation
	ppron	oov500	WP$	quant	hapax Disleg.	MD	i	VBZ	article	auxverb	filler
Range	present	NNP	VBG	WP	they	EX	past	SMOG	NNP	insight	NNPS
Top 10	(0.118-0.105)	(0.092-0.075)	(0.116-0.097)	(0.148-0.132)	(0.090-0.081)	(0.103-0.092)	(0.134-0.119)	(0.178-0.166)	(0.165-0.156)	(0.019-0.016)	(0.0001-0.0033)

Table 2.6: Top ten distinctive features and scores for each community.

Measure was one of two appearing as distinctive for 4 communities.

Globally distinctive features:

Table 2.6, also shows the most and least globally distinctive features (ΔG score). Verb categories, Brunét's W Measure and insight are amongst the most distinctive globally. As expected when we rank all features by the standard deviation of their mean scores across communities, 9 of the 10 highest deviation features are amongst the most globally distinct. These deviations range from 0.2445 to 0.2673. The one exception, auxiliary verbs (auxverb) - the 9th most distinct feature - was ranked 12 by deviation.

Consistency between community-specific and globally distinct features.

Not surprisingly, there are consistencies. For example, the most frequently occurring community-level distinctive feature is the use of present tense which is also globally distinctive. Similarly, globally the most distinctive feature, the use of present tense singular verbs (VBP) is also distinctive for 4 communities. At the tail end, the use of linebreaks is least distinctive (see Table 2.6). It is also least distinctive for 8 of our 9 communities (not shown in tables).

Difference between distinctive and representative features.

We observe that features which are extremely representative of communities generally are not distinctive, either globally or at the community level. Features such as Yule's K Measure were in the top 10 representative features for /4c/travel and /v/travel, however, globally they were amongst the least distinctive features. Even at a community level, Yule's K Measure is not a distinctive feature, ranking as the 12th and 18th *least* distinct feature for /4c/travel and /v/travel respectively. Similarly, while the use of swear words is a representative feature for /4c/politics and /4c/travel (Table 2.4), it is not a distinct feature for either community. Instead, it is a distinct feature for /v/politics. These indicate that representativeness and distinctiveness are different properties.

Takeaway:

In summary, we find that there is consistency between community-specific and globally distinct features. And representativeness and distinctiveness appear to capture different properties.

Case Study:

The two thematically similar communities /r/television and /4c/television have vastly different discourse styles. Here we analyse their language complexity and their use of negations.

Community	**Style**				**Content**			
	Accuracy(%)	Precision	Recall	F-score	Accuracy(%)	Precision	Recall	F-score
voat	92.10	0.901	0.921	0.911	85.47	0.832	0.855	0.843
/v/politics	98.00	0.987	0.980	0.983	92.00	0.993	0.920	0.955
/v/travel	83.96	0.802	0.840	0.820	77.57	0.654	0.776	0.709
/v/television	91.95	0.890	0.919	0.904	84.56	0.834	0.850	0.840
4chan 1%	97.93	0.983	0.979	0.981	97.27	0.953	0.973	0.963
/4c/politics	98.95	0.979	0.989	0.984	96.83	0.989	0.968	0.979
/4c/travel	97.72	0.972	0.977	0.975	97.73	0.869	0.977	0.920
/4c/television	97.22	0.995	0.972	0.984	97.30	1.00	0.973	0.986
reddit 10%	96.87	0.979	0.969	0.974	94.99	0.982	0.950	0.966
/r/politics	95.19	0.985	0.952	0.968	96.10	0.987	0.961	0.974
/r/travel	97.87	0.966	0.979	0.973	93.33	0.966	0.933	0.949
/r/television	97.37	0.987	0.974	0.980	95.61	0.995	0.956	0.975
Average	96.08	0.951	0.954	0.952	93.50	0.921	0.923	0.922

Table 2.7: Performance on predicting community membership (Full set).

/4c/television has the lowest average 'CLI' of all communities at 5.31, whereas /r/television has a relatively high average 'CLI' of 7.44. Compared to a user on /4c/television, a user on /r/television would require an additional 2 years of education to understand the discourse. Similarly, the 'SMOG' score (19.4 versus 14.5) and 'ARI' (5.4 versus 2.4) also indicate that the language on /r/television is more complex than on /4c/television.

Negations, on average, occur once every 4 sentences on /r/television, and less frequently, once every 8 sentences in /4c/television. The standard deviation of negations, for /4c/television is half that for /r/television, indicating more consistency in the rarer use of negations for the 4chan community. Drawing from Gonzales et al. [27] which indicates that a higher use of negations within a community is indicative of lower cohesion between its members, we suggest that there is less disagreement amongst /4c/television users. This in turn might indicate a stronger shared group identity amongst 4chan users.

This notion is further supported by differences in the use of first-person pronouns. We expect communities with a stronger sense of group identity to use these less frequently than communities where there is more individualism. We observed that users are 1.7 times more likely to use the first person pronoun on /r/television than on /4c/television. Possibly the increased emphasis on individualism also leads to greater disagreement (negation) on /r/television compared to /4c/television.

In sum, we note that users on /4c/television appear to have a stronger sense of group identity and they express themselves using a less complex language. This case study illustrates that by analyzing style we can gain deeper insights about specific communities and the differences between them.

Predicting Community Membership

Table 2.7, presents the results of our prediction experiments addressing our research question **RQ2**.

Community	Style		Drop		Content		Drop	
	Acc.(%)	F1	%Δ Acc.	%Δ F1	Acc.(%)	F1	%Δ Acc.	%Δ F1
voat	78.57	0.817	-14.69	-10.32	87.71	0.878	2.63	4.15
/v/politics	90.48	0.884	-7.68	-10.12	90.48	0.896	-1.66	-6.22
/v/travel	62.73	0.704	-25.29	-14.17	88.29	0.831	13.82	17.07
/v/television	78.52	0.823	-14.60	-8.89	84.56	0.900	0	7.14
4chan 1%	88.60	0.886	-9.53	-9.53	69.23	0.742	-28.8	-22.95
/4c/politics	89.73	0.855	-9.31	-13.06	69.19	0.703	-28.54	-28.13
/4c/travel	89.77	0.916	-8.14	-6.01	68.18	0.734	-30.23	-20.20
/4c/television	86.70	0.894	-10.83	-9.15	70.18	0.785	-27.87	-20.45
reddit 10%	92.89	0.907	-4.11	-6.89	91.57	0.866	-3.61	-10.36
/r/politics	94.14	0.915	-1.10	-5.53	95.05	0.885	-1.10	-9.14
/r/travel	93.01	0.924	-4.96	-4.98	86.90	0.834	-6.89	-12.09
/r/television	91.52	0.882	-6.01	-10.04	92.89	0.878	-2.85	-9.97
Average	87.89	0.868	-7.3	-8.83	82.85	0.831	-9.09	-9.97

Table 2.8: Performance on predicting community membership (10k subset).

Style is an excellent predictor of community membership.

With the exception of /v/travel, we observe accuracies higher than 90% and F-scores above 0.9. In almost all cases style results are numerically about the same or better than content for both accuracy and F-score. The biggest wins are seen in voat. For example, content (0.71 F-score) has an even harder time predicting for /v/travel than style (0.82 F-score). Community-level average scores are 0.95 for 4chan and reddit and above 0.90 for voat. While we expected some degree of success these strong results exceed our expectations. These again indicate that style is distinct to each community. Comparing average scores (last row of Table 2.7) we find content and style to be statistically equivalent ($p>0.05$)[8]. Thus style and content are equally excellent at predicting community membership. Thus, our answer to **RQ2** is that style excels in predicting community membership.

We note that substituting LIWC-$v2$ for LIWC-$v3$ drops performance an insignificant amount (accuracy 95.84, F-score 0.95). This tells us that prediction performance is unperturbed if we ignore the 'Affective' and 'Perceptual' process categories of LIWC. In particular this includes sentiment, commonly used in prior style research.

[8]We use Smucker et al.'s [74] bootstrap test of significance in this paper.

Analysis of misclassifications:

The biggest losses were from a content perspective: /v/travel was misclassified as /v/television (8.41%) as /r/travel (7.48%) and as /4c/travel (5.61%). The topical overlap between /v/television and /v/travel and between the various travel communities seems to be challenging for content-based prediction. Style also confused /v/travel but the errors were fewer. Style mapped /v/-travel to /v/television(7.56%), /r/travel(3.77%) and /4c/travel(0.94%). The style confusion between /v/travel and /v/television could be because of style similarities at the medium level, an aspect to probe in further research. As an aside, the politics communities are highly distinguishable across media via content which leads us to infer that they are sufficiently different in topic.

Results on a 10k-subset:

A possible explanation for the relatively weak results for content-based prediction in /v/travel could be the low volume of data (only 742 comments, Table 2.1). All other communities had at the least 5,000 comments or greater, with a majority having more than 100,000 comments. To test this, we repeated the prediction experiments by downsampling each dataset to around 10K posts (10k-subset). If our intuition is supported then content based performance should improve for /v/travel. With style classifiers, it is unknown what is likely to happen.

Table 2.1 describes the 10k-subset and Table 2.8 compares performance between the full and the 10k-subset experiments. When sizes are more comparable, we observe 14% and 17% improvements in accuracy and F-score respectively for content-based classification in /v/travel. Likewise, content performance also improves (but for F-score only) for /v/television - the community with the second smallest dataset. These support our intuition that content based prediction for voat was limited in the original experiment by the small dataset size (relative to the other communities).

Interestingly, the content classifier degrades markedly for 4chan in the reduced dataset, but not as much for reddit. Taking the politics communities as an example, we note that in the full set, /4c/politics shared only 1.90% of

Window Size	Style				Content			
	Accuracy(%)	Precision	Recall	F-score	Accuracy(%)	Precision	Recall	F-score
Half Month	93.81	0.916	0.908	0.912	93.08	0.911	0.914	0.912
One Month	*96.08*	*0.951*	*0.954*	*0.952*	*93.50*	*0.921*	*0.923*	*0.922*
Two Month	94.56	0.937	0.930	0.933	93.14	0.923	0.926	0.925

Table 2.9: Performance on predicting community membership for varying time windows.

its vocabulary with the /4c/television. In contrast, /r/politics shared 11.62% of its vocabulary with /r/television. In the 10k set, the former percentage increased by almost 11 times compared to just an increase of 2 times for the latter. This observation is consistent with the classic guideline that training data size matters when building content based classifiers.

In contrast, for style-based prediction, this dependence is not as severe. While performance drops with style for all communities, 4chan style based prediction is markedly more resilient to downsizing than 4chan content based prediction. With reddit, the drops in style-scores and content-scores are about even across the two measures. Overall, style performance drops are less for style (7.3% in accuracy and 8.83% in F-score) than for content (9.09% accuracy and 9.91% F-score). However, statistical tests on the average scores indicate once again that style and content are equivalent in predicting community membership.

Takeaway:

The answer to **RQ2** is that prediction based on style gives excellent results which are on average statistically equivalent to prediction using content.

Additionally, style classifiers appear less sensitive to reductions in training data.

2.6 Additional Analysis & Discussion

Feature Ablation Analysis

We find that even when ignoring any one of the four style categories listed in Table 2.2, style on average numerically outperforms the content-based model

(with >1.2 million features). The lowest performance obtained was when we excluded all LIWC-$v3$ features and used the remaining set of 148 features (accuracy 94.99%, F-score 94.18). Table 2.11 summarizes these results.

When used individually the word-level features (135 features) and LIWC-$v3$ (114 features) are the only ones to numerically beat content. The worst performer is readability (6 features), but the scores are still good: accuracy 86.97%, F-score 84.40. Table 2.12 summarizes these results with full feature set results given in the first two rows.

Effect of Window size on Results

In our main results we had used 1-month as the window length to create pseudo-documents. Table 2.9 shows the effect of using 2-month and half-month windows compared to the one-month windows. It can be seen that compared to the content classifier, the accuracy of the style classifier exhibits slight variations. However, even with varying window sizes, the style and content classifiers continue to be statistically equivalent.

Sport	**Community**	**Months**	**# of comments**	**Time Period**
Soccer	/r/soccer	51	3,449,070	Dec'14 - Feb'19
Football	/r/nfl	51	5,299,037	Dec'14 - Feb'19
Hockey	/r/nhl	62	32,815	Jan'15 - Feb'19
Basketball	/r/nba	51	4,050,545	Dec'15 - Feb'19

Table 2.10: Summary of the Sports Dataset

Thematically Similar Communities

We now ask if our findings hold for a seemingly harder problem - communities stem from the same broad topic and same platform. Data for four thematically similar sports-based communities on reddit (/r/nhl, /r/nfl, /r/nba, /r/soccer) is described in Table 2.10.

Again performance is excellent and not statistically different for style and content classifiers. Average accuracies and F-scores are 98.15% and 0.982 for

style and 98.72% and 0.987 for content classifiers. Thus our style classifier is capable of distinguishing between close communities.

Platform Level Style

Perhaps there are platform level (viewed as supra-communities) stylistic patterns that influence the member subreddits, subverses and 4chan boards. The greater the stylistic influence of a platform on a community the higher the number of shared representative features likely. We find that /4c/travel and /v/travel were stylistically most similar to their platforms since both shared 7 of their top-10 features with the parent community. In contrast, /r/television was stylistically the least similar to its parent reddit community with no shared feature. The remaining six communities were also distinct, with 2 or less features shared with the parent platform-level community.

Amongst the shared features we see that the use of 'swear words' is representative of 4chan in general, and also of both /4c/politics and /4c/travel. While the use of agreeable words ('assent' category) was representative for the voat platform and both /v/politics and /v/travel. In contrast reddit did not have a platform level feature which was shared by more than a single sub-community.

Category excluded	**# of features**	**F-Score**	**Accuracy**
Parts of Speech	184	96.02	96.68%
Word level features w/o LIWC	241	95.02	95.84%
Readability	256	94.65	95.60%
Character Level Features	219	94.56	95.48%
LIWC-$v3$	148	94.18	94.99%
Content	*1,211,385*	*92.20*	*93.50%*

Table 2.11: Performance when feature category is excluded.

Platform Level Readability Scores

Finally we explore nuances related language complexity at the platform level. Ranking the three platforms by language complexity we observed that rank-

Category included	# of features	F-Score	Accuracy
All (LIWC-$v3$)	262	95.20	96.08%
All (LIWC-$v2$)	282	95.02	95.84%
Word level features (includes LIWC-$v3$)	135	94.17	95.24%
LIWC-$v3$	114	93.54	94.51%
Content	*1,211,385*	*92.20*	*93.50%*
Word level features w/o LIWC	21	90.80	92.10%
Parts of Speech	78	90.49	91.78%
Character Level Features	43	88.88	90.59%
Readability	6	84.40	86.97%

Table 2.12: Performance when the category of features listed is the only one included.

ings were largely the same across the 7 readability measures we used (Table 2.2). From the most to least complex we have: reddit, voat then 4chan.

The Dale-Chall readability measure differed slightly, ranking voat as having the most complex language followed by reddit then 4chan. The ARI and CLI rely on character count to measure language complexity. The Dale-Chall readability index, SMOG and the Gunning Fog index all measure the ratio of complex words to the total words. While, both SMOG and Gunning Fog define complexity based on the number of syllables, Dale-Chall readability index defines complexity as based on a list of 3,000 predefined simple words. This difference in definitions and the small size of 3,000 explain why it gives a slightly different readability result. We note that this ordering is fairly consistent with the complexity based ordering of the individual communities analyzed. Voat showing the most variation (relatively) indicates lower platform-level consistency.

2.7 Conclusion

An individual's linguistic style is known to develop through subconscious processes while vocabulary to express content can be acquired through a deliberate conscious process. We suggest that community-style is likely also acquired or shaped through a 'subconscious' process that occurs through the interactions between community members. In contrast to most of prior research, our communities of interest are made up of individuals interacting because of

shared interests. We find that communities have distinctive style and that we can use 200+ style features to successfully predict community membership. Additionally, style based classifiers are able to predict community membership as well as content-based classifiers.

We made use of pseudo-documents as representative of the group's language, by doing so we assume that changes in style are minimal across time. Additionally, our study is limited to a few communities within 3 social media platforms. We will address these limitations in future research.

Our analysis focuses on traditional stylometric including lexicon based methods like LIWC. These methods do not explictly focus on the sensorial language dimension of style. In the subsequent chapters, we propose methods that overcome this limitation of traditional style and measure sensorial style more directly.

Chapter 3

EXPLORING SENSORIAL STYLE

3.1 Introduction

Sensory perceptions shape how we use language and communicate [61]. When we use sensorial words (i.e. words with meanings connected to our senses) like *fuzzy* or *stinky*, besides communicating sensorial experiences these also stimulate perceptual systems in the recipient's mind [77].

The space of senses – sometimes called the "Aristotelian" senses [75], include the five modalities: visual, auditory, haptic, gustatory, and olfactory. Relatively recently, linguistics and psychologists have added a sixth sense – interoception [19]. This refers to the perception of sensations from inside the body, both physical such as hunger and pain, and emotional, such as joy. This sensory space has been the basis of much prior research.

Sensorial Linguistics, emerging from psychology and cognitive science, is about studying how language relates to the senses. A key focus has been to study how different sensorial experiences and perceptions are packaged into linguistic units [87]. Researchers have looked at how some senses dominate in language [89], how sensorial language varies across lexical categories [44] and how sensory experiences influence sensorial language [20, 59]. However, the domain of sensorial linguistics is still nascent with many unexplored questions.

Style in sensorial language use:

Meanwhile, in computer science, stylometrics has developed as a computational approach to analyzing patterns in language use. A key limitation in stylometrics is that linguistic style around sensorial language has not been studied systematically. A person who wants to express being depressed has several word choices. She could use "sad" or the less frequent "downcast".

[0]This work has been published at the Proceedings of the 29th International Conference on Computational Linguistics 2022 [37]

Her propensity to choose one or the other may be considered as part of her linguistic style.

Consider a cloudy scene. A person may use visual language focusing on color, and say "*the clouds are white*". Another may use haptic language focusing on texture, "*the clouds are fluffy*". While sensorial language is clearly important for communication, we do not yet know if there are distinguishable patterns in sensory language use at the level of individuals, texts, etc. This gap in stylometrics motivates us to ask the following about sensorial language style:

- **RQ 1:** Is the notion of sensorial style meaningful or is it a product of random chance?
- **RQ 2:** How much data do we need to get a stable representation of sensorial style?
- **RQ 3:** Does sensorial style vary with time?
- **RQ 4:** Which features are representative and distinctive of the individuals within each genre?

3.2 Background

Sensorial language: Sensorial language is not uniformly distributed across the six sensory modalities as reflected in sensorial lexicons, such as the one we use [48]. This is also observed in large text collections. In their analysis of 8 million words from around 7,000 English texts Koblet and Purves [39] found over 28,000 visual descriptions and only 78 referring to the olfactory modality. Similar findings for multiple corpora are observed in Winter, Perlman, and Majid [88]. Our results are consistent with these prior works.

Sensorial language, the brain & emotion: The salience of sensorial words is known to be highly correlated with the volumes of cortical activation in the brain [68]. Lievers [41] show that there is directionality to how senses are substituted for each other.

Winter [86] found that gustatory and olfactory words (e.g., 'stinky', 'delicious') are on average more emotionally valenced than visual and auditory words and these also appear in more emotionally valenced sentences.

Bubl et al. [11], show that alterations in mental states have a direct effect on perception; specifically, that depression directly impacts how the color blue was perceived. Kernot et al. [35], found a decrease in the novelist Iris Murdoch's use of olfactory language following her diagnosis of depression and Alzheimer's. We credit this study for providing us with the hint that sensorial language may lead to a sensorial style. We take their sensorial style analysis forward with larger collections of authors, several genres and a more informative representation of sensorial style.

3.3 Representing Sensorial Language Style through Synaesthesia

Sensorial style may be represented at different levels of abstractions. At the lowest level, we can represent the proportion of an individual's language that is sensorial and also examine the frequencies of different sensorial words. At a higher level of abstraction, we can ask how frequent are different sense modalities (visual, auditory, etc.) in an individual's language. Alternatively, we can represent style by the extent to which a person's use of sensory modalities aligns with general expectations. This is related to synaesthesia, where one sense modality is used when another is expected - a well studied phenomenon in sensorial linguistics [41, 23]. As an example in their work, Lievers and Huang [42] developed a lexicon of perception that they used to automatically identify perception related synaesthetic metaphors. Similarly we also approach the problem of sensorial style through the lens of their synaesthetic usage.

In the 2010 animated film 'Despicable Me', the character of Agnes hugs a unicorn and says "It is so *fluffy*". One reason why this quote acquired somewhat of a meme status is because it subverted audience expectations of a more visual word like "*pretty*" or "*white*" to describe the unicorn. Instead she opts for the more unexpected haptic word "*fluffy*". This substitution of visual language for haptic, a synaesthesia, might indicate that Agnes' per-

ceives the world in a more tactile manner rather than in a visual way. In order to assess if this is a stylistic tendency, we can examine all of Agnes' language use and ask the general question: to what extent does she use haptic language in contexts where we generally expect visual language? We can ask similar questions related to each combination of expected versus observed sensory modalities. Observations for all combinations, including the homogeneous non-synaesthesia ones, are accumulated to form Agnes' (or any other individual's, group's or genre's) sensory style representation.

Terminology and notation: More formally, we consider a sentence to be a "sensorial sentence" if it has at least one word or phrase that appears in a sensorial lexicon. Further, we define a "sense-focused sentence" to be a sensorial sentence with a single sensorial term selected as focus term. Thus, if a sensorial sentence has n sensorial terms then we derive from it n sense-focused sentences.

Assume $S_i = \{S_{i1}, S_{i2}, ..., S_{in}\}$ is the set of n sense-focused sentences identified from the writings of individual $i \in I$. Let $C = \{H, V, I, O, G, A\}$ represent the modalities: Haptic, Visual, Interoceptive, Olfactory, Gustatory, Auditory respectively. Using $\bar{N}$ to represent concepts that are "not-sensorial" we define $\widetilde{C}$ as: $\widetilde{C} = C \cup \{\bar{N}\}$.

We also define two functions. $F(\hat{s}_{ij})$ is a sensory lexicon lookup function that returns the sensorial category $c \in C$ for the focused sensorial term $\hat{s}_{ij}$ in the sense focused sentence S_{ij}. E.g., given the sensorial sentence "*The unicorn is **white** and **fluffy***", we have two sense-focused sentences S_{ij} and S_{ik} corresponding to the two sensorial terms, $\hat{s}_{ij}$ and $\hat{s}_{ik}$. For S_{ij} with the focus term "***white***" $F(\hat{s}_{ij})$ will return V. For S_{ik} with focus term "***fluffy***" it will return H.

The second function we define is $M(S_{ij})$ which returns the "expected" modality $c \in \widetilde{C}$ for the same focus term in S_{ij}

We describe this function in Section 3.3.

Calculating observed to expected ratios: We represent an individual's sense-focused sentences as a list of length $|S_i|$. Each entry is a pair of expected and observed modalities of the form $[(M(S_{ij}), (F(\hat{s}_{ij})]$. For observed modality $y \in C$ and expected modality $x \in \widetilde{C}$, the observed to expected ratio α_i^{xy} for

individual i is:

$$\alpha_i^{xy} = \frac{|\{S_{it} : F(\hat{s}_{it}) = y \text{ and } M(S_{it}) = x\}|}{|\{S_{it} : M(S_{it}) = x\}|} \tag{3.1}$$

Note that these ratios are the informative units of sensorial style. For example, if the ratio is 1 when x and y are the same modality, a homogeneous combination, then the individual's use of that modality is highly aligned with general expectation. On the other hand, if it is close to 0 then she deviates considerably from the expected use of modality x, i.e. there is greater synaesthesia.

Style vectors: For each $x \in \widetilde{C}$, we then concatenate its 6 ratios into a vector of the form:

$$s_i^x = \underset{y \in C}{\sqcup} \alpha_i^{xy} \tag{3.2}$$

It follows that

$$\sum_{y \in C} \alpha_i^{xy} = \begin{cases} 1, & \text{if } |\{S_{it} : M(S_{it}) = x\}| \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

We can now define the sensorial style vector u_i of $i \in I$ as a concatenation of the seven vectors, one for each expected modality. The size of u_i is $x \times y = 42$.

$$u_i = \underset{x \in \widetilde{C}}{\sqcup} s_i^x \tag{3.3}$$

Implementing function $M(S_{ij})$

Given a sense-focused sentence, function M returns the expected modality $c \in \widetilde{C}$ of the sentence's focus sensorial term as per general expectation in English. We leverage RoBERTa-MLM[1] as a stand-in for general English language usage. RoBERTa is a transformer based language model pre-trained on 160 GB of data [45]. Prior works like Mosbach et al. [58] and Sinha et al. [73] have shown that language models learn the norms of the language on which they are trained.

This makes them ideal for our task. We mask the focus sensorial words in our sentences and input them to the model. RoBERTa returns the probabilities for all the words in its vocabulary at each masked location. Probabilities

[1] We experimented with BERT as well, however, RoBERTa gave us more accurate results.

represent likelihood of appearance of the words at that location. We use these probabilities to identify the expected sense modality at each masked location as follows.

Let $W = \{(w_1, p_1), (w_2, p_2) \dots (w_N, p_N)\}$ be the ranked set of words returned by RoBERTa for a masked location (location of focus sensorial term) in S_{ij}; top ranked has highest probability.

Using $F(w_k)$, we lookup the sense for each word in the top 100. We combine this information to get an aggregate probability score $\Pi(c, S_{ij})$ for each modality c as follows:

$$\Pi(c, S_{ij}) = \sum_{\substack{k \leq 100 \\ F(w_k) = c}} p_k \tag{3.4}$$

For greater confidence we only include S_{ij} in our analysis if its majority modality has $\Pi(c, S_{ij}) > 0.5$. We then define

$$M(S_{ij}) = \underset{c \in \widetilde{C}}{\operatorname{argmax}}\, \Pi(c, S_{ij}) \tag{3.5}$$

In essence, $M(S_{ij})$ returns the expected modality with the highest aggregate probability for the focused sense word in S_{ij} as determined using RoBERTa.

Sensorial Lexicon

	Lexicon			
	Original		Modified	
Modality	*N*	%	*N*	%
Visual	29552	75.0	9419	50.2
Interoceptive	3546	9.0	3449	18.4
Auditory	4528	11.5	3803	20.3
Haptic	675	1.7	972	5.2
Gustatory	890	2.3	890	4.7
Olfactory	216	0.5	216	1.2
Total	39407	100	18749	100

Table 3.1: Distribution of modalities in original Lynott et al. [48] lexicon and our modified subset lexicon.

We use the Lancaster sensorimotor norms lexicon [48], which has 39,954 concepts from the English Language. Brysbaert et al. estimates that the average adult lexicon is composed of approximately 42,000 words [10]. Therefore this lexicon approximates a significant majority of everyday English.

Each concept was rated by annotators along a 0-5 scale for the six modalities (Auditory, Gustatory, Haptic, Olfactory, Visual, Interoceptive) For example "*fluffy*" is rated 4.41 for Haptic, 0.29 for Gustatory, 3.77 for Visual, 0.35 for Interoceptive and 0 for Auditory and Olfactory.

The dimension with the highest rating is the dominant modality.

Dominance alone is not enough to ensure that a concept belongs to a particular sensorial modality since almost half of the concepts score less than 2.55 on any sense modality. Therefore, we filter the lexicon by ranking all concepts in a given modality by their rating and selecting only those in the top quartile. This ensures strong alignment to dominant modalities. Table 3.1 describes the lexicons[2].

3.4 Methods

RQ 1: Is Sensorial Style a Product of Random Chance?

In order to be meaningful our representation of sensorial style should not be a product of random chance.

If an individual chooses sensorial modalities randomly and not deliberately then we expect her observed and expected modality distributions to be independent of each other.

For example in "*the clouds are* ***white***" the expected modality may be visual but the individual randomly chooses from one of the six senses. We use this random model in our analysis.

As a first step, for any $i \in I$ with a set of n sense-focused sentences S_i, we define, $\Gamma(i)$, the distribution of the sense modalities in C observed in S_i. For each sense-focused term $\hat{s_{ij}}$ in S_i, we use a function $\bar{F}(\hat{s_{ij}}, \Gamma(i))$ that returns a random modality $c \in C$ with distribution $\Gamma(i)$.

For each $i \in I$, we create m random pseudo-documents $\mathcal{R}_i = \{\mathcal{R}_i^1, \mathcal{R}_i^2 \dots \mathcal{R}_i^m\}$. Each random pseudo-document has the same set of sense-focused sentences S_i. However, instead of using $F(\hat{s}_{ij})$ to look up the modality, we use $\bar{F}(\hat{s}_{ij}, \Gamma(i))$ to

[2]There are other sensory lexicons like Lievers and Huang [43] and Lynott and Connell [47]. Besides being the largest and most recent the Lynott et al. [48] sensorimotor lexicon is the only one to include ratings for interoception.

get a random modality. Equation 3.6, a modification of equation 3.1, gives us $\bar{\alpha}_{ik}^{xy}$ which is used to calculate the style vector u_i^k for random pseudo-document $\mathcal{R}_i^k$.

$$\bar{\alpha}_{ik}^{xy} = \frac{\left|\{S_{ij} : \bar{F}(\hat{s}_{ij}, \Gamma(i)) = y \text{ and } M(S_{ij}) = x\}\right|}{|\{S_{ij} : M(S_{ij}) = x\}|} \tag{3.6}$$

Thus, for each $i \in I$ with sensorial style vector u_i, we have m random style vectors $\{u_i^1, u_i^2 \dots u_i^m\}$ generated from the random pseudo-documents in $\mathcal{R}_i$.

Let $U_i = \{u_i\} \cup \{u_i^1, u_i^2 \dots u_i^m\}$. For each vector $v \in U_i$, we calculate its average cosine similarity with all other elements in U_i. Ranking the elements of U_i by decreasing order of average similarity we check whether the style vector $u_i \in U_i$ has lower average similarity than at least 95% of the vectors in U_i (i.e. p-value < 0.05). If so, we infer with 95% confidence that i's style vector, u_i, is not random and therefore likely a product of an individual stylistic choice.

RQ 2: How much data is needed to describe sensorial style?

Given the set of sense focused sentences S_i, where $|S_i| = n$, we randomly sample subsets from S_i of size k and compute the style vector from each sample. We explore how increasing the values for k affect style convergence.

For a given sentence set size k, we identify m random samples (with replacement) of S_i, each of size k. Thus, we create a set of sentence sets, $T_i^k = \{T_{i,1}^k, T_{i,2}^k, \dots, T_{i,m}^k\}$. For each sentence set $T_{i,j}^k \in T_i^k$, we use the method discussed in Section 3.3 to generate the corresponding style vector $u_{i,j}^k$. This gives us a set of m sensorial style vectors, $U_i^k = \{u_{i,1}^k, u_{i,2}^k, \dots, u_{i,m}^k\}$. We then use cosine similarity to calculate the average pairwise similarities between all elements in U_i^k. We recompute this average self-similarity $\text{sim}(U_i^k)$ for different values of k in increasing steps of r.

We say that the style of the individual has converged for a minimum of k sensorial sentences if $\text{sim}(U_i^k) \approx \text{sim}(U_i^{k+r})$, where $k + r$ is the next sentence set size tested.

RQ 3: Does sensorial style vary over time?

Here we investigate whether style vectors evolve over time spans. We segment the writings by time and consider how the average similarity in style varies with temporal distance. We first identify all pairs of time points t_a and t_b that are γ duration apart. We then build style vectors for each author with text anchored at t_a and for each author with text anchored at t_b. We then compute the average pairwise similarity between t_a and t_b style vectors. We repeat this for all values of γ that are of interest.

We use a notion of windowing around the time points (t_a and t_b) to reduce noise. Each window is of size δ and distributed equally around each time point. For example, for t_a we create an individual's style vector from all the sense-focused sentences that were published in the range $t_a - \frac{\delta}{2} < \tau_a < t_a + \frac{\delta}{2}$.

RQ 4: Which features are representative and distinctive of the individuals within each genre?

A genre can be represented by the set of sensorial style vectors of its members. Each style vector is composed of 42 features that explore synaesthesia. We are interested in exploring which features are representative of the members of a genre, and also features that make the members distinct.

We consider a sensorial style feature to be representative if the variation in its usage is low. This would indicate that the members use the feature in a consistent manner. Formally, a stylistic feature α^p is more representative for the members of a genre than another feature α^q if its standard deviation $\sigma(\alpha^p)$, across all the members is lower than the standard deviation $\sigma(\alpha^q)$. At the other end, a high variation would indicate that the feature is distinctive amongst the members.

3.5 Datasets

We analyze 3 literary genres — novels, poems, music lyrics. Compared to poems and lyrics, novels can span tens of thousands of sentences. Additionally, novels and poetry are generally associated with a single author. Lyrics are

Genre	# Authors	# Works	#Sentences	# Sensorial Sentences	# Sensorial Expressions
Novels	130	317	1,525,894	156,570 (10%)	474,299
Lyrics	5,321	20,785	1,007,090	754,572 (75%)	1,501,501
Poetry	1,246	3,315	85,236	4,979(6%)	8,209

Table 3.2: Dataset details for each genre. The percentage of total sentences that are sensorial is in parentheses.

sometimes collaborations, however, we assume an artist would not perform a song that is in a style they do not like. Thus, we assume music lyrics to be a reflection of the artist's style.

Novels: We collected English language novels from the Domestic fiction genre of Project Gutenberg[3]. There were 317 works written by 130 authors, with the earliest by Henry Fielding from the early 18^{th} century and the latest by Rebecca West from the mid-20^{th} century.

Lyrics: We collected songs that were listed on the Billboard Hot 100 charts, 1963 to 2021 (inclusive). This weekly chart ranking of song popularity is considered the industry standard [85]. We assume the first time a song is listed to be its year of production. We obtained song lyrics using the Genius API[4]. There are $20,785$ song lyrics.

Poetry: Following works like [46], we used the corpus of poems available on the Poetry Foundation's[5] website. To make this dataset more comparable to the lyrics dataset, we only included works published after 1963.

3.6 Results

We present our results in two parts. First, we make our general observations. Second, we present results related to our specific research questions.

[3]https://www.gutenberg.org/

[4]Some lyrics were not available on http://www.genius.com, because they were instrumentals like the "*Star Wars Theme*" which hit No. 1 in 1977, or were not in the Genius database.

[5]The Poetry Foundation was established in 2003, and one of its goals is to make "the best poetry" accessible (https://www.poetryfoundation.org/foundation/about).

Genre	$> 95^{th}$	N
Novels	112	123
Lyrics	701	735
Poetry	20	85

Table 3.3: Number of individuals with lower average similarity than 95% of random vectors.

General Observations

	Novels	Lyrics	Poetry
	$I-O$	$I-O$	$A-G$
	$\bar{N}-O$	$\bar{N}-O$	$A-O$
	$A-O$	$V-O$	$G-I$
	$I-G$	$A-O$	$H-O$
	$V-O$	$I-G$	$O-A$
	$A-G$	$V-G$	$O-H$
Range	(0.00,0.01)	(0.00,0.02)	(0.00,0.00)

Table 3.4: The top representative features for each genre.

	Novels	Lyrics	Poetry
	$G-G$	$H-H$	$A-A$
	$O-O$	$G-G$	$I-I$
	$H-H$	$A-A$	$H-H$
	$G-V$	$H-V$	$\bar{N}-V$
	$H-V$	$G-V$	$G-G$
	$A-A$	$A-I$	$\bar{N}-I$
Range	(0.43,0.15)	(0.44,0.19)	(0.47,0.27)

Table 3.5: The top distinctive features for each genre.

Each feature is an (expected,observed) pair e.g., $I-O$ in means that we observe Olfactory language when we expected Interoceptive language. This is the most representative feature for Novels and Lyrics as it has the lowest standard deviation. We include the range of variances of the top most distinctive and representative features.

Domination over lower senses: The five Aristotelian sensorial modalities have classically been thought of as part of a hierarchy with vision and audition dominating over the three so-called "lower senses"—Touch, Taste, Smell [30][6]. This hierarchy manifests in the frequency of language use with the visual and auditory modalities being used more often than the lower senses [50]. We have consistent results. Figure 3.2 shows that for all three genres visual and

[6]Note interoception is generally not considered in discussions of this hierarchy, possibly because of its relatively recent inclusion [18].

auditory dominate over the "lower senses". Concepts associated with haptic, gustatory and olfactory modalities — combined, form less than 10% of the total sensorial language.

While auditory dominates the "lower senses" in all three genres, it occurs less than half as often as visual. Going beyond the classical five senses, in all cases interoception dominates the three "lower senses" surpassing audition in this regard. Additionally, in lyrics, interoception is as common as visual. Clearly interoception with its emphasis on sensations within the body, both physical and emotional, is important in language.

Sensorial style across genres: We investigate how sensorial style varies across genres. Using the method described in Section 3.3 we calculate genre-level sensorial style vectors by combining sentence sets at the genre level. We show the distribution over 42 sensorial combinations in Figure 3.1 for just Lyrics. Each cell represents an expected-observed modality combination.

Figure 3.1: Distribution of expected-observed modalities in Lyrics. Note that we calculate proportions using equation 3.1, however, for illustrative purposes we show sensorial distribution as percentages.

Observed modalities are largely as expected with some exceptions: The diagonal values which are in the range 56 to 88% indicate that the observed modalities are generally consistent with expected modalities. That is, the individuals in our datasets select from the 6 sensorial modalities in a manner that is consistent with the general norms of language use. The highest consistency is for interoceptive and the lowest is for olfactory. We observe this trend across all genres (see Appendix A).

Looking at off diagonal values, we observe visual language used in 25% of the cases where we expected olfactory language, 19% and 11% of the cases where we expected gustatory and haptic language, respectively. This usage of visual language as a replacement for lower senses was observed across all 3 genres. This replacement or *cross-modal compensation* might be because the lower senses do not have a strong relation with the perceptual system and consequently individuals might be relying on visual language as a semantic scaffold to compensate for the weaker perceptual system of the lower senses [77].

We also observe (in all three genres) that in more than 90% of instances where we expected to see non-sensorial language we instead observed interoception. This might also be because interoception dominates in our data and is consistent with observations about higher senses in the literature [50].

Results for research questions

RQ1: Is sensorial style a product of random chance? We investigate whether sensorial style is motivated by the individual choices or whether it is a product of randomness.

Table 3.3 provides the results. If sensorial style is a non random phenomenon and a product of individual choice and intent, we expect the sensorial style vectors to be distinct from random vectors. That is, we expect them to have a lower average similarity as compared to random vectors (generated by our random model).

Methodological details are in Section 3.4).

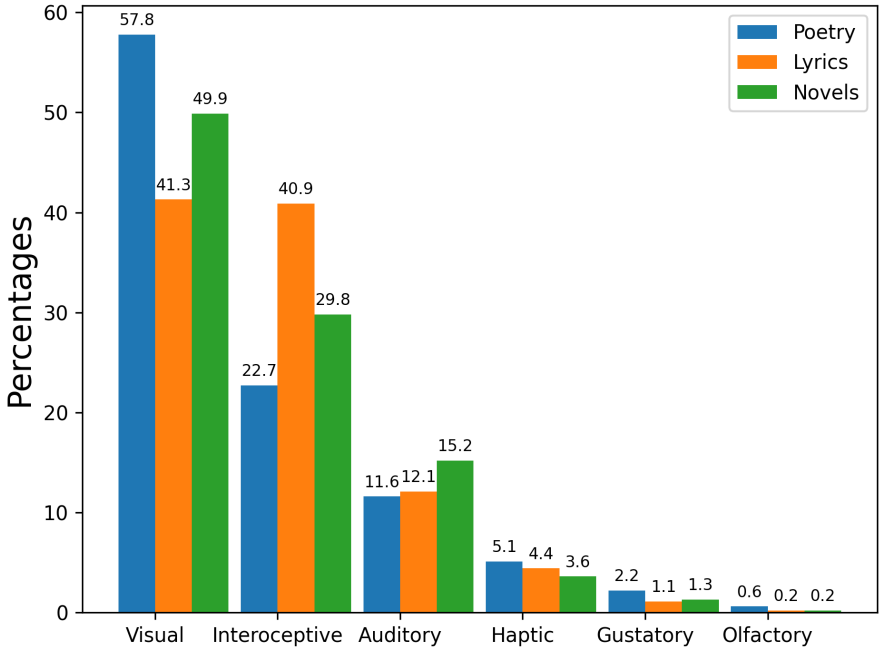


Figure 3.2: Distribution of observed modalities.

Considering all individuals with more than 10 sense-focused sentences[7], more than 90% of the individual sensorial style vectors in the novels and lyrics datasets are non-random. However, in the poetry dataset only around 23% of the individual vectors in the poetry dataset were non-random.

A possible reason for the difference in poetry is likely data sparsity — fewer sensorial expressions/ author (see Table 3.2). Exploring this intuition further we find that the non-random vectors in poetry have on average 159 sense-focused sentences while the remaining vectors that looked random had on average 24 sense-focused sentences. Similarly, 8 out of the 10 most prolific individuals had non-random vectors. However, none of the 10 least prolific individuals had non-random vectors. These support our intuition regarding data sparsity being the cause of the difference in poetry.

RQ2: How much data do we need to get a stable representation

[7]Because of the volume of lyrics, we limit the analysis to individuals with > 500 sense-focused sentences.

of sensorial style? We evaluate the average similarity for each individual with progressively larger samples sizes, k, of their sense-focused sentences. We chose a range of values of k from $k = 1$ to $k = 10$ with a granularity of 1, from $k = 10$ to $k = 100$, and $k = 100$ to $k = 1,000$ with granularities of 10 and 100. In Figure 3.3 we summarize these results with the median of average similarity across all individuals in each genre. We say the sensorial style vector has converged at a k value when the graph becomes more or less horizontal from that point onward. From the figure we see that as k increases similarity

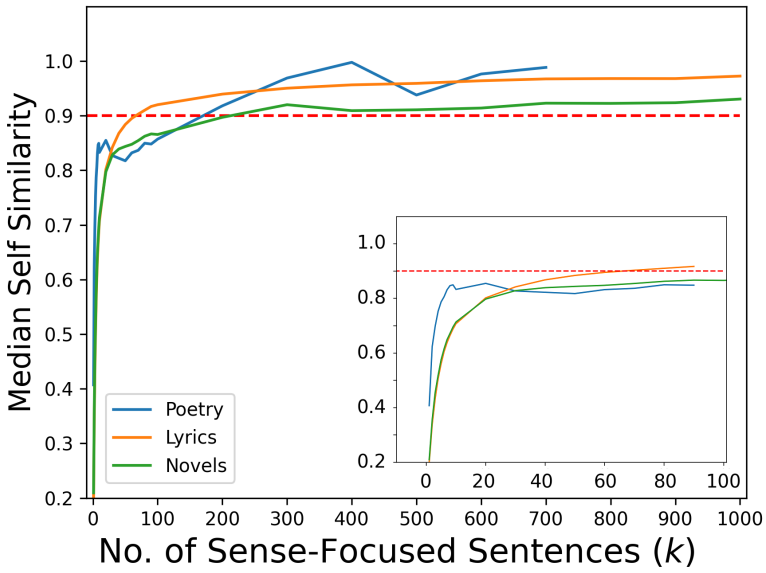


Figure 3.3: Convergence of style vectors as a function of k, the number of sense-focused sentences sampled.

increases (within the m samples of size k). We note that lyrics reaches a median average similarity of 0.90 with a sample of less than 100 sense focused sentences. In contrast, we need between 200 and 300 sentences to get the same 0.9 median average similarity for novels. Compared to novels and lyrics the plot for poetry has some fluctuations at $k \geq 400$, perhaps because there are only 7 poets with more than 100 sentences.

RQ 3: Does sensorial style vary over time?

We now explore whether sensorial style has changed over time in lyrics.

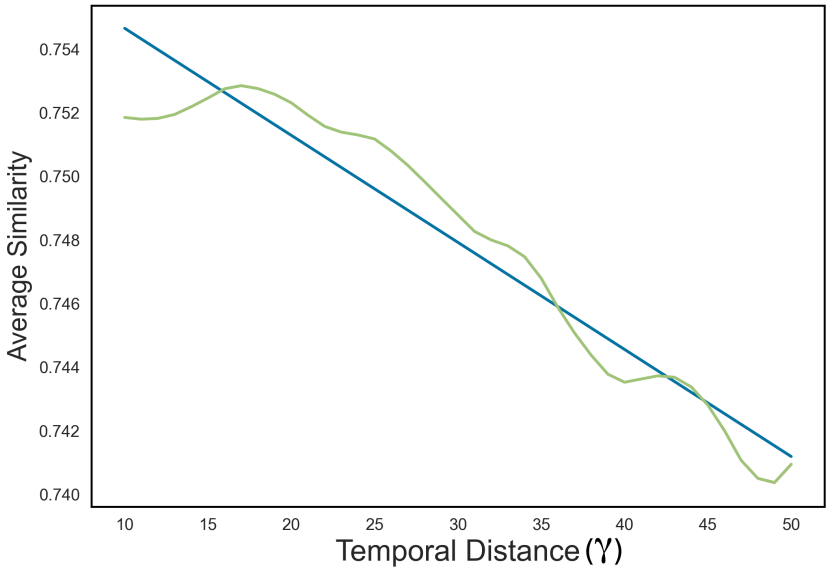


Figure 3.4: Average similarity between music lyrics as a function of their temporal distance. The blue line represents the linear approximation of the relationship. As temporal distance increases similarity decreases slightly.

As observed earlier, the poetry dataset is sparse so we do not include it. And with novels, it is sometimes challenging to pinpoint a single publication year, since some texts are written over multiple years[8].

Using the method described in Section 3.4, we measure average similarity in style of lyrics between pairs of individual sensorial style vectors that are a temporal distance γ away from each other. We use window length $\delta = 1.5$ years. Figure 3.4 illustrates our findings.

We observe that as temporal distance between works increases the similarity of sensorial style decreases. The augmented Dickey-Fuller test had a $p-$value $= 0.96$ meaning we cannot reject H_0, that the distribution is not-stationary [15]. However, the decrease is very slight; the 50 year drop in similarity is only 0.01.

Approximating the relationship between the average similarity and temporal distance as a linear process, we note that the average similarity decreases

[8]As an example, Louisa May Alcott's *Little Women* was published in two volumes spanning two years, 1868 and 1869.

very slowly at a rate of 3.37×10^{-4} per year.

RQ 4: Which features are representative and distinctive of the individuals within each genre? Table 4, shows the top-6 features that are representative of the members of each genre. We observe that these are all synaesthetic. Additionally the use of olfactory language in non-olfactory contexts is a representative feature in a majority of table cells (10/18). The standard deviations of top representative features is relatively low (between 0.00 to 0.02). This would indicate that the level of consistency for these top features is generally consistent for the three genres. One takeaway from these observations can be that in synaesthetic contexts, individuals are more prone to using lower senses (like olfaction) in a more consistent manner.

Table 5, shows the top-6 features that had the highest standard deviation between the members of each genre and were the most distinctive. The distinctive features were predominantly non-syneasthetic. In the cases where the distinctive feature was synaesthetic, the observed modality was either interoceptive or visual. This would indicate that there is greater diversity in expressions that rely on higher senses as a semantic scaffold.

For each genre, we rank all the sensorial style features by the standard deviation and compare them using Pearson's correlation. We observe that, the features are highly correlated. Lyrics had a high correlation with both novels (0.75) and poetry (0.81), while poetry and novels had a slightly lower correlation of 0.48.

Can sensorial style be used for prediction tasks?

We investigate whether sensorial style features can be used to identify genre. We compare against other standard style representations: LIWC [80] and content-free words (CFW) vectors [33]. We use standard 5 fold cross validation for each experiment to train and test a random forest classifier. We consider the most prolific 50 authors/genre.

Table 3.6 shows the results. We observe that sensorial style predicts genre with a high level of accuracy ($> 90\%$).

While the other representations achieve close to perfect accuracies, key to note is that our goal is less about beating baselines and more about understanding the kinds of signals conveyed by sensorial style.

Method	Baseline	Sensorial Style	LIWC	CFW
Features	—	42	73	307
Accuracy	0.33	0.91	0.99	0.99

Table 3.6: Prediction accuracy of the different features.

3.7 Case Study: Sensorial style in Lyrics

As a small illustration, we explore how sensorial style varies across different songs composed by the same artist. We consider all 962 artists who had at least 5 songs in the Hot 100 and extract a sensorial vector for each song. We then measure the average pair-wise cosine similarity (self similarity) amongst the songs of each artist. Almost 80% of the artists had an average self similarity ≥ 0.70. Only two artists had a self similarity < 0.50.

The rapper *NF* was the most consistent artist with an average self similarity of 0.93. Conversely, the least consistent artist was the rock musician *Tommy James* with an average similarity of 0.42. For example, in the song "When I Grow Up", *NF* used auditory language non-synaesthetically in 85.7% of the cases and for the song "NO NAME" this happened in 76.9% of the cases. The similarity between these two songs was over 95%.

In contrast, "Nothing to Hide" and "Ball and Chain" by *Tommy James'* had a similarity of 0.32. In "Ball and Chain", the artist uses visual language in all the instances where it was expected. However in "Nothing to Hide", he uses interoceptive language synaesthetically instead of visual in about 57% of the cases. This case study demonstrates a method for exploring sensorial style and their variations across writings at the individual level.

3.8 Conclusion

We have shown that individuals have sensorial language style and that this sensorial style is a non random phenomenon for novelists and musicians and therefore is likely developed intentionally. Interestingly, we also found that it takes just a few hundred sentences to extract stable sensorial style representations. We also show that sensorial style in lyrics largely stable over time;

the average similarity decreases at a rate of 3.37×10^{-4} per year.

Additionally, we show that sensorial style vectors seem to perform well at genre identification. The performance was high (> 0.90), however, it was not close to perfect as with other style representations. The question about how sensorial style representations can be improved to increase performance requires further investigation.

Our study is limited in that our method relies heavily on the underlying Lynott et al. [48] lexicon, and as with similar studies, is only as good as the lexicon. Additionally, we assume that each term is associated with a single sensorial modality. However, as research in psychology and neurology has shown, sensorial language is cross-modal. We leave this analysis to future work. In summary, we take a first step towards showing that sensorial style has a legitimate role in stylometrics research.

Chapter 4

EXTRAPOLATING SENSORIAL LEXICONS

4.1 Introduction

This chapter addresses a critical gap in the study of sensorial language, the limited availability of sensorial lexicons across languages. Representations of sensorial style like sensorial synaesthesia (Chapter 3) rely on a lexicon of sensorial words annotated for various sensory modalities. Unfortunately, these sensory lexicons are currently available for only a small set of languages. These languages include Russian [54], Italian [82], Dutch [76], and English [48]. Among these, only the Dutch and English lexicons cover all six senses, including interoception. This limited availability of lexicons significantly restricts the scope of sensorial style analysis beyond these few languages. To enable broader analysis across multiple languages and contexts (our goal in Chapter 6), we need a method to develop sensorial lexicons for other languages.

While developing sensorial lexicons from scratch for each language would be ideal, it is a resource-intensive undertaking requiring extensive annotation efforts by native speakers. A more practical approach is to systematically extrapolate existing lexicons, such as the English Sensorimotor norms [48], to other languages. Although this approach has its limitations, it provides a viable path forward for sensorial language analysis until comprehensive locally-developed lexicons become available. The goal of this chapter is to propose and evaluate methods for extrapolating sensorial lexicons from a source language (English) to multiple target languages while maintaining the reliability and validity of the sensory annotations.

Extrapolating lexicons across languages presents its own set of challenges, particularly in preserving semantic nuances and accounting for cultural variations in sensory expression. A simple translation-based approach fails to

account for polysemy[1]. For example, the English word 'car' can refer to both a four-wheeled automobile and a train carriage. In contrast, the Hindi word /gɑːɾiː/, which directly translates to 'car', encompasses a broader range of meanings, including motorbikes and various other forms of transportation like a bullock-cart, not implied by the English term. A simple translation-based method assumes a one-to-one correspondence between words in different languages, which does not capture the capacity of a word or phrase to have multiple meanings.

In this chapter, we propose a novel synset-based approach that uses BabelNet to extrapolate monolingual sensorial lexicons to multiple languages. These include 5 language groups: Germanic (English, German), Romance (French, Spanish), Balto-Slavic (Russian, Polish), Indo-Iranian (Hindi, Urdu) and Others (Arabic, Tamil). We first introduce BabelNet's multilingual lexical-semantic network (Section 4.2). We then present our method for expanding the sensorial lexicon to multiple target languages (Section 4.3) and evaluate the effectiveness of the proposed method through three key metrics: coverage analysis, dominant modality prediction, and user agreement studies (Section 4.5). In Chapter 6, we use this approach to analyze representations of sensorial style in multiple languages.

4.2 BabelNet: A Multilingual Lexical-Semantic Network

BabelNet [60] is a multilingual extension of WordNet [55] that represents meaning and ontological relations between words, for over 100 languages, using *synsets*. A *synset* (synonym set) represents a distinct concept through a collection of words or expressions that share the same meaning in a given context. Each synset contains equivalent words from multiple languages that express the same concept. For example, one synset might contain the English words 'car', 'automobile', and 'motor vehicle' along with their semantic equivalents in other languages like Spanish 'coche', 'automovil', Hindi '/gɑːɾiː/', and so on. Importantly, words can belong to multiple synsets to capture their different meanings - for instance, the Hindi word '/gɑːɾiː/' appears in both the

[1]Polysemy refers to a single word having multiple meanings.

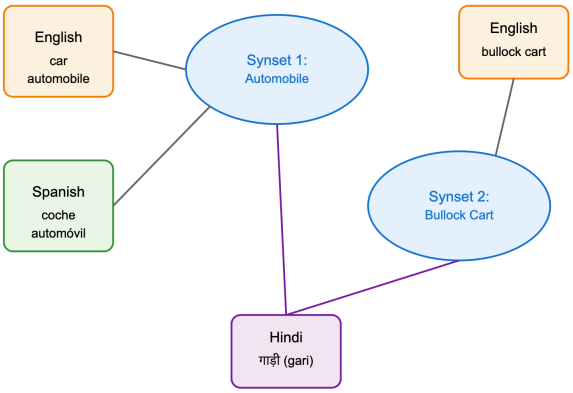


Figure 4.1: An illustration of how BabelNet's synset structure connects equivalent concepts across languages. The central synset 'Automobile' groups together semantically equivalent words from different languages that share the concept, such as English 'car'/'automobile', Spanish 'coche'/'automovil', and Hindi 'gari'. While the synset 'Bullock Cart' groups together English 'bullock cart' with the Hindi 'gari'.

'automobile' synset mentioned above and in a separate synset with 'bullock-cart', as illustrated in Figure 4.1. This structure allows BabelNet to capture how the same word can have different meanings across contexts. The representation of polysemy in BabelNet (and WordNet) facilitates more accurate comparisons across languages than methods relying on direct word-to-word translations. BabelNet's synset structure also allows for the preservation of multiple meanings across languages. For example, while a word embedding might conflate the automobile and bullock-cart meanings of '/gɑːɾiː/' into a single vector, BabelNet maintains these concepts as separate concepts, each with its own cross-lingual connections.

Synsets in BabelNet (and WordNet) serve as the fundamental units for describing semantic relationships between concepts. The semantic relationships between synsets are primarily represented through hierarchical structures, particularly as relationships of hyponyms[2] or hypernyms[3] as well as

[2]A hyponym is a word that denotes a more specific concept within a broader category, such as 'carrot' is a hyponym of 'vegetable'.

[3]A hypernym is a word that denotes a broader category that encompasses more specific

other semantic relations such as antonyms [4], meronyms [5], and domain categories[6].

4.3 Generating Multilingual Sensorial Lexicons

We start with English as our reference language and use BabelNet's synset structure to extrapolate sensorial ratings of target languages. In our reference lexicon (The Sensorimotor Lexicon [48]), each word has ratings along a 0-5 scale for six sensory modalities (Auditory, Gustatory, Haptic, Olfactory, Visual, Interoceptive). For example, 'sweet' might be rated as [0.35, 4.10, 0.35, 0.75, 1.65, 2.84] across these dimensions.

We first aggregate sensory information at the synset level. For each synset, we average the sensorial vectors of all its English member words from our reference lexicon. For example, if 'sweet' and 'sugar' belong to the same synset with gustatory ratings of 4.10 and 4.67, the synset maintains this strong gustatory association with an average rating of 4.37.

Next, we handle target language words by calculating their sensorial characteristics based on their synset memberships. Since words can belong to multiple synsets (capturing different meanings), we average the vectors of all synsets they belong to. For instance, if the Spanish word 'dulce' belongs to both a synset about taste (containing words 'sweet', 'sugary') and a synset about texture (containing words 'soft', 'gentle') with sensorial vectors with ratings [0.39, 4.37, 1.36, 0.89, 1.36, 3.03] and [0.42, 4.75, 4.42, 0.95, 1.42, 2.95] respectively. Its final sensorial vector [0.41, 4.56, 2.89, 0.92, 1.39, 2.99] would average the two synset vectors. This approach allows us to capture both sensory associations. This process is illustrated in Figure 4.2.

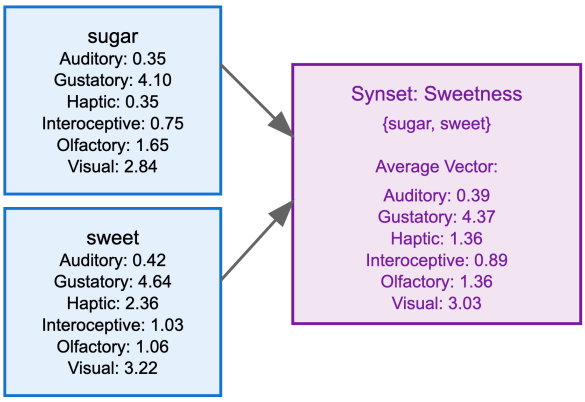
We formally define the computation of sensorial vectors for a target language as follows:
Given m words in the reference language (English) and k synsets, let $\mathbf{M} \in \mathbb{R}^{m \times 6}$ be the reference lexicon matrix where row i contains the sensory ratings

concepts, such as 'vegetable' is a hypernym of 'carrot'.

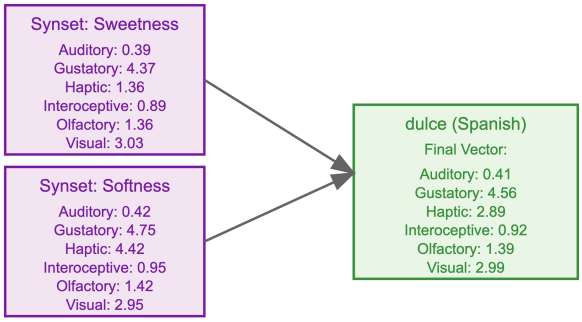
[4] opposite meanings, e.g., 'hot'-'cold'

[5] part-whole relationships, e.g., 'wheel'-'car'

[6] topical classifications like 'medicine' or 'sports'



(a) Step 1: Calculating the Sensorial Vector of a Synset from English words



(b) Step 2: Calculating the Sensorial Vector of a Spanish Word

Figure 4.2: Two-step process for calculating sensorial vectors of words in a target langauge.

for word w_i, $\mathcal{S} = \{s_1, ..., s_k\}$ be the set of all BabelNet synsets containing words from the reference lexicon and $\mathcal{W}_T = \{w_1, ..., w_n\}$ be the vocabulary of the target language.

We compute the target language sensory ratings matrix $\mathbf{N} \in \mathbb{R}^{n \times 6}$ in two steps:

1. Calculate synset-level ratings:

$$\text{synset\_ratings}(s_j) = \frac{1}{|\{w_i : w_i \in s_j\}|} \sum_{w_i \in s_j} \mathbf{M}_i$$

2. Calculate target word ratings:

$$\mathbf{N}_i = \frac{1}{|\{s_j : w_i \in s_j\}|} \sum_{s_j : w_i \in s_j} \text{synset\_ratings}(s_j)$$

More concisely, we define the computation of sensorial vectors for a target language as: $\mathbf{N} \in \mathbb{R}^{n \times 6}$ as

$$\mathbf{N} = \mathbf{C} \cdot \mathbf{G} \cdot \mathbf{D} \cdot \mathbf{S} \cdot \mathbf{M} \tag{4.1}$$

where:

- $\mathbf{S} \in \{0,1\}^{k \times m}$ maps reference words to synsets: $\mathbf{S}_{ij} = 1$ if word j belongs to synset i.
- $\mathbf{D} \in \mathbb{R}^{k \times k}$ is a diagonal matrix normalizing synset ratings: $\mathbf{D}_{ii} = 1/\sum_j \mathbf{S}_{ij}$.
- $\mathbf{G} \in \{0,1\}^{n \times k}$ maps target words to synsets: $\mathbf{G}_{ij} = 1$ if word i belongs to synset j.
- $\mathbf{C} \in \mathbb{R}^{n \times n}$ is a diagonal matrix normalizing target word ratings: $\mathbf{C}_{ii} = 1/\sum_j \mathbf{G}_{ij}$.

In our case, we have assumed a sensorial vector of 6 sensory modalities. However the proposed method can be extended to an arbitrary number of modalities without a loss in generality.

4.4 Target Language Sensorial Vocabulary

Our goal is to generate new sensorial lexicons, annotated for the 6 senses, in target languages that do not have them. We select 10 target languages across 5 families: Germanic (English, German[7]), Romance (French, Spanish), Balto-Slavic (Russian, Polish), Indo-Iranian (Hindi, Urdu) and Others (Arabic, Tamil). As a sanity check we also include English as a target language.. This selection encompasses both high-resource languages like Spanish

[7]We selected German, instead of Dutch, as a representative Germanic language because of ease of access to L1 German speakers compared to L1 Dutch speakers for the user agreement study in Sec 4.5. However, the method can be extended to include Dutch as well.

and English, as well as low-resource languages like Hindi, allowing us to test our extrapolation method across diverse linguistic contexts.

For each target language, we generate a vocabulary using WordFreq [78], a multilingual word frequency database. From WordFreq, we extract the 50,000 most frequent words per language. This frequency-based selection ensures we at least cover commonly used words. We label a word in the vocabulary as sensorial if it shares at least one synset with any of the words in the reference sensorial lexicon.

Once the sensorial vocabulary is selected, we quantify each word's sensorial properties across the six modalities (Auditory, Gustatory, Haptic, Olfactory, Visual, Interoceptive) using sensorial vectors derived from the method from Equation 4.1.

4.5 Results

To evaluate the effectiveness of our multilingual lexicon generation method, we use a Wikipedia parallel article dataset, where the same topic exists across multiple languages. It provides a controlled environment for cross-linguistic analysis. By selecting 3,000 random topics from English Wikipedia and their corresponding articles in all 10 target languages (yielding 30,000 total articles), we ensure that variations in our measurements stem from genuine linguistic differences rather than topical differences. The Wikipedia dataset covers a range of articles from scientific explanations to historical narratives, geographical descriptions, and culinary articles. This allows us to test our lexicons across varied contexts where sensorial language may be used differently. This diversity helps validate that our generated lexicons are robust and applicable across different domains, not just specialized contexts.

We evaluate the proposed lexicon generation method across 3 metrics (i) Coverage, (ii) Dominant Modality Prediction, (iii) User Agreement.

i. Coverage:

Almost all languages had 50,000 or more words in the WordFreq Corpus — with the exceptions of Urdu (48,289 words) and Hindi (27,330 words). The sensorial lexicons are subsets of these vocabulary lists.

	English	Spanish	French	German	Russian	Polish	Arabic	Hindi	Urdu	Tamil	avg (std)
Sensorial Vocabulary	9974	5288	5272	4807	3062	3256	3409	2496	1600	1201	4037 (2393)
Vocabulary Coverage	0.20	0.11	0.11	0.10	0.06	0.07	0.07	0.09	0.03	0.02	0.085 (0.040)
Corpus Coverage	0.10	0.08	0.07	0.06	0.05	0.07	0.08	0.07	0.07	**0.03**	0.069 (0.018)

Table 4.1: The vocabulary and corpus coverage of the lexicons. We use Wikipedia for the Corpus Coverage analysis. The coverages with a p-value <0.05 are shown in bold.

Table 4.1 presents the coverage of the lexicons for the 10 languages. Coverage measures the extent to which the lexicons capture sensorial language use within a text. We measure two types of coverage (a) Vocabulary Coverage (b) Corpus Coverage.

Vocabulary Coverage:

Vocabulary coverage is the proportion of a language's total vocabulary that is made up of sensorial words. It measures how many of the words in the target language vocabulary (As defined in 4.4) are sensorial. As an example, a vocabulary coverage of 6% for Russian would mean that around 3,000 of the 50,000 Russian words share at least one synset with a word in the reference sensorial lexicon.

We find that English had the highest vocabulary coverage at 20%, while Tamil and Urdu exhibited much lower coverage, around 2-3%. However, vocabulary coverage alone does not provide a complete picture of the use of sensorial language. It can have a skew in distribution and can over-represent rare words, that might not be frequently used in actual language contexts.To remedy this, we also measure corpus coverage that measures the proportion of the total text that is sensorial, using corpus coverage.

Corpus Coverage: While vocabulary coverage measures the proportion of unique sensory words in a language's vocabulary, corpus coverage measures how frequently sensory words are actually used in texts. We calculate corpus coverage as the proportion of words in a text that are sensorial. Using Wikipedia articles as our reference corpus, we measure corpus coverage by dividing the total count of sensorial words by the total word count in each language's articles.

On average, the sensorial lexicon coverage was 6.9%, with a small standard deviation of 1.8%. We find that the generated sensorial lexicons achieved relatively consistent corpus coverage across all languages. This suggests that, despite the differences in vocabulary size and sensorial word count, the actual use of sensorial language is relatively uniform across languages. Only the Tamil lexicon had a statistically lower coverage (p-value < 0.05). For the purposes of this work, we omit Tamil for subsequent evaluations.

The consistent corpus coverage across languages would perhaps suggest that our lexicon extension method is robust in identifying sensorial words, ensuring comprehensive and representative coverage for almost all the languages.

ii. Dominant Modality Prediction

In this task, the goal is to predict the dominant sensory modality of a given sensorial word. The dominant modality is defined as the modality with the highest sensorial rating. As an example, if we assume the Spanish word *dulce* has the sensorial vector [0.41, 4.56, 2.89, 0.92, 1.39, 2.99][8], then the dominant modality will be Gustatory with a rating of 4.56.

We use the human-annotated sensorial lexicons from Russian [54], Italian [82], Dutch [76] as our gold standards. We use our proposed method to infer the sensorial vectors for each word in these lexicons and compare the dominant modality of the inferred sensorial vectors to those of the original gold standard. We compare the proposed method with the baseline translation based method. This baseline method translates each sensory word from the target language into the reference language (English) and maps the corresponding sensory ratings from the reference language's lexicon to the target word. As an example, if the Spanish word 'dulce' is translated to the English word 'sweet', then the sensory ratings for 'sweet', [0.42, 4.64, 2.36, 1.03, 3.22], from the English lexicon would be assigned to 'dulce' as [0.42, 4.64, 2.36, 1.03, 3.22]. The dominantly modality of 'sweet', Gustatory is directly mapped to 'dulce'.

The scope of this analysis is constrained by the availability of sensorial lexicons. Although Dutch and Italian are not among our primary target lan-

[8]The sensorial vectors are ordered Auditory, Gustatory, Haptic, Olfactory, Visual, Interoceptive respectively.

Language	Lexicon Size	Accuracy	
		Proposed Method	Baseline Method
Dutch (NL)	24036	0.62	0.48
Italian (IT)	1121	0.88	0.73
Russian (RU)	506	0.57	0.53

Table 4.2: The accuracy of the two methods for the Dominant Modality Prediction task.

guages for cross-cultural analysis, we include them in this evaluation because they have existing sensorial lexicons that allow us to validate our methods. However, it's important to note that only the Dutch lexicon includes words labeled along all six sensory dimensions. In contrast, the Italian and Russian lexicons provide word ratings for only the five classical sensory modalities. As a result, in the case of Italian and Russian, if either method predicts Interoception as the dominant modality, it is considered an incorrect prediction. In the case of all three languages (Table 4.2), we find that our proposed synset-based method outperforms the baseline translation based method.

Error Analysis

An analysis of prediction errors reveals a consistent pattern across the three languages: most errors stem from the model's tendency to over-predict visual modality, particularly for terms that should be classified under other sensory categories.

For Dutch (Figure 4.3), we find that approximately between 35% to 62% of non-visual sensory terms were incorrectly classified as visual. This visual bias was most pronounced for haptic terms. Italian (Figure B.1) showed similar patterns. Even with 85% accuracy for visual terms, the errors were not evenly distributed across modalities. When the model made mistakes on gustatory or olfactory terms, it predominantly predicted them as visual. Russian (Figure B.2) exhibited the strongest visual bias in prediction errors.

iii. User Agreement:

We present a user agreement study to compare the agreement between the proposed method and L1-speaker judgements. For each language, we

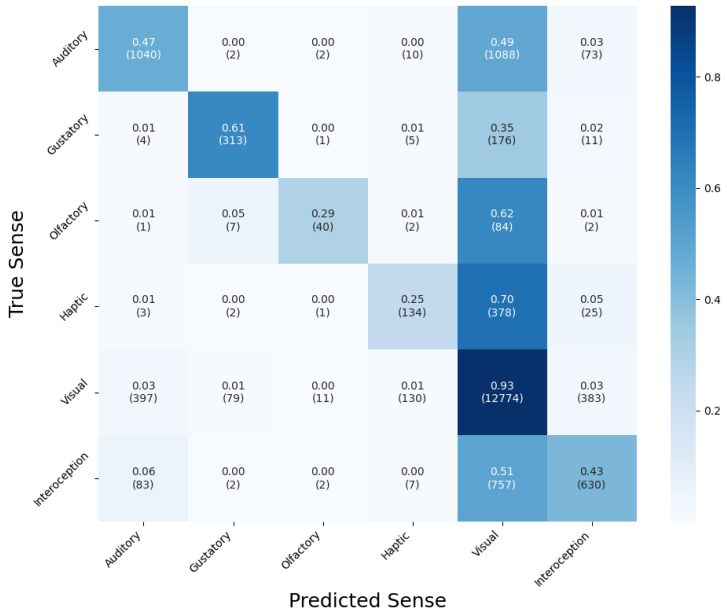


Figure 4.3: Dominant sense prediction accuracy for Dutch. Values show the probability of each prediction (0-1) with raw counts of occurrences shown in parentheses.

Language	English	Spanish	French	German	Arabic	Hindi	Urdu
Cohen's κ	0.97	0.78	0.80	0.89	0.79	0.94	0.72

Table 4.3: The Cohen's κ between the majority agreement and the method's prediction.

select 100 words that have been labeled for their dominant modality using our method. We use a stratified sampling approach and stratify the words based on their predicted dominant sensorial modality, such that we select a random set of 10 words for each modality. We also use a proportional sampling approach and randomly select 40 words from the set of words that have not been selected previously. This ensures that we have both a representation from each modality as well as a diverse and representative selection.

We presented these 100 words to at least three native speakers (L1) of each language. Each word was presented in a forced-choice task, where participants chose between two modalities: one assigned by our method and the other

randomly selected from the remaining set of modalities (Appendix B provides an example of the prompt). The choices were randomly shuffled to minimize order effects.

In Table 4.3, we report the Cohen's κ between our proposed method and the majority label of the speakers[9]. We observe that the Cohen's κ was greater than 0.7 for all languages suggesting high alignment of the method with human judgements.

The results of the three tasks (Coverage, Dominant Modality Prediction, User Agreement) demonstrate that the proposed lexicon extension method is effective in accurately predicting the dominant modalities of sensorial words across multiple target languages.

4.6 Conclusion

We have presented a systematic approach to extend monolingual sensorial lexicons to multiple languages using BabelNet's multilingual lexical-semantic networks. The proposed method uses BabelNet's synset structure to preserve multiple meanings of words (polysemy) when mapping sensorial information across languages.

We show the effectiveness of the proposed approach using three evaluation metrics. First, our coverage analysis showed that the generated lexicons achieved consistent corpus coverage (averaging 6.9%) across most languages, with only Tamil showing statistically lower coverage. Notably, while vocabulary sizes varied considerably across languages, the actual usage of sensorial terms remained relatively uniform, suggesting the robustness of our approach in capturing sensorial language across different linguistic contexts.

Second, in the dominant modality prediction task, our synset-based method outperformed the baseline translation approach across all tested languages. The method achieved accuracies of 0.88 for Italian, 0.62 for Dutch, and 0.57 for Russian, successfully preserving sensory categorizations even when extending to languages with different linguistic structures.

[9]Our experiments are still ongoing for the Slavic language pair (Russian and Polish)

Third, our user agreement study revealed high agreement between native speakers and our method's predictions, with Cohen's κ values exceeding 0.7 for all languages, demonstrating the validity of our generated lexicons and their alignment with native speakers' intuitions about sensory terms.

These results validate our approach for extending sensorial lexicons across languages while maintaining reliability and validity. While the method has shown promise, it is important to note that its effectiveness varies across languages, with some showing stronger results than others. This variability suggests opportunities for future refinement of the approach, particularly for languages that showed lower performance metrics.

The successful development of these multilingual sensorial lexicons enables broader cross-linguistic analysis of sensorial style. In Chapter 6, we proceed with studying texts from the different languages using the lexicons generated for this chapter.

Chapter 5

EXPANDING SENSORIAL STYLE DEFINITIONS

5.1 Introduction

In this chapter we expand our understanding of sensorial language use. In Chapter 3, we introduced one representation of sensorial style through the lens of synaesthesia, here we propose two additional representations — *Sensorial Prevalence* and *Sensorial Diversity*. Sensorial Prevalence captures how frequently different sensory modalities are used in language, while Sensorial Diversity measures the semantic richness and variety within each sensory modality. Together with synaesthesia, these dimensions provide complementary perspectives on how sensory experiences are expressed linguistically.

In Chapter 6, we use these expanded dimensions to investigate cross-cultural patterns in sensory language use, examining both universal tendencies and culture-specific variations. This broader framework for studying sensorial style allows us to move beyond individual languages and genres to understand sensory language use at a cross-cultural level, contributing to both stylometrics and sensorial linguistics.

5.2 Defining Sensorial Prevalence

Consider that in general our sensory experiences and the language we use to describe them are often dominated by sight, with visual words comprising approximately 70-80% of English sensory language [89]. In contrast, smell-related words account for less than 1%, reflecting the typically less prominent role of olfactory experiences in our daily lives.

However, this distribution of sensory modality prevalence is not necessarily consistent and can vary significantly depending on context. In some cases, the patterns are intuitive and predictable. For instance, recipe instructions are likely to employ more gustatory words compared to a computer science

textbook. A manual for fabric selection might emphasize haptic language to describe textures. Perfume reviews would naturally incorporate more olfactory vocabulary.

In other cases, the prevalence patterns can be surprisingly complex. Consider music - while it is fundamentally an auditory experience that might be expected to emphasize sound-related language, it is also deeply emotional. Music lyrics often focus more on interoceptive descriptions of feelings and bodily sensations than on auditory terms. For instance, lyrics frequently describe experiences like 'heartache' or 'love' rather than sounds, making it unclear whether auditory or interoceptive terms would dominate in musical discourse. Such complex interactions between sensory modalities demonstrate why we cannot always predict which sensory dimensions will be most prevalent in different contexts.

These variations in how sensory language is distributed across different contexts form the basis of what we term '*Sensorial Prevalence*'. This dimension of sensorial style quantifies the emphasis placed on different sensory modalities, allowing us to systematically investigate how sensory experiences are prioritized across different contexts and cultures.

A high prevalence of visual terms would indicate a greater emphasis on visual information in a culture or text. This could be seen in art criticism like "The vibrant hues and bold brushstrokes create a striking contrast against the muted background" or technical manuals as in the case "Ensure the blinking red LED is visible when the device is in standby mode". Similarly, a relatively higher prevalence of taste terms could reflect a rich culinary tradition where gustatory experiences are given preference, as in the case of "The dish offers a perfect balance of sweet and savory, with hints of umami and a subtle tang".

We measure Sensorial Prevalence by calculating the proportion of words associated with each sensory modality within a document or corpus. Formally, we can define Sensorial Prevalence $\text{Prev}(d, s)$ of a sense s in a document d as

$$\text{Prev}(d, s) = \frac{\sum\limits_{\substack{w \in W_d, \\ M(w)=s}} f_{w,d}}{\sum\limits_{w \in W_d} f_{w,d}}$$

Let D be the set of all documents in a given corpus. We define W_d as the set of words in document $d \in D$. The function $M(w)$ maps each word to its dominant sensory modality s. The frequency of word $w \in W_d$ in the document is defined as $f_{w,d}$. We represent the average Sensorial Prevalence of a sensory modality across all documents in the corpus as:

$$\text{Prev}(s) = \frac{1}{|D|} \sum_{d \in D} \text{Prev}(d, s) \tag{5.1}$$

5.3 Defining Sensorial Diversity

Consider that when describing a visual scene like Ansel Adams' famous photograph of the Grand Teton Mountains, one might use a diverse range of terms to capture visual elements such as the sky, the winding river, and the mountain itself. Their description might include phrases like 'blue sky', 'serpentine river', 'towering mountain', and 'expansive landscape'. In contrast, another person might employ more semantically cohesive language, focusing solely on the Grand Teton mountain and its specific features. They might describe the mountain's jagged peaks, snow-capped summit, rocky face, and steep slopes. Their description would include terms like 'craggy', 'imposing', 'majestic', and 'rugged' — all pertaining specifically to the mountain's characteristics. Notably, both descriptions engage with the same sensory modality — vision – but exhibit different ranges of diversity in their language use, with one spreading across multiple elements of the scene and the other focusing on a single element.

The semantic range within a sensory modality can also vary across languages and contexts. These variations in the semantic range within language and culture form the basis of what we term 'Sensorial Diversity'. We define this representation of sensorial style by measuring diversity in the semantics (meaning) of sensorial words in a text. If the semantic diversity of sensorial words used is higher, this would indicate a greater diversity in sensorial language for that modality.

We can define Sensorial Diversity, $\text{Div}(d, s)$ of a sense s in a document d. We adapt the 'Mean Shared Compounds' [2] metric for our definition of

Sensorial Diversity. Our definition of Sensorial Diversity considers the average pairwise semantic distance between words or concepts and normalizes this value across the document. Formally we define it as

$$\mathrm{Div}(d, s) = \sum_{\substack{w_1, w_2 \in W_d, \\ w_1 \neq w_2, \\ M(w_1)=M(w_2)=s}} \gamma(d, s) \cdot f_{w_1,d} f_{w_2,d} \cdot \mathrm{dist}(w_1, w_2)$$

where $\gamma(d, s)$ normalizes the diversity score based on the number of sensory words of modality s in document d and $\mathrm{dist}(w_1, w_2)$ is the semantic distance between words w_1 and w_2, the frequency of words w_1 and w_2 in the document are represented by $f_{w_1,d}$ and $f_{w_2,d}$ respectively.

We extend this definition and represent the Average Sensorial Diversity $\mathrm{Div}(s)$ of a sensory modality across all documents in the corpus as:

$$\mathrm{Div}(s) = \frac{1}{|D|} \sum_{d \in D} \mathrm{Div}(d, s) \tag{5.2}$$

Semantic Distance

We measure semantic distance between words using a synset-based notion of semantic distance, that uses the ontological relationships from BabelNet (See: Chapter 4). We also use this definition to define a word-based notion of semantic distance that allows us to compare words directly.

BabelNet can be represented as directed graph where nodes represent *synsets* and edges represent semantic relationships between them. The depth of a node represents its level of specificity and its distance from the root node. For example, in the path of synsets described as *entity* → *living_thing* → *animal* → *mammal* → *dog* → *German_shepherd*, *German_shepherd* would be at depth 5.

Using this graph structure, we define the semantic distance between two synsets (b_1 and b_2) as $\mathrm{dist}(b_1, b_2)$.

To find the distance between two words, we consider all synsets that each word belongs to and choose the minimum distance:

$$\mathrm{dist}(w_1, w_2) = \min_{w_1 \in b_1, w_2 \in b_2} \mathrm{dist}(b_1, b_2) \tag{5.3}$$

This notion of distance aligns with psycholinguistic research suggesting that humans tend to process word meanings through their closest semantic connections [17]. When words have multiple possible meanings (polysemy), people typically access the most relevant or closely related senses first.

Normalization Factor

We define the normalization factor $\gamma(d, s)$ for document d and sensory modality s as the reciprocal of ${}^{N}C_2$ and define it as:

$$\gamma(d, s) = \frac{2}{N(N-1)}$$

where N is the total number of words in document d of modality s :

This normalization ensures that the diversity measure is comparable across documents of different lengths and with varying numbers of sensory words. Since $\gamma(d, s) = \frac{1}{{}^{N}C_2}$, the range of γ is:

$$0 < \gamma(d, s) \leq 1$$

A large $\gamma(d, s) \approx 1$ value indicates fewer sensory words of modality s in the document. Conversely, a small $\gamma(d, s) \approx 0$ value suggests that there are many sensory words of modality s in the document.

5.4 Measuring the Contribution of Individual Words to Sensorial Style

The two representations of Sensorial Style — Sensorial Prevalence and Sensorial Diversity (as well as Sensorial Synaesthesia) measure style at the level of modalities. We extend our measurements to word-level by measuring the contribution of individual words or concepts towards these representations. The measure estimates the degree to which the presence of a particular word affects the magnitude of the representations.

Measuring the Contribution of Individual Words to Sensorial Prevalence

We define the contribution of a word w to Sensorial Prevalence of a sense s as:

$$\text{Cont}_{\text{Prev}}(w, s) = \text{Observed}_{\text{Prev}}(w, s) - \text{Expected}_{\text{Prev}}(w, s) \tag{5.4}$$

This measures the difference between the observed usage and the expected usage of a word in a specific sensory modality which we define as:

$$\text{Observed}_{\text{Prev}}(w, s) = \frac{1}{|D|} \sum_{d \in D} \frac{f_{w,d}}{\sum_{\substack{v \in W_d, \\ M(v)=s}} f_{v,d}}$$

$$\text{Expected}_{\text{Prev}}(w, s) = \frac{\sum_{d \in D} f_{w,d}}{\sum_{d \in D} \sum_{\substack{v \in W_d, \\ M(v)=s}} f_{v,d}}$$

Where D is the set of all documents, W_d is the set of words in document d, $M(v)$ provides the sensory modality of word v, $f_{w,d}$ is the frequency of word w in document d.

A positive contribution indicates that the word is used more frequently in the sensory modality than expected, while a negative contribution suggests its used less frequently than expected. Comparing contributions across languages will help reveal whether similar words contribute consistently to the prevalence of sensory modalities across languages, or whether their contributions vary due to cultural and linguistic differences. For example, we can examine if words related to 'meat' contribute similarly to gustatory prevalence across languages, or if certain cultures emphasize different taste dimensions in their sensory vocabulary like 'fruits'.

Measuring the Contribution of Individual Words to to Sensorial Diversity

Similarly, we define the contribution of a word w to Sensorial Diversity of a sense s as:

$$\text{Cont}_{\text{Div}}(w, s) = \text{Observed}_{\text{Div}}(w, s) - \text{Expected}_{\text{Div}}(w, s) \tag{5.5}$$

$$\text{Observed}_{\text{Div}}(w, s) = \frac{1}{|D|} \sum_{d \in D} \gamma(d, s) f_{w,d} \cdot \sum_{\substack{v \in W_d, \\ v \neq w, \\ M(w)=M(v)=s}} f_{v,d} \text{dist}(w, v)$$

$$\text{Expected}_{\text{Div}}(w, s) = \frac{2 \cdot \sum_{d \in D} f_{w,d}}{\sum_{d \in D} \sum_{\substack{v \in W_d, \\ M(v)=s}} f_{v,d}} \cdot \frac{1}{|D|} \sum_{d \in D} \frac{\sum_{\substack{v \in W_d, \\ M(v)=s}} f_{v,d} \cdot \text{dist}(w, v)}{\sum_{\substack{v \in W_d, \\ M(v)=s}} f_{v,d}}$$

A positive contribution indicates that the word contributes more to Sensorial Diversity than expected, while a negative contribution suggests it contributes less than expected. These contributions can reveal how different languages organize their sensory vocabulary through varying semantic connections.

5.5 Conclusion

In this chapter, we expanded the framework for representing sensorial style by introducing two new dimensions: Sensorial Prevalence and Sensorial Diversity. Together with the previously established synaesthesia-based representation, these provide complementary perspectives on how sensory experiences are encoded in language.

Sensorial Prevalence quantifies the relative emphasis placed on different sensory modalities, allowing us to examine how languages and contexts prioritize different types of sensory information. The prevalence measure can reveal both universal patterns (like the dominance of visual language) as well as context-specific variations (like increased gustatory language in culinary texts). Sensorial Diversity measures the semantic richness within each sensory modality, capturing how languages vary in their capacity to express nuanced sensory distinctions.

The contribution metrics introduced for both prevalence and diversity allow us to examine which specific words and concepts drive patterns in sensorial

style. This word-level analysis provides insights into how cultural and linguistic factors shape sensory vocabulary.

These representations, while informative, do not capture all aspects of sensorial style. Other potentially valuable representations exist, such as word-level approaches that examine patterns in individual sensory word usage, or approaches that analyze how sensory language combines

Chapter 6

EXTENDING ANALYSIS TO MULTIPLE LANGUAGES

6.1 Introduction

In our investigation of sensorial style, we aim to study different contexts, not just different literary genres but also different languages. We postulate that sensorial style can differ between cultures, stemming from differences in experiences, and traditions. As an example, consider a culture like that of the United States, where coffee is considered an integral part of daily life. The language used to describe coffee has become highly developed and nuanced. Coffee can be described as 'bitter' when the focus is on taste, 'aromatic' when it is smell, 'hot' when the focus is touch, 'dark' when the focus is sight, and 'relaxing' when the focus is on its effects as an early morning drink.

In contrast, in a culture like that of North India, where coffee does not hold the same cultural significance, the sensory language used to describe coffee would likely not have the same level of cultural significance that it does in the American context. Instead, a beverage like tea, which occupies a similar cultural space in North India that coffee does in the US, would elicit a range of varying sensory experiences and descriptions. For instance, tea might be described as fragrant (smell), soothing (interoception), golden or amber (sight), warm (touch), and spicy or malty (taste), reflecting its cultural importance and the variety of ways it is prepared and consumed.

These cultural differences in sensory language not only reflect varying culinary traditions but may also provide insights into broader cultural priorities and attentional focuses. By examining how different cultures use sensorial language, we can develop a better understanding of how they perceive and interact with their environment. A comparative approach allows us to uncover shared similarities and dissimilarities between different linguistic cultures and traditions, viewed through the lens of sensory experience. To this end, we investigate whether certain patterns in sensorial style hold true across different

languages and cultures.

The multilingual sensorial lexicons developed in Chapter 4, combined with the additional representations of Sensorial Style — Sensorial Prevalence and Sensorial Diversity defined in Chapter 5, provide tools for investigating patterns in sensorial language use across different languages. Using these lexicons and representations, we extend our investigation beyond English to examine how different languages may encode sensory experiences.

In Chapter 3, we showed that the frequency of the different sensorial modalities in English is hierarchically ordered, with visual and auditory language being used more frequently than the language of touch, taste, and smell. We extend this using the notion of *Sensorial Prevalence* ask the following questions:

RQ 1 (a): Does the hierarchy of sensory modalities exist across languages?

RQ 1 (b): Does this hierarchical relationship persist across different types of texts?

Building on our investigation of sensory modality prevalence, we also examine how languages vary in their semantic richness within each sensory modality through the lens of *Sensorial Diversity*. This leads us to ask:

RQ 2 (a): Do the different sensorial modalities exhibit similar kinds of patterns of diversity across languages?

RQ 2 (b): Are patterns of sensorial diversity consistent across different types of texts?

These questions probe the universality and variability of sensorial style across languages and text types.

6.2 Sensorial Style of Cross Cultural Data

We analyze cross-cultural patterns in sensorial language using a parallel Wikipedia dataset, where the same subject matter or 'topic' exists across multiple languages. The articles were selected from Wikidata's most popular articles, specifically those ranked between *Q1* and *Q100000*. In Wikidata, these Q-numbers represent the most fundamental and widely referenced concepts, with lower Q-numbers generally indicating more essential or frequently accessed

articles. For example, *Q1* is 'universe', *Q5* is 'human', and *Q222695* is the 'Grand Teton National Park'. Higher Q-numbers represent progressively more specific or less central concepts.

A topic represents a specific concept, entity, or subject that is discussed consistently across different language versions of Wikipedia. For example, the topic 'Grand Teton National Park' would be represented by articles titled '*Grand Teton National Park*' in English and '*Parque nacional de Grand Teton*' in Spanish. While the language of these articles differs, the underlying subject matter or 'topic' remains constant.

From this pool of Wikidata entries, we randomly selected 3,000 topics[1] that had parallel articles across all 10 target languages, yielding a total of 30,000 articles (The distribution of topics can be found in Appendix C.1). Parallel articles allows us to control for content while isolating linguistic and cultural differences in sensory expression. We complement this controlled analysis with non-parallel corpora of music lyrics and recipes, which offer more naturalistic examples of sensory language in cultural contexts. These datasets reflect how different cultures naturally express sensory experiences in everyday language use. Analyzing both controlled parallel texts and naturalistic cultural content, gives us a lens to look at both universal patterns and cultural specificities in how languages encode sensory experiences.

6.3 Sensorial Prevalence in Wikipedia

We calculate the Sensorial Prevalence of the Wikipedia dataset using equation 5.1. Table 6.1, presents the average prevalence of each sensory modality for the languages studied. We summarize the findings.

Visual Dominance: Visual language overwhelmingly dominates across all languages. The prevalence of visual terms ranges from 0.64 (Polish) to 0.89 (Urdu), with most languages showing a very high visual prevalence of 0.85 or more. This underscores the primacy of visual perception in human language and cognition.

[1]The entire list of the 3000 topics can be accessed at https://github.com/osama-khalid/multilingual_style/blob/main/wikipedia_topics.txt.

Language	Vision	Audition	Interoception	Gustation	Haptic	Olfactory
English	0.87	0.07	0.04	0.005	0.01	0.002
Spanish	0.87	0.08	0.02	0.004	0.02	0.002
French	0.85	0.10	0.02	0.005	0.01	0.002
German	0.78	0.15	0.04	0.004	0.01	0.003
Russian	0.86	0.08	0.03	0.006	0.02	0.002
Polish	0.64	0.04	0.28	0.005	0.01	0.001
Arabic	0.88	0.05	0.04	0.005	0.02	0.002
Hindi	0.87	0.06	0.04	0.005	0.02	0.001
Urdu	0.89	0.07	0.04	0.006	0.01	0.001

Table 6.1: Average prevalence of the sensory modalities in each language.

Consistent Sensory Hierarchy: We observe a remarkably consistent hierarchy of sensory modalities across languages. The general order of prevalence follows: Vision > Audition > Interoception > Touch > Taste > Smell This pattern holds true for almost all languages in our study, with Polish being the only exception where interoception surpasses audition in prevalence.

Language-Specific Variations: Despite the overall consistency, there are some notable variations. Polish shows a distinctly lower visual prevalence (0.64) compared to other languages, coupled with an order of magnitude higher interoceptive prevalence (0.28). German exhibits a higher auditory prevalence (0.15) compared to other languages.

Modality-Specific Observations: Prevalence of Audition ranges from 0.04 (Polish) to 0.15 (German). Most languages show an interoceptive prevalence around 0.02-0.04, with Polish being a notable outlier at 0.28. Gustation and Olfaction consistently show the lowest prevalence across all languages, typically below 0.01.

While this analysis highlights the general trends in Sensorial Prevalence, it does not explain the specific ways in which the sensorial style may vary or align across languages. Examining the specific contributions of individual synsets can with a better understanding of how different languages encode sensory experiences.

Contribution of Synsets to Prevalence

We adapt the Contribution metric introduced in Equation 5.4 to measure

the contribution of synsets to Sensorial Prevalence. The contribution of a synset is defined as the average contribution of all words within that synset. Formally, we define the contribution of a synset b to the prevalence of a sense s as

$$\text{Cont}_{\text{Prev}}(b, s) = \frac{1}{|W_b|} \sum_{\substack{w \in W_b, \\ M(w)=s}} \text{Cont}_{\text{Prev}}(w, s) \tag{6.1}$$

Where W_b is the set of words in synset b, $M(w)$ is the dominant modality of word w $\text{Cont}_{\text{Prev}}(w, s)$ is the contribution of word w to the prevalence of sense s calculated using equation 5.4.

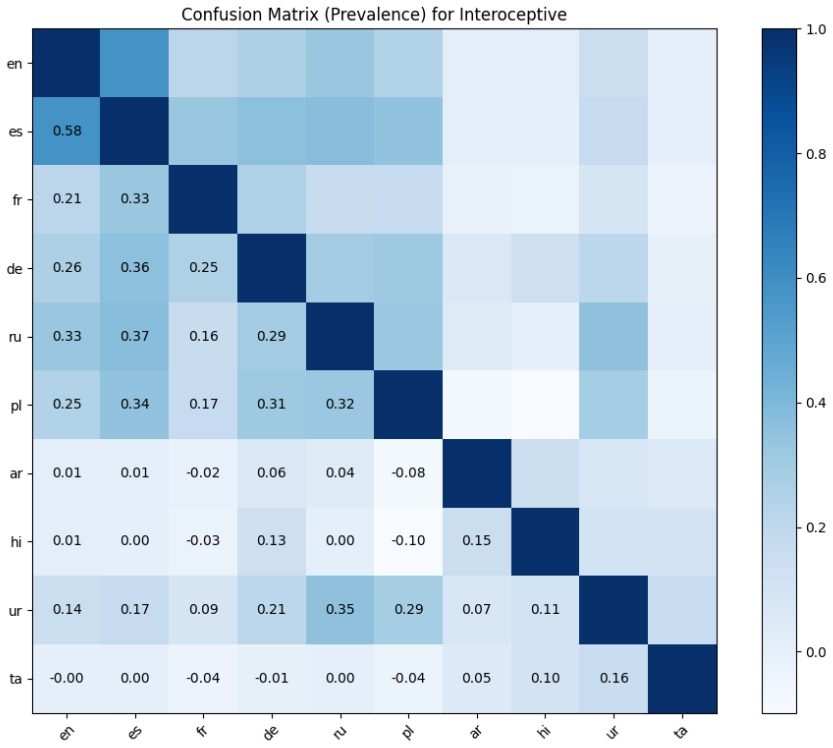


Figure 6.1: The cosine similarities of the vectors of contribution of synsets in Interoceptive Prevalence.

Finally, we create a contribution vector for each language. This vector

contains the contributions towards prevalence of all synsets to a particular sensory modality in that language. The dimension of this vector is equal to the total number of synsets in our lexicon. For a language L and sensory modality s, the contribution vector V_L,s is

$$V_L, s = [\text{Cont}_{\text{Prev}}(b_1, s), \text{Cont}_{\text{Prev}}(b_2, s), ..., \text{Cont}_{\text{Prev}}(b_n, s)] \tag{6.2}$$

Where $b_1, b_2, \cdots, b_n$ are all the synsets in our lexicon.

While examining patterns in sensorial language contribution across modalities, we found that language pairs generally showed low similarity (less than 0.5) across all senses. Here we present detailed results for gustation as an illustrative case since food practices and descriptions offer clear examples of how cultural differences manifest in language use. The contribution patterns in gustatory language are particularly revealing. For instance, from Fig 6.1 we find that the Spanish (es) and English (en) vectors had the highest interoceptive similarity (0.60). And from Fig 6.2, we can see that the Spanish vectors also had a high gustatory similarity with the Polish (pl) — 0.58 — and German (de) — 0.60 — vectors. Similarly, the Polish and German gustatory vectors also had a high similarity (0.52). In general, the language pairs had a low cosine similarity across all senses with no pair having a similarity greater than 0.5.

The high similarity between German, Spanish and Polish can be attributed to the shared culinary experiences between the language speakers. For instance, two of the synsets that contribute most significantly to the gustatory modality in both Polish and German are 'steak' and 'hamburger', reflecting the importance of meat in these cultures' cuisines. In contrast, the synsets with the highest contribution in Urdu are 'yogurt' and 'bananas', which would indicate a greater emphasis on vegetarian elements. This difference in culinary focus would explain the low similarity in gustatory language between Urdu and German (0.21) or Spanish (0.17).

This pattern becomes even more apparent when examining Table 6.2, which represents the top 10 synsets contributing to each language's gustatory prevalence. While European languages emphasize meat products and heavier foods in their top contributors (such as ham, bacon, and beef), Urdu

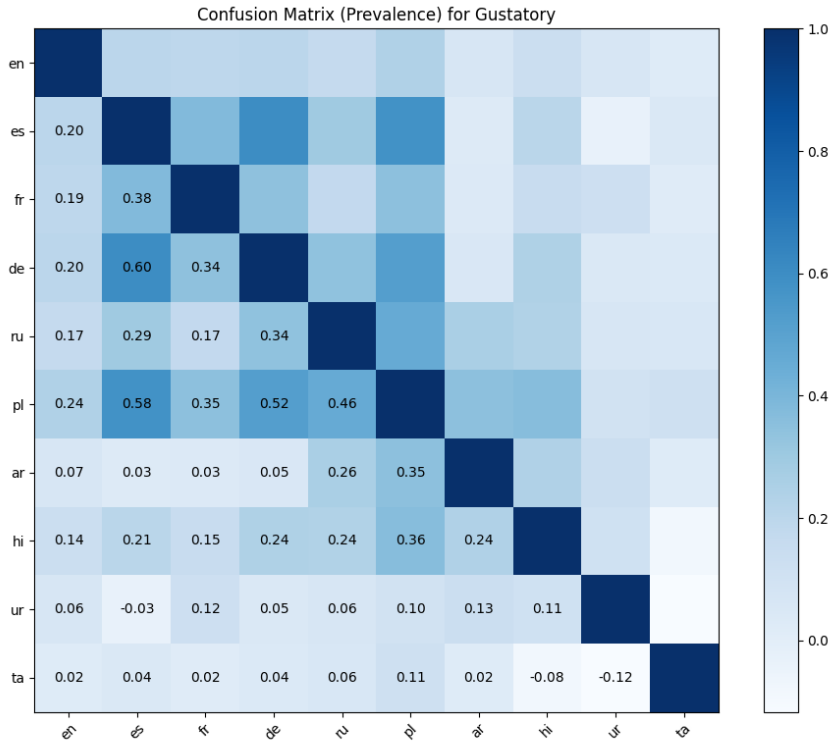


Figure 6.2: The cosine similarities of the vectors of contribution of synsets in Gustatory Prevalence.

and other South Asian languages show more emphasis on fruits, dairy products, and vegetarian items (like yogurt, bananas, and pomegranate). The European languages show relatively limited variation in their meat-based concepts, while South Asian languages and Arabic display greater diversity in their fruit-related terms, reflecting local cultural preferences and agricultural traditions.

RQ 1 (a): Does the hierarchy of sensory modalities exist across languages?

The sensory modality hierarchy is generally consistent across languages, with vision being the most dominant modality, followed by audition, interoception, haptic, gustation, and olfaction. This hierarchy aligns with the

widely held assumption that visual and auditory experiences are more prominently represented in language. Furthermore, while the overarching sensory hierarchy is maintained, across languages, it does not mean that the same types of words (synsets) contribute equally across languages. Different languages emphasize different sets of words for each modality based on cultural and contextual factors.

6.4 Sensorial Diversity in Wikipedia

We calculate the Sensorial Diversity of the Wikipedia dataset using equation 5.2. Table 6.3 presents the average diversity scores for each sensory modality across different languages in the Wikipedia corpus. We summarize the findings for the different sensory modalities.

Visual Diversity: Most languages show high diversity in visual language, with scores ranging from 0.75 to 0.90 for languages like Spanish, French, German, Russian, Polish, and Hindi. This suggests a rich and varied use of visual descriptors in these languages. However, English, Arabic, and Urdu show notably lower visual diversity (0.51, 0.48, and 0.48 respectively).

Auditory Diversity: There's considerable variation in auditory diversity across languages. Russian and Polish show the highest auditory diversity (0.79 and 0.75), while Arabic has the lowest (0.34). Most other languages fall in the middle range.

Interoceptive Diversity: Spanish, French, and German show relatively high

Rank	English	Spanish	French	German	Russian	Polish	Arabic	Hindi	Urdu	Tamil
1	salt	ham	olive	hamburger	vegetable	hamburger	salt	cola	yogurt	mango
2	eating	bacon	french toast	steak	cola	steak	meat	curry	apple	salt
3	consuming	barbecue	bacon	bacon	liquor	cola	condiment	jam	jam	whole milk
4	dinner	rum	cereal	beer	apple	salt	spice	pea	banana	olive
5	lunch	icing	wheat	ham	taste	jam	apple	millet	juice	ice cream
6	olive	meal	jam	jam	salt	macaroni	fruit	sugar	watermelon	buttermilk
7	drink	hamburger	salt	beef	icing	cocoa	dinner	almond	sandwich	tea
8	beverage	salt	millet	yam	meal	brussels sprouts	vegetable	sandwich	pomegranate	cashew
9	gulp	sugar	beef	honey	rum	spice	pomegranate	vinegar	breakfast	chocolate
10	toast	honey	soda	lunch	breakfast	honey	olive	spinach	cashew	mint

Table 6.2: Words representing the top 10 synsets that contributed the most to each language's gustatory prevalence. From each synset we pick a word that acts as an exemplar.

Language	Vision	Audition	Interoception	Gustation	Haptic	Olfactory
English	0.51	0.48	0.37	0.35	0.50	0.28
Spanish	0.89	0.65	0.65	0.49	0.57	0.19
French	0.88	0.56	0.65	0.59	0.67	0.25
German	0.90	0.59	0.62	0.42	0.71	0.31
Russian	0.88	0.79	0.52	0.49	0.64	0.34
Polish	0.87	0.75	0.09	0.47	0.60	0.22
Arabic	0.48	0.34	0.33	0.10	0.32	0.05
Hindi	0.75	0.65	0.57	0.33	0.57	0.11
Urdu	0.48	0.56	0.30	0.22	0.25	0.00

Table 6.3: Average diversity of the sensory modalities in each language.

interoceptive diversity (0.65, 0.65, and 0.62), while Polish shows an unusually low score (0.09). This could indicate significant differences in the role interoceptive concepts have across languages.

Gustatory Diversity: French shows the highest gustatory diversity (0.59), while Arabic and Urdu show the lowest (0.10 and 0.22). This might reflect differences in culinary traditions and the importance of taste descriptions in different cultures.

Haptic Diversity: German and French show the highest haptic diversity (0.71 and 0.67), while Urdu and Arabic show the lowest (0.25 and 0.32).

Olfactory Diversity: Olfactory language shows the lowest diversity scores across all languages, with Urdu and Arabic having extremely low diversity. Even the highest olfactory diversity score (Russian at 0.34) is lower than most scores in other modalities. This aligns with the general observation that olfactory vocabulary tends to be more limited across languages.

Language Specific Patterns: Some languages, like Spanish, French, and German, consistently show high diversity across most modalities. In contrast, Arabic and Urdu generally show lower diversity scores across the board. English shows an interesting pattern with moderate diversity in most modalities but high diversity in haptic sensations.

The variations in diversity can indicate differences in linguistic and cultural emphasis on certain sensory experiences. This becomes apparent when we compare the Wikipedia article for 'bat' in French, a language with relatively

high gustatory diversity, and Urdu, a language with low gustatory diversity.

In the Urdu article, we find that the gustatory terms primarily belong to synsets for fruits like 'mango' and 'banana,' used to describe the dietary habits of bats. In contrast, the French article not only covers the bat's diet but also includes references to semantically distinct concepts. For instance, it mentions Bacardi Rum, which uses a bat as a symbol on its rum bottles. This broader range of gustatory references likely contribute to the higher gustatory diversity in French.

This comparison illustrates how languages with higher sensory diversity tend to incorporate a wider range of concepts and contexts when discussing a particular topic, even when the primary subject (in this case, bats) is not directly related to that sensory modality.

Measuring the Contribution of Synsets to Sensorial Diversity

We adapt the Contribution metric introduced in Equation 5.5 to measure the contribution of synsets to Sensorial Diversity. The contribution of a synset is defined as the average contribution of all words within that synset. Formally, we define the contribution of a synset b to the diversity of a sense s as

$$\text{Cont}_{\text{Div}}(b, s) = \frac{1}{|W_b|} \sum_{\substack{w \in W_b, \\ M(w)=s}} \text{Cont}_{\text{Prev}}(w, s) \tag{6.3}$$

Where W_b is the set of words in synset b, $M(w)$ is the dominant modality of word w $\text{Cont}_{\text{Div}}(w, s)$ is the contribution of word w to the diversity of sense s calculated using equation 5.5.

Mirroring equation 6.2, we create a contribution vector for each language. This vector contains the contributions towards diversity of all synsets to a particular sensory modality in that language. The dimension of this vector is equal to the total number of synsets in our lexicon. For a language L and sensory modality s, the contribution vector V_L,s is

$$V_L, s = [\text{Cont}_{\text{Div}}(b_1, s), \text{Cont}_{\text{Div}}(b_2, s), ..., \text{Cont}_{\text{Div}}(b_n, s)] \tag{6.4}$$

Where $b_1, b_2, \cdots, b_n$ are all the synsets in our lexicon.

Conducting a synset-level contribution analysis for each sensory modality, we find that the average similarity between languages is relatively low (less

than 0.5) for the 'higher senses' — vision and audition. However, as we move toward the other senses, patterns in sensory language become more distinct and varied across languages.

We observe that interoceptive language shows higher overall similarity across languages, particularly among European languages (Appendix C). Spanish exhibits relatively high similarities with French (0.58) and English (0.54). In contrast, Arabic shows very low similarity with other languages. The Indo-Iranian languages Urdu and Hindi display moderate similarity with each other and with other languages, despite their linguistic relationship.

This is further evidenced when we consider Table 6.4, which shows the top 10 synsets contributing to interoceptive diversity. We observe distinct patterns in how different languages encode interoceptive experiences:

The European languages (English, Spanish, French, German) show notable variation in their expression of physical and emotional states. For example English emphasizes discomfort and relief with words like 'suffer', 'congested', 'hurt', 'relief' and Spanish focuses on relaxation states with words like 'relaxer', 'relaxant', 'relaxing'. While the Slavic languages show a strong focus on emotional experiences. As an example Russian emphasizes positive emotions with words like 'euphoria', 'elation', 'satisfaction' and Polish focuses on emotional states with words like 'lonely', and 'nervous'.

Arabic and the Indo-Iranian languages (Hindi and Urdu) show distinct patterns. Arabic focuses on physical discomfort 'sore', 'stressor', 'itch'. Hindi emphasizes bodily states and sensations ('dream', 'pain', 'indigestion') and Urdu combines emotional and physical states with words like 'cold', 'passion', 'pleasant'.

Rank	English	Spanish	French	German	Russian	Polish	Arabic	Hindi	Urdu
1	suffer	relaxer	internally	self	euphoria	sense	ulcer	dream	cold
2	repentance	relaxant	swallow	hunger	elation	lonely	sore	pain	passion
3	resting	relaxing	hunger	starve	satisfaction	nervous	stressor	indigestion	pleasant
4	congested	fun	hate	starvation	euphoria	bold	itch	defecate	health
5	hurt	delicious	itch	sting	sorrow	fret	love	migraine	happiness
6	comfortable	palsy	rejuvenate	forget	sad	starvation	nervousness	wakening	thirst
7	relief	paralyze	pride	slow	happy	sad	emotion	waking	hope
8	release	internally	empathetic	discontent	nostalgia	sorrow	sentimentalism	stress	bold
9	remedy	remorse	high	thrilled	homesickness	sadness	terror	itch	sad
10	risk	twinge	satisfaction	angry	sadness	heartache	awkwardness	scabies	digested

Table 6.4: Words representing the top 10 synsets that contributed the most to each language's interoceptive diversity.

Haptic diversity reveals varied patterns of similarity. We observe a high similarity between the Slavic languages Russian and Polish, with a score of 0.65, while Arabic and Urdu show relatively low similarity with other languages. Olfactory diversity exhibits low similarities across language pairs, with one exception: the Romance languages Spanish and French have a notable similarity score of 0.64. Finally, gustatory language shows higher overall similarity across languages compared to auditory and visual modalities.

Overall, linguistic families appear to influence similarities in some cases (e.g., Russian and Polish), but cultural and geographical factors also seem to play significant roles (e.g., Spanish and English). European languages generally show higher similarities across most modalities compared to non-European languages.

RQ 2 (a): How does sensory diversity vary across languages?

Our findings suggest that sensorial diversity is shaped by both linguistic and cultural factors, with some languages demonstrating broader sensory vocabularies and more diverse uses of sensorial descriptors. Languages like French, Spanish, and German consistently show higher diversity across multiple sensory modalities, whereas Arabic and Urdu show lower diversity. The varying levels of diversity across modalities and languages suggest that sensorial experiences are encoded differently in different linguistic and cultural contexts. For instance, the consistently low olfactory diversity across all languages aligns with previous research indicating a limited olfactory vocabulary in most cultures [49]. In contrast, the high visual and auditory diversity in many languages reflects the primacy of these senses in human perception and communication [88].

Sensorial Style Across Languages

We analyze the sensorial style of languages more holistically by combining their sensorial prevalence and diversity patterns into a single representation. We create a composite vector for each language that concatenates the synset contribution vectors for both prevalence and diversity across all sensory modalities. We focus on the Indo-European languages in this analysis.

The hierarchical clustering analysis in fig. 6.3 largely aligns with known phylogenetic relationships in languages. The Slavic languages (Russian and Polish) cluster closely together, similarly, the Romance languages (French and Spanish) form their own cluster. English — despite being a Germanic language, shows closer affinity to this Romance cluster than to German. This likely reflects the substantial Romance (particularly French) influence on English vocabulary following the Norman Conquest. The Indo-Iranian languages (Hindi and Urdu) show the greatest distance from the European language clusters, suggesting distinct patterns in how these languages encode sensory experiences.

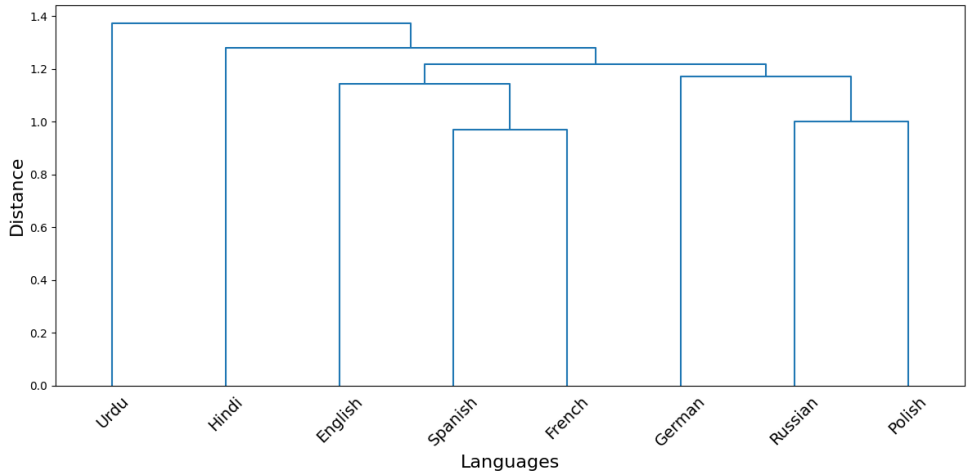


Figure 6.3: Hierarchical clustering of languages based on their composite sensorial style vectors. The y-axis represents the Ward distance metric used in the clustering algorithm. Greater vertical distances indicate more distinct sensorial style patterns between language clusters.

These phylogenetic patterns in sensorial style suggest that while some aspects of sensory language may be universal, the specific ways languages encode sensory experiences are influenced by both linguistic inheritance (Slavic and Romance language members being closer to each other) and cultural contact i.e. the position of English being sensorially closer to Romance languages.

Our examination of Sensorial Prevalence and Diversity in Wikipedia articles helps us understand how different languages encode sensory experiences

	Vision	Audition	Interoception	Gustation	Haptic	Olfactory
English	0.66	0.004	0.02	0.18	0.12	0.015
Spanish	0.77	0.007	0.02	0.18	0.02	0.012
French	0.75	0.004	0.03	0.16	0.05	0.003
German	0.56	0.056	0.05	0.21	0.07	0.047
Russian	0.62	0.007	0.16	0.18	0.02	0.002
Polish	0.32	0.002	0.48	0.13	0.04	0.028
Arabic	0.68	0.010	0.03	0.20	0.08	0.000
Hindi	0.67	0.032	0.01	0.15	0.14	0.001

Table 6.5: Average prevalence of the sensory modalities in Recipe Instructions.

in general encyclopedic content. However, language use can vary significantly depending on the context and purpose of communication. To gain a more comprehensive understanding of sensorial style across languages and cultures, we to expand our analysis to include more specialized forms of text. With this in mind, we now turn our attention to two distinct types of specialized corpora: Music Lyrics and Recipe Instructions.

6.5 Sensorial Style of Specialized Corpora: Music Lyrics and Recipe Instructions

To further understand the use of sensory language, we expand our exploration to include more topic-focused corpora, specifically music lyrics and recipe instructions. While both of these corpora focus on specialized topics, the datasets themselves are more heterogeneous across languages. This heterogeneity makes it impossible to have truly parallel datasets, as each language's corpus reflects different cultural contexts, artistic expressions, and culinary traditions.

Song Lyrics:

Songs are typically unique to each culture and language's tradition. Song translations can often lose the original meaning, rhythm and context, making it impossible to create a dataset of parallel song lyrics. Instead, we selected 1000 random songs independently for each language and used the Genius API to extract their lyrics. Both Tamil and Urdu had less than 1000 songs on

	Vision	Audition	Interoception	Gustation	Haptic	Olfactory
English	0.63	0.11	0.22	0.005	0.04	0.002
Spanish	0.76	0.07	0.15	0.003	0.01	0.002
French	0.78	0.10	0.08	0.005	0.03	0.001
German	0.54	0.14	0.29	0.008	0.01	0.003
Russian	0.74	0.09	0.13	0.012	0.02	0.008
Polish	0.60	0.09	0.28	0.004	0.03	0.001
Arabic	0.79	0.11	0.08	0.006	0.02	0.001
Hindi	0.59	0.14	0.24	0.003	0.02	0.002

Table 6.6: Average prevalence of the sensory modalities in Music Lyrics.

Genius and we exclude them from this part of our analysis.

Recipe Instructions:

Recipes, like songs, are a product of their cultural context. As with songs, we compiled a recipe instruction dataset by randomly selecting 1000 recipes for each language using Cookpad. We exclude both Tamil and Urdu, since they had less than 1000 recipes. Cookpad did not have recipes available in Russian[2]. We used Ovkuse[3] to collect Russian recipe instructions.

Results

We present the results for both Sensorial Prevalence and Sensorial Diversity for the specialized corpora below.

Sensorial Prevalence:

We find that visual language continues to dominate across all languages in both Recipe Instructions and Music Lyrics, mirroring the trend observed in the Wikipedia corpus. However, the prevalence of visual terms varies more widely. It ranges from 0.32 in Polish to 0.77 in Spanish Recipe Instructions. Gustatory prevalence is significantly higher compared to the Wikipedia corpus ranging from 0.13 in Polish to 0.21 in German. Similarly, haptic prevalence was higher for almost all languages.The increase in prevalence underscores the importance of taste and texture in culinary instructions.

[2]Cookpad removed Russian recipes (https://cookpad.com/ru) following the Russian invasion of Ukraine.

[3]https://ovkuse.ru/

Shifting to Music Lyrics, we note that visual language remains dominant. However, prevalence of interoceptive terms significantly increased for almost all languages. This likely reflects the emotional and introspective nature of musical expression.

Looking at the tables of interoceptive and auditory contributions in music lyrics reveals interesting patterns in how different languages encode sensory and emotional experiences in musical expression. For interoceptive language (Table 6.7), we see a strong emphasis on emotional and physical sensations across languages. English lyrics tend to use more aggressive or intense interoceptive terms (e.g., 'fuck', 'crush', 'comeback'), while Spanish and French favor terms related to fear and anxiety ('afraid', 'fear', 'trepidation'). German and Polish show similar patterns focusing on anxiety and pain. Russian presents an interesting contrast, including both negative sensations ('hurt', 'hunger') and states of calmness ('quiet', 'tranquility'). Arabic and Hindi demonstrate a mix of physical and emotional states, with Arabic including both negative ('tense', 'hunger') and positive ('love') terms, while Hindi focuses more on pain-related concepts.

In contrast, the auditory contributions (Table 6.8) show more variation in how languages encode sound in lyrics. English relies heavily on speech-related terms ('say', 'read', 'reckon'). Romance languages (Spanish and French) emphasize vocal sounds and music-related terms ('song', 'vocal', 'chatter', 'music'). German includes more onomatopoeic sounds ('squeak', 'squeal', 'creak') alongside vocal terms. Russian and Polish focus strongly on voice and speech-related concepts, while Arabic shows a specific focus on musical performance

Rank	English	Spanish	French	German	Russian	Polish	Arabic	Hindi
1	fuck	afraid	stress	panic	uneasy	scared	tense	you
2	comeback	fear	afraid	angst	hurt	hurt	hunger	pain
3	caress	trepidation	fear	fear	hunger	pain	passion	striving
4	ego	ache	scare	trepidation	famine	sore	sad	painfully
5	refuge	pain	care	afraid	seething	terror	love	hurting
6	crush	woe	uncertainty	anxiety	quiet	apprehension	sadly	ache
7	red	distress	doubt	terror	tranquility	fear	sleep	delight
8	redemption	sorrow	strain	suffering	calmness	health	slumber	rest
9	warrant	hurting	fuss	hurt	starvation	tense	infatuation	sleep
10	runaway	blissful	panic	apprehension	chill	stress	fuck	afraid

Table 6.7: Words representing the top 10 synsets that contributed the most to each language's Interoceptive prevalence in Music Lyrics.

('violinist', 'violin', 'jazz'). Hindi includes a mix of basic sound terms ('hum', 'noise') and human vocalizations ('laugh', 'snicker'). In general, we find that interoceptive terms tend to cluster around emotional and physical sensations fairly consistently across languages, auditory terms show more cultural and linguistic variation in how sound is encoded and emphasized in lyrics.

Rank	English	Spanish	French	German	Russian	Polish	Arabic	Hindi
1	no	music	chatter	squeak	voice	no	register	hum
2	sash	sound	click	squeal	vocalism	talk	sound	yeah
3	negation	song	snap	creak	dial	verbalize	voice	liar
4	viola	vocal	machine	no	telephone	echo	violinist	yes
5	say	talk	music	talk	word	brave	violin	noise
6	rain	discourse	release	drum	language	reply	vocalization	racket
7	read	say	laugh	laughter	quiet	response	vocalism	hiss
8	reckon	speak	vote	laugh	speak	hear	jazz	snicker
9	verify	beat	voice	loud	quietly	shriek	recitation	laugh
10	swear	tale	voice_over	vocalization	lie	cry	exclaim	echo

Table 6.8: Words representing the top 10 synsets that contributed the most to each language's Auditory prevalence in Music Lyrics.

RQ 1 (b): Does the hierarchical relationship exist across texts? A general sensory hierarchy in Sensorial Prevalence exists across texts and dominance of visual language persists across all contexts. However, the sensory hierarchy demonstrates significant flexibility. Gustatory terms become more prevalent in recipes, while interoceptive and auditory terms have higher prevalence in lyrics, reflecting the communicative needs of different text types and reflecting cultural variations in sensory emphasis.

Sensorial Diversity:

Examining the sensorial diversity in Music Lyrics (Table 6.9) and Recipe Instructions (Table 6.10), we observe patterns that both align with and diverge from those found in the Wikipedia corpus.

In Music Lyrics, visual diversity remains high across most languages, it decreased on average compared the Wikipedia corpus (0.69 in Music Lyrics compared to 0.74 in Wikipedia). There are some notable shifts within visual diveristy. As an example, English shows a significant decrease in visual diversity (0.37 in lyrics vs 0.51 in Wikipedia), while Polish maintains its high diversity (0.87 in lyrics vs 0.87 in Wikipedia). Interoceptive diversity in lyrics shows notable differences from Wikipedia. Overall, interoceptive diversity decreased from 0.45 in Wikipedia to 0.35 in Music Lyrics. However, not all

	Vision	Audition	Interoception	Gustation	Haptic	Olfactory
English	0.37	0.38	0.23	0.33	0.40	0.36
Spanish	0.79	0.53	0.39	0.22	0.24	0.05
French	0.82	0.58	0.50	0.50	0.56	0.21
German	0.81	0.69	0.27	0.19	0.48	0.22
Russian	0.72	0.46	0.51	0.24	0.28	0.18
Polish	0.87	0.49	0.36	0.09	0.42	0.00
Arabic	0.51	0.20	0.20	0.00	0.00	0.00
Hindi	0.69	0.35	0.38	0.13	0.18	0.00

Table 6.9: Average diversity of the sensory modalities in Music Lyrics.

	Vision	Audition	Interoception	Gustation	Haptic	Olfactory
English	0.51	0.15	0.14	0.40	0.35	0.03
Spanish	0.81	0.26	0.17	0.50	0.46	0.00
French	0.79	0.22	0.33	0.41	0.68	0.00
German	0.80	0.11	0.44	0.47	0.72	0.01
Russian	0.75	0.33	1.00	0.43	0.57	-
Polish	0.76	-	0.00	0.33	0.7	0.00
Arabic	0.33	-	0.26	0.34	0.39	-
Hindi	0.80	0.22	0.33	0.33	0.24	

Table 6.10: Average diversity of the sensory modalities in Food Recipes. The symbol '-' indicates cases where insufficient data was available to calculate diversity scores.

languages saw a decrease. Polish shows a dramatic increase (0.36 in lyrics vs 0.09 in Wikipedia), indicating a greater emphasis on expressing internal sensations and emotions in Polish lyrics compared to general text.

Shifting to Recipe Instructions, we observe some striking differences from the Wikipedia corpus. While visual diversity remains generally high, for most languages, we saw a decrease in visual language compared to Wikipedia (0.69 in recipes). Both gustatory and haptic diversity show significant variations across most languages compared to the Wikipedia set. Overall the average haptic and gustatory diversities remained similar to Wikipedia.

RQ 2 (b): Are patterns of sensorial diversity consistent across different types of texts?

Patterns of sensorial diversity vary across different text types. Visual di-

versity decreased in both Music Lyrics and Recipe Instructions compared to Wikipedia, while trends for other modalities were less clear. At the language level, some languages showed varying patterns - for example, Polish had higher interoceptive diversity in Music Lyrics but lower gustatory diversity in Food Recipes. Conversely, English showed lower interoceptive diversity in Music Lyrics but increased gustatory diversity in Recipes. Significance tests confirmed that most languages exhibited different sensorial diversity patterns in specialized corpora compared to Wikipedia.

6.6 Conclusion

This study set out to investigate sensorial language use across different languages and types of corpora, guided by two main research questions focused on Sensorial Prevalence and Sensorial Diversity. Our findings reveal both universal trends and significant variations, providing valuable insights into the complex relationship between language, culture, and sensory perception.

Revisiting our research questions:

RQ 1: Sensorial Prevalence Across Languages and Texts

(a) We found that a general hierarchy of sensory modalities does exist across languages, with vision consistently dominating, followed by audition, interoception, haptic, gustation, and olfaction. This hierarchy persists across different languages in the Wikipedia corpus, suggesting a universal trend in human perception and language.

(b) However, when examining specialized corpora, we observed that this hierarchy exhibits flexibility. In recipe instructions, gustatory and tactile terms became more prevalent, while in music lyrics, interoceptive language gained prominence. This demonstrates that the context of communication significantly influences the sensory emphasis in language use.

RQ 2: Sensorial Diversity Patterns

(a) Our analysis revealed that patterns of sensorial diversity vary across languages, reflecting both linguistic and cultural factors. Some languages, like French and German, consistently showed higher diversity across multiple

sensory modalities, while others, like Arabic and Urdu, generally exhibited lower diversity.

(b) The patterns of sensorial diversity were not consistent across different types of texts. For instance, visual diversity decreased in both music lyrics and recipe instructions compared to Wikipedia, while other modalities showed varied patterns depending on the language and text type.

One limitation of this work is its reliance on a lexicon derived primarily from Indo-European languages. Even though both Arabic and Tamil were studied, both these languages can be considered Indo-European adjacent because of their close cultural, historical and geographical proximity to the Indo-European languages (Hindi and Tamil, Spanish and Arabic). Which might introduced potential biases when extending the analysis to non other Indo-European languages. The lexicons used in this study, based on synset matching, might not fully capture the nuances of sensory language in non Indo-European languages like the Bantu or Sino-Tibetan languages.

Chapter 7

EXPLICATING STYLE

7.1 Introduction

Linguistic style includes traditional style features like sentence length, language complexity, sentiment, and syntactic structure that we discussed in Chapter 2, but these also include patterns in the language used to describe sensory experiences or sensorial style that we discussed in Chapters 3 and 6.

Stylometrics has largely overlooked patterns in the use of sensorial language, while standard stylometric lexicons, such as LIWC, include some sensorial terms, these are generally distributed across different LIWC subcategories and their coverage of the sensorial language space is sparse. Moreover, there has been no focused investigation of the *relationship* between traditional stylometrics and sensorial language. Thus, we do not know for example if these two major dimensions of linguistic style are independent of each other or related to some degree. Our goal is to investigate this relationship.

Our motivation for studying this relationship stems from theories in cognitive science. The interaction between different dimensions of linguistic style can be modeled using cognitive frameworks similar to the 'mental lexicon' proposed by [40], which posits a central repository of linguistic knowledge that mediates various aspects of language processing. We extend this idea to propose a *Central Language Processing Unit* (CLPU) that coordinates interactions between different representations of linguistic style.

This model of interactions within linguistic style also aligns with the grounded cognition theory [7], which suggests that linguistic processes are closely tied to the brain's perceptual, motor, and introspective systems. This theory implies that how we articulate sensory and bodily states influences our language use. However, previous research in this area has been limited to

[0]This work is under review for International Conference on Learning Representations 2025 [4]

small-scale studies [92, 66]. In this context, our work aims to bridge this gap in stylometric research by computationally modeling the relationship between traditional style features and sensorial style across large and diverse text collections. We propose a novel approach to modeling this relationship, drawing from stylometrics, sensorial linguistics as well as cognitive sciences.

Our work makes the following contributions to the field:

- We model the interactions within traditional LIWC-style and sensorial style using Reduced-Rank Ridge Regression (R4). We use R4 to identify low-rank group structures within LIWC-style.
- We introduce Stylometrically Lean Interpretable Models (SLIM-LLMs), which provide a more interpretable lens to study the relationship between traditional linguistic style and sensorial style.
- We conduct large-scale analysis across diverse text genres, providing empirical support for theoretical claims about the interaction between different aspects of linguistic style.

7.2 Methods

Representing Sensorial Style

Sensorial style is modeled and represented across a range of granularities. A synaesthesia-based approach has been used to model sensorial style at a high level (Chapter 3). A high-level approach to modeling sensorial style focuses on patterns of sensory language-use across broader linguistic units or entire texts, rather than on individual words. In contrast, we model sensorial style at the word-level, which focuses on individual sensorial words and their relationships to other linguistic style features.

We represent a sensorial sentence as a one-hot encoding of the sensorial vocabulary. In Chapter 3 we defined the sensorial vocabulary V as a subset of 18,749 words from the Lancaster Sensorimotor Lexicon [48]. We consider a sentence to be sensorial if it has one or more sensorial words in it. We use this criterion and consider a sensorial sentence to have just one sensorial

term. For example, 'it is a noisy room' has two sensorial words, the auditory 'noisy' and the visual 'room'. Assuming 'noisy' and 'room' are the second and fourth words in the sensorial vocabulary, this sentence constitutes two sensorial sentences represented as [0, 1, 0, 0, ..., 0] for 'noisy' and [0, 0, 0, 1, ..., 0] for 'room'. The length of the two vectors equals the size of our sensorial vocabulary; that is, $|V| = 18,749$.

We formalize the previous idea as follows. Let $V = \{w_1, w_2, ..., w_n\}$ be the sensorial vocabulary of size n. For a given sensorial word w in a sentence, we represent it as a vector $\mathbf{y} \in \{0,1\}^n$, where $y_i = 1$ if $w = w_i$ and 0 otherwise. A sentence S with m sensorial words is represented as a set of m n-vectors and $S = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m\}$, where $\mathbf{y}_j$ $(j = 1, \ldots, m)$ corresponds to the one hot encoding of the jth sensorial sentence.

We represent each sensorial sentence as a vector based on the LIWC-style [64]. Let $X = \{x_1, x_2, ..., x_m\}$ be the set of m LIWC categories. For a given sensorial sentence S, we exclude the sensorial term w_s and represent the style of the remaining sentence as a vector $\mathbf{s} \in \mathbb{R}^m$. Each element s_i of this vector corresponds to the proportion of words in S excluding w_s that belong to the i^{th} LIWC category x_i: $s_i = (\big|\{w \in S \setminus \{w_s\} : w \in x_i\}\big|)/(|S| - 1)$.

For example, given the sentence 'it is a noisy room' with two sensorial words 'noisy' and 'room', we create two style vectors. For 'room', the style vector will be based on ['it', 'is', 'a', 'noisy'], and for 'noisy' the style vector will be based on ['it', 'is', 'a', 'room']. This is comparable to the BERT masked language model setup, where each sensorial word is treated as the target word to be predicted, and the embedding is calculated from the remaining words in the sentence.

Linear Models for Style Interactions

We use regression to model the relation between traditional style and sensorial style. Let the style features of a sentence S be the LIWC vector $\mathbf{x} = (x_1, \ldots x_m)$ and let $\mathbf{y} = (y_1, y_2 \ldots y_n)$ be the one-hot sensorial vector of the sentence, where m is the number of style features and n is the size of the sensorial vocabulary $\mathcal{S}$ [48]. Then, $\mathbf{y}^\top = \mathbf{x}^\top \mathbf{B} + \mathbf{e}^\top$ models the relation

between linguistic style $\mathbf{x}$ and sensorial language use $\mathbf{y}$, with $\mathbf{e}$ denoting the errors independent of $\mathbf{x}$. The regression coefficient matrix is $\mathbf{B} \in \mathbb{R}^{m \times n}$, and its $(i, j)$the element b_{ij} is the mean increase in the sensorial word y_j for a unit increase in style feature x_i, given other features in $\mathbf{x}$ remain unchanged. The linear regression model is equivalent to a sensorial-word-prediction problem, where we predict the sensorial word w_s in a sentence from the linguistic style of the remaining text. This method is analogous to the masked word prediction task used to train LLMs like BERT [24].

We fit the regression model to the training data as follows. For a set of k sentences, the ith sentence has sensorial vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})$, and its corresponding style vector is $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$. The training data are represented as the $k \times n$ matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_k]^\top$ and $k \times m$ matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_k]^\top$. For a sufficiently large k, the least squares estimate of $\mathbf{B}$ is $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ [67]. Previous works have shown that LIWC features have a low-rank structure [26]. However, the standard least squares approach fails to capture this structure and the latent dependencies between the sensorial features and LIWC-style features, which correspond to the columns of $\mathbf{Y}$ and $\mathbf{X}$. This limitation is particularly significant because not all LIWC features capture the same amount of information. For example, the function category words are more informative than categories like fillers. The word categories have group behavior. For instance, in the LIWC features, first person singular is a subcategory of personal pronouns, whereas the ingestion category contains words like '*eat*' that also belong to the verb category.

Reduced-Rank Ridge Regression

We circumvent the previous limitations by assuming that $\mathbf{B}$ is a low-rank matrix. This assumption implies that the previous linear model becomes a reduced-rank regression model [3], which assumes that $\mathbf{B}$ has a rank r and $r \ll \min\{m, n\}$. In a sparse $\mathbf{B}$, a large fraction of the entries are 0, where $b_{ij} = 0$ denotes that x_i and y_j are not associated. Similarly, a row sparse $\mathbf{B}$ has $b_{ij} = 0$ for $j = 1, \ldots, n$ for many is. If the ith row of $\mathbf{B}$ is zero, then x_i is not associated with any sensorial word. To model a rank-r $\mathbf{B}$, we set $\mathbf{B} = \mathbf{U}\mathbf{V}^\top$,

where $\mathbf{U} = (u_1, u_2 \dots u_r) \in \mathbb{R}^{m \times r}$ and $\mathbf{V} = (v_1, v_2 \dots v_r) \in \mathbb{R}^{n \times r}$. By assuming row sparsity of $\mathbf{B}$, we can effectively select a subset of LIWC features that have the strongest associations with sensorial words across different contexts. This assumption is more appropriate for our goals of identifying the most influential LIWC features that contribute to sensorial language use.

$$\hat{\mathbf{U}}_s, \hat{\mathbf{V}}_s = \underset{\mathbf{U} \in \mathbb{R}^{m \times r},\, \mathbf{V}^\top \mathbf{V} = \mathbf{I}_r}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|_F^2 + \lambda \sum_{j=1}^{m} \|\mathbf{U}_j\|_2^2, \quad \hat{\mathbf{B}}_s = \hat{\mathbf{U}}_s \hat{\mathbf{V}}_s^\top \quad (7.1)$$

where $\hat{\mathbf{B}}_s$ is the SRRR estimate of $\mathbf{B}$, $\mathbf{I}_r$ is an $r \times r$ identity matrix, $\|\cdot\|_F$ is the Frobenius norm, and $\|\mathbf{U}_j\|_2$ is the group lasso penalty on the jth row of $\mathbf{U}$ [90]. Qian et al. [67] develop an efficient algorithm for estimating $\mathbf{U}$ and $\mathbf{V}$ using the alternative minimization algorithm, which estimates $\mathbf{U}$ given $\mathbf{V}$ and vice versa [67]. The group lasso norm on $\mathbf{U}$ rows implies that some of the $\mathbf{B}_s$ rows are zeros, but the estimation algorithm suffers from computational bottlenecks particularly when k and m are in the order of ten thousand.

We propose Reduced-Rank Ridge Regression (R4) as an efficient alternative to SRRR. The $\mathbf{B}$ matrix in our problem is not sparse because all stylistic features are associated with sensorial words, even when their magnitudes are small; therefore, we replace the group lasso penalty on the $\mathbf{B}$ rows by a ridge penalty to obtain the R4 estimates of $\mathbf{U}$ and $\mathbf{V}$ as

$$\hat{\mathbf{U}}, \hat{\mathbf{V}} = \underset{\mathbf{U} \in \mathbb{R}^{m \times r},\, \mathbf{V}^\top \mathbf{V} = \mathbf{I}_r}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|_F^2 + \lambda \sum_{j=1}^{m} \|\mathbf{U}_j\|_2^2, \quad \hat{\mathbf{B}} = \hat{\mathbf{U}}\hat{\mathbf{V}}^\top, \quad (7.2)$$

where $\hat{\mathbf{B}}$ is the R4 estimate of $\mathbf{B}$ and is obtained by a slight modification of the alternative minimization algorithm in [67]. The estimation algorithm of $\mathbf{V}$ given $\mathbf{U}$ remains the same in (7.1), but the estimation of $\mathbf{U}$ given $\mathbf{V}$ uses ridge regression. Unlike $\mathbf{B}_s$ in (7.1), $\hat{\mathbf{B}}$ is not sparse but has better predictive performance [29]. The columns of $\hat{\mathbf{U}}$ represent the latent factors or components that capture the shared structure between LIWC and sensorial features.

Modeling Non-Linear Style Interactions

The R4 model in (7.2) assumes a linear association between LIWC-style and sensorial style. The associations, however, are nonlinear from linguistic and cognitive perspectives. We model the relationship between LIWC-style and sensorial style as a phenomenon mediated by a *Central Language Processing Unit* (CLPU), using Large Language Models (LLMs) as a proxy for the of the *CLPU*. The *CLPU* is a similar construct to Levelt's 'mental lexicon' [40]. LLMs, trained on vast corpora of human language, encapsulate general language norms and patterns. They capture the complex interactions mediated by our broader linguistic knowledge and cognitive processes [52].

To model this interaction, we represent traditional stylistic features of a sentence using our LIWC-based representation. We then use an LLM for a masked language modeling task on the original sentence, with the sensorial words masked. Finally, we use the LLM's predictions for masked sensorial words, combined with the LIWC-style, to predict sensorial style. Formally, let S be the original sentence, and $m(S)$ be the sentence with sensorial words masked. Let f be the function represented by the LLM that takes the masked sentence $m(S)$ and returns the encoder embedding representation of the masked word. Then, the model relating sensorial words and LLM's encoder embeddings of the masked word is

$$\mathbf{y}_i = g(f(m(S_i)); \mathbf{x}_i) + \mathbf{e}_i, \quad \mathbf{e}_i \in \mathbb{R}^n, \quad i = 1, \ldots, k, \tag{7.3}$$

where S_i is the ith sentence, $\mathbf{y}_i$ and $\mathbf{x}_i$ remain the same as in (7.2), $\mathbf{e}_i$ is the ith error vector, and g is a classifier function that predicts sensorial language use from the combination of the LLM's encoder embeddings and the original stylistic features.

Stylometrically Lean Interpretable Models (SLIM-LLMs)

LLMs like BERT are often overparameterized [53]. This can obscure the relationship between LIWC-style and sensorial style due to redundancies in the model's training. To address this, we propose using dimensionality reduction techniques to create Stylometrically Lean Interpretable Models (SLIM-LLMs).

SLIM-LLMs are reduced versions of standard LLMs that aim to reveal the underlying relationships between LIWC-style and sensorial style more clearly. We create SLIM-LLMs using Singular Value Decomposition (SVD). Let $\mathbf{E} \in \mathbb{R}^{k \times d}$ be the encoder embedding matrix of our LLM, where d is the dimension of the hidden state and k is the number of sentences in our dataset.

The SLIM-LLM retain only the top r singular values and their corresponding singular vectors for the SVD of $\mathbf{E}$ and are denoted as $\mathbf{E}_{\text{slim}}$. Specifically, let $\mathbf{E} = \mathbf{U}\,\mathbf{V}^\top$ be the SVD of $\mathbf{E}$, where $\mathbf{U} \in \mathbb{R}^{k \times k}$ and $\mathbf{V} \in \mathbb{R}^{d \times d}$ are the left and right orthonormal matrices. Then, $\mathbf{E}_{\text{slim}} = \mathbf{U}_r\ {}_r\mathbf{V}_r^\top$, where $\mathbf{U}_r \in \mathbb{R}^{k \times r}$, ${}_r \in \mathbb{R}^{r \times r}$, and $\mathbf{V}_r \in \mathbb{R}^{d \times r}$. The nonlinear classification model relating sensorial words and LLMs in (7.3) is now rewritten for SLIM-LLMs as

$$\mathbf{y}_i = g(f_{\text{slim}}(m(S_i)); \mathbf{x}_i) + \mathbf{e}_{\text{slim}i}, \quad \mathbf{e}_{\text{slim}i} \in \mathbb{R}^n, \quad i = 1, \ldots, k, \tag{7.4}$$

where $\mathbf{e}_{\text{slim}i}$ is the ith error term, f_{slim} is the function represented by our SLIM-LLM that takes the masked sentence $m(S_i)$ as input and outputs a dimension-reduced embedding of $\mathbf{x}_i$, and g is a classifier function that predicts sensorial language use from the combination of the SLIM-LLM's reduced encoder embeddings and the original stylistic features. In this formulation, $f_{\text{slim}}(m(S_i))$ represents the projection of the masked sentence $m(S_i)$ onto the reduced-dimensional space defined by $\mathbf{U}_r$ so that $f_{\text{slim}}(m(S_i)) = \mathbf{U}_r^\top f(m(S_i))$, where $f(m(S_i))$ is the original LLM's encoder embedding for the masked sentence $m(S_i)$. By reducing the dimensionality of the encoder embeddings, we aim to maintain the benefits of using LLMs as proxies for the mental lexicon while revealing more interpretable relationships between the different aspects of linguistic style.

The choice of r, the number of singular values to retain, represents a trade-off between model complexity and interpretability. A smaller r results in a more interpretable model, but may lose some nuanced relationships, while a larger r retains more information but may be less interpretable. The optimal value of r can be determined through empirical analysis.

7.3 Experiments

We use BERT-base [24] to investigate the relationship between traditional style (LIWC-style) and sensorial style across diverse contexts. We study the style of 5 different text genres[1]. This section details the datasets and models used in our study.

Datasets

Language Genre	Datasets	Source	Sensorial Sentences
Critical	Business Reviews	Yelp.com	2,101,603
Literary	Novels	Project Gutenberg	1,929,260
Poetic	Music Lyrics	Genius.com	1,107,749
Persuasive	Advertisements	Airbnb Descriptions	1,442,050
Informative	Articles	Wikipedia	1,563,888

Table 7.1: Overview of text collections and genres

We analyze 5 different text genres. Each language genre represents a distinct way in which language is employed to achieve specific communicative goals or to serve particular purposes.

Critical Language: Reviews from the Yelp Dataset Challenge (2005-2013), encompassing approximately 42,000 businesses.

Literary Language: English novels from Project Gutenberg's Domestic fiction category, spanning works from 18^{th} century author Regina Maria Roche to 20^{th} century writer Lucy Maud Montgomery.

Poetic Language: Lyrics of songs featured on the Billboard Hot 100 charts (1963-2021), obtained via the Genius API. This chart is widely regarded as the music industry benchmark [85].

Persuasive Language: Airbnb property descriptions (2008-2022), showcasing accommodations, amenities, and local attractions to potential guests.

[1]Experiments using BERT-large, DistilBERT [**sanh2019distilbert**] and RoBERTa-base [45] gave comparable results (See: Appendix D.2), thus we only report BERT-base results.

Informative Language: Wikipedia articles, collected in July 2024. Unlike other datasets, these entries are subject to continuous updates, precluding precise dating.

Table 7.1 presents an overview of our text collections and genres, along with the specific number of sensorial sentences extracted from each collection. For our experiments, we randomly select a standardized sample of 300,000 sensorial sentences from each set to ensure consistency across all language aspects.

Results

Latent Representation of LIWC-Style

We investigate the relationship between the latent representation of LIWC-style and sensorial style. To find the optimal number of latent dimensions that best capture LIWC-style, we solve the Reduced-Rank Ridge Regression (R4) for a range of r values from 1 to 74.

Using the reconstructed $\mathbf{B} = \mathbf{U}\mathbf{V}^\top$ for this range of r, we calculate the mean squared error (MSE) on the test data. Figure 7.1 shows the MSE for the five datasets across different values of r.

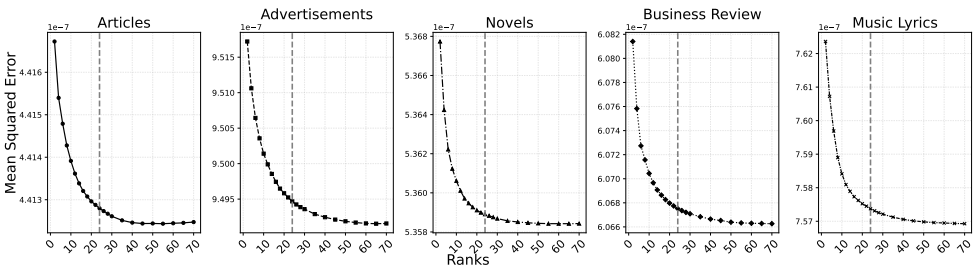


Figure 7.1: Mean Squared Error (MSE) for the five language aspect datasets (Articles, Advertisements, Novels, Business Reviews, and Music Lyrics) plotted against the number of latent dimensions (r) in the Reduced-Rank Ridge Regression (R4) model. The plot shows the decrease in reconstruction error as the number of latent dimensions increases from 1 to 74.

While the reconstruction errors vary in absolute terms between the five genres, we observe a general trend across all datasets. On average, we see the

greatest decrease in the reconstruction error within the first 20 dimensions. The error rate begins to asymptote for values of $r > 20$.

Based on this observation and the diminishing returns in error reduction, we empirically determine that $r \approx 24$ provides an optimal latent dimension representation for LIWC-style. This choice balances model complexity with performance, capturing most of the variance in the data while maintaining interpretability.

This finding suggests that the relationship between LIWC-style features and sensorial language use can be effectively represented in a relatively low-dimensional latent space across diverse language genres.

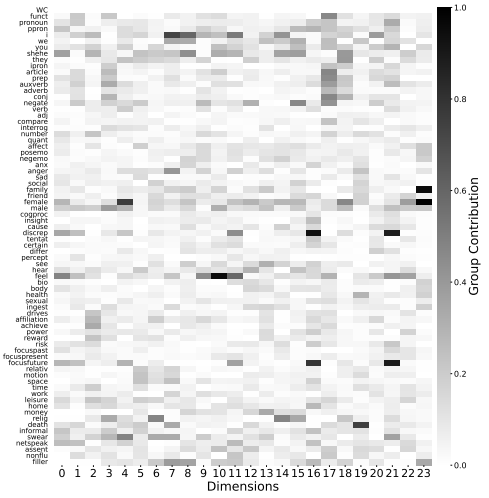


Figure 7.2: Heatmap showing the latent representation of LIWC categories across 24 dimensions for Wikipedia articles. The intensity indicates the strength of the contribution of each LIWC category to each latent dimension.

Group structure in LIWC-Style

In the original formulation of our model, $\mathbf{y}^\top = \mathbf{x}^\top \mathbf{B} + e^\top$, all dimensions of the LIWC features are treated as independent. However, our analysis of the $\mathbf{U} \in \mathbb{R}^{n \times r}$ matrix, which represents the latent dimensions of our Reduced-Rank Ridge Regression (R4) model, reveals group structures indicating inter-dependencies among LIWC features and their collective relationship with sensorial style.

Figure 7.2 illustrates the group structure in the $\mathbf{U} \in \mathbb{R}^{74 \times 24}$ latent representation for Wikipedia articles[2].

[2]See Appendix D.1 for the representations of other genres and more detailed visualiza-

We find similar group structures in the latent representations of other genres as well. From the figure, we note that some latent dimensions appear more influential than others, as indicated by stronger and more widespread contributions across LIWC categories, as an example the Discrepancy category *'discrep'* contributes to both groups 16 and 21. We also find that related LIWC categories often contribute strongly to the same latent dimensions, forming natural groupings. An example of this would be the contribution of function words, categories like *'i'*[3],*'we'*[4], *'shehe'*[5] in Group 17.

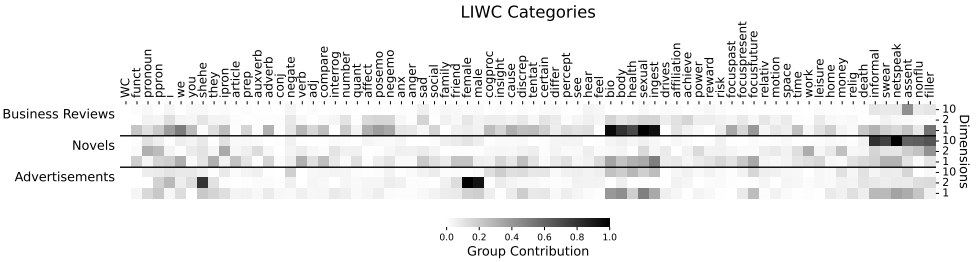


Figure 7.3: The heatmap shows the contribution of LIWC categories to specific latent dimensions, across three genres: Business Reviews, Novels, and Advertisements.

In Figure 7.3, we examine a sample of columns of 3 other genres. We observe that:

Business Reviews (Yelp): A group forms around categories of LIWC biological processes, including words focused on consumption. This aligns with the nature of restaurant reviews, where descriptions of food and eating experiences are central.

Novels (Gutenberg): We observe a group forming around informal language use, including categories related to fillers, non-fluencies, and netspeak. This clustering would reflect the author's attempt to mimic natural, conversational speech patterns in dialogue and narration.

tions.

[3] 1^{st} person pronouns.

[4] 3^{rd} person pronouns.

[5] 2^{nd} person pronouns.

Advertisements (Airbnb): We observe an emergent group that combines elements from disparate LIWC categories, specifically gendered words (masculine and feminine) from the social processes category and gendered pronouns (she/he) from the function word category. This grouping is not apparent in the standard LIWC classification but emerges in our analysis. Such a pattern suggests that Airbnb property descriptions may employ gender-specific language strategies that are not captured by LIWC's predefined categories. This finding demonstrates how our approach can reveal latent linguistic structures that are not immediately evident from simple LIWC groupings, potentially offering new insights into the stylistic techniques used in persuasive advertising language.

These groupings, emerging from the latent representation, reveal how different aspects of language use cluster together in genre-specific ways. They provide insights into the underlying structures of LIWC-style across various text types and how these relate to sensorial style. The presence of these group structures, not accounted for in the original independent dimension assumption, highlights the complexity of the relationship between LIWC-style features and sensorial style.

Exploring LIWC-Style using SLIM-BERT

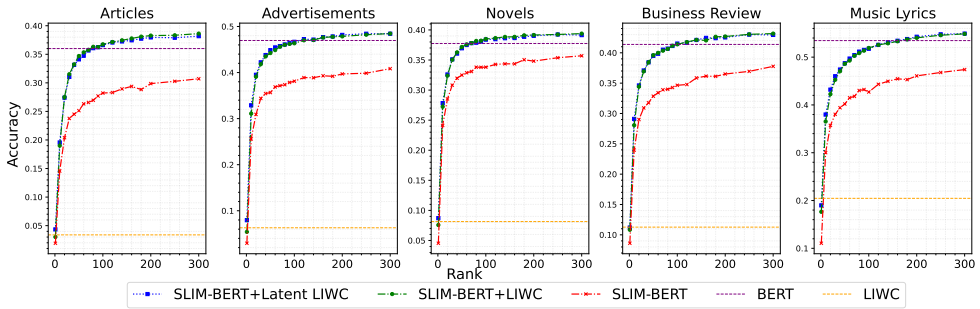


Figure 7.4: Accuracy of sensorial word prediction against the rank (number of dimensions) used in the SLIM-BERT model for different language aspects

We investigate the relationship between linguistic style and sensorial language use by using low-dimensional projections of LLMs — SLIM-LLMs model

augmented with LIWC features. We use these SLIM-LLMs for the sensorial word prediction task described in section 7.2. For each masked sensorial sentence, we extract the SLIM-LLM representation and use it (along with LIWC representations) as input to a fully connected Multi-Layered Perceptron (MLP) that is trained to predict the masked sensorial word. Figure 7.4 presents the performance of BERT-base for each language aspect. We focus on the first 240 dimensions of the SLIM-BERT model.

We compare the performance of three configurations of SLIM-BERT:
SLIM-BERT+Latent LIWC: SLIM-BERT augmented with latent LIWC features.
SLIM-BERT+LIWC: SLIM-BERT augmented with raw LIWC features.
SLIM-BERT: SLIM-BERT without LIWC features.

For reference, we also show the performance of the full BERT-base model and raw LIWC features (shown as horizontal lines).

Across all genres, we observe that augmenting SLIM-BERT with LIWC features (both latent and raw) consistently improves performance over SLIM-BERT alone. For instance, in Articles, SLIM-BERT+Latent LIWC achieves an accuracy of 0.380, compared to 0.299 for SLIM-BERT alone. This pattern is consistent across other categories, with SLIM-BERT+Latent LIWC reaching accuracies of 0.483 for Advertisements, 0.390 for Novels, 0.430 for Business Reviews, and 0.545 for Music Lyrics. These results suggest that linguistic style, as captured by LIWC, provides complementary information to the language model for predicting sensorial language use.

The SLIM-BERT with the Latent LIWC configuration performs as well as or slightly better than SLIM-BERT with the raw LIWC features. For example, in the Music Lyrics category, SLIM-BERT+Latent LIWC achieves 0.545 accuracy compared to 0.543 for SLIM-BERT+LIWC, indicating that the latent representation of LIWC features effectively captures the most relevant aspects of linguistic style for this task, while potentially reducing noise or redundancy in the raw LIWC features.

In most cases, our SLIM-BERT+Latent LIWC configuration approaches or even exceeds the performance of the full BERT model, while using a fraction of the parameters. For instance, in Novels, SLIM-BERT+Latent LIWC achieves

0.390 accuracy compared to 0.378 for the full BERT model. Similarly, for Business Reviews, SLIM-BERT+Latent LIWC reaches 0.430 accuracy, surpassing the full BERT model's 0.416. This demonstrates the effectiveness of our dimensionality reduction approach in capturing the most relevant features for this task. The dimensionality reduction filters out noise and less relevant information, focusing on the most salient features of sensorial language prediction. Additionally, the addition of latent LIWC features provides complementary stylistic information that enhances our model's predictive power.

These results demonstrate the effectiveness of our SLIM-BERT approach in modeling the relationship between linguistic style and sensorial language use. The consistent improvements from LIWC augmentation, particularly using our latent LIWC representation, suggest a strong link between stylometric features and sensorial language across various language aspects. This supports our hypothesis of a mediated interaction between linguistic style and sensorial language, as modeled by our SLIM-LLM framework.

7.4 Conclusion

Our results demonstrate that SLIM-LLMs, such as SLIM-BERT, and LIWC-style features capture complementary aspects of sensorial style across various language genres. The combination of these two representations consistently outperforms either representation alone, supporting our hypothesis of a mediated interaction between LIWC-style and sensorial style through a *Central Language Processing Unit* (CLPU).

For example, in the case of Articles and Advertisements, we observe that the combination of SLIM-BERT and Latent LIWC features achieves higher accuracy than the sum of their individual performances. Specifically, for Articles, SLIM-BERT+Latent LIWC with $r = 240$ achieves an accuracy of 0.48, compared to SLIM-BERT (0.41 at $r = 240$) and LIWC-style (0.06) alone.

While we focused on LIWC-style features in this work, our approach can be extended to incorporate other stylometric features such as ANEW, VADER, and measures of linguistic complexity like Readability and Hapax Legomenon. Such extensions would let us not only study the relationships between these

features and sensorial style, but also the interactions with the rest of the stylometric features.

One limitation of this study is its focus on English language texts. However, the dimensionality reduction technique used to create SLIM-LLMs is not inherently language-specific and is only limited by the underlying LLM's training data. This approach can be extended to other languages by creating SLIM versions of language-specific or multilingual models, such as SLIM-BETO for Spanish (based on the BETO model [13]) or SLIM-mBERT (based on the multilingual BERT model).

Chapter 8

CONCLUSION

This thesis expands our understanding of linguistic style by systematically investigating sensorial style — how individuals and communities encode sensory experiences through language. Through a series of interconnected studies, we have demonstrated that sensorial style is a meaningful and measurable dimension of linguistic style that provides valuable insights into how humans perceive and communicate their sensory experiences.

Our investigation began by establishing that linguistic style manifests beyond individuals, showing that online communities develop distinctive stylistic patterns that can predict community membership with accuracy comparable to content-based approaches (Chapter 2). This foundational work demonstrated the validity of studying style at aggregate levels, laying the groundwork for examining specific stylistic dimensions like sensorial language use. Future research could explore how these stylistic patterns emerge and evolve over time within communities, and how individual differences in sensory perception affect community-level patterns.

Building on this foundation, we introduced one of the first formal representations of sensorial style through the lens of synaesthesia patterns (Chapter 3). Our analysis revealed that these sensorial style patterns are non-random and stabilize with relatively small amounts of text, suggesting they reflect meaningful stylistic choices rather than chance variations. This work opened new avenues for studying how individuals and communities encode sensory experiences linguistically.

While our initial investigation of sensorial style yielded important insights, studying these patterns solely in English provided only a limited view of how humans encode sensory experiences through language. A critical contribution of this thesis was developing methods to extend sensorial analysis beyond English. By creating a systematic approach to extrapolate sensorial lexicons

to multiple languages using BabelNet (Chapter 4), we enabled broader cross-linguistic investigation of sensorial style. Despite including multiple languages, our analysis was primarily focused on Indo-European languages or those (like Arabic and Tamil) with significant cultural contact with Indo-European languages, limiting our understanding of how sensorial style manifests in linguistically and culturally more distant languages. Future research should expand this work to more diverse language families, like the Bantu, Indigenous American, Sino-Tibetan, or Austronesian language families, and potentially develop methods to automatically identify sensorial terms in new languages without relying on translation-based approaches.

This methodological advance from Chapter 4 was complemented by the introduction of additional theoretical frameworks - Sensorial Prevalence and Sensorial Diversity (Chapter 5) - which provided new tools for analyzing how different languages and cultures encode sensory experiences. Our work explored both high-level synaesthetic patterns and representations like prevalence and diversity. Future research could investigate more granular representations of sensorial style. For instance, researchers could explore representations at various levels — such as analyzing patterns in sensory phrases, investigating how multiple sensory terms combine within sentences.

Our cross-linguistic analysis (Chapter 6) revealed both universal patterns and cultural variations in sensorial language use. While we found a consistent hierarchy of sensory modalities across languages (with vision typically dominating), we also discovered significant cultural variations in how different sensory experiences are encoded and emphasized. These findings contribute to our understanding of both universal cognitive constraints and cultural influences on sensory language. This sets the stage for further studies examining how cultural and social factors influence sensorial style development and whether sensorial style analysis can be used to track cultural changes over time. Additionally, this opens up exciting possibilities for investigating how sensorial style develops during language acquisition and how it might be influenced by cultural as well as cognitive differences.

Finally, we investigated the relationship between traditional stylometric features and sensorial style through the development of Stylometrically Lean

Interpretable Models (SLIM-LLMs) (Chapter 7). This work demonstrated that sensorial style interacts with other stylistic dimensions in systematic ways, suggesting an integrated model of linguistic style that encompasses both traditional stylometric features and sensory language patterns. Future work could build on this by developing more sophisticated models of cross-modal sensory language and examining how sensorial style manifests in multimodal communication beyond just text.

This thesis makes several important theoretical contributions to both stylometrics and sensorial linguistics. First, it establishes sensorial style as a measurable dimension of linguistic style, expanding our understanding of how style manifests in language. Second, it demonstrates that patterns in sensory language use exist at multiple levels –- from individuals to communities to entire languages - suggesting a relation between cognitive, social, and cultural factors in shaping linguistic style.

Building on methods from stylometrics and sensorial linguistics, this thesis establishes sensorial style as a measurable dimension of linguistic style and demonstrates that patterns in sensory language use emerge across multiple levels, from individuals to entire languages. While the study focuses on Indo-European languages and related contexts, the methodological advances and theoretical frameworks developed here provide a foundation for future research exploring sensorial style across more diverse linguistic and cultural contexts.

Appendix A

EXPLORING SENSORIAL STYLE

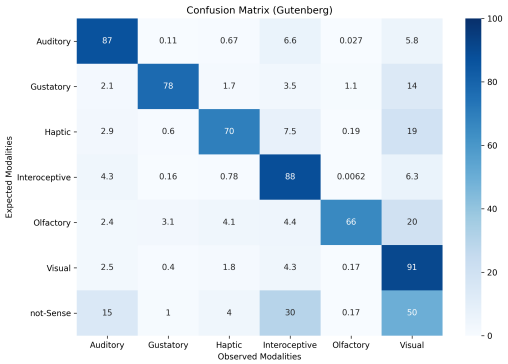


Figure A.1: Distribution of expected-observed modalities in the Novels Dataset. The heatmap shows the proportion of times each sensory modality was observed (columns) when a particular modality was expected (rows). Darker colors indicate higher proportions.

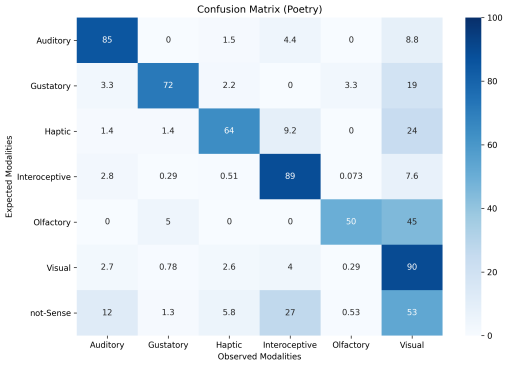


Figure A.2: Distribution of expected-observed modalities in the Poetry Dataset.

Appendix B

EXTRAPOLATING SENSORIAL LEXICONS

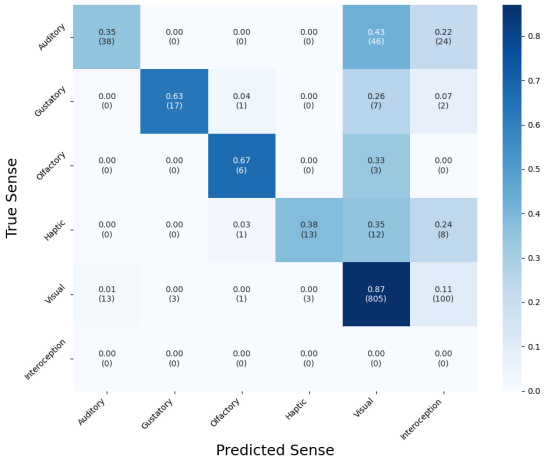


Figure B.1: Dominant sense prediction accuracy for Italian. Values show the probability of each prediction (0-1) with raw counts of occurrences shown in parentheses.

```
A word may align with one or more of 6 distinct sensorial dimensions: hearing,
     taste, touch, smell, sight and interoception
(the perception of sensations from inside the body, both physical such as
    hunger and pain, and emotional, such as joy).

Consider the following English words and their sensory alignments:
Barking: Hearing
Fluffy: Touch
Headache: Interoception
Unicorn: Sight

You are given a number of English words and possible sensorial modalities it
    can align with.
For each word, identify which sensorial modality is more likely to align with
    the word.
```

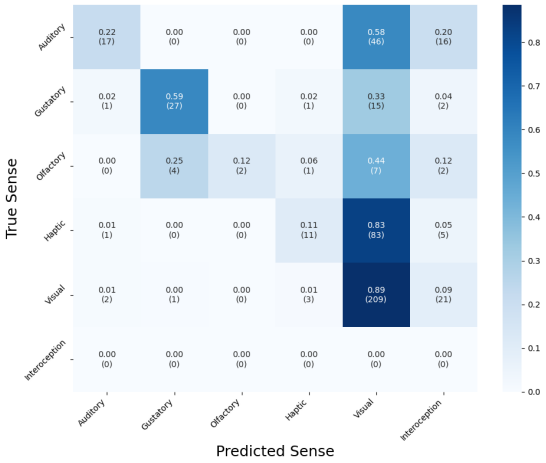


Figure B.2: Dominant sense prediction accuracy for Russian. Values show the probability of each prediction (0-1) with raw counts of occurrences shown in parentheses.

Listing B.1: English Prompted used for user study

Appendix C

EXPANDING SENSORIAL STYLE ANALYSIS TO MULTIPLE LANGUAGES

C.1 Wikipedia Topics

Table C.1: Distribution of Topics in the Wikipedia Dataset

Category	No. of Topics	Examples
STEM	828	Actinium, Escalator Krakatoa, Tropics Lizard
Geography	1020	Bolu Province, Pune England, Ottoman Empire Hephthalites
Culture	692	Nihilism, September 6 Shamanism, John Cena Xbox 360
History and Society	460	Isis, Sumer Volgograd, Rostov-on-Don Tony Blair

The complete list of the 3000 topics and their categories can be accessed at https://github.com/osama-khalid/multilingual_style/blob/main/wikipedia_topics.txt

C.2 Stylistic Similarities

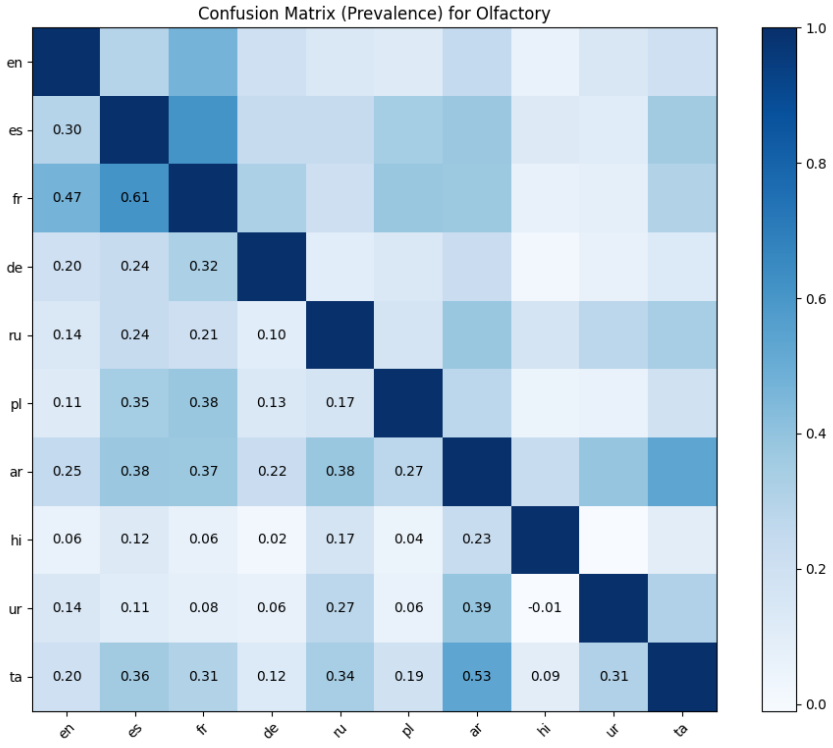


Figure C.1: The cosine similarities of the vectors of contribution of synsets in Olfactory Prevalence.

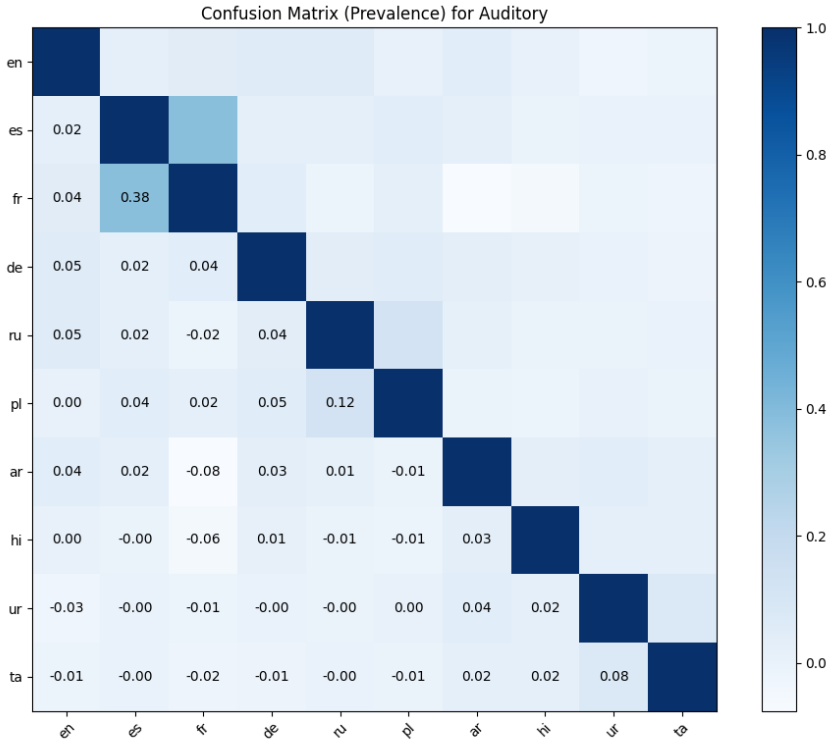


Figure C.2: The cosine similarities of the vectors of contribution of synsets in Auditory Prevalence.

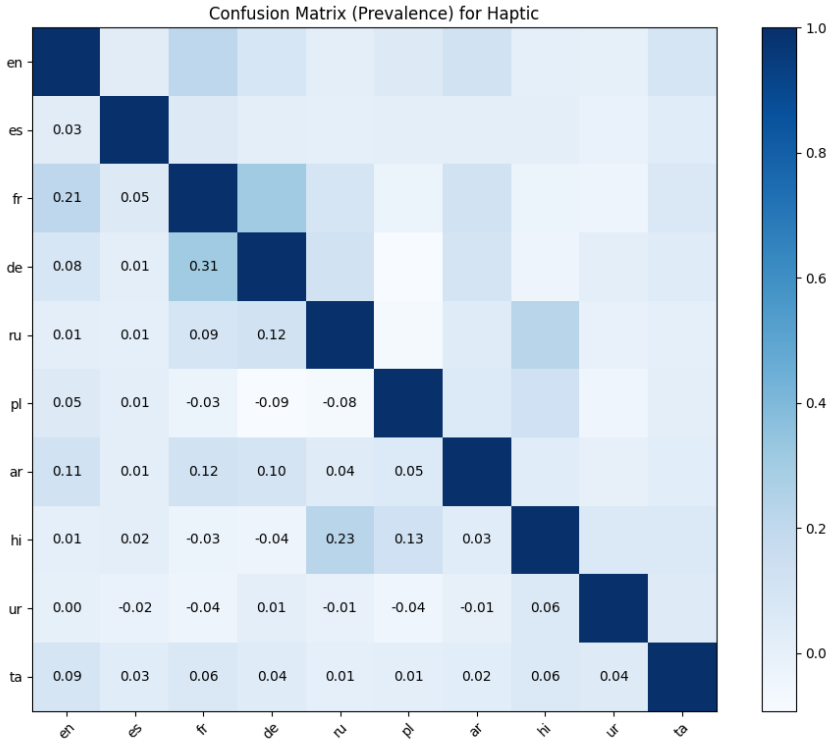


Figure C.3: The cosine similarities of the vectors of contribution of synsets in Haptic Prevalence.

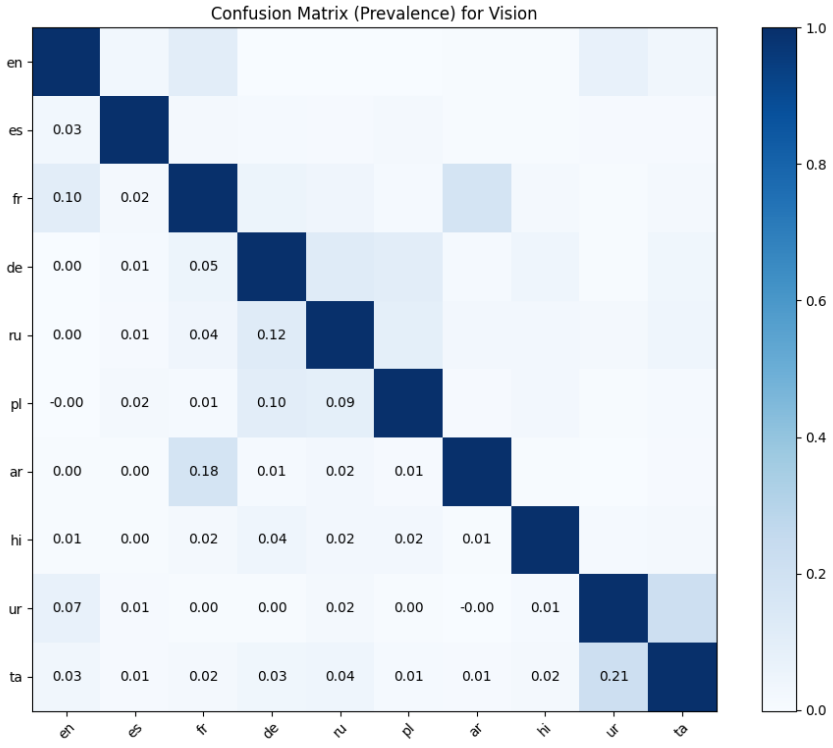


Figure C.4: The cosine similarities of the vectors of contribution of synsets in Visual Prevalence.

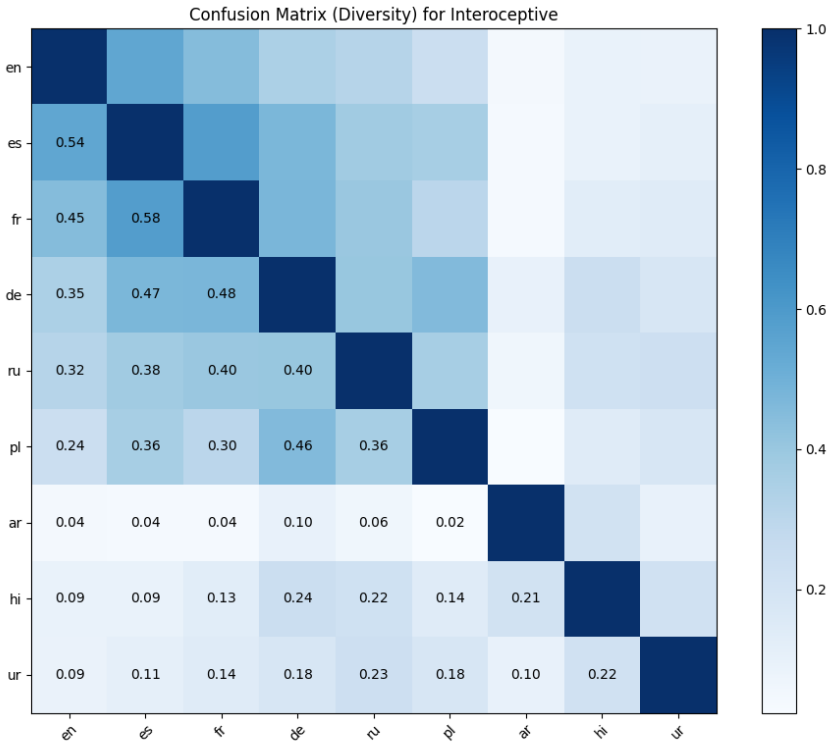


Figure C.5: The cosine similarities of the vectors of contribution of synsets in Interoceptive Diversity.

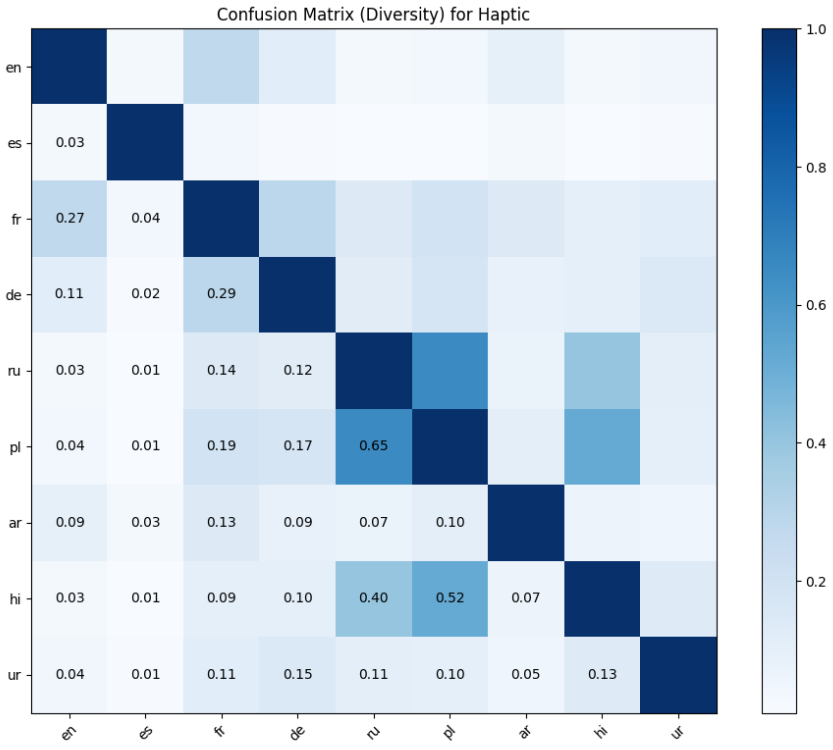


Figure C.6: The cosine similarities of the vectors of contribution of synsets in Haptic Diversity.

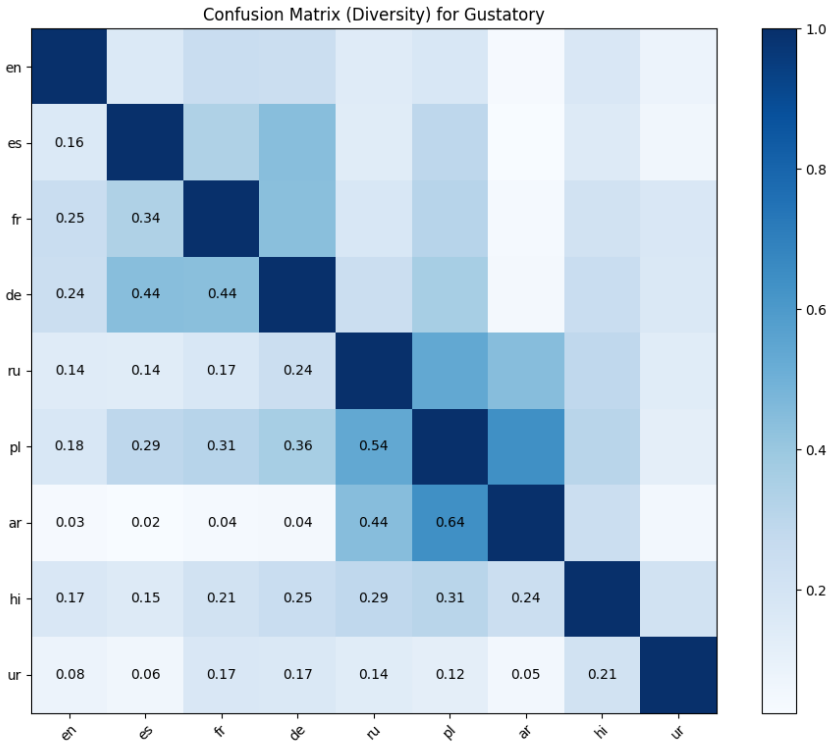


Figure C.7: The cosine similarities of the vectors of contribution of synsets in Gustatory Diversity.

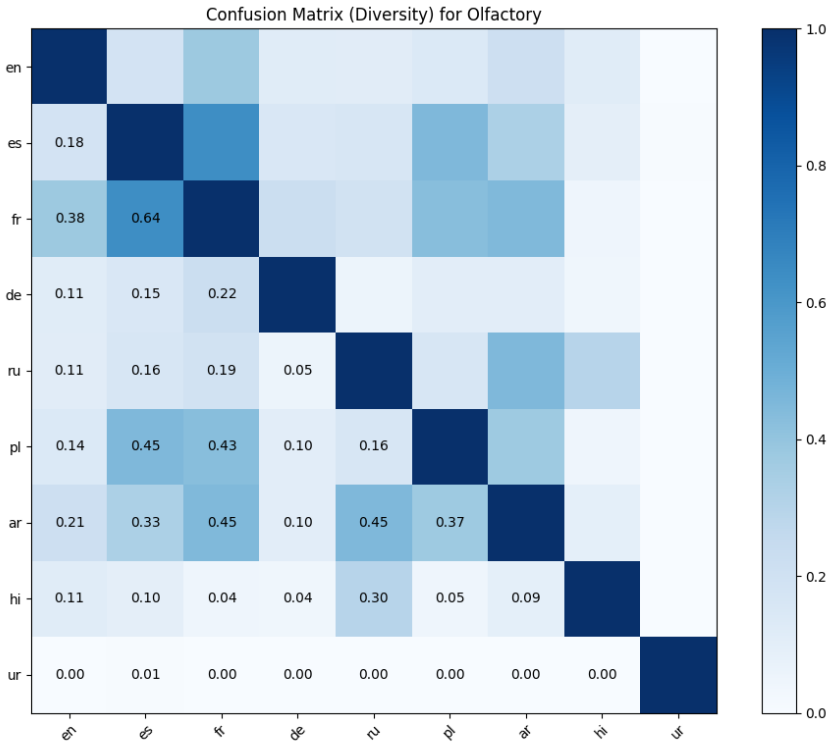


Figure C.8: The cosine similarities of the vectors of contribution of synsets in Olfactory Diversity.

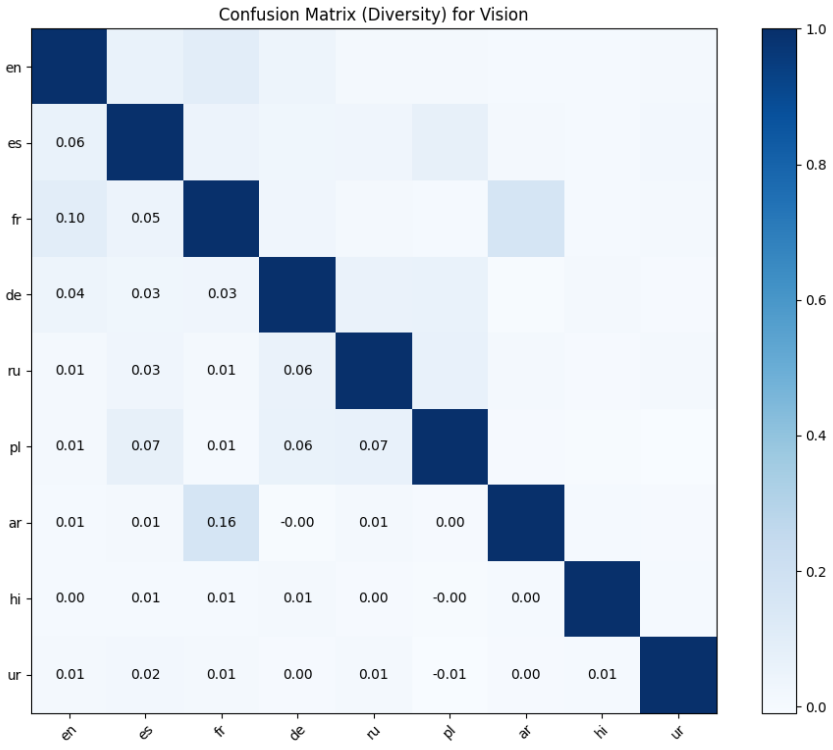


Figure C.9: The cosine similarities of the vectors of contribution of synsets in Visual Diversity.

Appendix D

EXPLICATING STYLE

D.1 Latent Representations of LIWC-Style Across Text Genres

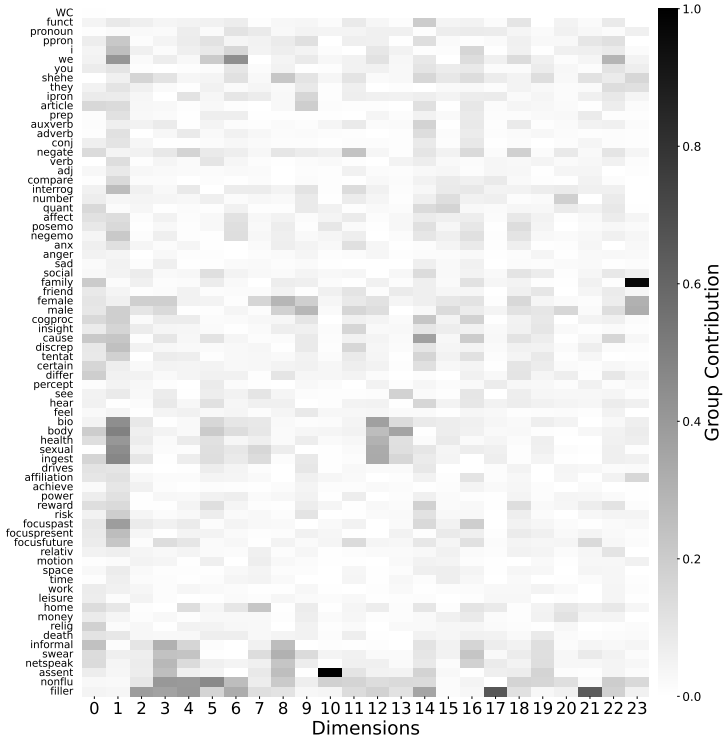


Figure D.1: Heatmap showing the latent representation of LIWC categories across 24 dimensions for Music Lyrics. The intensity indicates the strength of contribution of each LIWC category to each latent dimension.

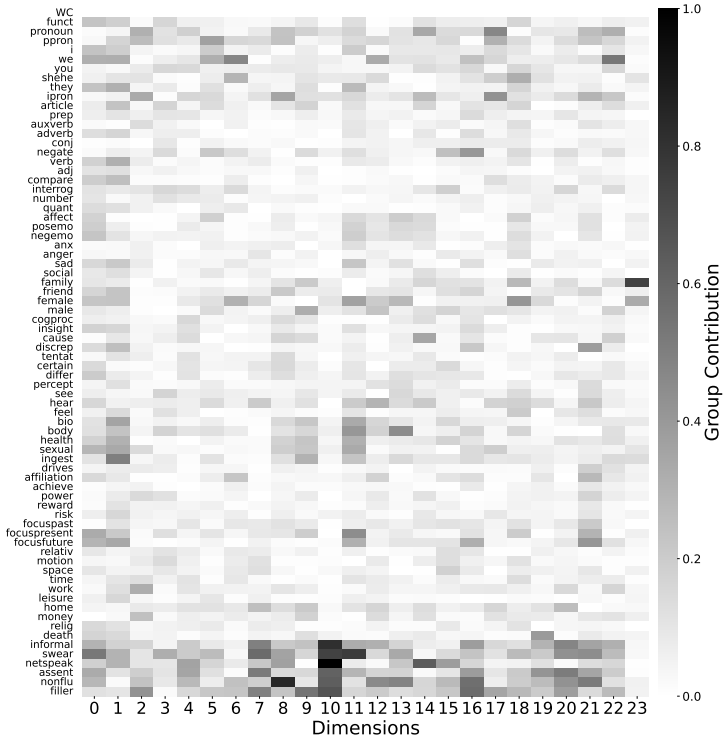


Figure D.2: Heatmap showing the latent representation of LIWC categories across 24 dimensions for Novels. The intensity indicates the strength of contribution of each LIWC category to each latent dimension.

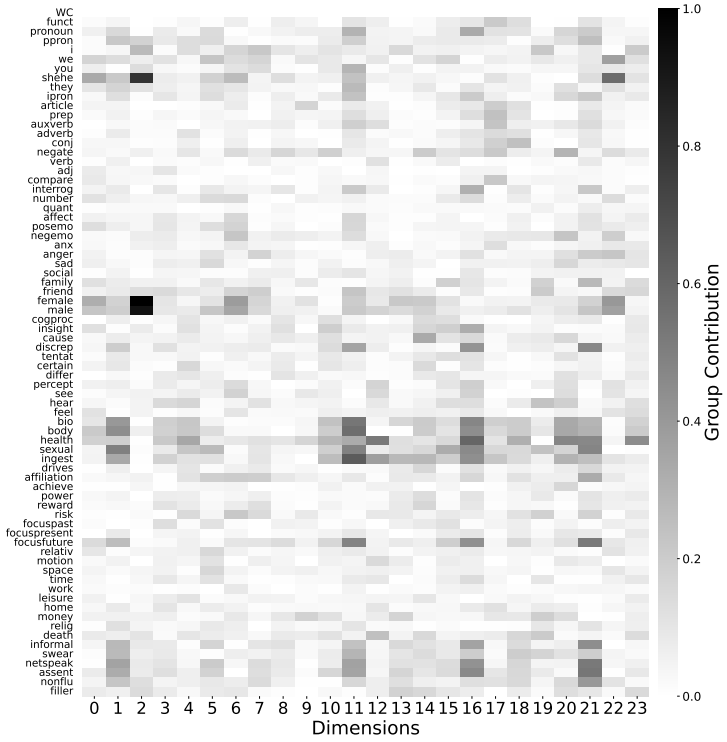


Figure D.3: Heatmap showing the latent representation of LIWC categories across 24 dimensions for Advertisements. The intensity indicates the strength of contribution of each LIWC category to each latent dimension.

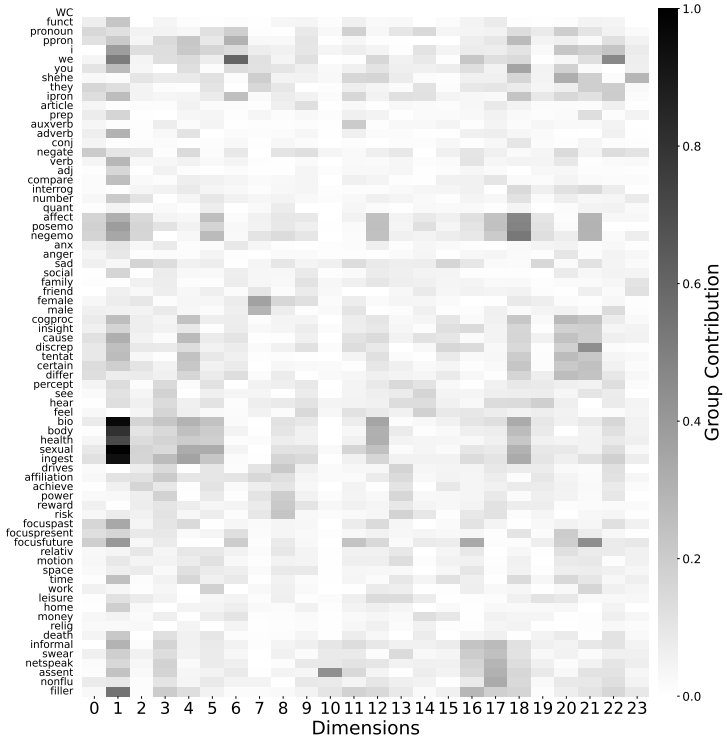


Figure D.4: Heatmap showing the latent representation of LIWC categories across 24 dimensions for Business Reviews. The intensity indicates the strength of contribution of each LIWC category to each latent dimension.

D.2 Performance Comparison of SLIM-LLMs Across Different Text Genres

Table D.1: Performance Comparison of SLIM-LLMs Across Different Text Genres

Model	Config.	Articles			Ads			Novels			Reviews			Lyrics		
		S-80	S-240	Full	S-80	S-240	Full	S-80	S-240	Full	S-80	S-240	Full	S-80	S-240	Full
BERT-base	Full	–	–	.360	–	–	.469	–	–	.378	–	–	.416	–	–	.533
	SLIM	.279	.299	–	.372	.403	–	.332	.352	–	.339	.368	–	.423	.465	–
	+LIWC	.362	**.385**	–	.457	.481	–	.381	**.391**	–	.407	.429	–	.511	.543	–
	+Lat LIWC	.357	.380	–	.462	**.483**	–	.379	.390	–	.409	**.430**	–	.510	**.545**	–
BERT-large	Full	–	–	.373	–	–	.473	–	–	.389	–	–	.420	–	–	.517
	SLIM	.289	.315	–	.380	.418	–	.348	.364	–	.355	.380	–	.434	.464	–
	+LIWC	.384	**.404**	–	.469	**.491**	–	.394	**.406**	–	.424	**.440**	–	.513	.540	–
	+Lat LIWC	.379	.400	–	.473	**.491**	–	.393	.405	–	.424	.439	–	.514	**.542**	–
RoBERTa	Full	–	–	.356	–	–	.499	–	–	.397	–	–	.465	–	–	.565
	SLIM	.242	.281	–	.386	.418	–	.327	.356	–	.367	.401	–	.438	.489	–
	+LIWC	.336	**.365**	–	.472	.501	–	.386	**.405**	–	.440	**.467**	–	.525	.566	–
	+Lat LIWC	.336	.363	–	.478	**.502**	–	.387	.403	–	.441	.466	–	.528	**.567**	–
DistilBERT	Full	–	–	.330	–	–	.454	–	–	.348	–	–	.391	–	–	.523
	SLIM	.237	.267	–	.345	.378	–	.290	.318	–	.305	.332	–	.397	.446	–
	+LIWC	.326	.347	–	.442	**.467**	–	.343	.359	–	.376	**.407**	–	.493	.532	–
	+Lat LIWC	.327	**.350**	–	.441	**.467**	–	.344	**.361**	–	.378	.400	–	.494	**.533**	–
LIWC	–	–	–	.033	–	–	.063	–	–	.083	–	–	.113	–	–	.203

Note: S-80 and S-240 refer to SLIM-LLM models with 80 and 240 dimensions respectively.
The best performing SLIM-LLM configuration for each model and genre is highlighted in bold.

BIBLIOGRAPHY

[1] Dominic Abrams and Michael A Hogg. *Social identifications: A social psychology of intergroup relations and group processes*. Routledge, 2006.

[2] Yong-Yeol Ahn et al. "Flavor network and the principles of food pairing." In: *Scientific reports* 1.1 (2011), p. 196.

[3] Theodore Wilbur Anderson. "Estimating linear restrictions on regression coefficients for multivariate normal distributions." In: *The Annals of Mathematical Statistics* (1951), pp. 327–351.

[4] Anonymous. "SLIM-LLMs: Low-Rank Models of Linguistic Style." In: *Submitted to The Thirteenth International Conference on Learning Representations*. under review. 2024. URL: https://openreview.net/forum?id=SzWvRzyk6h.

[5] Mónica G Ayuso. "" How lucky for you that your tongue can taste the'r'in'Parsley'": trauma theory and the literature of Hispaniola." In: *Afro-Hispanic Review* (2011), pp. 47–62.

[6] Richard W Bailey. "Authorship attribution in a forensic setting." In: *Advances in computer-aided literary and linguistic research* 9 (1979), pp. 87–106.

[7] Lawrence W Barsalou. "Grounded cognition." In: *Annu. Rev. Psychol.* 59.1 (2008), pp. 617–645.

[8] Michael Scott Bernstein et al. "4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community." In: *Fifth International AAAI Conference on Weblogs and Social Media*. 2011.

[9] Margaret M Bradley and Peter J Lang. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Tech. rep. Technical report C-1, the center for research in psychophysiology …, 1999.

[10] Marc Brysbaert et al. "How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age." In: *Frontiers in psychology* 7 (2016), p. 1116.

[11] Emanuel Bubl et al. “Seeing gray when feeling blue? Depression can be measured in the eye of the diseased.” In: *Biological psychiatry* 68.2 (2010), pp. 205–208.

[12] Mary Bucholtz and Kira Hall. “Language and identity.” In: *A companion to linguistic anthropology* 1 (2004), pp. 369–394.

[13] José Cañete et al. “Spanish Pre-Trained BERT Model and Evaluation Data.” In: *PML4DC at ICLR 2020*. 2020.

[14] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. “Author gender identification from text.” In: *Digital Investigation* 8.1 (2011), pp. 78–88.

[15] Yin-Wong Cheung and Kon S Lai. “Lag order and critical values of the augmented Dickey–Fuller test.” In: *Journal of Business & Economic Statistics* 13.3 (1995), pp. 277–280.

[16] Meri Coleman and Ta Lin Liau. “A computer readability formula designed for machine scoring.” In: *Journal of Applied Psychology* 60.2 (1975), p. 283.

[17] Allan M Collins and M Ross Quillian. “Retrieval time from semantic memory.” In: *Journal of verbal learning and verbal behavior* 8.2 (1969), pp. 240–247.

[18] Louise Connell, Dermot Lynott, and Briony Banks. “Interoception: the forgotten modality in perceptual grounding of abstract and concrete concepts.” In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1752 (2018), p. 20170143.

[19] Arthur D Craig. “How do you feel? Interoception: the sense of the physiological condition of the body.” In: *Nature reviews neuroscience* 3.8 (2002), pp. 655–666.

[20] Ilja Croijmans et al. “Measuring multisensory imagery of wine: The vividness of wine imagery questionnaire.” In: *Multisensory research* 32.3 (2019), pp. 179–195.

[21] Walter Daelemans. “Explanation in computational stylometry.” In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. 2013, pp. 451–462.

[22] Munmun De Choudhury et al. “Predicting depression via social media.” In: *Seventh international AAAI conference on weblogs and social media*. 2013.

[23] Stephen De Ullmann. "Romanticism and synaesthesia: A comparative study of sense transfer in Keats and Byron." In: *PMLA* 60.3 (1945), pp. 811–827.

[24] Jacob Devlin. "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv preprint arXiv:1810.04805* (2018).

[25] Song Feng, Ritwik Banerjee, and Yejin Choi. "Syntactic stylometry for deception detection." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics. 2012, pp. 171–175.

[26] Shuang Geng et al. "Understanding the focal points and sentiment of learners in MOOC reviews: A machine learning and SC-LIWC-based approach." In: *British Journal of Educational Technology* 51.5 (2020), pp. 1785–1803.

[27] Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. "Language style matching as a predictor of social dynamics in small groups." In: *Communication Research* 37.1 (2010), pp. 3–19.

[28] Robert Gunning. "The fog index after twenty years." In: *Journal of Business Communication* 6.2 (1969), pp. 3–13.

[29] Trevor Hastie. "Ridge regularization: An essential concept in data science." In: *Technometrics* 62.4 (2020), pp. 426–433.

[30] David Howes. *Sensual relations: Engaging the senses in culture and social theory*. University of Michigan Press, 2010.

[31] Tianran Hu et al. "What the language you tweet says about your occupation." In: *Tenth International AAAI Conference on Web and Social Media*. 2016.

[32] Yuheng Hu, Kartik Talamadupula, and Subbarao Kambhampati. "Dude, srsly?: The surprisingly formal nature of Twitter's language." In: *Seventh International AAAI Conference on Weblogs and Social Media*. 2013.

[33] James M Hughes et al. "Quantitative patterns of stylistic influence in the evolution of literature." In: *Proceedings of the National Academy of Sciences* 109.20 (2012), pp. 7682–7686.

[34] Clayton Hutto and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216–225.

[35] David Kernot, Terry Bossomaier, and Roger Bradbury. "The impact of depression and apathy on sensory language." In: *Open Journal of Modern Linguistics* 7.1 (2016), pp. 8–32.

[36] Osama Khalid and Padmini Srinivasan. "Style matters! Investigating linguistic style in online communities." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 360–369.

[37] Osama Khalid and Padmini Srinivasan. "Smells like Teen Spirit: An Exploration of Sensorial Style in Literary Genres." In: *Proceedings of the 29th International Conference on Computational Linguistics*. 2022, pp. 55–64.

[38] J Peter Kincaid et al. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Tech. rep. Naval Technical Training Command Millington TN Research Branch, 1975.

[39] Olga Koblet and Ross S Purves. "From online texts to Landscape Character Assessment: Collecting and analysing first-person landscape perception computationally." In: *Landscape and Urban Planning* 197 (2020), p. 103757.

[40] Willem JM Levelt. "Accessing words in speech production: Stages, processes and representations." In: *Cognition* 42.1-3 (1992), pp. 1–22.

[41] Francesca Strik Lievers. "Synaesthesia: A corpus-based study of cross-modal directionality." In: *Functions of language* 22.1 (2015), pp. 69–95.

[42] Francesca Strik Lievers and Chu-Ren Huang. "A lexicon of perception for the identification of synaesthetic metaphors in corpora." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 2270–2277. URL: https://aclanthology.org/L16-1360.

[43] Francesca Strik Lievers and Chu-Ren Huang. "A lexicon of perception for the identification of synaesthetic metaphors in corpora." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016, pp. 2270–2277.

[44] Francesca Strik Lievers and Bodo Winter. "Sensory language across lexical categories." In: *Lingua* 204 (2018), pp. 45–61.

[45] Yinhan Liu et al. "Roberta: A robustly optimized bert pretraining approach." In: *arXiv preprint arXiv:1907.11692* (2019).

[46] Andres Lou, Diana Inkpen, and Chris Tanasescu. "Multilabel subject-based classification of poetry." In: *The Twenty-Eighth International Flairs Conference*. 2015.

[47] Dermot Lynott and Louise Connell. "Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form." In: *Behavior research methods* 45.2 (2013), pp. 516–526.

[48] Dermot Lynott et al. "The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words." In: *Behavior Research Methods* 52.3 (2020), pp. 1271–1291.

[49] Asifa Majid and Niclas Burenhult. "Odors are expressible in language, as long as you speak the right language." In: *Cognition* 130.2 (2014), pp. 266–270.

[50] Asifa Majid et al. "Differential coding of perception in the world's languages." In: *Proceedings of the National Academy of Sciences* 115.45 (2018), pp. 11369–11376.

[51] Asifa Majid et al. "Olfactory language and abstraction across cultures." In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1752 (2018), p. 20170139.

[52] Christopher D Manning et al. "Emergent linguistic structure in artificial neural networks trained by self-supervision." In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30046–30054.

[53] Alexandre Matton and Luke de Oliveira. "Emergent Properties of Fine-tuned Language Representation Models." In: *arXiv preprint arXiv:1910.10832* (2019).

[54] Alex Miklashevsky. "Perceptual experience norms for 506 Russian nouns: Modality rating, spatial localization, manipulability, imageability and other variables." In: *Journal of psycholinguistic research* 47.3 (2018), pp. 641–661.

[55] George A Miller. "WordNet: a lexical database for English." In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[56] Jon F Miller. *Assessing language production in children: Experimental procedures*. Vol. 1. University Park Press, 1981.

[57] Ashley Montagu. "Toolmaking, hunting, and the origin of language." In: *The sociogenesis of language and human conduct*. Springer, 1983, pp. 3–14.

[58] Marius Mosbach et al. “A Closer Look at Linguistic Knowledge in Masked Language Models: The Case of Relative Clauses in American English.” In: *arXiv preprint arXiv:2011.00960* (2020).

[59] Claire Murphy. “Olfactory and other sensory impairments in Alzheimer disease.” In: *Nature Reviews Neurology* 15.1 (2019), pp. 11–24.

[60] Roberto Navigli and Simone Paolo Ponzetto. “BabelNet: Building a very large multilingual semantic network.” In: *Proceedings of the 48th annual meeting of the association for computational linguistics*. 2010, pp. 216–225.

[61] Carita Paradis. “Is the notion of linguistic competence relevant in Cognitive Linguistics.” In: *Annual Review of Cognitive Linguistics* (2003), pp. 247–271.

[62] James W Pennebaker. “The secret life of pronouns.” In: *New Scientist* 211.2828 (2011), pp. 42–45.

[63] James W Pennebaker, Martha E Francis, and Roger J Booth. “Linguistic inquiry and word count: LIWC 2001.” In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001.

[64] James W Pennebaker et al. “The development and psychometric properties of LIWC2015.” In: (2015).

[65] Martin Potthast et al. “A stylometric inquiry into hyperpartisan and fake news.” In: *arXiv preprint arXiv:1702.05638* (2017).

[66] Friedemann Pulvermüller. “How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics.” In: *Trends in cognitive sciences* 17.9 (2013), pp. 458–470.

[67] Junyang Qian et al. “Large-scale multivariate sparse regression with applications to UK Biobank.” In: *The annals of applied statistics* 16.3 (2022), p. 1891.

[68] Jamie Reilly, Maurice Flurie, and Jonathan E Peelle. “The English lexicon mirrors functional brain activation for a sensory hierarchy dominated by vision and audition: Point-counterpoint.” In: *Journal of neurolinguistics* 55 (2020), p. 100895.

[69] Matt Reynolds. *The wheels are falling off the alt-right’s version of the internet*. July 2018. URL: https://www.wired.co.uk/article/alt-right-internet-is-a-ghost-town-gab-voat-wrongthink.

[70] Elizabeth Rochon et al. “Quantitative analysis of aphasic sentence production: Further development and new data.” In: *Brain and language* 72.3 (2000), pp. 193–218.

[71] Kamil Safin and Aleksandr Ogaltsov. “Detecting a change of style using text statistics.” In: *Working Notes of CLEF* (2018).

[72] Upendra Sapkota et al. “Cross-topic authorship attribution: Will out-of-topic data help?” In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 1228–1237.

[73] Koustuv Sinha et al. “Unnatural language inference.” In: *arXiv preprint arXiv:2101.00010* (2020).

[74] Mark D Smucker, James Allan, and Ben Carterette. “A comparison of statistical significance tests for information retrieval evaluation.” In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 2007, pp. 623–632.

[75] Richard Sorabji. “Aristotle on demarcating the five senses.” In: *The Philosophical Review* 80.1 (1971), pp. 55–79.

[76] Laura J Speed and Marc Brybaert. “Dutch sensory modality norms.” In: *Behavior research methods* 54.3 (2022), pp. 1306–1318.

[77] Laura J Speed and Asifa Majid. “Grounding language in the neglected senses of touch, taste, and smell.” In: *Cognitive neuropsychology* 37.5-6 (2020), pp. 363–392.

[78] Robyn Speer. *rspeer/wordfreq: v3.0*. Version v3.0.2. Sept. 2022. DOI: 10.5281/zenodo.7199437. URL: https://doi.org/10.5281/zenodo.7199437.

[79] Szabolcs Számadó. “Pre-hunt communication provides context for the evolution of early human language.” In: *Biological Theory* 5 (2010), pp. 366–382.

[80] Yla R Tausczik and James W Pennebaker. “The psychological meaning of words: LIWC and computerized text analysis methods.” In: *Journal of language and social psychology* 29.1 (2010), pp. 24–54.

[81] Hans Van Halteren et al. “New machine learning methods demonstrate the existence of a human stylome.” In: *Journal of Quantitative Linguistics* 12.1 (2005), pp. 65–77.

[82] Alessandra Vergallito, Marco Alessandro Petilli, and Marco Marelli. "Perceptual modality norms for 1,121 Italian words: A comparison with concreteness and imageability scores and an analysis of their impact in word processing tasks." In: *Behavior Research Methods* 52.4 (2020), pp. 1599–1616.

[83] Åke Viberg. "The verbs of perception: A typological study." In: (1983).

[84] Yilin Wang et al. "Understanding and discovering deliberate self-harm content in social media." In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2017, pp. 93–102.

[85] Joel Whitburn. *The Billboard Book of Top 40 Hits*. Billboard Book of Top 40 Hits, 2010.

[86] Bodo Winter. "Taste and smell words form an affectively loaded and emotionally flexible part of the English lexicon." In: *Language, Cognition and Neuroscience* 31.8 (2016), pp. 975–988.

[87] Bodo Winter. *Sensory linguistics: Language, perception and metaphor*. Vol. 20. John Benjamins Publishing Company, 2019.

[88] Bodo Winter, Marcus Perlman, and Asifa Majid. "Vision dominates in perceptual language: English sensory vocabulary is optimized for usage." In: *Cognition* 179 (2018), pp. 213–220.

[89] Bodo Winter, Marcus Perlman, and Asifa Majid. "Vision dominates in perceptual language: English sensory vocabulary is geared towards usage." In: ().

[90] Ming Yuan and Yi Lin. "Model selection and estimation in regression with grouped variables." In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68.1 (2006), pp. 49–67.

[91] Anis Zaman et al. "Detecting Low Self-Esteem in Youths from Web Search Data." In: *The World Wide Web Conference*. ACM. 2019, pp. 2270–2280.

[92] Rolf A Zwaan and Carol J Madden. "Embodied sentence comprehension." In: *Grounding cognition: The role of perception and action in memory, language, and thinking* 22 (2005).