A Computational Account of Selected Patterns of Linguistic Variation and Change

by

Jian Zhu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Linguistics and Scientific Computing)
in the University of Michigan
2022

Doctoral Committee:

Professor Patrice Speeter Beddor, Co-Chair
Assistant Professor David Jurgens, Co-Chair
Associate Professor Steven Abney
Professor Robin Queen
Assistant Teaching Professor Will Styler, University of California San Diego

Jian Zhu

lingjzhu@umich.edu

ORCID iD: 0000-0002-7849-1060

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help of many amazing people throughout my academic adventure at U-M.

Words cannot express my gratitude to my advisor Pam Beddor. To me, Pam is more than an advisor, but also a respected colleague, a good friend and an invaluable guide. She helped me navigate through the grueling Ph.D. journey and the daunting job market. As an advisor, she never backs down from her highest research standard. As a friend, she is caring, emphatic and supportive. And I know she was and is always willing to shield me from the adversaries that I have encountered during my Ph.D. years. Thank you, Pam! You set a role model that I can only hope to emulate.

I could not have undertaken this journey in computational sociolinguistics without the guidance of my co-advisor, David Jurgens. Even though I did not start working with David until my third year, David had exposed me to the intriguing world of computational social science and helped me successfully transition from an experimental linguist to a computational linguist. I really appreciated his kindness, professionalism, and expertise in computational social science.

I am also deeply indebted to Will Styler, who has been a great mentor and a great friend. Will is always willing to listen and share his thoughts, even to some of my wildest ideas. I was grateful that Will had encouraged me to start my first deep learning project.

I am also thankful to my committee members, Steve Abney and Robin Queen for their insightful feedback on this dissertation. Throughout these years at U-M Linguistics, I had benefitted immensely from interactions with Jelena Krivokapić, Andries Coetzee, Acrisio Pires, Marlyse Baptista, San Duanmu, Sam Epstein and Jonathan Brennan. I am particularly indebted to Jelena, who not only introduced me to the world of articulatory phonology but also offered many helpful career advice. Jen Nguyen, Sandie Petee and Talisha Reviere-Winston were excellent in helping me navigate through all the complex administrative processes. Thank you to you all!

Thanks should also go to everyone at PhonDi as well as Pam's and Jelena's Phonetic Group Meetings. Thank you to Kate Sherwood, Dave Ogden, Jiseung Kim, Hayley Heaton, Dominique Bouavichith, Dominique Canning, Steve Tobin, Justin Craft, Fahad Alrashed, Kelly Wright, Wyatt Barnes, Lauretta Cheng and Rachel Weissler. I really enjoyed the inspiring and joyful discussions we had. My gratitude also extends to everyone in the department. You all enriched my time here. In particular, I thank Aliaksei Akimenka, Lucy Chiang, Wilkinson Daniel Wong Gonzales, Yushi

Sugimoto and Tzu-Yun Tung for the time we hung out together and for the fun conversations we had. I would like to extend my sincere thanks to many great people I met at U-M: Minje Choi, Aparna Ananthasubramaniam, Julia Mendelsohn, Allie Lahnala and Warren Li.

I would also like to thank everyone who runs the Joint Scientific Computing Degree Program at The Michigan Institute for Computational Discovery and Engineering (MICDE). This degree program has enabled me to receive extensive training in mathematics, statistics and engineering at the various amazing departments at U-M. Such training empowers me to pursue an interdisciplinary research program that spans both linguistics and computational science.

I had the pleasure of collaborating with my brilliant collaborators, Will Styler, Ian Calloway, Zuoyu Tian, Cong Zhang, Yadong Liu, Chia-wen Lo and Jiaxin Pei. I have been fortunate enough to be able to work on many fascinating projects with you and hope that our ongoing and future collaborations will be fruitful.

Special thanks to people who dedicate themselves to the open-source projects and to the democratization of machine learning, especially contributors of HuggingFace Transformers. Without you, this dissertation could not have been completed. My gratitude and respect also go to Alexandra Elbakyan, who has been working courageously to tear down the (unnecessary) barriers of academia.

I'd like to acknowledge the gaming industry that has brought so much joy to me during the hardest times of my Ph.D. and helped me survive the Pandemic. A big thank you to Nintendo, From Software, SIE, Square Enix and Ubisoft.

In the past four years at Ann Arbor, I have also received great support from my close friends Xianing Zheng and Chen Sun. I'll always remember the times we spent together and the things we have been through.

Lastly, I would be remiss in not mentioning my loving parents, who are always behind me and support me unconditionally. My deepest gratitude goes to my wife, Yifan Hu, and our two lovely cats, Misty and Hippy. It is wonderful to meet you all.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

Language variation and change are ubiquitous, and one aim of linguistic research is to understand synchronic variation and how it contributes to change over time. This dissertation takes a computationally intensive approach to the investigation of language variation and change, with the goals of 1) understanding the complex linguistic landscape in online communities as a result of variation and change; and 2) developing machine learning-based methods to facilitate the processing of large-scale language data in the form of both texts and speech. The current dissertation reports three case studies on selected patterns of variation and change, which span lexical, stylistic, and speech variation in various contexts.

Study 1 centers on the hypothesis that lexical change in online communities is partially shaped by the structure of the community's underlying social network. To investigate the relationship between social networks and lexical change, I conducted a large-scale analysis of over 80k neologisms in 4420 online communities spanning more than a decade. Using Poisson regression and survival analysis, this study uncovers several associations between a community's network structure and lexical change within the community. In addition to overall community size, network properties including dense connections, the lack of local clusters, and more external contacts are shown to promote lexical innovation and retention. Unlike offline communities, these topic-based communities do not experience strong lexical leveling despite increased contact but rather tend to accommodate more niche words. The analysis not only confirms the influence of social networks on lexical change but also uncovers findings specific to online communities.

Study 2 takes a deep learning-based approach to studying individual stylistic variation in written texts. The proposed neural models achieve strong performance on authorship identification for short texts and are therefore used as a proxy to extract representations of idiolectal styles. Extensive analyses were conducted to assess how idiolectal styles were encoded by the data-driven neural model. Using an analogy-based probing task, the study shows that the learned latent spaces exhibit surprising regularities that encode qualitative and quantitative shifts of idiolectal styles. Through text perturbation, I quantify the relative contributions of different linguistic elements to idiolectal variation. Furthermore, I characterize idiolects through measuring inter- and intra-author variation, showing that variation in idiolects is often both distinctive and consistent.

Study 3 moves beyond textual variation and addresses a methodological bottleneck in speech analysis, that is, aligning continuous and highly variable speech signals to discrete phones. Two Wav2Vec2-based models for both text-dependent and text-independent phone-to- audio alignment are proposed. The proposed Wav2Vec2-FS, a semi-supervised model, directly learns phone-to-audio alignment through contrastive learning and a forward sum loss and can be coupled with a pretrained phone recognizer to achieve text-independent alignment. The other model, Wav2Vec2-FC, is a frame classification model trained on forced aligned labels that can perform both forced alignment and text-independent segmentation. Evaluation results suggest that, even when transcriptions are not available, both proposed methods generate results that are very close to those of existing forced alignment tools. A phonetic aligner for Mandarin Chinese with the same method is also reported. This work presents a neural pipeline of fully automated phone-to-audio alignment to facilitate the processing of the highly variable speech data.

This dissertation demonstrates that the abundance of publicly available language data and the advancement of machine learning methods can be effectively harnessed to inform linguistic theories of variation and change.

CHAPTER 1

Introduction

In this dissertation, I aim to investigate language variation across individuals and communities through computationally intensive methods. This research aims to model and understand the physical aspect of human speech and the social dimension of language, focusing on three themes: i) understanding speech and textual variations in human languages through computational modeling; ii) incorporating language variation into speech and NLP technologies; iii) democratizing machine learning technologies for the linguistic community by providing publicly available toolkits for computational analysis.

1.1 Background

Language as a complex dynamic system is constantly undergoing variation and change, which are manifested in all structural, psychological, physiological, physical, and social dimensions. The study of variation within a particular language seeks to not only describe the structural aspects of variation and change but also to understand the causes of variation and change. For example, sociolinguists investigate how variation is related to social factors and used to convey social meaning. Phoneticians study the physical and perceptual underpinnings of variation in sound systems. For both lines of inquiry, variation and change are inherently intertwined: investigating synchronic variation in language structures has the potential to inform diachronic language change. Variation provides the raw material for change or may reflect the ongoing change (Ohala, 1981; Chambers, 2013) but only inferring the process of change from historical records is subject to survival bias as only successful changes are well documented (Nevalainen et al., 2011). Investigating variation in individual languages from an across-language, typological perspective also provides insights into the commonalities and uniqueness of world languages and into the cognitive capacity of language users.

Traditionally, studies of variation and change have been challenged by the daunting problem of data sparsity (Conde-Silvestre, 2012; Britain, 2012). While tracking variation and ongoing changes is possible within a small, focused group of speakers or a small number of resource-rich languages, this approach may not provide the full landscape of variation and change (Dodsworth, 2019). The structuralist linguist Bloomfield once envisioned that “fluctuations in the frequency of forms could be accurately observed if we had a record of every utterance that was made in a speech-community during whatever period we wanted to study” (Bloomfield, 1933, page 394). Although such experiments have generally not been feasible until recently, the recent advent and wide spread of online communities, as well as the collection of large-scale multilingual datasets, have made available detailed records of language use within and across communities of practice (Meyerhoff and Strycharz, 2013).

Language variation and change happen in social contexts. The mainstream sociolinguistics community tends to focus on the social contexts in offline communities within which variation and change unfold. In the past decade, the emergence of the online register with an abundance of data has not only enabled sociolinguists to undertake a more in-depth analysis of variation with respect to classic social factors such as gender, age, personality, geographical areas, and social stratification but also provides emerging social contexts within which more language variation can arise, such as social networks, online communities, and live chats. The latter have not yet been systematically incorporated into current theories of variation and change.

In addition to text, online communities are becoming increasingly rich in multimodal data, especially speech. Variation in human speech—variation due in part to the phonetic context of speech—also poses great challenges to language researchers. In an age of large-scale speech corpora, it is now becoming ever more possible to study speech variation at scale. One of the paramount challenges in analyzing speech variation is the lack of a satisfactory method for extracting a wide range of phonetic variation. Many aspects of sociolinguistic and phonetic variation and ongoing change cannot be analyzed without fine-grained segmentation of the raw speech signals, and achieving that segmentation can be extremely costly and time-consuming. It will be necessary to explore emerging methods to segment, represent and analyze speech variation that encompass as many languages as possible.

In contrast to these linguistic approaches to the study of variation and change, Natural Language Processing (NLP) and Spoken Language Processing (SLP) communities tend to view language as a homogeneous whole, excluding linguistic variation at various levels as random noise (Flek, 2020). The statistically driven paradigm of language processing (Manning and Schutze, 1999), guided by the principle of maximum likelihood, naturally biases NLP models toward the most frequently occurring linguistic patterns in the corpora, and therefore towards the ‘standard language’ or the language used by a particular social group (Garimella et al., 2019; Hovy et al., 2020; Koenecke et al., 2020). Moreover, these computational methods are often developed for resource-rich languages, which might not apply to low resource languages where linguistic data are hard to collect and a written system might not even exist. Sociolinguistic and phonetic findings and theories may serve to improve the existing NLP and SLP tools, either in terms of enhancing model robustness to language variation and change, strengthening model capacity to process low resource languages, or ensuring the fairness of speech and language technology (Nguyen et al., 2016; Flek, 2020).

In an era of big data and machine learning, linguistics and NLP research are still conducted in relatively independent research communities. This dissertation shows that the cross-fertilization between the traditional quest for language variation and change and the data-driven NLP research can be effectively combined to yield more insights into language. As demonstrated through three case studies, the application of speech technology and NLP can benefit the linguistic study of variation and change. First, computational methods can inform the theory of variation and change research by enabling the study of under-explored communities, such as online communities, which have not received wide attention from the linguistics community due to the (perceived) lack of data processing methods. Secondly, computational methods can also facilitate and speed up the investigation of some of the classic research questions by providing state-of-the-art tools of automation. The three studies presented in this dissertation, while centering around very different aspects of variation and change (words, style, and speech), collectively demonstrate the immense benefits brought by the cross-fertilization of language technology and traditional linguistic inquiries. Furthermore, they also serve to emphasize that the development of sophisticated methodologies to track variation and change is no less important than—and is essential to—the theories of language and change themselves.

1.2 The current dissertation

This dissertation delves into both the social and physical aspects of language variation. In the first two studies, I investigate textual variation in online communities within the framework of computational sociolinguistics. Online communities have already become an integral part of modern society. In the past decade, numerous works in NLP and computational linguistics have focused their attention on online registers (Nguyen et al., 2016), yet traditional sociolinguistics still lags behind in formulating theories centering on variation and change in online language use.

- **Study 1: How does language change in online communities?** Language changes rapidly in online communities. In order to understand how the social structures of these communities of practice are related to language change, I investigated the influence of social networks on lexical change in Reddit communities. By using large-scale computational methods and network analysis, the study looks at 80,000 neologisms used by more than 100 million Reddit users across 12 years. The results suggest that the rate of lexical change is differentiated in different subcommunities due to the structure of the underlying networks.

- **Study 2: How does language vary in online communities?** To understand idiolects in online communities, I probe idiolectal variation in two online registers, product reviews and forum posts. In addition to proposing a computational model that achieves high accuracy in authorship verification, I use this method to quantify the contributions of lexical, syntactic, or discourse elements to individual variation and the degree to which individual styles are distinctive and consistent against a very large background population. As idiolect is less well studied, my work provides empirical evidence for the characterization of idiolects.

In Study 3, I move beyond variation within texts and focus on speech variation across two languages, English and Mandarin Chinese. In this line of research, I seek to understand speech variation and develop publicly available tools to facilitate data acquisition and processing for the speech community studying language variation and change. One of the goals is to democratize machine learning for the linguistics community by developing open-source speech and language technologies in the hands of linguists.

- **Study 3: How can speech analysis be facilitated by deep learning methods?** Aligning phones to speech signals has remained a bottleneck that limits the preprocessing of speech data from (socio)phonetic studies, which is due to the difficulty of segmenting speech. In order to facilitate the speech annotation pipeline, I present two neural network models that can both perform forced alignment and text-independent alignment, which empowers linguists with tools that speed up the otherwise time-consuming annotation process.

These topics cover various aspects of language variation, ranging from how the dynamic variation of words is shaped by community network structures (*Study 1*), to how variation between individual language users can be quantified and analyzed computationally (*Study 2*), and to how phonetic variation in two languages can be annotated and analyzed (*Study 3*). The dissertation demonstrates how the cross-fertilization between traditional sociolinguistics and phonetics on the one hand and computational linguistics on the other can expand the methodological scope of language research, which will have implications for future quantitative studies of language variation and change. This work addresses a set of specific research questions concerning variation and change for Study 1-3, and these research questions are given in Chapters 2, 3 and 4 respectively.

1.3 Broader implications

In addition to addressing the specific research questions, I anticipate that the current dissertation will have broader implications for the field of computational linguistics in general.

- **An understanding of language variation and change in the online register**. While traditional sociolinguistic studies primarily focus on offline communities, this dissertation draws attention to language use in online platforms, which have already permeated people's daily lives. The current studies will illustrate how language adapts to and is shaped by online communities.

- **Novel computational methods to quantify language variation and change**. This dissertation also proposes a set of new computational methods for analyzing and labeling a massive amount of textual and speech data, thereby providing methodological tools and insights for subsequent studies and facilitating the processing of massive data sets.

CHAPTER 2

Study 1: Social Network and Lexical Change

2.1 Introduction[1]

Lexical change is a prevalent process, as new words are added, thrive, and decline in day-to-day usage. While there is a certain randomness at play in word creation and adoption (Newberry et al., 2017), there are also psychological, social, linguistic, and evolutionary factors that systematically affect lexical change (Christiansen and Kirby, 2003; Labov, 2007; Lupyan and Dale, 2010). In sociolinguistics, one structural factor that has long been recognized as influencing lexical and other types of change is the language community's social network.

Mathematically, a network, or a graph, is defined as the representation of a set of points and the connections between points (West et al., 2001). As a specific subcategory of a network, a social network is the aggregation of relationships between individuals (Milroy and Llamas, 2013). In social network analysis, individuals are usually treated as a node in a network whereas ties between people can be represented as edges connecting between nodes which, taken together, provide a characterization of the interactional dynamics between people. Using social networks as a means to capture social interactions, sociolinguists have found that structures of networks are related to synchronic language variation and diachronic change (e.g., Milroy and Milroy, 1985; Nevalainen, 2000; Tieken-Boon van Ostade, 2000; Marshall, 2004; Fitzmaurice, 2007; Sairio, 2009).

The current study extends network-based sociolinguistic research to the study of lexical change in online communities (specifically, Reddit), which remain understudied despite their expansion in past decades. While I make comparisons between offline and online communities, my data are exclusively from online communities. In particular, my focus is on online communities of practice. Communities of practice (Eckert and McConnell-Ginet, 1992; Holmes and Meyerhoff, 1999; Schwen and Hara, 2003) have been defined as "*an aggregate of people who come together*

[1] Portions of this chapter have appeared in *The structure of online social networks modulates the rate of lexical change* (Zhu and Jurgens, 2021b) in the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

around mutual engagement in an endeavor" (Eckert and McConnell-Ginet, 1992, page 464), rather than offline speech communities. In a community of practice,

> "*[w]ays of doing things, ways of talking, beliefs, values, power relations – in short, practices – emerge in the course of this mutual endeavor. As a social construct, a CofP is different from the traditional community, primarily because it is defined simultaneously by its membership and by the practice in which that membership engages*" (Eckert and McConnell-Ginet, 1992, page 464).

Members of the same community of practice engage in activities that signify their identities and the extent to which they belong, including linguistic activities (Holmes and Meyerhoff, 1999). The presence of online communities (e.g., Reddit, Twitter) as communities of practices provides a playground for studying sociolinguistic variation within a subset of a large speech community (mainly the English-speaking community). The lexical changes of interest are the emergence and propagation of neologisms in online communities. While these changes might not propagate across an entire language community (e.g., English communities), the changes are "locally meaningful" (Eckert, 2019) in communities of practices, specifically in Reddit sub-communities, which are relatively independent of each other (Singer et al., 2014). Within sociolinguistics, Zimman and Hayworth (2020), for example, have treated the online trans community as a community of practice because of some shared practices among users. Studies in computational social science have shown that different online communities also exhibit community-specific collective behaviors in language use (Danescu-Niculescu-Mizil et al., 2013; Zhang et al., 2018), providing further empirical foundations to treat – as I do in this study – each subreddit community as relatively independent and to compare across these sub-communities.

In this study, I examine how network structures affect lexical **innovation**, **retention** and **levelling** in online communities. Specifically, I ask 1) how network structure contributes to the introduction of new words to online communities (innovation), 2) how structural properties affect the survival of these newly introduced words (retention) and 3) whether the increased inter-connectedness causes online communities to adopt a similar set of new words (levelling). This work offers the following contributions. First, using a massive longitudinal dataset of 4420 communities, I precisely quantify the structural mechanisms that drive these lexical processes. This work adds to network studies in sociolinguistics focusing on in-person observations of local communities (Conde-Silvestre, 2012; Sharma and Dodsworth, 2020) and shows that conclusions drawn from offline communities are insufficient to account for behavior seen in online social networks. And these topic-based online communities also do not experience strong levelling due to increased contact. Second, emerging studies in online communities (Danescu-Niculescu-Mizil et al., 2013; Stewart and Eisenstein, 2018; Del Tredici and Fernández, 2018) focus exclusively on lexical change at the individual or word level. A new contribution of this study is its investigation of how global network properties affect lexical change at the community level. Finally, although sampling offline networks presents practical difficulties, I extract complete networks for thousands of online communities, providing a large-scale dataset to explore the structural factors of lexical change.

In the remainder of this chapter, I begin with a discussion of the literature on language change and social networks in Section 2.2. The details of constructing a corpus of Reddit communities and neologisms are described in Section 2.3. Sections 2.4 and 2.5 are about the numerical measurements and the empirical validation of subreddit networks. Investigations of lexical innovation, survival and levelling are provided in Sections 2.6, 2.7 and 2.8 respectively. Finally, the results are discussed in Section 2.9. The code is available at https://github.com/lingjzhu/reddit_network.

2.2 Background

2.2.1 Language change and social networks

How new words emerge and change are long-standing questions that have received extensive attention, especially in the literature of corpus linguistics and lexicography. While there are many social factors contributing to lexical change, there remains relatively little research on these factors, especially from the perspective of social networks. The broader literature on language change, though, has a distinguished history of investigating the connection between social networks and change.

Since the landmark study of sound change in the Belfast community by Milroy and Milroy (1985), the impact of network structures on change has been a key consideration in sociolinguistics. As an initial illustration of the role of social networks in variation and change, a classic finding of Milroy and Milroy (1985) is that loose-knit networks with mostly weak ties are more conducive to information diffusion, thereby facilitating innovation and change, while close-knit networks with strong bonds impose norm-enforcing pressure on language usage, strengthening the localized linguistic norms. The reason is that close-knit social networks are present in isolated, small communities where communications with the external communities were few, thereby limiting the introduction of linguistic innovations (norm-enforcement mechanism). Drawing on pioneering works on social networks (Granovetter, 1977, 1983), Milroy and Milroy (1985) proposed the *weak tie model of change*, which holds that the structural properties of social networks can account for the general tendency of some language communities to be more resistant to linguistic change than others (Milroy and Milroy, 1985, 1992; Milroy and Llamas, 2013).

One compelling observation in favor of this argument concerns the comparison between two Germanic languages, Icelandic and English. Icelandic has changed little since the late thirteenth century, which could be due to the norm-enforcing pressure inherent in the strong kinship and friendship ties. In contrast, in Early Modern London English, the loosening of network ties, accompanied by the rise of the mobile merchant class, was argued to be responsible for some radical change in the language (Milroy and Milroy, 1985).

In general, weak ties in social networks usually mean more social mobility and exposure to more diverse linguistic variation, which facilitates the introduction of linguistic variants (Milroy and Milroy, 1985, 1992; Milroy, 2002; Milroy and Llamas, 2013). For example, using historical data, Nevalainen (2000) compared the replacement of the second-person pronoun "ye" by the oblique form "you" throughout several centuries in three dialect areas, London, the North, and East Anglia. The morphological change diffused much faster in the London area than in the other two areas, mostly because the loose network ties between people in big cities are more conducive to the diffusion of variants, a finding consistent with the weak-tie hypothesis.

Another finding of the influence of social networks on ongoing variation and change is that people with similar connections tend to speak similarly (Dodsworth and Benton, 2017; Dodsworth, 2019). Dodsworth (2019) constructed relatively large bipartite networks (189 people) across three generations between individuals and schools in the Southern U.S. city of Raleigh, North Carolina. They found that close network distances between two individuals as measured by Jaccard distance generally imply less acoustic distance, though the network variables interacted a lot with other social variables such as gender and local identity. For example, the adoption of the regional vowel variants is not determined by network distances alone as, independent of network positions, blue-collar women and white-collar women in Raleigh adopted different variants due to contrasting symbolic needs. It is not uncommon for social networks to interact with gender and cultural groups differentially (Wei, 1994; Dubois and Horvath, 1998; Lanza and Svendsen, 2007; Sharma, 2017). For example, speakers in British Asian communities, depending on their connections to different cultural groups tend to adopt different variants in speech (post-alveolar /t/, an indicator of Asianness) and have different language choices (in multilingual communities) (Sharma, 2017).

These examples of findings of the influences of social network ties on sound change and lexical change have been reinforced by numerous other studies, and several general findings have emerged that are especially relevant to the current investigation. First, the general finding is that network homophily (McPherson et al., 2001) holds for language use in that speakers connected by the same network tend to adopt similar linguistic variants (e.g., Sharma, 2017; Dodsworth, 2019). Secondly, the weak-tie theory of change (Milroy and Milroy, 1985, 1992) has generally been confirmed by subsequent studies, showing that loose-knit networks are conducive to linguistic diffusion whereas close-knit networks are norm-enforcing (e.g., Milroy and Milroy, 1985; Nevalainen, 2000). However, social networks generally interact with other social variables such as gender and cultural groups (Lanza and Svendsen, 2007; Sharma, 2017), such that the influence of social networks is not deterministic nor mechanistic (Eckert, 2019; Sharma and Dodsworth, 2020), but is filtered through the specific social and ideological contexts.

Although traditional sociolinguistic studies have delineated many aspects of the role of networks in change, methodologically network studies in sociolinguistics have much to gain from computational network analysis. Due to the difficulty of collecting widespread network data, most social network studies within the sociolinguistic community focus predominantly on speakers in local, less mobile communities where ties between people tend to be strong (Milroy and Milroy, 1985; Dodsworth and Benton, 2017; Dodsworth, 2019; Sharma, 2017; Conde-Silvestre, 2012; Sharma and Dodsworth, 2020). These networks are typically constructed through qualitative interviews or questionnaires but can be very limited in scope and population coverage, which can introduce sampling biases in the data. Another dimension for improvement is that most sociolinguistic studies consider a limited set of network variables, such as size, tie strength, and density. These sociolinguistic inquiries could benefit from the development of modern graph theory and network mining techniques developed in computational science.

Many sociolinguistic studies have also tended to emphasize the role of the individual, and individual-to-individual connections, in language change. Such a micro-scale view is valuable for revealing how individual behaviors ultimately result in collective change. However, the study of how language patterns within a community change under the influence of structural factors can complement studies on individuals and shed light on language variation and change at another scale. Except for a few recent simulation studies (Reali et al., 2018, see Section 2.2.2), researchers have rarely explored how the global properties of social networks systematically affect lexical change, although the weak tie model does predict the influence of social networks at the macro-level. In addition, while some lexicographic studies attempt to enumerate factors that affect the acceptance of neologisms (Metcalf, 2004; Barnhart, 2007), how network structures affect acceptance is rarely taken into consideration. A key limitation of previous studies has been accessing a large longitudinal dataset of communities with different network properties as well as a precise estimate of the network structure of larger communities, which are limitations this study overcomes.

2.2.2 Language change at the community level

Individual communities have their own structure and interconnectedness and multiple studies have investigated how these properties could potentially affect lexical change. It has been suggested that the rate of language change might not be constant across different language communities, as smaller communities might experience a higher rate of change (Nettle, 1999a; Bowern, 2010; Bromham et al., 2015) and changes might happen in bursts rather than continuously (Atkinson et al., 2008).

Several relevant analyses have been conducted on real-world language communities. Bromham et al. (2015) examined the gain and loss of cognate words in multiple Polynesian languages, which were well documented in the comparative linguistic database. After controlling for external factors like age, geographical area, and language relationship, they showed through Poisson regression that communities with a larger population tend to gain more new (cognate) words while communities with a smaller population lose more (cognate) words. These findings, as suggested by Bromham et al. (2015), were consistent with the predictions of evolutionary theories, or specifically population genetic theory, which predicted that

> larger populations should have higher rates of adaptation because there are more individuals to generate novelty and fewer disruptive influences of random sampling on the process of fixation of new variants. (page 2100)

However, follow-up work has suggested that the effect of population size is not universal. Given that previous studies only focused on languages from small-scale communities (e.g., Polynesian languages), Greenhill et al. (2018) further tested the population hypothesis on three families of languages, namely, Indo-European, Austronesian, and Bantu Languages, totaling 153 pairs of closely related languages. For Indo-European languages, Poisson regression indicated that small language communities did experience significantly more word loss than large communities. However, this finding did not hold for Austronesian or Bantu languages.

Given the empirical challenges of testing community-level variation and change, some researchers have turned to artificial environments to investigate this question (Fagyal et al., 2010; Nettle, 1999b; Blythe and Croft, 2009; Baxter et al., 2009; Reali et al., 2018; Lev-Ari, 2018). Nettle (1999b) simulated language change within a social network paradigm in which artificial learners randomly picked up linguistic variants in the ambient environment but variants were weighted differently depending on the speakers' social position in a network. Results suggested that some speakers have greater social influence in propagating linguistic variants, though functional biases also played a role in language adoption.

Other studies of simulated communities have investigated the influence of population size on the propagation of change. Nettle (1999a) argued that simulation results indicated that language changes were more likely in small-scale communities. Reali et al. (2018) suggested that population size but not other network parameters was one of the most important factors accounting for the growth of vocabulary in a community. In their simulations, agents could replicate linguistic conventions from other agents but those conventions could either be easy or hard to learn, depending on the number of encounters. The environments in which the agents interact were assigned to various network structures to control for the interactions between agents. Results showed that when the network size is large, only easy to learn linguistic structures (e.g., words) tend to propagate widely. But network structures did not seem to interact with sizes. These results lead Reali et al. (2018) to suggest that these results might explain the real-world observation that languages spoken by large populations tend to have a larger vocabulary but simpler grammatical structure.

Another approach is to test these community-level hypotheses about change on real humans who interact in a controlled setting. Using an artificial learning paradigm, Raviv et al. (2019) tested the role of population size in shaping languages in the laboratory. Participants were partitioned into large (8 people) or small (4 people) groups in the laboratory and were instructed to create an artificial language with limited symbols. In the experiments, in-group participants interacted with each other by describing the visual scenes with the artificial languages they create and these communication games were repeated in multiple interactions. The post-hoc analysis shows that large groups created more systematic languages whereas small groups tended to exhibit more linguistic variation. Raviv et al. (2019) argued that large groups faced more communication challenges as participants were more heterogeneous and had less shared history. It was this communication pressure that shaped more systematic linguistic structures in the lab. This experimental study exemplifies an appealing approach to investigating how social factors can shape language formation. Yet Raviv et al. (2020), using the same experimental paradigm, further found that social network structures did not contribute to the formation of linguistic structures (but see Lev-Ari, 2018, for a different outcome with a similar paradigm). However, the scale of the experiments was limited by the costs of working with humans. Only small-scale groups were investigated, so it was uncertain whether the results would still hold if the community size continues to expand.

The general findings of these studies suggest that community structures, especially population size, do play a role in shaping language, specifically lexical, variation and change. However, the role of social networks in these studies remains unclear. Studies on real language communities did not consider social networks (Atkinson et al., 2008; Bromham et al., 2015; Greenhill et al., 2018), as such data were impossible to collect. Simulation studies report conflicting results regarding the role of networks (Nettle, 1999b; Reali et al., 2018), which might be due to the different parameter settings of the artificial environments. Moreover, simulations might deviate significantly from the statistical properties of actual populations, as simulated environments are usually a much-simplified version of the real world. Due to limitations of empirical data collection, though, structural factors other than population size are rarely investigated. To date, these findings are rarely tested on large-scale real networks. This chapter addresses some of the unanswered questions in the literature, as the existence of Reddit provides a detailed record of speaker interactions and the underlying social structures. With large-scale data in online communities, it is possible to revisit community-level variation and change with more detailed measurements of social structures and more ecological validity.

2.2.3 Lexical variation and change in online communities

While most studies on variation and change in traditional sociolinguistics focus on offline communities, the rise of social media and the proliferation of Internet speech has drawn increasing attention to the lexical change in online communities, including Twitter (Eisenstein et al., 2014; Goel et al., 2016; Würschinger, 2021), Reddit (Altmann et al., 2011; Stewart and Eisenstein, 2018; Del Tredici and Fernández, 2018) and review sites (Danescu-Niculescu-Mizil et al., 2013).

This line of investigation is usually pursued in the corpus linguistics and computational linguistics literature. Corpus linguists have made many efforts to track the dynamics of online neologisms through the construction and the analysis of large-scale web-based corpora (Renouf, 2007; Renouf et al., 2007; Kerremans et al., 2012; Gérard, 2017; Cartier, 2017). However, in corpus linguistics, the research question centers on describing the emergence and the decline of neologisms as they are, rather than probing into the possible social factors that induce these lexical changes. For example, Renouf (2007) provided case studies on the life-cycles of several media and Internet neologisms and the lexical productivity of neologisms. She found that most neologisms tend to follow certain patterns of life cycles, starting from

> birth or re-birth, followed by gentle or steeper upward trajectories in the frequency of use and leading to brief or lengthier moments at the zenith of popularity, after which they take faster or slower downward paths until they reach a stable level of use. (page 23)

While these studies provide detailed descriptions of neologisms, they fall short of explaining the potential causes of word change.

For offline communities, the role of word frequency in word survival and spread has clearly emerged in the literature (e.g., Bybee, 2010; Pagel et al., 2007). For example, for Indo-European languages over the centuries, word frequency has exerted a differentiated effect on word change, as high-frequency words tend to evolve slowly but low-frequency words change more rapidly (Pagel et al., 2007). Turning to online communities, though, Altmann et al. (2011, page 1) analyzed longitudinal data from online chatrooms and showed that the word niche – "the range of individuals using the word and the range of topics it is used to discuss" – played a more important role than lexical frequency, at least for short-term changes in online environments.

Accumulating evidence increasingly suggests that language use in online communities is nonetheless constrained and influenced by many social factors. In addition to language-internal factors such as frequency and context, language-external social factors, notably community norms, social networks, and topics, have been linked to language variation and change in online communities. For example, it has been shown that the usage of certain words is associated with community loyalty and norms (Zhang et al., 2017; Bhandari and Armstrong, 2019) and is indicative of user behaviors (Danescu-Niculescu-Mizil et al., 2013; Noble and Fernández, 2015; Chang and Danescu-Niculescu-Mizil, 2019; Klein et al., 2019). For example, Zimman and Hayworth (2020) investigate the short-scale change of gender-related lexical items in online transgender communities, which reflects the changing sociopolitical attitudes towards transgender communities.

Stewart and Eisenstein (2018) investigate the survival of lexical items in Reddit. Their study looked at two factors, linguistic dissemination, or the diversity of linguistic contexts in which a word is used, and social dissemination, a measure of word niche similar to Altmann et al. (2011). After rigorous statistical analysis, they concluded that a word's occurrence in more diverse linguistic contexts was the strongest predictor of its survival while social dissemination was a comparatively weaker predictor. Yet social factors are still consistently found to affect lexical variation and change in the virtual environment. Royalty to community identities is reported to affect the adoption of word variants on review sites. In two targeted beer review sites under investigation, new users usually actively adopted new words to engage with the community identity, yet they became more conservative in word adoption as they were gradually leaving the community (Danescu-Niculescu-Mizil et al., 2013). Such sociolinguistic variation also manifested itself at the community level. For example, in the large online community Reddit, each subreddit community norm resulted in a preference for different word usage, though sometimes the lexical variation was only short-term changes (Zhang et al., 2017; Bhandari and Armstrong, 2019).

Specifically for social networks, studies along this line also tend to focus on the role of individual language users within a social network (Paolillo, 1999; Paradowski and Jonak, 2012). Paolillo (1999) was one of the earliest studies on the sociolinguistics of online networks. Through analyzing the log files of online chat rooms, Paolillo reported that a user's position in a social network has a structured relationship to their use of linguistic variants, such that different linguistic variables were localized in different regions of a social network. Paradowski and Jonak (2012) examined the exposure threshold for which users adopted a hashtag based on the adoption of users in their networks and found that, interestingly, few exposures were needed for users to adopt a given hashtag. Another study close to this current study is Kershaw et al. (2016), which investigates word innovations in two social media platforms by looking at a variety of grammatical and topical factors. Taking longitudinal text samples from both Reddit and Twitter, they measured longitudinal variation in frequency, word form, and word meaning that naturally emerged in online interactions. The findings show that lexical change in these two social media platforms exhibited slightly different patterns: lexical variation in Twitter seemed to be more geographically bounded whereas lexical variation in Reddit tended to be topically bounded. However, Kershaw et al. (2016) only used network information to partition the dataset without exploring the role of these structural attributes in depth. Del Tredici and Fernández (2018) examined the use of neologisms in 20 subreddit communities.To test the weak tie theory of language change, they quantified the tie strength by computing the overlap of adjacent neighborhoods between user pairs, based on which users were divided into weak-tie users and strong-tie users. Weak-tie users were found to be more likely to be innovators, that is, to introduce a new word to the community than strong-tie users. While not being in the majority, strong-tie users played a more important role in innovation spread, as sustained adoption of neologisms by strong-tie users predicted innovation spread. The finding that weak-tie users tend to innovate whereas strong-tie users tend to propagate is consistent with the *weak tie theory of language change*. Despite prior studies, less is known about how network structures are systematically related to the community-level lexical change in online communities, which I address here.

2.3 The Reddit Network Corpus

To analyze lexical innovation in a network setting across long time scales, I use comments made to Reddit, one of the most popular social media sites. There, 330M users are active in about 1M distinct topic-based sub-communities (subreddits). Here I define each subreddit as a community of practice (Schwen and Hara, 2003), as each subreddit is relatively independent with various norms formed through interactions. The subreddit communities span a wide range of social network structures (Hamilton et al., 2017) and linguistic use patterns (Zhang et al., 2017), making them ideal for studying the propagation of sociolinguistic variations in online communities.

2.3.1 Neologisms

Neologisms are newly emerging language norms that fall along a continuum from the common words known to the overwhelming majority of users to nonce words that are mostly meaningless and rarely adopted. I only focus on Internet neologisms, e.g. *lol, lmao, idk*, as community slangs in Reddit communities. Such neologisms are abundant in ever-evolving online communications as people use them for convenience or to signify in-group identity. The non-standard, idiosyncratic spelling patterns of Internet neologisms also make them easier to track than nuanced meaning shifts.

I obtained Internet slangs from two online dictionary sources, `NoSlang.com` and `Urban Dictionary`. The neologisms in `NoSlang.com` have been used in a previous study (Del Tredici and Fernández, 2018). After filtering some lexical entries, I ended up with approximately 80K Internet neologisms for subsequent analysis. I set the minimum frequency threshold of neologisms to 10 over the entire dataset; this low setting ensures that the analysis is not biased by selectively looking only at surviving words, which may obscure the lexical change process.

Many of these neologisms were not first coined on Reddit but were coined elsewhere and sub-

Frequency	Neologisms
Most frequent	lol, /r, kinda, bitcoin, idk, lmao, tbh
	tl;dr, alot, /s, omg, lvl, hahaha, iirc
Least frequent	thugmonster, blein, sotk, f'tang
	yobbish, ferranti, sonse, yampy

Table 2.1: Examples of neologisms.

sequently introduced into subreddits by users. Since it was neither feasible nor possible to trace the exact origins of these words, I instead focused on how words were introduced and adopted. This approach is also consistent with previous studies of lexical change (Altmann et al., 2011; Grieve et al., 2017; Del Tredici and Fernández, 2018).

Table 2.1 shows some samples of the most frequent and least frequent neologisms in Reddit. These linguistic innovations were collected from `NoSlang.com` and `Urban Dictionary`. I filtered out lexical entries that: 1) span more than one word, 2) can be found as an entry in an English dictionary after lemmatization, 3) are identified as person names, 4) contain non-alphabetical characters, numbers, or emojis and 5) do not show up in the Reddit dataset.

I set loose criteria for word inclusion. Many of the frequent neologisms have already been incorporated into the daily lexicon, such as *wiki*, *google* and *instagram*. I manually filtered out these words in the wordlist and the number of such words is less than 100. I also keep typos in the curated list, as these words often carry special meanings. For example, *alot*, *atleast* and *recieve* are the typos that are used more than 1 million times, so frequent that they carry some special meanings and functions such as identity assertion.

One caveat is that neologisms with different spelling variants, *tl;dr, tldr*, could potentially be treated as two separate entries or could have only one of the all variants captured. While it is possible to correct for some most common examples, correcting for all such lexical entries is infeasible on such a scale. Some of the neologisms were originally typos (e.g., *recieve*) but they became so widespread that some users began to use them intentionally. I decided to leave these entries as they are in the dataset. Initial analysis in Figure 2.1 indicates that the frequency distribution of the neologisms conforms almost perfectly to a power-law fit, $p(x) \propto x^{-\alpha}$, with empirical $\alpha = 2.065$.

After automatic filtering, I manually inspected the 5000 most frequent words with greater care so as to filter out some invalid entries. In addition, I also sampled a few hundred words at different frequency bins for close inspection. For the rest of the words, I only scanned them for a quick sanity check. As an additional validation, I replicated some of the results on full list of 170k words, which were created by setting the frequency threshold to 2. I obtained similar main results. This result is not included in the main text because I consider that analyzing neologisms with such a low-frequency threshold may include too many unreliable lexical tokens.

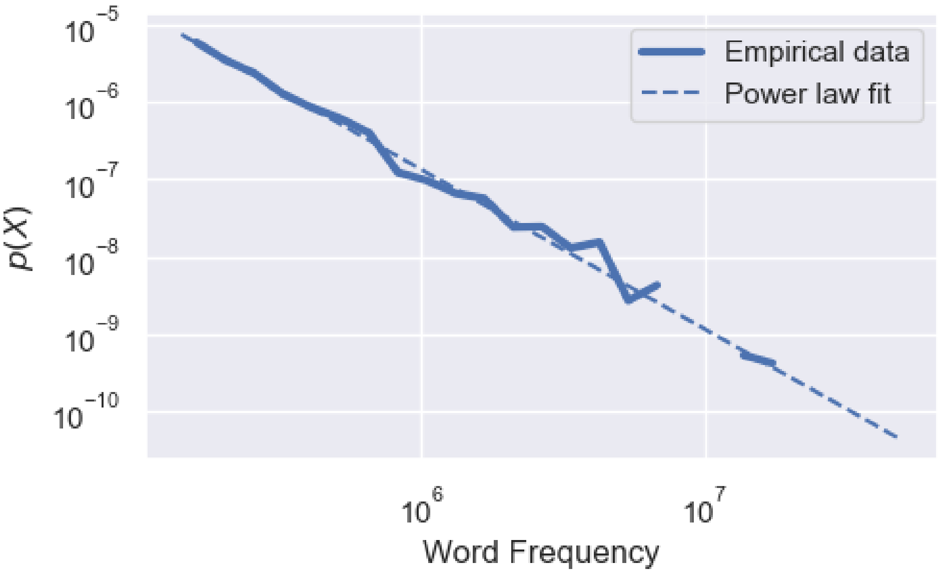


Figure 2.1: Distribution of word frequency.

2.3.2 Network construction

To strike a balance between acquiring active subreddits and preserving the diversity of these communities, I initially select the top 4.5K subreddits based on their overall size from their inception to October 2018 via the `Convokit` package (Chang et al., 2020b). Let $\mathcal{C}_{Reddit} = \{C_1, C_2, \ldots, C_n\}$ be the set of subreddit communities included in the corpus. A subreddit community C_n is further discretized into multiple monthly subreddit communities $c_n(t)$ based on its actual life span in the monthly time step t, such that $C_n = \{c_n(1), c_n(2), \ldots, c_n(t_{max})\}$. For each $c_n(t)$, I extracted all individual comments except those marked as `[deleted]` and performed tokenization via SpaCy. During text cleaning, I removed numbers, emojis, urls, punctuations and stop words, and set a cutoff frequency of 10 over the entire dataset to exclude infrequent typos or misspellings. Only those monthly subreddits $c_n(t)$ with more than 500 words or 50 users after preprocessing are retained. Some communities known for their content in foreign languages are also removed. After preprocessing, 4420 subreddits were left in the analysis.

2.3.2.1 Community networks

For a community c from month $t = 1, 2, \ldots, t_{max}$, its temporal network can be represented as a discrete-time sequence of network snapshots $\mathcal{G}_c = \{G_c(1), G_c(2), \ldots, G_c(t_{max})\}$. Each snapshot network at time t, $G_c(t) = \{V_c(t), E_c(t)\}$ consists of a set of user nodes $V_c(t)$ and a set of edges $E_c(t)$ characterizing direct interactions between users. $G_c(t)$ is initiated as an undirected and unweighted graph under the assumption that these commenting communications are mutual and bi-directional.

A user u_i is represented as a node if this user has posted at least one comment at month t. An edge e_{ij} exists between user u_i and user u_j if these two users have interacted in close proximity in a common discussion thread, that is, separated by at most two comments (Hamilton et al., 2017; Del Tredici and Fernández, 2018). Since online communications are asynchronous, a discussion thread created at time t may still have active comments from users at time $t+1$ or later. For such threads, I only included interactions at time t in $G_c(t)$ and grouped later interactions into the future time steps at which these interactions happened. Users marked as `[deleted]` or `AutoModerater` were all removed. Networks with less than 50 nodes were excluded from the current analysis. As communities may have a wide range of life spans, $t = 1, 2, \ldots, t_{max}$ are only defined with respect to the life span of a community instead of absolute times in months. After filtering, a total of 289.8k community networks have been extracted for all 4420 communities.

2.3.2.2 Inter-community networks

While the temporal network $\mathcal{G}_c$ represents the interaction dynamics within a community, I also identify the network dynamics between communities. I created temporal network $\mathcal{G}_{IC}$ to characterize the connections between communities at consecutive months $t = 1, 2, \ldots, t_{max}$, $\mathcal{G}_{IC} = \{G_{IC}(1), G_{IC}(2), \ldots, G_{IC}(t_{max})\}$, in which $G_{IC}(t) = \{V_{IC}(t), E_{IC}(t)\}$. $V_{IC}(t)$ contains the set of nodes whereas $E_{IC}(t)$ is the set of edges between communities. A community is represented as a node u_i in $G_{IC}(t)$, except for communities that do not exist or are no longer active at time t. Two communities are determined to be connected if they share active users, that is, users who had posted at least 2 comments in both communities during that month. Each network snapshot is initiated as a weighted and undirected network with the edge weights set to the numbers of shared users, as an approximation of connection strength. Finally, 152 inter-community networks have been constructed since the inception of Reddit in 2005 until October 2018.

2.4 Network statistics

Communities in Reddit can be defined in terms of how their members relate within the community (intra) and how the community relates to other communities (inter) through multi-community memberships by its users (Tan and Lee, 2015). I formalize both as potential influences. As network attributes may be affected by the hyperparameters for network construction, I additionally validate this approach in Section 2.5.

In this section, I describe the extraction of certain network features that are used to characterize the global structure of networks.

2.4.1 Intra-community features

I take the following network measurements for each $G_c(t)$ to characterize the global properties of community networks: number of nodes, number of edges, density, average local clustering coefficient, transitivity, average degree, maximum degree, degree assortativity, the fraction of the largest connected components and fraction of singletons. These network measures can characterize the size and connectedness of Reddit networks (Hamilton et al., 2017; Cunha et al., 2019). Specifically, the number of nodes mainly measures the network size. Network connectedness is characterized by the number of edges, density, average local clustering coefficient, transitivity, average degree, maximum degree, degree assortativity, the fraction of the largest connected components, and the fraction of singletons. For example, the clustering coefficient and transitivity measure whether there are local clusters (isolated groups) in the network, whereas the fraction of singletons measures the number of users that are not connected in the network. The number of edges, average degree, and density characterize the global connection of the networks. Maximum degree, degree assortativity, and the fraction of the largest connected components mainly describe whether the connections are concentrated in a small group of users. Altogether, these network measures provide a detailed characterization of the network structures numerically.

Parameters like average local clustering coefficient, transitivity, and assortativity are highly influenced by the underlying degree distribution (Hamilton et al., 2017). I adjusted these parameters by computing their relative differences with respect to the mean values of five random baseline networks, which were generated by randomly rewiring the original network for 10 $\times$ edge count iterations and preserving the original degree sequence. These features are referred to as adjusted local clustering coefficient, adjusted transitivity, and adjusted assortativity in the following text.

In a close-knit community, the majority of users should form strong ties with each other. I should expect to observe higher density, a higher fraction of the largest connected components, and a lower fraction of singletons, since these measures quantify well-connectedness of online communities. The average clustering coefficient measures the shared neighbors between users, which should also be higher for close-knit communities.

2.4.2 Inter-community features

Reddit is a community that is highly connected internally, with users constantly moving between sub-communities. In addition to the intra-community network features, it is also necessary to measure a community's external connections to other communities. User mobility and external influence have been found to play a role in the process of lexical change (Conde-Silvestre, 2012). For each between-community network snapshot $G_{IC}(t)$ at time t, I focus on the network properties of individual nodes (communities). To quantify their structural roles, I computed the degree centrality, closeness centrality, eigenvector centrality, betweenness centrality, and PageRank centrality for each community node. These centrality measures quantify the connectedness of a community to other communities, which can be used as an indicator of their degrees of external contact and user mobility (Newman, 2018). Centrality is a common measure in network analysis and it measures the importance of a network node in the whole network. For example, the importance here can be defined as whether a node is connected to many other nodes (degree centrality), whether a node is connected to other important nodes (Pagerank centrality) or whether a node connects two relatively isolated clusters in the network (betweenness centrality). Each centrality measure only encodes partial information of node importance, so combining these measures in the model could provide a more holistic characterization of the inter-community networks.

2.5 Validation of the constructed networks

2.5.1 Intra-community networks

I constructed the network representations of Reddit communities with the same method as that used by Hamilton et al. (2017) and Del Tredici and Fernández (2018), so that this study is consistent and comparable with previous works. The rationale behind this setting is that ”two users who comment in such proximity interacted with each other, or at least directly with the same material” (Hamilton et al., 2017).

Here I compare the inter-community networks in this study with two types of baseline networks extracted from the same Reddit communities. I randomly sampled 100 networks from the data and created the following two baseline networks.

- **DRG**: The Direct-reply Graph (DRG) was constructed by treating every user as a node. An edge was created between two users if one user directly replied to the other. This network could *underestimate* the user interactions as users are likely to read nearby posts in the same comment chain when replying.

- **TG**: The Thread Graph (TG) was constructed by setting each user as a node and two users were connected by an edge if they had commented in the same thread. This network might *overestimate* the user interactions because, in some mega threads that span hundreds or thousands of posts, users might not interact with all the people in the same thread but only with nearby users.

As these two baseline networks might either underestimate or overestimate the connections, I used these two networks to provide an estimate of the possible errors of the constructed networks.

Variables	Kendall τ correlations	
	TG	DRG
# Nodes	1 (0.0)	1 (0.0)
# Edges	0.81 (0.27)	0.957 (0.05)
density	0.50 (0.34)	0.928 (0.08)
assortativity	0.04 (0.26)	0.189 (0.25)
local clustering	0.40 (0.18)	0.21 (0.36)
global clustering	0.07 (0.29)	0.56 (0.25)
average degree	0.54 (0.30)	0.84 (0.14)
max degree	0.75 (0.12)	0.66 (0.22)
min degree	0.34 (0.03)	0.44 (0.09)
LCC %	0.60 (0.19)	0.55 (0.19)
Singletons %	0.53 (0.22)	0.49 (0.22)

Table 2.2: Correlations between two baseline intra-community networks and the intra-community networks used in this study. The reported numbers are mean correlations with standard deviations inside the parentheses.

The results are presented in Table 2.2. Despite the different settings, most of the network parameters have correlations ranging from moderate to strong. But the correlations for assortativity and clustering coefficients are weaker. However, TG is not considered a good indication of the connections on Reddit as users are unlikely to interact with all users in a long thread. DRG and the constructed networks are more similar to each other. Hamilton et al. (2017) had noted that changing the original networks to DRG did not significantly change their analysis results of Reddit networks.

2.5.2 Inter-community networks

In order to validate this approach to construct the inter-community graph, I constructed different inter-community graphs by setting the posting threshold of active users to 2, 3 and 4. One concern is that setting the threshold too low ($>= 1$) results in extremely dense graphs, which are challenging to process.

After extracting the network features from these networks, I compared them by computing the Kendall rank correlation coefficients between these features. The results in Table 2.3 show that these networks are highly correlated in structural features, especially for the degree, eigenvector, and PageRank centralities. The correlations for betweenness and closeness are more unstable but still moderately correlated. So adjusting the threshold does not significantly bias the results qualitatively.

Centrality	**Kendall τ correlations**	
	Threshold: 3	**Threshold: 4**
Betweenness	0.62 (0.26)	0.48 (0.37)
Closeness	0.64 (0.17)	0.50 (0.25)
Degree	0.85 (0.05)	0.79 (0.06)
Eigenvector	0.92 (0.03)	0.88 (0.05)
Pagerank	0.90 (0.04)	0.86 (0.06)

Table 2.3: Correlations between two baseline inter-community networks and the inter-community networks used in this study. The reported numbers are mean correlations with standard deviations inside the bracket.

2.5.3 A summary of network statistics

Table 2.4 provides some general statistics of the whole Reddit Network Corpus.

	Total
Months	152
Subreddits	4420
Inter.Networks	152
Intra.Networks	289170
Users	$>$ 50 millions
Neologisms	80071

Table 2.4: Summary of the full corpus.

	Median	Inter-quartile range
Nodes	765	[351, 1776]
Density	0.0059	[0.0025, 0.0121]
Average Degree	4.19	[3.04, 6.29]
Largest Connected Component	88.6%	[81.9%, 92.4%]
Singletons	9.4%	[6.5%, 13.9%]
Inter-Community Degree	1351	[492, 2322]

Table 2.5: Statistical summary of 289170 Subreddit networks.

The average duration of the 4420 subreddit communities is 65 months. Statistical summaries of all 289170 networks are presented in Table 2.5.

2.6 Lexical innovations

In what types of communities are neologisms likely to be introduced? Here, I investigate the extent to which the number of innovations introduced per month can be predicted with only the structural properties of community networks.

2.6.1 Experiment setup

Given a set of communities $C = \{c_1, c_2, \ldots, c_n\}$ spanning time steps $T = \{1, 2, \ldots, t_{max}\}$, I aim to predict the count of monthly lexical innovations for each community $Y = \{y_1^{c_1}, y_2^{c_1}, \ldots, y_{t_{max}}^{c_n}\}$ from the corresponding network attributes $\mathbf{X} = \{\mathbf{x}_1^{c_1}, \mathbf{x}_2^{c_1}, \ldots, \mathbf{x}_{t_{max}}^{c_n}\}$. The predicted variable $y_t^{c_n}$ is computed by counting only innovations first introduced into community c_n at month t. Any subsequent usage of the same innovations after their first introduction is not counted as innovations in community c_n. The feature vector $\mathbf{x}_t^{c_n}$ is the structural features of the network at time t for c_n. After removing about 0.03% invalid data points and outliers, I ended up with 289.1k samples for the task.

2.6.2 Feature preprocessing

I used mean-variance normalization to normalize all prediction features. Since the distribution of some features was highly skewed, before normalization, I log-transformed the following intra-community features: *number of nodes, the number of edges, density, average degree, maximum degree*, and the following inter-community features: *degree centrality, closeness centrality, Pagerank centrality, betweenness centrality, eigenvector centrality*. The rest of the features were directly normalized. Whether to perform log transformation was determined by visual inspection of the density plot. A small number 10^{-6} was added before taking the logarithm to improve numerical stability. I found that such a practice improved the performance during cross-validation relative to directly normalizing all features.

The following features were used to predict the number of innovations per month. Some of the features were correlated and the correlations varied from weak to strong.

- **Inter.**: *degree centrality, closeness centrality, Pagerank centrality, betweenness centrality, eigenvector centrality*

- **Intra.**: *number of nodes, number of edges, density, average degree, maximum degree, the proportion of the largest connected components, the proportion of singletons, adjusted assortativity, adjusted transitivity, adjusted clustering coefficients.*

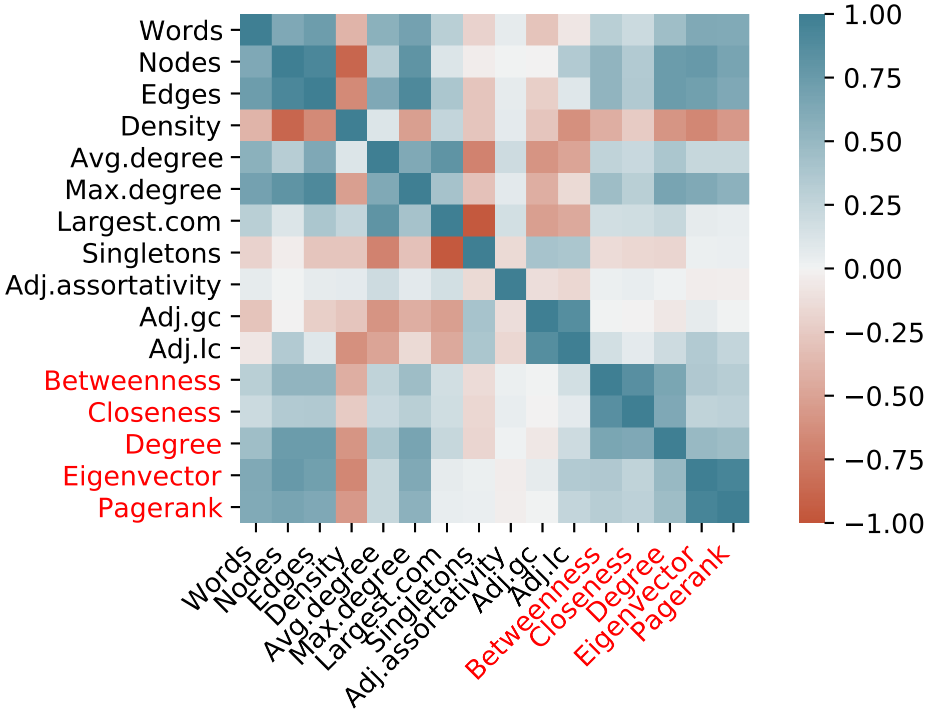


Figure 2.2: The correlation matrix between variables. Correlation coefficients were computed using the Spearman correlation, as some variables are related to log-linear relations. Variable names in black are within-community features while variables in red are inter-community features.

2.6.3 Correlations between variables

For empirical networks, some network attributes are often correlated. Here I present the correlation matrix between variables used in innovation prediction in Figure 2.2 for illustration. The correlation matrix for features in survival analysis also exhibits a similar pattern of correlations. Directly using these correlated features will hinder the interpretation of individual variables.

To eliminate correlations between variables, PCA with whitening was applied to decompose all of the features into principal components. I did consider the delta features, which were the change in these variables with respect to the last month. However, these added temporal features did not improve the performance. So I assumed that changes in each month might not be highly relevant.

2.6.4 Implementation

All models were implemented in `sklearn`. The baseline was the mean number of innovations across all time and all subreddits as the prediction. For the rest of the models, I performed tenfold cross-validation to select the best parameters. After parameter selection, the regularization parameter for the Poisson regression was 10^{-2} and the maximum number of iterations was 300. For the histogram-based gradient boosting trees, the maximum number of splits was set to 256 and the loss was the Poisson loss. Otherwise, I kept the default hyperparameters.

I used both intra-community and inter-community features for innovation prediction. However, in empirical networks, certain structural features tend to be correlated. For example, network size and density are usually strongly correlated on a log-log scale in online social networks (Backstrom et al., 2012), which is also apparent in this dataset (Spearman ρ=-0.87). Such correlations may confound the interpretation of the feature contributions (see Section 2.6.3). To generate orthogonal features, I first standardized all 15 network features and then used principal component analysis (PCA) with whitening to decompose them into principal components (PCs). Standardization was necessary as it could prevent a few variables with a large range of variance from dominating the PCs. I found that the first five PCs accounted for 87% of the total variance and 10 PCs explained 99% of the total variance.

Since counts of innovations are non-negative integers, Poisson regression and Histogram-based Gradient Boosted Trees (HGBT) with Poisson loss were used to predict the number of innovations with PCs. The model parameters were selected through ten-fold cross-validation. The data were randomly partitioned into training and test sets with a ratio of 90%/10%. I report the mean absolute error (MAE) and the mean Poisson deviance (MPD) averaged across 20 runs with different random partitions of data. Both metrics should be minimized by the models.

Model	**MAE**	**MPD**
Baseline (mean)	19.37	30.16
Poisson reg. (PCs=5)	11.79	12.29
Poisson reg. (PCs=10)	11.14	11.03
Poisson reg. (raw feat.)	11.72	12.21
HGBT (PCs=5)	10.57	9.63
HGBT (PCs=10)	9.65	8.19
HGBT (raw feat.)	**9.24**	**7.49**

Table 2.6: Results of lexical innovation prediction.

2.6.5 Results

As summarized in Table 2.6, all models outperformed the mean baseline by a significant margin, suggesting that the internal network structures and the external connections to other communities are systematically correlated to the count of lexical innovations per month. The three largest coefficients of the Poisson model with 5 PCs correspond to the first three PCs (see Figure 2.3 and

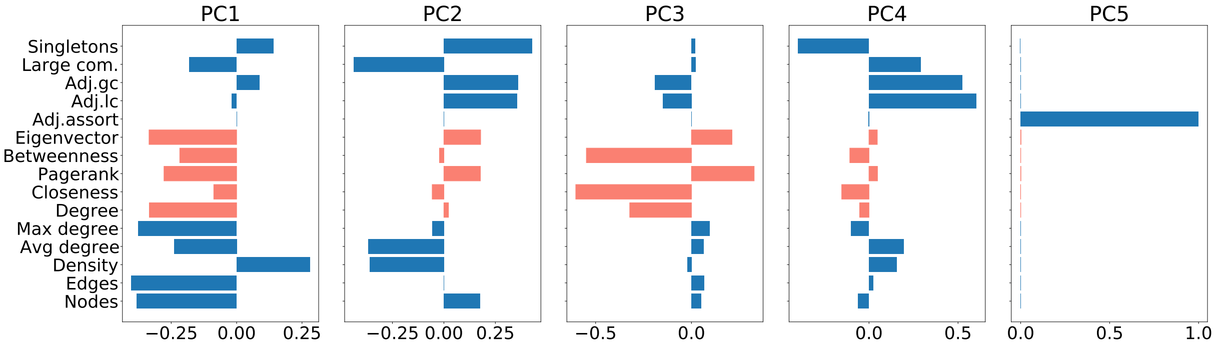


Figure 2.3: The decomposition of the PCs is used in predicting innovations. Inter-community features are highlighted in orange bars. Adj.lc, Adj.gc, and Adj.assort are local clustering coefficient, global clustering coefficient, and assortativity adjusted with respect to a random network. PC1 represents the overall size, PC2 the density of intra-community connections, and PC3 the inter-community connections.

Variables	Coefficients
PC1	-0.877
PC2	-0.195
PC3	0.193
PC4	-0.024
PC5	-0.003

Table 2.7: **[Predicting innovations]** Coefficients for the Poisson regression with first 5 PCs.

Table 2.7) [2]. PC1 represents the overall size of the network, such that the Poisson model predicts that networks having a larger overall size tend to have more innovations (Coefficient: -0.87). PC2 indicates the fragmentation and the local clusteredness of the network and contributes negatively to lexical innovation (Coefficient.: -0.20). In other words, fragmented networks with local clusters tend to have fewer innovations as this structure inhibits the spread of information. PC3 is generally related to inter-community connections with a positive correlation to innovation (Coefficient.: 0.19). Yet what matters is not the number of communities connected (degree centrality) but the quality of those connections (Pagerank centrality). High Pagerank centrality suggests that the network might be connected to many influential communities, as these connections are weighted higher in the Pagerank algorithm (Page et al., 1999).

While structural properties can account for many regularities in the creation of lexical innovations, there are also surges of innovations that cannot be explained by structural factors alone. Inspection of the data suggests that the surges of innovations at the tail of empirical distributions are often related to some factors beyond network structures, including topical variations or external

[2] Note that the coefficient sign for a PC must be interpreted with respect to its loading on structural components.

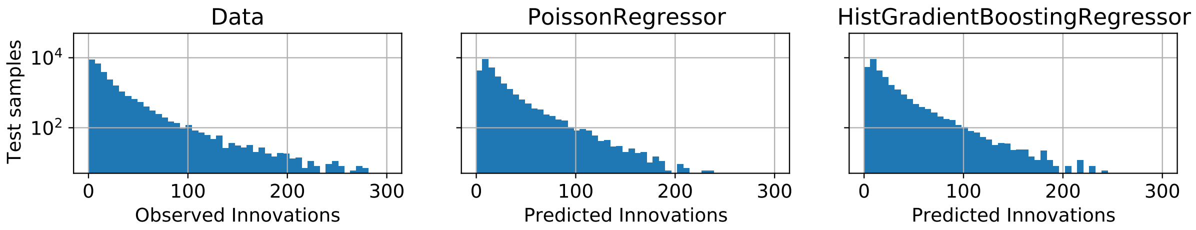


Figure 2.4: **[Predicting innovations]** Distribution of observed innovations in test samples (left) and predicted distributions by Poisson regression with all network features (middle) and by HGBT with all network features (right). Both models well approximate the empirical distribution of lexical innovation counts but fall short of predicting the trailing long tail.

events, such as community migration or new game releases for some game communities.

2.7 Survival Analysis of Neologisms

Not all lexical innovations survive through time, with only a few neologisms eventually becoming widely adopted by community members. Here, I test the structural factors that systematically affect the survival of words in online communities.

Survival analysis models the elapsed time before a future event happens (Kleinbaum and Klein, 2010), which has been used to predict word survival (Stewart and Eisenstein, 2018). Unlike the traditional Cox model, deep survival analysis approximates the risk (hazard) with neural networks, thereby achieving improved performance. Yet traditional models like the Cox model are highly interpretable. To combine the strengths of both models, I fitted both Cox models and deep survival analysis models in this section.

2.7.1 Deep survival analysis

I estimated word survival with the Logistic Hazard model (LH) proposed by Kvamme and Borgan (2019). Given samples $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and time steps $\{1, 2, \ldots, T\}$, the LH method estimates $h(t|\mathbf{x})$, the hazard function of the death event with respect to time t, with a deep neural network. The hazard function can be interpreted as the word's "danger of dying" at t. In this section, I describe the details of deep survival analysis.

2.7.1.1 Model specification

I adopted the Logistic Hazard model developed by Kvamme and Borgan (2019) and Kvamme et al. (2019). The original derivation comes from Kvamme and Borgan (2019).

In survival analysis, given a set of discrete time steps $T = \{t_1, t_2, \dots, t_n\}$ and the event time t^*, the goal is to estimate the probability mass distribution (PMF) of the event time $f(t)$ and the survival function $S(t)$.

$$\begin{aligned} f(t) &= P(t^* = t_i), \\ S(t) &= P(t^* > t_i) = \sum_{j>i} f(t_j) \end{aligned} \tag{2.1}$$

The model can also be expressed as the hazard function $h(t)$.

$$\begin{aligned} h(t) &= P(t^* = t_i | t^* > t_{i-1}) \\ &= \frac{f(t_i)}{S(t_{i-1})} \\ &= \frac{S(t_{i-1}) - S(t_i)}{S(t_{i-1})} \end{aligned} \tag{2.2}$$

With the above equations, the survival function can be rewritten as follows.

$$\begin{aligned} f(t_i) &= h(t_i)S(t_{i-1}) \\ S(t_i) &= [1 - h(t_i)]S(t_{i-1}) \end{aligned} \tag{2.3}$$

It then follows that

$$S(t_i) = \prod_{k=1}^{i} [1 - h(t_k)] \tag{2.4}$$

For each individual i, the likelihood function can be formulated as

$$L_i = f(t_i)^{d_i} S(t_i)^{1-d_i} \tag{2.5}$$

The above equation can be rewritten with respect to the hazard function.

$$\begin{aligned} L_i =& f(t_i)^{d_i} S(t_i)^{1-d_i} \\ =& [h(t_i)S(t_{i-1})]^{d_i} \Big([1 - h(t_i)]S(t_{i-1})\Big)^{1-d_i} \\ =& h(t_i)^{d_i} [1 - h(t_i)]^{1-d_i} S(t_{i-1}) \\ =& h(t_i)^{d_i} [1 - h(t_i)]^{1-d_i} \prod_{k=1}^{i-1} [1 - h(t_k)] \end{aligned} \tag{2.6}$$

The loss function is the negative log likelihood function, the negative of the sum of $log(L_i)$ over all samples. After some algebraic operations, the loss function of the Logistic Hazard model can be formulated as the common binary cross-entropy function.

$$\begin{aligned}\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{T}\big(&y_{ij}log[h(t_j|\mathbf{x}_i)] \\ &+(1-y_{ij})log[1-h(t_j|\mathbf{x}_i)]\big)\end{aligned} \tag{2.7}$$

where y_{ij} is the binary event indicator for sample i at time t.

Let $\mathbf{x}$ be an input feature vector and $\phi(\mathbf{x}) \in \mathbb{R}^h$ is the neural network that transforms input $\mathbf{x}$ into h output vectors. Each output vector corresponds to a discrete time step such that $\phi(\mathbf{x}) = \{\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \ldots, \phi_h(\mathbf{x})\}$. The hazard function then can be approximated by the sigmoid function.

$$h(t_i|\mathbf{x}) = \frac{1}{1+exp[-\phi_i(\mathbf{x})]} \tag{2.8}$$

After the model is trained, the survival function $S(t|\mathbf{x}_i)$ for sample $\mathbf{x}_i$ can be computed as

$$S(t|\mathbf{x}_i) = \prod_{t=1}^{T}[1-h(t|\mathbf{x}_i)] \tag{2.9}$$

$S(t|\mathbf{x}_i)$ can be interpreted as the chance of survival at time t for sample $\mathbf{x}_i$, that is, the survival probability of a word given the corresponding network features at time t.

2.7.2 Baseline models

For comparison and interpretation, I also ran baseline Cox's proportional hazard models (Kleinbaum and Klein, 2010) with the same data partitions and discretization scheme. Unlike the deep survival analysis which uses complex non-linear functions to estimate the hazard, the Cox model assumes a linear relationship between the log hazard of an individual word and the individual network features (Kleinbaum and Klein, 2010). This assumption slightly restricts the predictive capacity of the Cox model, in comparison with the deep survival analysis. However, this also makes the Cox model highly interpretable, because each independent variable's contributions can be directly assessed through regression coefficients β_i. The Cox model estimates the hazard function $h(t_i|\mathbf{x})$ with the following equations.

$$h(t_i|\mathbf{x}) = \beta_0(t_i) \cdot \exp\Big(\sum_{i=1}^{n}\beta_i(\mathbf{x}-\overline{\mathbf{x}})\Big) \tag{2.10}$$

2.7.3 Data coding

I consider only communities that have existed longer than six months and words that survived more than three months. The subreddit duration restriction avoids right-censoring of the data from new communities forming and quickly dying (a common event), which would skew estimates of word survival. A word's survival time is defined as the total number of months a word persists in a community, excluding the intervening month in which the word is not used. The last time step t at which the word shows up is considered the "death" event. However, if this last time step is also the last three recorded months, this word is considered right-censored such that a death event has not happened. This three-month buffer period is added to avoid false negatives. The network features for predictions were derived from averaging all the monthly features for the months that a particular word has existed. After preprocessing, I ended up with 1.47M samples with 69,683 distinct words. All features were then transformed into 10 orthogonal principal components using PCA with whitening. The first 5 PCs accounted for 90% of the total variance whereas all 10 PCs explained 99% of the variance.

2.7.4 Implementation

Models of deep survival analysis were implemented using the package `pycox` (Kvamme et al., 2019) [3]. The network features were normalized and partitioned in the same way as described in Section 2.6.2.

The actual survival time for these neologisms varied from 3 to 152 months. First, I discretized the survival time measured in actual months into 100 intervals based on the distribution of the event times, with the assumption that each interval has the same decrease in the survival probability. The resulting grid was denser during months with more event times and sparser during months with fewer event times. Such a practice is recommended by Kvamme and Borgan (2019), as it reduces parameters and stabilizes training.

I trained a three-layered Logistic Hazard model. For each of the first two layers, I used a linear layer with 256 hidden dimensions and ReLU activation function, followed by batch normalization and a dropout with a probability of 0.1. The last layer was a linear layer with an output dimension of 100 followed by a sigmoid activation function.

During training, I used the Adam optimizer with a learning rate of 0.001 and a batch size of 2048 samples. All hyperparameters were tuned with a simple grid search on the development set. Each model was trained for 5 epochs and was run 10 times with different random seeds and different partitions of data each time. The performance metrics were averaged over all 10 runs.

[3] https://github.com/havakv/pycox

Model	**Concordance**	**IBS**
Random baseline	0.50	0.25
Cox Model (PCs=5)	0.600	0.297
Cox Model (PCs=10)	0.662	0.289
Cox Model (raw)	0.665	0.209
LH (PCs=5)	0.584	0.245
LH (PCs=10)	0.691	0.192
LH (raw)	**0.718**	**0.152**

Table 2.8: Survival analysis results. All models outperform the concordance baseline.

These models were trained on an Nvidia V100 GPU and each run took about less than a minute to complete.

In each run, the data were randomly partitioned into around 80%, 10% and 10% portions as training, development and test sets, respectively, with different random seeds. In order to avoid information leaking, I ensured that samples in these three sets were from distinct subreddits. Each model was run for 3 epochs and was run 10 times with different data partitioning. The performance metrics were averaged.

I also ran baseline Cox models under the same conditions for comparison. I ran each Cox model ten times and report the average performance. All baseline Cox's models were implemented using the `CoxPHFitter` function via the package `lifelines`.

The performance is evaluated with *time-dependent concordance* (Antolini et al., 2005) and *Integrated Brier Score* (IBS) (Kvamme et al., 2019). Concordance measures the model's capacity to provide a reliable ranking of individual risk scores. A good concordance score should be above the 0.5 random baseline and close to 1. The IBS is the average squared distances between the observed survival events and the predicted survival probability and should be minimized by the model.

2.7.5 Results

The results in Table 2.8 show that structural factors of the community in which a neologism is introduced can predict its chance of survival or death, with all models outperforming the baseline by a large margin. Both models correctly predict the ranking of survival probability across time for candidate words. While the Logistic Hazard model outperforms the Cox model in general, it also suffers from unstable performance over multiple runs and tends to overfit the training data. Since samples in training and test sets do not overlap in subreddits, such performance indicates that there are strong associations between network structures and word survival such that the proposed models can generalize across communities.

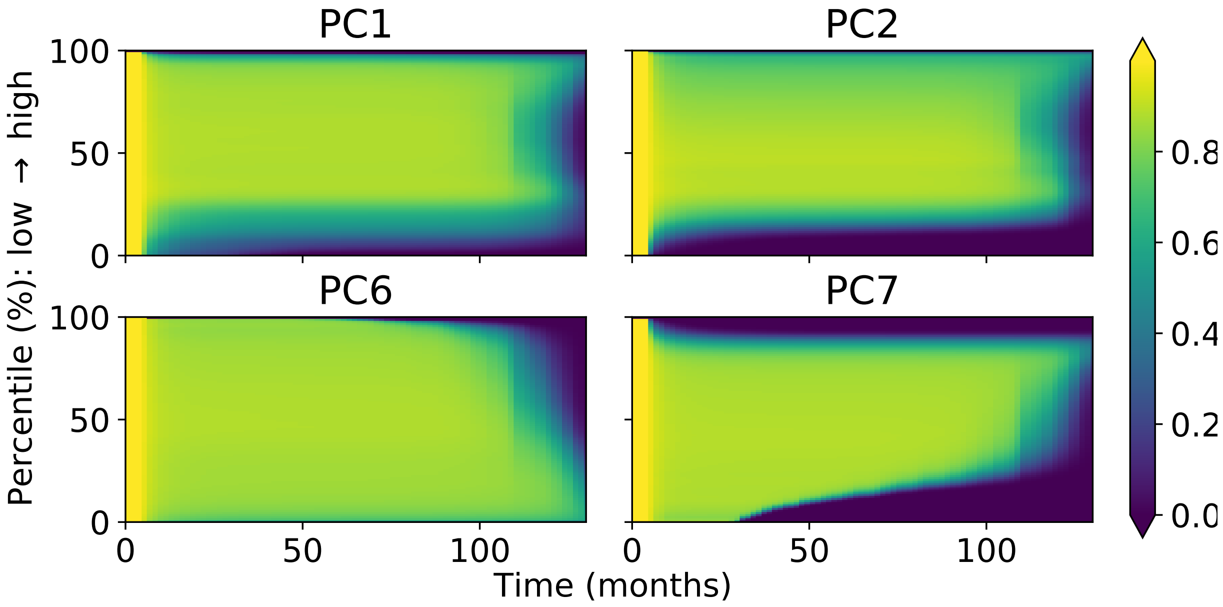


Figure 2.5: **[Predicting survival]** The contribution of predictors (PC1, PC2, PC6, and PC7) to the survival probability $S(t|\mathbf{x})$ with remaining features fixed. Brighter regions indicate high survival rates.

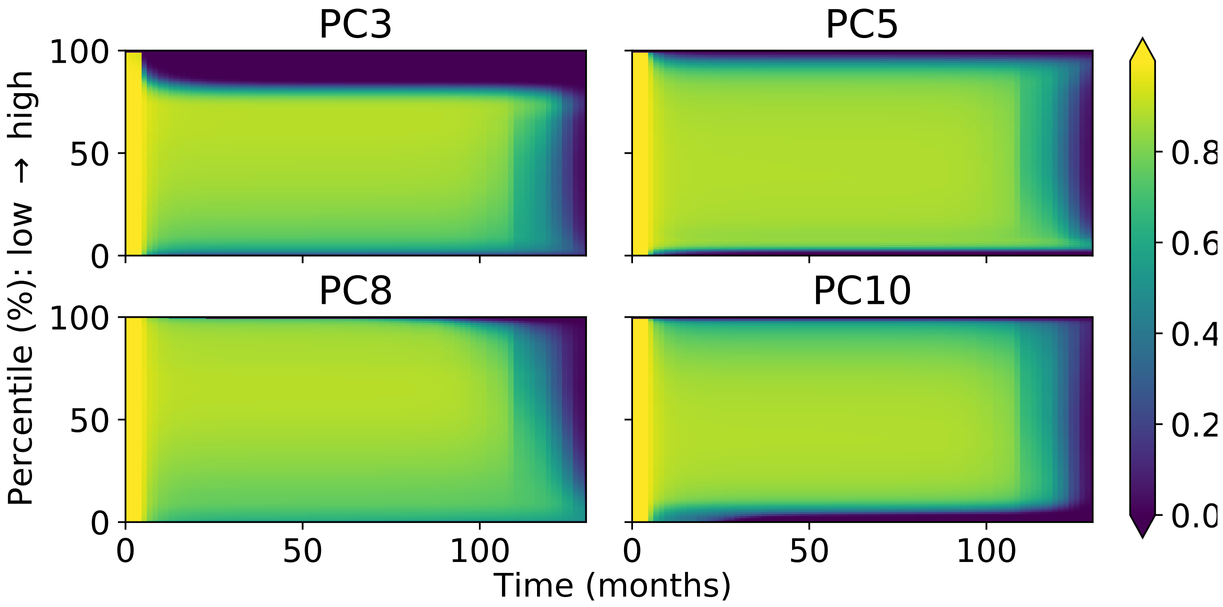


Figure 2.6: **[Predicting survival]** The contribution of the predictors (PC3, PC5, PC8 and PC10) to the survival probability $S(t|\mathbf{x})$ with remaining features fixed.

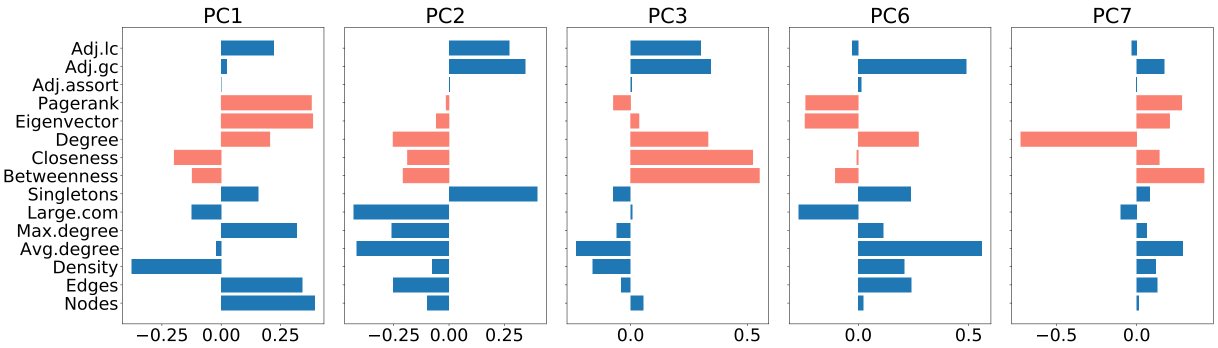


Figure 2.7: **[Predicting survival]** The five highest weighted PCs used by the survival model. Inter-community features are indicated by orange bars. Adj.lc, Adj.gc and Adj.assort are local clustering coefficients, global clustering coefficients and assortativity adjusted to a random network. PC1 represents the overall size, PC2 the within-community connections, and PC3 the inter-community connections. PC6 and PC7 are specific combinations of intra- and inter-community connections.

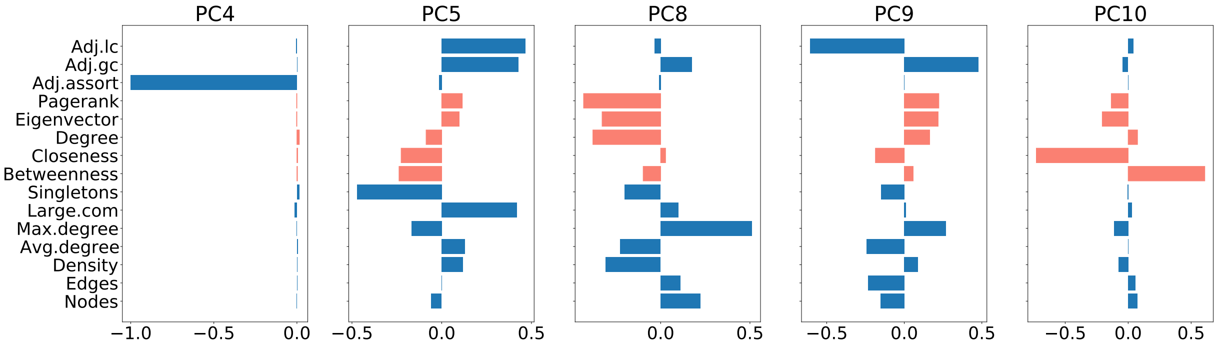


Figure 2.8: **[Predicting survival]** Remaining PCs for the network features. Inter-community features are highlighted in orange bars.

To interpret the LH model with 10 PCs, I generate the survival function $S(t|x)$ by varying a single feature from low to high but keep the remainder fixed at their median value (Figures 2.5 and 2.6). While the Cox model predicts the hazard (death rate) and the LH model predicts $S(t|x)$ (the survival rate) (in the reverse direction), I found that both models were highly consistent in assessing the input PCs, both in terms of relative weights and directions.

The relative contributions of each PC correspond to the magnitude of coefficients of the Cox model, shown in Table 2.9. As PCs are abstract dimensions, they should be interpreted as a weighted combination of individual network attributes. The contributions of network structural attributes to each PC are plotted in Figure 2.7 and Figure 2.8. A large overall size (PC1) tends to preserve neologisms, as large communities provide a basic threshold population for words to be used. In addition to sheer size, global network topology also contributes to neologism survival.

Variables	Coef.	Exp(coef)	S.E.
PC1	-0.122***	0.885	0.002
PC2	-0.072***	0.930	0.002
PC3	0.170***	1.186	0.003
PC4	0.009***	1.001	0.001
PC5	-0.017***	0.984	0.001
PC6	-0.160***	0.852	0.001
PC7	-0.516***	1.675	0.002
PC8	-0.048***	0.953	0.001
PC9	-0.004***	1.004	0.001
PC10	-0.054***	0.947	0.002

Table 2.9: Results of the Cox model. All coefficients are highly significant. Exp(coef) refers to the hazard, or the probability of death. Lower Exp(coef) suggests that this variable is protective. S.E. refers to the standard error of the regression coefficients.

PC2, PC3, PC6 and PC7 (see Figure 2.7 and 2.8) correspond to three different network structures. PC3 represents networks that have many external connections but are split into multiple clusters within the community, which contributes negatively to the survival probabilities. In contrast, less clustered networks with dense edges and rich external connections (PC2) increase word survival rates. Both PC6 and PC7 boost word survival rate and they both represent networks that are relatively densely connected, but PC6 has high connections to many external communities and is more fragmented whereas PC7 is more isolated in the inter-community network (low degree centrality) but its external connections are influential communities (high Pagerank and Betweenness centrality). This may suggest that inter- and intra-community connections complement each other. In general, within a community, dense connections in the network keep words alive whereas local clusters in the network are adverse to word survival. In the multi-community landscape, more external connections tend to promote word survival.

2.8 Lexical levelling

Levelling refers to the gradual replacement of localized linguistic features (*marked*) by mainstream linguistic features (*unmarked*) over the whole community (Kerswill, 2003), which has been observed in a wide range of offline linguistic communities due to increasing mobility and external contacts (Milroy, 2002; Kerswill, 2003).

The subreddit communities have become increasingly interconnected over time, as the average inter-community degree has increased from 6 in January 2008 to 2,323 in October 2018 (Figure 2.9). While some of these could be accounted for by the simultaneous growth in the number

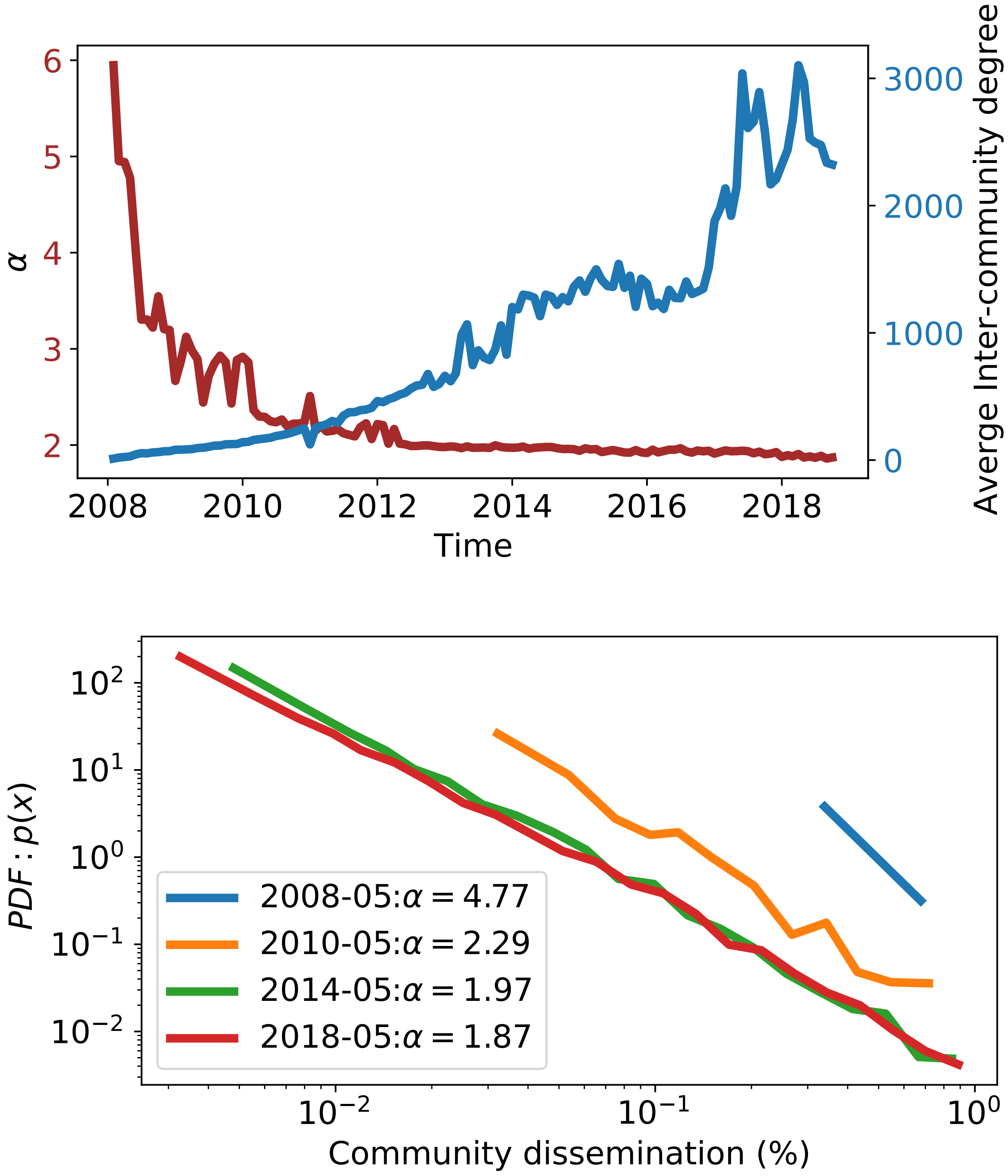


Figure 2.9: [**Top**] Change of average community degree and the shape parameter α of the power law fit $p(x) \propto x^{-\alpha}$ over time. The average community degree is increasing, indicating that more communities are connected to each other. The α is decreasing, suggesting that the tail of the distribution has become thicker, or more community-specific words have emerged. [**Bottom**] Snapshots of PDFs of dissemination across communities over time.

of subreddits, the growth in connectedness is also apparent. Such an increase in contact could promote the spread of neologisms across Reddit. In the same period, the number of variants that spread to more than 60% of the communities has grown slightly from 7 to 22. Some of the notable examples include words like *lol*, *alot*, *imao* and *cuz*. Meanwhile, the variants that are only confined to one community grew rapidly from 1992 in 2008 to 23,397 in 2018. The widespread use of some neologisms does not necessarily cause the loss of local expressions, as in offline communities. Instead, the community-specific terms and community-general terms develop in tandem. Many community-specific terms are nested within topic-based communities with little meaning overlap with those widespread variants and are therefore unlikely to be replaced by more general terms through levelling.

Figure 2.9 also shows that the probabilistic density distribution (PDF) of word dissemination (the percentage of communities sharing a neologism) conforms to the power-law fit $p(x) \propto x^{-\alpha}$, as a few words spread to most communities while most words are confined to a few communities. Further, the shape parameter α decreases asymptotically despite the growth of average inter-community degree (Figure 2.9), which implies that as the size of Reddit grows, more community-specific words, as well as more widespread words, emerge.

The number of community-specific words grew rapidly despite increased inter-community connectedness, which seems to go against the levelling trend observed in offline networks (Conde-Silvestre, 2012). In contrast to offline communities, these subreddit networks are of a different nature, as they are topic-based groups bounded by common interests. By joining these communities, users opt for fragmentation into some niche groups. Such segregation in topics and interests naturally brings in more community-specific words. In other words, there is not strong evidence for lexical levelling; instead, online communities go in the reverse direction, by developing more niche neologisms.

2.9 Discussion

In general, I find that global network properties significantly predict the innovation and retention of neologisms in online communities, indicating strong relations between global network properties and the process of lexical change.

In traditional sociolinguistics, as discussed in Section 2.2.1, weak ties within a social network have been linked to innovation and language change. Yet most studies only use indirect evidence to infer the underlying network types (Milroy and Milroy, 1985; Nevalainen, 2000; Dodsworth, 2019) and only broadly classify networks into loose-knit or close-knit ones (Milroy and Milroy, 1985; Milroy and Llamas, 2013; Sharma and Dodsworth, 2020). This might risk ignoring some fine-grained structural differences The quantitative analysis suggests that multiple structural properties at the community level play a role in online lexical change. Overall network size is the most prominent factor in lexical innovation and survival, as large communities provide the base population to create and use those neologisms. The effect of network size has also been emphasized in other network studies of language (Reali et al., 2018; Raviv et al., 2019; Laitinen et al., 2020). However, sheer size is only part of the story, as dense edges between users, the lack of separate local clusters, and rich external connections also promote both lexical innovation and survival. Dense connections within and across communities increase the visibility of neologisms so that they can be imitated by other users, as exposure alone predicts users' information spreading behavior (Bakshy et al., 2012). In contrast, local clustering tends to separate networks into disconnected parts, slowing the spread of new words. These structural attributes are also found to facilitate information spread in online social networks (Lerman and Ghosh, 2010). On a broader scale, the current results suggest that the lexical change process in online social networks may be similar to other information spread processes (Lerman and Ghosh, 2010; Guille et al., 2013). However, the findings that external contacts and connections promote lexical change are in contrast to some findings in offline communities. In her survey of the Toronto area, Nagy (2018) found no relationship between the strength of out-group ties and contact-influenced linguistic variation. While it is possible that Nagy (2018) only investigated relatively small samples of participants and linguistic variables, this difference implies that offline and online communities operate differently and more studies are needed to understand the differences.

The results show that conclusions drawn from offline communities might be insufficient to account for the behavior observed in online social networks. While the classic weak tie model emphasizes the role of loose social networks in offline language change (Milroy and Milroy, 1985; Nevalainen, 2000) and has been confirmed in studies of online communities that have investigated the role of individuals (Del Tredici and Fernández, 2018), this work further extends this model by showing that a variety of community-level network structural attributes also play a role in language change. For example, Figure 2.10 shows that larger size, denser connections, lack of local clustering and greater external contacts promote lexical innovation and retention in online communities, while density, as discussed most in offline studies, could be an emergent byproduct of network size. However, it is still unclear whether it is possible to draw an analogy between online social networks and offline networks since online communities like Reddit tend to be more loose-knit than offline communities and the migration is much easier in online communities. But here again the difference between online and offline networks merits further investigation.

The quantitative analysis also suggests a different levelling process in online communities with implications for sociolinguistic theories. The analysis of lexical growth and decline provides a window into how language use is related to people's activities. These results suggest that online communities operate differently from offline communities and deserve more attention from the

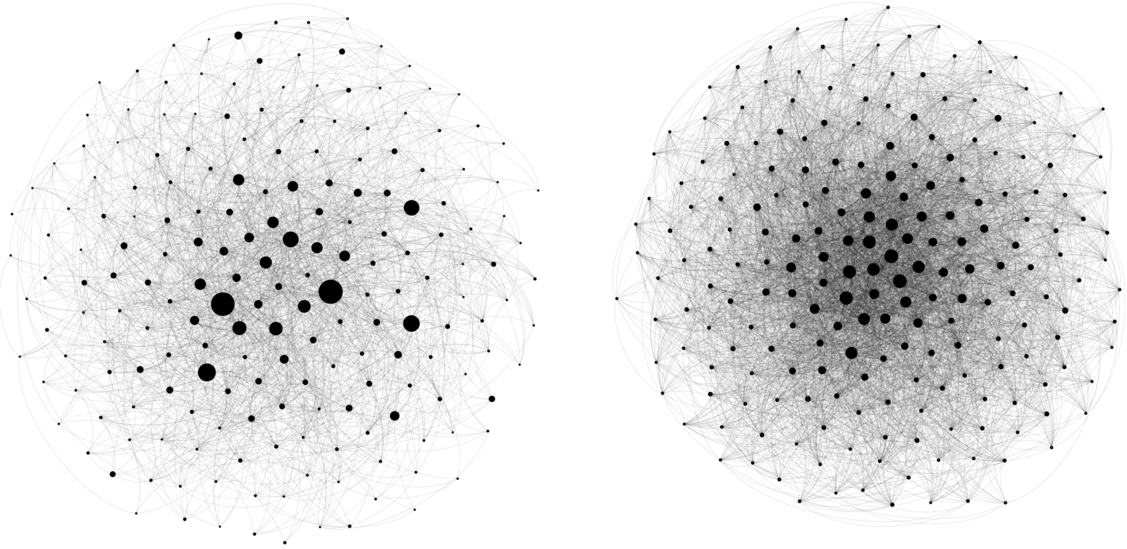


Figure 2.10: Applying the hypothesis of Milroy and Milroy (1985) to these two gaming subreddits of similar size suggests that the network with lower density (top; `r/masseffect`) will be more innovative than the more closely-connected community shown right (`r/F13thegame`). However, after controlling for size, the one with higher average degree (more inner-connections) (bottom: `r/F13thegame`) tends to develop more lexical innovations.

research community.

2.9.1 Limitations and future work

One limitation of this study is that topical variation is not explored in-depth, because I aimed to look at the contributions of networks alone by smoothing out topical variation with diverse communities. Yet topics have been found to affect users' posting behavior in online communities (Mathew et al., 2019) and niche topics do affect word retention (Altmann et al., 2011). In Reddit, communities involving certain niche or foreign topics, such as `r/pokemon`, might inherently introduce more lexical innovations than others. Secondly, I only focus on Internet neologisms on Reddit, which might not reflect the whole landscape of lexical change in all online communities. How these neologisms propagate across multiple social media platforms like Twitter, Facebook, Quora, and Mastodon, and how online and offline neologisms interact remain important questions to be addressed. Thirdly, while this study reveals the general patterns of lexical change, there are multiple sub-categories of neologisms such as discourse markers and name entities. It is of interest to ask whether different sub-categories may exhibit different patterns of usage in online communities. These research questions are worth exploring in future work.

CHAPTER 3

Study 2: Idiolectal Variations in the Online Register

3.1 Introduction[1]

In this chapter, I move away from linguistic change at the community-level and instead focus on how variation exists at the level of individual language users. Linguistic variation is ubiquitous and is in part a result of language users manifesting their linguistic identities through their linguistic choices. The notion that language functions as stylistic resources for the construction and performance of social identity rests upon two theoretical constructs: sociolect and idiolect (Grant and MacLeod, 2018). The term 'sociolect' refers to the socially structured variation at the group level, whereas 'idiolect' denotes language variation associated with individuals (Wardhaugh, 2011; Grant and MacLeod, 2018). Variationist sociolinguistics studies the systematic variation of sociolects that is linked to social indices within a linguistic community such as gender, ethnicity, and socioeconomic stratification (Labov, 1972). While a central concept in sociolinguistics, idiolect has received far more research attention in forensic linguistics (Wright, 2018; Grant and MacLeod, 2018).

Although idiolects have played a central role in stylometry and forensic linguistics, which seek to quantify and characterize individual textual features to separate authors (Grant, 2012; Coulthard et al., 2016; Neal et al., 2017), the theory of idiolect remains comparatively underdeveloped (Grant and MacLeod, 2018, also see Section 3.2.1). Yet an in-depth understanding of the nature and the variation of idiolect would not only shed light on the theoretical discussion of language variation but would also aid practical and forensic applications of linguistic science.

The study reported in this chapter characterizes the idiolectal variation of linguistic styles through a computational analysis of a large-scale corpus of short online texts. Specifically, I ask the following questions: 1) to what extent can I extract distinct styles from short texts, even for

[1] Portions of this chapter have appeared in *Idiosyncratic but not Arbitrary: Learning Idiolects in Online Registers Reveals Distinctive yet Consistent Individual Styles* (Zhu and Jurgens, 2021a) in the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

unseen authors; 2) what are the core stylistic dimensions along which individuals vary; and 3) to what extent is a given language user's idiolect consistent and distinctive?

This study introduces a new set of probing tasks for testing the relative contributions of different types of linguistic variation to idiolectal styles. By using deep metric learning, I first show that idiolects are, in fact, systematic and can be quantified separately from sociolects. Secondly, I show that the learned representations for idiolect also encode some stylistic dimensions with surprising regularity, analogous to linguistic regularity found in the word embeddings. Thirdly, using the proposed metrics for style distinctiveness and consistency, I show that individuals vary considerably in their internal consistency and distinctiveness in idiolect, which has implications for the limits of authorship recognition and the practice of forensic linguistics. For replication, I make the code available at `https://github.com/lingjzhu/idiolect`.

3.2 Idiolectal variation

3.2.1 Theoretical questions

'Idiolect' remains a fundamental yet elusive construct in sociolinguistics (Dittmar, 1996; Turell, 2010; Wright, 2018). The term has been defined as the totality of the possible utterances one could say (Bloch, 1948). A more recent definition by Wright (2018, online encyclopedia), perhaps narrower than the last one, is:

> "Idiolect" refers to an individual's unique variety and/or use of language, from the level of the phoneme to the level of discourse....The theory holds, therefore, that no two people who share a common language have exactly the same linguistic repertoire. In the same way that the variation exhibited in a person's language production is influenced by their dialect(s), sociolect(s) and by register, so too is it influenced by their personal, idiosyncratic, often habitual linguistic preferences—their idiolect.

In the context of forensic linguistics , 'idiolectal style' is practically described as

> a) how this [linguistic] system, shared by lots of people, is used in a distinctive way by a particular individual; b) the speaker/writer's production, which appears to be 'individual' and 'unique'. (Turell, 2010, page 217).

Yet some linguists deny the presence of idiolects (Barthes, 1968; Jakobson, 1971) from a theoretical point of view (see Turell, 2010; Barlow, 2013, for a summary of the history of defining idiolects). This study will not go into the theoretical debate of the existence of idiolects but adopts the definition of 'idiolectal style' used in the forensic linguistics context, that is, that individual styles—in this study, individual writing styles—appear to be unique.

Idiolect as a combination of one's cognitive capacity and sociolinguistic experiences (Grant and MacLeod, 2018) raises many interesting linguistic questions. First, what are idiolects composed of? Forensic studies often focus on a few linguistic features as capturing a person's idiolect – for example, as uniquely characterizing authorship of a text (Coulthard, 2004; Barlow, 2013; Wright, 2013) – but few have explicitly offered an explanation as to how and why, say, particular word sequences were useful or not in identifying authors (Wright, 2017). Statistical methods that rank the discriminating linguistic features between authors for the purpose of identifying individual variation have been proposed (Grant, 2012; Grant and MacLeod, 2020), but they are typically applied to a small number of authors. Grant and MacLeod (2020) have proposed a checklist of linguistic features at all levels for analyzing online texts to track criminals, for example. However, these features still rely on a combination of statistical and manual analysis on a case-by-case basis. Contributions of textual elements to idiolectal variation against a large background population are seldom systematically measured.

The perceived idiosyncrasies of idiolect often render it second to sociolect as an object of study (Labov, 1989). For example, Labov (1989) suggested that language should be the property of the community rather than individuals. Yet multiple scholars have suggested that idiolects are the building blocks of various sociolects (Eckert, 2012; Barlow, 2013; Wright, 2018). Hudson (1996) argued that language resides in individuals rather than communities, so the community aggregation of idiolects leads to social varieties (sociolects). Yet, similar to forensic study of idiolects, the exploration of individual variation in sociolinguistics tends to focus on a small number of linguistic variables (Schilling-Estes, 1998; Meyerhoff and Walker, 2007), making it hard to get an overall picture of individual variation and how it is related to social variation. It is important to note also the different goals of sociolinguistic and forensic studies of idiolect variation, with the former emphasizing its relations to performative identities and the latter (and this study) focusing on the uniqueness of textual style that distinguishes one author from other authors. From this latter perspective, though, it has been suggested that idiolect includes sociolinguistic variation that can potentially associate authors with their communities or by groups defined gender, age, profession and socio-economic status (Turell, 2010). I test this hypothesized relationship in Section 3.9 and do find preliminary quantitative evidence that idiolectal representations contain some information about sociolects.

Another theoretical question relevant to cognitive science as well as forensic linguistics is the extent to which an individual's idiolect is *distinctive* and *consistent* against a background population (Grant, 2012; Grant and MacLeod, 2018), two constructs that are the foundations of using language as evidence in forensic investigation. Studies in the domain of forensic linguistics (Johnson and Wright, 2014; Wright, 2013, 2017) provide evidence of both distinctiveness and consistency, yet, again, these forensic studies only focus on a specific set of linguistic features for a small group of authors. Yet it has also been suggested that idiolect is not stable in its entirety, as idiolect is a reflection of a person's linguistic experience, which might change over time (Wright, 2018). At the other extreme, the high performance of applying machine learning on authorship verification and attribution (Kestemont et al., 2018, 2019, 2020, 2021) stems from placing more emphasis on separating authors (distinctiveness) than consistency. An empirical investigation of these two concepts in a relatively large population remains to be conducted (Grant, 2012; Wright, 2017), given the daunting challenge of aggregating and analyzing large-scale data. But a deeper understanding of how individual styles are distinctive and consistent can offer guidance to the practice of forensic linguistics and provide a statistical foundation to weight linguistic evidence. To address this issue, here I propose two new metrics (see Section 3.10) to quantify these two linguistic constructs in large-scale textual datasets.

3.2.2 Stylometry and stylistic similarities

Stylomery is a subfield devoted to the study of individual styles whose goal is to analyze linguistic styles to determine the authorship of texts. There are usually two different but nonetheless related tasks, authorship verification and authorship attribution, though other tasks such as authorship clustering and profiling also exist (Neal et al., 2017). Authorship verification requires the determination of whether two texts are written by the same author or not, whereas authorship attribution means attributing the authorship of a text from a list of given authors. Traditional stylometry often relies on painstaking manual analysis of textual styles for a closed set of authors, yet statistical methods have gradually taken over in recent decades (Holmes, 1998; Koppel and Schler, 2003, 2004; Stamatatos, 2009). In the statistical paradigm, surface linguistic features, especially function words or character n-grams (n-consecutive characters), have been found to be effective in authorship analysis (Stamatatos, 2009; Kestemont, 2014; Neal et al., 2017). In a classic pipeline, linguistic features are first extracted for each text using linguistic rules written by linguists or statistical methods such as POS-taggers and parsers. Then these extracted features are used as numerical inputs to machine learning models such as logistic regression or support vector machines, which learn to associate linguistic features to authors (Stamatatos, 2009; Neal et al., 2017). The basic assumption in stylometry is the Stylome Hypothesis or the presence of 'stylome', "*a set of measurable traits of language products*", that can distinguish pairs of authors (Van Halteren et al., 2005, page 66). In this traditional approach, extracting reliable linguistic features is particularly important, as they determine the quality of the authorship model. The current features that are found to be reliably associated with individual styles span all linguistic levels, including lexical, syntactic, semantic and discourse levels. For example, some features such as type-token ratio, character/word n-grams (frequency of n consecutive units), average sentence length, average word length, and consistent misspellings (see Neal et al., 2017, for a summary). Despite the overwhelming success of deep learning in authorship identification and attribution in recent years, traditional linguistic features are still highly effective in authorship analysis (Kestemont et al., 2018, 2019, 2020, 2021).

Yet the wide application of machine learning and deep learning in recent years has greatly advanced the state-of-the-art performance in authorship verification (Boenninghoff et al., 2019a,b; Weerasinghe and Greenstadt, 2020), as deep neural networks can model the complex interactions between linguistic elements more than other traditional statistical methods. For example, Weerasinghe and Greenstadt (2020) and Weerasinghe et al. (2021) trained deep neural networks on linguistic feature vectors, greatly improving the accuracy of authorship verification. Yet the end-to-end neural networks have yielded the state-of-the-art results in authorship verification (e.g., Boenninghoff et al., 2019a,b, 2021; Peng et al., 2021; Futrzynski, 2021). In the end-to-end approach, raw texts are directly fed into a neural network without any feature extractions, and the neural network predicts the authorship of the text or outputs textual similarity between two texts. This state-of-the-art performance of end-to-end models show that neural networks are extremely good at figuring out implicitly the linguistic features related to one's unique style directly from texts, better than the handcrafted features written by linguists and stylometrists with domain knowledge. This is because many complex or long-range interactions between linguistic elements still cannot be easily captured by current knowledge of linguistic styles. Despite the impressive accuracy, interpreting neural networks with tens of millions of parameters is very challenging (Belinkov et al., 2020; Alishahi et al., 2019), making it hard to understand which linguistic elements contribute to the unique individual styles. Recent PAN Authorship Verification shared tasks suggest that characterizing individual styles in long texts can be solved with almost perfect accuracy (Kestemont et al., 2020); as a result, stylometric studies have increasingly focused on short texts in social media or online communications for authorship profiling or verification (Brocardo et al., 2013; Vosoughi et al., 2015; Boenninghoff et al., 2019b). Yet studies on the authorship of short social media texts are still limited. Prior studies in computational stylometry have demonstrated the power of statistical methods in analyzing individual styles. Building on the statistical and deep learning paradigm, I make use of deep neural networks as a method to extract representations of idiolectal styles from short texts. While the neural networks are difficult to interpret, I also develop a set of probing tasks to interpret them, tasks that use controlled textual stimuli to understand the behaviors of neural networks in response to the texts. This study points to a promising application for authorship detection in social media, as the proposed models can extract idiolect variation accurately in texts as short as 100 tokens.

3.3 Learning representations of idiolects

In forensic linguistics, textual similarity across authors has traditionally been quantified as the proportion of shared vocabulary and the number and length of shared phrases or characters (n-grams) (Coulthard, 2004). More sophisticated statistical methods to perform textual comparison have been developed over the years, notably deep learning methods (Neal et al., 2017; Kestemont et al., 2020).

To learn representations of idiolectal style, I propose using a proxy task of authorship verification, where, given two input texts, a model must determine if they were written by the same author or not (Neal et al., 2017). The identification is performed by scoring the two texts under comparison with a linguistic similarity measure and, if this measure exceeds a certain threshold, the two texts are judged to be written by the same author.

3.3.1 Task overview

In this task, deep neural networks are trained to identify authorship by learning from a large collection of texts. The texts are collected from online sources including Amazon reviews and Reddit posts, all of which are labeled with authorship. A neural network converts texts into numerical representations (vectors, or embeddings) and compares the distance between vectors for different texts. The distance between two text embeddings can be viewed as a proxy score of style similarity between two original texts. Models are trained in the contrastive learning paradigm, in which models learn the nuances of styles by comparing a massive amount of texts. In contrastive learning, the goal is to compare a text to another text by the same author (positive sample) and to many other texts by different authors (negative samples), such that neural networks learn to maximize the similarity between the same-author text pairs and minimize the similarity between different-author text pairs. Neural networks can only learn to do so by figuring out the statistical regularities inherent in idiolectal style. The loss function or the objective function guides the model during training by penalizing the errors models make and models are trained to minimize the penalty or minimize the loss. Training the model usually involves many parameters to set and the common practice is to run a model with different combinations of these hyperparameters to select the model with the best performance (hyperparameter tuning). The detailed mathematical formulation of the above description can be found below.

3.3.2 Task definition

Given a collection of text pairs by multiple authors $\mathcal{X} = \{(\mathbf{x}_1^{p_1}, \mathbf{x}_2^{p_1}), \ldots, (\mathbf{x}_{n-1}^{p_{t-1}}, \mathbf{x}_n^{p_t})\}$ from domain $\mathcal{P} = \{p_1, p_2, \ldots, p_t\}$ and labels $\mathcal{Y} = \{y_i, y_2, \ldots, y_n\}$, I aim to identify a function f_θ that can determine whether two text samples $\mathbf{x}_i$ and $\mathbf{x}_j$ are written by the same author ($y = 1$) or different authors ($y = 0$).

3.3.3 Stylometric similarity learning

My model for extracting stylistic embeddings from input texts is the same as the Sentence RoBERTa or BERT network (SBERT/SRoBERTa) (Reimers and Gurevych, 2019). For a text pair $(\mathbf{x}_i, \mathbf{x}_j)$, the Siamese model f_θ maps both text samples into embedding vectors $(\mathbf{z_i}, \mathbf{z_j})$ in the latent space such that $\mathbf{z_i} = f_\theta(\mathbf{x}_i)$ and $\mathbf{z_j} = f_\theta(\mathbf{x}_j)$. Rather than using the `[cls]` token as the representation, I use attention pooling to merge the last hidden states $[\boldsymbol{h}_0, \ldots, \boldsymbol{h}_k]$ into a single embedding vector to represent textual style.

$$\begin{aligned} \boldsymbol{h}_o =& \text{AttentionPool}([\boldsymbol{h}_0, \ldots, \boldsymbol{h}_k]) \\ \boldsymbol{z} =& \boldsymbol{W}_1 \cdot \sigma(\boldsymbol{W}_2 \cdot \boldsymbol{h}_o + \boldsymbol{b}_2) + \boldsymbol{b}_1 \end{aligned} \tag{3.1}$$

where W_1 and W_2 are learnable parameters and $\sigma(\cdot)$ is the ReLU activation function. Stylometric similarity between the text pair is then measured by a distance function $d(\mathbf{z}_i, \mathbf{z}_j)$. Here I mainly consider the cosine similarity.

The underlying models are RoBERTa (Liu et al., 2019) and BERT (Devlin et al., 2019). Specifically, I used the `roberta-base` or `bert-base-uncased` as the encoder.[2]

3.3.4 Loss function

The classic max-margin loss for deep metric learning was shown to be effective in previous work on stylometry (Boenninghoff et al., 2019a,b). Inspired by Kim et al. (2020), I used a continuous approximation of the max-margin loss to learn the stylometric distance between users. The additional hyperparameter in this loss allows the fine-grained control of the penalty magnitude for hard samples, or confusable samples.

The loss function is an adaptation from the proxy-anchor loss proposed by Kim et al. (2020). Given a text pair $\{\mathbf{x}_i, \mathbf{x}_j\}$, stylometric similarity between the text pair is then measured by a distance function $d(\mathbf{z}_i, \mathbf{z}_j)$. Here I mainly consider the cosine similarity. To minimize the distance for same-author pairs and maximize the distance between different-author pairs, the model was trained with a contrastive loss with pre-defined margins $\{\tau_s, \tau_d\}$ for the set of positive samples P^+

[2] The performance of `bert-base-cased` was almost identical to `bert-base-uncased`, so it was not included in the main results.

and negative samples P^- is given below.

$$\mathcal{L}_{(s)} = \frac{1}{|P^+|} \sum_{i,j \in P^+} \text{Softplus}\Big(\text{LogSumExp}\big(\alpha \cdot [d(\mathbf{z}_i, \mathbf{z}_j) - \tau_s]\big)\Big) \tag{3.2}$$

$$\mathcal{L}_{(d)} = \frac{1}{|P^-|} \sum_{i,j \in P^-} \text{Softplus}\Big(\text{LogSumExp}\big(\alpha \cdot [\tau_d - d(\mathbf{z}_i, \mathbf{z}_j)]\big)\Big) \tag{3.3}$$

$$\mathcal{L} = \mathcal{L}_{(s)} + \mathcal{L}_{(d)} \tag{3.4}$$

where Softplus(z) $= \log(1 + e^z)$ is a continuous approximation of the max function. α is a scaling factor that scales the penalty of the out-of-the-margin samples. The out-of-margin samples are exponentially weighted through the *Log-Sum-Exp* operation such that hard examples are assigned exponentially growing weights, prompting the model to learn hard samples harder.

During inference, I compare the textual distance $d(\mathbf{x}_1, \mathbf{x}_2)$ with the threshold τ_t, the average of the two margins, $\tau_t = \frac{\tau_s + \tau_d}{2}$. I set $\{\tau_s = 0.6, \tau_d = 0.4\}$ and $\alpha = 30$. Details about hyperparameters searching can be found in Appendix 3.3.5.

3.3.5 Hyperparameter tuning

In this section, I report the results from the hyperparameter tuning process. Table 3.1 reports some additional results obtained during hyperparameter tuning. Changing the masking probability or the margins for the contrastive loss has an impact on the final accuracy. This search is manual and not exhaustive. I used the best parameters in the main text.

For actual implementation, I used an effective batch size of 256. The default optimizer was the Adam optimizer with a learning rate of $1e - 5$. All models were trained on a single Nvidia V100 GPU with 16GB memory. The models were set to train for 5 epochs but I applied early stopping when the validation accuracy stopped to increase. Each epoch took about 2 hours to complete. For each model, I limited the maximum length of text samples to 100 tokens but the actual definition of tokens depended on the tokenizer used.

3.4 Baseline methods

I also compare the current models against several baseline methods in authorship verification.

1. **GLAD**. Groningen Lightweight Authorship Detection (GLAD) (Hürlimann et al., 2015) is a binary linear classifier using multiple handcrafted linguistic features.

Model	**Accuracy**	**F1**	**AUC**
$\tau_s = 0.7$, $\tau_d = 0.3$, $\alpha = 30$			
SRoBERTa - Amazon	0.744	0.676	0.901
SRoBERTa - Reddit	0.723	0.713	0.804
$\tau_s = 0.8$, $\tau_d = 0.2$, $\alpha = 30$			
SRoBERTa - Amazon	0.809	0.814	0.889
SRoBERTa - Reddit	0.708	0.722	0.784
$\tau_s = 0.8$, $\tau_d = 0.2$, $\alpha = 30$			
SRoBERTa - Amazon	0.809	0.814	0.889
SRoBERTa - Reddit	0.708	0.722	0.784
$\tau_s = 0.6$, $\tau_d = 0.4$, $\alpha = 10$			
SRoBERTa - Amazon	0.822	0.826	0.903
SRoBERTa - Reddit	0.73	0.734	0.81
$\tau_s = 0.6$, $\tau_d = 0.4$, $\alpha = 5$			
SRoBERTa - Amazon	0.819	0.825	0.901
SRoBERTa - Reddit	0.72	0.723	0.798

Table 3.1: Results of authorship verification on the develop / test sets. $\tau_s = 0.6$, $\tau_d = 0.4$ and $\alpha$are hyperparameters of the model.

2. **FVD**. The Feature Vector Difference (FVD) method (Weerasinghe and Greenstadt, 2020) is a deep learning method for authorship verification using the absolute difference between two traditional stylometric feature vectors.

3. **AdHominem**. AdHominem is an LSTM-based Siamese network for authorship verification in social media domains (Boenninghoff et al., 2019a).

4. **BERT$_{\textbf{Concat}}$/RoBERTa$_{\textbf{Concat}}$**. The model is fine-tuned BERT/RoBERTa by concatenating two texts under comparison. Authorship was determined by performing the binary classification task on the `<cls>` token (Ordoñez et al., 2020).

When setting up these models, I tried to make minimal changes to the original implementation. Details of changes are provided below.

3.4.1 GLAD

I used the original code for GLAD[3]. The linguistic features were extracted using the `combo4` options in the code, which covers 23 linguistic features. While support vector machine (SVM) was used as the classifier in their paper (Hürlimann et al., 2015), I found SVM did not scale to the size of the current data. Instead, I ran a logistic regression model on the features as this allowed me

[3] https://github.com/pan-webis-de/huerlimann15

to interpret the feature importance. The performance of logistic regression was very close to the Random Forest classifier in their code.

3.4.2 FVD

The model was trained with the code released by the original authors[4] (Weerasinghe and Greenstadt, 2020). I kept the original feature extraction methods and the model architecture. The input to the neural network was a 2314-dimensional feature vector, which was computed by taking the absolute difference between the linguistic feature vectors of the two authors under comparison. The two-layered fully connected neural network was trained for 100 epochs and the model with the best validation accuracy was kept.

3.4.3 AdHomenin

I used the original implementation[5] provided by the author. While this implementation was slightly different from that described in Boenninghoff et al. (2019a), no modification was made to the code other than adapting the code to work on the current data. The same pre-processing method, model architecture, and parameters were kept. Yet the evaluation code was not used as it ignores uncertain samples, which is a standard practice in PAN 20 (Kestemont et al., 2020). The model was trained for 5 epochs and I only kept the model with the best validation results.

3.4.4 BERT$_{\textbf{Concat}}$/RoBERTa$_{\textbf{Concat}}$

I mostly followed the description by Ordoñez et al. (2020). While Ordoñez et al. (2020) used a variant of RoBERTa known as Longformer (Belainine et al., 2020), I re-implemented the model using the original pre-trained RoBERTa, so that the model can be directly compared to the Siamese version. Since the Longformer is highly similar to RoBERTa and BERT, I do not expect a significant performance gap between them.

3.4.5 Evaluation metrics

To ensure consistency, all evaluation metrics were computed by the functions in `Sklearn`: `accuracy_score` for accuracy, `F1_score` for F1 and `roc_auc_score` for AUC.

[4] https://github.com/pan-webis-de/weerasinghe20

[5] https://github.com/boenninghoff/AdHominem

3.5 Data

3.5.1 Amazon reviews

This dataset was extracted from the release of the full Amazon review dataset up to 2018 (Ni et al., 2019). I filtered out reviews that were shorter than 50 words to ensure sufficient text to reveal stylistic variation. I only retained users that have reviews in at least two product domains (e.g., Electronics and Books) and at least five reviews in each domain. After text cleaning, the dataset contained 128,945 users. I partitioned 40%, 10%, and 50% of users into training, development, and test sets. There were 51398, 12849, and 64248 unique users in each set respectively. As one of the goals is to analyzed stylistic variation, I reserved the majority of the data (50% of users) for model evaluation and for subsequent linguistic analysis. The maximum length of all samples was limited to 100 tokens.

3.5.2 Negative sampling

For each user, I randomly sampled six pairs of texts written by the same user as positive samples of same-authorship (SA). For negative samples, I randomly sampled six texts from the rest of the data and paired them with the original text. In order to improve generalizability across domains, I enforced a sampling scheme that half of the positive/negative samples were matched in domain while the other half were cross-domain.

3.5.3 Reddit posts

To test generalizability, I additionally constructed a second dataset from the online community Reddit and ran a subset of experiments on it. The top 200 subreddits were extracted via the `Convokit` package (Chang et al., 2020a). Only users who had posted texts longer than 50 words in more than 10 subreddits were selected in the dataset, resulting in 55,368 unique users. I partitioned the 60%, 10%, and 30% of users as training, development, and test sets respectively. Each user's idiolect is represented by 10 posts from 10 different subreddits. The binary labels were then generated by randomly sampling from SA and DA pairs, using the same negative sampling procedure used in the creation of the Amazon dataset.

3.6 Evaluation on the proxy task

3.6.1 Results

To test whether the current model does recover stylistic features, I first test its performance on the proxy task of author verification, contextualizing the performance with other models specifically designed for that task. The performance of authorship verification is evaluated by accuracy, F1 and the area-under-the-curve score (AUC) (Kestemont et al., 2020). F1 measures the performance of a classifier by weighting true positive, true negative, false positive and false negative altogether, and is therefore a better metric than accuracy, especially when the data is imbalanced. A more detailed definition of F1 is given in Section 4.8. Similarly, AUC measures the tradeoff between true positives and false positives, and it provides a comprehensive evaluation of performance across all possible classification thresholds. If the AUC is closer to 1, it means the model performance is better at predicting authorship. The results in Table 3.2 suggest that all models are able to at least recover some distinctive aspects of individual styles even in these short text. Deep learning-based methods generally achieve better verification accuracy than GLAD (Hürlimann et al., 2015) and FVDNN (Weerasinghe and Greenstadt, 2020), models based on traditional linguistic features such as POS tags, average sentence length, lexical diversity, ngrams and tf-idf features. Siamese architecture demonstrates its usefulness in the authorship verification task, as AdHominem (Boenninghoff et al., 2019a) and SRoBERTa perform better than the pre-trained transformer RoBERTa$_{\text{Concat}}$. These results confirm that models recognize authors' stylistic variation. The error analysis also shows that identifying the same author across domains or different authors in the same domain poses a greater challenge to these models in general, though different model choices may exhibit various inductive biases. As SRoBERTa is shown to be the most effective architecture in this task, I will use these models to examine idiolectal variation for the rest of the study.

3.6.2 Error analysis

I also analyzed the error distributions across different conditions, shown in Table 3.3. Given a pair of texts, I categorized them into four different categories, same-author (SA)/different author (DA), or same domain (SD)/different domain (DD). Unsurprisingly, most methods still struggling with SA-DD and DA-SD pairs, suggesting that domain-specific/topic information partially interferes with the extraction of writing styles. The cases with RoBERTa$_{\text{Concat}}$ and BERT$_{\text{Concat}}$ are particularly interesting, as both models consistently performed worse on SA pairs but outperformed the rest of the models on DA pairs. Cosine distance-based models seem to better balance the trade-off across conditions. This shows that model architectures also exhibit inductive biases of their own, which

Model	Accuracy	F1	AUC
Amazon reviews			
Random	50%	0.5	0.5
GLAD	67.1%	0.667	0.738
FVDNN	65.1%	0.671	0.714
AdHominem	73.3%	0.781	0.811
BERT$_{Concat}$	71.3%	0.664	0.802
SBERT	76.7%	0.768	0.850
RoBERTa$_{Concat}$	73.9%	0.686	0.838
SRoBERTa	**82.9%**	**0.831**	**0.909**
Reddit posts			
BERT$_{Concat}$	65.0%	0.600	0.721
SBERT	66.3%	0.669	0.727
RoBERTa$_{Concat}$	71.0%	0.695	0.794
SRoBERTa	**73.0%**	**0.737**	**0.812**

Table 3.2: Results of authorship verification on the Amazon (top) and Reddit (bottom) test samples

Model	Accuracy			
	SA-SD	SA-DD	DA-SD	DA-DD
GLAD	73.7%	58.4%	62.9%	73.6%
FVDNN	75%	67.5%	55.7%	62.3%
AdHominem	81.2%	73.1%	63.6%	75.4%
BERT$_{Concat}$	61.8%	51.6%	85.3%	86.5%
SBERT	84.1%	77.2%	67.7%	78%
RoBERTa$_{Concat}$	64.4%	49.6%	89.2%	92.5%
SRoBERTa	88.4%	82.5%	74.8%	83.2%

Table 3.3: Accuracy by domain and authorship. All experiments were run on the Amazon reviews.

may bias them to be more or less effective in certain conditions.

3.7 Linguistic Analysis

In this section, I seek to quantify the idiolectal variation at different linguistic levels.

3.7.1 Ordering

Hierarchical models of linguistic identities hold that authorial identities are reflected at all linguistic levels (Herring, 2004; Grant, 2012; Grant and MacLeod, 2020), yet the relative importance of these elements is seldom empirically explored. In order to understand the contributions of lexical distribution, syntactic ordering, or discourse coherence, I test the contributions of different linguistic features to authorship verification by perturbing the input texts. To force the model to only use lexical information, I randomly permute all tokens in the data, removing information about syntactic ordering (`lexical model`). The organization of discourse might also provide cues to idiolectal style. To test this, I preserve the word order within a sentence but permute sentences within the text to disrupt discourse information (`lexico-syntactic model`). Then I ran experiments on these datasets using the same set of hyperparameters to compare the model performance on these perturbing inputs.

- *Lexical permutation*
  - **Original sentence**: "Music is very good and provides great atmosphere for the music tracks."
  - **Permuted sentence**: "provides is the Music very atmosphere tracks. great very good and for"
- *Sentence permutation* (Sentence numbers were only added for illustration and were not present in the original review.)
  - **Original paragraph**: "[1] Elden Ring is an RPG that features deep systems and an open world that only gets better the longer you play. [2] It starts with basics; you are low level with shabby gear. [3] You can't even level up until you meet someone. [4] You are a nobody. [5] This is your journey to become something more than you were at the start. [6] It's going to be hard, but everything is at first. [7] Keep trying, it gets easier as you get stronger and better equipped. [8] I am having a blast and I am terrible at Souls genre games."[6]
  - **Permuted paragraph**: "[6] It's going to be hard, but everything is at first. [7] Keep trying, it gets easier as you get stronger and better equipped. [3] You can't even level up until you meet someone. [1] Elden Ring is an RPG that features deep systems and an open world that only gets better the longer you play. [2] It starts with basics; you are low level with shabby gear. [8] I am having a blast and I am terrible at Souls genre games. [5] This is your journey to become something more than you were at the start. [4] You are a nobody."

[6] Text sample from: https://www.amazon.com/Elden-Ring-Xbox-One/product-reviews/B07SMBNTSJ/ref=cm_cr_getr_d_paging_btm_prev_3?ie=UTF8&reviewerType=all_reviews&pageNumber=3

3.7.2 Content and function words.

The use of function words has long been recognized as an important stylometric feature. A small set of function words is disproportionately frequent, relatively stable across content, and seems less under authors' conscious control (Kestemont, 2014). Yet few studies have empirically compared the relative contributions between function words and content words. To test this, I masked out all content words in the original texts with a masked token `<mask>`, which was recognized by the transformer models. For comparison, I also created masked texts with only content words. Punctuation and relative positions between words were retained as this allows the model to maximally exploit the spatial layout of content/function words.

For example, a sentence can be manipulated as follows.

- *Masking content words*
  - **Original sentence**: "Just put in about fifty hours and so far Elden Ring is very impressive."
  - **Masked sentence**: "Just `<mask>` in about `<mask>` `<mask>` and so `<mask>` `<mask>` `<mask>` is very `<mask>`."
- *Masking function words*
  - **Original sentence**: "Just put in about fifty hours and so far Elden Ring is very impressive."
  - **Masked sentence**: "`<mask>` put `<mask>` `<mask>` fifty hours `<mask>` `<mask>` far Elden Ring `<mask>` `<mask>` impressive."

3.7.3 Results

While the importance of lexical information in authorship analysis has been emphasized, it is suggested that only using lexical information is insufficient in forensic linguistics (Grant and MacLeod, 2020). The results in Table 3.4 suggest that, even with only lexical information, the model performance is only about 4% lower than for models with access to all information. Syntactic and discourse ordering do contribute to author identities, yet the contributions are relatively minor. In forensic linguistics, it is commonly the case that only fragmentary texts are available (Grant, 2012), and the findings suggest that even without broader discourse information, it is still possible to estimate author identity with good confidence. The weak contribution of discourse coherence to authorship analysis highlights that the high level organization of texts is only somewhat consistent within authors, which has been mentioned but rarely tested in forensic linguistics (Grant and MacLeod, 2020). Yet it is possible that the short length of texts ($<= 100$) limits the utility of coherence information.

		Amazon		**Reddit**	
		Test (Permuted)	Test (Original)	Test (Permuted)	Test (original)
Lexical	SBERT	0.716 / 0.798	0.736 / 0.794	0.639 / 0.668	0.657 / 0.667
	SRoBERTa	0.781 / 0.856	0.751 / 0.850	0.654 / 0.734	0.686 / 0.714
Lexico-syntactic	SBERT	0.767 / 0.851	0.770 / 0.852	0.681 / 0.722	0.685 / 0.724
	SRoBERTa	**0.820 / 0.895**	**0.822 / 0.896**	**0.730 / 0.798**	**0.732 / 0.800**

Table 3.4: Results (F1/AUC) on permuted examples show that models are largely insensitive to syntactic and ordering variation, and, instead, idiolect is mostly captured through lexical variation.

		Amazon	Reddit
Function	SBERT	0.775 / 0.856	0.680 / 0.744
	SRoBERTa	0.786 / 0.858	0.683 / 0.755
Content	SBERT	0.735 / 0.812	0.641 / 0.689
	SRoBERTa	**0.795 / 0.870**	**0.708 / 0.768**

Table 3.5: Results (F1/AUC) show that function or content words alone are reliable authorial cues. For SRoBERTa, content words seem to convey slightly more idiolectal cues despite topical variation.

From Table 3.5, it is apparent that, even with half of the words masked out, the transformed texts still contain an abundance of reliable stylometric cues to individual writers, such that the overall accuracy is not significantly lower than models with full texts. While the importance of function words in authorship analysis has been emphasized (Kestemont, 2014), content words seem to convey slightly more idiolectal cues despite the topical variation. Both SBERT and SRoBERTa achieve similar performance on Amazon and Reddit data, yet SRoBERTa better exploits the individual variation in content words. These results strongly suggest that there are unique individual styles that are stable across topics, and the additional probing also reveals that topic information is significantly reduced in the learned embeddings.

3.8 Analysis of tokenization methods

I hypothesized that the large performance gap between BERT and RoBERTa ($\tilde{}5\%$) could be caused by the discrepancy in the tokenization methods. Both models share the same transformer architecture but they differ in how they encode words.

These tokenization methods are illustrated with the examples below.

- *Word-based tokenization*
  - **Original sentence**: I trained several lightweight Siamese LSTM models from scratch. →
  - **Tokenized sentence**: [we, trained, several, lightweight, `<UNK>`, `<UNK>`, models, from, scratch, .]
- *BERT-base-uncased tokenization*
  - **Original sentence**: I trained several lightweight Siamese LSTM models from scratch. →
  - **Tokenized sentence**: [`[CLS]`, we, trained, several, lightweight, siam, ##ese, l, ##st, ##m, models, from, scratch, ., `[SEP]`]
- *BERT-base-cased tokenization*
  - **Original sentence**: I trained several lightweight Siamese LSTM models from scratch. →
  - **Tokenized sentence**: [`[CLS]`, We, trained, several, lightweight, Si, ##ame, ##se, L, ##ST, ##M, models, from, scratch, ., `[SEP]`]
- *RoBERTa-base tokenization*
  - **Original sentence**: I trained several lightweight Siamese LSTM models from scratch. →
  - **Tokenized sentence**: [`<s>`, We, `<space>`trained, `<space>`several, `<space>`lightweight, `<space>`Si, ames, e, `<space>`L, ST, M, `<space>`models, `<space>`from, `<space>`scratch, ., `</s>`]

In the word-based tokenization, which dominated NLP in the past decades, a sentence is segmented into individual words from the vocabulary. While the method is simple and intuitive, it suffers from inefficiency, as not all words can be encoded in a pre-defined vocabulary and having a large vocabulary takes up a huge amount of memory. In the example above, 'Siamese' and 'LSTM' are treated as out-of-vocabulary tokens (represented as `<UNK>`). Yet failure to encode these two key words could change the sentence style and meaning significantly. The current method to handle the out-of-vocabulary tokens is the Byte Pair Encoding (BPE) tokenization (Sennrich et al., 2016), which learns the best segmentation units directly from the data without human supervision. The learned units are generally called ‘wordpiece’ or ‘subwords’, units that sometimes only consist of part of a word. For example, the `RoBERTa-base` tokenizer tokenizes the word “Siamese” into a list of subwords: “Si”, “ames” and “e”. The use of subwords can reduce vocabulary size and effectively eliminate out-of-vocabulary tokens, as any unknown words can be segmented into character strings. Yet subword units might not be linguistically meaningful units, which could potentially make learning less efficient (though this does not seem to be a problem for large transformers). The BPE tokenizer preseves the word boundaries by attaching ## to signify word internal units or using space marker `<space>` to indicate word beginning. In the example above, both BERT and RoBERTa dynamically tokenize textual inputs into subword units, while keeping most words as their original forms. While `BERT-base-uncased` converts all words into lower case, `BERT-base-cased` and `RoBERTa-base` preserves the casing information.

While both BERT and RoBERTa use the BPE tokenization, they still differ in their actual implementation. The BERT tokenizer is learned after preprocessing the texts with heuristic rules (Devlin et al., 2019), whereas the BPE tokenizer for RoBERTa is learned without any additional preprocessing or tokenization of the input (Liu et al., 2019). As a result, the RoBERTa tokenizer should be able to encode more textual variation and language-external variation, such as formatting, emojis and line breaks. The additional variation encoded could potentially contribute to the identification of individual styles. In the next section, I test this hypothesis by controlling for model hyperparameters and training data.

3.8.1 Method

To verify the role of tokenizers, I trained several lightweight Siamese LSTM models from scratch that only differed in tokenization methods: 1) word-based tokenizer with the vocabulary size set to either 30k or 50k to match the sizes of BPE encodings; 2) pre-trained wordpiece tokenizer for `bert-base-uncased` and `bert-base-cased` with 30522 and 28996 wordpieces in the vocabulary respectively; 3) pre-trained tokenizer for `roberta-base` with a vocabulary size of 50265.

3.8.1.1 Tokenization methods

For the word-based tokenization, I made use of the `word_tokenizer` function in `NLTK`. Either 30k or 50k most frequent lexical tokens were kept as the vocabulary for training the LSTM model, plus a padding token and an OOV token. As for the BPE tokenizer, I directly used the pre-trained tokenizers for BERT and RoBERTa accessed through HuggingFace’s `Transformers` (Wolf et al., 2020).

3.8.1.2 Model specification

The underlying model is an LSTM-based Siamese network. The model consists of two bidirectional LSTM layers with 300 hidden states for each direction. The last hidden states of the last layer in both forward and backward directions were concatenated as the representation of the whole input text, which was then passed to a two-layer fully connected network with 300 hidden states in each layer. The similarity between the two paired texts was computed with the cosine distance function. The hyperparameters for the loss function were $\tau_d = 0.4$ and $\tau_s = 0.6$. No pre-trained word embedding weights were used and all weights were trained from scratch.

3.8.2 Training details

The training, development, and test data were the same as those in the main experiments. The model was optimized by the Adam optimizer with a learning rate of 0.001. In order to avoid gradient explosion and gradient vanishing, gradient clipping was applied to stabilize training with the maximum gradient norm set to 1. The model was trained on an 11GB RTX 2080Ti with an effective batch size of 256 for 10 epochs. The average training time for each model was about 3 hours.

Tokenization	**Accuracy**	**F1**	**AUC**
`Word-30k`	67.8%	0.675	0.745
`Word-50k`	67.8%	0.679	0.744
`BERT-uncased`	67.1%	0.680	0.737
`BERT-cased`	67.2%	0.677	0.737
`RoBERTa`	**73.1%**	**0.734**	**0.804**

Table 3.6: The effect of tokenization methods on the model performance with respect to Amazon reviews. The pre-trained RoBERTa BPE tokenizer encodes more textual variations than the rest.

3.8.3 Results

As shown in Table 3.6, the RoBERTa tokenizer outperforms other tokenizers by a substantial margin, even though it has similar numbers of parameters to `Word-50k`. Interestingly, the pre-trained BERT tokenizer is not superior to the word-based tokenizer, despite better handling of out-of-vocabulary (OOV) tokens. For word-based tokenizers, increasing the vocabulary from 30k to 50k does not bring any improvements, indicating that many tokens were unused during evaluation. The results strongly suggest that choosing the expressive tokenizer, such as the RoBERTa BPE tokenizers directly trained on raw texts, can effectively encode more stylistic features. For example, for the word `cat`, the RoBERTa tokenizer gives different encodings for its common variants such as

`Cat`, `CAT`, `[space]CAT` or `caaats`, but these are all treated as the OOV tokens in word-based tokenizations. While the BERT tokenizer handles most variants, it fails to encode formatting variation such as `CAT`, `[space]CAT` and `[space][space]CAT`. Such nuances in formatting are an essential dimension of idiolectal variation in forensic analysis of electronic communications (Grant, 2012; Grant and MacLeod, 2020).

3.9 Characterizing sociolects

Language varies at both individual and collective levels (Eckert, 2012). Previous studies that have linked textual features to social variables found that textual readability measures are reliably associated with socio-economic status (Flekova et al., 2016; Basile et al., 2019). While those studies successfully predict social variables directly from texts, I investigate how social variation is encoded in idiolectal embeddings, numerical representations of individual styles. Diagnostic classification is employed to probe to what extent the collective language variation is retained in the stylistic embeddings and whether the style-independent topical information is removed in the embeddings. In diagnostic classification, I use a classifier to predict the socio-economic statuses of the authors using their style representations learned by the authorship verification model. The classification accuracy (e.g. how accurately the classifier can predict the socio-economic statuses) can be used to detect how much social information is encoded in the idiolectal representations. Higher accuracy implies that more social information is preserved in the style representations.

3.9.1 Dataset compilation

From the test set, I created a small subset of high socioeconomic status (SES) users and low SES users by using the prices of the reviewed products as a proxy. I verified that there is a clear distinction in readability between high SES and low SES groups, which has previously been found to be a reliable linguistic indicator of SES (Flekova et al., 2016; Basile et al., 2019). This dataset is used to probe sociolect in idiolect representations.

I compiled this sociolect dataset as a subset of the test data, which contained unseen speakers and samples by the trained model. The core idea is to select users that fall into distinct socioeconomic statuses by utilizing the price tag of their reviewed products. If a user consistently reviews expensive products, it is more likely that this user is associated with high socioeconomic status. This method has been previously used in a study that surveyed socio-economically related variations (Basile et al., 2019).

The meta-information was provided together with the original Amazon dataset.[7] For each prod-

[7] `http://deepyeti.ucsd.edu/jianmo/amazon/index.html`

uct, I acquired the product title and its price from the product meta-information based on its unique identifier. However, the meta-information was incomplete for a sizable fraction of data, either missing certain attributes or in the wrong format. I only kept the products with complete meta-information.

Then for each product domain, I discretized the price distribution by categorizing product prices into ten quantiles. As a proxy metric for price ranking, the quantile into which a product fell was used as an approximation of the relative expensiveness of the product. This was done for each domain separately rather than for the whole dataset, otherwise, a few domains such as appliances, luxury products, or electronics will dominate the tail of the distribution. After categorizing the data, I averaged the rankings of all the products associated with a user, the result of which was treated as an approximation of a user's socioeconomic status. I kept the top 10% and the bottom 10% of users as high SES users and low SES users respectively, so as to maximize the differences between these two groups. This approach resulted in 6567 users with 72335 reviews in the high SES group and 6939 users with 79190 reviews in the low SES groups. The dataset is relatively balanced (48% vs. 52%) so I did not further resample the data. The distribution of product domains is displayed in Figure 3.1.

3.9.2 Label validation

The dataset is created using meta-information associated with each review, but there is no guarantee that the extracted socio-economic statuses reflect the real socio-economic statuses of the individual authors. Precautions must be taken to verify these generated labels. After dataset creation, I verified that there was linguistic variation conditioned on socio-economic status in the dataset by measuring several readability metrics, linguistic indicators shown to correlate reliably with socio-economic status (Flekova et al., 2016; Basile et al., 2019). A readability score is an aggregation index of measures including average sentence lengths, syllable counts and number of high frequency words, etc[8]. The readability scores were computed by the functions provided in `textstat`[9]. The differences are all statistically significant, implying that reviews written by high SES users tend to be more linguistically complex than those by low SES users. These results are consistent with results reported in previous studies (see Table 1 in Flekova et al. (2016) and Table 3 in Basile et al. (2019)). These results in Table 3.7 imply that there are significant style differences between high SES and low SES groups. Therefore, the data and the SES labels can be used to probe the sociolects.

[8] https://quanteda.io/reference/textstat_readability.html

[9] https://github.com/shivam5992/textstat

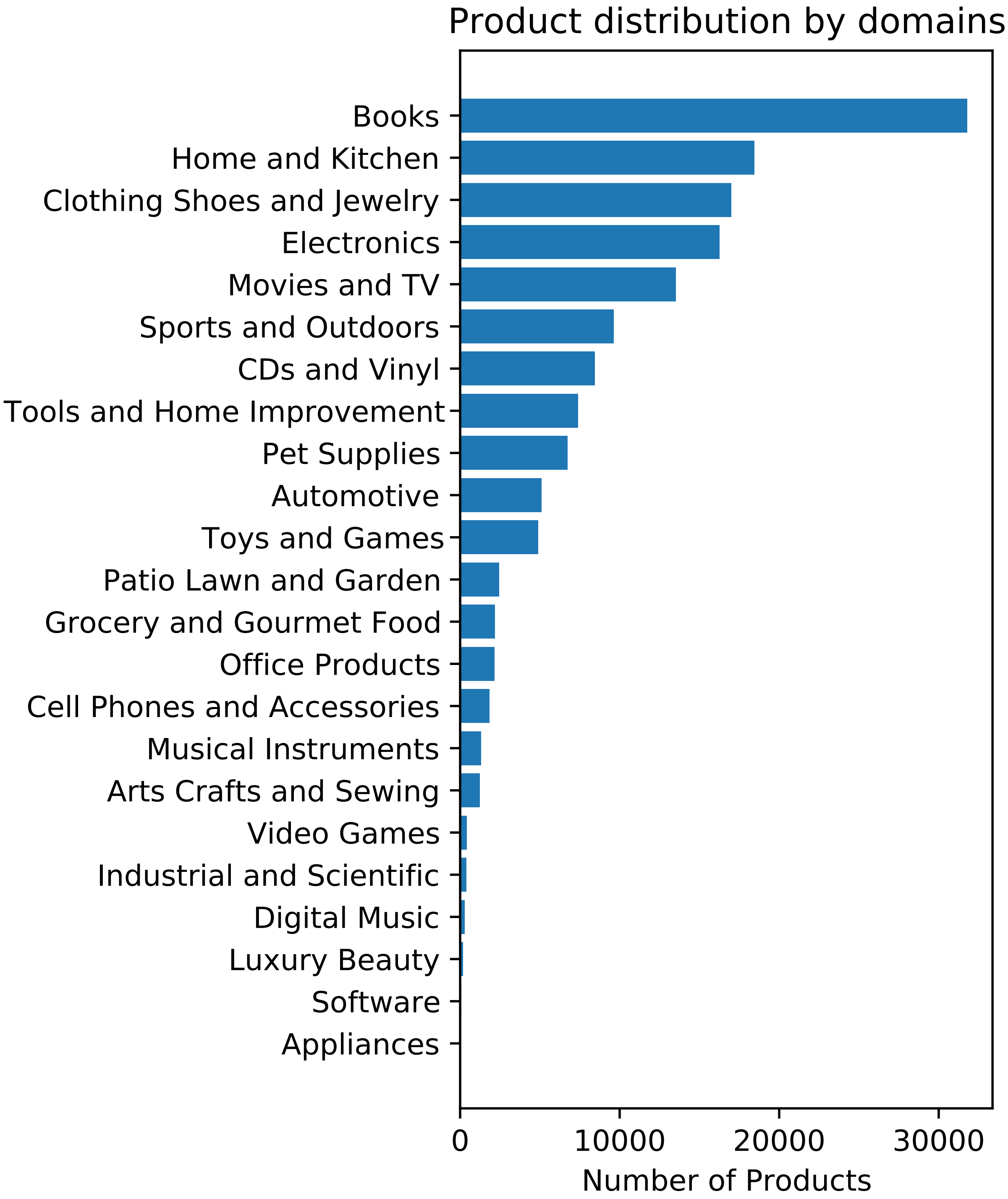


Figure 3.1: Distributions of product domains

Metrics	Low SES	High SES
ARI	11.8	13.1
Coleman-Liau	7.26	7.61
Dale-Chall	7.21	7.49
Flesch-Reading	65.69	61.19
Flesch-Kincaid	10.0	11.1
Gunning-Fog	12.04	13.14

Table 3.7: The median values of various readability metrics.

3.9.3 Experiments

Five models were trained to predict SES based on language. I also used the same models and data to predict the product domain of each short text.

- **TF-IDF**. TF-IDF, or term frequency-inverse document frequency, is a classic method to convert textual information to numerical representations for machine learning models (Ramos et al.; Aizawa, 2003; Leskovec et al., 2020). The intuition of this algorithm is that the importance of a term is inversely related to its frequency across documents. If a word is used in a specific and restricted context (or by some specific authors), then it receives a higher weight. But if a word is used widely across documents (or authors), this word is assigned less weight. TF-IDF encodes mostly topical information.

- **Handcrafted stylometric features** These features include character n-grams, POS-Tag n-grams, special characters, frequency of function words, number of characters, number of words, average number of characters per word, distribution of word-lengths, vocabulary richness (the ratio of hapax-legomenon and dis-legomenon), POS-tag chunks, NP and VP constructions (Weerasinghe and Greenstadt, 2020).

- **RoBERTa**. This method makes use of both topical and linguistic features as RoBERTa takes whole texts as input.

- **SRoBERTa embeddings**. If the proposed authorship verification learns solely idiolect representations, then this model only predicts with idiolectal features.

- **Random baseline (BL)**. In this case, the predictions are made by randomly predicting the social categories. No textual information is used by this baseline method. This is primarily used to simulate the worst case scenario.

For TD-IDF and Stylometric features, I used logistic regression as the base model. The stylometric features were extracted using the FVD method (Weerasinghe and Greenstadt, 2020), one of the baseline methods for authorship verification. For RoBERTa and SRoBERTa, I added a two-layered neural network on top of the `[cls]` token with cross-entropy loss. The only difference was that, for SRoBERTa, the base RoBERTa was frozen during training. I ran each model 3 times with different random seeds. For each time, I randomly split the data into 75% and 25% partitions for training and testing. The averaged results were reported.

Model	TF-IDF	Stylo.	RoBERTa	SRoBERTa	BL
SES	0.633	0.588	**0.644**	0.592	0.50
Domain	0.601	0.492	**0.681**	0.343	0.03

Table 3.8: F1 scores for SES and domain predictions.

3.9.4 Results

For the challenging task of SES prediction, all models attain moderate performance that is consistently above chance level, as shown in Table 3.8, echoing previous findings (Flekova et al., 2016; Basile et al., 2019). This implies that even though the style embeddings only contain partial information of the original texts (mainly the idiolectal styles), they still contain sociolectal variation. Compared to the fine-tuned RoBERTa, the idiolectal features have filtered out some SES-related variation, which could be related to domain-specific information. Notably, that the style embeddings were especially poor at predicting product domain indicates that idiolectal style is not simply capturing product domain as a proxy for SES (e.g., learning more expensive domains). The SRoBERTa's high performance on SES and low performance on domain suggest that the task setup and sampling strategy forced the model to smooth out a significant portion of variation associated with topics. As noted by Boenninghoff et al. (2019a), even if surface linguistic features are not highly content-related, they still achieve moderate performance, suggesting that variation across domains may be more than topical. The fact that SES variation is present in the idiolectal embeddings suggests that at least some SES variation is nested within idiolectal variation (Eckert, 2012).

3.10 Characterizing idiolectal styles

While the last section focuses on charactering sociolects, in this section, I turn my attention to distinctiveness and consistency in writing styles, both of which are key theoretical assumptions in forensic linguistics (Grant, 2012). Specifically, I propose two metrics to quantify the inter-author distinctiveness and the intra-author consistency based on the style representations learned by the authorship verification model.

3.10.1 Distinctiveness

Not all authors have a distinctive writing style, as there is huge heterogeneity in the dataset. I examine inter-author variation through **inter-author distinctiveness** by constructing a graph that connects users with similar style embeddings, described next. The underlying approach is that, if an author's style is found to be similar to that of many authors, their style is determined to be not distinctive, otherwise their style is distinctive, or different from many authors. Distinctiveness can be easily measured with a user-to-user network where users are connected to each other if they are similar in writing style (above a certain similarity threshold). Then I can count how many other authors a author is similar to, which is used as a proxy for distinctiveness.

For each user in the test set, I randomly sampled one text sample and extracted its embedding through the Siamese models. Then I created the pairwise similarity matrix M by computing the pairwise similarity between each text pair. Then M is pruned by removed entries below a threshold τ_{cutoff}, the same threshold τ_t that is used to determined SA or DA pairs. The pruned matrix $\hat{M}$ is treated as the graph adjacency matrix from which a network G is constructed.

$$S_i = 1 - \frac{\sum_j^N \mathbb{I}[j \in V_i]}{N} \tag{3.5}$$

where V_i is the set of neighbors of node i in G, N the total node count, and $\mathbb{I}[\ ]$ the indicator function. $\sum_j^N \mathbb{I}[j \in V_i]$ is the degree centrality of node i. I found that features from the unweighted graph are perfectly correlated with the ones from the weighted graph. The unweighted graph is kept for computational efficiency. The scores were averaged over 5 runs. The intuition is that, since authors are connected to similar authors, the more neighbors an author has, the less distinctive their style is. A distinctiveness of 0.6 implies that this author is different from 60% of authors in the dataset.

3.10.2 Consistency

Consistency is another attribute that is important in the study of idiolect, especially in forensic linguistics. A certain degree of consistency provides the foundation for forensic linguistics, as linguists assume that one's idiolect must be consistent in style within a reasonable time window. Therefore I measured the **intra-author consistency** in styles by quantifying the self-similarity of each other's writings. If an author maintains a similar style across multiple writings, this author is considered consistent in their idiolect. To quantify this, I calculated the average similarity across all text by an author and this was done for all authors.

The self-similarity in style can be quantified by the conicity, a measure of the averaged vector alignment to mean (ATM), as in the following equation (Chandrahas et al., 2018).

$$Conicity(\mathbf{V}) = \frac{1}{|\mathbf{V}|} \sum_{\mathbf{v} \in \mathbf{V}} ATM(\mathbf{v}, \mathbf{V}) \tag{3.6}$$

$$ATM(\mathbf{v}, \mathbf{V}) = cosine\Big(\mathbf{v}, \frac{1}{|\mathbf{V}|} \sum_{\mathbf{x} \in \mathbf{V}} \mathbf{x}\Big) \tag{3.7}$$

where $\mathbf{v} \in \mathbf{V}$ is a latent vector in the set of vectors $\mathbf{V}$. The ATM measures the cosine distance between $\mathbf{v}$ to the centroid of $\mathbf{V}$ whereas the conicity indicates the overall clusteredness of vectors in $\mathbf{V}$ around the centroid. If all texts written by the same user are highly aligned around their centroid with a conicity close to 1, this suggests that this user is highly consistent in writing style.

3.10.3 Analysis

The distributions of style distinctiveness and consistency both conform to a normal distribution (Figures 3.2 and 3.3), yet no meaningful correlation exists between these two measures (`Amazon`: Spearman's ρ=0.078; `Reddit`: Spearman's ρ=0.11). In general, users are highly consistent in their writing styles even in such a large population, with an average of 0.8, much higher than that for random samples (˜0.4). Users are also quite distinctive from one another, as on average a user's style is different from 80% of users in the population pool. Yet individuals do differ in their degrees of distinctiveness and consistency, which may be taken into consideration in forensic linguistics. This is because inconsistency or indistinctiveness may weaken the strength of linguistic evidence in forensic investigations.

In Table 3.9, the least distinctive text is characterized by plain language, proper formatting, and typical content, which reflects the unmarked style of stereotypical Amazon reviews. Yet this review itself is still quite distinctive as it differs from 60% of the total reviews. The most distinctive review exhibits multiple deviations from the norm of this genre. The style is unconventional with uncapitalized letters, run-on sentences, typos, the lack of periods, and the use of colloquial alternative spellings such as "haft", "lil" and "wore", all of which make this review highly marked. For style consistency, the most consistent writers incline towards using similar formatting, emojis, and narrative perspectives across reviews, whereas the least consistent users tend to shift across registers and perspectives in writings.

More text samples of Amazon reviews with polarizing distinctiveness (as determined by different models) are given in Table 3.10 and Table 3.11 to illustrate what the model has learned. Full reviews are given in the tables, though only the first 100 words are used by the model during inference. Even on the same data, different models single out reviews with wide-ranging stylistic traits, suggesting that tokenization methods and model architecture could impose inductive biases

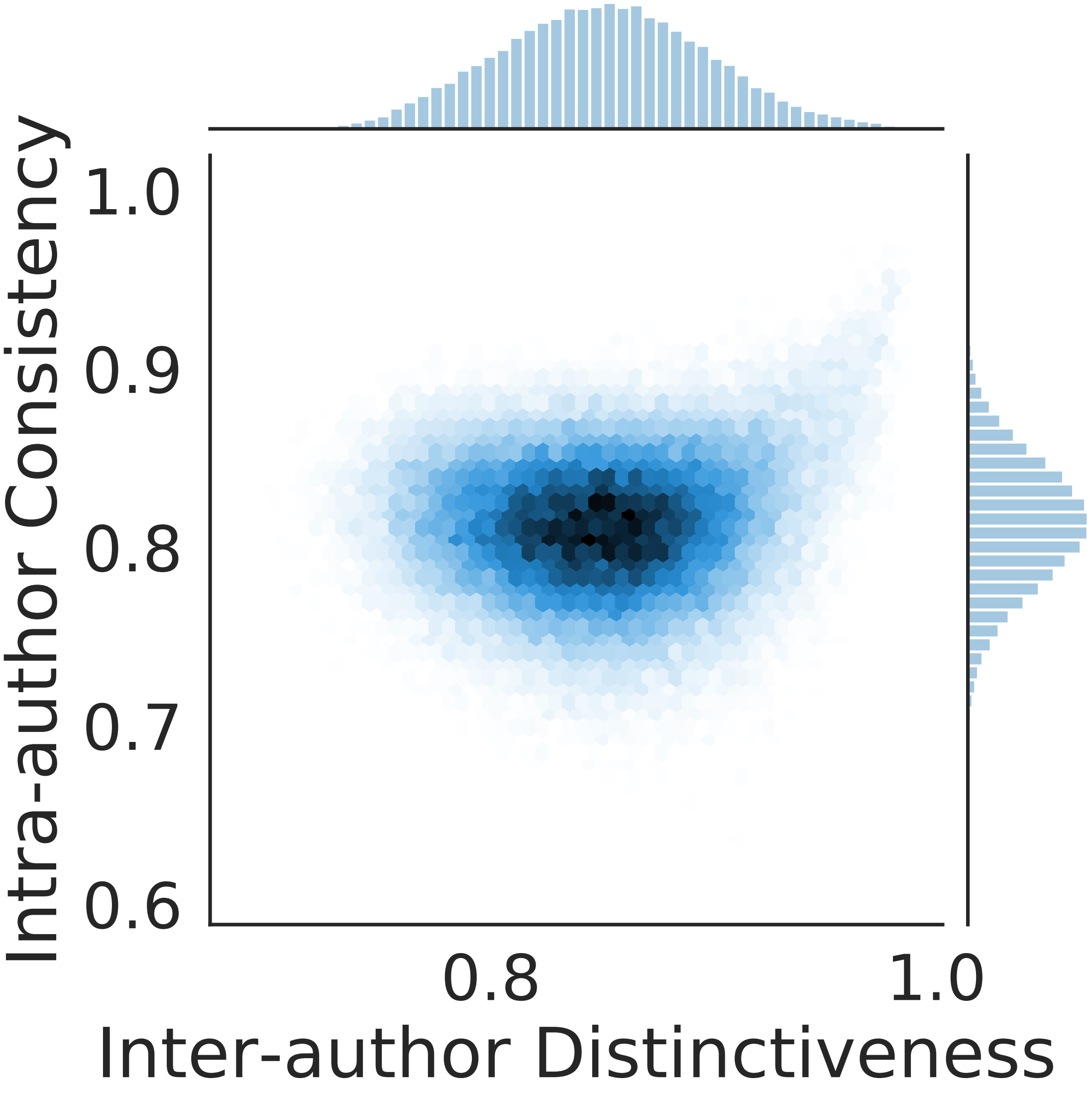


Figure 3.2: The joint distribution of distinctiveness and consistency on the Amazon reviews as computed with SRoBERTa embeddings. Both follow normal distributions yet there is no meaningful correlation between them, suggesting that these two dimensions may vary independently of each other.

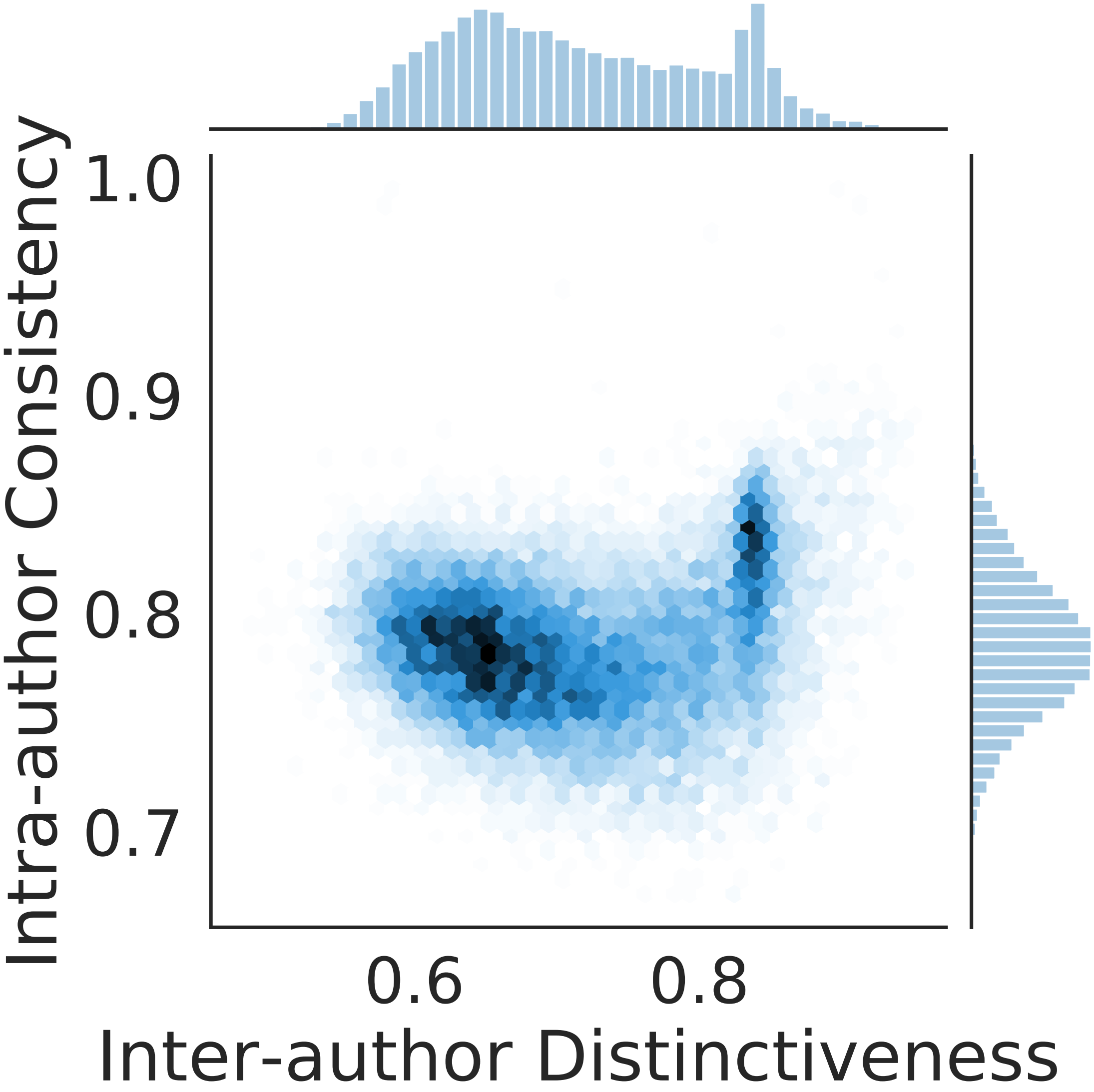


Figure 3.3: The joint distribution of distinctiveness and consistency as computed with Reddit posts.

Most distinctive	Least distinctive
its ok seems like a reprint i mean its not horrible but i was expecting a lil better qaulity but if i wore to do it again yes i would still buy this poster its not blurry or anything but if you have a good eye it seems a lil like a reprint	Nice, thinner style plates that are well suited for building Lego projects. They hold Lego pieces securely and match up perfectly. Also, as a big PLUS for this company you get amazing customer service.

Table 3.9: Sample Amazon review excerpts with the most and the least distinctive style as predicted by SRoBERTa.

on learning idiolect traits. The least distinctive reviews tend to be the reviews that conform to the norm of reviews with grammatical sentences and without unusual formatting. The most distinctive reviews often have some pretty deviant style markers, including systematic misuse of punctuation (e.g., "," for "."), frequent use of non-alphabetical symbols and random uses of capital letters (see Table 3.10 and Table 3.11).

Consistency across reviews is illustrated in Table 3.12 and Table 3.13, which showcase the text samples from the most and the least consistent authors in terms of their writing styles. For each model, each column presents reviews written by the same author. Consistent authors often maintain persistent use of certain style markers, such as the smiling emoji in Table 3.12 and the lack of punctuation in Table 3.13. For least consistent authors, they tend shift they styles drastically across reviews, e.g., formal style with complex noun phrases and embedded clauses to colloquial conversational style with short, often incomplete sentences. These samples also illustrate the challenge of identifying authorship. Some authors are inherently more inconsistent across their writings as they are more capable of switching between styles than average writers. Tracking idolectal invariance will be more challenging for these authors, especially for authors with few text samples.

I tested how various authors affect the verification performance. To avoid circular validation resulting from repeatedly using the same training data, I retrained the model with the repartitioned test data and tested them using the development set. The original test set was repartitioned into three disjoint chunks of equal size, each chunk containing authors solely from either the top, middle or bottom 33% in terms of distinctiveness or consistency. Results in Table 3.14 suggest that, while most models performed similarly, models trained on inconsistent or indistinctive authors significantly underperformed. This result may have implications for comparative authorship analysis in that it is desirable to control the number of inconsistent or indistinctive authors in the dataset.

3.10.4 Distribution of distinctiveness and consistency

Here I also show the joint distribution of distinctiveness and consistency given by SBERT in Figure 3.3. The shape of the distribution is a bivariate normal distribution and these two metrics are

Model	Most distinctive	Least distinctive
$\text{LSTM}_{\text{BERT}}$	The late 8́0s were a golden age for CD reissues, especially of tracks from the 5́0s and 6́0s, since the new digital format was just gaining popularity, there was a retro-1960s revival going on, and record companies realized they had whole new revenue stream from people buying (or re-buying) back-catalog material for their new players. The compilations issued then were full of quality stuff, unlike later bottom-of-the-barrel reissues.	This Urban Fantasy series pulls you right in and the more you know the characters the more you want to know. Hailey Edwards will make you smirk, bite your nails, cry and hope, hope, hope because her characters become (our) friends. As fantastic as the characters origins and abilities are their personalities are so appealing that I found myself hoping in the goodness of even some of the meanies.
$\text{LSTM}_{\text{RoBERTa}}$	**UPDATE 4/19/16** apparently got a bad cable Couldn't figure out why I was having issues connected to Ethernet Thought it might be a network driver issue or a modem issue But after replacing this cable with a shorter one had laying around come to conclusion its this cable that was bad Not a big deal it happens only out a couple $$$, disappointed but not to upset I need a 15 foot+ Ethernet cable it works , really not much to review ends snap in ok , no twists in cable works good	My cats don't like to be brushed. But when I can get several strokes in, this works well. I use the dog brush on my dog; the cat brush is a little smaller that the dog brush and weighs less which are good changes to make for the kitty models.

Table 3.10: Sample review excerpts with the most and the least distinctive style as given by LSTM models.

Model	Most distinctive	Least distinctive
SBERT	A Great Forza, Serafin conducts with wonderful pace, warmth and subtly for such an unsubtle opera making this a real beauty, and so easy to listen to. Callas is quite magnificent with a fine supporting cast. Disregard many of the somewhat breathless negatives, gushing with crushes and arguments for other favorite sopranos, so juvenile, the fact is there are many great female opera singers all suited to different operas some more than others, Callas happens to be one of the greatest in emotional commitment and inner depth of feeling,	So I hold a bachelors and masters in Speech Language Pathology and really have limited background in computer programing. Even with my good command of the English Language, I found this book difficult to follow and found myself rereading sections of it. I had to get through a third of the book, just to have an idea of what it was about. I learned about the history of computer programing and the need for there to be a better system for programmers and managers to communicate and produce better outcomes.
SRoBERTa	WHAT a WASTE of TIME !!! The LARGEST Funnel ... Is ” MAYBE 3 Inch WIDE ” & the TUBE Part, MIGHT can FIT a # 2 PENCIL in IT ??? The SMALL ONE has a TUBE With LESS Then 1/4 Inch ??? AS a COOK , THESE are ” A TOTAL JOKE ” (SOMETHING , I SEE at a FLEA MARKET) !!! WHAT..... ”FOODS CANYOU FIT 1/4 - 1/3 INCH OPENING ” ???	I wanted a simple steel men’s ring without a design and that wouldn’t show fingerprints. This ring is perfect. One great thing that I enjoy is that the interior is rounded and polished, making it feel like silk when I put it on. Very affordable, too! Just goes to show, you don’t have to break the bank to get attractive quality.

Table 3.11: Sample review excerpts with the most and the least distinctive style as given by transformer models.

Model	Most consistent	Least consistent
BERT	**Author A:** capcom is the greatest video game company in the universe there true genius's the best of the best capcom rocks all the games capcom made from the 80s,90s,2000s,2010's and 2015 are the greatest video games in the universe there true classics the best of the best all the games capcom made from the 80s,90s,2000s,2010;s and 2015 rocks 2015 is the greatest year for capcom a perfect year the best of the best 2015 for capcom rocks ˆ_ˆ	**Author B:** I haven't read the novel. I can't say whether this is a good adaptation. There is no plot as such; just a random collection of events dictated by fate. When it became clear that all characters are pawns of fate, what happens to them became uninteresting. Likely, this follows the novels intention. I watched for 30 min and stopped.
	Author A: the star wars prequel trilogy is the greatest movie trilogy in the universe there true classics the best of the best the star wars prequel trilogy rocks the star wars charecter anakin skywalker is the greatest movie charecter in the universe its pure genius the best of the best the star wars charecter anakin skywalker rocks ˆ_ˆ	**Author B:** A true feat of alchemy, turning base metal (a script worth itś weight in manure) into piles of cash. Or more specifically, this is one of the dumbest, least plausible, movies wev́e watched in a long time. And yet not without comic relief. Now, who was it that said, ”Nothing will come of nothing”? Silly old bard.

Table 3.12: The most and the least consistent authors as identified by SBERT.

Model	**Most consistent**	**Least consistent**
RoBERTa	**Author C:** these our not real instruction tapes but introductions to who Larry really is you can learn from them some really good stuff kenpo is a marshal art that is based on common sense any one who really understand his marshal art will be doing kenpo with out knowing kenpo our even taking a class all marshal artist will run it to these principles for they our the principle of the sword	**Author D:** Some of these songs have only appeared on CBS Christmas compliation albums in the 60ś. I have not run across the Mike Douglas ”Touch Hands on Christmas Morning” since itś appearance on a CSP LP for Grantś Department Stores in 1967. Audio quality is pretty decent for the age of the recordings. The shear quantity of the music makes it well worth $14. Happy Holidays.
	Author C: i set down with the book try to read it and this is a book that takes a lot of time to read but in it i find nothing really new our enlightening just same old new age dribble it sounds like a psychologist is writing the book i find the same old thing in new age books now there is a lot of people that would say this is a good book	**Author D:** Well worth the price. Helps greatly compared to a junky rubber duckie antenna that comes with portables. Got mine with the SMA connector which works on Yaesu portables. You will need an SMA female to female SMA adapter to use with the cheap Chinese portables. Even works on mobiles if you keep the power down and use an adapter. A good choice for scanner use on VHF & UHF too

Table 3.13: The most and the least consistent authors as identified by SRoBERTa.

	Range	Amazon	Reddit
Consistent	Random	**0.806** / 0.875	0.677 / 0.761
	High	0.801 / **0.879**	**0.680** / **0.767**
	Moderate	0.805 / 0.874	0.672 / 0.757
	Low	0.797 / 0.867	0.661 / 0.716
Distinctive	Random	0.808 / 0.876	0.695 / **0.754**
	High	0.803 / 0.874	**0.709** / 0.750
	Moderate	**0.808** / **0.880**	0.663 / 0.731
	Low	0.793 / 0.861	0.611 / 0.68

Table 3.14: Performance (F1/AUC) on different partitions of data. Models trained on inconsistent or indistinctive authors significantly underperformed.

Comparison	Spearman' r
Distinctiveness	
SRoBERTa vs. SBERT	0.76
$\text{LSTM}_{\text{RoBERTa}}$ vs. $\text{LSTM}_{\text{BERT}}$	0.37
SRoBERTa vs. $\text{LSTM}_{\text{RoBERTa}}$	0.12
SBERT vs. $\text{LSTM}_{\text{BERT}}$	0.08
Consistency	
SRoBERTa vs. SBERT	0.76
$\text{LSTM}_{\text{RoBERTa}}$ vs. $\text{LSTM}_{\text{BERT}}$	0.61
SRoBERTa vs. $\text{LSTM}_{\text{RoBERTa}}$	0.68
SBERT vs. $\text{LSTM}_{\text{BERT}}$	0.58

Table 3.15: Correlations for distinctiveness and consistency across model types. The results were based on Amazon reviews.

not correlated.

The overall distributions of distinctiveness and consistency computed using different models are given in Figure 3.4. The distribution of style distinctiveness conforms to a normal distribution regardless of models (Figure 3.4), though the distribution is more peaked for better models. For style consistency, its distribution also conforms to a normal distribution but the distributions predicted by different models are highly similar.

3.10.5 Correlations across models

I used Spearman correlation coefficients to assess to what extent different models assign similar rankings of distinctiveness and consistency. Results are presented in Table 3.15. The consistency scores given by all models are moderately correlated yet the correlations for distinctiveness are generally weak. Even trained on the same data, different model architectures still give slightly different rankings of distinctiveness and consistency, implying that the inductive biases of models themselves may also influence the results of authorship verification.

3.11 Compositionality of styles

3.11.1 Motivation

Finally, I sought to understand how stylistic variations are encoded. At least for certain stylistic features, there is *additive stylistic compositionality* in the latent space onto which the texts are projected (Figure 3.5).

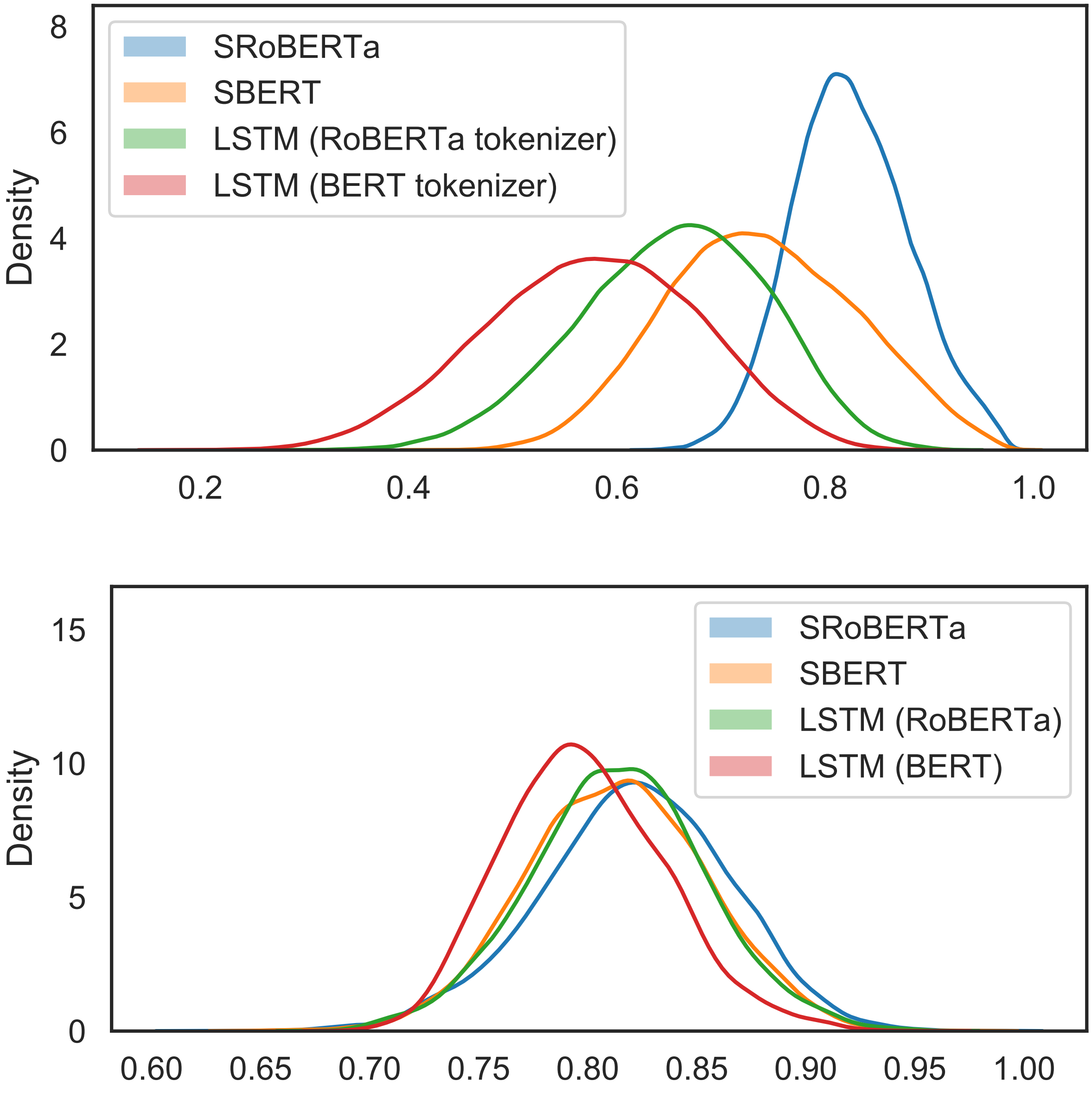


Figure 3.4: Distributions of style distinctiveness. [**Top**] and distributions of style consistency [**Bottom**] in Amazon reviews. The x-axis stands for distinctiveness or consistency, whereas the y-axis is the probability.

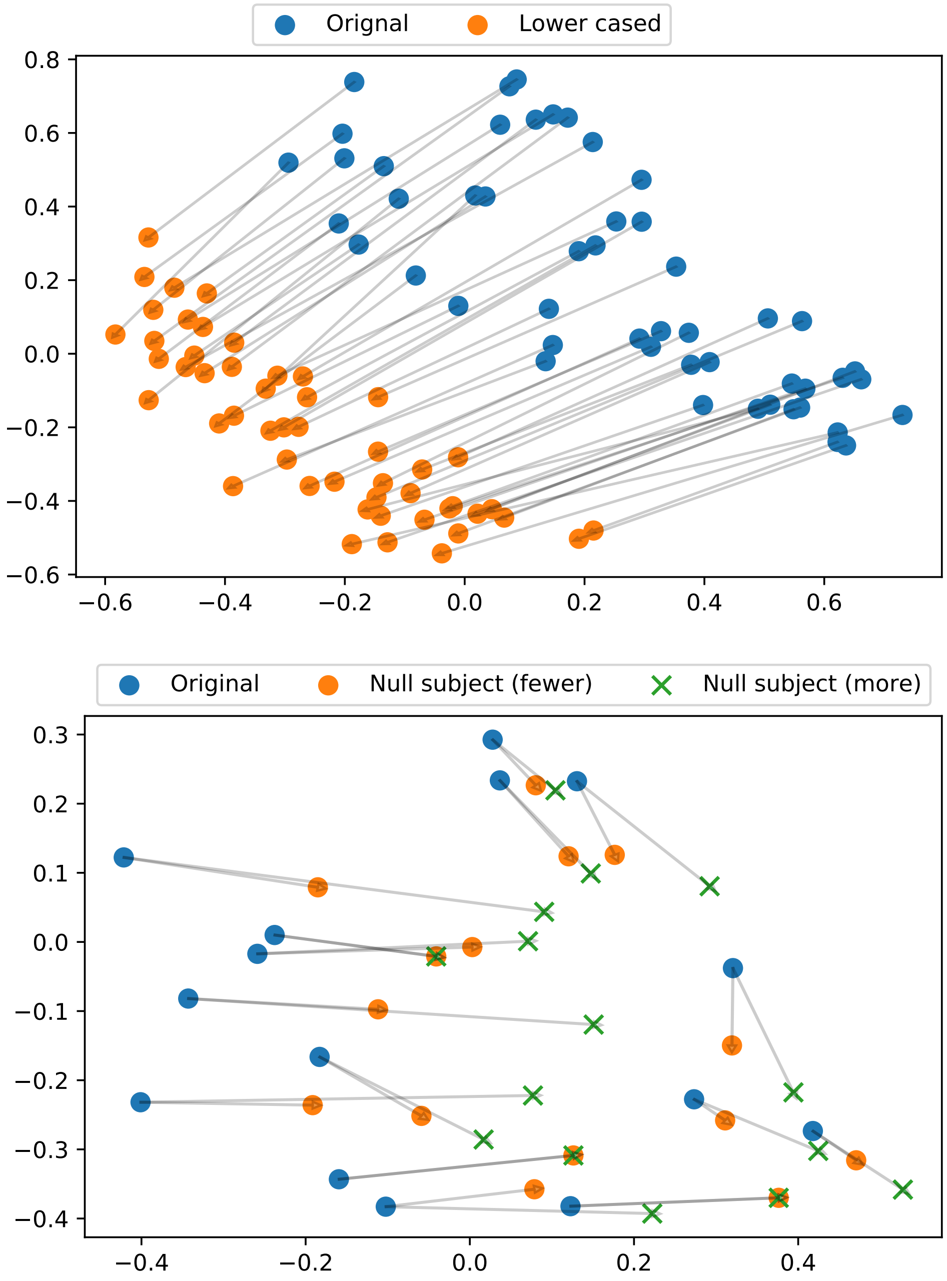


Figure 3.5: Compositionality in stylistic embeddings encoded by SRoBERTa, projected into the first two principal components. Both x and y axes are principal components. **[Top]** Lower-casing all letters shifts all original texts in the same direction (blue dots → orange dots; e.g., “I love Cantonese BBQ!”→“i love cantonese bbq!”). **[Bottom]** The magnitude of movement in one direction is proportionate to the number of null-subject sentences in texts (blue dots → orange dots → green crosses; e.g., “I went out. I bought durians.”→“Went out. I bought durians.”→“Went out. Bought durians”.).

Additive stylistic compositionality is analogous to the distributed representations of words, which have been found to exhibit striking syntactic and semantic regularities in the vector space (Mikolov et al., 2013a,b). The classic example is the operation `king` - `man` + `woman` $\approx$ `queen`, which demonstrates the additive semantic compositionality learned by the distributed model. Such regularities are often assessed by the analogy task. For vector representations of word pairs $(\vec{a}, \vec{b})$ and $(\vec{c}, \vec{d})$ of the same linguistic relations within the pair, it is expected that $\vec{a} - \vec{b} \approx \vec{c} - \vec{d}$. Let f be a function that converts a word to an numerical embedding. According to Mikolov et al. (2013b), it is found that

$$f(\texttt{king}) - f(\texttt{man}) \approx f(\texttt{queen}) - f(\texttt{woman}) \tag{3.8}$$

For the distributed representations of stylometry learned by the Siamese network, I also found such parallels in terms of writing styles. At least for certain stylistic features, there is **additive stylistic compositionality** in the latent space onto which the texts are projected. If the same stylistic element is manipulated in the same way in two different sentences, their vector representations will shift in the same direction and length. For example, making sentences null subject will cause them to undergo the same change in the vector space, as follows.

$$f(\texttt{I went out.}) - f(\texttt{went out.}) \approx f(\texttt{I bought durians.}) - f(\texttt{bought durians.}) \tag{3.9}$$

An example is shown in Figure 3.5. Style embeddings were extracted from controlled texts and principal component analysis was done on the embeddings to reduce them in the two dimensional space for visualization. When I converted all letters in the original review to lower case, such operations caused the representations of these reviews to move collectively in approximately the same direction in the latent space. Apart from systematic style shift, finer-grain style change can also be detected by the vector length. I selected reviews that contain 10 instances of the pronoun "I" as the subject. I either removed the first three instances or removed all the occurrences of "I" from the original reviews. The resulting representations both shift in the same direction but differ in the magnitude of the shift.

3.11.2 Method

In light of the word analogy task for word embeddings (Mikolov et al., 2013b; Linzen, 2016), I designed a series of linguistic stimuli that vary systematically in styles to probe the structure of the stylistic embeddings. For each stylistic dimension, I created n text embedding pairs $\mathcal{P} = [(\mathbf{p}_r^1, \mathbf{p}_m^1), \ldots, (\mathbf{p}_r^n, \mathbf{p}_m^n)]$ where $\mathbf{p}_r^i$ is the embedding of a randomly sampled text and $\mathbf{p}_m^i$ is the embedding of the modified version of $\mathbf{p}_r^i$ so that it differs from $\mathbf{p}_r^i$ in only one stylistic aspect. For sample i and j from $\mathcal{P}$, I quantified

$$S_{ij} = cosine(\mathbf{p}_r^i - \mathbf{p}_m^i, \mathbf{p}_r^j - \mathbf{p}_m^j) \tag{3.10}$$

Like the word analogy task, if this stylistic dimension is encoded only in one direction, I should expect S_{ij} close to 1.

For a target qualitative style shift, I randomly sampled 1000 texts and modified the text to approximate the target stylistic dimension. For example, if null subject is the target feature, I remove the subjective “I” from “I recommend crispy pork!” to “Recommend crispy pork!”. Then I compute S_{ij} for each pair, totaling 499500 possible comparisons. Here I selected 10 stylistic markers of textual formality for evaluation (MacLeod and Grant, 2012; Biber and Conrad, 2019) and these textual modifications cover insertion, deletion, and replacement operations.

For quantitative style shifts, I measure S_k between samples as well as the difference in length with the following equation. I compare two embedding pairs $(\mathbf{p}_r^k, \mathbf{p}_s^k)$ and $(\mathbf{p}_r^k, \mathbf{p}_l^k)$, where both $\mathbf{p}_s^k$ and $\mathbf{p}_l^k$ differ from $\mathbf{p}_r^i$ in only one stylistic dimension. But $\mathbf{p}_l^k$ is further along that dimension than $\mathbf{p}_s^k$. For instance, compared to the original review $\mathbf{p}_r^k$, $\mathbf{p}_s^k$ contains five more “!!!” whereas $\mathbf{p}_l^k$ contains ten more such tokens. Here I surveyed four stylistic markers of formality.

$$\begin{aligned} S_k =& cosine(\mathbf{p}_l^k - \mathbf{p}_r^k, \mathbf{p}_s^k - \mathbf{p}_r^k) \\ \Delta norm_k =& ||\mathbf{p}_l^k - \mathbf{p}_r^k||_2 - ||\mathbf{p}_s^k - \mathbf{p}_r^k||_2 \end{aligned} \tag{3.11}$$

For each stylistic shift, I collected 2000 samples, each containing at least 8 markers of that style. Then I modified the original review $\mathbf{p}_r^k$ to the target style by incrementally transforming the keywords into the target keywords by 50% ($\mathbf{p}_s^k$) and 100% ($\mathbf{p}_l^k$). If these styles are highly organized, I should expect S_k to be close to 1, suggesting that changes in the same dimension point to the same direction. Yet I also expect that quantitative changes should also be reflected in the significant length difference (magnitude of $\Delta norm_k$) and the direction of the difference ($\Delta norm_k$ being positive or negative) of the style vectors.

3.11.3 Results

Results in Table 3.16 suggest that both models outperform the random baseline, S_{ij}, generated with the same samples by randomly replacing some words. Like word embeddings, stylistic embeddings also exhibit a linear additive relationship between various stylistic attributes. In Figure 3.5, converting all letters to lower case causes textual representations to move collectively in approximately the same direction. Despite such regularities, the offset vectors for style shifts were not perfectly aligned in all instances, which may be attributed to the variations across texts.

For quantitative changes, SRoBERTa on both Amazon and Reddit data encode the same type of change in the same direction, as the vector offsets are highly aligned to each other (Table 3.17). Yet, greater degrees of style shift relative to the original text translate to a larger magnitude of change along that direction ($\Delta norm_k$ in Table 3.17). In Figure 3.5, after removing the first three

Style Shift	**Amazon**	**Reddit**
Random	0.032	0.002
and → &	0.698	0.484
. → space	0.389	0.352
. → !!!!	0.569	0.396
lower-cased	0.679	0.660
upper-cased	0.463	0.470
I → ∅	0.420	0.306
going to → gonna	0.398	0.390
want to → wanna	0.478	0.505
-ing → -in’	0.522	0.484
w/o elongation → w/	0.410	0.343

Table 3.16: $\overline{S_{ij}}$ for qualitative style shifts, averaged across all comparisons. The offset vectors for style shifts are highly aligned.

Style shift	**Amazon**		**Reddit**	
	$\overline{S_k}$	$\overline{\Delta Norm}$	$\overline{S_k}$	$\overline{\Delta Norm}$
Random	0.686	0.059	0.692	0.059
and → &	0.957	0.176	0.922	0.124
. → !!!!	0.964	0.085	0.953	0.072
I → ∅	0.860	0.257	0.831	0.273
-ing → -in’	0.933	0.103	0.928	0.097

Table 3.17: $\overline{S_k}$ and $\overline{\Delta Norm}$ for quantitative style shifts. The direction and the magnitude of change encode the type and the degree of style shift. $\overline{\Delta Norm}$ is positive, suggesting that more manipulations result in a longer offset vector along that direction.

instances or all the occurrences of “I” from the original text, the resulted representations both shift in the same direction but differ in magnitude. Such changes cannot be explained by random variation, suggesting that both models learn to encode fine-grained stylistic dimensions in the latent space through the proxy task.

While I only examine several stylistic markers, I was aware that the learned style representations also exhibit regularities for other lexical manipulations, as long as the manipulation is systematic and regular across samples. An explanation is that the model is systematically tracking the fine-grained variations in lexical statistics. Yet the proposed model must also encode more abstract linguistic features because it outperformed GLAD (Hürlimann et al., 2015) and FVDNN (Weerasinghe and Greenstadt, 2020), which also track bag-of-words or bag-of-ngram features. Previous research on word embeddings attributes the performance of the analogy task to the occurrence patterns (Pennington et al., 2014; Levy et al., 2015; Ethayarajh et al., 2019). The fact that this variation is systematically encoded beyond random variation and in such a fine-grain manner indicates that they are stylistic dimensions along which individual choices vary frequently and regularly.

3.12 Discussion

The relatively unconstrained nature of online genres tolerates a much wider range of stylistic variation than conventional genres (Hovy et al., 2015). Online genres are often marked by unconventional spellings, heavy use of colloquial language, extensive deviations in formatting, and the relaxation of grammatical rules, providing rich linguistic resources to construct and perform one’s identity. The current analysis of idiolects in online registers has highlighted that idiolectal variations permeate all linguistic levels, present in both surface lexico-syntactic features and high-level discourse organization. Traditional sociolinguistic research often regards idiolects as idiosyncratic and unstable and not as regular as sociolects (Labov, 1989; Barlow, 2018); here, I show that idiolectal variation is not only highly distinctive but also consistent, even in a relatively large population. These findings suggest that individuals may differ considerably in degrees of consistency and distinctiveness across multiple text samples, which sheds light on the theoretical discussions and practical applications in forensic linguistics. In previous studies on forensic linguistics, distinctiveness and consistency have been raised and discussed, but usually for a small group of authorsor comparison between specific individuals (Grant, 2012; Coulthard et al., 2016). In the past, population-based comparisons were criticized due to the lack of population-level data (Grant, 2012). This study has filled in this gap by measuring distinctiveness and consistency against a large background population ($>$100 thousands authors), showing that even among a large population most authors are still distinctive in their writing styles. The findings provide further validation of the use of large-scale authorship analysis to tackle, for example, several online criminal activities through language tracking (Grant and Macleod, 2016; Grant and MacLeod, 2020; MacLeod and Grant, 2021). These findings also have methodological implications for sociolinguistics. Compared to conventional methods of manually analyzing a handful of variations from a limited sample, this study has developed an effective method for discovering, understanding and exploiting sociolinguistic variation in large scale text data. The new methodology can expand the scope of traditional sociolinguistics to the massive data available in online communities.

This study also finds that the learned embeddings for idiolectal styles exhibit regularities through vector arithmetic at the discourse level, similar to the linear structures within word embeddings (Levy and Goldberg, 2014; Mikolov et al., 2013b). Linguistically controlled probing tasks were designed to understand the structure of the neural embeddings, revealing both qualitative and quantitative regularities of styles. The continuous and gradient nature of style were encoded in the geometry of the latent space. While such regularities are shown by linguistic probes, it remains unclear how these regularities emerge during contrastive learning. The stylistic traits we surveyed are relatively simple (at the surface level), and therefore it will be of interest to investigate how subtle styles such as tone or concreteness are potentially encoded. As recent work has shown how deep learning models encode regular syntactic and discourse information (Tenney et al., 2019; Manning et al., 2020; Chiang et al., 2020), future research is needed to understand the embedding spaces for linguistic styles.

In Chapters 2 and 3, I have investigated variation and change in written texts. Variation in another important modality of language, speech, also deserves in-depth investigation. In the next chapter, I will move on to present a machine learning-based tool to aid the tracking of variation in speech.

CHAPTER 4

Study 3: Semi-supervised Learning of Phone-to-audio Alignment

4.1 Introduction[1]

In this chapter, I move from variation in the textual domain in previous chapters to variation in the speech domain. To study speech variation, I develop a tool to automate phone-to-audio alignment. Aligning phones to audio has been a fundamental task in speech research, as many speech analyses depend on knowing the exact timing of phones. Precise phone-to-audio alignment has a variety of applications in speech science and technologies, including for speech synthesis (e.g., Ren et al., 2019; Ren et al., 2020), pronunciation assessments (e.g., Chen et al., 2009; Plantinga and Fosler-Lussier, 2019; Mathad et al., 2021), sociophonetics (e.g., Yuan and Liberman, 2009; Baranowski, 2013; Bailey, 2016; Fromont and Watson, 2016; Gonzalez et al., 2020; MacKenzie and Turton, 2020), language documentation (e.g., Gonzalez et al., 2018; Johnson et al., 2018; Tang and Bennett, 2019; Babinski et al., 2019; Barth et al., 2020) and speech pathology (e.g., Terbeh et al., 2016; Yeung et al., 2015; Riad et al., 2020). Figure 4.1 presents a graphical illustration of phone segmentation as represented in a Praat textgrid, in which a continuous speech signal is segmented into individual, non-overlapping intervals that correspond to a phone or a word. The segmentation, therefore, allows speech researchers to measure, or manipulate the acoustic signal at the fine-grained phone or word level. Currently, the segmentation has been mostly done by trained speech scientists, but the time and effort taken to perform manual segmentation are tremendous. A conservative estimate is that 2 seconds of speech can take about 1 minute to annotate (Mahr, 2021) but the actual annotation time could be far longer for noisy, unclear, or spontaneous speech. The Buckeye Corpus, one of the most widely used speech corpora with phone-level segmentation, took multiple years to annotate by a large group of phoneticians (Pitt et al., 2005). The time-consuming

[1] Portions of this work have appeared in *Phone-to-audio alignment without text: A Semi-supervised Approach* (Zhu et al., 2022b) in the proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

process of segmentation imposes severe restrictions on scaling up the construction and the analysis of speech corpora.

In this chapter, I present my experiments to develop neural network models to perform forced alignment and textless alignment. After outlining the goals of this study (Section 4.2), I review the relevant literature on phone-to-audio alignment (Section 4.3) and describe my approaches to achieve accurate phone-to-audio alignment in both English and Mandarin through neural networks (Sections 4.4 through 4.11).

4.2 Goals

Since the advent of end-to-end automatic speech recognition (Hannun et al., 2014; Amodei et al., 2016; Chan et al., 2016), researchers have concentrated on predicting phones or texts directly from the speech signal. In this paradigm, the timing information of individual phones is usually not preserved, as the final output that researchers are aiming for is simply a list of strings of phones or characters. However, for many speech applications, including most phonetic research, speech synthesis, and pronunciation assessments, for example, the precise timing of phones or words is needed. For generating precise phone segmentations from raw speech signals, the conventional phone recognition task needs to be time-aligned phone recognition, a task that jointly performs phone recognition and segmentation (see Figure 4.2 for an illustration).

Given that modern ASR has increasingly shifted to deep neural networks, which have achieved state-of-the-art performance, this study focuses on deep learning methods for phone-to-audio alignment. Directly training neural networks to perform time-aligned phone recognition poses daunting challenges, including that the models require a large amount of human segmented speech labels to learn from. Except for some relatively small-scale corpora like TIMIT (Garofolo et al., 1993) and the Buckeye Corpus (Pitt et al., 2005), such annotated speech corpora are generally unavailable and, as mentioned, take considerable time and effort to create. In most scenarios, researchers only have access to collections of speech recordings and their transcriptions aligned at the sentence level. In this study, I tackle this challenge using weakly supervised learning, by training a forced alignment model coupled with a phone recognizer and training the time-align phone recognition model on forced-aligned labels.

In the rest of the chapter, I present two neural network-based methods for both text-dependent and text-independent phone-to-audio alignment through weak supervision. Compared to the classic HMM models for forced alignment, neural networks are far more powerful in ASR performance and have dominated ASR research and applications. My contribution can be summarized as follows. First, I present a semi-supervised model that performs forced alignment, but can also be combined with a phone recognizer for text-independent alignment. This model achieves re-

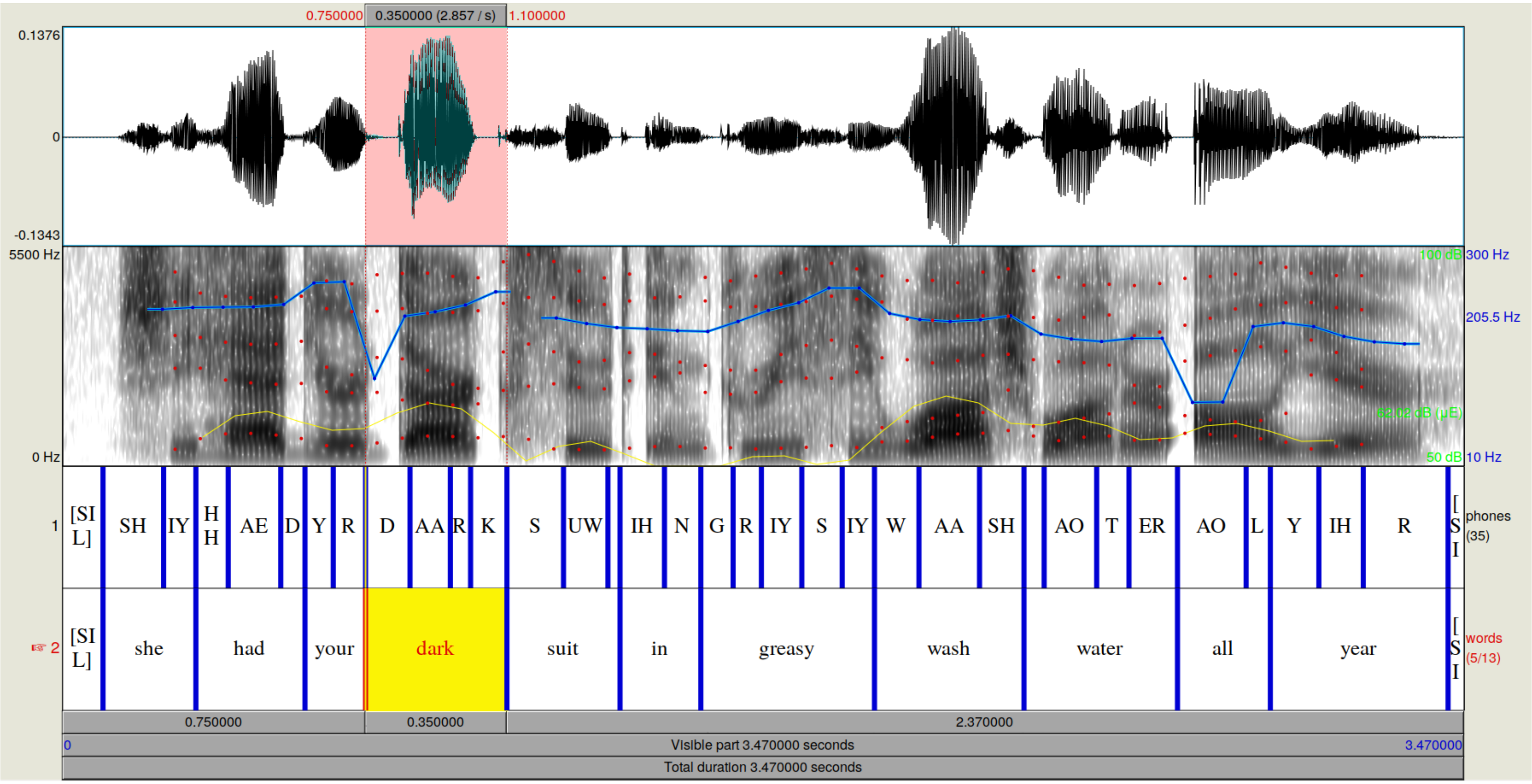
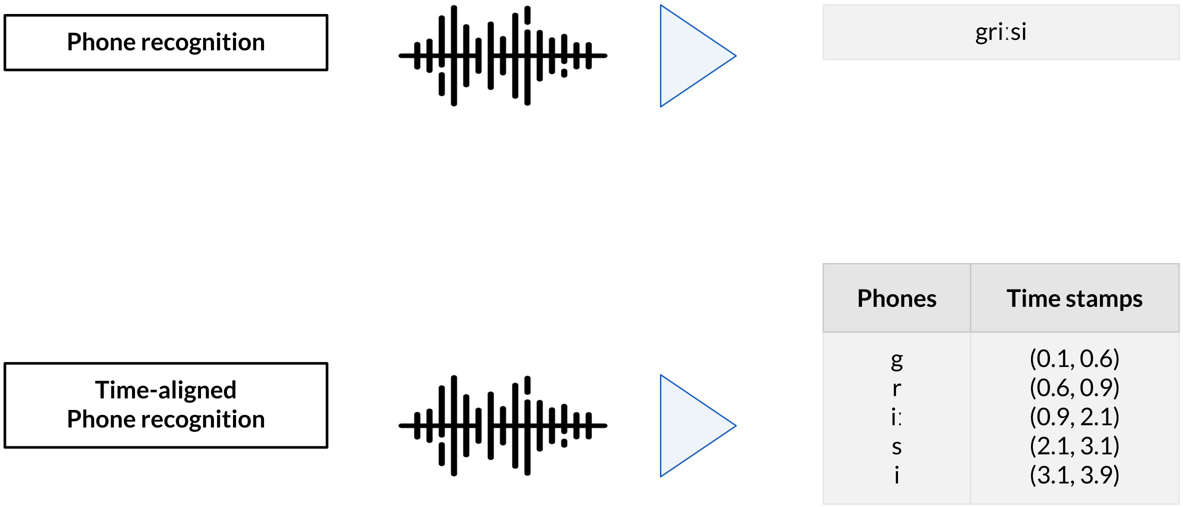


Figure 4.1: Illustration of phone segmentation for a sentence from the TIMIT corpus.



Phones	Time stamps
g	(0.1, 0.6)
r	(0.6, 0.9)
i:	(0.9, 2.1)
s	(2.1, 3.1)
i	(3.1, 3.9)

Figure 4.2: Illustration of the difference in the output of phone recognition and time-aligned phone recognition models.

sults that are comparable to those of four existing forced alignment tools. Second, I present a frame classification model for text-independent alignment and show that this model achieves performance that is close to that of text-dependent models. Extensive analyses show that the proposed methods maintain good performance across different settings. In contrast to currently available tools, my work provides speech and language researchers with a strong deep learning-based tool to obtain phone segmentations even when textual transcriptions are not available.

4.3 Background

4.3.1 Forced alignment

In recent decades, advancements in automatic speech recognition (ASR) have introduced a wide range of statistical methods for automatic segmentation. The most widely used technology used by the linguistics community is *forced alignment*, which can align speech recordings to their orthographic transcriptions with an acoustic model, a pronunciation dictionary that maps orthographic words onto phones, and the Viterbi algorithm to search for the optimal alignment path.

There are currently several existing tools for performing forced alignment over phone sequences, including the Montreal Forced Aligner (McAuliffe et al., 2017), ProsodyLab (Gorman et al., 2011), Penn Forced Aligner (FAVE) (Rosenfelder et al., 2011), WebMAUS (Kisler et al., 2012), Gentle (Ochshorn and Hawkins, 2017), EasyAlign (Goldman, 2011), DARLA (Coto-Solano et al., 2021) and SPPAS (Bigi, 2012). The majority of these are based on the classic HMM system built upon Kaldi (Povey et al., 2011) or the HTK toolkit (Young et al., 2002). Though neural ASR systems have shown better performance than HMM systems (Baevski et al., 2020), phone-to-audio alignment has not shown many benefits from the predictive power of neural networks. One reason is that modern ASR has increasingly shifted towards end-to-end training using loss functions like connectionist temporal classification (CTC) (Graves et al., 2006), that is, training an ASR model that directly maps acoustic signals to graphemes, bypassing phones altogether. The end-to-end training pipeline greatly simplifies the ASR training and deployment, yet the CTC loss necessary in this process disregards precise frame alignment. A few studies, though, have explored using neural networks to perform segmentation of sentences (Kürzinger et al., 2020) and phones (Kelley and Tucker, 2018; Schulze-Forster et al., 2020; Teytaut and Roebel, 2021) and these approaches show great potential for neural forced alignment, although they still require text transcriptions.

Despite the success of forced alignment tools, forced alignment itself also faces many engineering challenges. First, due to its "forced" nature, a forced aligner will only align strictly to the given phone sequences, even if a given phone sequence does not reflect all that is spoken in the speech signal. When converting the textual transcriptions to phone representations, the output phones are always limited by the grapheme-to-phoneme conversion (G2P) tool, which might not reflect the actual pronunciation. The G2P process can introduce errors when there are out-of-vocabulary words not included in the dictionary. Recent research in statistical G2P might mitigate the errors to some extent (e.g., Bisani and Ney, 2008; Novak et al., 2012; Toshniwal and Livescu, 2016; Park and Lee, 2020), but G2P itself is a challenging problem (e.g., Deri and Knight, 2016; Peters et al., 2017; Gorman et al., 2020; Vesik et al., 2020), mainly because the mapping between phones and graphemes is not regular for many languages and pronunciation dictionaries are lacking except for a small number of high resource languages.

However, even if G2P were to achieve perfect accuracy in G2P, problems would remain. The generated phone sequence can only reflect the canonical pronunciations recorded in a dictionary. In production, speech usually undergoes a variety of phonetic and phonological processes, making phonetic realizations of phones deviate from their dictionary pronunciation to varying degrees. The lack of invariance refers to the fact that the acoustic realization of a phone varies considerably in different phonetic environments (Klatt, 1979). Physiological and social factors such as emotions, accents, gender, dialects, etc. add to the variability of connected speech (Benzeghiba et al., 2007). Aligning these variable phonetic realizations to the dictionary pronunciation risks eliminating a significant portion of linguistic variability, which itself is the subject of many phonetic studies.

Thirdly, forced alignment is also a difficult task in itself, as is ASR in general. Forced alignment is inherently an ASR task, though transcriptions are available. Most publicly available forced aligners are based on the HMM frameworks, which, albeit still reliable, are increasingly being taken over by the more powerful neural networks in recent ASR research. HMM-based hybrid systems are limited in several aspects. HMM is based upon the Markov assumption that the current acoustic frame only depends on its previous frame (Bishop, 2006). This greatly limits the model's capacity to extract long contextual information given that speech is, as just discussed, highly contextualized. In contrast, neural networks, notably RNNs and transformers, are capable of modeling very long speech contexts from both the past and the future, making them extremely effective in modeling ASR (Jurafsky and Martin, 2018). There are currently few studies of neural network-based forced aligners (but see Kelley and Tucker, 2018; Schulze-Forster et al., 2020; Teytaut and Roebel, 2021). Given that neural ASR has surpassed HMM hybrid systems in performance (Jurafsky and Martin, 2018), it is expected that neural networks can further improve the accuracy of forced alignment.

4.3.2 Text-independent phone segmentation

While forced alignment provides an effective solution to phone segmentation with orthographic transcriptions, most audio data in the wild are not transcribed into texts. In order to perform phonetic analysis on these audio data, it is necessary to perform phone segmentation directly on raw audio signals—that is, to perform text-independent phone segmentation. Compared to forced alignment, text-independent phone segmentation has received much less attention but is being actively researched in recent years, especially using unsupervised methods.

The problem of locating phone boundaries in the acoustic signal has been explored by many researchers (e.g., Brent, 1999; Sakran et al., 2017). Machine learning methods are highly accurate in segmenting phones when trained on human segmentation data (Keshet et al., 2005; King and Hasegawa-Johnson, 2013; Kreuk et al., 2020b). Even in the absence of human annotations, a few unsupervised or self-supervised phone segmentation methods have also been proposed (Qiao et al., 2008; Rasanen, 2014; Franke et al., 2016; Wang et al., 2017; Kreuk et al., 2020a; Kamper and van Niekerk, 2021; Bhati et al., 2021). The use of deep learning-based methods in recent years has been closing the gap between fully supervised methods, including Contrastive Predictive Coding (CPC) (Kreuk et al., 2020a; Bhati et al., 2021) and quantized Variational AutoEncoder (qVAE) (Kamper and van Niekerk, 2021). However, most of these methods only output phone boundaries but do not jointly predict *both* boundaries and phones, making them less practical in phonetic research.

Very recently, though, unsupervised methods for joint segmentation and recognition of phones have been proposed that achieve competitive results to supervised ASR (Liu et al., 2022c; Baevski et al., 2021; Liu et al., 2022a). First proposed by Baevski et al. (2021), this line of work adopts the framework of generative adversarial networks (Goodfellow et al., 2014), in which ASR models can be trained with unpaired speech and text labels. In adversarial training, a generator takes a speech signal and predicts the frame-level phones. Given that the ground truth phone labels are not available in the unsupervised setting, it is impossible to judge whether or not the predictions are correct. Instead, an adversarial classifier ("discriminator") is used to discriminate whether the generated phone sequence looks like a real phone sequence in that language or not, with real phone sequences coming from the unpaired texts. By outputting predicted phone sequences that are so similar to real phone sequences that a discriminator cannot distinguish between them, the generator learns to convert a speech signal into its phone representations.

It is also possible to perform phone segmentation into latent units directly learned from data. Inspired by the training of HuBERT (Hsu et al., 2021) and speech unit discovery (Brent, 1999), Lakhotia et al. (2021) also proposed a method for unsupervised segmentation of speech signals, though the segmented units were based on clustering centroids in the latent space that encode phonetic information, rather than actual phones. For many engineering applications that do not require interpretations, these latent units work effectively as representations of speech structure. But for most phonetic research, the uninterpretable nature of these latent units renders them less useful.

As these studies aim to perform segmentation and recognition jointly at the frame level, the predictions given by machine learning models can also be utilized to derive phones and their timestamps from raw audio. Compared to forced alignment, text-independent phone alignments can generate time-aligned textgrids directly from the audio, further cutting down the costs of labeling speech signals. However, textless alignment methods that directly segment the audio are not currently available to the speech community. Given the advantages of text-independent segmentation and the lack of such a tool, I also developed a method to perform text-independent alignments directly on audio, as transcriptions might not always be available in many practical applications. The motivations and details are given in the following sections.

4.4 Neural forced alignment

In forced alignment, the goal is to align acoustic frames with phones given in the transcription. Given a speech signal $X^S \in \mathbb{R}^{1\times T_{raw}}$ and a sequence of phonetic symbols $Y^P \in \mathbb{R}^{1\times N}$, some neural networks are used to encode them into hidden representations Φ, such that $\boldsymbol{Y} = \Phi_S(Y_P)$ and $\boldsymbol{X} = \Phi_P(X_S)$, where $\boldsymbol{X} \in \mathbb{R}^{K\times T} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_t]$ and $\boldsymbol{Y} \in \mathbb{R}^{K\times N} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_n]$, $T \geq N$. I seek to learn a transformation f to encode these features into the monotonic alignment $S = f(\boldsymbol{X}, \boldsymbol{Y})$, so that it aligns each phonetic symbol y^P_m to a subset of consecutive acoustic frames $\boldsymbol{x}_i$: $S = [\boldsymbol{x}_i = y^P_m, \boldsymbol{x}_j = y^P_m, \dots, \boldsymbol{x}_k = y^P_n]$. This approach requires phone transcriptions. When transcriptions are not available, phone labels $\hat{Y}^P$ can be estimated with a phone recognizer from audio.

The proposed neural forced alignment model learns the phone-to-audio alignment through the self-supervised task of reconstructing original speech with both heavily corrupted speech representations and phonemic information (Schulze-Forster et al., 2020). This could be implemented as the same pretraining task of Wav2Vec2 (Baevski et al., 2020), in which masked speech is used by a transformer model to predict the quantized embedding of the original speech. Meanwhile, I relied on the forward sum loss to guide the attention matrix to be monotonic alignments (Shih et al., 2021; Badlani et al., 2021). The technical formulation of the model is detailed in the following sections.

A graphical illustration of the proposed model is presented in Figure 4.3. It is assumed that a phone sequence will be derived from the textual transcription with a grapheme-to-phoneme conversion tool. The speech encoder, which is a transformer encoder, converts a given corrupted speech signal into a sequence of speech embeddings, or numerical vectors representing the features of a segment of speech within a sliding window. The phone encoder, another transformer encoder, converts the corresponding phone sequence into a series of phone embeddings, each of which is a numerical representation of a phone. The phone embeddings and the speech embeddings are of different lengths, so they are aligned with the attention mechanism (Bahdanau et al., 2015), which, in this case, can be conceptualized as the pairwise similarities between the two sequences of embeddings. Each aligned speech and phone embeddings are concatenated together to be used as numerical features to reconstruct the original uncorrupted speech embeddings. The reconstruction task, as well as numerical constraints to form diagonal alignment (one-to-one monotonic phone-to-audio mapping), will incrementally refine the alignments between speech and phones during model training.

4.4.1 Encoders

For model architecture, I used two neural network encoders to convert speech signals and phone sequences into numerical representations, so as to align them.

I used the pretrained `wav2vec2-base` model (Baevski et al., 2020) with a convolutional layer on top as the speech encoder Φ_S. Wav2Vec2 is a transformer-based speech model that has been pretrained on about 60k hours of English speech signals. The pretraining enables it to induce powerful speech representations from a massive amount of data, such that it achieves state-of-the-art performance in multiple speech recognition benchmarks (Baevski et al., 2020). Using the pretrained Wav2Vec2 as the speech encoder can reduce the training time and also improve the alignment performance relative to using Mel Frequency Cepstral Coefficients (MFCC) or training a neural network from scratch.

The frame rate of the speech encoder has a non-negligible impact on the final performance. Currently, most HMM-based forced aligners adopt a 10ms frameshift, so that they can encode temporal difference that is longer than 10ms (e.g., McAuliffe et al., 2017; Ochshorn and Hawkins, 2017). As the original frame rate of Wav2Vec2 (98Hz, or about 20ms per frame) could be too coarse to derive precise timing, I increased the frame rate by two methods, either by modifying the speech model or upsampling the speech signals. To modify the original Wav2Vec2 architecture, I reduced the stride of the last convolutional layer in its feature extractor from 2 to 1 (denoted as `-10ms`, which double the length of the output speech frames. Given that Wav2Vec2 only accepts input speech signals at a sampling rate of 16k, the other approach is to upsample the raw speech signal to 32kHz (denoted as `-32k`), so that the resulting acoustic frames are twice as long. Both approaches effectively doubled the frame rate to 98Hz (about 10ms per frame), while allowing the model to load all the pretrained parameters of `wav2vec2-base`. The downside of doubling the frame rate is that computational costs and memory consumption are also increased, causing the model to perform inference more slowly and take up more memory space. Compared to the original Wav2Vec2, upsampling the raw speech signals will incur more computational costs at every layer of the model and at the data preprocessing step. In contrast, increasing the convolution stride only costs more computations at a subset of layers and no additional steps for preprocessing are needed.

The phone encoder Φ_T is a reduced version of the BERT model (Devlin et al., 2019) with a convolutional layer on top. It has 4 hidden layers, each of which has 12 self-attention heads of 32 hidden dimensions. The model was pretrained with masked language modeling (Devlin et al., 2019) on phone sequences generated from the whole Book Corpus (Zhu et al., 2015), which amounts to more than 2 billion phone tokens. The procedure of pretraining was similar to that of BERT. All sentences in Book Corpus were converted to phone representations in batch using a grapheme-to-phoneme conversion tool, `g2p_en`[2]. Then 20% of the phone tokens in a batch were sampled for masked language modeling, among which 80% were replaced by a `<mask>` token, 10% were replaced by a random phone, and the rest 10% retained the original phones. The phone encoder was then pretrained to predict the original phone from the masked/randomly replaced tokens with sentence contexts. Through extensive studies, this pretraining procedure has been shown to improve downstream tasks by learning effective feature presentations (Devlin et al., 2018; Peters et al., 2018; Liu et al., 2019). It is expected that pretraining the phone encoder on large-scale data could help the model better learn contextual phone representations, which can accelerate model convergence during the subsequent training of the neural forced aligner.

[2] The tool is available at: https://github.com/Kyubyong/g2p

4.4.2 Loss function

The neural forced aligner was optimized with two loss functions. The self-supervised training objective is to reconstruct the original speech through masked (corrupted) speech signals and the phone sequence. Yet it is hoped that the model can learn the implicit phone-to-audio alignment through learning the proxy task. In the original Wav2Vec2, the speech signals were masked in the temporal dimension at a probability of 7.5% (Baevski et al., 2020). Given that speech signals are redundant, at such a low masking probability, the model can learn to reconstruct the original signals based on corrupted signals alone without making use of the phone transcription. To force the model to make use of phone transcriptions, I set the masking probability up to 40%. Now that the corruption is well beyond the redundancy in speech, the model has to learn the correct alignment between phone and speech to optimize the task.

The model learns a matrix $\boldsymbol{A} \in \mathbb{R}^{N \times T}$ that aligns the phone sequence $\boldsymbol{Y}$ and the masked speech representation $\hat{\boldsymbol{X}}$.

$$\boldsymbol{D}_{ij} = f_y(\boldsymbol{y}_i)^T f_x(\hat{\boldsymbol{x}}_j) \tag{4.1}$$

$$\boldsymbol{A} = softmax(\boldsymbol{D}, \texttt{dim=0}) \tag{4.2}$$

$$\boldsymbol{H} = concatenate[\hat{\boldsymbol{X}}, \boldsymbol{Y}\boldsymbol{A}] \tag{4.3}$$

where f_y and f_x are two dense layers. $\boldsymbol{D} \in \mathbb{R}^{N \times T}$ is the (unnormalized) similarities between $\boldsymbol{Y}$ and $\boldsymbol{X}$. For the hidden states $\boldsymbol{H} \in \mathbb{R}^{2K \times T} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T]$ at each time step, the model maximizes the similarity between each $\boldsymbol{h}_t$ and the quantized version of the acoustic feature at the same time step. The quantized embeddings can be selected from a codebook using Gumbel softmax parameterized by a neural network, as in the original Wav2Vec2 (Baevski et al., 2020). Let $\boldsymbol{Q}_t$ represent the set consisting of a true quantized embedding and $n = 50$ negative samples at step t. The loss function is formulated as follows.

$$\mathcal{L}_m = \frac{1}{T}\sum_{t=1}^{T} -\log \frac{\exp\left(sim(\boldsymbol{h}_t, \boldsymbol{q}_t)/\kappa\right)}{\sum_{\boldsymbol{q}_k \sim \boldsymbol{Q}_t} \exp(sim(\boldsymbol{h}_t, \boldsymbol{q}_k)/\kappa)} \tag{4.4}$$

The masking is performed with spectral augmentation with both time and feature masking (Park et al., 2019). If the time masking probability is too low, the model might simply ignore the phone information. However, if the time masking probability is too high, it will create a gap between training and inference inputs, as the input audio for inference is usually not corrupted. I sampled the time-masking probability $p/100$ for each batch from a discrete uniform distribution of $[p_l, p_h]$. The original weights of the Wav2Vec2 quantizer were kept fixed throughout training.

However, the attention matrices might not necessarily align acoustic frames and phone tokens.

Therefore I added the forward-sum loss used in HMM systems to constrain the attention matrix to be monotonic and diagonal (Shih et al., 2021; Badlani et al., 2021).

$$\mathcal{L}_{FS} = -\sum_{\boldsymbol{X},\boldsymbol{Y} \in S} \log P(\boldsymbol{Y}|\boldsymbol{X}) \tag{4.5}$$

where S is the optimal alignment. All probability quantities are estimated in the attention matrix $\boldsymbol{A}$. This loss function can be implemented using the off-the-shelf PyTorch CTC loss by setting the probability of the blank symbol to an impossibly low value. I adapted the implementation (Shih et al., 2021; Badlani et al., 2021). Thus the final loss is a weighted sum of the two loss functions.

$$\mathcal{L} = \mathcal{L}_m + \lambda \mathcal{L}_{FS} \tag{4.6}$$

where λ was set to 1 based on empirical validation.

4.4.3 Incremental training

The alignment pattern in the neural model is notoriously hard to train. If the neural forced aligner is directly trained on all data, the model tends to collapse into random alignment patterns. Presumably, this is because inducing a single ground truth alignment out of so many potential alignments is extremely challenging, especially when the acoustic sequence is long. If the model makes misalignment mistakes at the early stages of training, these early deviations from the ground truth only get reinforced in subsequent iterations (Zeyer et al., 2017). Techniques such as guided attention loss (Tachibana et al., 2018) or diagonal prior (Shih et al., 2021; Badlani et al., 2021) or external supervision with available alignments (Ren et al., 2020) have been used to enable the formation of text-to-speech alignment. The central idea is to constrain the possible alignment patterns, so as to eliminate a large number of impossible alignments, prompting the model to converge faster.

Guided by the same idea to facilitate alignment learning, I trained the model incrementally in the curriculum learning paradigm (Bengio et al., 2009) on increasingly longer audio samples. The model is first trained on audio samples that are shorter than 2 seconds, as these samples are easier to align than longer samples. After the model has formed a diagonal alignment pattern for these short audio samples, longer audio samples are fed into the model to increase the training difficulty. The process iterates several rounds, with longer audio samples at each round, until the neural forced aligner forms a diagonal alignment pattern. Like noisy student training (Park et al., 2020) and the iterative training in Kaldi (Povey et al., 2011), a series of models are generated in the process and the last model with the best performance is retained. The algorithmic formulation is presented as follows.

1. Initialize model M with pretrained weights M_0. Set $M = M_0$.

2. Partition the length-sorted data C into k different chunks $C_1, C_2, \dots, C_k$ by duration, such that each chunk contains audio longer than the previous chunk. Then train M on each chunk sequentially with a frame shift of 20ms.

3. Upsample the frame shift of M to 10ms. Retrain M using the same data and procedure as in 2.

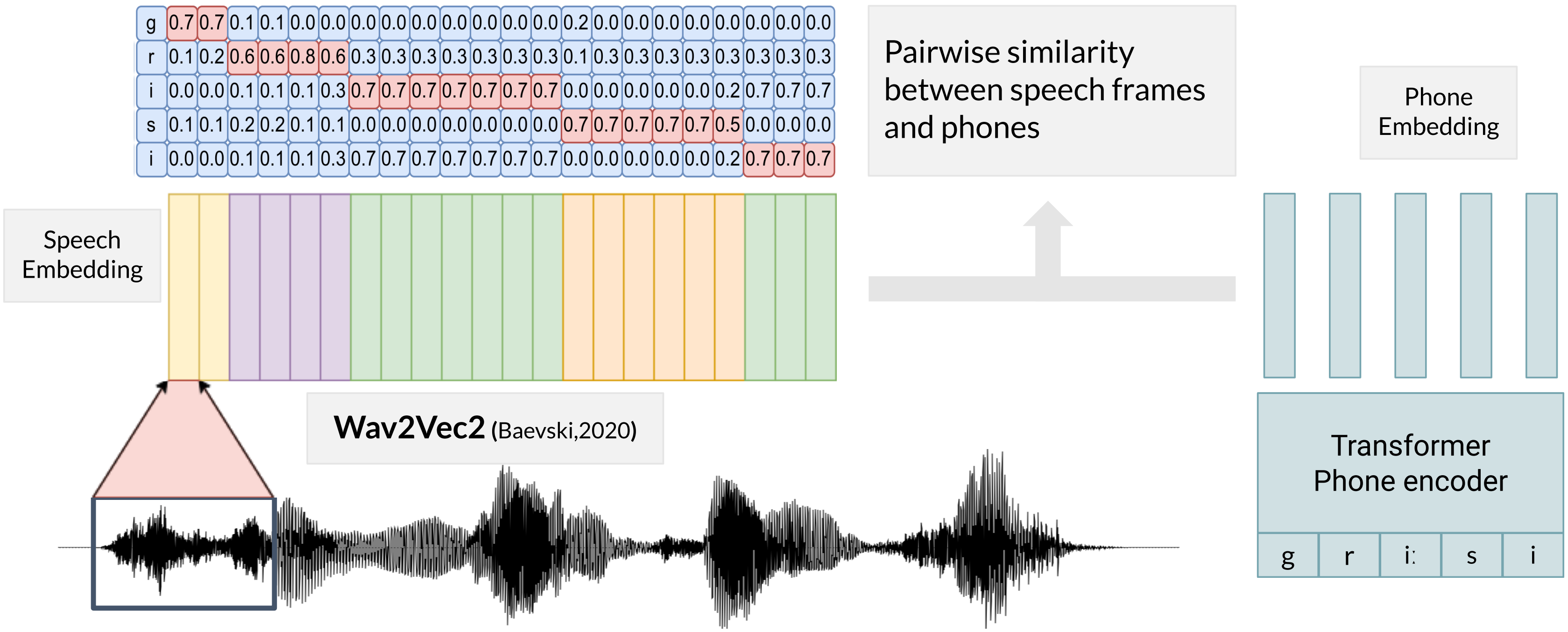
4. (*Optional*) Train a frame classification model M' from scratch with alignments generated from the last step. Set $M = M'$ and go back to 2.

Reinitialization with a frame classification model speeds up the convergence as it has been noted that a prior is essential for obtaining good alignment (Zeyer et al., 2017). As the learned attention alignments are sharp, the decoding can be performed simply using the `argmax` function without resorting to the Viterbi decoding. This forced alignment is referred to as `W2V2-FS`.

4.5 Frame classification model

For text-independent alignment, I fine-tuned a Wav2Vec2 model to perform frame-wise phone classification (referred to as `W2V2-FC-Libris`), that is, the time-aligned phone recognition shown in Figure 4.2. In this task, the model discretizes speech into acoustic frames of 20ms or 10ms with a series of convolutional layers, transforms the resulting frames into speech embeddings with a 12-layer transformer model, and uses a fully connected layer to predict a probability distribution of phone categories for each acoustic frame (Baevski et al., 2020), which can later be post-processed to derive precise time stamps. The challenge of the task is that frame-wise phone labels are required for training, but those are generally not available on a large scale due to the complexity and the cost of manual segmentation. Had such data existed, the task could have been greatly simplified. In the semi-supervised setting, the frame classification model can be trained on pseudo-labels generated by forced alignment labels, and therefore no human segmentations are required. For the sake of convenience, the training labels were alignments of the 960h *Librispeech* (Panayotov et al., 2015) obtained through MFA used Lugosch et al. (2019). However, any one of the forced alignment tools could work for creating such pseudo-labels. I also trained additional frame classification models (`W2V2-FC`) on the alignments generated from the semi-supervised method above. A graphical illustration of the frame classification model is shown in Figure 4.4.

The forced alignments were discretized into frame-level pseudo-labels at the temporal resolution of 10ms or 20ms and the model was trained to classify each frame into a phone category using the cross-entropy loss. The downside of this approach is that forced alignments are not always



g	0.7	0.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
r	0.1	0.2	0.6	0.6	0.8	0.6	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.1	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
i	0.0	0.0	0.1	0.1	0.1	0.3	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.0	0.0	0.0	0.0	0.0	0.2	0.7	0.7	0.7
s	0.1	0.1	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.7	0.7	0.7	0.7	0.5	0.0	0.0	0.0
i	0.0	0.0	0.1	0.1	0.1	0.3	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.0	0.0	0.0	0.0	0.0	0.2	0.7	0.7	0.7

Pairwise similarity between speech frames and phones
Speech Embedding
Phone Embedding
Wav2Vec2 (Baevski,2020)
Transformer Phone encoder
g
r
iː
s
i

Figure 4.3: Illustration of the Wav2Vec2-FS model. The Wav2Vec2-FS model has two separate transformer encoders. The speech encoder converts speech signals into a sequence of embeddings whereas the phone encoder maps phones into phone embeddings. The attention mechanism is used to align these two embeddings.

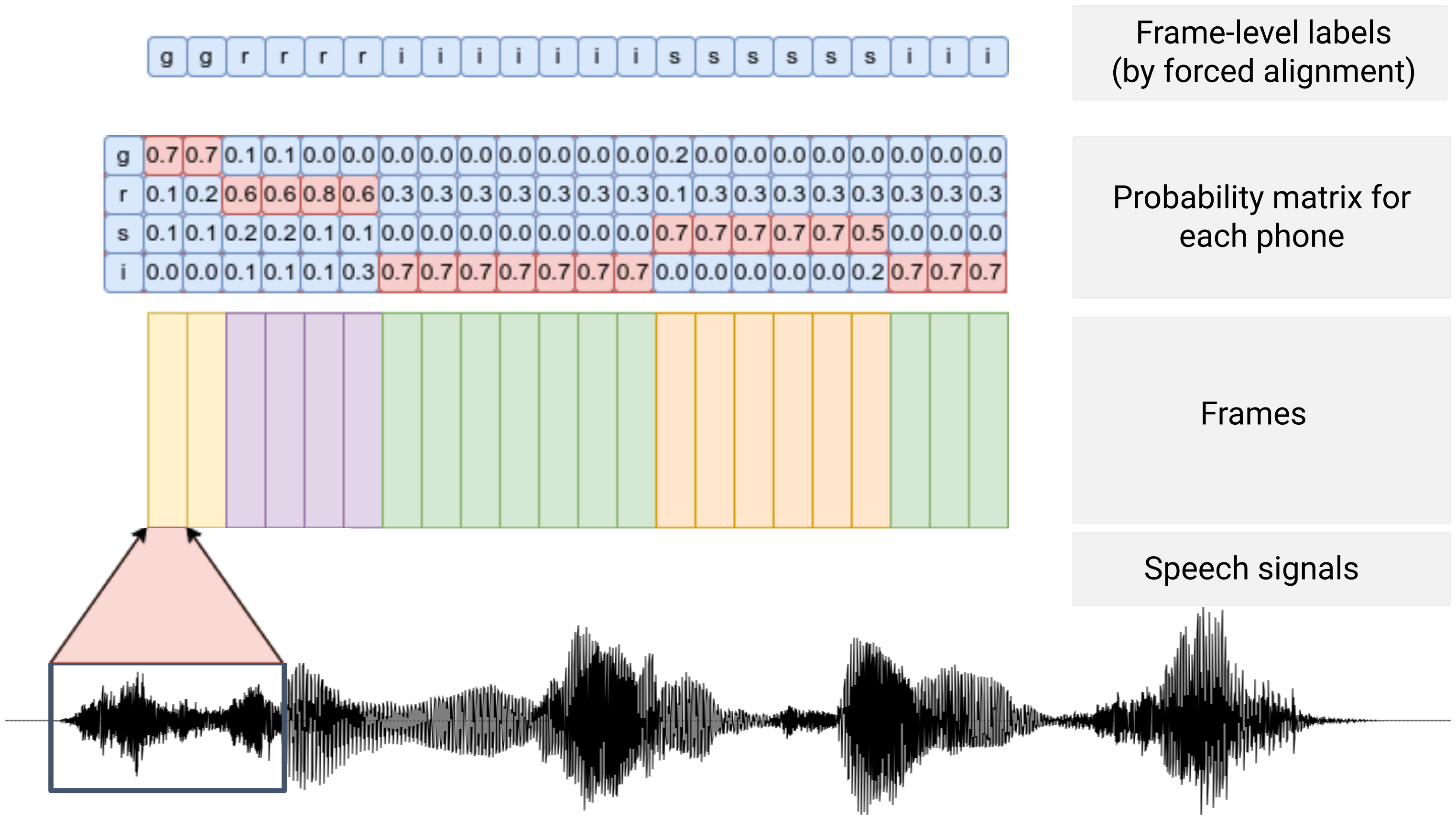
accurate, such that the biases of the forced aligner can be propagated to the frame classification model, limiting its ultimate segmentation performance. Given the scarcity of human-annotated labels, the use of pseudo-labels is a compromise. During inference, the segmentation can be performed in two ways.

1. As illustrated in Figure 4.4, textless phone alignment can be derived from model predictions directly with audio by taking the framewise phone labels with the maximum probability (highlighted in red), without the help of any transcriptions. Then the consecutive frames with the same phones are merged and the duration of each phone is a product of the number of consecutive frames and the temporal resolution. For example, in Figure 4.4, the phone [s] has been predicted for 6 consecutive frames (in orange) and the temporal resolution is 10ms, then its duration is 6 frames $\times$ 10ms $=$ 60ms.

2. Forced alignment was also performed using a Dynamic Time Warping (DTW) algorithm based on an output probability matrix and its transcription. The distance matrix for DTW is constructed by reformatting the model output, such that the resultant probability matrix has rows in the order of the transcription and the columns represent the output probability. The DTW algorithm seeks to identify the optimal alignment that maximizes the total sum of probabilities with monotonic constraints.

4.6 Baseline phone recognizers

A Wav2Vec2-based phone recognizer with the connectionist temporal classification (CTC) loss function (Graves and Jaitly, 2014) was also trained on about 2500 hours of speech. The CTC loss function is one of the most popular methods in mainstream ASR models. It directly optimizes the ASR output strings without requiring framewise alignments, thereby tremendously simplifying the pipeline of ASR models, making end-to-end training possible. While a CTC-based model still makes framewise predictions of phone labels, it only does so when the model is confident enough about the current phone, otherwise, it will skip the prediction by outputting a `[blank]` symbol. The `[blank]` symbol will simply be ignored during the decoding, in which CTC outputs are converted into the actual phone strings. Given this nature, the CTC model tends to avoid making predictions immediately at phone boundaries but waits until it has gathered enough acoustic evidence (Kurata and Audhkhasi, 2018; Plantinga and Fosler-Lussier, 2019). For unidirectional models, the predicted phones tend to be delayed from the actual phone boundaries, whereas for bidirectional models, the predicted phones shift somewhat arbitrarily either to the left or the right of the actual boundary (Plantinga and Fosler-Lussier, 2019). While CTC loss is favorable for its



g	0.7	0.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
r	0.1	0.2	0.6	0.6	0.8	0.6	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.1	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
s	0.1	0.1	0.2	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.7	0.7	0.7	0.7	0.5	0.0	0.0	0.0
i	0.0	0.0	0.1	0.1	0.1	0.3	0.7	0.7	0.7	0.7	0.7	0.7	0.7	0.0	0.0	0.0	0.0	0.0	0.2	0.7	0.7	0.7

Figure 4.4: Illustration of the Wav2Vec2-FC model. The Wav2Vec2-FC model converts speech signals into a series of speech frames (embeddings) and predicts the forced-aligned segmentation for each frame.

simplicity and convenience in ASR, it is not the optimal method for deriving precise temporal alignments.

In this study, the CTC-based phone recognizer serves two purposes. It could be deployed in the alignment pipeline to recognize phones directly from speech signals, thereby providing transcriptions to the neural forced aligner, achieving textless alignment. Secondly, as coarse phone-to-audio alignments can be derived from CTC outputs, the phone recognizer could also be used as a baseline for textless alignment.

4.7 Experiments

For the English dataset, I used the Librispeech (Panayotov et al., 2015) and the 1600h *Common Voice* 6.1 (Ardila et al., 2020), both of which were widely used ASR corpora of read speech. The TIMIT dataset with human annotations (Garofolo et al., 1993) was used for evaluation. The TIMIT was annotated with a phone set of 61 distinct phones. However, most forced aligners and G2P models are based on the CMU pronunciation dictionary, which has a phone set of 39. Therefore the original TIMIT 61 phone annotations were collapsed into the 39 CMU phone set. The flap [ɾ] `DX` did not have a one-to-one mapping, so I kept all occurrences. While this slightly degraded the overall accuracy, this was the same for all forced alignment methods under comparison except `WebMAUS`, which used more phones than those in the CMU phone set. All data and their train/test partitions were based on the HuggingFace `datasets` package (Lhoest et al., 2021). All textual transcriptions were converted into phone sequences using an open-source grapheme-to-phoneme (G2P) converter (Park and Kim, 2019).

In the experiments, I ran four iterations of training. I set the time masking probability to $[0.05, 0.2]$ for the 20ms models and $[0.05, 0.4]$ for 10ms models. The first two iterations were trained on Common Voice. I partitioned the Common Voice dataset into three sets $\{C1, C2, C3\}$, $C1$ containing sentences below 3 seconds, $C2$ with sentences between 3-5 seconds, and $C3$ consisting of sentences up to 10 seconds. The third iteration was trained on Librispeech and in the fourth iteration, the model was fine-tuned on the TIMIT training set. The model generally converged within 1,000 training steps so training on the full dataset was not needed. Annotated boundary information was never used during training. An evaluation was performed on the TIMIT test set, which does not overlap with any of the training data. The proposed model was coded in `pytorch` based on the Wav2Vec2 implementations in the `transformers` package (Wolf et al., 2020). By default, I used an effective batch size of 32 and the Adam optimizer with a learning rate of 1e-6 and rate decay of 1e-6. The training was done distributedly on three GPUs with 12 GB of memory.

I compared the proposed method with four publicly available forced aligners: `Montreal Forced Aligner (MFA)` (McAuliffe et al., 2017), `Penn Forced Aligner (FAVE)`

(Rosenfelder et al., 2011), WebMAUS (Kisler et al., 2012) and Gentle (Ochshorn and Hawkins, 2017)[3]. Specifically for MFA, I tested its performance both using a pretrained Librispeech model and performing forced alignment from scratch on TIMIT.

4.8 Evaluation

To assess the performance of the proposed models, I evaluated the forced alignment results with precision, recall, F1, and R-value (Kreuk et al., 2020a). For each predicted phone boundary, if the timing was within tolerance τ and the predicted phone matches, it was considered a hit, or True Positive (TP). If the model predicts a non-existent boundary, it is considered a False Positive (FP). Failure to predict an existing boundary is called False Negative (FN). The calculation of precision and recall are detailed below.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.8}$$

High precision means that a model tends to correctly predict a boundary when it makes a prediction, though it may refrain from predicting all boundaries. High recall implies that a model successfully finds most of the positive boundaries, but it could over-predict phone boundaries. The combination of precision and recall provides an overall picture of model performance.

The F1 score is the harmonic mean of precision and recall, which can give a comprehensive assessment of the classification quality. Let P be precision and R be recall. The F1 score can be calculated as follows.

$$\text{F1} = \frac{2(P \times R)}{P + R} \tag{4.9}$$

Sometimes F1 score can give overly optimistic results, as high precision and low recall can still result in high F1 (Michel et al., 2017). So it has been suggested that the R-value is a relatively robust measure to assess the segmentation quality (Räsänen et al., 2009; Kreuk et al., 2020b,a). Let $OS = R/P - 1$ and the R-value is defined as follows.

$$r_1 = \sqrt{(1 - OS)^2 + OS^2} \tag{4.10}$$

$$r_2 = \frac{-OS + R - 1}{\sqrt{2}} \tag{4.11}$$

[3] The alignments from FAVE and WebMAUS were kindly done by Cong Zhang.

$$\text{R-value} = 1 - \frac{|r_1| + |r_2|}{2} \tag{4.12}$$

Each boundary marked the onset and the offset of consecutive phones, so I only evaluated the phone onsets with a tolerance of 20ms. The above measures only assess the accuracy of phone boundaries. I also measured the overall quality of segmentation with the percentage of correctly predicted frame labels. The human annotations were discretized into frame-level labels at a temporal resolution of 10ms and the accuracy of predicting these phones at each frame is calculated. The rationale behind this approach is that if the boundaries are not correctly predicted but the midpoints or stable parts of the phone are correctly marked, these delineations are still useful for phoneticians to take acoustic measurements for certain sounds, notably fricatives and vowels.

4.8.1 Forced alignment

As shown in Table 4.1, all methods can perform forced alignment with good accuracy, though `Gentle` seems to lag slightly behind other methods. It should also be noted that the results often depend on the test materials. For example, Gonzalez et al. (2020) found that `MFA` produced the highest accuracy for their sociolinguistic interview data, whereas `WebMAUS` showed the poorest performance. These results conflict with the current results here, which could be due to the discrepancy in testing materials. The proposed forced alignment method also shows comparable performance, comparable to all existing forced alignment methods, if not better. In general, increasing the frameshift from 20ms to 10ms improved the alignment accuracy. However, phones in continuous speech often undergo a variety of phonetic processes such as deletion and insertion, resulting in pronunciation variants deviating from dictionary pronunciations. Yet most G2P converters can only provide dictionary pronunciations, which may lead to decreased performance. I also quantify the influence of a pretrained G2P converter on alignment accuracy. In Table 4.1, it is clear that using the dictionary pronunciations predicted by the G2P converter decreases the alignment performance by a large margin (5% $\sim$ 8%), yet the performance is still close to other public forced aligners. This also highlights the limitations of forced alignment, since it can only align to the canonical pronunciations while ignoring the actual phonetic variation in connected speech.

As different speech varieties may have varying phone inventories (like American English vs. Received Pronunciation), it would be desirable for the pretrained `W2V2-FS` model to adapt to different phone inventories. To test this, I can simply fine-tune the model after modifying the phone embeddings of the text encoder. For TIMIT, I expanded the existing 39 phone embeddings of the original `W2V2-FS` model to include the full TIMIT 61 phone set while keeping the rest of the pretrained weights intact, and fine-tuned the `W2V2-FS` model. The model converged quickly after a few hundred iterations and achieved performance comparable to that of the original model

Table 4.1: Evaluation results of text-dependent alignment

Model	P	R	F1	R-val	Overlap
Baseline models					
`FAVE`	0.57	0.59	0.58	0.64	74.3%
`MFA-Libris`	0.61	0.61	0.61	0.67	73.5%
`MFA`	0.62	0.63	0.63	0.68	75.0%
`Gentle`	0.49	0.46	0.48	0.56	67.7%
`WebMAUS`	**0.70**	**0.70**	**0.70**	**0.75**	78.8%
Proposed models					
`W2V2-FC-20ms-Libris`	0.49	0.47	0.48	0.56	73.8%
`W2V2-FC-10ms-Libris`	0.57	0.54	0.55	0.62	76.4%
`W2V2-FC-32k-Libris`	0.66	0.63	0.64	0.69	79.3%
`W2V2-FS-20ms`	0.47	0.49	0.48	0.55	71.6%
`W2V2-FS-10ms`	0.68	0.68	0.68	0.73	**80.4**%
`W2V2-FS-32k`	0.63	0.65	0.64	0.69	79.3%
Pretrained G2P converter					
`W2V2-FS-20ms`	0.40	0.42	0.41	0.49	65.1%
`W2V2-FS-10ms`	0.56	**0.58**	0.57	0.63	72.5
`W2V2-FC-32k-Libris`	**0.58**	0.57	0.58	**0.64**	**73.0%**
Phone set adaptation (`TIMIT-61`)					
`W2V2-FS-20ms`	0.49	0.53	0.51	0.57	70.5%
`W2V2-FS-10ms`	**0.66**	**0.70**	**0.68**	**0.72**	**79.7%**

Table 4.2: Evaluation results of text-independent alignment

Model	P	R	F1	R-val	Overlap
`W2V2-CTC-10ms`	0.31	0.29	0.30	0.42	43.9%
`W2V2-CTC-20ms`	0.31	0.30	0.31	0.42	46.6%
Phone recognition + `W2V2-FS`					
`W2V2-FS-20ms`	0.40	0.42	0.41	0.48	64.2%
`W2V2-FS-10ms`	0.56	0.58	0.57	0.63	71.5%
`W2V2-FC-32k-Libris`	0.57	0.57	0.57	0.64	72.2%
Direct inference from audio signals					
`W2V2-FC-20ms-Libris`	0.57	0.59	0.58	0.63	72.7%
`W2V2-FC-10ms-Libris`	0.55	0.58	0.56	0.62	72.5%
`W2V2-FC-32k-Libris`	**0.60**	**0.63**	**0.61**	**0.66**	**74.3%**

(see last two rows in Table 4.1). Yet `W2V2-FC` models might be less flexible in this regard, as they require pre-existing alignments to train.

4.8.2 Text-independent alignment

For text-independent alignment, I trained a phone recognizer on both Librispeech and Common Voice by fine-tuning `wav2vec2-base` using the CTC loss. This phone recognizer was used to derive phone transcriptions first before performing forced alignment with `W2V2-FS`. Using a phone recognizer works reasonably well (Table 4.2), as the alignment is only 1% less accurate than when using a pretrained G2P converter (see Table 4.1 for results on using a pretrained G2P converter). As expected, the alignments generated by the CTC loss alone (first two rows) deviated significantly from the ground truth, since the CTC loss does not encourage time-synchronous alignment. In contrast, `W2V2-FC` models achieved better performance (Table 4.2) than other text-independent models, highly comparable to that of text-dependent alignment methods. This shows that the frame classification model used in HMM systems is still robust with neural networks. Overall, these encouraging results suggest that both text-independent models are practical means to derive phone-to-audio alignments.

When the errors of G2P are taken into account, the current forced alignment methods might not possess a strong advantage over existing tools. The discrepancy could be accounted for by two causes. First, silence poses challenges to forced alignment as it is not encoded in the transcription. Without preprocessing silence, phones are often misaligned when a long silent interval is present. Secondly, I did not explicitly model pronunciation variants. However, HMM-based systems (Povey et al., 2011) often have built-in methods to alleviate these two issues. This will be the next step in my research.

4.9 Ablation analysis

I performed ablation analysis to examine the effectiveness of the `W2V2-FS` model. Ablation analysis is a common technique to assess neural network models in which individual components are deliberately removed to analyze their specific contributions to the final performance (Meyes et al., 2019). To assess whether the proposed training method and loss functions were helpful for the formation of diagonal alignment, I trained multiple models under different settings: 1) no curriculum training; 2) no $\mathcal{L}_{FS}$; 3) no $\mathcal{L}_m$. These were compared to the full model. As shown in Fig 4.5, both the contrastive loss and the forward-sum loss are necessary to facilitate the learning of sharp attention alignment. When the audio and the transcription were aligned using only the forward sum loss $\mathcal{L}_{FS}$, the alignment can still be learned but the overall quality was noisy, making it hard to extract unambiguous alignments from it. Removing either the curriculum learning or the contrastive loss generally causes the model to fail to converge to a diagonal alignment pattern. When curriculum learning was not implemented, the model converged to random patterns, mainly because directly learning alignments from long audios is challenging. If the forward sum loss $\mathcal{L}_{FS}$ is not used to guide the model, the model learns the alignment patterns that are not diagonal, though the alignments were sharp and clean. In summary, the ablation analysis shows that curriculum training and the forward sum loss are necessary for the formation of monotonic alignments whereas the contrastive self-reconstruction task is helpful for reducing the noises in the alignment patterns.

Specifically for the curriculum training, multiple iterations of training are generally needed to boost performance (see Table 4.3), as initial alignments tend to be suboptimal, especially for noisy crowd-sourced datasets like Common Voice. The model performance gradually increased in later iterations. During training, I found that the forced alignment model was *very* sensitive to initialization, and poor alignment was reinforced but not corrected in subsequent training. So using a good frame classification model to provide a good prior estimation was important. Audio quality had a large impact on the quality of alignment. Training on TIMIT or Librispeech-clean from scratch resulted in better alignments, faster convergence rates, and fewer iterations than training on the crowd-sourced Common Voice. The increasing accuracy in Table 4.3 illustrates how model performance can be incrementally improved through multiple iterations of fine-tuning.

4.10 Summary

The English aligners implemented using two deep learning-based methods show impressive performance in aligning audio to phones. With the powerful neural networks, the proposed methods outperformed most publicly available forced aligners in most evaluation metrics. The ablation analysis further demonstrates that the proposed model architecture (the forward-sum loss and the

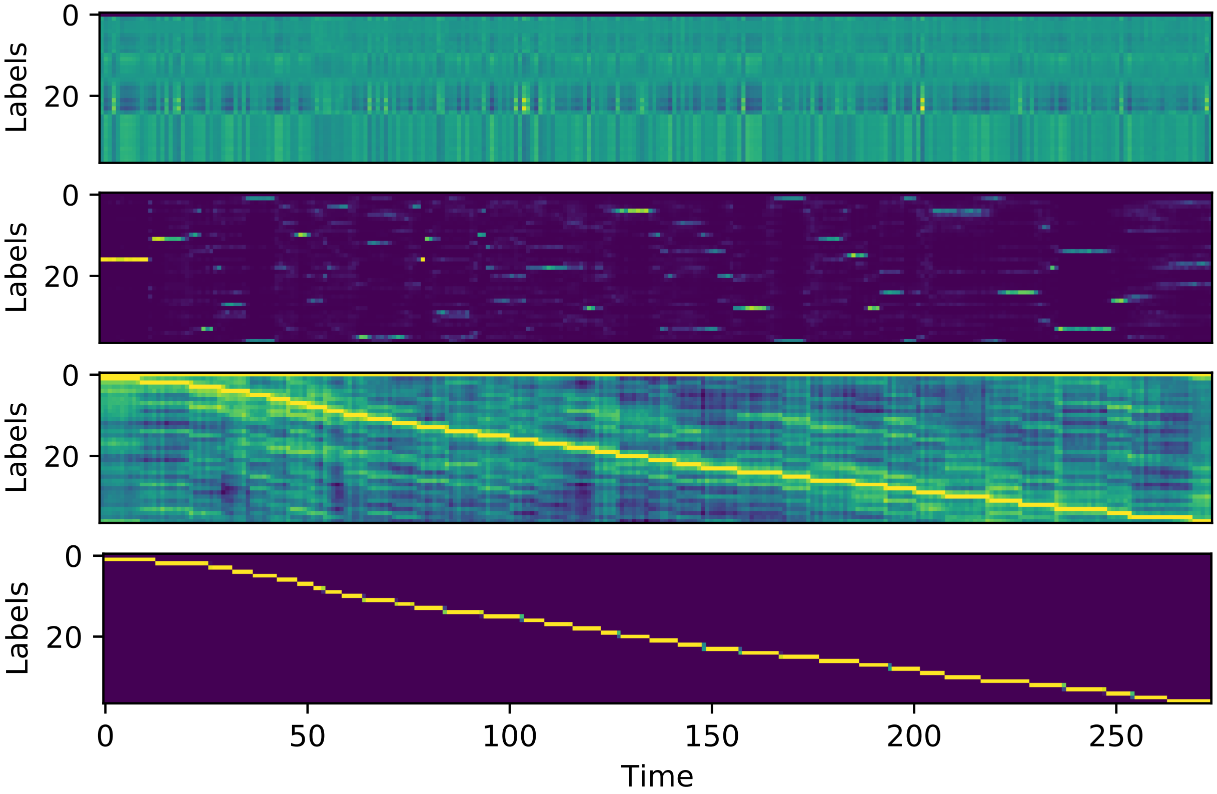


Figure 4.5: Sample alignments from the results of ablation studies in Section 4.9. From top to bottom: 1) No curriculum training; 2) No $\mathcal{L}_{FS}$; 3) No $\mathcal{L}_m$; 4) Full model.

Table 4.3: Evaluation of `W2V2` models from different iterations

Model	Iteration	Training Data	R-val	Overlap
`FS-10ms`	1	Common Voice	0.41	60.7%
`FC-10ms(w/o text)`	1	Common Voice	0.37	55.8%
`FS-10ms`	2	Common Voice	0.51	68.1%
`FC-10ms(w/o text)`	2	Common Voice	0.53	66.2%
`FS-10ms`	3	Librispeech	0.60	74.1%
`FC-10ms(w/o text)`	3	Librispeech	0.52	68.1%
`FC-10ms(w/ text)`	3	Librispeech	**0.63**	**75.3%**

self-reconstruction task) and the curriculum learning paradigm are necessary for the formation of alignment patterns during training. All these results strongly suggest that the proposed models are promising methods for constructing accurate and accessible phone-to-alignment systems.

4.11 Developing a textless aligner for Mandarin Chinese

4.11.1 Overview

Most forced aligners are developed for English but, for many languages, such a tool is lacking. In this section, I extended the frame classification model in the above sections to Mandarin Chinese. A few adjustments were made compared to the English phonetic aligner. For Mandarin, only the frame classification model (Wav2Vec2-FC) is adopted, as this model can perform both forced alignment and textless alignment. Wav2Vec2-FS, which can still be effectively applied to Mandarin, consumes more computational resources than the Wav2Vec2-FC because a phone encoder must be trained in addition to the speech encoder and multiple iterations with curriculum training are needed. Wav2Vec2-FS also relies on an external phone recognizer to perform textless alignment and this makes it even more costly in practical use. Given the above concerns, the relatively lightweight and robust Wav2Vec2-FC seems to be a better choice for Mandarin. In the following sections, training details of the Mandarin textless aligner are described.

4.11.2 Data

To create a large dataset for model training, I aggregated over 1000 hours of Mandarin speech from multiple Mandarin datasets, including MagicData (∼755 hours) (Magic Data Technology Co., 2019), Aishell-1 (∼150 hours) (Hui Bu, 2017), Aishell-3 (∼50 hours) (Shi et al., 2020), ST-CMDS (∼100 hours) (Surfingtech), Datatang (∼200 hours) (Beijing DataTang Technology Co., 2019), THCHS-30 (∼30 hours) (Dong Wang, 2015) and PrimeWords (∼100 hours) (Primewords Information Technology Co., 2018). All these data were resampled to 16kHz and utterances longer than 15 seconds were removed to avoid memory issues. Due to the uncontrolled nature of the dataset, spoken utterances were collected from a variety of recording devices under different recording conditions with varying degrees of background noise. The data encompass different genders, multiple age groups, and accents from almost all Chinese provinces. It is assumed that the diversity of speech utterances can make the resultant model more generalizable to unseen speakers, as scaling up the dataset generally increase ASR performance (Zhang et al., 2020; Xiao et al., 2021; Chan et al., 2021).

4.11.3 Grapheme-to-phoneme conversion

Mandarin is represented orthographically by ideogram-based characters that are phonetically irregular. A pronunciation dictionary is often used to convert Chinese characters to their phonetic transcriptions, *Pinyin Romanization*. Pinyin is preferred over IPA because it is the most widely adopted phonetic alphabet in the Chinese speech processing community and in other forced aligners such as FAVE (Rosenfelder et al., 2011) and Kaldi (Povey et al., 2011). As all Chinese characters are monosyllabic, the Pinyin Romanization only divides a syllable into the onset and the rhyme. Within a syllable, the onset is the consonant but is optional. The rhyme is the vocalic part of the syllable, including glides, monophthong, diphthong, and an optional nasal consonant coda. Rhymes are considered as stand-alone units themselves, and individual phones in the rhyme are not separated. All sentences were converted to Pinyin with a publicly available G2P converter g2pM[4] (Park and Lee, 2020).

4.11.4 Experiment details

Since Wav2Vec2 has only been pretrained in English and no pretrained speech transformer for Mandarin is available, I selected the multilingual pretrained XLS-R (Babu et al., 2021) as the speech encoder. As the multilingual and extra-large version of Wav2Vec2, XLS-R has been pretrained on about 372k hours of speech encompassing 128 different languages. Trained with the same methods as Wav2Vec2, XLS-R has exhibited significant improvements in a variety of speech-related downstream tasks, suggesting that it has learned effective feature representations for human speech in general. The superior performance of XLS-R motivates me to select it as the speech encoder for Mandarin. The pretrained XLS-R was wav2vec2-xls-r-300m, which is publicly available via HuggingFace Hub[5]. Due to computational constraints, I only select the smallest version of XLS-R with 300 million parameters.

The frame classification model must be trained in a supervised manner, so frame-level phone labels or pseudo-labels must be available. Pseudo-labels from forced alignments were created for all speech utterances in the aggregated corpus iteratively. The initial alignments for training the XLSR-FC were from the 85-hour Aishell-3 (Shi et al., 2020), which were obtained through Montreal Forced Aligner (McAuliffe et al., 2017) and released publicly[6]. First, the XLS-R model was trained to predict the forced aligned pseudo-labels on the whole Aishell-3 speech corpus. Then forced alignment was performed on the rest of the speech corpus with the pretrained XLS-R frame classifier using the same method in Section 4.5. In this way, pseudo-labels for ∼1300 hours

[4] https://github.com/kakaobrain/g2pM

[5] https://huggingface.co/facebook/wav2vec2-xls-r-300m

[6] https://github.com/ming024/FastSpeech2

of Mandarin were created to train a more powerful frame-level phone classifier. For the forced aligned corpus, 500 utterances were partitioned as the test set and the rest were used for training.

With $\sim$1300 hours of forced aligned speech, I experimented with three models with different hyperparameters for the Mandarin phonetic aligner.

- **XLSR-FC-20ms**. This model is the same as the original XLS-R, except that a classifier head is added on top of the transformer and the quantizer was removed. The model was initialized with weights from XLS-R.
- **XLSR-FC-10ms**. This model has the same settings as the XLSR-FC-20ms described above. The only difference is that the stride of the last convolutional layer was set to 1 instead of 2. This change of hyperparameter upsamples the temporal resolution of the model to about 10ms while allowing the pretrained weights to be reused without any conflicts.
- **Wav2Vec2-FC-tiny-10ms**. The large size of XLS-R renders it slow in making inferences, especially when running on CPUs. To provide a lightweight model, I have also trained a tiny version of Wav2Vec2. The model was first pretrained on all the Mandarin speech in the same pretraining methods as the original Wav2Vec2 (Baevski et al., 2020), as pretrained weights tend to increase downstream performance and reduce the training time. Before pretraining, all weights were randomly initialized but the quantizer reused the pretrained weights from XLS-R (Babu et al., 2021) and remained frozen throughout training. After pretraining, the quantizer was discarded, leaving only the encoder weights. For this model, the hidden dimension was set to 386 and the number of hidden layers to 6, while the rest of the hyperparameters were the same as `wav2vec2-base`.

The effective batch size was set to 256 through gradient accumulation. The AdamW optimizer with a learning rate of $3e - 4$ was used with a warm-up step of 1000 iterations. All models were trained on a single A100 with 48GB of memory for up to 2 epochs. On average, it took about two days to complete a training loop.

4.11.5 Results

Model performance was evaluated with the overlap with forced aligned labels, as manual annotation of Mandarin utterances was not publicly available. The overlap measures the percentage of correctly predicted frame-level phone labels, which were generated with the initial XLS-R forced aligner and discretized with a temporal resolution of 10ms. Table 4.4 shows that models with more parameters are better at recognizing and segmenting phones, even when the large model has lower temporal resolution than the small model (XLSR-FC-20ms vs. Wav2Vec2-FC-tiny-10ms). Overall, XLSR-FC-10ms achieves the best performance in time-aligned phone recognition, suggesting

Table 4.4: Evaluation results for the Mandarin frame classification models.

	% Overlap
XLSR-FC-20ms	93.7%
XLSR-FC-10ms	94.7%
Wav2Vec2-FC-tiny-10ms	91.5%

that a large-scale model and high temporal resolution are important for improving recognition results. Yet the tiny model can segment audio samples much faster than large models with only a minor loss in accuracy. Compared to the English results, results for Mandarin might seem much better. However, this is because the evaluation labels were from the forced-aligned segmentations, not from human segmentations due to the lack of data. In addition, the Mandarin aligner only segments the speech into onsets and rhymes while rhymes can consist of multiple vowels or glides, or a nasal coda. The English aligner, in contrast, segments the speech into individual phones, which is more challenging, thereby showing lower accuracy.

A sample textgrid illustrating different segmentations is shown in Figure 4.6. Inspection indicates that automatic segmentation, whether it is forced alignment or text-independent phone segmentation, is very close to human segmentation. Errors do occur but they mostly occur at the boundaries, whereas the stable part of phones is usually correctly recognized. For phonetic studies that aim to analyze vowels, misplacement of phone boundaries probably does not matter too much if most of the vowel segment is correctly identified. The first two syllables present a particularly interesting case to illustrate the difference between forced alignment and text-independent phone segmentation. The original syllable undergoes tone-3 sandhi due to the presence of the third syllable, such that the Pinyin ‘ka3 er3 pu3’ [ka˨˩˦ ɜɹ˨˩˦ pʰuː˨˩˦] becomes ‘ka2 er2 pu3’ [ka˧˥ ɜɹ˧˥ pʰuː˨˩˦]. Moreover, the actual production is highly rhoticized to the extent that the vowel ‘a3’ has almost completely merged into the rhotic sound ‘er3’ [ɜɹ˨˩˦]. Such deviations from the dictionary pronunciations pose some challenges for forced alignment, which can only align speech signals to the canonical pronunciations. The forced aligner attempted to align the speech for ‘ka2 er2’ [ka˧˥ ɜɹ˧˥] to the dictionary pronunciation ‘ka3 er3’ [ka˨˩˦ ɜɹ˨˩˦], only to result in inaccurate segmentation (the ‘er3’ [ɜɹ˨˩˦] was only aligned to one frame, hence not visible in the textgrid). In comparison, the textless aligner predicts the strong rhotic sound in the acoustic signal as ‘r’ [ɹ], which is closer to the actual pronunciation. Without the constraint of transcriptions, the textless aligner is more flexible in dealing with such phonetic variation in connected speech.

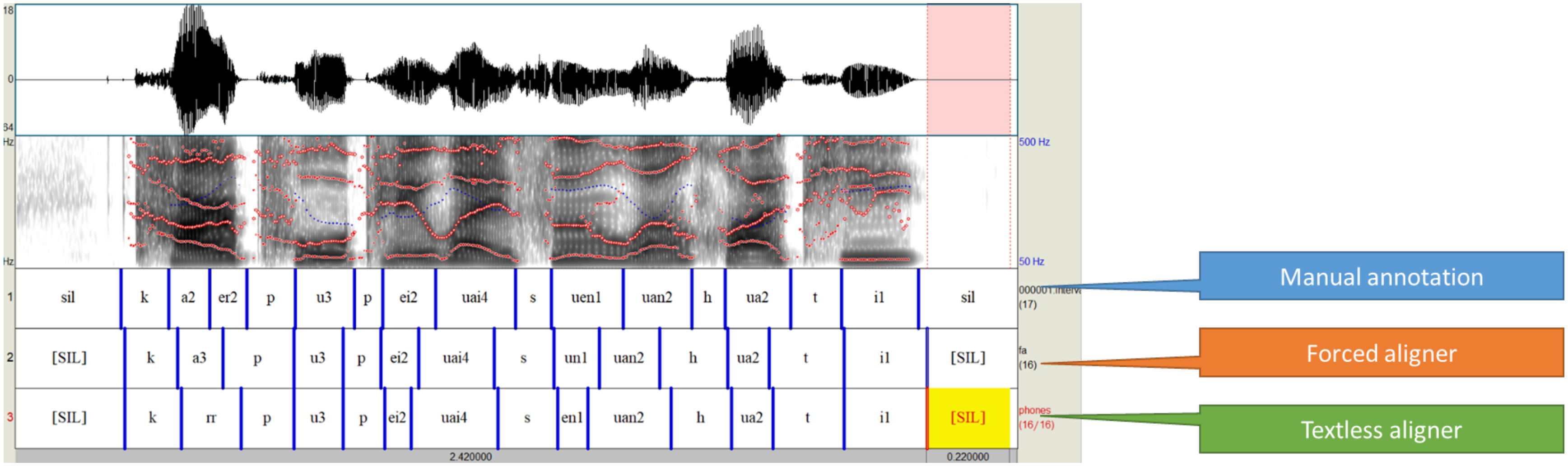


Figure 4.6: Sample textgrids of human annotation (**top**), forced aligned segmentation (**middle**) and predicted segmentation (**bottom**).

4.12 Discussion

In this study, I present two deep learning-based methods to align audio to phones and achieve comparable, if not better, performance against several existing forced alignment tools. In comparison with the publicly available forced aligners, the proposed neural aligners are more accurate in aligning phones to speech. In addition to forced alignment, the proposed model can also perform text-independent phone segmentation (textless alignment) to directly segment raw audio signals, an end-to-end annotation method that has not been widely adopted in the field due to the lack of appropriate tools. The development of the current phonetic aligners will provide a powerful toolkit that can potentially facilitate the processing of data for phoneticians and speech engineers.

The example in Figure 4.6 also shows that textless alignment is more flexible in coping with phonetic variation, as it directly predicts what is actually articulated rather than what should be articulated. Compared to forced alignment which can only map speech onto dictionary pronunciations, the textless alignment can further be used to spot and track phonetic variations in the wild. Altogether, these models can be deployed as a pipeline to bootstrap phone labels from naturalistic audio data in the wild such as YouTube, Podcasts, and radios, all of which contain massive audio data but remain under-exploited in speech research. Given the good performance of `W2V2-FC-Libris`, this method can also be combined with MFA(McAuliffe et al., 2017) to train a large-scale text-independent frame classification model to get rid of text in inference. The availability of easy-to-use and accurate aligners will have tremendous implications for speech corpus creation, phonetic research, and speech technology. The W2V2 models can be accelerated by harnessing the computing power of GPU. For example, in the inference mode, the proposed model aligned the 1600 hours Common Voice dataset in about 10 hours on a 2080Ti GPU without batching. Given the good performance on phone segmentation that requires much precision, the model can also perform well on related tasks such as word and sentence segmentation, which partially depend on obtaining good phone segmentation first.

The aligner developed in this chapter is available at https://github.com/lingjzhu/charsiu as freely available software. A tutorial for linguists can also be found at https://colab.research.google.com/github/lingjzhu/charsiu/blob/development/charsiu_tutorial.ipynb.

4.12.1 Limitations

Most automatic alignment tools are developed based on speech produced by healthy adults with a particular accent. It remains a challenge to transfer the knowledge in these tools to speech produced by specific population groups, particularly children, older speakers, and people with speech pathologies. For example, when processing child speech, the current forced alignment tools only work well on certain classes of sounds, such as vowels and fricatives, and they only work for older children (Knowles et al., 2018; Mahr, 2021). The bias in the training data might also have profound implications for social justice in deployment. It has been shown that current ASR engines tend to bias against the speech patterns of racial minorities (Koenecke et al., 2020; Mengesha et al., 2021). The proposed model in this study is no exception, as it is mostly trained on "mainstream" dialects and accents with dictionary pronunciation, so it may bias towards certain accents. For example, both Mandarin and English aligners rely on the phone set from the "standard" accent, such that sounds not present in these accents will be forcefully mapped onto the closest phones in the phone set. This could smooth out some meaningful sociolinguistic variation and cause biases in phonetic research.

4.12.2 Future directions

The diversity of human speech across languages poses a great challenge to speech processing. Developing computational methods that can work with crosslinguistic data will have important implications for studying low resource languages and language varieties. Yet simply recognizing phones in an utterance might not be the most helpful tool for phonetic studies, as a fine-grained acoustic analysis requires precise time-alignment between symbolic phonetic units and acoustic signals. While my proposed alignment tools have enabled deriving such phone-to-audio alignments easily, these methods only work well for resource-rich languages such as English and Mandarin Chinese, because neural networks are usually data-inefficient, requiring a large amount of data to train. For low resource languages, the lack of data makes it hard to train any machine learning models. Therefore, more innovative and data-efficient methods are needed to develop speech technology for low-resource languages. In the face of these issues, devising an effective scheme to annotate and represent multilingual speech variation is the goal of continued studies. In the future, I will continue to improve the alignment accuracy and extend it to different languages.

CHAPTER 5

General Discussion

5.1 Discussion

This dissertation investigates how large-scale data and computational methods can be harnessed to address issues of sociolinguistic and phonetic variation in text and speech. In this section, I will discuss the implications of the individual studies that comprise this dissertation. Building on these case studies, I also propose some future directions that can advance the field of computational sociolinguistics and computational phonetics.

5.1.1 Summary

This section summarizes the three individual studies conducted in this dissertation.

In Study 1 (Chapter 2), I focus on the understudied effect of social networks on lexical change. I conducted a large-scale analysis of over 80k neologisms in 4420 online communities across a decade. Using Poisson regression and survival analysis, this study demonstrates that a community's network structure plays a significant role in lexical change. In addition to overall community size, network properties including dense connections, the lack of local clusters, and more external contacts are shown to promote lexical innovation and retention. Unlike offline communities, these topic-based communities do not experience strong lexical leveling despite increased contact but rather tend to accommodate more niche words. The analysis provides support for the broad sociolinguistic hypothesis that lexical change is partially shaped by the structure of the underlying network but also uncovers findings specific to online communities.

In Study 2 (Chapter 3), I introduce a new approach to studying idiolects through a massive cross-author comparison to identify and encode stylistic features in written texts. The neural model achieves strong performance on authorship identification for short texts and through an analogy-based probing task, showing that the learned latent spaces exhibit surprising regularities that encode qualitative and quantitative shifts of idiolectal styles. Through text perturbation, I quantify the relative contributions of different linguistic elements to idiolectal variation. Furthermore, I provide a description of idiolects by measuring inter- and intra-author variation, showing that variation in idiolects is often both distinctive and consistent.

In Study 3 (Chapter 4), which addresses a foundational issue in phonetic variation and analysis, I present two Wav2Vec2-based models for both text-dependent and text-independent phone-to-audio alignment. The proposed Wav2Vec2-FS, a semi-supervised model, directly learns phone-to-audio alignment through contrastive learning and a forward sum loss, and can be coupled with a pretrained phone recognizer to achieve text-independent alignment. The other model, Wav2Vec2-FC, is a frame classification model trained on forced aligned labels that can perform both forced alignment and text-independent segmentation. Evaluation results suggest that, even when transcriptions are not available, both proposed methods generate results that are very close to those of existing forced alignment tools. In addition to the English phonetic aligner, I also report the development of a phonetic aligner for Mandarin Chinese with the same method. This work presents a neural pipeline of fully automated phone-to-audio alignment.

5.1.2 Implications

The abundance of large-scale texts and human behavior data in online communities can inform theories of sociolinguistic variation and change, particularly in newly emerging online communities. In traditional sociolinguistics, age, gender, geographical region, and social class are among the most widely studied variables. However, in online communities, researchers have access to other social variables that might also be related to language use, including social networks, user activity patterns, online platforms, and more. In Studies 1 (Chapter 2) and 2 (Chapter 3), the availability of the whole Reddit community and Amazon reviews allows one to observe lexical change at scale in real-time over a decade, thereby providing a panoramic view of sociolinguistic change in the online space. Bloomfield (1933) envisioned that the study of change can greatly benefit if we have complete records of how every word is used in a speech community. In light of this vision, this dissertation shows that the detailed records of web data provide a perfect testing ground for theories of variation and change at a scale that could not be matched in traditional sociolinguistics. For language change, Study 1 (Chapter 2) observes real-time lexical change at the scale of months among tens of millions of users, achieving a finer time scale than most studies of change in historical sociolinguistics, which are typically only able to study change on a decade-long—or much longer—scale. The huge repertoire of review databases also enables me to study individual variation on a background population of more than a hundred thousand users, which could provide a more detailed and comprehensive view of idiolectal variation than studies that, due to practical constraints, only survey a dozen or so users. While I only investigated the specific topics of lexical change, social networks, and individual variation, these data, and this approach can also be applied to more theoretical questions in sociolinguistics, such as how online users learn and adapt to language in their life cycles (Danescu-Niculescu-Mizil et al., 2013), how sociolinguistic interactions take place among different users (Zhang et al., 2018), and how morphological productivity is actively used by various users and communities (Hofmann et al., 2020). The potential list of topics is far from being exhausted and the full potential of big data to play a more important role in sociolinguistics has yet to be unleashed.

The advancement of speech processing and NLP technologies has tremendous potential to enable fine-grained linguistic analysis. In traditional corpus linguistics, frequency and POS tagging have become commonly used techniques to analyze linguistic structures (McEnery and Hardie, 2011), but they still can only extract surface-level linguistic structures. However, the recent development of NLP has made available more toolkits that allow for the analysis of "deeper" linguistic structures, such as dependency parsing, discourse parsing, dialog act labeling, and pretrained language models, among others. A few million words or one hundred hours of speech were considered a large amount of data in traditional linguistic inquiry. But parallel computing and fast methods of tokenizing, parsing and digital speech processing have made it possible to process large-scale data from Reddit (2TB), Amazon (10GB), and Common Voice (500GB), reducing the computing time from months to days. In Study 2 (Chapter 3), the use of deep learning models was also capable of extracting more complex statistical linguistic patterns in idiolects, compared to frequency-based statistical modeling. All three studies presented in this dissertation illustrate that state-of-the-art technologies in speech and language processing, if applied appropriately, can yield in-depth linguistic analyses that had not been thought possible in many traditional linguistic inquiries.

This dissertation also has the potential to extend the analytic toolkits of computational sociolinguistics. Currently, most studies on computational sociolinguistics focus predominantly on textual variation, especially variation in online social networks, including Twitter, Reddit, and Wikipedia (Nguyen et al., 2016; Hovy and Yang, 2021). Yet phonetic and phonological variation has received comparatively little attention in the computational paradigm (e.g., Suni et al., 2019; Strycharczuk et al., 2020; Gittelson et al., 2021; Bartelds et al., 2022), though speech is the most widely studied linguistic element in traditional sociolinguistics. The under-representation of speech in computational sociolinguistics is, presumably, partially caused by the challenge of collecting and annotating speech data. Constructing a large-scale text dataset is relatively easy, since texts are easy to store, abundant on the web, and do not require complicated annotations. In contrast, speech data take up much more storage space and computing resources and are harder to collect and analyze. Unlike texts, speech, as represented by a list of numerical arrays, is semantically opaque and therefore requires time-consuming annotations to discrete into analyzable formats. The development of phonetic aligners that can generate textgrid segmentation has the potential to greatly simplify the pipeline of audio data preprocessing, making it possible to extend the phonetic analysis to many unexploited audio sources that have become an ineligible part of our society, including YouTube videos and Podcasts (Clifton et al., 2020). These user-generated contents are highly varied in sociolinguistic style, topic and affective state, which could be a great source to study sociolinguistic variation and self-presentation in online communities. Deriving time-aligned annotation has been a fundamental yet time-consuming step in speech research, and current phonetic aligners provide a flexible and powerful tool to facilitate this process, thereby making it possible to create and analyze large-scale speech datasets easily. In contrast to currently available tools, my work provides speech and language researchers with a strong deep learning-based tool to obtain phone segmentation even when textual transcriptions are not available.

5.2 Limitations and ethical concerns

This dissertation is inherently limited in several aspects.

5.2.1 Data

Traditional sociolinguistics tends to suffer from insufficient data and the challenge to collect real-life interactions without the influence of the observer's paradox (Labov et al., 1981). The big data paradigm in computational sociolinguistics can effectively increase the statistical power and reliability of analysis compared to more traditional sociolinguistic inquiries. However, this by no means implies that computational sociolinguistics is less biased and less restricted than traditional sociolinguistics. Online texts such as Tweets, reviews, and forum posts are the predominant sources of data in most computational sociolinguistics studies because these data are easily accessible and in huge quantities (Nguyen et al., 2016). This dissertation is no exception.

The scope of the current dissertation is unavoidably constrained by the sources of data. It is important to be aware that online communities are only a small fraction of human societies and the language used in these online contexts may not generalize to offline linguistic interactions, nor even to other online communities not covered in the study. Study 1 relies on the historical data from Reddit, but Reddit users are more likely to be young, urban, white, and male (Duggan and Smith, 2013) than the rest of the population, not to mention that there are many other online platforms that work differently from Reddit. The conclusions drawn in Chapter 2 merit more validation from other online platforms to understand whether platform-specific designs implemented will affect the network dynamics and hence the language dynamics. The same sampling biases are also faced in Study 2. Amazon users whose opinions are polarized towards the products under review are more likely to leave reviews than users who only hold moderate opinions, such that sentiments in Amazon reviews exhibit a ”J-shape” distribution (slightly more positive than negative) (Hu et al., 2009). Only a small fraction of users who authored at least five reviews were selected for analysis. It is assumed that this small fraction of users can generalize to other users, and more evidence is needed to test this. For Study 3 (Chapter 4), the training data for ASR models tend to be sociolinguistically homogeneous, making them less generalizable to diverse variations related to social variables including age, gender, race, and geographical locations (Koenecke et al., 2020; Martin, 2021; Feng et al., 2021). Yet it should be noted that these are not the problems of this dissertation alone, but problems of almost all neural speech and NLP models (see Hovy and Prabhumoye, 2021, for a summary).

Language diversity is also an issue. Most of the experiments were conducted in English, except for Study 3, in which Mandarin Chinese was also investigated. The limited diversity of languages also restricts the conclusions drawn in this dissertation, as a change in language structure could also change the dynamics of language use. It is expected that the statistical methods developed in this dissertation do generalize to multiple languages, as the common structures of language can be well captured by neural networks optimized for structured sequence processing (e.g, Libovický et al., 2020; Zhao et al., 2021), such as Wav2Vec2 and RoBERTa models. However, it remains unclear whether other languages interact with social networks as English does, whether idiolects exhibit different patterns of linguistic variations cross-linguistically, and whether and how diverse phonological structures and phonetic variation affect the accuracy of forced alignment. While the choice of English as the main focus is due to the lack of resources and feasibility considerations, the conclusions of this dissertation would be greatly strengthened if more languages were surveyed.

5.2.2 Interpretability

Interpretability of machine learning models, especially neural networks, has remained a challenging topic that has been actively investigated in the research community (Murdoch et al., 2019). Despite their huge success in many complex tasks, deep learning models are notorious for being black boxes, due to their huge parameters and the complex interactions in the intermediate layers (Zhang and Zhu, 2018; Zhang et al., 2021). For most NLP tasks, model interpretability is sacrificed for performance boost and the recent trend to scale up pretrained language models also scales up the difficulty for interpretability (Bommasani et al., 2021). In this dissertation, while many state-of-the-art machine learning models were employed to analyze language data, it still remains hard to understand how these models make decisions based on the complex interactions between language-internal and language-external factors.

In Study 1 (Chapter 2), gradient boosting trees and deep survival analysis are highly effective in predicting lexical innovation and survival, yet I have to rely on the weaker but more interpretable

Poission regression and Cox models to understand the contributions of individual network variables. However, these interpretable models often make simplifying assumptions about the data, so they might fail to fully capture the complex relationships between network attributes and language change. Not being able to analyze the best-performing model could imply that some interactions in network variables are not receiving full attention in the discussion. The challenge is greater for Study 2 (Chapter 3), where a deep neural network is the main approach to extracting idiolects. The underlying models for SBERT and SRoBERTa have 110 million and 125 million parameters respectively, making it almost impossible to isolate linguistic aspects that contribute most to idiolectal styles. To interpret the model, I made controlled manipulations on the input to probe the models' responses to these structurally controlled inputs. While this method did allow me to gain some understanding of the inner workings of the proposed models, the resultant interpretation still remains indirect and simplified. Despite the extensive experiments, our understanding of idiolectal styles is still limited because of the black-box nature of the computational method. It is still difficult to identify the individual linguistic variants that incrementally build up the unique idiolectal style. Study 3 (Chapter 4) took a purely speech engineering approach to model phones in context, so the interpretability issue received little attention throughout the study. There is little understanding, though, as to how the Wav2Vec2 speech encoder extracts phonetic information and which phonetic properties are used in the recognition processes. Yet understanding the inner workings of the model not only enables us to design better models but also allows us to learn about the interactions among phonetic properties in highly variable connected speech.

Despite the immense power of computational methods to analyze large-scale language data, their lack of transparency has greatly limited the wide application of computational methods to questions in theoretical linguistics (Baroni, 2020; Linzen and Baroni, 2021). While some attempts were made to interpret neural networks, including input manipulation and probing, interpreting large-scale models still remains a huge barrier for such research. It is hoped that the future development of interpretability studies will address these issues and make machine learning even more conducive to the study of linguistic theories.

5.2.3 Correlation and causality

For any computational analysis of naturalistic data, it is important to be aware that correlations do not necessarily imply causation. Most machine learning methods exploit the statistical correlations between variables such that one set of variables can be used to predict the other set of variables. But these predictive performances in NLP tasks do not always warrant causal interpretation and making casual inference is particularly challenging for language data (Wood-Doughty et al., 2018; Keith et al., 2020; Feder et al., 2021), which could lead to difficulties in computational sociolinguistics, a discipline that aims to understand language variation. The studies presented in this dissertation (especially Study 1) are based on observational data collected in real communities with little experimental control. The relations between network structures and lexical change could be correlational and I must rely on prior literature on variation and change to interpret the results. That is, I assume that networks should have an impact on language rather than the other way round, as networks have been found to play a role in information propagation (Lerman and Ghosh, 2010; Guille et al., 2013). Despite support from prior work, strong causal interpretation should generally be avoided as much as possible. Rather, it would be more appropriate to claim that the current evidence is in favor of the hypothesis that networks partially shape language change. The application of causal analysis could partially address this issue (Stewart and Eisenstein, 2018; Feder et al., 2021), yet the tension between correlation and causation persists in computational sociolinguistics and computational social science in general. It is hoped that future research in causal analysis (Wood-Doughty et al., 2018; Keith et al., 2020; Feder et al., 2021) and experimental methods (e.g., Centola, 2010) could facilitate the inference of causal relations in big data.

5.2.4 Ethical considerations

NLP research involves large data and powerful models often carry some ethical concerns (Leidner and Plachouras, 2017; Bender and Friedman, 2018; Jobin et al., 2019; Bender et al., 2021). The neural network models proposed in this dissertation are mainly used as a tool to understand language or to facilitate linguistics research. However, applying these models in various settings could have unintended consequences, such as reinforcing existing current demographic biases (Hovy and Prabhumoye, 2021). As mentioned in section 4.12.1, it is known that ASR systems tend to bias against non-mainstream accents such as African American Vernacular English (AAVE), exhibiting higher error rates when dealing with phonetic variations related to races and age, for example (Koenecke et al., 2020; Martin, 2021; Feng et al., 2021). One major source of the biases is the existing training data, which were mainly collected for "standard" English and clear speech. Training ASR models on these varieties will make models less robust to the phonetic and grammatical variations in real life, thereby working less well for some under-represented groups. The phonetic aligner in Study 3 is trained on such data and may not perform well on other varieties of English. Applying the phonetic aligner to some varieties may give biased results, such as smoothing out some of the existing sociophonetic variation. If this aligner is also applied in the pipeline of practical applications like speech synthesis and pronunciation, the biases could potentially propagate to downstream models without proper supervision.

Chapter 3 also raises some privacy concerns if the neural models are used to actively track online users' identities, especially given that the proposed SRoBERTa model has achieved state-of-the-art performance in identifying personal identities from texts against other models. While computational stylometry has some forensic applications including tracking criminals in online activities (Grant and MacLeod, 2020), authorship verification in online social networks, if put to malicious use, may weaken the anonymity of some users, leading to potential privacy issues. The results also show that caution should be taken when deploying these models in forensic scenarios, as different models or tokenizers might show different inductive biases that may bias toward certain types of users. Another potential bias is that I only selected a small group of the most productive writers from the pool (less than 20% of all data), but this sample might not necessarily represent all populations. I urge that caution should be exercised when using these models in real-life settings.

All of the experiments were performed on public data, in accordance with the terms of service. All social media users in the datasets were anonymized. In Chapter 3, the term “Siamese” may be considered offensive if interpreted as referencing some groups of people. The use of this term follows the research naming of a mathematical model in machine learning literature. I use the word here to refer to a neural network model and make no reference to particular population groups. In Study 1, a great number of low-frequency neologisms collected from `Urban Dictionary` may be considered offensive to specific groups of populations. I collected the word usage data as they were in order to recover as realistic of a lexical landscape in Reddit as possible. However, these offensive words by no means reflect the author’s values. Nor do I endorse the use of these words.

Given these ethical concerns, I still consider that the benefits of these studies outweigh the potential dangers. For deep learning-based stylometry in Study 2, while many studies focus on improving performance, I provide insights into how some of these models make decisions and expose some of the models’ biases. The interpretability and bias analysis could be used to guide the proper use of these methods in decision-making processes. The analysis could also be useful in developing adversarial techniques that guard against the malicious use of such technologies. Although the phonetic aligner in Study 3 might have higher error rates for some under-represented speech varieties, it nonetheless provides a potentially useful tool to facilitate phonetic research. Ideally, in speech research, the automatic alignment results will undergo rigorous checks to verify their fairness and correct mistakes. Yet these models will not be the end of my research. It is my hope that these demographic biases will be mitigated in my future research.

5.3 Future research

This dissertation seeks to understand selected patterns of language variation and change through computational methods. While it answers some research questions, it also leaves open many more questions in computational phonetics and computational sociolinguistics. These unanswered questions will be a good initial step for future research. Here I discuss several future research directions that build upon the current dissertation.

5.3.1 Cross-linguistic and unsupervised phone recognition and alignment

The phonetic aligner in Study 3 works well for languages with rich language resources. Yet the proposed method might not work well for many low resource languages across the world, in which speech data and their transcriptions are generally lacking. One of the paramount challenges of analyzing speech variation in the world's languages is the lack of a unified dataset and a common annotation method for representing speech sounds. The recent development of a cross-linguistic grapheme-to-phoneme converter (e.g., Zhu et al., 2022a) can help mitigate the phone transcription issue and the availability of the UCLA Phonetic Corpus (Li et al., 2021) and the Common Voice (Ardila et al., 2020), a crowd-sourced multilingual speech corpus covering more than 72 languages, could provide some initial multilingual data. Building on these data and the current phone-to-audio alignment models, I will continue to explore the possibility of multilingual phone-to-audio alignment methods through bootstrapping from existing language data.

Given the challenges of speech data annotation, I hope to develop unsupervised methods to exploit the inherent structures of speech to perform automatic phone segmentation. Traditionally, speech recognition datasets require intensive human transcriptions to create paired speech and text data. Yet in recent years, unsupervised speech recognition, that is, training speech recognition on speech signals without paired text or phone labels has shown great potential for low resource and multilingual speech recognition (Yeh et al., 2019; Liu et al., 2018, 2021; Baevski et al., 2021; Liu et al., 2022b). Unsupervised speech recognition is based on the framework of generative adversarial networks (Goodfellow et al., 2014). If phones can be induced from raw acoustic signals, the next step is to explore grouping the segmented phones into spoken word units. In this way, it is possible to build a vocabulary for languages without a writing system.

Many linguistic variations and ongoing changes can be subphonemic or paralinguistic, which might not be captured by phonemic transcriptions alone. Another promising approach is to develop computational models to induce phonetic representations in a data-driven and self-supervised manner by designing various tasks for speech pretraining and representation. The data-driven phonetic representations may provide more nuanced characterizations of sociophonetic variation. The multilingual model could theoretically deal with any language variety. Studying unsupervised phone segmentation will enable us to better understand the rich linguistic structures encoded in speech signals and empower (socio-)phoneticians with tools to facilitate the analysis and documentation of low-resource and under-represented languages.

5.3.2 Sociolinguistic variation in multimodal communication

In real-life communication, multiple aspects of language, be it textual, acoustic, or visual, are used jointly to achieve a communication goal. Speech and textual communications are at the center of current sociolinguistic analysis, yet research on how socially structured variation is reflected in multimodal behaviors and how these cues are employed to convey and shape one's social identity has only begun to take shape. For example, in an ongoing study of multimodal communication, I seek to capture the interactions between social categories (e.g., gender race) of comedians and their stand-up comedy performances by taking into consideration textual, acoustic, and visual aspects. By taking my research in this direction, I will extend my research experience in text and speech analysis to visual features. The goal is to see how classic sociolinguistic variables like gender, race, age, social class, and power will affect the use of multimodal cues in communication and, as in the third-wave paradigm (Eckert, 2012), how these cues are jointly employed to construct social identities. To work with large-scale naturalistic data, I am planning to go beyond the traditional hypothesis-testing paradigm and explore causal inference in the analysis of large-scale naturalistic data.

5.3.3 Incorporating sociolinguistic variation in NLP

As discussed in Section 5.2, most of the current speech and language technologies suffer from biases, especially demographic biases (Hovy and Prabhumoye, 2021). A reason for the existence of these biases is that NLP models do not explicitly take into account sociolinguistic variation, but assume that language is a homogeneous whole (Hovy and Yang, 2021). The neglect of sociolinguistic variation permeates the NLP research pipeline, ranging from data collection to experiment design. Consistent with this dissertation's theme of sociolinguistic (and phonetic) variation and change, it is hoped that sociolinguistic variation could be incorporated into NLP research. Sociolinguistic knowledge could be used to guide the collection of more balanced and fair language data that are representative of different demographic groups. Such datasets can also be used to evaluate the robustness of NLP models with respect to language variation and social biases. On the system side, it is important to develop NLP systems that are socially aware, such as language generation or question answering systems that can output sociolinguistically diverse texts. This dissertation proposes some methods for analyzing sociolinguistic variations and they may be of use in developing socially aware technologies. It is believed that these efforts will eventually make speech and language technologies more equitable.

BIBLIOGRAPHY

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Afra Alishahi, Grzegorz Chrupała, and Tal Linzen. 2019. Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *Natural Language Engineering*, 25(4):543–557.

Eduardo G Altmann, Janet B Pierrehumbert, and Adilson E Motter. 2011. Niche as a determinant of word fate in online groups. *PloS one*, 6(5):e19009.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA. PMLR.

Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. 2005. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24):3927–3944.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Quentin D Atkinson, Andrew Meade, Chris Venditti, Simon J Greenhill, and Mark Pagel. 2008. Languages evolve in punctuational bursts. *Science*, 319(5863):588–588.

Sarah Babinski, Rikker Dockum, J Hunter Craft, Anelisa Fergus, Dolly Goldenberg, and Claire Bowern. 2019. A robin hood approach to forced alignment: English-trained algorithms and their use on australian languages. *Proceedings of the Linguistic Society of America*, 4(1):3–1.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale.

Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. 2012. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42.

Rohan Badlani, Adrian Łancucki, Kevin J Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro. 2021. One tts alignment to rule them all. *arXiv preprint arXiv:2108.10447*.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021. Unsupervised speech recognition. *arXiv preprint arXiv:2105.11084*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

George Bailey. 2016. Automatic detection of sociolinguistic variation using forced alignment. In *University of Pennsylvania Working Papers in Linguistics: Selected Papers from New Ways of Analyzing Variation (NWAV 44)*, pages 10–20. York.

Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528.

Maciej Baranowski. 2013. Sociophonetics. In *The Oxford handbook of sociolinguistics*, pages 403–424. Oxford University Press.

Michael Barlow. 2013. Individual differences and usage-based grammar. *International Journal of Corpus Linguistics*, 18(4):443–478.

Michael Barlow. 2018. The individual and the group from a corpus perspective. *The Corpus Linguistic Discourse: In Honour of Wolfgang Teubert. Amsterdam*, pages 163–184.

David K Barnhart. 2007. A calculus for new words. *Dictionaries: Journal of the Dictionary Society of North America*, 28(1):132–138.

Marco Baroni. 2020. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307.

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. 2022. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137.

Danielle Barth, James Grama, Simon Gonzalez, and Catherine Travis. 2020. Using forced alignment for sociophonetic research on a minority language. *University of Pennsylvania Working Papers in Linguistics*, 25(2):2.

Roland Barthes. 1968. *Elements of semiology*, volume 4. Macmillan.

Angelo Basile, Albert Gatt, and Malvina Nissim. 2019. You write like you eat: Stylistic variation as a predictor of social stratification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2583–2593, Florence, Italy. Association for Computational Linguistics.

Gareth J. Baxter, Richard A. Blythe, William Croft, and Alan J. McKane. 2009. Modeling language change: An evaluation of trudgill's theory of the emergence of new zealand english. *Language Variation and Change*, 21(2):257–296.

Ltd Beijing DataTang Technology Co. 2019. aidatatang$_200zh$.

Billal Belainine, Fatiha Sadat, Mounir Boukadoum, and Hakim Lounis. 2020. Towards a multi-dataset for complex emotions learning based on deep neural networks. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 50–58, Marseille, France. European Language Resources Association.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. 2007. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786.

Abhinav Bhandari and Caitrin Armstrong. 2019. Tkol, httt, and r/radiohead: High affinity terms in Reddit communities. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 57–67, Hong Kong, China. Association for Computational Linguistics.

Saurabhchand Bhati, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velázquez, and Najim Dehak. 2021. Segmental Contrastive Predictive Coding for Unsupervised Word Segmentation. In *Proc. Interspeech 2021*, pages 366–370.

Douglas Biber and Susan Conrad. 2019. *Register, Genre, and Style*. Cambridge University Press.

Brigitte Bigi. 2012. SPPAS: a tool for the phonetic segmentation of speech. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1748–1755, Istanbul, Turkey. European Language Resources Association (ELRA).

Maximilian Bisani and Hermann Ney. 2008. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451.

Christopher M Bishop. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Bernard Bloch. 1948. A set of postulates for phonemic analysis. *Language*, 24(1):3–46.

Leonard Bloomfield. 1933. *Language*. The University of Chicago Press.

Richard A Blythe and William A Croft. 2009. The speech community in evolutionary language dynamics. *Language Learning*, 59:47–63.

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M Nickel. 2019a. Explainable authorship verification in social media via attention-based similarity learning. In *2019 IEEE International Conference on Big Data*, pages 36–45. IEEE.

Benedikt Boenninghoff, Robert M Nickel, and Dorothea Kolossa. 2021. O2d2: Out-of-distribution detector to capture undecidable trials in authorship verification. *arXiv preprint arXiv:2106.15825*.

Benedikt Boenninghoff, Robert M Nickel, Steffen Zeiler, and Dorothea Kolossa. 2019b. Similarity learning for authorship verification in social media. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2457–2461. IEEE.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Claire Bowern. 2010. Correlates of language change in hunter-gatherer and other 'small'languages. *Language and Linguistics Compass*, 4(8):665–679.

Michael R Brent. 1999. Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, 3(8):294–301.

David Britain. 2012. 24 innovation diffusion in sociohistorical linguistics. *The handbook of historical sociolinguistics*, 68:451.

Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang. 2013. Authorship verification for short messages using stylometry. In *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6. IEEE.

Lindell Bromham, Xia Hua, Thomas G Fitzpatrick, and Simon J Greenhill. 2015. Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, 112(7):2097–2102.

Joan Bybee. 2010. *Language, usage and cognition*. Cambridge University Press.

Emmanuel Cartier. 2017. Neoveille, a web platform for neologism tracking. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 95–98.

Damon Centola. 2010. The spread of behavior in an online social network experiment. *science*, 329(5996):1194–1197.

JK Chambers. 2013. Patterns 14 of variation including change. *The handbook of language variation and change*, 129:297.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. 2021. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv:2104.02133*.

Chandrahas, Aditya Sharma, and Partha Talukdar. 2018. Towards understanding the geometry of knowledge graph embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 122–131, Melbourne, Australia. Association for Computational Linguistics.

Jonathan Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of blocked community members: Redemption, recidivism and departure. In *The World Wide Web Conference*, pages 184–195.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020a. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020b. Convokit: A toolkit for the analysis of conversations. In *Proceedings of SIGDIAL*.

Lei Chen, Klaus Zechner, and Xiaoming Xi. 2009. Improved pronunciation features for construct-driven assessment of non-native spontaneous speech. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for*

Computational Linguistics, pages 442–449, Boulder, Colorado. Association for Computational Linguistics.

Hsiao-Yu Chiang, Jose Camacho-Collados, and Zachary Pardos. 2020. Understanding the source of semantic regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 119–131, Online. Association for Computational Linguistics.

Morten H Christiansen and Simon Kirby. 2003. Language evolution: Consensus and controversies. *Trends in cognitive sciences*, 7(7):300–307.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 podcasts: A spoken English document corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Juan Camilo Conde-Silvestre. 2012. The role of social networks and mobility in diachronic sociolinguistics. *The Handbook of Historical Sociolinguistics*, pages 332–352.

Rolando Coto-Solano, James N. Stanford, and Sravana K. Reddy. 2021. Advances in completely automated vowel analysis for sociophonetics: Using end-to-end speech recognition systems with darla. *Frontiers in Artificial Intelligence*, 4.

Malcolm Coulthard. 2004. Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics*, 25(4):431–447.

Malcolm Coulthard, Alison Johnson, and David Wright. 2016. *An introduction to forensic linguistics: Language in evidence*. Routledge.

Tiago Cunha, David Jurgens, Chenhao Tan, and Daniel Romero. 2019. Are all successful communities alike? characterizing and predicting the success of online communities. In *The World Wide Web Conference*, pages 318–328.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318. ACM.

Marco Del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1591–1603, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 399–408, Berlin, Germany. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Norbert Dittmar. 1996. Explorations in'idiolects'. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 109–128.

Robin Dodsworth. 2019. Bipartite network structures and individual differences in sound change. *Glossa: a journal of general linguistics*, 4(1).

Robin Dodsworth and Richard A Benton. 2017. Social network cohesion and the retreat from southern vowels in raleigh. *Language in Society*, 46(3):371–405.

Zhiyong Zhang Dong Wang, Xuewei Zhang. 2015. Thchs-30 : A free chinese speech corpus.

Sylvie Dubois and Barbara M Horvath. 1998. Let's tink about dat: Interdental fricatives in cajun english. *Language Variation and Change*, 10(3):245–261.

Maeve Duggan and Aaron Smith. 2013. 6% of online adults are reddit users. *Pew Internet & American Life Project*, 3:1–10.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Penelope Eckert. 2019. The individual in the semiotic landscape. *Glossa: a journal of general linguistics*, 4(1).

Penelope Eckert and Sally McConnell-Ginet. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual review of anthropology*, 21(1):461–488.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11).

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Zsuzsanna Fagyal, Samarth Swarup, Anna María Escobar, Les Gasser, and Kiran Lakkaraju. 2010. Centers and peripheries: Network roles in language change. *Lingua*, 120(8):2061–2079.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Susan Fitzmaurice. 2007. Questions of standardization and representativeness in the development of social networks-based corpora: The story of the network of eighteenth-century english texts. In *Creating and digitizing language corpora*, pages 49–81. Springer.

Lucie Flek. 2020. Returning the N to NLP: Towards contextually personalized classification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Lucie Flekova, Daniel Preoţiuc-Pietro, and Lyle Ungar. 2016. Exploring stylistic variation with age and income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 313–319, Berlin, Germany. Association for Computational Linguistics.

Joerg Franke, Markus Mueller, Fatima Hamlaoui, Sebastian Stueker, and Alex Waibel. 2016. Phoneme boundary detection using deep bidirectional lstms. In *Speech Communication; 12. ITG Symposium*, pages 1–5. VDE.

Robert Fromont and Kevin Watson. 2016. Factors influencing automatic segmental alignment of sociophonetic corpora. *Corpora*, 11(3):401–431.

Romain Futrzynski. 2021. Author classification as pre-training for pairwise authorship verification.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon Technical Report n*, 93:27403.

Christophe Gérard. 2017. The logoscope: Semi-automatic tool for detecting and documenting the context of french new words. In *The dynamics of lexical diffusion. Data, methods, models.*

Ben Gittelson, Adrian Leemann, and Fabian Tomaschek. 2021. Using crowd-sourced speech data to study socially constrained variation in nonmodal phonation. *Frontiers in Artificial Intelligence*, 3:112.

Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *International Conference on Social Informatics*, pages 41–57. Springer.

Jean-Philippe Goldman. 2011. Easyalign: an automatic phonetic alignment tool under praat. In *Proc. Interspeech 2011*, pages 3233–3236.

Simon Gonzalez, James Grama, and Catherine E Travis. 2020. Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, 6(1).

Simon Gonzalez, Catherine Travis, James Grama, Danielle Barth, and Sunkulp Ananthanarayan. 2018. Recursive forced alignment: A test on a minority language. In *Proceedings of the 17th Australasian International Conference on Speech Science and Technology*, volume 145, page 148.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 40–50, Online. Association for Computational Linguistics.

Kyle Gorman, Jonathan Howell, and Michael Wagner. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193.

Mark Granovetter. 1983. The strength of weak ties: A network theory revisited. *Sociological theory*, pages 201–233.

Mark S Granovetter. 1977. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier.

Tim Grant. 2012. Txt 4n6: method, consistency, and distinctiveness in the analysis of sms text messages. *Journal of Law and Policy*, 21:467.

Tim Grant and Nicci Macleod. 2016. Assuming identities online: experimental linguistics applied to the policing of online paedophile activity. *Applied linguistics*, 37(1):50–70.

Tim Grant and Nicci MacLeod. 2020. *Language and Online Identities: The Undercover Policing of Internet Sexual Crime*. Cambridge University Press.

Tim Grant and Nicola MacLeod. 2018. Resources and constraints in linguistic identity performance–a theory of authorship. *Language and Law/Linguagem e Direito*, 5(1):80–96.

A. Graves and N. Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *ICML*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Simon J Greenhill, Xia Hua, Caela F Welsh, Hilde Schneemann, and Lindell Bromham. 2018. Population size and the rate of language evolution: a test across indo-european, austronesian, and bantu languages. *Frontiers in psychology*, 9:576.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2017. Analyzing lexical emergence in modern american english online 1. *English Language & Linguistics*, 21(1):99–127.

Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. 2013. Information diffusion in online social networks: A survey. *ACM Sigmod Record*, 42(2):17–28.

William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Eleventh International AAAI Conference on Web and Social Media*.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition.

Susan C. Herring. 2004. Computer-Mediated Discourse Analysis: An Approach to Researching Online Behavior. Learning in doing, pages 338–376. Cambridge University Press, New York, NY, US.

Valentin Hofmann, Hinrich Schütze, and Janet Pierrehumbert. 2020. A graph auto-encoder model of derivational morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1127–1138, Online. Association for Computational Linguistics.

David I Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, 13(3):111–117.

Janet Holmes and Miriam Meyerhoff. 1999. The community of practice: Theories and methodologies in language and gender research. *Language in society*, 28(2):173–183.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on World Wide Web*, pages 452–461.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*.

Nan Hu, Paul A Pavlou, and Jie Jennifer Zhang. 2009. Why do online product reviews have a j-shaped distribution? overcoming biases in online word-of-mouth communication. *Communications of the ACM*, 52(10):144–147.

Richard Anthony Hudson. 1996. *Sociolinguistics*. Cambridge university press.

Xingyu Na Bengu Wu Hao Zheng Hui Bu, Jiayu Du. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Oriental COCOSDA 2017*, page Submitted.

Manuela Hürlimann, Benno Weck, Esther van den Berg, Simon Suster, and Malvina Nissim. 2015. Glad: Groningen lightweight authorship detection. In *CLEF (Working Notes)*.

Roman Jakobson. 1971. Studies on child language and aphasia. In *Studies on Child Language and Aphasia*. De Gruyter Mouton.

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Alison Johnson and David Wright. 2014. Identifying idiolect in forensic authorship attribution: an n-gram textbite approach. *Language and Law*, 1(1).

Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data.

Daniel Jurafsky and James H Martin. 2018. Speech and language processing (draft). *In preparation [cited 2022 April 20] Available from: https://web. stanford. edu/˜ jurafsky/slp3*.

Herman Kamper and Benjamin van Niekerk. 2021. Towards Unsupervised Phone and Word Segmentation Using Self-Supervised Vector-Quantized Neural Networks. In *Proc. Interspeech 2021*, pages 1539–1543.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.

Matthew C. Kelley and Benjamin V. Tucker. 2018. A Comparison of Input Types to a Deep Neural Network-based Forced Aligner. In *Proc. Interspeech 2018*, pages 1205–1209.

Daphné Kerremans, Susanne Stegmayr, and Hans-Jörg Schmid. 2012. The neocrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. *Current methods in historical semantics*, 73:59.

Daniel Kershaw, Matthew Rowe, and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 553–562.

Paul Kerswill. 2003. Dialect levelling and geographical diffusion in british english. *Social dialectology: in honour of Peter Trudgill*, pages 223–243.

Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer, and Dan Chazan. 2005. Phoneme alignment based on discriminative learning. In *Proc. Interspeech 2005*, pages 2961–2964.

Mike Kestemont. 2014. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden. Association for Computational Linguistics.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at pan 2020. In *CLEF*.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2021. Overview of the cross-domain authorship verification task at pan 2021. In *CLEF (Working Notes)*.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. 2018. Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs*, volume 2125 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. 2020. Proxy anchor loss for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3238–3247.

Sarah King and Mark Hasegawa-Johnson. 2013. Accurate speech segmentation by mimicking human auditory processing. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8096–8100.

Thomas Kisler, Florian Schiel, and Han Sloetjes. 2012. Signal processing via web services: the use case webmaus. In *Digital Humanities Conference 2012*.

Dennis H Klatt. 1979. Speech perception: A model of acoustic–phonetic analysis and lexical access. *Journal of phonetics*, 7(3):279–312.

Colin Klein, Peter Clutton, and Adam G Dunn. 2019. Pathways to conspiracy: The social and linguistic precursors of involvement in reddit's conspiracy theory forum. *PloS one*, 14(11):e0225098.

David G Kleinbaum and Mitchel Klein. 2010. *Survival analysis*. Springer.

Thea Knowles, Meghan Clayards, and Morgan Sonderegger. 2018. Examining factors influencing the viability of automatic acoustic analysis of child speech. *Journal of Speech, Language, and Hearing Research*, 61(10):2487–2501.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Moshe Koppel and Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, pages 72–80.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62.

Felix Kreuk, Joseph Keshet, and Yossi Adi. 2020a. Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation. In *Proc. Interspeech 2020*, pages 3700–3704.

Felix Kreuk, Yaniv Sheena, Joseph Keshet, and Yossi Adi. 2020b. Phoneme boundary detection using learnable segmental features. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8089–8093. IEEE.

Gakuto Kurata and Kartik Audhkhasi. 2018. Improved knowledge distillation from bi-directional to uni-directional lstm ctc for end-to-end speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 411–417. IEEE.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *International Conference on Speech and Computer*, pages 267–278. Springer.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. 2019. Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20(129):1–30.

Håvard Kvamme and Ørnulf Borgan. 2019. Continuous and discrete-time survival prediction with neural networks. *arXiv preprint arXiv:1910.06724*.

William Labov. 1972. *Sociolinguistic patterns*. 4. University of Pennsylvania Press.

William Labov. 1989. Exact description of the speech community: Short a in philadelphia. *Language Change and Variation*, pages 1–57.

William Labov. 2007. Transmission and diffusion. *Language*, 83(2):344–387.

William Labov et al. 1981. Field methods of the project on linguistic change and variation.

Mikko Laitinen, Masoud Fatemi, and Jonas Lundberg. 2020. Size matters: Digital social networks and language change. *Frontiers in Artificial Intelligence*, 3:46.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, et al. 2021. Generative spoken language modeling from raw audio. *arXiv preprint arXiv:2102.01192*.

Elizabeth Lanza and Bente Ailin Svendsen. 2007. Tell me who your friends are and i might be able to tell you what language (s) you speak: Social network analysis, multilingualism, and identity. *International journal of bilingualism*, 11(3):275–300.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Kristina Lerman and Rumi Ghosh. 2010. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM), 2010*.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of massive data sets*. Cambridge university press.

Shiri Lev-Ari. 2018. Social network size can influence linguistic malleability and the propagation of linguistic change. *Cognition*, 176:31–39.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Q Lhoest et al. 2021. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*.

Xinjian Li, David R Mortensen, Florian Metze, and Alan W Black. 2021. Multilingual phonetic dataset for low resource speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6958–6962. IEEE.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Tal Linzen. 2016. Issues in evaluating semantic spaces using word analogies. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 13–18, Berlin, Germany. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Alexander H Liu, Wei-Ning Hsu, Michael Auli, and Alexei Baevski. 2022a. Towards end-to-end unsupervised speech recognition. *arXiv preprint arXiv:2204.02492*.

Alexander H. Liu, Wei-Ning Hsu, Michael Auli, and Alexei Baevski. 2022b. Towards end-to-end unsupervised speech recognition.

Da-Rong Liu, Kuan-Yu Chen, Hung yi Lee, and Lin shan Lee. 2018. Completely unsupervised phoneme recognition by adversarially learning mapping relationships from audio embeddings. In *Proc. Interspeech 2018*, pages 3748–3752.

Da-rong Liu, Po-chun Hsu, Yi-chen Chen, Sung-feng Huang, Shun-po Chuang, Da-yi Wu, and Hung-yi Lee. 2021. Learning phone recognition from unpaired audio and phone sequences based on generative adversarial network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:230–243.

Da-rong Liu, Po-chun Hsu, Yi-chen Chen, Sung-feng Huang, Shun-po Chuang, Da-yi Wu, and Hung-yi Lee. 2022c. Learning phone recognition from unpaired audio and phone sequences based on generative adversarial network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:230–243.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech Model Pre-Training for End-to-End Spoken Language Understanding. In *Proc. Interspeech 2019*, pages 814–818.

Gary Lupyan and Rick Dale. 2010. Language structure is partly determined by social structure. *PloS one*, 5(1).

Laurel MacKenzie and Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of british english. *Linguistics Vanguard*, 6(s1).

Nicci MacLeod and Tim Grant. 2012. Whose tweet? authorship analysis of micro-blogs and other short-form messages. In *Proceedings of The International Association of Forensic Linguists' Tenth Biennial Conference*.

Nicci MacLeod and Tim Grant. 2021. Assuming identities online: How linguistics is helping the policing of online grooming and the distribution of abusive images. In *Rethinking Cybercrime*, pages 87–104. Springer.

Ltd. Magic Data Technology Co. 2019. Magicdata mandarin chinese read speech corpus.

T. Mahr. 2021. Performance of forced-alignment algorithms on children's speech. *Journal of Speech, Language, and Hearing Research*.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*.

Jonathan Marshall. 2004. *Language change and sociolinguistics: Rethinking social networks*. Springer.

Joshua L Martin. 2021. Spoken corpora data, automatic speech recognition, and bias against african american language: The case of habitual'be'. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 284–284.

Vikram C Mathad, Tristan J Mahr, Nancy Scherer, Kathy Chapman, Katherine C Hustad, Julie Liss, and Visar Berisha. 2021. The impact of forced-alignment errors on automatic pronunciation evaluation. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 176–180. International Speech Communication Association.

Binny Mathew, Ritam Dutt, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Deep dive into anonymity: Large scale analysis of Quora questions. In *International Conference on Social Informatics*, pages 35–49. Springer.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Tony McEnery and Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "i don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on african americans. *Frontiers in Artificial Intelligence*, 4.

Allan A Metcalf. 2004. *Predicting new words: The secrets of their success*. Houghton Mifflin Harcourt.

Miriam Meyerhoff and Anna Strycharz. 2013. Communities of practice. *The handbook of language variation and change*, pages 428–447.

Miriam Meyerhoff and James A Walker. 2007. The persistence of variation in individual grammars: Copula absence in 'urban sojourners' and their stay-at-home peers, bequia (st vincent and the grenadines) 1. *Journal of Sociolinguistics*, 11(3):346–366.

Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. 2019. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*.

Paul Michel, Okko Rasanen, Roland Thiollière, and Emmanuel Dupoux. 2017. Blind phoneme segmentation with temporal prediction errors. In *Proceedings of ACL 2017, Student Research Workshop*, pages 62–68, Vancouver, Canada. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

James Milroy and Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics*, 21(2):339–384.

Lesley Milroy. 2002. Introduction: Mobility, contact, and language change–working with contemporary speech communities. *Journal of Sociolinguistics*, 6(1):3–15.

Lesley Milroy and Carmen Llamas. 2013. Social networks. *The Handbook of Language Variation and Change*, pages 407–427.

Lesley Milroy and James Milroy. 1992. Social network and social class: Toward an integrated sociolinguistic model. *Language in society*, 21(1):1–26.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Naomi Nagy. 2018. Linguistic attitudes and contact effects in toronto's heritage languages: A variationist sociolinguistic investigation. *International Journal of Bilingualism*, 22(4):429–446.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)*, 50(6):1–36.

Daniel Nettle. 1999a. Is the rate of linguistic change constant? *Lingua*, 108(2-3):119–136.

Daniel Nettle. 1999b. Using social impact theory to simulate language change. *Lingua*, 108(2-3):95–117.

Terttu Nevalainen. 2000. Mobility, social networks and language change in Early Modern England. *European Journal of English Studies*, 4(3):253–264.

Terttu Nevalainen, Helena Raumolin-Brunberg, and Heikki Mannila. 2011. The diffusion of language change in real time: Progressive and conservative individuals and the time depth of change. *Language Variation and Change*, 23(1):1–43.

Mitchell G Newberry, Christopher A Ahern, Robin Clark, and Joshua B Plotkin. 2017. Detecting evolutionary forces in language change. *Nature*, 551(7679):223–226.

Mark Newman. 2018. *Networks*. Oxford university press.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Bill Noble and Raquel Fernández. 2015. Centre stage: How social network position shapes linguistic coordination. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 29–38, Denver, Colorado. Association for Computational Linguistics.

Josef R. Novak, Nobuaki Minematsu, and Keikichi Hirose. 2012. WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding. In *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, pages 45–49, Donostia–San Sebastián. Association for Computational Linguistics.

R. Ochshorn and M. Hawkins. 2017. Gentle. https://lowerquality.com/gentle/.

John J Ohala. 1981. The listener as a source of sound change. In Carrie S Masek, Roberta A Hendrick, and Mary Frances Miller, editors, *Papers from the Parasession on Language and Behavior*, pages 178–203. Chicago Linguistic Society, Chicago.

Juanita Ordoñez, Rafael Rivera Soto, and Barry Y Chen. 2020. Will longformers pan out for authorship verification. *Working Notes of CLEF*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Mark Pagel, Quentin D Atkinson, and Andrew Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, 449(7163):717–720.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

John C Paolillo. 1999. The virtual speech community: Social network and language variation on irc. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, pages 10–pp. IEEE.

Michał B Paradowski and Łukasz Jonak. 2012. Diffusion of linguistic innovation as social coordination. *Psychology of Language and Communication*, 16(2):131–142.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Daniel S. Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V. Le. 2020. Improved Noisy Student Training for Automatic Speech Recognition. In *Proc. Interspeech 2020*, pages 2817–2821.

Kyubyong Park and Jongseok Kim. 2019. g2pE. https://github.com/Kyubyong/g2p.

Kyubyong Park and Seanie Lee. 2020. g2pM: A Neural Grapheme-to-Phoneme Conversion Package for Mandarin Chinese Based on a New Open Benchmark Dataset. In *Proc. Interspeech 2020*, pages 1723–1727.

Zeyang Peng, Leilei Kong, Zhijie Zhang, Zhongyuan Han, and Xu Sun. 2021. Encoding text information by pre-trained model for authorship verification. In *CLEF*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ben Peters, Jon Dehdari, and Josef van Genabith. 2017. Massively multilingual neural grapheme-to-phoneme conversion. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 19–26, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

P. Plantinga and E. Fosler-Lussier. 2019. Towards real-time mispronunciation detection in kids' speech. In *ASRU*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Ltd. Primewords Information Technology Co. 2018. Primewords chinese corpus set 1. `https://www.primewords.cn`.

Yu Qiao, Naoya Shimomura, and Nobuaki Minematsu. 2008. Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3989–3992. IEEE.

Juan Ramos et al. Using tf-idf to determine word relevance in document queries. Citeseer.

Okko Rasanen. 2014. Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36.

Okko Johannes Räsänen, Unto Kalervo Laine, and Toomas Altosaar. 2009. An improved speech segmentation quality measure: the r-value. In *Tenth Annual Conference of the International Speech Communication Association*. Citeseer.

Limor Raviv, Antje Meyer, and Shiri Lev-Ari. 2019. Larger communities create more systematic languages. *Proceedings of the Royal Society B*, 286(1907):20191262.

Limor Raviv, Antje Meyer, and Shiri Lev-Ari. 2020. The role of social network structure in the emergence of linguistic structure. *Cognitive Science*, 44(8):e12876.

Florencia Reali, Nick Chater, and Morten H Christiansen. 2018. Simpler grammar, larger vocabulary: How population size affects language. *Proceedings of the Royal Society B: Biological Sciences*, 285(1871):20172586.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*.

Y. Ren et al. 2019. Fastspeech: fast, robust and controllable text to speech. In *NeurIPS*.

Antoinette Renouf. 2007. Tracing lexical productivity and creativity. *Lexical creativity, texts and contexts*, 58:61.

Antoinette Renouf, Andrew Kehoe, and Jayeeta Banerjee. 2007. Webcorp: an integrated system for web text search. In *Corpus linguistics and the web*, pages 47–67. Brill.

Rachid Riad, Anne-Catherine Bachoud-Lévi, Frank Rudzicz, and Emmanuel Dupoux. 2020. Identification of primary and collateral tracks in stuttered speech. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1681–1688, Marseille, France. European Language Resources Association.

Ingrid Rosenfelder, Josef Fruehwald, Keelan Evanini, and Jiahong Yuan. 2011. FAVE (forced alignment and vowel extraction) program suite. *URL http://fave. ling. upenn. edu*.

Anni Sairio. 2009. Language and letters of the bluestocking network. sociolinguistic issues in eighteenth-century epistolary english. *Neuphilologische Mitteilungen*, 110(4):526–528.

Alaa Ehab Sakran, Sherif Mahdy Abdou, Salah Eldeen Hamid, and Mohsen Rashwan. 2017. A review: Automatic speech segmentation. *International Journal of Computer Science and Mobile Computing*, 6(4):308–315.

Natalie Schilling-Estes. 1998. Investigating “self-conscious” speech: The performance register in ocracoke english. *Language in society*, 27(1):53–83.

Kilian Schulze-Forster, Clement SJ Doire, Gaël Richard, and Roland Badeau. 2020. Joint phoneme alignment and text-informed speech separation on highly corrupted speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7274–7278. IEEE.

Thomas M Schwen and Noriko Hara. 2003. Community of practice: A metaphor for online design? *The Information Society*, 19(3):257–270.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Devyani Sharma. 2017. Scalar effects of social networks on language variation. *Language Variation and Change*, 29(3):393–418.

Devyani Sharma and Robin Dodsworth. 2020. Language variation and social networks. *Annual Review of Linguistics*, 6.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2020. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*.

Kevin J Shih, Rafael Valle, Rohan Badlani, Adrian Lancucki, Wei Ping, and Bryan Catanzaro. 2021. Rad-tts: Parallel flow-based tts with robust alignment learning and diverse synthesis. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*.

Philipp Singer, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. 2014. Evolution of reddit: from the front page of the internet to a self-referential community? In *Proceedings of the 23rd international conference on world wide web*, pages 517–522.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Ian Stewart and Jacob Eisenstein. 2018. Making “fetch” happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4370, Brussels, Belgium. Association for Computational Linguistics.

Patrycja Strycharczuk, Manuel López-Ibáñez, Georgina Brown, and Adrian Leemann. 2020. General northern english. exploring regional variation in the north of england with machine learning. *Frontiers in Artificial Intelligence*, page 48.

Antti Suni, Marcin Włodarczak, Martti Vainio, and Juraj Šimko. 2019. Comparative Analysis of Prosodic Characteristics Using WaveNet Embeddings. In *Proc. Interspeech 2019*, pages 2538–2542.

Surfingtech. St-cmds-20170001$_1, freestchinesemandarincorpus$.

Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara. 2018. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4784–4788. IEEE.

Chenhao Tan and Lillian Lee. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1056–1066.

Kevin Tang and Ryan Bennett. 2019. Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (mayan). In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, pages 1719–1723.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline.

Naim Terbeh, Ayman Trigui, Mohsen Maraoui, and Mounir Zrigui. 2016. Arabic speech analysis to identify factors posing pronunciation disorders and to assist learners with vocal disabilities. In *2016 International Conference on Engineering & MIS (ICEMIS)*, pages 1–8. IEEE.

Yann Teytaut and Axel Roebel. 2021. Phoneme-to-Audio Alignment with Recurrent Neural Networks for Speaking and Singing Voice. In *Proc. Interspeech 2021*, pages 61–65.

Ingrid Tieken-Boon van Ostade. 2000. Social network analysis and the history of english. *European Journal of English Studies*, 4(3):211–216.

Shubham Toshniwal and Karen Livescu. 2016. Jointly learning to align and convert graphemes to phonemes with neural attention models. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 76–82. IEEE.

M Teresa Turell. 2010. The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *International Journal of Speech, Language & the Law*, 17(2).

Hans Van Halteren, Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77.

Kaili Vesik, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2020. One model to pronounce them all: Multilingual grapheme-to-phoneme conversion with a transformer ensemble. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 146–152, Online. Association for Computational Linguistics.

Soroush Vosoughi, Helen Zhou, and Deb Roy. 2015. Digital stylometry: Linking profiles across social networks. In *International Conference on Social Informatics*, pages 164–177. Springer.

Yu-Hsuan Wang, Cheng-Tao Chung, and Hung-Yi Lee. 2017. Gate activation signal analysis for gated recurrent neural networks and its correlation with phoneme boundaries. In *Proc. Interspeech 2017*, pages 3822–3826.

Ronald Wardhaugh. 2011. *An introduction to sociolinguistics*, volume 28. John Wiley & Sons.

Janith Weerasinghe and Rachel Greenstadt. 2020. Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification—Notebook for PAN at CLEF 2020. In *CLEF 2020 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Janith Weerasinghe, Rhia Singh, and Rachel Greenstadt. 2021. Feature vector difference based authorship verification for open-world settings. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 2201–2207. CEUR-WS.org.

Li Wei. 1994. *Three generations, two languages, one family: Language choice and language shift in a Chinese community in Britain*, volume 104. Multilingual Matters.

Douglas Brent West et al. 2001. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4598, Brussels, Belgium. Association for Computational Linguistics.

David Wright. 2013. Stylistic variation within genre conventions in the enron email corpus: developing a textsensitive methodology for authorship research. *International Journal of Speech, Language & the Law*, 20(1).

David Wright. 2017. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem. *International Journal of Corpus Linguistics*, 22(2):212–241.

David Wright. 2018. Idiolect. *Oxford Bibliographies in Linguistics*.

Quirin Würschinger. 2021. Social networks of lexical innovation. investigating the social dynamics of diffusion of neologisms on twitter. *Frontiers in Artificial Intelligence*, page 106.

Alex Xiao, Weiyi Zheng, Gil Keren, Duc Le, Frank Zhang, Christian Fuegen, Ozlem Kalinli, Yatharth Saraf, and Abdelrahman Mohamed. 2021. Scaling asr improves zero and few shot learning. *arXiv preprint arXiv:2111.05948*.

Chih-Kuan Yeh, Jianshu Chen, Chengzhu Yu, and Dong Yu. 2019. Unsupervised speech recognition via segmental empirical output distribution matching. In *International Conference on Learning Representations*.

Yu Ting Yeung, Ka Ho Wong, and Helen Meng. 2015. Improving automatic forced alignment for dysarthric speech transcription. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. 2002. The htk book. *Cambridge university engineering department*, 3(175):12.

Jiahong Yuan and Mark Liberman. 2009. Investigating/l/variation in english through forced alignment. In *Tenth Annual Conference of the International Speech Communication Association*. Citeseer.

Albert Zeyer, Eugen Beck, Ralf Schlüter, and Hermann Ney. 2017. CTC in the Context of Generalized Full-Sum HMM Training. In *Proc. Interspeech 2017*, pages 944–948.

Justine Zhang, Cristian Danescu-Niculescu-Mizil, Christina Sauper, and Sean J Taylor. 2018. Characterizing online public discussions through patterns of participant interactions. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–27.

Justine Zhang, William Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Quan-shi Zhang and Song-Chun Zhu. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39.

Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. 2021. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 229–240, Online. Association for Computational Linguistics.

Jian Zhu and David Jurgens. 2021a. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jian Zhu and David Jurgens. 2021b. The structure of online social networks modulates the rate of lexical change. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2201–2218, Online. Association for Computational Linguistics.

Jian Zhu, Cong Zhang, and David Jurgens. 2022a. Byt5 model for massively multilingual grapheme-to-phoneme conversion.

Jian Zhu, Cong Zhang, and David Jurgens. 2022b. Phone-to-audio alignment without text: A semi-supervised approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8167–8171. IEEE.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Lal Zimman and Will Hayworth. 2020. Lexical change as sociopolitical change in trans and cis identity labels: New methods for the corpus analysis of internet data. *University of Pennsylvania Working Papers in Linguistics*, 25(2):17.