# Computational Sociolinguistics: A Survey

Dong Nguyen
University of Twente

A. Seza Doğruöz
Tilburg University/
Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS)

Carolyn P. Rosé
Carnegie Mellon University

Franciska de Jong
University of Twente/
Erasmus University Rotterdam

*Language is a social phenomenon and variation is inherent to its social nature. Recently, there has been a surge of interest within the computational linguistics (CL) community in the social dimension of language. In this article we present a survey of the emerging field of 'Computational Sociolinguistics' that reflects this increased interest. We aim to provide a comprehensive overview of CL research on sociolinguistic themes, featuring topics such as the relation between language and social identity, language use in social interaction and multilingual communication. Moreover, we demonstrate the potential for synergy between the research communities involved, by showing how the large-scale data-driven methods that are widely used in CL can complement existing sociolinguistic studies, and how sociolinguistics can inform and challenge the methods and assumptions employed in CL studies. We hope to convey the possible benefits of a closer collaboration between the two communities and conclude with a discussion of open challenges.*

## 1. Introduction

Science has experienced a paradigm shift along with the increasing availability of large amounts of digital research data (Hey, Tansley, and Tolle 2009). In addition to the traditional focus on the description of natural phenomena, theory development and computational science, data-driven exploration and discovery have become a dominant ingredient of many methodological frameworks. In line with these developments, the field of computational linguistics (CL) has also evolved.

Human communication occurs in both verbal and nonverbal form. Research on computational linguistics has primarily focused on capturing the informational dimension of language and the structure of verbal information transfer. In the words of Krishnan and Eisenstein (2015), computational linguistics has made great progress in modeling language's informational dimension, but with a few notable exceptions, computation has had little to contribute to our understanding of language's social dimension. The recent increase in interest of computational linguists to study language in social contexts is partly driven by the ever increasing availability of social media data. Data from social media platforms provide a strong incentive for innovation in the CL research agenda and the surge in relevant data opens up methodological possibilities for studying text as social data. Textual resources, like many other language resources, can be seen as a data type that is signaling all kinds of social phenomena. This is related to the fact that language is one of the instruments by which people construct

their online identity and manage their social network. There are challenges as well. For example, social media language is more colloquial and contains more linguistic variation, such as the use of slang and dialects, than the language in datasets that have been commonly used in CL research (e.g., scientific articles, newswire text and the Wall Street Journal) (Eisenstein 2013b). However, an even greater challenge is that the relation between social variables and language is typically fluid and tenuous, while the CL field commonly focuses on the level of literal meaning and language structure, which is more stable.

The tenuous connection between social variables and language arises because of the symbolic nature of the relation between them. With the language chosen a social identity is signaled, which may buy a speaker[1] something in terms of footing within a conversation, or in other words: for speakers there is room for choice in how to use their linguistic repertoire in order to achieve social goals. This freedom of choice is often referred to as the agency of speakers and the linguistic symbols chosen can be thought of as a form of social currency. Speakers may thus make use of specific words or stylistic elements to represent themselves in a certain way. However, because of this agency, social variables cease to have an essential connection with language use. It may be the case, for example, that on average female speakers display certain characteristics in their language more frequently than their male counterparts. Nevertheless, in specific circumstances, females may choose to de-emphasize their identity as females by modulating their language usage to sound more male. Thus, while this exception serves to highlight rather than challenge the commonly accepted symbolic association between gender and language, it nevertheless means that it is less feasible to predict how a female will sound in a randomly selected context.

Speaker agency also enables creative violations of conventional language patterns. Just as with any violation of expectations, these creative violations communicate indirect meanings. As these violations become conventionalized, they may be one vehicle towards language change. Thus, agency plays a role in explaining the variation in and dynamic nature of language practices, both within individual speakers and across speakers. This variation is manifested at various levels of expression – the choice of lexical elements, phonological variants, semantic alternatives and grammatical patterns – and plays a central role in the phenomenon of linguistic change. The audience, demographic variables (e.g., gender, age), and speaker goals are among the factors that influence how variation is exhibited in specific contexts. Agency thus increases the intricate complexity of language that must be captured in order to achieve a social interpretation of language.

Sociolinguistics investigates the reciprocal influence of society and language on each other. Sociolinguists traditionally work with spoken data using qualitative and quantitative approaches. Surveys and ethnographic research have been the main methods of data collection (Eckert 1989; Milroy and Milroy 1985; Milroy and Gordon 2003; Tagliamonte 2006; Trudgill 1974; Weinreich, Labov, and Herzog 1968). The datasets used are often selected and/or constructed to facilitate controlled statistical analyses and insightful observations. However, the resulting datasets are often small in size compared to the standards adopted by the CL community. The massive volumes of data that have become available from sources such as social media platforms have provided the opportunity to investigate language variation more broadly. The opportunity for

1 We use the term 'speaker' for an individual who has produced a message, either as spoken word or in textual format. When discussing particular social media sites, we may refer to 'users' as well.

the field of sociolinguistics is to identify questions that this massive but messy data would enable them to answer. Sociolinguists must then also select an appropriate methodology. However, typical methods used within sociolinguistics would require sampling the data down. If they take up the challenge to instead analyze the data in its massive form, they may find themselves open to partnerships in which they may consider approaches more typical in the field of CL.

As more and more researchers in the field of CL seek to interpret language from a social perspective, an increased awareness of insights from the field of sociolinguistics could inspire modeling refinements and potentially lead to performance gains. Recently, various studies (Hovy 2015; Stoop and van den Bosch 2014; Volkova, Wilson, and Yarowsky 2013) have demonstrated that existing NLP tools can be improved by accounting for linguistic variation due to social factors, and Hovy and Søgaard (2015) have drawn attention to the fact that biases in frequently used corpora, such as the Wall Street Journal, cause NLP tools to perform better on texts written by older people. The rich repertoire of theory and practice developed by sociolinguists could impact the field of CL also in more fundamental ways. The boundaries of communities are often not as clear-cut as they may seem and the impact of agency has not been sufficiently taken into account in many computational studies. For example, an understanding of linguistic agency can explain why and when there might be more or less of a problem when making inferences about people based on their linguistic choices. This issue is discussed in depth in some recent computational work related to gender, specifically Bamman, Eisenstein, and Schnoebelen (2014) and Nguyen et al. (2014) who provide a critical reflection on the operationalization of gender in CL studies.

The increasing interest in analyzing and modeling the social dimension of language within CL encourages collaboration between sociolinguistics and CL in various ways. However, the potential for synergy between the two fields has not been explored systematically so far (Eisenstein 2013b) and to date there is no overview of the common and complementary aspects of the two fields. This article aims to present an integrated overview of research published in the two communities and to describe the state-of-the-art in the emerging multidisciplinary field that could be labeled as '*Computational Sociolinguistics*'. The envisaged audiences are CL researchers interested in sociolinguistics and sociolinguists interested in computational approaches to study language use. We hope to demonstrate that there is enough substance to warrant the recognition of '*Computational Sociolinguistics*' as an autonomous yet multidisciplinary research area. Furthermore, we hope to convey that this is the moment to develop a research agenda for the scholarly community that maintains links with both sociolinguistics and computational linguistics.

In the remaining part of this section, we discuss the rationale and scope of our survey in more detail as well as the potential impact of integrating the social dimensions of language use in the development of practical NLP applications. In Section 2 we discuss *Methods for Computational Sociolinguistics*, in which we reflect on methods used in sociolinguistics and computational linguistics. In Section 3, *Language and Social Identity Construction*, we discuss how speakers use language to shape perception of their identity and focus on computational approaches to model language variation based on gender, age and geographical location. In Section 4 on *Language and Social Interaction*, we move from individual speakers to pairs, groups and communities and discuss the role of language in shaping personal relationships, the use of style-shifting, and the adoption of norms and language change in communities. In Section 5 we discuss *Multilingualism and Social Interaction*, in which we present an overview of tools for processing multilingual communication, such as parsers and language identification systems. We

will also discuss approaches for analyzing patterns in multilingual communication from a computational perspective. In Section 6 we conclude with a summary of major challenges within this emerging field.

### 1.1 Rationale for a Survey of Computational Sociolinguistics

The increased interest in studying a social phenomenon such as language use from a data-driven or computational perspective exemplifies a more general trend in scholarly agendas. The study of social phenomena through computational methods is commonly referred to as 'Computational Social Science' (Lazer et al. 2009). The increasing interest of social scientists in computational methods can be regarded as illustrating the general increase of attention for cross-disciplinary research perspectives. 'Multidisciplinary', 'interdisciplinary', 'cross-disciplinary' and 'transdisciplinary' are among the labels used to mark the shift from monodisciplinary research formats to models of collaboration that embrace diversity in the selection of data and methodological frameworks. However, in spite of various attempts to harmonize terminology, the adoption of such labels is often poorly supported by definitions and they tend to be used interchangeably. The objectives of research rooted in multiple disciplines often include the ambition to resolve real world or complex problems, to provide different perspectives on a problem, or to create cross-cutting research questions, to name a few (Choi and Pak 2006).

The emergence of research agendas for (aspects of) computational sociolinguistics fits in this trend. We will use the term *Computational Sociolinguistics* for the emerging research field that integrates aspects of sociolinguistics and computer science in studying the relation between language and society from a computational perspective. This survey article aims to show the potential of leveraging massive amounts of data to study social dynamics in language use by combining advances in computational linguistics and machine learning with foundational concepts and insights from sociolinguistics. Our goals for establishing Computational Sociolinguistics as an independent research area include the development of tools to support sociolinguists, the establishment of new statistical methods for the modeling and analysis of data that contains linguistic content as well as information on the social context, and the development or refinement of NLP tools based on sociolinguistic insights.

### 1.2 Scope of Discussion

Given the breadth of this field, we will limit the scope of this survey as follows. First of all, the coverage of sociolinguistics topics will be selective and primarily determined by the work within computational linguistics that touches on sociolinguistic topics. For readers with a wish for a more complete overview of sociolinguistics, we recommend the introductory readings by Bell (2013), Holmes (2013) and Meyerhoff (2011).

The availability of social media and other online language data in computer-mediated formats is one of the primary driving factors for the emergence of computational sociolinguistics. A relevant research area is therefore the study of Computer-Mediated Communication (CMC) (Herring 1996). Considering the strong focus on speech data within sociolinguistics, there is much potential for computational approaches to be applied to spoken language as well. Moreover, the increased availability of recordings of spontaneous speech and transcribed speech has inspired a revival in the study of the social dimensions of spoken language (Jain et al. 2012), as well as in the analysis of the relation between the verbal and the nonverbal layers in spoken dialogues (Truong et al. 2014). As online data increasingly becomes multimodal, for example

with the popularity of vlogs (video blogs), we expect the use of spoken word data for computational sociolinguistics to increase. Furthermore, we expect that multimodal analysis, a topic that has been the focus of attention in the field of human-computer interaction for many years, will also receive attention in computational sociolinguistics.

In the study of communication in pairs and groups, the individual contributions are often analyzed in context. Therefore, much of the work on language use in settings with multiple speakers draws from foundations in discourse analysis (De Fina, Schiffrin, and Bamberg 2006; Hyland 2004; Martin and White 2005; Schegloff 2007), pragmatics (such as speech act theory (Austin 1975; Searle 1969)), rhetorical structure theory (Mann and Thompson 1988; Taboada and Mann 2006) and social psychology (Giles and Coupland 1991; Postmes, Spears, and Lea 2000; Richards 2006). For studies within the scope of computational sociolinguistics that build upon these fields the link with the foundational frameworks will be indicated. Another relevant field is computational stylometry (Daelemans 2013; Holmes 1998; Stamatatos 2009), which focuses on computational models of writing style for various tasks such as plagiarism detection, author profiling and authorship attribution. Here we limit our discussion to publications on topics such as the link between style and social variables.

### 1.3 NLP Applications

Besides yielding new insights into language use in social contexts, research in computational sociolinguistics could potentially also impact the development of applications for the processing of textual social media and other content. For example, user profiling tools might benefit from research on automatically detecting the gender (Burger et al. 2011), age (Nguyen et al. 2013), geographical location (Eisenstein et al. 2010) or affiliations of users (Piergallini et al. 2014) based on an analysis of their linguistic choices. The cases for which the interpretation of the language used could benefit most from using variables such as age and gender are usually also the ones for which it is most difficult to automatically detect those variables. Nevertheless, in spite of this kind of challenge, there are some published proofs of concept that suggest potential value in advancing past the typical assumption of homogeneity of language use embodied in current NLP tools. For example, incorporating how language use varies across social groups has improved word prediction systems (Stoop and van den Bosch 2014), algorithms for cyberbullying detection (Dadvar et al. 2012) and sentiment-analysis tools (Hovy 2015; Volkova, Wilson, and Yarowsky 2013). Hovy and Søgaard (2015) show that POS taggers trained on well-known corpora such as the English Penn Treebank perform better on texts written by older authors. They draw attention to the fact that texts in various frequently used corpora are from a biased sample of authors in terms of demographic factors. Furthermore, many NLP tools currently assume that the input consists of monolingual text, but this assumption does not hold in all domains. For example, social media users may employ multiple language varieties, even within a single message. To be able to automatically process these texts, NLP tools that are able to deal with multilingual texts are needed (Solorio and Liu 2008b).

## 2. Methods for Computational Sociolinguistics

As discussed, one important goal of this article is to stimulate collaboration between the fields of sociolinguistics in particular and social science research related to communication at large on the one hand, and computational linguistics on the other hand. By addressing the relationship with methods from both sociolinguistics and the social sci-

ences in general we are able to underline two expectations. First of all, we are convinced that sociolinguistics and related fields can help the field of computational linguistics to build richer models that are more effective for the tasks they are or could be used for. Second, the time seems right for the CL community to contribute to sociolinguistics and the social sciences, not only by developing and adjusting tools for sociolinguists, but also by refining the theoretical models within sociolinguistics using computational approaches and contributing to the understanding of the social dynamics in natural language. In this section, we highlight challenges that reflect the current state of the field of computational linguistics. In part these challenges relate to the fact that in the field of language technologies at large, the methodologies of social science research are usually not valued, and therefore also not taught. There is a lack of familiarity with methods that could easily be adopted if understood and accepted. However, there are promising examples of bridge building that are already occurring in related fields such as learning analytics. More specifically, in the emerging area of discourse analytics there are demonstrations of how these practices could eventually be observed within the language technologies community as well (Rosé in press; Rosé and Tovares 2015; Rosé et al. 2008).

At the outset of multidisciplinary collaboration, it is necessary to understand differences in goals and values between communities, as these differences strongly influence what counts as a contribution within each field, which in turn influences what it would mean for the fields to contribute to one another. Towards that end, we first discuss the related but distinct notions of reliability and validity, as well as the differing roles these notions have played in each field (Section 2.1). This will help lay a foundation for exploring differences in values and perspectives between fields. Here, it will be most convenient to begin with quantitative approaches in the social sciences as a frame of reference. In Section 2.2 we discuss contrasting notions of theory and empiricism as well as the relationship between the two, as that will play an important and complementary role in addressing the concern over differing values. In Section 2.3 we broaden the scope to the spectrum of research approaches within the social sciences, including strong quantitative and strong qualitative approaches, and the relationship between CL and the social disciplines involved. This will help to further specify the concrete challenges that must be overcome in order for a meaningful exchange between communities to take place. In Section 2.4 we illustrate how these issues come together in the role of data, as the collection, sampling, and preparation of data are of central importance to the work in both fields.

### 2.1 Validation of Modeling Approaches

The core of much research in the field of computational linguistics, in the past decade especially, is the development of new methods for computational modeling, such as probabilistic graphical models and deep learning within a neural network approach. These novel methods are valued both for the *creativity* that guided the specification of novel model structures and the corresponding requirement for new methods of inference as well as the achievement of *predictive accuracy* on tasks for which there is some notion of a correct answer.

Development of new modeling frameworks is part of the research production cycle both within sociolinguistics (and the social sciences in general) and the CL community, and there is a lot of overlap with respect to the types of methods used. For example, logistic regression is widely employed by variationist sociolinguists using a program called VARBRUL (Tagliamonte 2006). Similarly, logistic regression is widely used in the

CL community, especially in combination with regularization methods when dealing with thousands of variables, for example for age prediction (Nguyen et al. 2013). As another example, latent variable modeling approaches (Koller and Friedman 2009) have grown in prominence within the CL community for dimensionality reduction, managing heterogeneity in terms of multiple domains or multiple tasks (Zhang, Ghahramani, and Yang 2008), and approximation of semantics (Blei, Ng, and Jordan 2003; Griffiths and Steyvers 2004). Similarly, it has grown in prominence within the quantitative branches of the social sciences for modeling causality (Glymour et al. 1987), managing heterogeneity in terms of group effects and subpopulations (Collins and Lanza 2010), and time series modeling (Rabe-Hesketh and Skrondal 2012; Rabe-Hesketh, Skrondal, and Pickles 2004).

The differences in reasons for the application of similar techniques are indicative of differences in values. While in CL there is a value placed on creativity and predictive accuracy, within the social sciences, the related notions of *validity* and *reliability* underline the values placed on conceptual contributions to the field. Validity is primarily a measure of the extent to which a research design isolates a particular issue from confounds so that questions can receive clear answers. This typically requires creativity, and frequently research designs for isolating issues effectively are acknowledged for this creativity in much the same way a novel graphical model would be acknowledged for the elegance of its mathematical formulation. Reliability, on the other hand, is primarily a measure of the reproducibility of a result and might seem to be a distinct notion from predictive accuracy. However, the connection is apparent when one considers that a common notion of reliability is the extent to which two human coders would arrive at the same judgment on a set of data points, whereas predictive accuracy is the extent to which a model would arrive at the same judgment on a set of data points as a set of judgements decided ahead of time by one or more humans.

While at some deep level there is much in common between the goals and values of the two communities, the differences in values signified by the emphasis on creativity and predictive accuracy on the one side and reliability and validity on the other side nevertheless poses challenges for mutual exchange. Validity is a multi-faceted notion, and it is important to properly distinguish it from the related notion of reliability. If one considers shooting arrows at a target, one can consider reliability to be a measure of how much convergence is achieved in location of impact of multiple arrows. On the other hand, validity is the extent to which the point of convergence centers on the target. Reproducibility of results is highly valued in both fields, which requires reliability wherever human judgment is involved, such as in the production of a gold standard (Carletta 1996; Di Eugenio and Glass 2004). However, before techniques from CL will be adopted by social science researchers, standards of validation from the social sciences will likely need to be addressed (Krippendorff 2013). We will see that this notion requires more than the related notion of creativity as appreciated within the field of CL.

One aspect that is germane to the notion of validity that goes beyond pure creativity is the extent to which the essence that some construct actually captures corresponds to the intended quantity. This aspect of validity is referred to as *face validity*. For example, the face validity of a sentiment analysis tool could be tested as follows. First, an automatic measure of sentiment would be applied to a text corpus. Then, texts would be sorted by the resulting sentiment scores and the data points from the end points and middle compared with one another. Are there consistent and clear distinctions in sentiment between beginning, middle, and end? Is sentiment the main thing that is captured in the contrast, or is something different really going on? While the CL

community has frequently upheld high standards of reliability, it is rare to find work that deeply questions whether the models are measuring the right thing. Nevertheless, this deep questioning is core to high quality work in the social sciences, and without it, the work may appear weak.

Another important notion is *construct validity*, or the extent to which the experimental design manages extraneous variance effectively. If the design fails to do so, it affects the interpretability of the result. This notion applies when we interpret the learned weights of features in our models to make statements about language use. When not controlling for confounding variables, the feature weights are misleading and valid interpretation is not possible. For example, many studies on gender prediction (see Section 3) ignore extraneous variables such as age, while gender and age are known to interact with each other highly. Where confounds may not have been properly eliminated in an investigation, again the results may appear weak regardless of the numbers associated with the measure of predictive accuracy.

Another important methodological idea is triangulation. Simply put, it is the idea that if you look at the same object through different lenses, each of which is designed to accentuate and suppress different kinds of details, you get more information than if you looked through just one, analogous to the value obtained through the use of ensemble methods like *bagging*. Triangulation is thus an important way of strengthening research findings in the social sciences by leveraging multiple views simultaneously rather than just using one in addressing a question. Sentiment analysis can again be used for illustration purposes. Consider a blog corpus for which the age of each individual blogger is available. Let's assume that a model for predicting age allocated high weights to some sentiment-related words. This may be considered as evidence that the model is consistent with previous findings that older people use more words that express a positive sentiment. Another method could measure sentiment for each blog individually. If the measured sentiment would correlate with the age of bloggers across the corpus, the two methods for investigating the connection between age and sentiment would tell the same story and the confidence in the validity of the story would increase. This type of confirming evidence is referred to as an indication of convergent validity.

Another form of triangulation is where distinctions known to exist are confirmed. For this example, assume that a particular model for predicting political affiliation placed high weights on some sentiment-related words in a corpus related to issues for which those affiliated with one political perspective would take a different stance than those affiliated with another perspective, and this affiliation is known for all data points. The experimenters may conclude that this evidence is consistent with previous findings suggesting that voters express more positive sentiment towards political stances they are in favor of. If this is true, then if the model is applied to a corpus where both parties agree on a stance, the measure of sentiment should become irrelevant. Assuming the difference in the role of sentiment between the corpora is consistent with what is expected, the interpretation is strengthened. This is referred to as divergent validity since an expected difference in relationship is confirmed. Seeking convergent and divergent validity is a mark of high quality work in the social sciences, but it is rare in evaluations in the field of CL, and without it, again, the results may appear weak from a social science perspective. In order for methods from CL to be acceptable for use within the social sciences, these perceived weaknesses must be addressed.

### 2.2 Theory versus Empiricism

Above we discussed the importance placed on validity within the social sciences that stems from the goal of isolating an issue in order to answer questions. In order to clarify why that is important, it is necessary to discuss the value placed on theory versus empiricism.

Within the CL community, a paradigm shift took place after the middle of the 1990s. Initially, approaches that combined symbolic and statistical methods were of interest (Klavans and Resnik 1996). But with the focus on very large corpora and new frameworks for large-scale statistical modeling, symbolic- and knowledge-driven methods have been largely left aside, though the presence of linguistics as an active force can still be seen in some areas of computational linguistics, such as tree banking. Along with older symbolic methods that required carefully crafted grammars and lexicons, the concept of knowledge source has become strongly associated with the notion of theory, which is consistent with the philosophical notion of linguistic theory advocated by Chomskyan linguistics and other formal linguistic theories (Backofen and Smolka 1993; Green 1992; Schneider, Dowdall, and Rinaldi 2004; Wintner 2002). As knowledge-based methods have to a large extent been replaced with statistical models, a grounding in linguistic theory has become less and less valued. A desire to replace theory with empiricism dominated the Zeitgeist and drove progress within the field. Currently, the term *theory* seems to be associated with old and outdated approaches. It often has a negative connotation in contrast to the positive reception of empiricism, and contemporary modeling approaches are believed to have a greater ability to offer insights into language than symbolic modeling frameworks.

In contrast, in the social sciences the value of a contribution is measured in terms of the extent to which it contributes towards theory. Theories may begin with human originated ideas. But these notions are only treated as valuable if they are confirmed through empirical methods. As these methods are applied, theoretical models gain empirical support. Findings are ratified and then accumulated. Therefore, theories become storehouses for knowledge obtained through empirical methods. Atheoretical empiricism is not attractive within the social sciences where the primary value is on building theory and engaging theory in the interpretation of models.

As CL seeks to contribute to sociolinguistics and the social sciences, this divide of values must be addressed in order to avoid the fields talking at cross purposes. To stimulate collaboration between fields, it is important to not only focus on task performance, but also to integrate existing theories into the computational models and use these models to refine or develop new theories.

### 2.3 Quantitative versus Qualitative Approaches

The social sciences have both strong qualitative and quantitative branches. Similarly, sociolinguistics has branches in qualitative research (e.g., interactional sociolinguistics) and quantitative research (variationist sociolinguistics). From a methodological perspective, most computational sociolinguistics work has a strong resemblance with quantitative and therefore variationist sociolinguistics, which has a strong focus on statistical analysis to uncover the distribution of sociolinguistic variables (Tagliamonte 2006). So far we have mostly reflected on methods used in CL and their commonality with the methods used in the quantitative branches in sociolinguistics and the social sciences, but the time is right for a greater focus on how qualitative methods may also be of use. Some thoughts about what that might look like can be found in the work of

Rosé and Tovares (2015), who explore the productive tension between the two branches as it relates to interaction analysis. The field of computational linguistics could benefit from exploring this tension to a greater degree in its own work, for example by taking a deeper look at data through human eyes as part of the validation of constructed models.

The tension between qualitative and quantitative branches can be illustrated with the extent to which the agency of speakers is taken into account. As explained in the introduction, linguistic agency refers to the freedom of speakers to make choices about how they present themselves in interaction. A contrasting notion is the extent to which social structures influence the linguistic choices speakers make. Regardless of research tradition, it is acknowledged that speakers both have agency and are simultaneously influenced by social structures. The question is which is emphasized in the research approach. Quantitative researchers believe that the most important variance is captured by representation of the social structure. They recognize that this is a simplification, but the value placed on quantification for the purpose of identifying causal connections between variables makes the sacrifice of accuracy worth it. In the field of CL, this valuing is analogous to the well-known saying that all models are wrong, but some are nevertheless useful. On the other side are researchers committed to the idea that the most important and interesting aspects of language use are the ones that violate norms in order for the speaker to achieve a goal. These researchers may doubt that the bulk of choices made by speakers can be accounted for by social structures. We see the balance and tension between the ideas of language reflecting established social structures and language arising from speaker agency within current trends in variationist sociolinguistics. Much of that work focused on the ways in which language variation can be accounted for by reference to social structures (Bell 2013). On the other hand, more recently the agency of speakers is playing a more central role as well in variationist sociolinguistics (Eckert 2012).

While in CL qualitative research is sometimes dismissed as being quantitative work that lacks rigor, one could argue that high quality qualitative research has a separate notion of rigor and depth that is all its own (Morrow and Brown 1994). An important role for qualitative research is to challenge the operationalizations constructed by quantitative researchers. To achieve the adoption of CL methods and models by social science researchers, the challenges from the qualitative branches of the social sciences will become something to consider carefully.

As computational linguistics shares more values with variationist sociolinguistics, many studies within computational sociolinguistics also focus on the influence of social structures. For example, work on predicting social variables such as gender (Section 3) is built on the idea that gender determines the language use of speakers. However, such research ignores the agency of speakers: Speakers use language to construct their identity and thus not everyone might write in a way that reflects their biological sex. Moving forward, it would make sense for researchers in computational sociolinguistics to reflect on the dominant role of social structures over agency. Some work in CL has already begun to acknowledge the agency of speakers when interpreting findings (Bamman, Eisenstein, and Schnoebelen 2014; Nguyen et al. 2014).

One way of conceptualizing the contrast between the usage of computational models in the two fields is to reconsider the trade-off between maximizing interpretability — typical of the social sciences and sociolinguistics —, and maximizing predictive accuracy, typical of CL. Both fields place a premium on rigor in evaluation and generalization of results across datasets. To maintain a certain standard of rigor, the CL community has produced practices for standardization of metrics, sampling, and avoidance of overfitting or overestimation of performance through careful separation of

training and testing data at all stages of model development. Within the social sciences, the striving for rigor has also produced statistical machinery for analysis, but most of all it has resulted in an elaborate process for validation of such modeling approaches and practices for careful application and interpretation of the results.

One consequence of the focus on interpretability within the social sciences is that models tend to be kept small and simple in terms of the number of parameters, frequently no more than 10, or at least no more than 100. Because the models are kept simple, they can be estimated on smaller datasets, as long as sampling is done carefully and extraneous variance is controlled. In the CL community, it is more typical for models to include tens of thousands of parameters or more. For such large models, massive corpora are needed to prevent overfitting. As a result, research in the CL community is frequently driven by the availability of large corpora, which explains the large number of recent papers on data from the web, such as Twitter and Wikipedia. Because of this difference in scale, a major focus on parallelization and approximate inference has been an important focus of work in CL (Heskes, Albers, and Kappen 2002), whereas interest in such methods has only recently grown within the social sciences.

### 2.4 Spotlight on Corpora and Other Data

Data collection is a fundamental step in the research cycle for researchers in both sociolinguistics and computational linguistics. Here we will reflect on the differences in the practices and traditions within both fields and on the emerging use of online data. In the subsequent sections of this survey, there will be dedicated subsections about the data sources used in the specific studies relevant to the discussed themes (e.g., on identity construction).

Traditionally, sociolinguists have been interested in datasets that capture informal speech (also referred to as the '*vernacular*'), i.e., the kind of language used when speakers are not paying attention (Tagliamonte 2006). A variety of methods have been used to collect data, including observation, surveys and interviews (Mallinson, Childs, and Herk 2013; Tagliamonte 2006). The sociolinguistic datasets are carefully prepared to enable in-depth analyses of how a speech community operates, carefully observing standards of reliability and validity as discussed previously. Inevitably, these data collection methods are labor-intensive and time-consuming. The resulting datasets are often small in comparison to the ones used within computational linguistics. The small sizes of these datasets made the work in sociolinguistics of limited interest to the field of CL.

The tide began to turn with the rise of computer mediated communication (CMC). Herring (2007) defines CMC as '*predominantly text-based human-human interaction mediated by networked computers or mobile telephony*'. The content generated in CMC, and in particular when generated on social media platforms, is a rich and easy to access source of large amounts of informal language coming together with information about the context (e.g., the users, social network structure, the time or geolocation at which it was generated) that can be used for the study of language in social contexts on a large scale. Examples include microblogs (Eisenstein et al. 2014; Kooti et al. 2012), web forums (Garley and Hockenmaier 2012; Nguyen and Rosé 2011) and online review sites (Danescu-Niculescu-Mizil et al. 2013b; Hovy, Johannsen, and Søgaard 2015). For example, based on data from Twitter (a popular microblogging site) dialectal variation has been mapped using a fraction of the time, costs and effort that was needed in traditional studies (Doyle 2014). However, data from CMC is not always easy to collect. As an example, while text messaging (SMS) is widely used, collecting SMS data has

been difficult due to both technical and privacy concerns. The SMS4science project (Dürscheid and Stark 2011) aims to overcome these difficulties by asking people to donate their messages, collaborating with the service providers for the collection of the messages, and applying anonymization to ensure privacy.

A complicating issue in data collection in sociolinguistics is that participants might adjust their language use towards the expectations of the data collector. This phenomenon is known as the 'observer's paradox', a term first coined by Labov (1972): "*the aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain these data by systematic observation*". In social media, the observer's paradox could potentially be argued to have lost much of its strength, making it a promising resource to complement traditional data collection methods. While a convenient source of data, the use of social media data does introduce new challenges that must be addressed regardless of field, and this offers a convenient beginning to a potential exchange between fields.

First, social media users are usually not representative of the general population (Mislove et al. 2011; Nguyen et al. 2013). A better understanding of the demographics could aid the interpretation of findings, but often little is known about the users. Collecting demographic information requires significant effort, or might not even be possible in some cases due to ethical concerns. Furthermore, in many cases the complete data is not fully accessible through an API, requiring researchers to apply a sampling strategy (e.g., randomly, by topic, time, individuals/groups, phenomenon (Androutsopoulos 2013; Herring 2004)). Sampling may introduce additional biases or remove important contextual information. These problems are even more of a concern when datasets are reused for secondary analysis by other researchers whose purposes might be very different from those who performed the sampling.

Social media data also introduces new units of analysis (such as messages and threads) that do not correspond entirely with traditional analysis units (such as sentences and turns) (Androutsopoulos 2013). This raises the question about valid application of findings from prior work. Another complicating factor is that in social media the target audience of a message is often not explicitly indicated, i.e., multiple audiences (e.g., friends, colleagues) are collapsed into a single context (Marwick and boyd 2011). Some studies have therefore treated the use of hashtags and user mentions as proxies for the target audience (Nguyen, Trieschnigg, and Cornips 2015; Pavalanathan and Eisenstein 2015a). Furthermore, while historically the field of sociolinguistics started with a major focus on phonological variation, e.g., Labov (1966), the use of social media data has led to a higher focus on lexical variation in computational sociolinguistics. However, there are concerns that a focus on lexical variation without regard to other aspects may threaten the validity of conclusions. Phonology does impact social media orthography at both the word level and structural level (Eisenstein 2013a), suggesting that studies on phonological variation could inform studies based on social media text data and vice versa. For example, Eisenstein (2013a) found that consonant cluster reduction (e.g., *just* vs. *jus*) in Twitter is influenced by the phonological context, in particular, reduction was less likely when the word was followed by a segment that began with a vowel.

There are practical concerns as well. First, while both access and content have often been conceptualized as either public or private, in reality this distinction is not as absolute, for example, a user might discuss a private topic on a public social media site. In view of the related privacy issues, Bolander and Locher (2014) argue for more awareness regarding the ethical implications of research using social media data.

Automatically processing social media data is more difficult compared to various other types of data that have been used within computational linguistics. Many de-

veloped tools (e.g., parsers, named entity recognizers) do not work well due to the informal nature of many social media texts. While the dominant response has been to focus on text normalization and domain adaptation, Eisenstein (2013b) argues that doing so is throwing away meaningful variation. For example, building on work on text normalization, Gouws et al. (2011) showed how various transformations (e.g., dropping the last character of a word) vary across different user groups on Twitter. As another example, Brody and Diakopoulos (2011) find that lengthening of words (e.g., *cooooll*) is often applied to subjective words. They build on this observation to detect sentiment-bearing words. The tension between normalizing and preserving the variation in text also arises in the processing and analysis of historical texts (see Piotrowski (2012) for an overview), which also contain many spelling variations. In this domain, normalization is often applied as well to facilitate the use of tools such as parsers. However, some approaches first normalize the text, but then replace the modernized word forms with the original word forms to retain the original text. Another issue with social media data is that many social media studies have so far focused primarily on one data source. A comparison of the online data sources in terms of language use has only been done in a few studies (Baldwin et al. 2013; Hu, Talamadupula, and Kambhampati 2013).

Another up and coming promising resource for studying language from a social perspective is crowdsourcing. So far, crowdsourcing is mostly used to obtain large numbers of annotations, e.g., Snow et al. (2008). However, 'crowds' can also be used for large-scale perception studies, i.e., to study how non-linguists interpret messages and identify social characteristics of speakers (Clopper 2013), and for the collection of linguistic data, such as the use of variants of linguistic variables. Within sociolinguistics, surveys have been one of the instruments to collect data and crowdsourcing is an emerging alternative to traditional methods for collecting survey data.

Crowdsourcing has already been used to obtain perception data for sociolinguistic research, for example, to study how English utterances are perceived differently across language communities (Makatchev and Simmons 2011) and to obtain native-likeness ratings of speech samples (Wieling et al. 2014). For some studies, games have been developed to collect data. Nguyen et al. (2014) studied how Twitter users are perceived based on their tweets by asking players to guess the gender and age based on displayed tweets. Leemann et al. (2016) developed a mobile app that predicted the user's location based on a 16-question survey. By also collecting user feedback on the predictions, the authors compared their data with the Linguistic Atlas of German-speaking Switzerland, which was collected about 70 years before the crowdsourcing study. The mismatches between the Atlas data and self-reported data from the mobile app were seen to suggest linguistic change in progress.

Crowdsourcing also introduces challenges. For example, the data collection method is less controlled and additional effort for quality control is often needed. Even more problematic is that usually little is known is about the workers, such as the communities they are part of. For example, Wieling et al. (2014) recruited participants using e-mail, social media and blogs, which resulted in a sample that was likely to be biased towards linguistically interested people. However, they did not expect that the possible bias in the data influenced the findings much. Another concern is that participants in crowdsourcing studies might modulate their answers towards what they think is expected, especially when there is a monetary compensation. In the social sciences in general, crowdsourcing is also increasingly used for survey research. Behrend et al. (2011) compared the data collected using crowdsourcing with data collected from a traditional psychology participant pool (undergraduates) in the context of organizational psychology research and concluded that crowdsourcing is a potentially viable resource

to collect data for this research area. While thus promising, the number of studies so far using crowdsourcing for sociolinguistic research is small and more research needs to be done to study the strengths and weaknesses of this data collection method for sociolinguistic research.

## 3. Language and Social Identity

We now turn to discussing computational approaches for modeling language variation related to social identity. Speakers use language to construct their social identity (Bucholtz and Hall 2005). Being involved in communicative exchange can be functional for the transfer of information, but at the same it functions as a staged performance in which users select specific codes (e.g., language, dialect, style) that shape their communication (Wardhaugh 2011). Consciously or unconsciously speakers adjust their performance to the specific social context and to the impression they intend to make on their audience. Each speaker has a personal linguistic repertoire to draw linguistic elements or codes from. Selecting from the repertoire is partially subject to 'identity work', a term referring to the range of activities that individuals engage in to create, present, and sustain personal identities that are congruent with and supportive of the self-concept (Snow and Anderson 1987).

Language is one of the instruments that speakers use in shaping their identities, but there are limitations (e.g., physical or genetic constraints) to the variation that can be achieved. For example, somebody with a smoker's voice may not be able to speak with a smooth voice but many individual characteristics still leave room for variation. Although traditionally attributed an absolute status, personal features (e.g., age and gender) are increasingly considered social rather than biological variables. Within sociolinguistics, a major thrust of research is to uncover the relation between social variables (e.g., gender, age, ethnicity, status) and language use (Eckert 1997; Eckert and McConnell-Ginet 2013; Holmes and Meyerhoff 2003; Wagner 2012). The concept of sociolects, or social dialects, is similar to the concept of regional dialects. While regional dialects are language varieties based on geography, sociolects are based on social groups, e.g., different groups according to social class (with labels such as 'working class' and 'middle class'), or according to gender or age. A study by Guy (2013) suggests that the cohesion between variables (e.g., nominal agreement, denasalization) to form sociolects is weaker than usually assumed. The unique use of language by an individual is an idiolect, and this concept is in particular relevant for authorship attribution (e.g., Grieve (2007)).

Recognizing that language use can reveal social patterns, many studies in computational linguistics have focused on automatically inferring social variables from text. This task can be seen as a form of automatic metadata detection that can provide information on author features. The growing interest in trend analysis tools is one of the drivers for the interest in the development and refinement of algorithms for this type of metadata detection. However, tasks such as gender and age prediction do not only appeal to researchers and developers of trend mining tools. Various public demos have been able to attract the attention of the general public (e.g., TweetGenie[2] (Nguyen, Trieschnigg, and Meder 2014) and Gender Guesser[3]), which can be attributed to a widespread interest in the entertaining dimension of the linguistic dimension of identity work. The

2 http://www.tweetgenie.nl
3 http://www.hackerfactor.com/GenderGuesser.php

automatic prediction of individual features such as age and gender based on only text is a nontrivial task. Studies that have compared the performance of humans with that of automatic systems for gender and age prediction based on text alone found that automatic systems perform better than humans (Burger et al. 2011; Nguyen et al. 2013). A system based on aggregating guesses from a large number of people still predicted gender incorrectly for 16% of the Twitter users (Nguyen et al. 2014). While most studies use a supervised learning approach, a recent study by Ardehaly and Culotta (2015) explored a lightly supervised approach using soft constraints. They combined unlabeled geotagged Twitter data with soft constraints, like the proportion of people below or above 25 years in a county according to Census data, to train their classifiers.

Within computational linguistics, linguistic variation according to gender, age and geographical location have received the most attention, compared to other variables such as ethnicity (Ardehaly and Culotta 2015; Pennacchiotti and Popescu 2011; Rao et al. 2011) and social class. Labels for variables like social class are more difficult to obtain and use because they are rarely made explicit in online user profiles that are publically available. Only recently this direction has been explored, with occupation as a proxy for variables like social class. Occupation labels for Twitter users have been extracted from their profile description (Preoţiuc-Pietro, Lampos, and Aletras 2015; Preoţiuc-Pietro et al. 2015; Sloan et al. 2015). Preoţiuc-Pietro et al. (2015) then mapped the derived occupations to income and Sloan et al. (2015) mapped the occupations to social class categories. However, these studies were limited to users with self-reported occupations in their profiles.

Many studies have focused on individual social variables, but these variables are not independent. For example, there are indications that linguistic features that are used more by males increase in frequency with age as well (Argamon et al. 2007). As another example, some studies have suggested that language variation across gender tends to be stronger among younger people and to fade away with older ages (Barbieri 2008). Eckert (1997) notes that the age considered appropriate for cultural events often differs for males and females (e.g., getting married), which influences the interaction between gender and age. The interaction between these variables is further complicated by the fact that in many uncontrolled settings the gender distribution may not be equal for different age ranges (as observed in blogs (Burger and Henderson 2006) and Twitter (Nguyen et al. 2013)). Therefore, failing to control for gender while studying age (and vice versa) can lead to misinterpretation of the findings.

In this section an overview will be presented of computational studies of language variation related to social identity. This section will first focus on the datasets that have been used to investigate social identity and language variation in computational linguistics (Subsection 3.1). After surveying computational studies on language variation according to gender (Subsection 3.2), age (Subsection 3.3) and location (Subsection 3.4), we conclude with a discussion of how various NLP tasks, such as sentiment detection, can be improved by accounting for language variation related to the social identity of speakers (Subsection 3.5).

### 3.1 Data Sources

Early computational studies on social identity and language use were based on formal texts, such as the British National Corpus (Argamon et al. 2003; Koppel, Argamon, and Shimoni 2002), or datasets collected from controlled settings, such as recorded conversations (Singh 2001) and telephone conversations (Boulis and Ostendorf 2005; Garera and Yarowsky 2009; Van Durme 2012) where protocols were used to coordinate

the conversations (such as the topic). With the advent of social media, a shift is observed towards more informal texts collected from uncontrolled settings. Much of the initial work in this domain focused on blogs. The Blog Authorship Corpus (Schler et al. 2006), collected in 2004 from blogger.com, has been used in various studies on gender and age (Argamon et al. 2007; Gianfortoni, Adamson, and Rosé 2011; Goswami, Sarkar, and Rustagi 2009; Nguyen, Smith, and Rosé 2011; Sap et al. 2014). Others have created their own blog corpus from various sources including LiveJournal and Xanga (Burger and Henderson 2006; Mukherjee and Liu 2010; Nowson and Oberlander 2006; Rosenthal and McKeown 2011; Sarawgi, Gajulapalli, and Choi 2011; Yan and Yan 2006).

More recent studies are focusing on Twitter data, which contains richer interactions in comparison to blogs. Burger et al. (2011) created a large corpus by following links to blogs that contained author information provided by the authors themselves. The dataset has been used in various subsequent studies (Bergsma and Van Durme 2013; Van Durme 2012; Volkova, Wilson, and Yarowsky 2013). Others created their own Twitter dataset (Eisenstein, Smith, and Xing 2011; Kokkos and Tzouramanis 2014; Liao et al. 2014; Rao et al. 2010; Zamal, Liu, and Ruths 2012). While early studies focused on English, recent studies have used Twitter data written in other languages as well, like Dutch (Nguyen et al. 2013), Spanish and Russian (Volkova, Wilson, and Yarowsky 2013), and Japanese, Indonesian, Turkish, and French (Ciot, Sonderegger, and Ruths 2013). Besides blogs and Twitter, other web sources have been explored, including LinkedIn (Kokkos and Tzouramanis 2014), IMDb (Otterbacher 2010), YouTube (Filippova 2012), e-mails (Corney et al. 2002), a Belgian social network site (Peersman, Daelemans, and Vaerenbergh 2011) and Facebook (Rao et al. 2011; Sap et al. 2014; Schwartz et al. 2013).

Two aspects can be distinguished that are often involved in the process of creating datasets to study the relation between social variables and language use.

*Labeling.* Datasets derived from uncontrolled settings such as social media often lack explicit information regarding the identity of users, such as their gender, age or location. Researchers have used different strategies to acquire adequate labels:

- *User-provided information*. Many researchers utilize information provided by the social media users themselves, for example based on explicit fields in user profiles (Burger et al. 2011; Schler et al. 2006; Yan and Yan 2006), or by searching for specific patterns such as birthday announcements (Zamal, Liu, and Ruths 2012). While this information is probably highly accurate, such information is often only available for a small set of users, e.g., for age, 0.75% of the users in Twitter (Liao et al. 2014) and 55% in blogs (Burger and Henderson 2006). Locations of users have been derived based on geotagged messages (Eisenstein et al. 2010) or locations in user profiles (Mubarak and Darwish 2014).
- *Manual annotation*. Another option is manual annotation based on personal information revealed in the text, profile information, and public information on other social media sites (Ciot, Sonderegger, and Ruths 2013; Nguyen et al. 2013). In the manual annotation scenario, a random set of authors is annotated. However, the required effort is much higher resulting in smaller datasets and biases of the annotators themselves might influence the annotation process. Furthermore, for some users not enough information may be available to even manually assign labels.

- *Exploiting names*. Some labels can be automatically extracted based on the name of a person. For example, gender information for names can be derived from census information from the US Social Security Administration (Bamman, Eisenstein, and Schnoebelen 2014; Prabhakaran, Reid, and Rambow 2014), or from Facebook data (Fink, Kopecky, and Morawski 2012). However, people who use names that are more common for a different gender will be incorrectly labeled in these cases. In some languages, such as Russian, the morphology of the names can also be used to predict the most likely gender labels (Volkova, Wilson, and Yarowsky 2013). However, people who do not provide their names, or have uncommon names, will remain unlabeled. In addition, acquiring labels this way has not been well studied yet for other languages and cultures and for other types of labels (such as geographical location or age).

*Sample selection.* In many cases, it is necessary to limit the study to a sample of persons. Sometimes the selected sample is directly related to the way labels are obtained, for example by only including people who explicitly list their gender or age in their social media profile (Burger et al. 2011), who have a gender-specific first name (Bamman, Eisenstein, and Schnoebelen 2014), or who have geotagged tweets (Eisenstein et al. 2010). Restricting the sample, e.g., by only including geotagged tweets, could potentially lead to biased datasets. Pavalanathan and Eisenstein (2015b) compared geotagged tweets with tweets written by users with self-reported locations in their profile. They found that geotagged tweets are more often written by women and younger people. Furthermore, geotagged tweets contain more geographically specific non-standard words. Another approach is random sampling, or as random as possible due to restrictions of targeting a specific language (Nguyen et al. 2013). However, in these cases the labels may not be readily available. This increases the annotation effort and in some cases it may not even be possible to obtain reliable labels. Focused sampling is used as well, for example by starting with social media accounts related to gender-specific behavior (e.g., male/female hygiene products, sororities) (Rao et al. 2010). However, such an approach has the danger of creating biased datasets, which could influence the prediction performance (Cohen and Ruths 2013).

### 3.2 Gender

The study of gender and language variation has received much attention in sociolinguistics (Eckert and McConnell-Ginet 2013; Holmes and Meyerhoff 2003). Various studies have highlighted gender differences. According to Tannen (1990), women engage more in 'rapport' talk, focusing on establishing connections, while men engage more in 'report' talk, focusing on exchanging information. Similarly, according to Holmes (1995), in women's communication the social function of language is more salient, while in men's communication the referential function (conveying information) tends to be dominant. Argamon et al. (2003) make a distinction between involvedness (more associated with women) and informational (more associated with men). However, with the increasing view that speakers use language to construct their identity, such generalizations have also been met with criticism. Many of these studies rely on small sample sizes and ignore other variables (such as ethnicity, social class) and the many similarities between genders. Such generalizations contribute to stereotypes and the view of gender as an inherent property.

**3.2.1 Modeling Gender.** Within computational linguistics, researchers have focused primarily on automatic gender classification based on text. Gender is then treated as a binary variable based on biological characteristics, resulting in a binary classification task. A variety of machine learning methods have been explored, including SVMs (Boulis and Ostendorf 2005; Ciot, Sonderegger, and Ruths 2013; Corney et al. 2002; Fink, Kopecky, and Morawski 2012; Gianfortoni, Adamson, and Rosé 2011; Mukherjee and Liu 2010; Nowson and Oberlander 2006; Peersman, Daelemans, and Vaerenbergh 2011; Rao et al. 2010; Zamal, Liu, and Ruths 2012), logistic regression (Bergsma and Van Durme 2013; Otterbacher 2010), Naive Bayes (Goswami, Sarkar, and Rustagi 2009; Mukherjee and Liu 2010; Yan and Yan 2006) and the Winnow algorithm (Burger et al. 2011; Schler et al. 2006). However, treating gender as a binary variable based on biological characteristics assumes that gender is fixed and is something people *have*, instead of something people *do* (Butler 1990), i.e., such a setup neglects the agency of speakers. Many sociolinguists, together with scholars from the social sciences in general, view gender as a social construct, emphasizing that gendered behavior is a result of social conventions rather than inherent biological characteristics.

**3.2.2 Features and Patterns.** Rather than focusing on the underlying machine learning models, most studies have focused on developing predictive features. Token-level and character-level unigrams and n-grams have been explored in various studies (Bamman, Eisenstein, and Schnoebelen 2014; Burger et al. 2011; Fink, Kopecky, and Morawski 2012; Sarawgi, Gajulapalli, and Choi 2011; Yan and Yan 2006). Sarawgi, Gajulapalli, and Choi (2011) found character-level language models to be more robust than token-level language models. Grouping words by meaningful classes could improve the interpretation and possibly the performance of models. Linguistic Inquiry and Word Count (LIWC, Pennebaker, Francis, and Booth (2001)) is a dictionary-based word counting program originally developed for the English language. It also has versions for other languages, such as Dutch (Zijlstra et al. 2005). LIWC has been used in experiments on Twitter data (Fink, Kopecky, and Morawski 2012) and blogs (Nowson and Oberlander 2006; Schler et al. 2006). However, models based on LIWC alone tend to perform worse than unigram/ngram models (Fink, Kopecky, and Morawski 2012; Nowson and Oberlander 2006). By analyzing the developed features, studies have shown that males tend to use more numbers (Bamman, Eisenstein, and Schnoebelen 2014), technology words (Bamman, Eisenstein, and Schnoebelen 2014) and URLs (Schler et al. 2006; Nguyen et al. 2013), while females use more terms referring to family and relationship issues (Boulis and Ostendorf 2005). A discussion of the influence of genre and domain on gender differences is provided later in this section.

Various features based on grammatical structure have been explored, including features capturing individual POS frequencies (Argamon et al. 2003; Otterbacher 2010) as well as POS patterns (Argamon et al. 2003, 2009; Bamman, Eisenstein, and Schnoebelen 2014; Schler et al. 2006). Males tend to use more prepositions (Argamon et al. 2007, 2009; Otterbacher 2010; Schler et al. 2006) and more articles (Argamon et al. 2007; Nowson and Oberlander 2006; Otterbacher 2010; Schler et al. 2006; Schwartz et al. 2013), however Bamman, Eisenstein, and Schnoebelen (2014) did not find these differences to be significant in their Twitter study. Females tend to use more pronouns (Argamon et al. 2003, 2007, 2009; Bamman, Eisenstein, and Schnoebelen 2014; Otterbacher 2010; Schler et al. 2006; Schwartz et al. 2013), in particular first person singular (Nguyen et al. 2013; Otterbacher 2010; Schwartz et al. 2013). A measure introduced by Heylighen and Dewaele (2002) to measure formality based on the frequencies of different word classes has been used in experiments on blogs (Mukherjee and Liu 2010; Nowson, Oberlander,

and Gill 2005). Sarawgi, Gajulapalli, and Choi (2011) experimented with probabilistic context-free grammars (PCFGs) by adopting the approach proposed by Raghavan, Kovashka, and Mooney (2010) for authorship attribution. They trained PCFG parsers for each gender and computed the likelihood of test documents for each gender-specific PCFG parser to make the prediction. Bergsma, Post, and Yarowsky (2012) experimented with three types of syntax features and found features based on single-level context-free-grammar (CFG) rules (e.g., NP→ PRP) to be the most effective. In some languages such as French, the gender of nouns (including the speaker) is often marked in the syntax. For example, a male would write '*je suis allé*', while a female would write *'je suis allée'* ('I went'). By detecting such *'je suis'* constructions, Ciot, Sonderegger, and Ruths (2013) improved performance of gender classification in French.

Stylistic features have been widely explored as well. Studies have reported that males tend to use longer words, sentences and texts (Goswami, Sarkar, and Rustagi 2009; Otterbacher 2010; Singh 2001), and more swear words (Boulis and Ostendorf 2005; Schwartz et al. 2013). Females use more emotion words (Bamman, Eisenstein, and Schnoebelen 2014; Nowson and Oberlander 2006; Schwartz et al. 2013), emoticons (Bamman, Eisenstein, and Schnoebelen 2014; Gianfortoni, Adamson, and Rosé 2011; Rao et al. 2010; Volkova, Wilson, and Yarowsky 2013), and typical social media words such as *omg* and *lol* (Bamman, Eisenstein, and Schnoebelen 2014; Schler et al. 2006).

Groups can be characterized by their attributes, for example females tend to have maiden names. Bergsma and Van Durme (2013) used such distinguishing attributes, extracted from common nouns for males and females (e.g., granny, waitress), to improve classification performance. Features based on first names have also been explored. Although not revealing much about language use itself, they can improve prediction performance (Bergsma and Van Durme 2013; Burger et al. 2011; Rao et al. 2011).

*Genre.* So far, not many studies have analyzed the influence of genre and domain (Lee 2001) on language use, but a better understanding will aid the interpretation of observed language variation patterns. Using data from the British National Corpus, Argamon et al. (2003) found a strong correlation between characteristics of male and non-fiction writing and likewise, between female and fiction writing. Based on this observation, they trained separate prediction models for fiction and non-fiction (Koppel, Argamon, and Shimoni 2002). Building on these findings, Herring and Paolillo (2006) investigated whether gender differences would still be observed when controlling for genre in blogs. They did not find a significant relation between gender and linguistic features that were identified to be associated with gender in previous literature, however the study was based on a relatively small sample. Similarly, Gianfortoni, Adamson, and Rosé (2011) revisited the task of gender prediction on the Blog Authorship Corpus. After controlling for occupation, features that previously were found to be predictive for gender on that corpus were not effective anymore.

Studies focusing on gender prediction have tested the generalizability of gender prediction models by training and testing on different datasets. Although models tend to perform worse when tested on a different dataset than the one used for training, studies have shown that prediction performance is still higher than random, suggesting that there are indeed gender-specific patterns of language variation that go beyond genre and domain (Sap et al. 2014; Sarawgi, Gajulapalli, and Choi 2011). Gianfortoni, Adamson, and Rosé (2011) proposed the use of 'stretchy patterns', flexible sequences of categories, to model stylistic variation and to improve generalizability across domains.

*Social Interaction.* Most computational studies on gender-specific patterns in language use have studied speakers in isolation. As the conversational partner[4] and social network influence the language use of speakers, several studies have extended their focus by also considering contextual factors. For example, this led to the finding that speakers use more gender-specific language in same-gender conversations (Boulis and Ostendorf 2005). On the Fisher and Switchboard corpus (telephone conversations), classifiers dependent on the gender of the conversation partner improve performance (Garera and Yarowsky 2009). However, exploiting the social network of speakers on Twitter has been less effective so far. Features derived from the friends of Twitter users did not improve gender classification (but it was effective for age) (Zamal, Liu, and Ruths 2012). Likewise, Bamman, Eisenstein, and Schnoebelen (2014) found that social network information of Twitter users did not improve gender classification when enough text was available.

Not all computational studies on gender in interaction contexts have focused on gender classification itself. Some have used gender as a variable when studying other phenomena. In a study on language and power, Prabhakaran, Reid, and Rambow (2014) showed how the gender composition of a group influenced how power is manifested in the Enron corpus, a large collection of emails from Enron employees (described in more detail in Section 4.1). In a study on language change in online communities, Hemphill and Otterbacher (2012) found that females write more like men over time in the IMDb community (a movie review site), which they explain by men receiving more prestige in the community. Jurafsky, Ranganath, and McFarland (2009) automatically classified speakers according to interactional style (awkward, friendly, or flirtatious) using various types of features, including lexical features based on LIWC (Pennebaker, Francis, and Booth 2001), prosodic, and discourse features. Differences, as well as commonalities, were observed between genders, and incorporating features from both speakers improved classification performance.

**3.2.3 Interpretation of Findings.** As mentioned before, most computational approaches adopt a simplistic view of gender as an inherent property based on biological characteristics. Only recently, the computational linguistics community has noticed the limitations of this simplistic view by acknowledging the agency of speakers. Two of these studies based their argumentation on an analysis of the social networks of the users. Automatic gender predictions on YouTube data correlated more strongly with the dominant gender in a user's network than the user-reported gender (Filippova 2012). Likewise, in experiments by Bamman, Eisenstein, and Schnoebelen (2014), incorrectly labeled Twitter users also had fewer same-gender connections. In addition, they identified clusters of users who used linguistic markers that conflicted with general population-level findings. Another study was based on data collected from an online game (Nguyen et al. 2014). Thousands of players guessed the age and gender of Twitter users based on their tweets, and the results revealed that many Twitter users do not tweet in a gender-stereotypical way.

Thus, language is inherently social and while certain language features are *on average* used more by males or females, individual speakers may diverge from the stereotypical images that tend to be highlighted by many studies. In addition, gender is shaped differently depending on the culture and language, and thus presenting gender

4 An individual who participates in a conversation, sometimes also referred to as interlocutor or addressee

as a universal social variable can be misleading. Furthermore, linguistic variation within speakers of the same gender holds true as well.

### 3.3 Age

Aging is a universal phenomenon and understanding the relation between language and age can provide interesting insights in many ways. An individual at a specific time represents both a place in history as well as a life stage (Eckert 1997), and thus observed patterns can generate new insights into language change as well as how individuals change their language use as they move through life. Within computational linguistics, fewer studies have focused on language variation according to age compared to studies focusing on gender, possibly because obtaining age labels requires more effort than gender labels (e.g., the gender of people can often be derived from their names; cf. Section 3.1). Most of these studies have focused on absolute chronological age, although age can also be seen as a social variable like gender.

Sociolinguistic studies have found that adolescents use the most non-standard forms, because at a young age the group pressure to not conform to established societal conventions is the largest (Eckert 1997; Holmes 2013). In contrast, adults are found to use the most standard language, because for them social advancement matters and they use standard language to be taken seriously (Bell 2013; Eckert 1997). These insights can explain why predicting the ages of older people is harder, e.g., distinguishing between a 15- and a 20-year old person based on their language use is easier than distinguishing between a 40- and a 45-year old person (Nguyen et al. 2013, 2014). Thus, age is an important variable to consider, especially when we consider processes relevant for language evolution, since the degree of language innovation varies by age (Labov 2001).

**3.3.1 Modeling Age.** A fundamental question is *how* to model age, and so far researchers have not reached a consensus yet. Eckert (1997) distinguishes between chronological age (number of years since birth), biological age (physical maturity) and social age (based on life events). Speakers are often grouped according to their age, because the amount of data is in many cases not sufficient to make more fine-grained distinctions (Eckert 1997). Most studies consider chronological age and group speakers based on age spans (Barbieri 2008; Labov 1966; Trudgill 1974). However, chronological age can be misleading since persons with the same chronological age may have had very different life experiences. Another approach is to group speakers according to 'shared experiences of time', such as high school students (Eckert 1997).

Within computational linguistics the most common approach is to model age-specific language use based on the chronological age of speakers. An exception is Nguyen et al. (2013) who explored classification into life stages. However, even when focusing on chronological age, the task can be framed in different ways as well. Chronological age prediction has mostly been approached as a *classification* problem, by modeling the chronological age as a *categorical* variable. Based on this task formulation, various classical machine learning models have been used, such as SVMs (Peersman, Daelemans, and Vaerenbergh 2011; Rao et al. 2010), logistic regression (Nguyen et al. 2013; Rosenthal and McKeown 2011) and Naive Bayes (Tam and Martell 2009).

The boundaries used for discretizing age have varied depending on the dataset and experimental setup. Experiments on the blog authorship corpus (Schler et al. 2006) used categories based on the following age spans: 13-17, 23-27, and 33-47, removing the age ranges in between to simplify the task. Rangel et al. (2013) adopted this approach in the Author Profiling task at PAN 2013. The following year, the difficulty of the task at PAN

2014 was increased by considering the more fine-grained categories of 18-24, 25-34, 35-49, 50-64 and 65+ (Rangel et al. 2014). Zamal, Liu, and Ruths (2012) classified Twitter users into 18-23 and 25-30. Other studies explored boundaries at 30 (Rao et al. 2010), at 20 and 40 (Nguyen et al. 2013), at 40 (Garera and Yarowsky 2009) and at 18 (Burger and Henderson 2006).

In several studies experiments have been done by varying the classification boundaries. Peersman, Daelemans, and Vaerenbergh (2011) experimented with binary classification and boundaries at 16, 18 and 25. Tam and Martell (2009) experimented with classifying teens versus 20s, 30s, 40s, 50s and adults. Not surprisingly, in both studies a higher performance was obtained when using larger age gaps (e.g., teens versus 40s/50s) than when using smaller age gaps (e.g., teens versus 20s/30s) (Peersman, Daelemans, and Vaerenbergh 2011; Tam and Martell 2009). Rosenthal and McKeown (2011) explored a range of splits to study differences in performance when predicting the birth year of blog authors. They related their findings to pre- and post social media generations.

For many applications, modeling age as a categorical variable might be sufficient. However, it does have several limitations. First, selecting age boundaries has proven to be difficult. It is not always clear which categories are meaningful. Secondly, researchers have used different categories depending on the age distribution of their dataset, which makes it difficult to make comparisons across datasets.

Motivated by such limitations, recent studies have modeled age as a *continuous* variable, removing the need to define age categories. Framing age prediction as a regression task, a frequently used method has been linear regression (Nguyen, Smith, and Rosé 2011; Nguyen et al. 2013; Sap et al. 2014; Schwartz et al. 2013). Liao et al. (2014) experimented with a latent variable model that jointly models age and topics. In their model, age-specific topics obtain low standard deviations of age, while more general topics obtain high standard deviations. Another approach that would remove the need to define age categories is the unsupervised induction of age categories. Analyzing the discovered age groups could shed more light on the relation between language use and age, but we are not aware of existing research in this area.

**3.3.2 Features and Patterns.** The majority of studies on age prediction have focused on identifying predictive features. While some features tend to be effective across domains, others are domain-specific (Nguyen, Smith, and Rosé 2011). Features that characterize male speech have been found to also increase with age (Argamon et al. 2007), thus simply said, males tend to sound older than they are.

Unigrams alone already perform well (Nguyen, Smith, and Rosé 2011; Nguyen et al. 2013; Peersman, Daelemans, and Vaerenbergh 2011). Features based on part of speech are effective as well. For example, younger people tend to use more first and second person singular pronouns (e.g., *I, you*), while older people more often use first person plural pronouns (e.g., *we*) (Barbieri 2008; Nguyen et al. 2013; Rosenthal and McKeown 2011). Older people also use more prepositions (Argamon et al. 2009; Nguyen et al. 2013), determiners (Argamon et al. 2009; Nguyen, Smith, and Rosé 2011) and articles (Schwartz et al. 2013). Most of these studies focused on English and therefore some of these findings might not be applicable to other languages. For example, the effectiveness of pronoun-related features should also be studied in pro-drop languages (e.g., Turkish and Spanish).

Various studies have found that younger people use less standard language. They use more alphabetical lengthening (e.g., *niiiice*) (Nguyen et al. 2013; Rao et al. 2010), more contractions without apostrophes (e.g., *dont*) (Argamon et al. 2009), more Internet

acronyms (e.g., *lol*) (Rosenthal and McKeown 2011), more slang (Barbieri 2008; Rosenthal and McKeown 2011), more swear words (Barbieri 2008; Nguyen, Smith, and Rosé 2011), and more capitalized words (e.g., *HAHA*) (Nguyen et al. 2013; Rosenthal and McKeown 2011). Specific words such as *like* are also highly associated with younger ages (Barbieri 2008; Nguyen, Smith, and Rosé 2011). Younger people also use more features that indicate stance and emotional involvement (Barbieri 2008), such as intensifiers (Barbieri 2008; Nguyen et al. 2013) and emoticons (Rosenthal and McKeown 2011). Younger people also use shorter words and sentences and write shorter tweets (Burger and Henderson 2006; Nguyen et al. 2013; Rosenthal and McKeown 2011).

**3.3.3 Interpretation of Findings.** Age prediction experiments are usually done on datasets collected at a specific point in time. Based on such datasets, language use is modeled and compared between users with different ages. Features that are found to be predictive or that correlate highly with age are used to highlight how differently 'younger' and 'older' people talk or write. However, the observed differences in language use based on such datasets could be explained in multiple ways. Linguistic variation can occur as an individual moves through life (*age grading*). In that case the same trend is observed for individuals at different time periods. Linguistic variation can also be a result of changes in the community itself as it moves through time (*generational change*) (Bell 2013; Sankoff 2006). For example, suppose we observe that younger Twitter users include more smileys in their tweets. This could indicate that smiley usage is higher at younger ages, but that when Twitter users grow older they decrease their usage of smileys. However, this could also indicate a difference in smiley usage between generations (i.e., the generation of the current younger Twitter users use more smileys compared to the generation of the older Twitter users). This also points to the relation between synchronic variation and diachronic change. Synchronic variation is variation across different speakers or speech communities at a particular point in time, while diachronic change is accumulation of synchronic variation in time and frequency. To have a better understanding of change, we need to understand the spread of variation across time and frequency. As is the case for gender, age can be considered a social variable and thus when only modeling chronological age, we are ignoring the agency of speakers and that speakers follow different trajectories in their lives.

### 3.4 Location

Regional variation has been extensively studied in sociolinguistics and related areas such as dialectology (Chambers and Trudgill 1998) and dialectometry (Wieling and Nerbonne 2015). The use of certain words, grammatical constructions, or the pronunciation of a word, can often reveal where a speaker is from. For example, '*yinz*' (a form of the second-person pronoun) is mostly used around Pittsburgh, which can be observed on Twitter as well (Eisenstein 2015). Dialectology traditionally focuses on the geographical distribution of individual or small sets of linguistic variables (Chambers and Trudgill 1998). A typical approach involves identifying and plotting *isoglosses*, lines that divide maps into regions where specific values of the variable predominate. The next step involves identifying bundles of isoglosses, often followed by the identification of dialect regions. While these steps have usually been done manually, computational approaches have recently been explored as well. For example, Grieve, Speelman, and Geeraerts (2011) demonstrated how methods from spatial analysis can be used for automating such an analysis.

The study of regional variation has been heavily influenced by new statistical approaches, such as from computational linguistics, machine learning and spatial analysis. A separate branch has also emerged, referred to as dialectometry (Wieling and Nerbonne 2015). In contrast to dialectology, which focuses on individual linguistic variables, dialectometry involves aggregating linguistic variables to examine linguistic differences between regions. Nerbonne (2009) argues that studies that focus on individual variables are sensitive to noise and that therefore aggregating linguistic variables will result in more reliable signals. This aggregation step has led to the introduction of various statistical methods, including clustering, dimensionality reduction techniques and regression approaches (Heeringa and Nerbonne 2013; Nerbonne and Wieling 2015; Wieling and Nerbonne 2010). Recently, researchers within dialectometry have explored the automatic identification of characteristic features of dialect regions (Wieling and Nerbonne 2010), a task which aligns more closely with the approaches taken by dialectologists.

While the datasets typically used in dialectology and dialectometry studies are still small compared to datasets used in computational linguistics, similar statistical methods have been explored. This has created a promising starting point for closer collaboration with computational linguistics.

**3.4.1 Modeling Geographical Variation.** Within CL, we find two lines of work on computationally modeling geographical variation.

*Supervised.* The first approach starts with documents labeled according to their dialect, which can be seen as a supervised learning approach. Most studies taking this approach focus on automatic dialect identification, which is a variation of automatic language identification, a well-studied research topic within the field of computational linguistics (Baldwin and Lui 2010; Hughes et al. 2006). While some have considered automatic language identification a solved problem (McNamee 2005), still many outstanding issues exist (Hughes et al. 2006), including the identification of dialects and closely related languages (Zampieri et al. 2014, 2015). In studies on automatic dialect identification, various dialects have been explored, including Arabic (Darwish, Sajjad, and Mubarak 2014; Elfardy and Diab 2013; Huang 2015; Zaidan and Callison-Burch 2013), Turkish (Doğruöz and Nakov 2014), Swiss German (Scherrer and Rambow 2010) and Dutch (Trieschnigg et al. 2012) dialects.

*Unsupervised.* An alternative approach is to start with location-tagged data to automatically identify dialect regions. While the models are given labels indicating the locations of speakers, the dialect labels themselves are not observed. In the context of modeling dialects, we consider it an unsupervised approach (although it can be considered a supervised approach when the task is framed as a location prediction task). The majority of the work in this area has used Twitter data, because it contains fine-grained location information in the form of GPS data for tweets or user-provided locations in user profiles.

Much of the research that starts with location-tagged data is done with the aim of automatically predicting the location of speakers. The setup is thus similar to the setup for the other tasks that we have surveyed in this section (e.g., gender and age prediction). Eisenstein et al. (2010) developed a topic model to identify geographically coherent linguistic regions and words that are highly associated with these regions. The model was tested by predicting the locations of Twitter users based on their tweets. While the topic of text-based location prediction has received increasing attention (Han,

Cook, and Baldwin 2012; Wing and Baldridge 2011), using these models for the discovery of new sociolinguistic patterns is an option that has not been fully explored yet, since most studies primarily focus on prediction performance.

Various approaches have been explored to model the location of speakers, an aspect that is essential in many of the studies that start with location-tagged data. In Wing and Baldridge (2011), locations are modeled using geodesic grids, but these grids do not always correspond to administrative or language boundaries. Users can also be grouped based on cities (Han, Cook, and Baldwin 2012), but such an approach is not suitable for users in rural areas or when the focus is on more fine-grained geographical variation (e.g., within a city). Eisenstein et al. (2010) model regions using Gaussian distributions, but only focus on the United States and thus more research is needed to investigate the suitability of this approach when considering other countries or larger regions.

**3.4.2 Features and Patterns.** Word and character n-gram models have been frequently used in dialect identification (King, Radev, and Abney 2014; Trieschnigg et al. 2012; Zaidan and Callison-Burch 2013). Similarly, many text-based location prediction systems make use of unigram word features (Eisenstein et al. 2010; Han, Cook, and Baldwin 2012; Wing and Baldridge 2011).

Features inspired by sociolinguistics could potentially improve performance. Darwish, Sajjad, and Mubarak (2014) showed that for identifying Arabic dialects a better classification performance could be obtained by incorporating known lexical, morphological and phonological differences in their model. Scherrer and Rambow (2010) also found that using linguistic knowledge improves over an n-gram approach. Their method is based on a linguistic atlas for the extraction of lexical, morphological and phonetic rules and the likelihood of these forms across German-speaking Switzerland. Doğruöz and Nakov (2014) explored the use of light verb constructions to distinguish between two Turkish dialects.

To support the discovery of new sociolinguistic patterns and to improve prediction performance, several studies have focused on automatically identifying characteristic features of dialects. Han, Cook, and Baldwin (2012) explored various feature selection methods to improve location prediction. The selected features may reflect dialectal variation but this was not the focus of the study. The method by Prokić, Çöltekin, and Nerbonne (2012) was based on in-group and out-group comparisons using data in which linguistic varieties were already grouped (e.g., based on clustering). Peirsman, Geeraerts, and Speelman (2010) compared frequency-based measures, such as chi-square and log-likelihood tests, with distributional methods. Automatic methods may identify many features that vary geographically such as topic words and named entities, and an open challenge is to separate this type of variation from the more sociolinguistically interesting variations. For example, the observation that the word '*beach*' is used more often near coastal areas or that '*Times Square*' is used more often in New York is not interesting from the perspective of a sociolinguist.

Making use of location-tagged data, several studies have focused on analyzing patterns of regional variation. Doyle (2014) analyzed the geographical distribution of dialectal variants (e.g., the use of double modals like '*might could*') based on Twitter data, and compared it with traditional sociolinguistic data collection methods. Starting with a query-based approach, he uses baseline queries (e.g., '*I*') for estimating a conditional distribution of data given metadata. His approach achieved high correlations with data from sociolinguistic studies. Jørgensen, Hovy, and Søgaard (2015) studied the use of three phonological features of African American Vernacular English using manually selected word pairs. The occurrence of the features was correlated with location data

(longitude and latitude) as well as demographic information obtained from the US census bureau. While these approaches start with attested dialect variants, automatic discovery of unknown variation patterns could potentially lead to even more interesting results. To study how a word's meaning varies geographically, Bamman, Dyer, and Smith (2014) extended the skip gram model by Mikolov et al. (2013) by adding contextual variables that represent states from the US. The model then learns a global embedding matrix and additional matrices for each context (e.g., state) to capture the variation of a word's meaning.

The increasing availability of longitudinal data has made it possible to study the spreading of linguistic innovations geographically and over time on a large scale. A study by Eisenstein et al. (2014) based on tweets in the United States indicates that linguistic innovations spread through demographically similar areas, in particular with regard to race.

**3.4.3 Interpretation of Findings.** Labeling texts by dialect presumes that there are clear boundaries between dialects. However, it is not easy to make absolute distinctions between language varieties (e.g., languages, dialects). Chambers and Trudgill (1998) illustrate this with the example of traveling from village to village in a rural area. Speakers from villages at larger distances have more difficulty understanding each other compared to villages that are closer to each other, but there is no clear-cut distance at which speakers are no longer mutually intelligible. A computational approach was taken by Heeringa and Nerbonne (2001) to shed more light on this puzzling example. Besides linguistic differences, boundaries between language varieties are often influenced by other factors such as political boundaries (Chambers and Trudgill 1998). Therefore, deciding on the appropriate labels to describe linguistic communication across different groups of speakers (in terms of language, dialect, minority language, regional variety, etc.) is an on-going issue of debate. The arbitrariness of the distinction between a language and dialect is captured with the popular expression "*A language is a dialect with an army and navy*" (Bright 1997). Methods that do not presume clear dialect boundaries are therefore a promising alternative. However, such methods then rely on location-tagged data, which is usually only available for a portion of the data.

### 3.5 Text Classification Informed by Identity Information

So far, we have focused on automatically predicting the variables themselves (e.g., gender, age, location) but linguistic variation related to the identity of speakers can also be used to improve various other NLP tasks. Dadvar et al. (2012) trained gender-specific classifiers to detect instances of cyberbullying, noticing that language used by harassers varies by gender. To improve the prediction performance of detecting the power direction between participants in emails, Prabhakaran, Reid, and Rambow (2014) incorporated the gender of participants in e-mail conversations and the overall 'gender environment' as features in their model. Volkova, Wilson, and Yarowsky (2013) studied gender differences in the use of subjective language on Twitter. Representing gender as a binary feature was not effective, but the use of features based on gender-dependent sentiment terms improved subjectivity and polarity classification. Hovy (2015) found that training gender- or age-specific word embeddings improved tasks such as sentiment analysis and topic classification.

## 4. Language and Social Interaction

The previous section explored computational approaches to the study of identity construction through language. We discussed variables such as gender, age and geographical location, thereby mostly focusing on the influence of social structures on language use. However, as we also pointed out in the previous section, speaker agency enables violations of conventional language patterns. Speakers do not act in isolation, but they are part of pairs, groups and communities. Social interaction contexts produce the opportunity for variation due to agency. In response to the particulars of these social settings and encounters (e.g., the addressee or audience, topic, and social goals of the speakers), there is thus much variation within individual speakers. The variation that is related to the context of interaction will be the focus of this section.

We start this section with a discussion of data sources for large-scale analyses of language use in pairs, groups and communities (Section 4.1). Next, we discuss computational approaches to studying how language reflects and shapes footing within social relationships (Section 4.2). Much of this work has revolved around the role of language in power dynamics by studying how speakers use language to maintain and change power relations (Fairclough 1989). We will continue with a discussion on style-shifting (i.e., the use of different styles by a single speaker) in Section 4.3. We will discuss two prominent frameworks within sociolinguistics, Audience Design (Bell 1984) and Communication Accommodation Theory (Giles, Coupland, and Coupland 1991), and discuss how these frameworks have been studied within the computational linguistics community. Finally, we will move our focus to the community level and discuss computational studies on how members adapt their language to conform to or sometimes diverge from community norms. One might speculate about how these micro-level processes might eventually become conventional, and therefore consider how these processes may lead to language change over time (Section 4.4).

### 4.1 Data Sources

Many of the types of data that are relevant for the investigation of concepts of social identity, are also relevant for work on communication dynamics in pairs, groups and communities. The availability of detailed interaction recordings in online data has driven and enabled much of the work on this topic within computational linguistics. A variety of online discussion forums have been analyzed, including online cancer support communities (Nguyen and Rosé 2011; Wang, Reitter, and Yen 2014), a street gang forum (Piergallini et al. 2014), and more recently discussion forums in Massive Open Online Courses (MOOCs) (Wen, Yang, and Rosé 2014b, 2014a). Review sites, such as TripAdvisor (Michael and Otterbacher 2014), IMDb (Hemphill and Otterbacher 2012) and beer review communities (Danescu-Niculescu-Mizil et al. 2013b), have also been used in studies on language in online communities.

The Enron email corpus is another frequently used data source. The Enron corpus is a large email corpus with messages from Enron employees, which was made public during the legal investigation of the Enron corporation. The corpus has been used in various studies, for example, investigations related to email classification (Klimt and Yang 2004) and structure of communication networks (Diesner and Carley 2005). In particular, in studies on language and social dynamics, the Enron email corpus has featured in analyses of power relationships (Diehl, Namata, and Getoor 2007; Gilbert 2012; Prabhakaran, Rambow, and Diab 2012b; Prabhakaran, Reid, and Rambow 2014), since Enron's organizational structure is known and can be integrated in studies on

hierarchical power structures connected with quantitative capacity theories of power. Such theories treat power as a stable characteristic that inheres in a person. An example theory within this space is Resource Dependency Theory (Pfeffer and Salancik 1978).

For studies that involve more dynamic notions of power (e.g., identifying individuals who are pursuing power), other resources have also been explored, including Wikipedia Talk Pages (Bender et al. 2011; Bracewell, Tomlinson, and Wang 2012; Danescu-Niculescu-Mizil et al. 2012; Swayamdipta and Rambow 2012), transcripts of political debates (Prabhakaran, John, and Seligmann 2013; Prabhakaran, Arora, and Rambow 2014) and transcripts of Supreme Court arguments (Danescu-Niculescu-Mizil et al. 2012).

### 4.2 Shaping Social Relationships

Language is not only a means to exchange information but language also contributes to the performance of action within interaction. Language serves simultaneously as a reflection of the relative positioning of speakers to their conversation partners as well as actions that accompany those positions (Ribeiro 2006). Sometimes distributions of these actions can be considered to cohere to such a degree that they can be thought of as defining conversational roles (Yang, Wen, and Rosé 2015). At a conceptual level, this work draws heavily from a foundation in linguistic pragmatics (Grice 1975; Levinson 1983) as well as sociological theories of discourse (Gee 2011; Tannen 1993), which each provide a complementary view. Concepts related to expectations or norms that provide the foundation for claiming such positions may similarly be described either from a philosophical perspective or a sociological one (Postmes, Spears, and Lea 2000). In viewing interaction as providing a context in which information and action may flow towards the accomplishment of social goals, speakers position themselves and others as sources or recipients of such information and action (Martin and Rose 2003). When performatives, i.e., speech acts used to perform an action, break norms related to social positions, they have implications for relational constructs such as politeness (Brown and Levinson 1987), which codifies rhetorical strategies for acknowledging and managing relational expectations while seeking to accomplish extra-relational goals. In the remaining part of this section, we focus on computational studies within this theme. We first discuss the general topic of automatic extraction of social relationships from text, and then focus on power and politeness.

*Automatic Extraction of Social Relationships.* Recognizing that language use may reveal cues about social relationships, studies within CL have explored the automatic extraction of different types of social relationships based on text. One distinction that has been made is between weak ties (e.g., acquaintances) and strong ties (e.g., family and close friends) (Granovetter 1973). Gilbert and Karahalios (2009) explored how different types of information (including messages posted) can be used to predict tie strength on Facebook. In this study, the predictions were done for ties within a selected sample. Bak, Kim, and Oh (2012) studied differences in self-disclosure on Twitter between strong and weak ties using automatically identified topics. Twitter users disclose more personal information to strong ties, but they show more positive sentiment towards weak ties, which may be explained by social norms regarding first-time acquaintances on Twitter.

Other studies have automatically extracted social relationships from more extensive datasets, enabling analyses of the extracted network structures. These studies have focused on extracting signed social networks, i.e., networks with positive and negative edges, for example based on positive and negative affinity between individuals or

formal and informal relationships. Work within this area has drawn from Structural Balance Theory (Heider 1946), which captures intuitions such as that when two individuals have a mutual friend, they are likely to be friends as well, and from Status Theory (Leskovec, Huttenlocher, and Kleinberg 2010), which involves edges that are directed and reflect status differences. Hassan, Abu-Jbara, and Radev (2012) developed a machine learning classifier to extract signed social networks and found that the extracted network structure mostly agreed with Structural Balance Theory. Krishnan and Eisenstein (2015) proposed an unsupervised model for extracting signed social networks, which they used to extract formal and informal relations in a movie-script corpus. Furthermore, their model also induced the social function of address terms (e.g., *dude*). To infer edge signs in a social network, West et al. (2014) formulated an optimization problem that combined two objectives, capturing the extent to which the inferred signs agreed with the predictions of a sentiment analysis model, and the extent to which the resulting triangles corresponded with Status and Structural Balance Theory.

*Power.* Work on power relations draws from social psychological concepts of relative power in social situations (Guinote and Vescio 2010), in particular aspects of relative power that operate at the level of individuals in relation to specific others within groups or communities. Relative power may be thought of as operating in terms of horizontal positioning or vertical positioning: Horizontal positioning relates to closeness and related constructs such as positive regard, trust and commitment, while vertical positioning relates to authority and related constructs such as approval and respect among individuals within communities. Within the areas of linguistics and computational linguistics, investigations have focused on how speakers use language to maintain and change power relations (Fairclough 1989). Operationalization and computational modeling of these two dimensions has important applications in the field of learning sciences (Howley, Mayfield, and Rosé 2013).

Within computational linguistics, much of the work related to analysis of power as it is reflected through language has focused on automatically identifying power relationships from text. Though some of the literature cited above is referenced in this work, the engagement between communities has remained so far at a simple level. Fine-grained distinctions between families of theories of power, and subtleties about the relationship between power and language are frequently glossed over. One way in which this is visible is in the extent to which the locus of meaning is treated as though it is in the text itself rather than an emergent property of the interaction between speakers. Though some references to external power structures and transient power relationships are mentioned, much room remains for deeper reflection on the connection between power and language.

Research in the computational linguistics community related to these issues is normally centered around classification tasks. Earlier studies have focused on hierarchical power relations based on the organizational structure, thereby frequently making use of the Enron corpus. Bramsen et al. (2011) extracted messages between pairs of participants and developed a machine learning classifier to automatically determine whether the messages of an author were UpSpeak (directed towards a person of higher status) or DownSpeak (directed towards a person of lower status). With a slightly different formulation of the task, Gilbert (2012) used logistic regression to classify power relationships in the Enron corpus and identified the most predictive phrases. Besides formulating the task as a classification task, ranking approaches have been explored as well (Diehl, Namata, and Getoor 2007; Nguyen et al. 2014; Prabhakaran, John, and

Seligmann 2013). For example, Prabhakaran, John, and Seligmann (2013) predicted the ranking of participants in political debates according to their relative poll standings.

Studies based on external power structures, such as the organizational structure of a company, treat power relations as static. Recent studies have adopted more dynamic notions of power. For example, Prabhakaran, Rambow, and Diab (2012b) discuss a setting with an employee in a Human Resources department who interacts with an office manager. The HR employee has power over the office manager when the situation is about enforcing a HR policy, but the power relation will be reversed when the topic is allocation of new office space. In their study using the Enron corpus, they compared manual annotations of situational power with the organization hierarchy and found that these were not well aligned. Other studies have focused on a more dynamic view of power as arising through asymmetries with respect to needed resources or other goals, as characterized in consent-based theories of power such as exchange theory (Guinote and Vescio 2010). This would include such investigations as identifying persons who are pursuing power (Bracewell, Tomlinson, and Wang 2012; Swayamdipta and Rambow 2012) and detecting influencers (Biran et al. 2012; Huffaker 2010; Nguyen et al. 2014; Quercia et al. 2011). This could also include studying how language use changes when users change their status in online communities (Danescu-Niculescu-Mizil et al. 2012).

Depending on the conceptualization of power and the used dataset, labels for the relations or roles of individuals have been collected in different ways, such as based on the organizational structure of Enron (Bramsen et al. 2011; Gilbert 2012), the number of followers in Twitter (Danescu-Niculescu-Mizil, Gamon, and Dumais 2011), standings in state and national polls to study power in political debates (Prabhakaran, John, and Seligmann 2013), admins and non-admins in Wikipedia (Bender et al. 2011; Danescu-Niculescu-Mizil et al. 2012), and manual annotation (Biran et al. 2012; Nguyen et al. 2014; Prabhakaran and Rambow 2013).

Many computational approaches within this sphere build on a foundation from pragmatics related to speech act theory (Austin 1975; Searle 1969), which has most commonly been represented in what are typically referred to as conversation, dialog or social acts (Bender et al. 2011; Ferschke, Gurevych, and Chebotar 2012). Such categories can also be combined into sequences (Bracewell, Tomlinson, and Wang 2012). Other specialized representations are also used, such as features related to turn taking style (Prabhakaran, John, and Seligmann 2013; Swayamdipta and Rambow 2012), topic control (Nguyen et al. 2014; Prabhakaran, Arora, and Rambow 2014; Strzalkowski et al. 2012), and 'overt displays of power', which Prabhakaran, Rambow, and Diab (2012a) define as utterances that constrain the addressee's actions beyond what the underlying dialog act imposes.

*Politeness.* Polite behavior contributes to maintaining social harmony and avoiding social conflict (Holmes 2013). Automatic classifiers to detect politeness have been developed to study politeness strategies on a large scale. According to politeness theory by Brown and Levinson (1987), three social factors influence linguistically polite behavior: social distance, relative power, and ranking of the imposition (i.e., cost of the request). Drawing from this theory, Peterson, Hohensee, and Xia (2011) performed a study on the Enron corpus by training classifiers to automatically detect formality and requests. Emails that contained requests or that were sent to people of higher ranks indeed tended to be more formal. According to politeness theory, speakers with greater power than their addressees are expected to be less polite (Brown and Levinson 1987). Danescu-Niculescu-Mizil et al. (2013a) developed a politeness classifier and found that in Wikipedia polite editors were more likely to achieve higher status, but once

promoted, they indeed became less polite. In StackExchange, a site with an explicit reputation system, users with a higher reputation were less polite than users with a lower reputation. Their study also revealed new interactions between politeness markings (e.g., '*please*') and morphosyntactic context.

### 4.3 Style Shifting

According to Labov (1972), there are no single-style speakers since speakers may switch between styles (style-shifting) depending on their communication partners (e.g., addressee's age, gender and social background). Besides the addressee, other factors such as the topic (e.g., politics vs. religion) or the context (e.g., a courtroom vs. family dinner) can contribute to style shifting. In early studies, Labov stated that "*styles can be arranged along a single dimension, measured by the amount of attention paid to speech*" (Labov 1972), which thus views style shifting as mainly something responsive. The work by Labov on style has been highly influential, but not everyone agreed with his explanation for different speech styles. We will discuss two theories (Communication Accommodation Theory and Audience Design) that have received much attention in both sociolinguistics and computational linguistics and that focus on the role of audiences and addressees on style. Even more recent theories are emphasizing the agency of speakers as they employ different styles to represent themselves in a certain way or initiate a change in the situation. Besides switching between styles, multilingual speakers may also switch between languages or dialects. This is discussed in more depth in Section 5.

*Communication Accommodation Theory.* Communication Accommodation Theory (CAT) (Giles, Taylor, and Bourhis 1973; Giles, Coupland, and Coupland 1991; Soliz and Giles 2014) seeks to explain why speakers accommodate[5] to each other during conversations. Speakers can shift their behavior to become more similar (convergence) or more different (divergence) to their conversation partners. Convergence reduces the social distance between speakers and converging speakers are often viewed as more favorable and cooperative. CAT has been developed in the 1970s and has its roots in the field of social psychology. While CAT has been studied extensively in controlled settings, e.g., Gonzales, Hancock, and Pennebaker (2010), only recently studies have been performed in uncontrolled settings such as Twitter conversations (Danescu-Niculescu-Mizil, Gamon, and Dumais 2011), online forums (Jones et al. 2014), Wikipedia Talk pages and Supreme Court arguments (Danescu-Niculescu-Mizil et al. 2012), and even movie scripts (Danescu-Niculescu-Mizil and Lee 2011).

Speakers accommodate to each other on a variety of dimensions, ranging from pitch and gestures, to the words that are used. Within computational linguistics, researchers have focused on measuring linguistic accommodation. LIWC has frequently been employed in these studies to capture stylistic accommodation, for example as reflected in the use of pronouns (Danescu-Niculescu-Mizil and Lee 2011; Danescu-Niculescu-Mizil, Gamon, and Dumais 2011; Jones et al. 2014; Niederhoffer and Pennebaker 2002). Speakers do not necessarily converge on all dimensions (Giles, Coupland, and Coupland 1991), which has also been observed on Twitter (Danescu-Niculescu-Mizil, Gamon, and Dumais 2011). Although earlier studies used correlations of specific features between participants, on turn-level or overall conversation-level (Levitan, Gravano, and

5 The phenomenon of adapting to the conversation partner has also been known as 'alignment', 'coordination' and 'entrainment'.

Hirschberg 2011; Niederhoffer and Pennebaker 2002; Scissors et al. 2009), these correlations fail to capture the temporal aspect of accommodation. The measure developed by Danescu-Niculescu-Mizil, Gamon, and Dumais (2011) is based on the increase in probability of a response containing a certain stylistic dimension given that the original message contains that specific stylistic dimension. Wang, Reitter, and Yen (2014) used a measure based on repetition of words (or syntactic structures) between target and prime posts. Jones et al. (2014) proposed a measure that takes into account that speakers differ in their tendency to accommodate to others. Similarly, Jain et al. (2012) used a Dynamic Bayesian Model to induce latent style states that group related style choices together in a way that reflects relevant styles within a corpus. They also introduce global accommodation states that provide more context in identification of style shifts in interactions that extend for more than a couple of turns.

Social roles and orientations taken up by speakers influence how conversations play out over time and computational approaches to measure accommodation have been used to study power dynamics (Danescu-Niculescu-Mizil, Gamon, and Dumais 2011; Danescu-Niculescu-Mizil et al. 2012; Jones et al. 2014). In a study on power dynamics in Wikipedia Talk pages and Supreme court debates, Danescu-Niculescu-Mizil et al. (2012) found that people with a lower status accommodated more than people with a higher status. In addition, users accommodated less once they became an admin in Wikipedia. Using the same Wikipedia data, Noble and Fernández (2015) found that users accommodated more towards users that occupied a more central position, based on eigenvector and betweenness centrality, in the social network. Furthermore, whether a user was an admin did not had a significant effect on the amount of coordination that highly central users received. From a different angle, Gweon et al. (2013) studied transactive exchange in debate contexts. Transactivity is a property of an assertion that requires that it displays reasoning (e.g., a causal mechanism) and refers to or integrates an idea expressed earlier in the discussion. In this context, high concentrations of transactivity reflect a balance of power in a discussion. In their data, higher levels of speech style accommodation were correlated with higher levels of transactivity.

*Audience Design.* In a classical study set in New Zealand, Allan Bell found that newsreaders used different styles depending on which radio station they were talking for, even when they were reporting the same news on the same day. Bell's audience design framework (Bell 1984) explains style shifting as a response to audiences and shares similarities with CAT. One of the differences with CAT is that different types of audiences are defined from the perspective of the speaker (ranging from addressee to eavesdropper) and thus can also be applied to settings in which there is only a one-way interaction (such as broadcasting). Social media provides an interesting setting to study how audiences influence style. In many social media platforms, such as Twitter or Facebook, multiple audiences (e.g., friends, colleagues) are collapsed into a single context. Users of such platforms often imagine an audience when writing messages and they may target messages to different audiences (Marwick and boyd 2011).

Twitter has been the focus of several recent large-scale studies on audience design. In a study on how audiences influence the use of minority languages on Twitter, Nguyen, Trieschnigg, and Cornips (2015) showed how characteristics of the audience influence language choice on Twitter by analyzing tweets from multilingual users in the Netherlands using automatic language identification. Tweets directed to larger audiences were more often written in Dutch, while within conversations users often switched to the minority language. In another study on audience on Twitter, Bamman and Smith (2015) showed that incorporating features of the audience improved

sarcasm detection. Furthermore, their results suggested that users tend to use the hashtag #sarcasm when they are less familiar with their audience. Pavalanathan and Eisenstein (2015a) studied two types of non-standard lexical variables: those strongly associated with specific geographical regions of the United States and variables that were frequently used in Twitter but considered non-standard in other media. The use of non-standard lexical variables was higher in messages with user mentions, which are usually intended for smaller audiences, and lower in messages with hashtags, which are usually intended for larger audiences. Furthermore, non-standard lexical variables were more often used in tweets addressed to individuals from the same metropolitan area. Using a different data source, Michael and Otterbacher (2014) showed that reviewers on the TripAdvisor site adjust their style to the style of preceding reviews. Moreover, the extent to which reviewers are influenced correlates with attributes such as experience of the reviewer and their sentiment towards the reviewed attraction.

### 4.4 Community Dynamics

As we just discussed, people adapt their language use towards their conversation partner. Within communities, norms emerge over time through interaction between members, such as the use of slang words and domain-specific jargon (Danescu-Niculescu-Mizil et al. 2013b; Nguyen and Rosé 2011), or conventions for indicating retweets in Twitter (Kooti et al. 2012). Community members employ such markers to signal their affiliation. In an online gangs forum, for example, graffiti style features were used to signal group affiliation (Piergallini et al. 2014). To become a core member of a community, members adopt such community norms. As a result, often a change in behavior can be observed when someone joins a community. Multiple studies have reported that members of online communities decrease their use of first person singular pronouns (e.g., '*I*') over time and increase their use of first person plural pronouns (e.g., '*we*') (Cassell and Tversky 2005; Danescu-Niculescu-Mizil et al. 2013b; Nguyen and Rosé 2011), suggesting a stronger focus on the community. Depending on the frequency of use and social factors, local accommodation effects could influence how languages change in the long term (Labov 1994, 2001). Fine-grained, large-scale analyses of language change are difficult in offline settings, but the emergence of online communities has enabled computational approaches for analyzing language change within communities.

Early investigations of this topic were based on data from non-public communities, such as email exchanges between students during a course (Postmes, Spears, and Lea 2000) and data from the Junior Summit '98, an online community where children from across the world discussed global issues (Cassell and Tversky 2005; Huffaker et al. 2006). In these communities, members joined at the same time. Furthermore, the studies were based on data spanning only several months.

More recent studies have used data from public, online communities, such as online forums and review sites. Data from these communities typically span longer time periods (e.g., multiple years). Members join these communities intermittently and thus, when new users join, community norms have already been established. Nguyen and Rosé (2011) analyzed an online breast cancer community, in which long-time members used forum-specific jargon, highly informal style, and showed familiarity and emotional involvement with other members. Time periods were represented by the distribution of high frequency words and measures such as Kullback-Leibler divergence were used to study how language changed over time. Members who joined the community showed increasing conformity to community norms during the first year of their participation. Based on these observations, a model was developed to determine

membership duration. Hemphill and Otterbacher (2012) also studied how members adopt community norms over time but focused specifically on gender differences. They studied changes in the use of various characteristics, such as hedging, word/sentence complexity and vocabulary richness, in IMDb (the Internet Movie Database), a community in which males tend to receive higher prestige than females.

Not only members change their behavior over time as they participate in a community, communities themselves are also constantly evolving. Kershaw, Rowe, and Stacey (2016) identified and analyzed word innovations in Twitter and Reddit based on variation in frequency, form and meaning. They performed their analyses on a global level, i.e., the whole dataset, and on a community level, based on applying a community detection algorithm to the Reddit data and grouping the geotagged tweets by geopolitical units.

Language change on both member-level and community-level was analyzed by Danescu-Niculescu-Mizil et al. (2013b) in two beer review communities. Language models were created based on monthly snapshots to capture the linguistic state of a community over time. Cross-entropy was then used to measure how much a certain post deviated from a language model. Members in these communities turned out to follow a two-stage lifecycle: They first align with the language of the community (innovative learning phase), however at some point they stop adapting their language (conservative phase). The point at which members enter the conservative phase turned out to be dependent on how long a user would end up staying in the community.

These studies illustrate the potential of using large amounts of online data to study language change in communities in a quantitative manner. However, in such analyses biases in the data should be considered carefully, especially when the dynamics and content of the data are not understood fully. For example, Pechenick, Danforth, and Dodds (2015) call into question the findings on linguistic change based on the Google books corpus, due to its bias towards scientific publications. Furthermore, they point out that prolific authors in the dataset can influence the findings as well.

## 5. Multilingualism and Social Interaction

Languages evolve due to the interaction of speakers within and outside their speech communities. Within sociolinguistics, multilingual speakers and speech communities have been studied widely with respect to the contexts and conditions of language mixing and/or switching across languages. We use the term 'multilingual speaker' for someone who has a repertoire of various languages and/or dialects and who may mix them depending on contextual factors like occasion (e.g., home vs. work) and conversation partners (e.g., family vs. formal encounters). This section is dedicated to computational approaches for analyzing multilingual communication in relation to the social and linguistic contexts. We first start with a brief introduction into multilingual communication from a sociolinguistic point of view. Later, we expand the discussion to include the analysis of multilingual communication using computational approaches.

Human mobility is one of the main reasons for interaction among speakers of different languages. Weinreich (1953) was one of the first to explain why and how languages come into contact and evolve under each other's influence in a systematic manner. Sociolinguists (Auer 1988; Gumperz 1982; Myers-Scotton 2002; Poplack, Sankoff, and Miller 1988) have studied various aspects of language contact and mixing across different contact settings.

Language mixing and code-switching are used interchangeably and there is not always a consensus on the terminology. According to Gumperz (1982), language mixing

refers to the mixing of languages within the same text or conversation. Wei (1998) describes language alternations at or above the clause level and calls it code-mixing. Romaine (1995) differentiates between inter-sentential (i.e., across sentences) and intra-sentential (i.e., within the same sentence) switches. Poplack, Sankoff, and Miller (1988) refer to complete languages shifts of individual users as code-switching.

Language mixing spans across a continuum ranging from occasional switches (e.g., words or fixed multi-word expressions) to more structural ones (e.g., morphological, syntactic borrowings). The duration and intensity of interaction between speakers of contact languages influence the types of switches. When the frequency of switched words increases in use, they may get established in the speech community and become borrowed/loan words (e.g., hip hop-related Anglicisms in a German hip hop forum (Garley and Hockenmaier 2012)).

Earlier studies on language mixing were mostly based on multilingual spoken data collected in controlled or naturalistic settings (Auer 1988; Myers-Scotton 1995). Nowadays, the wide-spread use of internet in multilingual populations provides ample opportunities for large-scale and in-depth analyses of mixed language use in online media (Danet and Herring 2007; Hinnenkamp 2008; Hinrichs 2006; Paolillo 2001; Tsaliki 2003). Still most of these studies focus on qualitative analyses of multilingual online communication with limited data in terms of size and duration.

The rest of this section presents a discussion of data sources for studying multilingual communication on a large scale (Section 5.1). Consequently, we discuss research on adapting various NLP tools to process mixed-language texts (Section 5.2). We conclude this section with a discussion of studies that analyze, or even try to predict, the use of multiple languages in multilingual communication (Section 5.3).

### 5.1 Data Sources

In sociolinguistics, conversational data is usually collected by the researchers themselves, either among small groups of speakers at different times (Doğruöz and Backus 2007, 2009) or from the same group of speakers longitudinally (Milroy and Milroy 1978; Trudgill 2003). The manual transcription and annotation of data is time-intensive and costly. Multilingual data from online environments is usually extracted in small volumes and for short periods. Automatic analysis of this type of data has been difficult for most languages, especially when resources or technical support are lacking.

Within computational linguistics, there is a growing interest in the automatic processing of mixed-language texts. Lui, Lau, and Baldwin (2014) and Yamaguchi and Tanaka-Ishii (2012) studied automatic language identification in mixed-language documents from Wikipedia by artificially concatenating texts from monolingual sources into multilingual documents. However, such approaches lead to artificial language boundaries. More recently, social media (such as Facebook (Vyas et al. 2014), Twitter (Jurgens, Dimitrov, and Ruths 2014; Peng, Wang, and Dredze 2014; Solorio et al. 2014) and online forums (Nguyen and Doğruöz 2013)) provide large volumes of data for analyzing multilingual communication in social interaction. Transcriptions of conversations have been explored by Solorio and Liu (2008b), however their data was limited to three speakers. Language pairs that have been studied for multilingual communication include English-Hindi (Vyas et al. 2014), Spanish-English (Peng, Wang, and Dredze 2014; Solorio and Liu 2008a, 2008b), Turkish-Dutch (Nguyen and Doğruöz 2013), Mandarin-English (Adel, Vu, and Schultz 2013; Peng, Wang, and Dredze 2014), and French-English (Jurgens, Dimitrov, and Ruths 2014). Besides being a valuable resource for studies on multilingual social interaction, multilingual texts in social media have also been used to

improve general purpose machine translation systems (Huang and Yates 2014; Ling et al. 2013).

Processing and analyzing mixed-language data often requires identification of languages at the word level. Language identification is a well-researched problem in CL and we discussed it in the context of dialect identification in Section 3.4.1. Here, we discuss language identification for mixed-language texts. Several datasets are publicly available to stimulate research on language identification in mixed-language texts, including data from the shared task on Language Identification in Code-Switched Data (Solorio et al. 2014) covering four different language pairs on Twitter, romanized Algerian Arabic and French texts from the comments section of an online Algerian newspaper (Cotterell et al. 2014), Turkish-Dutch forum posts (Nguyen and Doğruöz 2013) and web documents in different languages (King and Abney 2013).

Annotation on a fine-grained level such as individual words has introduced new challenges in the construction of datasets. More fine-grained annotations require more effort and sometimes the segments are so short that they can no longer be clearly attributed to a particular language. For example, annotating the language of named entities remains a challenge in mixed-language texts. Named entities have been labeled according to the context (King and Abney 2013), ignored in the evaluation (Elfardy and Diab 2012b; Nguyen and Doğruöz 2013) or treated as a separate category (Elfardy and Diab 2012a; Solorio et al. 2014). Annotation at sentence-level is also challenging. For example, Zaidan and Callison-Burch (2013) annotated a large corpus for Arabic dialect identification using crowdsourcing. Their analysis indicated that many annotators over-identify their native dialect (i.e., they were biased towards labeling texts as written in their own dialect). Elfardy and Diab (2012a) presented guidelines to annotate texts written in dialectal variants of Arabic and Modern Standard Arabic on a word level.

### 5.2 NLP Tools for Multilingual Data

Most of the current NLP tools, such as parsers, are developed for texts written in a single language. Therefore, such tools are not optimized for processing texts containing multiple languages. In this section, we discuss the development of NLP tools that specifically aim to support the processing of multilingual texts. We start with research on automatic language identification, which is an important step in the preprocessing pipeline of many language-specific analysis tasks. Mixed-language documents have introduced new challenges to this task. We then continue with a discussion of work on various other NLP tools (e.g., parsers, topic modeling).

*Automatic Language Identification.* Automatic language identification is often the first step for systems that process mixed-language texts (Vyas et al. 2014). Furthermore, it supports large-scale analyses of patterns in multilingual communication (Jurgens, Dimitrov, and Ruths 2014; Kim et al. 2014; Papalexakis, Nguyen, and Doğruöz 2014). Most of the earlier research on automatic language identification focused on document-level identification of a single language (Baldwin and Lui 2010). To handle mixed-language texts, more fine-grained approaches have been explored, ranging from language identification at the sentence (Elfardy and Diab 2013; Zaidan and Callison-Burch 2013; Zampieri et al. 2014) and word level (Elfardy and Diab 2012b; King and Abney 2013; Nguyen and Doğruöz 2013; Solorio et al. 2014; Voss et al. 2014), approaches for text segmentation (Yamaguchi and Tanaka-Ishii 2012), and estimating the proportion of the various languages used within documents (Lui, Lau, and Baldwin 2014; Prager 1999). Depending on the application, different approaches may be suitable, but studies

that analyze patterns in multilingual communication have mostly focused on word-level identification (Nguyen and Doğruöz 2013; Solorio et al. 2014). Off-the-shelf tools developed for language identification at the document-level (e.g., the TextCat program (Cavnar and Trenkle 1994)) are not effective for word-level identification (Alex 2005; Nguyen and Doğruöz 2013). Language models (Elfardy and Diab 2012b; Nguyen and Doğruöz 2013) and dictionaries (Alex 2005; Elfardy and Diab 2012b; Nguyen and Doğruöz 2013), which are also commonly used in automatic language identification at the document level, have been explored. Furthermore, the context around the words has been exploited using Conditional Random Fields to improve performance on language identification at the word level (King and Abney 2013; Nguyen and Doğruöz 2013).

*Parsing.* Early studies on language mixing within computational linguistics focused on developing grammars to model language mixing (e.g., Joshi (1982)). However, the models developed in these early studies were not tested on empirical data. The more recently developed systems have been validated on large, real-world data. Solorio and Liu (2008b) explored various strategies to combine monolingual taggers to parse mixed-language texts. The best performance was obtained by including the output of the monolingual parsers as features in a machine learning algorithm. Vyas et al. (2014) studied the impact of different preprocessing steps on POS tagging of English-Hindi data collected from Facebook. Language identification and transliteration were the major challenges that impacted POS performance.

*Language and Topic Models.* Language models have been developed to improve speech recognition for mixed-language speech, by adding POS and language information to the language models (Adel, Vu, and Schultz 2013) or by incorporating syntactic inversion constraints (Li and Fung 2012). Peng, Wang, and Dredze (2014) developed a topic model that infers language-specific topic distributions based on mixed-language text. The main challenge for their model was aligning the inferred topics across languages.

### 5.3 Analysis and Prediction of Multilingual Communication

According to Thomason (2001), Gardner-Chloros and Edwards (2004), and Bhatt and Bolonyai (2011), social factors (e.g., attitudes and motives of the speakers, social and political context) are as important as linguistic factors in multilingual settings. Large-scale analysis of social factors in multilingual communication has only recently been possible with the availability of automatic language identification tools.

Twitter is frequently used as a resource for such studies. Focusing on language choice at the user level, researchers have extracted network structures, based on followers and followees (Eleta and Golbeck 2014; Kim et al. 2014), or mentions and retweets (Hale 2014), and analyzed the relation between the composition of such networks and the language choices of users. Users tweeting in multiple languages are often found to function as a bridge between communities tweeting in one language. Besides analyzing language choice at the user level, there is also an interest in the language choices for individual tweets. Jurgens, Dimitrov, and Ruths (2014) studied tweets written in one language but containing hashtags in another language. Automatic language identification was used to identify the languages of the tweets. However, as they note, some tweets were written in another language because they were automatically generated by applications rather than being a conscious choice of the user. Nguyen, Trieschnigg, and Cornips (2015) studied users in the Netherlands who tweeted in a minority language (Limburgish or Frisian) as well as in Dutch. Most tweets were written in Dutch, but

during conversations users often switched to the minority language (i.e., Limburgish or Frisian). Mocanu et al. (2013) analyzed the geographic distribution of languages in multilingual regions and cities (such as New York and Montreal) using Twitter.

In addition to the analysis of patterns in multilingual communication, several studies have explored the automatic prediction of language switches. The task may seem similar to automatic language identification, yet there are differences between the two tasks. Rather than determining the language of an utterance, it involves predicting whether the language of the next utterance is the same *without* having access to the next utterance itself. Solorio and Liu (2008a) were the first to predict whether a speaker will switch to another language in English-Spanish bilingual spoken conversations based on lexical and syntactic features. The approach was evaluated using standard machine learning metrics as well as human evaluators who rated the naturalness/human-likeness of the sentences the system generated. Papalexakis, Nguyen, and Doğruöz (2014) predicted when multilingual users switch between languages in a Turkish-Dutch online forum using various features, including features based on multi-word units and emoticons.

## 6. Research Agenda

Computational sociolinguistics is an emerging multidisciplinary field. Closer collaboration between sociolinguists and computational linguists could be beneficial to researchers from both fields. In this article, we have outlined some challenges related to differences in data and methods that must be addressed in order for synergy to be effective. In this section, we summarize the main challenges for advancing the field of computational sociolinguistics. These fall under three main headings, namely, expanding the scope of inquiry of the field, adapting methods to increase compatibility, and offering tools.

### 6.1 Expanding the Scope of Inquiry

The field of computational linguistics has begun to investigate issues that overlap with those of the field of sociolinguistics. The emerging availability of data that is of interest to both communities is an important factor, but in order for real synergy to come out of this, additional angles in the research agendas and tuning of the methodological frameworks in the respective communities would be needed.

*Going beyond lexical and stylistic variation.* Many studies within CL focus on lexical variation (e.g., Section 3 on social identity), possibly driven by the focus on prediction tasks. Stylistic variation has also received attention. Several of the discussed studies focus on variation in the usage of functional categories. For example, they zoom in on the usage of determiners, prepositions and pronouns for studying linguistic style accommodation (in Section 4.3). Others employ measures such as average word and sentence length (e.g., in Section 3). Advances in the area of stylometry (Stamatatos 2009) could inspire the exploration of more fine-grained features to capture style. Besides lexical and stylistic variation, linguistic variation also occurs on many other levels. Some computational studies have focused on phonological (Eisenstein 2013a; Jain et al. 2012; Jørgensen, Hovy, and Søgaard 2015) and syntactic (Doyle 2014; Gianfortoni, Adamson, and Rosé 2011; Johannsen, Hovy, and Søgaard 2015; Wiersma, Nerbonne, and Lauttamus 2010) variation, but so far the number of studies is limited. In combination

with the surge in availability of relevant data, these examples suggest that there seems to be ample opportunities for an extended scope.

*Extending focus to other social variables.* A large body of work exists on the modeling of gender, age and regional variation (Cf. Section 3). Other variables, like social class (Labov 1966), have barely received any attention so far within computational sociolinguistics. Although it is more difficult to obtain labels for some social variables, they are essential for a richer understanding of language variation and more robust analyses.

*Going beyond English and monolingual data.* The world is multilingual and multicultural, but English has received much more attention within computational sociolinguistics than other languages. There is a need for research to validate the generalizability of findings based on English data for other languages (Danet and Herring 2007). Furthermore, most studies within computational linguistics generally assume that texts are written in one language. However, these assumptions may not hold, especially in social media. A single user may use multiple languages, sometimes even within a syntactic unit, while most NLP tools are not optimized to process such texts. Tools that are able to process mixed-language texts will support the analysis of such data and shed more light on the social and linguistic factors involved in multilingual communication.

*From monomodal to multimodal data.* Another recommendable shift in scope would be a stronger focus on multimedia data. Video and audio recordings with a speech track encapsulate a form of language in which the verbal and nonverbal dimensions of human communication are available in an integrated manner and they represent a rich source for the study of social behavior. Among the so-called paralinguistic aspects for which detection models and evaluation frameworks exist are age, gender and affect (Schuller et al. 2010). The increasing volumes of recordings of spoken dialogue and aligned transcriptions, e.g., in oral history collections (Boyd 2013; De Jong et al. 2014), meeting recording archives (Janin et al. 2003), and video blogs (Biel et al. 2013), can add new angles to the investigation of sociolinguistic variation. In particular, the study of the interaction between (transcribed) speech, non-speech (laughter, sighs, etc.), facial expression and gestures is a promising area for capturing and predicting social variables as well as the related affective layers.

### 6.2 Adapting Methodological Frameworks to Increase Compatibility

To make use of the rich repertoire of theory and practice from sociolinguistics and to contribute to it, we have to appreciate the methodologies that underlie sociolinguistic research, e.g., the *rules of engagement* for joining into the ongoing scientific discourse. However, as we have highlighted in the methodology discussion earlier in the article, the differences in values between the communities can be perceived as a divide. While the CL community has experienced a history in which theory and empiricism are treated as the extreme ends of a spectrum, in the social sciences there is no such dichotomy, and empiricism contributes substantially to theory. Moving forward, research within computational sociolinguistics should build on and seek to partner in extending existing sociolinguistic theories and insights. This requires placing a strong focus on the interpretability of the developed models. The feasibility of such a shift in attention can be seen when observing successes of applied computational sociolinguistics work that has been adopted in other fields like health communication (Mayfield et al. 2014) and education (Rosé et al. 2008).

*Controlling for multiple variables.* Sociolinguistic studies typically control for multiple social variables (e.g., gender, age, social class, ethnicity). However, many studies in computational sociolinguistics focus on individual variables (e.g., only gender, or only age), which can be explained by the focus on social media data. The uncontrolled nature of social media makes it challenging to obtain data about the social backgrounds of the speakers and to understand the various biases that such datasets might have. The result is that models are frequently confounded, which results in low interpretability as well as limited justification for generalization to other domains.

On the other hand, much work in the CL community has focused on structured modeling approaches that take a step towards addressing these issues (Joshi et al. 2012, 2013). These approaches are very similar to the hierarchical modeling approaches used in sociolinguistic research to control for multiple sources of variation and thus avoid misattributing weight to extraneous variables. A stronger partnership within the field of CL between researchers interested in computational sociolinguistics and researchers interested in multi-domain learning would be valuable for addressing some of the limitations mentioned above. In this regard, inferring demographic variables automatically (see Section 3) may also help, since predicted demographic variables could be used in structuring the models. Another approach is the use of census data when location data is already available. For example, Eisenstein et al. (2014) studied lexical change in social media by using census data to obtain demographic information for the geographical locations. They justified their approach by assuming that lexical change is influenced by the demographics of the population in these locations, and not necessarily by the demographics of the particular Twitter users in these locations.

*Developing models that generalize across domains.* Many of the studies within the area of computational sociolinguistics have focused on a single domain. However, domain effects can influence the findings, such as which features are predictive for gender (e.g., Herring and Paolillo (2006)). Studies considering multiple domains enable distinguishing variables that work differently in different contexts, and therefore improve the interpretation of the findings. Recently, several studies within the area of computational sociolinguistics have performed experiments across domains (Sap et al. 2014; Sarawgi, Gajulapalli, and Choi 2011) and explored the effectiveness of domain adaptation approaches (Nguyen, Smith, and Rosé 2011; Piergallini et al. 2014). Another approach involves reconsidering the features used in an attempt to include more features with a deep connection with the predicted variable of interest. For example, Gianfortoni, Adamson, and Rosé (2011) show that features such as n-grams, usually reported to be predictive for gender classification, did not perform well after controlling for occupation in a blog corpus, but pattern-based features inspired by findings related to gender-based language practices did.

*Using sociolinguistics and the social sciences as a source of inspiration for methodological reflection.* Going forward, we need to appreciate where our work stands along an important continuum that represents a fundamental tension in the social sciences: qualitative approaches that seek to preserve the complexity of the phenomena of interest, versus quantitative approaches that discretize (but thereby also simplify) the phenomena to achieve more generalizability. For computational linguistics, a primarily quantitative field, work from research areas with a less strong or less exclusive focus on quantitative measures, such as sociolinguistics and the social sciences, could serve as a source of inspiration for methodological reflection. In this survey, we have questioned the operationalizations of the concepts of gender (Section 3.2), age (Section 3.3) and

language variety (Section 3.4) as discrete and static categories, based on insights from sociolinguistics. More critical reflection on such operationalizations could lead to a deeper insight into the limitations of the developed models and the incorrect predictions that they sometimes make.

### 6.3 Tuning NLP Tools to Requirements of Sociolinguistics Research

As a final important direction, we should consider what would be required for NLP tools to be supportive for sociolinguistic work.

*Developing models that can guide users of data analysis systems in taking next steps.* Sociolinguists are primarily interested in new insights about language use. In contrast, much of the work within CL is centered around highly specific analysis tasks that are isolated from scenarios of use, and the focus on the obtained performance figures for such tasks is fairly dominant. As Manning (2015) mentions: "*[..], there has been an over-focus on numbers, on beating the state of the art*". Only for few analysis methods, validation of the outcomes has been pursued (e.g., have we measured the right thing?) in view of the potential for integration of the models outside lab-like environments. Furthermore many of the models developed within CL make use of thousands of features. As a result, their value for practical data exploration tasks is therefore often limited. Sparse models, such as used in Eisenstein, Smith, and Xing (2011), that identify small sets of predictive features would be more suited for exploratory analysis. However, when the focus is on interpretability of the models, we must consider that the resulting average prediction performance of interpretable models may be lower (Piergallini et al. 2014).

*Developing pre-processing tools to support the analysis of language variation.* The performance of many developed NLP tools is lower on informal text. For example, POS taggers perform less well on texts written by certain user groups (e.g., younger people (Hovy and Søgaard 2015)) or on texts in certain language varieties (e.g., African American Vernacular English (Jørgensen, Hovy, and Søgaard 2015)). One of the approaches to improve the performance of tools has been to normalize the texts, but as Eisenstein (2013b) argues, doing so is removing the variation that is central to the study of sociolinguistics. To support deeper sociolinguistic analyses and to go beyond shallow features, we thus need pre-processing tools, such as POS taggers, that are able to handle the variation found in informal texts and that are not biased towards certain social groups.

## 7. Conclusion

While the computational linguistics field has historically emphasized interpretation and manipulation of the propositional content of language, another valid perspective on language is that it is a dynamic, social entity. While some aspects of language viewed from a social perspective are predictable, and thus behave much like other aspects more commonly the target of inquiry in the field, we must acknowledge that linguistic agency is a big part of how language is used to construct social identities, to build and maintain social relationships, and even to define the boundaries of communities. The increasing research on social media data has contributed to the insight that text can be considered as a data source that captures multiple aspects and layers of human and social behavior. The recent focus on text as social data and the emergence of computational social science are likely to increase the interest within the computational linguistics community on sociolinguistic topics. In this article, we have defined and set out a research agenda

for the emerging field of '*Computational Sociolinguistics*'. We have aimed to provide a comprehensive overview of studies published within the field of CL that touch upon sociolinguistic themes in order to provide an overview of what has been accomplished so far and where there is room for growth. In particular, we have endeavored to illustrate how the large-scale data-driven methods of our community can complement existing sociolinguistic studies, but also how sociolinguistics can inform and challenge our methods and assumptions.

**Acknowledgments**

Thanks to Mariët Theune, Dirk Hovy and Marcos Zampieri for helpful comments on the draft. Thanks also to the anonymous reviewers for their valuable and detailed feedback. This work was funded in part through NSF grant ACI-1443068 and ARL grant W911NF-11-2-0042. The first author was supported by the Netherlands Organization for Scientific Research (NWO) grant 640.005.002 (CATCH project FACT). The second author was supported by the Digital Humanities Research Grant from Tilburg University and a fellowship from the Netherlands Institute of Advanced Study in Humanities and Social Sciences.

**References**

Adel, Heike, Ngoc Thang Vu, and Tanja Schultz. 2013. Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 206–211, Sofia, Bulgaria.

Alex, Beatrice. 2005. An unsupervised system for identifying English inclusions in German text. In *Proceedings of the ACL Student Research Workshop*, pages 133–138, Ann Arbor, Michigan.

Androutsopoulos, Jannis. 2013. Online data collection. In Christine Mallinson, Becky Childs, and Gerard Van Herk, editors, *Data Collection in Sociolinguistics: Methods and Applications*. Routledge.

Ardehaly, Ehsan Mohammady and Aron Culotta. 2015. Inferring latent attributes of Twitter users with label regularization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 185–195, Denver, Colorado.

Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Auer, Peter. 1988. A conversation analytic approach to code-switching and transfer. In Monica Heller, editor, *Codeswitching: Anthropological and sociolinguistic perspectives*. Berlin: Mouton de Gruyter, pages 187–213.

Austin, John Langshaw. 1975. *How to do things with words*. Oxford University Press.

Backofen, Rolf and Gert Smolka. 1993. A complete and recursive feature theory. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 193–200, Columbus, Ohio.

Bak, JinYeong, Suin Kim, and Alice Oh. 2012. Self-disclosure and relationship strength in Twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 60–64, Jeju Island, Korea. Association for Computational Linguistics.

Baldwin, Timothy, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan.

Baldwin, Timothy and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, California.

Bamman, David, Chris Dyer, and Noah A. Smith. 2014. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834,

Baltimore, Maryland.
Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.
Bamman, David and Noah A. Smith. 2015. Contextualized sarcasm detection on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 574–577, Oxford, UK.
Barbieri, Federica. 2008. Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics*, 12(1):58–88.
Behrend, Tara S., David J. Sharek, Adam W. Meade, and Eric N. Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43(3):800–813.
Bell, Allan. 1984. Language style as audience design. *Language in Society*, 13(2):145–204.
Bell, Allan. 2013. *The guidebook to sociolinguistics*. John Wiley & Sons.
Bender, Emily M., Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57. Portland, Oregon.
Bergsma, Shane, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337, Montreal, Canada.
Bergsma, Shane and Benjamin Van Durme. 2013. Using conceptual class attributes to characterize social media users. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720, Sofia, Bulgaria.
Bhatt, Rakesh M. and Agnes Bolonyai. 2011. Code-switching and the optimal grammar of bilingual language use. *Bilingualism: Language and Cognition*, 14(04):522–546.
Biel, Joan-Isaac, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. 2013. Hi youtube!: Personality impressions and verbal content in social video. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, pages 119–126, Sydney, Australia.
Biran, Or, Sara Rosenthal, Jacob Andreas, Kathleen McKeown, and Owen Rambow. 2012. Detecting influencers in written online conversations. In *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, pages 37–45, Montreal, Canada.
Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
Bolander, Brook and Miriam A. Locher. 2014. Doing sociolinguistic research on computer-mediated data: A review of four methodological issues. *Discourse, Context & Media*, 3(0):14 – 26.
Boulis, Constantinos and Mari Ostendorf. 2005. A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 435–442, Ann Arbor, Michigan.
Boyd, Doug. 2013. *Special Issue: Oral History in the Digital Age*, volume 40. Oral History Review.
Bracewell, David B., Marc Tomlinson, and Hui Wang. 2012. A motif approach for identifying pursuits of power in social discourse. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 1–8, Palermo, Italy.
Bramsen, Philip, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 773–782, Portland, Oregon, USA.
Bright, William. 1997. Notes. *Language in Society*, 26(03):469–470.
Brody, Samuel and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, Scotland, UK.
Brown, Penelope and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*, volume 4 of *Studies in Interactional Sociolinguistics*. Cambridge University Press.
Bucholtz, Mary and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.
Burger, John D., John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In

*Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK.

Burger, John D. and John C. Henderson. 2006. An exploration of observable features related to blogger age. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 15–20, Menlo Park, California.

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

Carletta, Jean. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, 22(2):249–254.

Cassell, Justine and Dona Tversky. 2005. The language of online intercultural community formation. *Journal of Computer-Mediated Communication*, 10(2).

Cavnar, William B. and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, Las Vegas, US.

Chambers, Jack K. and Peter Trudgill. 1998. *Dialectology*. Cambridge University Press.

Choi, Bernard C. K. and Anita W. P. Pak. 2006. Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clin Invest Med*, 29(6):351–364.

Ciot, Morgane, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Seattle, Washington, USA.

Clopper, Cynthia G. 2013. Experiments. In Christine Mallinson, Becky Childs, and Gerard Van Herk, editors, *Data Collection in Sociolinguistics: Methods and Applications*. Routledge.

Cohen, Raviv and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy! In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 91–99, Cambridge, Massachusetts, USA.

Collins, Linda M. and Stephanie T. Lanza. 2010. *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. John Wiley & Sons.

Corney, Malcolm, Olivier de Vel, Alison Anderson, and George Mohay. 2002. Gender-preferential text mining of e-mail discourse. In *Proceedings of the 18th Annual Computer Security Applications Conference (ACSAC '02)*, pages 282–289, Las Vegas, Nevada.

Cotterell, Ryan, Adithya Renduchintala, Naomi Saphra, and Chris Callison-Burch. 2014. An Algerian Arabic-French code-switched corpus. In *LREC-2014 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*, Reykjavik, Iceland.

Dadvar, Maral, Franciska M. G. de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, pages 23–25, Ghent, Belgium.

Daelemans, Walter. 2013. Explanation in computational stylometry. In *Proceedings of the 14th international conference on Computational Linguistics and Intelligent Text Processing (CICLing'13) - Volume 2*, pages 451–462, Samos, Greece.

Danescu-Niculescu-Mizil, Cristian, Michael Gamon, and Susan Dumais. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th International Conference on World Wide Web*, pages 745–754, Hyderabad, India.

Danescu-Niculescu-Mizil, Cristian and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA.

Danescu-Niculescu-Mizil, Cristian, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708, Lyon, France.

Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013a. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria.

Danescu-Niculescu-Mizil, Cristian, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013b. No country for old members: User lifecycle and linguistic

change in online communities. In *Proceedings of the 22nd international conference on World Wide Web (WWW '13)*, pages 307–318, Rio de Janeiro, Brazil.

Danet, Brenda and Susan C. Herring, editors. 2007. *The multilingual Internet: Language, culture, and communication online*. Oxford University Press.

Darwish, Kareem, Hassan Sajjad, and Hamdy Mubarak. 2014. Verifiably effective Arabic dialect identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1465–1468, Doha, Qatar.

De Fina, Anna, Deborah Schiffrin, and Michael Bamberg, editors. 2006. *Discourse and Identity*. Cambridge University Press.

De Jong, Franciska, Arjan van Hessen, Tanja Petrovic, and Stef Scagliola. 2014. Croatian memories: Speech, meaning and emotions in a collection of interviews on experiences of war and trauma. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 26–31, Reykjavik, Iceland.

Di Eugenio, Barbara and Michael Glass. 2004. The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Diehl, Christopher P., Galileo Namata, and Lise Getoor. 2007. Relationship identification for social network discovery. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, pages 546–552, Vancouver, British Columbia, Canada.

Diesner, Jana and Kathleen M Carley. 2005. Exploration of communication networks from the Enron email corpus. In *Proceedings of SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security*, pages 3– 14, Newport Beach, CA, USA.

Doğruöz, A. Seza and Ad Backus. 2007. Postverbal elements in immigrant Turkish: Evidence of change? *International Journal of Bilingualism*, 11(2):185–220.

Doğruöz, A. Seza and Ad Backus. 2009. Innovative constructions in Dutch Turkish: An assessment of ongoing contact-induced change. *Bilingualism: Language and Cognition*, 12(01):41–63.

Doğruöz, A. Seza and Preslav Nakov. 2014. Predicting dialect variation in immigrant contexts using light verb constructions. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1391–1395, Doha, Qatar.

Doyle, Gabriel. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 98–106, Gothenburg, Sweden.

Dürscheid, Christa and Elisabeth Stark, 2011. *Digital Discourse. Language in the New Media*, chapter sms4science: An International Corpus-Based Texting Project and the Specific Challenges for Multilingual Switzerland. Oxford: Oxford University Press.

Eckert, Penelope. 1989. *Jocks and burnouts: Social categories and identity in the high school*. Teachers College Press.

Eckert, Penelope. 1997. Age as a sociolinguistic variable. In Florian Coulmas, editor, *The handbook of sociolinguistics*. Blackwell Publishers.

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

Eckert, Penelope and Sally McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

Eisenstein, Jacob. 2013a. Phonological factors in social media writing. In *Proceedings of the Workshop on Language Analysis in Social Media (LASM 2013)*, pages 11–19, Atlanta, Georgia.

Eisenstein, Jacob. 2013b. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia.

Eisenstein, Jacob. 2015. Written dialect variation in online social media. In Charles Boberg, John Nerbonne, and Dom Watt, editors, *Handbook of Dialectology*. Wiley.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287, Cambridge, MA.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE*, 9(11):e113114, 11.

Eisenstein, Jacob, Noah A. Smith, and Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:*

*Human Language Technologies-Volume 1*, pages 1365–1374, Portland, Oregon, USA.
Eleta, Irene and Jennifer Golbeck. 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41:424 – 432.
Elfardy, Heba and Mona Diab. 2012a. Simplified guidelines for the creation of large scale dialectal Arabic annotations. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey.
Elfardy, Heba and Mona Diab. 2012b. Token level identification of linguistic code switching. In *Proceedings of COLING 2012: Posters*, pages 287–296, Mumbai, India.
Elfardy, Heba and Mona Diab. 2013. Sentence level dialect identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Sofia, Bulgaria.
Fairclough, Norman. 1989. *Language and power*. London: Longman.
Ferschke, Oliver, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in Wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786, Avignon, France.
Filippova, Katja. 2012. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1478–1488, Jeju Island, Korea.
Fink, Clay, Jonathon Kopecky, and Maksym Morawski. 2012. Inferring gender from the content of tweets: A region specific example. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 459–462, Dublin, Ireland.
Gardner-Chloros, Penelope and Malcolm Edwards. 2004. Assumptions behind grammatical approaches to code-switching: When the blueprint is a red herring. *Transactions of the Philological Society*, 102(1):103–129.
Garera, Nikesh and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 710–718, Suntec, Singapore.
Garley, Matt and Julia Hockenmaier. 2012. Beefmoves: Dissemination, diversity, and dynamics of English borrowings in a German hip hop forum. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 135–139, Jeju Island, Korea.
Gee, James Paul. 2011. *An Introduction to Discourse Analysis: Theory and Method*. New York: Routledge, third edition.
Gianfortoni, Philip, David Adamson, and Carolyn P. Rosé. 2011. Modeling of stylistic variation in social media with stretchy patterns. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 49–59, Edinburgh, Scotland.
Gilbert, Eric. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1037–1046, Seattle, Washington, USA.
Gilbert, Eric and Karrie Karahalios. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220, Boston, Massachusetts, USA.
Giles, Howard and Nikolas Coupland. 1991. *Language: Contexts and consequences*. Mapping Social Psychology Series. Brooks/Cole Publishing Company.
Giles, Howard, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In Howard Giles, Justine Coupland, and Nikolas Coupland, editors, *Contexts of Accommodation*. Cambridge University Press, pages 1–68.
Giles, Howard, Donald M. Taylor, and Richard Bourhis. 1973. Towards a theory of interpersonal accommodation through language: some Canadian data. *Language in Society*, 2(2):177–192.
Glymour, Clark, Richard Scheines, Peter Spirtes, and Kevin Kelly. 1987. *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press.
Gonzales, Amy L., Jeffrey T. Hancock, and James W. Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19.
Goswami, Sumit, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *Proceedings of the Third International ICWSM Conference*, pages 214–217, San Jose, California.

Gouws, Stephan, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 20–29, Portland, Oregon.

Granovetter, Mark S. 1973. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.

Green, Neil. 1992. Meaning-text theory: Linguistics, lexicography, and implications. *Machine Translation*, 7(3):195–198.

Grice, H. Paul. 1975. *Logic and Conversation, Syntax and Semantics*, volume 3. Academic Press.

Grieve, Jack. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23:193–221.

Griffiths, Thomas L. and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Guinote, Ana and Theresa K. Vescio, editors. 2010. *The Social Psychology of Power*. The Guilford Press.

Gumperz, John J. 1982. *Discourse strategies*. Cambridge University Press.

Guy, Gregory R. 2013. The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? *Journal of Pragmatics*, 52:63 – 71.

Gweon, Gahgene, Mahaveer Jain, John McDonough, Bhiksha Raj, and Carolyn P. Rosé. 2013. Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. *International Journal of Computer-Supported Collaborative Learning*, 8(2):245–265.

Hale, Scott A. 2014. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 833–842, Toronto, Canada.

Han, Bo, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062, Mumbai, India.

Hassan, Ahmed, Amjad Abu-Jbara, and Dragomir Radev. 2012. Detecting subgroups in online discussions by modeling positive and negative relations among participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 59–70, Jeju Island, Korea.

Heeringa, Wilbert and John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change*, 13(03):375–400.

Heeringa, Wilbert and John Nerbonne, 2013. *Language and Space. An International Handbook of Linguistic Variation, Volume III: Dutch*, chapter Dialectometry. De Gruyter Mouton.

Heider, Fritz. 1946. Attitudes and cognitive organization. *The Journal of Psychology: Interdisciplinary and Applied*, 21(1):107–112.

Hemphill, Libby and Jahna Otterbacher. 2012. Learning the lingo?: Gender, prestige and linguistic adaptation in review communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 305–314, Seattle, Washington, USA.

Herring, Susan C., editor. 1996. *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*, volume 39 of *Pragmatics & Beyond New Series*. John Benjamins Publishing.

Herring, Susan C. 2004. Computer-mediated discourse analysis: An approach to researching online behavior. In Sasha Barab, Rob Kling, and James H. Gray, editors, *Designing for Virtual Communities in the Service of Learning*. New York: Cambridge University Press, pages 338 – 376.

Herring, Susan C. 2007. A faceted classification scheme for computer-mediated discourse. *Language@Internet*, 4(1).

Herring, Susan C. and John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Heskes, Tom, Kees Albers, and Bert Kappen. 2002. Approximate inference and constrained optimization. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 313–320, Acapulco, Mexico.

Hey, Tony, Stewart Tansley, and Kristin Tolle, editors. 2009. *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research.

Heylighen, Francis and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure.

*Foundations of Science*, 7(3):293–340.
Hinnenkamp, Volker. 2008. Deutsch, Doyc or Doitsch? Chatters as languagers–The case of a German–Turkish chat room. *International Journal of Multilingualism*, 5(3):253–275.
Hinrichs, Lars. 2006. *Codeswitching on the Web: English and Jamaican Creole in E-mail Communication*. John Benjamins Publishing Company.
Holmes, David I. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117.
Holmes, Janet. 1995. *Women, men and politeness*. Routledge.
Holmes, Janet. 2013. *An introduction to sociolinguistics*. Routledge.
Holmes, Janet and Miriam Meyerhoff, editors. 2003. *The handbook of language and gender*. Wiley-Blackwell.
Hovy, Dirk. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China.
Hovy, Dirk, 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 (WWW '15)*, pages 452–461, Florence, Italy.
Hovy, Dirk and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China.
Howley, Iris, Elijah Mayfield, and Carolyn P. Rosé, 2013. *The International Handbook of Collaborative Learning*, chapter Linguistic Analysis Methods for Studying Small Groups. Routledge.
Hu, Yuheng, Kartik Talamadupula, and Subbarao Kambhampati. 2013. Dude, srsly?: The surprisingly formal nature of Twitter's language. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 244–253, Boston, Massachusetts USA.
Huang, Fei. 2015. Improved Arabic dialect classification with social media data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2118–2126, Lisbon, Portugal.
Huang, Fei and Alexander Yates. 2014. Improving word alignment using linguistic code switching data. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–9, Gothenburg, Sweden.
Huffaker, David. 2010. Dimensions of leadership and social influence in online communities. *Human Communication Research*, 36(4):593–617.
Huffaker, David, Joseph Jorgensen, Francisco Iacobelli, Paul Tepper, and Justine Cassell. 2006. Computational measures for language similarity across time in online communities. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, pages 15–22, New York City, New York.
Hughes, Baden, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew Mackinlay. 2006. Reconsidering language identification for written language resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 485–488, Genoa, Italy.
Hyland, Ken. 2004. *Disciplinary Discourses: Social Interactions in Academic Writing*. The University of Michigan Press.
Jain, Mahaveer, John McDonough, Gahgene Gweon, Bhiksha Raj, and Carolyn P. Rosé. 2012. An unsupervised dynamic bayesian network approach to measuring speech style accommodation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 787–797, Avignon, France.
Janin, Adam, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, pages 364–367.
Johannsen, Anders, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China.
Jones, Simon, Rachel Cotterill, Nigel Dewdney, Kate Muir, and Adam Joinson. 2014. Finding Zelig in text: A measure for normalising linguistic accommodation. In *Proceedings of COLING 2014, the 25th International Conference on Computational*

*Linguistics: Technical Papers*, pages 455–465, Dublin, Ireland.

Jørgensen, Anna, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China.

Joshi, Aravind K. 1982. Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*, pages 145–150, Prague, Czechoslovakia.

Joshi, Mahesh, William W. Cohen, Mark Dredze, and Carolyn P. Rosé. 2012. Multi-domain learning: when do domains matter? In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1302–1312, Jeju Island.

Joshi, Mahesh, Mark Dredze, William W. Cohen, and Carolyn P. Rosé. 2013. What's in a domain? Multi-domain learning for multi-attribute data. In *Proceedings of NAACL-HLT 2013*, pages 685–690, Atlanta, Georgia.

Jurafsky, Dan, Rajesh Ranganath, and Dan McFarland. 2009. Extracting social meaning: Identifying interactional style in spoken conversation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 638–646, Boulder, Colorado.

Jurgens, David, Stefan Dimitrov, and Derek Ruths. 2014. Twitter users #codeswitch hashtags! #moltoimportante #wow. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 51–61, Doha, Qatar.

Kershaw, Daniel, 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, San Francisco, CA, USA.

Kim, Suin, Ingmar Weber, Li Wei, and Alice Oh. 2014. Sociolinguistic analysis of Twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 243–248, Santiago, Chile.

King, Ben and Steven Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1110–1119, Atlanta, Georgia.

King, Ben, Dragomir Radev, and Steven Abney. 2014. Experiments in sentence language identification with groups of similar languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 146–154, Dublin, Ireland.

Klavans, Judith L. and Philip Resnik, editors. 1996. *The balancing act: Combining symbolic and statistical approaches to language*. MIT press.

Klimt, Bryan and Yiming Yang. 2004. The Enron corpus: A new dataset for email classification research. In *Proceedings of the 15th European Conference on Machine Learning (ECML 2004)*, pages 217–226, Pisa, Italy.

Kokkos, Athanasios and Theodoros Tzouramanis. 2014. A robust gender inference model for online social networks and its application to LinkedIn and Twitter. *First Monday*, 19(9).

Koller, Daphne and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Kooti, Farshad, Haeryun Yang, Meeyoung Cha, Krishna Gummadi, and Winter Mason. 2012. The emergence of conventions in online social networks. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 194–201, Dublin, Ireland.

Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Krippendorff, Klaus, 2013. *Content Analysis: An Introduction to Its Methodology*, chapter Validity. SAGE Publications.

Krishnan, Vinodh and Jacob Eisenstein. 2015. "You're Mr. Lebowski, I'm the dude": Inducing address term formality in signed social networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1616–1626, Denver, Colorado.

Labov, William. 1966. *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1994. *Principles of Linguistic Change, Volume I, Internal Factors*. Wiley-Blackwell.

Labov, William. 2001. *Principles of Linguistic Change, Volume II, Social Factors*. Wiley-Blackwell.

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Life in the network: the coming age of computational social science. *Science*, 323(5915):721–723.

Lee, David Y.W. 2001. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3):37–72.

Leemann, Adrian, Marie-José Kolly, Ross Purves, David Britain, and Elvira Glaser. 2016. Crowdsourcing language change with smartphone applications. *PLoS ONE*, 11(1).

Leskovec, Jure, Daniel Huttenlocher, and Jon Kleinberg. 2010. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370, Atlanta, GA, USA.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge University Press.

Levitan, Rivka, Agustin Gravano, and Julia Hirschberg. 2011. Entrainment in speech preceding backchannels. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 113–117, Portland, Oregon, USA.

Li, Ying and Pascale Fung. 2012. Code-switch language model with inversion constraints for mixed language speech recognition. In *Proceedings of COLING 2012: Technical Papers*, pages 1671–1680, Mumbai, India.

Liao, Lizi, Jing Jiang, Ying Ding, Heyan Huang, and Ee-Peng Lim. 2014. Lifetime lexical variation in social media. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1643–1649, Québec City, Québec, Canada.

Ling, Wang, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, Sofia, Bulgaria.

Lui, Marco, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2(1):27–40.

Makatchev, Maxim and Reid Simmons. 2011. Perception of personality and naturalness through dialogues by native speakers of American English and Arabic. In *Proceedings of the SIGDIAL 2011 Conference*, pages 286–293, Portland, Oregon.

Mallinson, Christine, Becky Childs, and Gerard Van Herk, editors. 2013. *Data Collection in Sociolinguistics: Methods and Applications*. Routledge.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Manning, Christopher D. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Martin, James R. and David Rose. 2003. *Working with Discourse: Meaning Beyond the Clause*. Continuum.

Martin, James R. and Peter R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan.

Marwick, Alice E. and danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133.

Mayfield, Elijah, M. Barton Laws, Ira B . Wilson, and Carolyn P. Rosé. 2014. Automating annotation of information-giving for analysis of clinical conversation. *Journal of the American Medical Informatics Association*, 21(1):122–128.

McNamee, Paul. 2005. Language identification: a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges*, 20(3):94–101.

Meyerhoff, Miriam. 2011. *Introducing Sociolinguistics*. Routledge.

Michael, Loizos and Jahna Otterbacher. 2014. Write like I write: Herding in the language of online reviews. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pages 356–365, Ann Arbor, Michigan, USA.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop*.

Milroy, James and Lesley Milroy, 1978. *Belfast: change and variation in an urban*

*vernacular*, pages 19–36.
Milroy, James and Lesley Milroy. 1985. Linguistic change, social network and speaker innovation. *Journal of linguistics*, 21(2):339–384.
Milroy, Lesley and Matthew Gordon. 2003. *Sociolinguistics: Method and Interpretation*. Wiley-Blackwell.
Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the demographics of Twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 554–557, Barcelona, Catalonia, Spain.
Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE*, 8(4):e61981.
Morrow, Raymond A. and David D. Brown, 1994. *Contemporary Social Theory: Critical Theory and Methodology*, chapter Deconstructing the Conventional Discourse of Methodology: Quantitative versus Qualitative Methods. SAGE Publications.
Mubarak, Hamdy and Kareem Darwish. 2014. Using Twitter to collect a multi-dialectal corpus of Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7, Doha, Qatar.
Mukherjee, Arjun and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA.
Myers-Scotton, Carol. 1995. *Social motivations for codeswitching: Evidence from Africa*. Oxford: Clarendon.
Myers-Scotton, Carol. 2002. *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford: Oxford University Press.
Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.
Nerbonne, John and Martijn Wieling. 2015. Statistics for aggregate variationist analyses. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *Handbook of Dialectology*. Boston: Wiley.
Nguyen, Dong and A. Seza Doğruöz. 2013. Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 857–862, Seattle, Washington, USA.
Nguyen, Dong, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How old do you think I am?" A study of language and age in Twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 439–448, Boston, Massachusetts, USA.
Nguyen, Dong and Carolyn P. Rosé. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 76–85, Portland, Oregon.
Nguyen, Dong, Noah A. Smith, and Carolyn P. Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123, Portland, Oregon.
Nguyen, Dong, Dolf Trieschnigg, and Leonie Cornips. 2015. Audience and the use of minority languages on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 666–669, Oxford, United Kingdom.
Nguyen, Dong, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961, Dublin, Ireland.
Nguyen, Dong, Dolf Trieschnigg, and Theo Meder. 2014. Tweetgenie: Development, evaluation, and lessons learned. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 62–66, Dublin, Ireland.
Nguyen, Viet-An, Jordan Boyd-Graber, Philip Resnik, Deborah A. Cai, Jennifer E. Midberry, and Yuanxin Wang. 2014. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95(3):381–421.
Niederhoffer, Kate G. and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*,

21(4):337–360.
Noble, Bill 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.
Nowson, Scott and Jon Oberlander. 2006. The identity of bloggers: Openness and gender in personal weblogs. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 163–167, Palo Alto, California.
Nowson, Scott, Jon Oberlander, and Alastair J. Gill. 2005. Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1666–1671, Stresa, Italy.
Otterbacher, Jahna. 2010. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 369–378, Toronto, ON, Canada.
Paolillo, John C. 2001. Language variation on Internet Relay Chat: A social network approach. *Journal of sociolinguistics*, 5(2):180–213.
Papalexakis, Evangelos E., Dong Nguyen, and A. Seza Doğruöz. 2014. Predicting code-switching in multilingual communication for immigrant communities. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 42–50, Doha, Qatar.
Pavalanathan, Umashanthi and Jacob Eisenstein. 2015a. Audience-modulated variation in online social media. *American Speech*, 90(2):187–213.
Pavalanathan, Umashanthi and Jacob Eisenstein. 2015b. Confounds and consequences in geotagged Twitter data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2138–2148, Lisbon, Portugal.
Pechenick, Eitan Adam, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Characterizing the Google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*, 10(10):1–24.
Peersman, Claudia, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on search and mining user-generated contents (SMUC '11)*, pages 37–44, Glasgow, UK.
Peirsman, Yves, Dirk Geeraerts, and Dirk Speelman. 2010. The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(04):469–491.
Peng, Nanyun, Yiming Wang, and Mark Dredze. 2014. Learning polylingual topic models from code-switched social media documents. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 674–679, Baltimore, Maryland.
Pennacchiotti, Marco and Ana-Maria Popescu. 2011. A machine learning approach to Twitter user classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 281–288, Barcelona, Spain.
Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. Linguistic Inquiry and Word Count: LIWC 2001. *Mahwah, NJ: Lawrence Erlbaum*.
Peterson, Kelly, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the Enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, Portland, Oregon.
Pfeffer, Jeffrey and Gerald R. Salancik. 1978. *The External Control of Organizations: A Resource Dependence Perspective*. New York: Harper & Row.
Piergallini, Mario, Seza A. Doğruöz, Phani Gadde, David Adamson, and Carolyn P. Rosé. 2014. Modeling the use of graffiti style features to signal social relations within a multi-domain learning paradigm. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 107–115, Gothenburg, Sweden.
Piotrowski, Michael. 2012. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*.
Poplack, Shana, David Sankoff, and Christopher Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1):47–104.
Postmes, Tom, Russell Spears, and Martin Lea. 2000. The formation of group norms in computer-mediated communication. *Human Communication Research*, 26(3):341–371.
Prabhakaran, Vinodkumar, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An indicator of power in political

debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1481–1486, Doha, Qatar.

Prabhakaran, Vinodkumar, Ajita John, and Dorée D. Seligmann. 2013. Who had the upper hand? Ranking participants of interactions based on their relative power. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 365–373, Nagoya, Japan.

Prabhakaran, Vinodkumar and Owen Rambow. 2013. Written dialog and social power: Manifestations of different types of power in dialog behavior. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 216–224, Nagoya, Japan.

Prabhakaran, Vinodkumar, Owen Rambow, and Mona Diab. 2012a. Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522, Montréal, Canada.

Prabhakaran, Vinodkumar, Owen Rambow, and Mona Diab. 2012b. Who's (really) the boss? Perception of situational power in written interactions. In *Proceedings of COLING 2012*, pages 2259–2274, Mumbai, India.

Prabhakaran, Vinodkumar, Emily E. Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976, Doha, Qatar.

Prager, John M. 1999. Linguini: Language identification for multilingual documents. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences*, Maui, HI, USA.

Preoţiuc-Pietro, Daniel, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1754–1764, Beijing, China.

Preoţiuc-Pietro, Daniel, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PLoS ONE*, 10(9).

Prokić, Jelena, Çağrı Çöltekin, and John Nerbonne. 2012. Detecting shibboleths. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 72–80, Avignon, France.

Quercia, Daniele, Jonathan Ellis, Licia Capra, and Jon Crowcroft. 2011. In the mood for being influential on Twitter. In *Proceedings of the 3rd IEEE International Conference on Social Computing*, pages 307 – 314, Boston, MA.

Rabe-Hesketh, Sophia and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata*. Stata Press.

Rabe-Hesketh, Sophia, Anders Skrondal, and Andrew Pickles. 2004. GLLAMM Manual. *U.C. Berkeley Division of Biostatistics Working Paper Series, Paper 160*.

Raghavan, Sindhu, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 38–42, Uppsala, Sweden.

Rangel, Francisco, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the 2nd author profiling task at PAN 2014. In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*.

Rangel, Francisco, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at PAN 2013. *Notebook Papers of CLEF*.

Rao, Delip, Michael J. Paul, Clayton Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 598–601, Barcelona, Spain.

Rao, Delip, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd international workshop on search and mining user-generated contents (SMUC '10)*, pages 37–44, Toronto , ON, Canada.

Ribeiro, Branca Telles. 2006. Footing, positioning, voice. are we talking about the same things? In Anna De Fina, Deborah Schiffrin, and Michael Bamberg, editors, *Discourse and Identity*. Cambridge University Press, pages 48–82.

Richards, Keith. 2006. *Language and Professional Identity: Aspects of Collaborative*

*Interaction*. Palgrave Macmillan.
Romaine, Suzanne. 1995. Bilingualism (2nd edition). *Malden, MA: Blackwell Publishers*.
Rosé, Carolyn P., in press. *International Handbook of the Learning Sciences*, chapter Learning analytics in the Learning Sciences. Taylor & Francis.
Rosé, Carolyn P. and Alla Tovares, 2015. *Socializing Intelligence Through Academic Talk and Dialogue*, chapter What sociolinguistics and machine learning have to say to one another about interaction analysis. Washington, DC: American Educational Research Association.
Rosé, Carolyn P., Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3(3):237–271.
Rosenthal, Sara and Kathleen McKeown. 2011. Age prediction in blogs: a study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 763–772, Portland, Oregon.
Sankoff, Gillian, 2006. *Encyclopedia of Language and Linguistics*, chapter Age: Apparent time and real time. Amsterdam: Elsevier.
Sap, Maarten, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and Andrew Hansen Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151, Doha, Qatar.
Sarawgi, Ruchita, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: Tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 78–86, Portland, Oregon, USA.
Schegloff, Emanuel A. 2007. *Sequence Organization in Interaction: A Primer in Conversation Analysis*, volume 1. Cambridge University Press.
Scherrer, Yves and Owen Rambow. 2010. Word-based dialect identification with georeferenced rules. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1151–1161, Cambridge, MA.
Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pages 199–205, Menlo Park, California.
Schneider, Gerold, James Dowdall, and Fabio Rinaldi. 2004. A robust and hybrid deep-linguistic theory applied to large-scale parsing. In *Proceedings of the 3rd workshop on RObust Methods in Analysis of Natural Language Data (ROMAND 2004)*, pages 14–23, Geneva, Switzerland.
Schuller, Björn, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2010. The INTERSPEECH 2010 paralinguistic challenge. In *INTERSPEECH*, pages 2794–2797, Makuhari, Chiba, Japan.
Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791.
Scissors, Lauren E., Alastair J. Gill, Kathleen Geraghty, and Darren Gergle. 2009. In CMC we trust: The role of similarity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*, pages 527–536, Boston, MA, USA.
Searle, John R. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.
Singh, Sameer. 2001. A pilot study on gender differences in conversational speech on lexical richness measures. *Literary and Linguistic Computing*, 16(3):251–264.
Sloan, Luke, Jeffrey Morgan, Pete Burnap, and Matthew Williams. 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLoS ONE*, 10(3).
Snow, David A. and Leon Anderson. 1987. Identity work among the homeless: The verbal construction and avowal of personal identities. *American Journal of Sociology*, 92(6):1336–1371.
Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap

and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii.

Soliz, Jordan and Howard Giles. 2014. Relational and identity processes in communication: A contextual and meta-analytical review of Communication Accommodation Theory. In Elisia L. Cohen, editor, *Communication Yearbook 38*. Routledge.

Solorio, Thamar, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar.

Solorio, Thamar and Yang Liu. 2008a. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981, Honolulu, Hawaii.

Solorio, Thamar and Yang Liu. 2008b. Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060, Honolulu, Hawaii.

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

Stoop, Wessel and Antal van den Bosch. 2014. Using idiolects and sociolects to improve word prediction. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 318–327, Gothenburg, Sweden.

Strzalkowski, Tomek, Samira Shaikh, Ting Liu, George Aaron Broadwell, Jenny Stromer-Galley, Sarah Taylor, Umit Boz, Veena Ravishankar, and Xiaoai Ren. 2012. Modeling leadership and influence in multi-party online discourse. In *Proceedings of COLING 2012*, pages 2535–2552, Mumbai, India.

Swayamdipta, Swabha and Owen Rambow. 2012. The pursuit of power and its manifestation in written dialog. In *Proceedings of 2012 IEEE Sixth International Conference on Semantic Computing (ICSC)*, pages 22–29, Palermo, Italy.

Taboada, Maite and William C. Mann. 2006. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567–588.

Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge University Press.

Tam, Jenny and Craig H. Martell. 2009. Age detection in chat. In *ICSC '09. IEEE International Conference on Semantic Computing*, pages 33–39, Berkeley, CA.

Tannen, Deborah. 1990. *You just don't understand: Women and men in conversation*. Ballantine Books.

Tannen, Deborah. 1993. *Framing in Discourse*. Oxford University Press.

Thomason, Sarah G. 2001. *Language contact: an introduction*. Edinburgh: Edinburgh University Press.

Trieschnigg, Dolf, Djoerd Hiemstra, Mariët Theune, Franciska Jong, and Theo Meder. 2012. An exploration of language identification techniques for the Dutch folktale database. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage*, pages 47–51, Istanbul, Turkey.

Trudgill, Peter. 1974. *The social differentiation of English in Norwich*. Cambridge: Cambridge University Press.

Trudgill, Peter. 2003. *The Norfolk Dialect*. Norfolk Origins 7. Poppyland Publishing.

Truong, Khiet P., Gerben J. Westerhof, Sanne M. A. Lamers, and Franciska de Jong. 2014. Towards modeling expressed emotions in oral history interviews: Using verbal and nonverbal signals to track personal narratives. *Literary and Linguistic Computing*, 29(4):621–636.

Tsaliki, Liza. 2003. Globalization and hybridity: The construction of Greekness on the Internet. In Karim Haiderali Karim, editor, *The Media of Diaspora*. Routledge.

Van Durme, Benjamin. 2012. Streaming analysis of discourse participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 48–58, Jeju Island, Korea.

Volkova, Svitlana, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA.

Voss, Clare, Stephen Tratz, Jamal Laoudi, and Douglas Briesch. 2014. Finding romanized Arabic dialect in code-mixed tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2249–2253, Reykjavik, Iceland.

Vyas, Yogarshi, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979, Doha, Qatar.

Wagner, Suzanne E. 2012. Age grading in sociolinguistic theory. *Language and Linguistics Compass*, 6(6):371–382.

Wang, Yafei, David Reitter, and John Yen. 2014. Linguistic adaptation in conversation threads: Analyzing alignment in online health communities. In *Proceedings of the 2014 ACL Workshop on Cognitive Modeling and Computational Linguistics*, pages 55–62, Baltimore, Maryland, USA.

Wardhaugh, Ronald. 2011. *An Introduction to Sociolinguistics*. Wiley-Blackwell.

Wei, Li. 1998. The 'why' and 'how' questions in the analysis of conversational codeswitching. In Peter Auer, editor, *Codeswitching in conversation: Language, interaction and identity*. London: Routledge, pages 156–176.

Weinreich, Uriel. 1953. Languages in contact. findings and problems. *New York, Linguistic Circle of New York*.

Weinreich, Uriel, William Labov, and Marvin I. Herzog. 1968. Empirical foundations for a theory of language change. In Winfred P. Lehmann and Yakov Malkiel, editors, *Directions for Historical Linguistics: A Symposium*. Austin: University of Texas Press, pages 95–188.

Wen, Miaomiao, Diyi Yang, and Carolyn P. Rosé. 2014a. Linguistic reflections of student engagement in massive open online courses. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pages 525–534, Ann Arbor, Michigan, USA.

Wen, Miaomiao, Diyi Yang, and Carolyn P. Rosé. 2014b. Sentiment analysis in MOOC discussion forums: What does it tell us? In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 130–137, London, UK.

West, Robert, Hristo S. Paskov, Jure Leskovec, and Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association of Computational Linguistics – Volume 2, Issue 1*, pages 297–310.

Wieling, Martijn, Jelke Bloem, Kaitlin Mignella, Mona Timmermeister, and John Nerbonne. 2014. Measuring foreign accent strength in English. *Language Dynamics and Change*, 4(2):253–269.

Wieling, Martijn and John Nerbonne. 2010. Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features. In *Proceedings of TextGraphs-5 - 2010 Workshop on Graph-based Methods for Natural Language Processing*, pages 33–41, Uppsala, Sweden.

Wieling, Martijn and John Nerbonne. 2015. Advances in dialectometry. *Annual Review of Linguistics*, 1(1):243–264.

Wiersma, Wybo, John Nerbonne, and Timo Lauttamus. 2010. Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing*, 26(1):107–124.

Wing, Benjamin P. and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 955–964, Portland, Oregon, USA.

Wintner, Shuly. 2002. Formal language theory for natural language processing. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 71–76, Philadelphia, PA.

Yamaguchi, Hiroshi and Kumiko Tanaka-Ishii. 2012. Text segmentation by language using minimum description length. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 969–978, Jeju Island, Korea.

Yan, Xiang and Ling Yan. 2006. Gender classification of weblog authors. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 228–230, Palo Alto, California.

Yang, Diyi, Miaomiao Wen, and Carolyn P. Rosé. 2015. Weakly supervised role identification in teamwork interactions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1671–1680, Beijing, China.

Zaidan, Omar F. and Chris Callison-Burch. 2013. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Zamal, Faiyaz Al, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, pages 387–390, Dublin, Ireland.

Zampieri, Marcos, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland.

Zampieri, Marcos, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria.

Zhang, Jian, Zoubin Ghahramani, and Yiming Yang. 2008. Flexible latent variable models for multi-task learning. *Machine Learning*, 73(3):221–242.

Zijlstra, Hanna, Henriët van Middendorp, Tanja van Meerveld, and Rinie Geenen. 2005. Validiteit van de Nederlandse versie van de Linguistic Inquiry and Word Count (LIWC). *Netherlands Journal of Psychology*, 60(3):55–63.