TEXT AS SOCIAL AND CULTURAL DATA

A Computational Perspective on Variation in Text

Dong Nguyen

Text as Social and Cultural Data:
A Computational Perspective on Variation in Text

Dong Nguyen

Graduation committee:

Chairman:	Prof. dr. P.M.G. Apers
Promotors:	Prof. dr. F.M.G. de Jong
	Prof. dr. A.P.J. van den Bosch
Co-promotor:	Dr. M. Theune

Members:

dr. J. Eisenstein	Georgia Institute of Technology
Prof. dr. D.K.J. Heylen	University of Twente
Prof. dr. T. Meder	Meertens Institute
Prof. dr. ir. J. Nerbonne	University of Groningen
Prof. dr. A. Søgaard	University of Copenhagen
Prof. dr. ir. B.P. Veldkamp	University of Twente

CTIT Ph.D. Thesis Series No. 17-421
Centre for Telematics and Information Technology
University of Twente, The Netherlands
P.O. Box 217, 7500 AE Enschede

SIKS Dissertation Series No. 2017-09
The research reported in this dissertation has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

The research reported in this dissertation has been carried out within the Folktales as Classifiable Texts (FACT) project, part of the CATCH programme funded by NWO (grant number 640.005.002).

The research reported in this dissertation has been carried out at the Human Media Interaction group of the University of Twente.

ISBN: 978–90–365–4300–2
ISSN: 1381-3617 (CTIT Ph.D. Thesis Series No. 17-421)
Available online at https://doi.org/10.3990/1.9789036543002

Typeset with LaTeX. Printed by Ipskamp Printing Enschede.
Cover design by Annelien Dam.

TEXT AS SOCIAL AND CULTURAL DATA: A COMPUTATIONAL PERSPECTIVE ON VARIATION IN TEXT

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. T.T.M. Palstra,
on account of the decision of the graduation committee,
to be publicly defended
on Friday March 10, 2017 at 16:45.

by

Dong-Phuong Nguyen

born on March 23, 1987
in Nieuwegein, The Netherlands

This dissertation has been approved by:

Prof. dr. F.M.G. de Jong (promotor)
Prof. dr. A.P.J. van den Bosch (promotor)
Dr. M. Theune (co-promotor)

Acknowledgements

This dissertation marks the end of my PhD. Although I started my PhD in 2012, this dissertation is a result of a journey that started much earlier. Eight years ago I was searching for a Bachelor's project and got in touch with researchers at the Human Media Interaction group of the University of Twente. The Bachelor's project was my first encounter with language technologies, resulted in my first academic publication, my first conference trip, but maybe most importantly, it sparked my interest in academic research. I then moved to the States to pursue a Masters's degree at Carnegie Mellon University, a place that had a profound influence on me both academically and personally. At CMU, my interest in social media research was raised and I published my first papers in, what I would now call, 'computational sociolinguistics'. Little did I know back then, that this would have such a big influence on the topic of my dissertation. In 2012 I returned to the Netherlands. As part of a Dutch national project, I started as a PhD student among many familiar faces from my Bachelor's project. I thoroughly enjoyed my PhD and I am deeply grateful to have been able to work with and learn from so many people. I would like to thank:

. . . my advisors
I would like to express my deepest appreciation to Franciska and Mariët, my advisors in Twente, who both have been incredibly supportive throughout my PhD. Franciska gave me the freedom to pursue my research interests and always had wise advice to offer about academia. Without Mariët's keen eye for detail, many of my papers would have looked differently. I also very much appreciate that she always made time to help out with various PhD-related matters. I thank Antal for his invaluable feedback on drafts of this thesis.

. . . my PhD dissertation committee
I would like to extend my sincere gratitude to the members of my PhD dissertation committee for reviewing this dissertation.

. . . the FACT team
I thank Dolf for the enjoyable collaborations, for making the train journeys to Amsterdam more fun, and for the cup of teas (and also coffee later on) across the road. I thank Iwe for making sure I could run my experiments. Djoerd has been incredibly supportive and I enjoyed his always-present enthusiasm. I thank Theo for introducing me to the wonderful world of folktales and Marianne for the pleasant conversations.

. . . my collaborators from across the ocean
From Jacob I learned more about machine learning and I enjoyed our conversations about the field of computational sociolinguistics. Furthermore, I thank him for hosting me at Georgia Tech and for his tips on biking in Atlanta. During my Master's, Carolyn introduced me to sociolinguistics and strengthened my interest in social media. She stimulated me to think about the bigger story when writing papers and made valuable contributions to the survey article.

. . . my collaborators from another discipline
I learned lots from engaging with researchers outside of computer science during my PhD. Seza sparked my interest in multilingualism and more specifically code-switching and I enjoyed the fun conversations we had about academia and life. Leonie explained many sociolinguistic concepts to me patiently and made me look critically at the limitations of computational approaches. Tijs gave my research a push towards truly 'computational social science' by bringing social science theories into my research.

. . . my collaborators from TREC Fedweb, TINPOT, Twidentity and the Twitter Data Grant
Besides the FACT project, I also had the opportunity to be involved in several other projects. With Djoerd, Thomas, Dolf, and Adam, I organized the TREC Fedweb track. I have fond memories of our brainstorming session in a farm in the middle of nowhere, and of our reunion last year in London. The TINPOT project turned out to have a big influence on my research, with the TweetGenie demo being one of the highlights. The follow-up project, Twidentity, was fun as well. Anna, Dolf, Jolie, Leonie, Lysbeth, Rilana, and Theo, thanks! I learned a lot from interacting with social scientists in the Twitter Data Grant project. I thank Tijs, Anna, Michel, Ariana, Djoerd and Nugroho for the pleasant collaborations.

. . . my internship hosts
I thank Gabriella and Milad for hosting me at Microsoft Research. Thanks to the internship I gained experience in collecting data using crowdsourcing, which proved useful in two of the studies in this dissertation. I could have not wished for a better mentor at Google. Antonio made sure I could work on interesting projects and had valuable advice about programming.

. . . my colleagues at the University of Twente
I would like to thank my (ex-)office mates (Alejandro, Danish, Jelte, Merijn and Robby) for being very supportive throughout my PhD. During my PhD I was part of the Human Media Interaction group, an incredibly welcoming group. I thank Meiru for the fun times outside the office, and Khiet for the pleasant conversations and being a fellow-supervisor on several student projects. I would also like to thank the members of the Databases group, in particular Djoerd, Robin and Zhemin for insightful conversations on machine learning and information retrieval and Jan for the technical support.

. . . my colleagues at the Meertens Institute
During my PhD I spent about one day a week at the Meertens Institute in Amsterdam. Being around researchers who study language and culture stimulated me to look at research questions from different perspectives. Various variationist linguists from the Meertens Institute have at some point given me feedback or directed me to the right resources. I thank Folgert for insightful conversations on folk narratives.

. . . data contributors
I thank the Crowdflower workers for contributing to this thesis with their annotations and the visitors of the TweetGenie demo for trying out the demo and providing candid feedback.

. . . my academic friends
Thanks to Uma I quickly felt at home in Atlanta. With Julia, I was able to share many of the lows and highs of a PhD. I have fond memories of post-conference trips with Katja in Hawaii and China. I also thank her for helping me to get settled in Cambridge. Sofia and Dongwook made Cambridge lots of fun.

. . . my friends, partner and family
Most of all, I would like to thank my friends, partner and family. A special thanks to the ones who took care of my horse during my trips abroad. This thesis is dedicated to my parents, who have been extremely supportive throughout my studies, while reminding me to enjoy life. *Dank jullie wel, cám ơn*.

Dong Nguyen, London, January 2017.

Contents

1 Introduction

1.1 Text as Social and Cultural Data

The explosion of digital data has transformed the world. People create content through social media sites, track their health and movements through mobile apps, generate data by searching, browsing and clicking online, and so on. The increasing availability of these massive datasets – the rise of so-called **big data** – has transformed industry and policy making [Manyika et al., 2011, McAfee and Brynjolfsson, 2012]. Furthermore, it has led to a paradigm shift in science. In addition to the traditional focus on the description of natural phenomena, theory development and computational science (e.g., through simulations), data-driven exploration and discovery are becoming increasingly important in various scientific disciplines [Hey et al., 2009].

This dissertation focuses on two types of (big) data: **social** and **cultural data**. Within the social sciences and the humanities the potential of massive datasets, such as social media data and cultural heritage collections, to study social and cultural phenomena is increasingly being recognized [Golder and Macy, 2014]. Not only have there been significant efforts in increasing digitization and developing infrastructures to handle the larger datasets, but, as boyd and Crawford [2012, p. 665] argue, these large amounts of data have created "*a radical shift in how we think about research*". Big data has impacted the research process, raised new ethical questions, has given rise to radically new research directions and even new research fields. In line with these developments, the field of **computational social science** is emerging, in which computational approaches are used to analyze large datasets for social science research [Lazer et al., 2009]. Similarly, recently terms like **computational humanities**[1] and **cultural analytics** [Manovich, 2007, 2016] have been used to refer to the use of computational methods and large datasets for the study of human culture.

Texts are usually written *by* and *for* people and texts can thus be used to study all kinds of social and cultural phenomena. Texts often reflect the ideas, values and

[1] See for example the titles of events such as 'Computational Humanities - bridging the gap between Computer Science and Digital Humanities' (Dagstuhl Seminar 14301) and research programmes such as KNAW's Computational Humanities programme (2011-2016) [F. Willekens et al., 2010], and the following blogpost: http://lab.softwarestudies.com/2012/03/computational-humanities-vs-digital.html.

beliefs of their authors and target audiences. Texts also describe actions and events occurring over time. The increasing recognition of **text as social and cultural data** in computationally driven research is reflected in the increasing number of workshops and conferences that focus on this topic[2].

Textual data has always been a resource for studying social and cultural phenomena. Approaches such as discourse analysis and (qualitative and quantitative) content analysis are frequently used in both the humanities and the social sciences [Holsti, 1969, Johnstone, 2007]. With content analysis, text is broken down into units (e.g., sentences or phrases) and the units are coded according to a coding scheme described in a codebook. However, because the coding is typically done manually, this step is often time-consuming. Therefore, the use of computational methods has the potential to scale up analyses to larger datasets. For example, Bravo and Hoffman-Goetz [2016] conducted a content analysis of 4,222 Canadian tweets posted during the Movember campaign (a health campaign) in 2013. They manually coded the topics of the tweets and whether tweets were health or non-health related. Building on this work, Dwi Prasetyo et al. [2015] expanded the scope by considering more countries and a larger number of tweets. Considerably scaling up the analyses, over 406k tweets were automatically categorized according to whether they focused on health topics or the social aspects of the campaign using a machine learning classifier.

There are many more examples of large-scale text analysis studies that focus on social and cultural phenomena. Text analysis of online data has helped social scientists to study questions such as "*why do some health campaign participants raise more money than others?*" [Nguyen et al., 2015b]. Twitter has been used to study questions such as "*how do rumors and beliefs circulate among people?*" [Meder et al., 2015], "*what does language use tell us about the identity of speakers?*" [Nguyen et al., 2013a], and "*can TV ratings be predicted based on tweets?*" [Sommerdijk et al., 2016], and Wikipedia pages to study power dynamics [Danescu-Niculescu-Mizil et al., 2012].

Social media in particular has shown to be a rich resource to study social and cultural phenomena. Compared to data sources such as newswire texts that have been frequently used in computational linguistics (CL), and data collected using observations, interviews and surveys in the social sciences and the humanities, social media offers the following advantages: it is (i) large-scale, longitudinal data; (ii) rich contextual data, such as social network information; it offers (iii) the opportunity to study language use and human behavior in a multitude of social situations; and, perhaps the most valuable, it is a (iv) means to overcome much of the *observer's paradox*. This term, coined by Labov [1972], refers to the paradox of the need to observe a phenomenon as it would have been if it was not being observed. Tangherlini [2016, p. 6] describes the value of online data for overcoming this problem with "*fieldwork can now be carried out on and among (as opposed to with) groups and individuals who are not necessarily aware they are participating in an ethnographic project*". Social media thus provides a rich resource to study social and cultural phenomena.

[2]For example, workshops such as '*NLP and computational social science*' (EMNLP 2016, WebSci 2016), '*Computational approaches to code switching*' (EMNLP 2014, EMNLP 2016), '*Language Technology for Cultural Heritage, Social Sciences, and Humanities*' (LaTeCH at ACL 2016, now in its tenth year) and the '*ACL joint workshop on social dynamics and personal attributes*' (ACL 2014), and conferences such as *ACL*, *EMNLP*, *ICWSM*, *SocInfo* and the *New Directions in Analyzing Text as Data Conference*.

1

The research in this dissertation fits into two emerging areas in which questions about social and cultural phenomena are studied with computational means and digital texts: **computational sociolinguistics** and **computational folkloristics**. Language is one of the main instruments by which people construct their identity and manage their social network. The study of the social role of language has received much attention in the field of sociolinguistics, which focuses on the reciprocal influence of society and language. However, within the field of computational linguistics the social role of language has traditionally not received much attention. With the rise of social media and the increasing interest in using text to study social phenomena, the area of computational sociolinguistics (see also Chapter 2) is emerging, which uses computational approaches to study the relation between language and society. Another emerging area is computational folkloristics [Abello et al., 2012, Tangherlini, 2016], in which large datasets and computational approaches are leveraged to study folklore (e.g., songs, urban legends, clothing, dance, etc., that are transmitted through communication and behavioral example[3]). Efforts have ranged from digitization of resources to the design of methods for computational analyses, visualizations, and pattern extraction in the datasets. For example, automatic detection of similar folk narratives (e.g., urban legends and fairy tales) can be used to study how these narratives develop over time (see Chapter 3 for more background on this area).

Large datasets require the use of computational methods to analyze and process the data, forcing researchers to rethink basic concepts and tools from the social sciences and the humanities that were used for smaller datasets [Manovich, 2016, Tangherlini, 2016] and thus stimulating interaction between computer scientists (who usually develop these methods) and researchers from the social sciences and the humanities (who use and interpret the output of these methods). Furthermore, when analyzing big cultural and social datasets, aspects from the humanities and the social sciences are often both relevant [Manovich, 2016]. Thus, a trend can be observed towards computational and data-driven analyses in which multiple disciplines converge to study social and cultural phenomena.

The fields of computational sociolinguistics and computational folkloristics are inherently *interdisciplinary*[4]. As argued by Nissani [1997], interdisciplinary research has the potential to lead to creative breakthroughs and to prevent cross-disciplinary oversights and disciplinary cracks (e.g., neglecting important research problems that do not fall within disciplinary boundaries). However, interdisciplinary research also introduces challenges, due to differences in terminology, data collection methods, validation methods, etc. Thus, besides understanding the involved disciplines, interdisciplinary research also requires understanding how to connect them [Karlqvist, 1999]. In my view, these emerging research areas should not be seen as a replacement of the existing research areas. Rather, research in these areas should be considered as *complementing* the more traditional research methods and data collection methods that are used within sociolinguistics and folkloristics.

[3]Different definitions for the term folklore exist, see http://www.afsnet.org/?page=WhatIsFolklore.

[4]There is no agreement on the exact definition of the term *interdisciplinarity*, with various sources (e.g., Institute of Medicine and National Academy of Sciences and National Academy of Engineering [2004], Nissani [1997] and Aboelela et al. [2007]) providing slightly different definitions. Which of these definitions is adopted does not influence the argumentation in this dissertation.

1.2 Variation in Text

The theme of this dissertation and a common theme in both computational sociolinguistics and computational folkloristics research is **variation in text**. While variation in itself is a broad term, in this dissertation, variation in text refers to the phenomenon that *the same can be said in different ways*. From the perspective of folkloristics, or more specifically folk narrative research, this refers to telling the same story in different ways. From the perspective of sociolinguistics, this refers to variation in language (e.g., language choice, word choice, grammar, etc.). Variation can be random, but variation can also be a result of conscious choices to achieve a certain goal and such variation often exhibits structural patterns. In this section the perspectives of sociolinguistics and folk narrative research on variation are described in more detail.

This dissertation first considers variation from the perspective of sociolinguistics. In social media, language use tends to be informal and variation in language use is therefore abundant. For example, orthographic variations of *cool* are *coooool* (alphabetical lengthening) and *kewl*. These variations are also sometimes combined with intensifiers like *hella*, resulting in variations such as *hellakewl* and *hellacool*. As another example of variation, social media users may use multiple language varieties in their social media messages (e.g., English, Dutch and a Dutch dialect) and the social context (e.g., audience, goal of the message) often influences which language variety is selected. Language in social media is often being referred to as 'noisy', because its informal nature makes it more challenging to be processed by various NLP tools than, say, newswire texts. However, much of this variation exhibits regular patterns and carries social meaning [Eisenstein, 2013a]. This kind of variation plays a central role in linguistic change and is studied in the areas of sociolinguistics and computational sociolinguistics. Thus, the variation should be seen as part of the signal rather than noise, and modeling and understanding variation in language is therefore key towards more refined analyses of social phenomena as reflected in language use.

Variation is also an inherent feature of folk narrative data. Much of this data comes from historical sources, and for example orthographic variations occur frequently in historical texts. Being able to handle such variation is therefore important for processing and analyzing historical texts [Piotrowski, 2012]. However, variation in text can also be considered at a more abstract level when analyzing the structure and content of folk narratives from the perspective of folkloristics. Different variants of a story appear over time due to oral and written transmission. When asking people to recall a specific story, everyone tells his or her own version. Sometimes details may be left out simply because of recall problems. Other types of modifications are more profound, including adding details, exchanging character roles, specialization (e.g., a bird becomes a specific bird, like a sparrow, or a car gets a specific brand), adding repetition of events, and so on [Thompson, 1951]. Such modifications are often motivated by social reasons. For example, when fairy tales became increasingly popular in the 19th century, cruelties in these tales (e.g., the cannibalistic mother in the tale of Snow White) were often removed or softened to make them suitable for children. In adaptations intended for more adult audiences, often erotic and horrifying elements are introduced or emphasized [Joosen, 2012]. As in sociolinguistics, variation in folk

narratives often leads to change over time. Understanding the variation within folktale data could therefore shed more light on how narratives develop over time and geographically. For example, Karsdorp and Van den Bosch [2016] analyzed a corpus of 427 Dutch literary Little Red Riding Hood retellings to study the processes through which stories are retold. The increasing digitization of folk narratives enables the use of computational approaches to analyze variation in folk narratives and automatically enrich narratives with metadata to support information access.

1.3 Thesis Statement

While researchers in the humanities and the social sciences have long recognized that variation in text can reveal social and cultural patterns, such variation has traditionally not been considered in the development of computational approaches to processing and analyzing text. With the increasing recognition of *text as social and cultural data* and the emergence of areas such as *computational social science* and *computational sociolinguistics*, it is important to place a larger focus on the study of variation in text within computational frameworks. Besides potentially leading to new insights into social and cultural phenomena, computational modeling of variation in text could also lead to more effective text processing tools.

Research focusing on text analysis to study cultural and social phenomena is inherently interdisciplinary, drawing from different research disciplines, such as computational linguistics, information retrieval, machine learning, statistics, anthropology, linguistics, etc. While these disciplines make increasing use of the same data sources (e.g., social media data) and study similar research questions, so far the interaction between researchers from these different disciplines has been limited. Yet, the biggest progress may be made when these disciplines join forces and, thus, I advocate for more interaction between computer scientists and researchers from the social sciences and humanities. The work described in this dissertation benefited from interdisciplinary collaborations with researchers from various disciplines.

The interdisciplinary character of this dissertation is reflected in the goals of the studies presented. Computer science studies often focus on *prediction*. The term 'prediction' is often used interchangeably with terms such as 'forecasting'. In this dissertation, prediction refers to the use of data points from a sample to predict the values of other data points[5]. *Forecasting* involves making predictions into the future. The performance of a certain prediction model is usually estimated with a quantitative metric. In contrast, social science and humanities research often focus on *explanation*, e.g., obtaining new insights into a social or cultural phenomenon, hypothesis testing, theory development, etc. As a consequence, aspects such as *interpretability* of models have traditionally been valued differently in the different research communities and a further reflection on this issue is presented in Subsection 12.2.3. Thus, although computer science studies often focus on performance on specific tasks, the goals of the studies presented in this dissertation were often two-fold: not only doing 'well' on a certain task, but also generating new insights into the data and the social or cultural phenomenon that was studied.

[5] See a blogpost by Prof. Galit Shmueli: http://www.bzst.com/2011/09/predict-or-forecast.html.

1.4 Research Questions

This section presents the research questions that are addressed in this dissertation. The research questions are grouped according to the two main research themes: computational sociolinguistics and computational folkloristics. The chapters that address the research questions are indicated in parentheses.

Research Theme: Computational Sociolinguistics (Part II). Language is a social phenomenon and variation is inherent to its social nature. Speakers use language as a resource to construct their social identities. They may choose to use certain words, phrases, style elements, etc. to represent themselves in a certain way, thus giving rise to variation in language use. Within sociolinguistics, variation according to gender, age and location have been well studied (see Chapter 2). However, it is only since the rise of social media that computational linguists have become interested in this kind of variation. Building on the insight that language use can sometimes reveal aspects of an author's identity, computational approaches have been explored to predict such aspects based on the language use of the authors (a task often referred to as latent attribute prediction). The first two research questions focus on predicting the gender and age of authors, and more specifically Twitter users, based on their language use.

> **RQ1**. *To what extent can the age of Twitter users be predicted based on their tweets?* (Chapter 4)

Being able to automatically predict the age of authors based on their language use has many practical applications, such as more fine-grained analyses of social phenomena in social media, or personalized advertisements. It may also generate new insights into the relation between age and language use. So far research on automatically predicting the age of authors has been sparse (see also Subsection 2.3.3). For example, how age should be operationalized in prediction studies (e.g., as a categorical variable, a continuous variable, or based on life stages?) has not been explored yet. This study uses Twitter data to analyze the relation between language use and age. To what extent is it possible to predict the age of Twitter users based on their language use? Does the accuracy of the model depend on the age of Twitter users? And, what characterizes the language use of younger and older Twitter users?

> **RQ2**. *What are limitations and consequences of the typical operationalizations of gender and age in latent attribute studies?* (Chapter 5)

Although early studies considered gender and age as static, biological variables, they are increasingly considered social, fluid variables within the social sciences and the humanities. The way in which variables such as age and gender are operationalized has far reaching consequences, ranging from data collection to interpretation of the results and expectations regarding the performance that prediction models can attain on these tasks. We therefore explore consequences of these operationalizations for the task of gender and age prediction, using a novel way of data collection based on crowdsourcing.

The previous research questions focused on prediction tasks. However, as discussed earlier, studies in the social sciences and the humanities often focus on explanation instead. The next research question therefore focuses on methods for analysis of linguistic variation using computational approaches. More specifically, the next research question is about variation according to the location of the speakers.

> **RQ3**. *What is a suitable method to test for geographical language variation?* (Chapter 6)

Identifying linguistic variables (corresponding to the sets of variants which mean the same thing) that exhibit geographical variation is an essential step in many studies on regional dialects. For example, two different ways to refer to *french fries* in the Netherlands (*patat* versus *friet*) may be studied. Furthermore, automatically identifying such variables could potentially help in tasks such as predicting the location of social media users. Until recently, the selection of such variables was mostly done manually. While various statistical methods exist to test for geographical variation, it is not clear whether these methods are suitable for the domain of linguistic variation. For example, some of these methods may make assumptions that do not hold (e.g., a linear relationship between geographical and linguistic distance) or are only suitable for certain types of data (e.g., frequency data).

Language may vary according to the location of the speakers, and speakers from different regions may employ different language varieties. As speakers move, language varieties may come into contact and evolve under each other's influence. Most speakers are multilingual (e.g., in the Netherlands someone may speak Dutch, English and a minority language) and as a consequence, multiple language varieties may be used in a single conversation. Which language variety is chosen, depends on various factors, including social factors such as the conversation partner and the audience. Multilingual communication has been well studied within linguistics. However, the study of online multilingual communication, and in particular using computational approaches, has been little explored so far (see also Section 2.5). The following research questions therefore focus on analyzing variation in online multilingual communication.

> **RQ4**. *How can automatic language identification be performed at the word level?* (Chapter 7)

Multilingual people often employ multiple languages within a single conversation or document. Texts, therefore, may also contain multiple language varieties, but NLP tools are usually designed for texts written in a single language. The following is an example from an online forum for Turkish(TR)-Dutch(NL) speakers: "*<TR>agazina saglik,</TR><NL>ben helemaal met je mee eens</NL>*" ("nicely said, I totally agree with you"). Automatic language identification can help process such texts, but so far automatic language identification has mainly focused on identification at the level of documents. To facilitate the processing of texts with multiple language varieties and to enable studying language choice patterns on a larger scale, this dissertation explores different methods to automatically identify languages at the word level.

RQ5. *How does the target audience influence the language choice of social media users?* (Chapter 8)

Contextual factors, such as the audience, influence the language use of speakers. For example, when speaking with her boss, a speaker may use the standard language (e.g., Dutch), but while being at home with family she might use a minority language. However, studies that analyze the influence of audiences on language choices in multilingual communication are usually confined to small datasets. Building on the theory of audience design [Bell, 1984] and automatic language identification, the influence of audiences on whether Twitter users in the Netherlands use a minority language or Dutch is analyzed.

Research Theme: Computational Folkloristics (Part III). In this theme we focus on variation in folk narrative data. Variations of stories appear as stories are transmitted across time and space. Folklorists have developed categorization schemes based on the concept of tale types, which group similar stories together. For example, different variants of Little Red Riding Hood are grouped into one tale type. The studies presented in this theme therefore focus on similarity between narratives (see also Chapter 3 for more background). Like for the previous theme, we will consider both prediction and analysis. Building on the concept of tale types and the developed categorization schemes (i.e., tale type indexes), we explore the following two research questions:

RQ6. *Can the tale types of folk narratives be automatically predicted?* (Chapter 9)

Tale types are frequently used in folklore research to organize and analyze stories, however manually assigning them to stories is time consuming and is a bottleneck in the digitization of folk narrative collections. (Semi-)Automatically assigning tale types to folk narratives facilitates and speeds up the digital curation step. While tale types and the corresponding tale type indexes are well known and frequently used in the folkloristics community, critics have pointed out several limitations regarding these categorizations, as explained in Chapter 3. Automating the categorization process is also a way to investigate these criticisms by analyzing the robustness and consistency of the categorizations.

RQ7. *How is folk narrative similarity perceived by experts and non-experts?* (Chapter 10)

While the previous research question started from the assumption that tale types are appropriate to group similar folk narratives, this research question revisits the concept of tale types by studying how non-experts perceive folk narrative similarity. For example, do non-experts indeed assign narratives from the same tale type a higher similarity? Which aspects do non-experts consider when judging the similarity between folk narratives? And how does this differ from how experts make their judgement? A better understanding of folk narrative similarity could guide the development of more suitable similarity metrics.

1.5 Scientific Methodology

As a consequence of the novel research design taken, two aspects of the methodological framework adopted deserve special attention: evaluation and interdisciplinarity.

Evaluation. This dissertation builds on and contributes to natural language processing and information retrieval approaches for studying social and cultural phenomena through large-scale text analysis. Research in natural language processing and information retrieval is often evaluated using well-known evaluation metrics and benchmark datasets. However, for many of the topics in this dissertation, no suitable existing datasets were available. For example, for language identification at the word level, a new dataset was released containing annotations of posts in a Dutch-Turkish online community[6] (Chapter 7). As another example, a substantive annotation effort was carried out to collect data with gender and age information of Dutch Twitter users (Chapter 4). In case no empirical datasets with ground truth data could be collected, synthetic data was generated to evaluate the approaches (Chapter 6).

Interdisciplinarity. This dissertation builds mostly on research from areas within computer science. However, I believe that interdisciplinary collaborations are essential to make the biggest progress in this area of research. Thus, in many of the presented studies, I collaborated with researchers from outside the field of computer science. The interdisciplinary character of this dissertation is reflected in the motivation, formulation and evaluation of the various studies, which are heavily guided by insights from sociolinguistics, folk narrative research, and the social sciences and the humanities at large. Furthermore, while computer science studies tend to focus on prediction tasks, this dissertation features both prediction tasks (Chapters 4, 6, 7, 9) and analysis studies (Chapters 4, 5, 6, 8 and 10). Six out of the nine publications on which this dissertation is based have at least one co-author from the humanities or social sciences. Interacting with researchers from these different disciplines was an enjoyable experience, as Nissani [1997, p. 211] describes: *"To reach the pinnacle of their profession, they [monodisciplinary researchers] often end up exploring one interesting feature of a single atoll. Interdisciplinarians, by contrast, are forever treating themselves to the intellectual equivalent of exploring exotic lands*".

However, interdisciplinary research is also challenging. Differences in language use can occur as a result of jargon, but also different conventions exist in writing (such as the use of *I* versus *we*) because the role of the investigator in the research process tends to be viewed differently [Bracken and Oughton, 2006]. Moreover, appropriate publication venues are not abundant and research published at computer science venues is often not found by social science researchers and vice versa. While the majority of the work has been published within the NLP and IR communities (e.g., CIKM, EMNLP, ICWSM), later on I also presented at non-computer science conferences (e.g., New Ways of Analyzing Variation and the Language in the Media conference) and I co-authored articles published in non-computer science journals (e.g., the Journal of American Folklore [Meder et al., 2016]).

[6] In parallel with this work, other researchers also released a new dataset [King and Abney, 2013].

1.6 Main Contributions

This dissertation provides the following contributions:

A comprehensive overview of the emerging area of computational sociolinguistics (Chapter 2). In recent years a surge of interest can be observed in the study of the social dimension of language using computational approaches. However, researchers working in this area come from a variety of disciplines (e.g., computational linguistics, social computing, sociolinguistics), and as a result, research articles are scattered across various venues. In this dissertation, a comprehensive survey is provided of research in the emerging area of computational sociolinguistics. Furthermore, the survey explores commonalities and differences between the two main disciplines involved, computational linguistics and sociolinguistics, and identifies the primary research challenges in this area.

Methods for new natural language processing tasks (Chapters 7 and 9). Inspired by research questions from sociolinguistics and folk narrative research, this dissertation explores two new NLP tasks.

- *Language identification at the word level*. While automatic language identification is a well-researched problem, until recently the dominant focus was on document-level classification. However, multilingual speakers may use multiple language varieties in a single conversation, sometimes even in the same sentence or word (a phenomenon often referred to as code-switching). Drawing from sociolinguistics, where code-switching is a well-researched topic, we were among the first to study automatic identification of languages at the *word level* [Nguyen and Doğruöz, 2013] and this particular task is now an active research area [Solorio et al., 2014]. We show that incorporating context leads to improved performance compared to only considering individual words and draw attention to the different angles from which the performance can be measured.

- *Automatic identification of tale types*. In folk narrative research, tale types (collections of similar stories, e.g., based on plot) are frequently used to organize and study the narratives. In this dissertation, the task of automatically determining the tale type of a narrative is explored using a learning-to-rank approach.

Statistical testing for linguistic variation (Chapter 6). Identifying linguistic features that exhibit geographical variation (e.g., *pop/soda/coke* to refer to the drink) is an essential step in many dialect studies. However, existing approaches make assumptions about the nature of variation (e.g., Gaussian or aligned to geopolitical units) that often do not hold in practice. Furthermore, they are not applicable to all types of linguistic data. In this dissertation a non-parametric approach is explored that builds on kernel methods from machine learning. The proposed approach is compared with several existing methods using synthetic data. Furthermore, the approaches are applied to three empirical datasets.

Reflection on operationalizations used in computational linguistics approaches (Chapters 4, 5 and 10). To analyze and model social and cultural phenomena quantitatively, they have to be represented in a digital form, leading to simplification and discretization of the phenomena. In this dissertation we reflect on operationalizations used in CL research by focusing on three different tasks and drawing from insights from sociolinguistics and folk narrative research.

- *Automatic age prediction.* Within computational linguistics, age prediction has usually been approached as a multiclass classification problem. However, formulating the age prediction task in this way has several challenges. For example, age boundaries have been determined heuristically and vary between different datasets, making comparisons across datasets difficult. We were the first to study age prediction from three different perspectives: based on age categories, based on age as a continuous variable, and based on life stages.

- *Automatic gender prediction.* Within computational linguistics, gender is usually treated as a binary variable. However, this has the risk of reinforcing stereotypes and furthermore raises questions about the feasibility of the task to achieve an errorless performance, since not everyone employs language in a gender-stereotypical way. We provide a reflection on the operationalization of gender in latent attribute prediction studies based on data collected using an online game.

- *Automatic tale type prediction.* Tale types are a frequently used concept in folk narrative research. However, the concept is not well-defined and many existing catalogues are not consistent in their use of tale types. For example, within the ATU, the most frequently used catalogue, some tale types are specifically about a certain story (e.g., Little Red Riding Hood), while others group narratives belonging to a broad theme (e.g., stories about certain groups of people). We reflect on the operationalization of folk narrative similarity, and in particular on the concept of a tale type, based on similarity ratings of non-experts (crowdworkers) and experts (folk narrative researchers).

New insights in language variation in social media (Chapters 4 and 8). The studies presented in this dissertation also resulted in new insights regarding language variation in social media.

- *Language variation according to age on Twitter.* Based on manual annotation, the (estimated) age of Twitter users is obtained. The study shows how differences in language between ages decrease at older ages, and highlights linguistic features that are characteristic of age differences.

- *Language choice on Twitter.* By employing automatic language identification, a quantitative analysis is presented on the influence of audience on language choices in Twitter. While the target audiences of social media posts are usually unknown, this study focuses on two types of tweets (tweets with hashtags and user mentions) for which an indication of the target audience is obtained based on the tweet content.

1.7 Structure

The remainder of this dissertation is organized as follows:

Background (Part I). This part provides background on the two relevant research fields that are the focus of this dissertation. The chapters are not necessary for understanding the presented research in the remaining chapters, but they are useful for understanding the context of the research in this dissertation. **Chapter 2** provides background on computational sociolinguistics. A comprehensive overview of research in this area is provided, as well as a reflection on methods and a discussion of the main research challenges in this area. **Chapter 3** provides background on computational folkloristics. The most relevant concepts in folkloristics are explained and related work is discussed.

Computational Sociolinguistics (Part II). This part presents research in which variation is considered from the perspective of sociolinguistics. The first chapters focus on the role of language in social identity construction. **Chapter 4** presents research on automatically identifying the gender and age of social media users based on their tweets. **Chapter 5** reflects on the operationalizations of gender and age in computational linguistics studies using data collected using the TweetGenie demo. In **Chapter 6** a non-parametric approach is explored to test whether linguistic variables exhibit geographical variation. The focus then shifts to variation in the choice of language in online multilingual communication. **Chapter 7** develops an automatic approach to identify languages at the word level and **Chapter 8** studies the influence of the target audience on language choice in Twitter.

Computational Folkloristics (Part III). This part presents research in which variation is considered from the perspective of folk narrative research. **Chapter 9** presents a learning-to-rank approach to identify the tale types of folk narratives. **Chapter 10** takes a closer look at narrative similarity and compares how non-experts and experts perceive folk narrative similarity using a crowdsourcing experiment.

Discussion and Conclusion (Part IV). **Chapter 11** reflects on the presented research by discussing aspects related to ethical considerations and potential biases in the collected data. The dissertation ends with a summary and outlook for future work **(Chapter 12)**.

Part I

Background

2

Computational Sociolinguistics

This chapter is based on D. Nguyen, A.S. Doğruöz, C.P. Rosé, and F. de Jong, "Computational Sociolinguistics: A Survey", In Computational Linguistics, 42(3), pages 537-593, 2016 [Nguyen et al., 2016]

2.1 Introduction

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, 2013a]. 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 et al., 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 quantitatively in a multitude of social situations. The opportunity for 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.

[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.

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 et al., 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 et al. [2014b] and Nguyen et al. [2014a] 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, 2013a] and to date there is no overview of the common and complementary aspects of the two fields. This chapter 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 as well as the potential impact of integrating the social dimensions of language use in the development of practical NLP applications. In Section 2.2 (*Methods for Computational Sociolinguistics*), we reflect on methods used in sociolinguistics and computational linguistics. In Section 2.3 (*Language and Social Identity Construction*), we discuss computational approaches to model language variation based on gender, age and geographical location. In Section 2.4 (*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 2.5 (*Multilingualism and Social Interaction*), we present an overview of tools for processing multilingual communication and discuss approaches for analyzing patterns in multilingual communication from a computational perspective. In Section 2.6 we conclude with a summary of major challenges within this emerging field.

2.1.1 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 goes hand in hand with the 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 chapter 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.

2.1.2 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 lay-

ers 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 et al., 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 et al., 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.

2.1.3 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., 2013a], 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 et al., 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].

2.2 Methods for Computational Sociolinguistics

As discussed, one important goal of this chapter 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 sciences 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 (Subsection 2.2.1). 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 Subsection 2.2.2 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 Subsection 2.2.3 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 Subsection 2.2.4 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.

2.2.1 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. 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, e.g., for age prediction [Nguyen et al., 2013a]. 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 et al., 2008], and approximation of semantics [Blei et al., 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 et al., 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 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 2.3) 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.

2.2.2 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 et al., 2004, Wintner, 2002]. As CL has become increasingly data-driven and focused on improving prediction performance on various tasks, 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* is often associated with old and outdated approaches and it seems to have a negative connotation in contrast to the positive reception of empiricism.

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.

2.2.3 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 some quantitative researchers may dismiss qualitative research 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 2.3) 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 et al., 2014b, Nguyen et al., 2014a].

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 et al., 2002], whereas interest in such methods has only recently grown within the social sciences.

2.2.4 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 et al., 2013, Tagliamonte, 2006]. The sociolinguistic datasets are carefully prepared to enable in-depth analyses, 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 et al., 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., 2013a]. 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, 2013b, 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, 2013b]. 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 et al., 2015a, Pavalanathan and Eisenstein, 2015b]. 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, a focus on lexical variation without regard to other aspects may threaten the validity of conclusions. For example, phonology does impact social media orthography at both the word level and structural level [Eisenstein, 2013b], suggesting that studies on phonological variation could inform studies based on social media text data and vice versa. Eisenstein [2013b] 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 developed 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 [2013a] 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. Comparisons of the online data sources in terms of language use have been limited so far [Baldwin et al., 2013, Hu et al., 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. [2014a] 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 limited and more research needs to be done to study the strengths and weaknesses of this approach.

2.3 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 et al., 2014b] and Gender Guesser[3]), which can be attributed to a widespread interest in the entertaining dimension of the linguistic dimension of identity work. The automatic prediction of individual features such as age and gender based on only text is a nontrivial task. Studies that have compared the performance

[2]http://www.tweetgenie.nl

[3]http://www.hackerfactor.com/GenderGuesser.php

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., 2013a]. 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., 2014a]. 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 or income as a proxy for variables like social class. Occupation labels for Twitter users have been extracted from their profile description [Preoţiuc-Pietro et al., 2015a,b, Sloan et al., 2015]. Preoţiuc-Pietro et al. [2015b] 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. Eisenstein et al. [2011] used income data of each ZIP Code Tabulation Area from the US census for analyzing language variation in tweets. However, in their experiments income was less important than most other features (e.g., race and geography) in predicting text.

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). 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., 2013a]). 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 2.3.1). After surveying computational studies on language variation according to gender (Subsection 2.3.2), age (Subsection 2.3.3) and location (Subsection 2.3.4), 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 2.3.5).

2.3.1 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 et al., 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 et al., 2011, Goswami et al., 2009, Nguyen et al., 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 et al., 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 et al., 2013]. Others created their own Twitter dataset [Eisenstein et al., 2011, Kokkos and Tzouramanis, 2014, Liao et al., 2014, Rao et al., 2010, Zamal et al., 2012]. While early studies focused on English, recent studies have used Twitter data written in other languages as well, like Dutch [Nguyen et al., 2013a], Spanish and Russian [Volkova et al., 2013], and Japanese, Indonesian, Turkish, and French [Ciot et al., 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 et al., 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 et al., 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 et al., 2013, Nguyen et al., 2013a]. 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 et al., 2014b, Prabhakaran et al., 2014b], or from Facebook data [Fink et al., 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 et al., 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 et al., 2014b], 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 [2015a] 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., 2013a]. 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].

2.3.2 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 en-

gage 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.

2.3.2.1 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 et al., 2013, Corney et al., 2002, Fink et al., 2012, Gianfortoni et al., 2011, Mukherjee and Liu, 2010, Nowson and Oberlander, 2006, Peersman et al., 2011, Rao et al., 2010, Zamal et al., 2012], logistic regression [Bergsma and Van Durme, 2013, Otterbacher, 2010], Naive Bayes [Goswami et al., 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 processes rather than inherent biological characteristics [Cheshire, 2002].

2.3.2.2 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 et al., 2014b, Burger et al., 2011, Fink et al., 2012, Sarawgi et al., 2011, Yan and Yan, 2006]. Sarawgi et al. [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 et al. [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 et al., 2012] and blogs [Nowson and Oberlander, 2006, Schler et al., 2006]. However, models based on LIWC alone tend to perform worse than unigram/n-gram models [Fink et al., 2012, Nowson and Oberlander, 2006]. By analyzing the developed features, studies have shown that males tend to use more numbers [Bamman et al., 2014b], technology words [Bamman et al., 2014b] and URLs

[Nguyen et al., 2013a, Schler et al., 2006], 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 et al., 2014b, 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 et al. [2014b] 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 et al., 2014b, Otterbacher, 2010, Schler et al., 2006, Schwartz et al., 2013], in particular first person singular [Nguyen et al., 2013a, 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 et al., 2005]. Sarawgi et al. [2011] experimented with probabilistic context-free grammars (PCFGs) by adopting the approach proposed by Raghavan et al. [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 et al. [2012b] 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 et al. [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 et al., 2009, Otterbacher, 2010, Singh, 2001], and more swear words [Boulis and Ostendorf, 2005, Schwartz et al., 2013]. Females use more emotion words [Bamman et al., 2014b, Nowson and Oberlander, 2006, Schwartz et al., 2013], emoticons [Bamman et al., 2014b, Gianfortoni et al., 2011, Rao et al., 2010, Volkova et al., 2013], and typical social media words such as *omg* and *lol* [Bamman et al., 2014b, 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 et al., 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 et al. [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 et al., 2011]. Gianfortoni et al. [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 et al., 2012]. Likewise, Bamman et al. [2014b] 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 et al. [2014b] 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 Subsection 2.4.1). 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 et al. [2009] automatically classified speakers according to interactional style (awkward, friendly, or flirtatious) using various types of features, including lexical features based on LIWC [Pennebaker et al.,

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

2001], prosodic, and discourse features. Differences, as well as commonalities, were observed between genders, and incorporating features from both speakers improved classification performance.

2.3.2.3 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 et al. [2014b], 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., 2014a]. 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 as a universal social variable can be misleading. Furthermore, linguistic variation within speakers of the same gender holds true as well.

2.3.3 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. Subsection 2.3.1). 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., 2013a, 2014a]. 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].

2.3.3.1 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. [2013a] 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 et al., 2011, Rao et al., 2010], logistic regression [Nguyen et al., 2013a, 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 et al. [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., 2013a], 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 et al. [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 et al., 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 et al., 2011, 2013a, 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.

2.3.3.2 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 et al., 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 et al., 2011, 2013a, Peersman et al., 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., 2013a, Rosenthal and McKeown, 2011]. Older people also use more prepositions [Argamon et al., 2009, Nguyen et al., 2013a], determiners [Argamon et al., 2009, Nguyen et al., 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., 2013a, 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 et al., 2011], and more capitalized words (e.g., *HAHA*) [Nguyen et al., 2013a, Rosenthal and McKeown, 2011]. Specific words such as *like* are also highly associated with younger ages [Barbieri, 2008, Nguyen et al., 2011]. Younger people also use more features that indicate stance and emotional involvement [Barbieri, 2008], such as intensifiers [Barbieri, 2008, Nguyen et al., 2013a] 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., 2013a, Rosenthal and McKeown, 2011].

2.3.3.3 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.

2.3.4 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 et al. [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 linguis-

tic 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.

2.3.4.1 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 et al., 2014, Elfardy and Diab, 2013, Huang, 2015, Zaidan and Callison-Burch, 2014], 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 et al., 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 et al., 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.

2.3.4.2 Features and Patterns

Word and character n-gram models have been frequently used in dialect identification [King et al., 2014, Trieschnigg et al., 2012, Zaidan and Callison-Burch, 2014]. Similarly, many text-based location prediction systems make use of unigram word features [Eisenstein et al., 2010, Han et al., 2012, Wing and Baldridge, 2011].

Features inspired by sociolinguistics could potentially improve performance. Darwish et al. [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 et al. [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 et al. [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 et al. [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 et al. [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 et al. [2014a] 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.

2.3.4.3 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.

2.3.5 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 et al. [2014b] incorporated the gender of participants in e-mail conversations and the overall 'gender environment' as features in their model. Volkova et al. [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.

2.4 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 (Subsection 2.4.1). Next, we discuss computational approaches to studying how language reflects and shapes footing within social relationships (Subsection 2.4.2). 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 Subsection 2.4.3. We will discuss two prominent frameworks within sociolinguistics, Audience Design [Bell, 1984] and Communication Accommodation Theory [Giles et al., 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 (Subsection 2.4.4).

2.4.1 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 et al., 2014], a street gang forum [Piergallini et al., 2014], and more recently discussion forums in Massive Open Online Courses (MOOCs) [Wen et al., 2014a,b]. 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 et al., 2007, Gilbert, 2012, Prabhakaran et al., 2012a, 2014b], 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 et al., 2012, Danescu-Niculescu-Mizil et al., 2012, Swayamdipta and Rambow, 2012], transcripts of political debates [Prabhakaran et al., 2013, 2014a] and transcripts of Supreme Court arguments [Danescu-Niculescu-Mizil et al., 2012].

2.4.2 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 et al., 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 et al., 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 et al. [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 et al., 2010], which involves edges that are directed and reflect status differences. Hassan et al. [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].

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 et al., 2007, Nguyen et al., 2014d, Prabhakaran et al., 2013]. For example, Prabhakaran et al. [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 et al. [2012a] 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 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]. For example, studies on identifying persons who are pursuing power [Bracewell et al., 2012, Swayamdipta and Rambow, 2012], detecting influencers [Biran et al., 2012, Huffaker, 2010, Nguyen et al., 2014d, Quercia et al., 2011] and 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 et al., 2011], standings in state and national polls to study power in political debates [Prabhakaran et al.,

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., 2014d, 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 et al., 2012]. Such categories can also be combined into sequences [Bracewell et al., 2012]. Other specialized representations are also used, such as features related to turn taking style [Prabhakaran et al., 2013, Swayamdipta and Rambow, 2012], topic control [Nguyen et al., 2014d, Prabhakaran et al., 2014a, Strzalkowski et al., 2012], and 'overt displays of power', which Prabhakaran et al. [2012b] 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 et al. [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.

2.4.3 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 them-

selves 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 2.5.

Communication Accommodation Theory. Communication Accommodation Theory (CAT) [Giles et al., 1973, 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 et al. [2010], only recently studies have been performed in uncontrolled settings such as Twitter conversations [Danescu-Niculescu-Mizil et al., 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 et al., 2011, Jones et al., 2014, Niederhoffer and Pennebaker, 2002]. Speakers do not necessarily converge on all dimensions [Giles et al., 1991], which has also been observed on Twitter [Danescu-Niculescu-Mizil et al., 2011]. Although earlier studies used correlations of specific features between participants, at turn level or overall conversation level [Levitan et al., 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 et al. [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 et al. [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 et al., 2011, 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

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

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 et al. [2015a] 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 [2015b] 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.

2.4.4 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 et al., 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 et al. [2016] identified and analyzed word innovations in Twitter and Reddit based on variation in frequency, form and meaning. They performed their analyses at a global level, i.e., the whole dataset, and at a community level, based on applying a community detection algorithm to the Reddit data and grouping the geotagged tweets by geopolitical units.

Language change at 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 et al. [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.

2.5 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 et al., 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 et al. [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 (Subsection 2.5.1). Consequently, we discuss research on adapting various NLP tools to process mixed-language texts (Subsection 2.5.2). We conclude this section with a discussion of studies that analyze, or even try to predict, the use of multiple languages in multilingual communication (Subsection 2.5.3).

2.5.1 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]. Multilingual data from online environments is usually extracted in small volumes and for short periods, often because resources or technical support are lacking.

Within computational linguistics, there is a growing interest in the automatic processing of mixed-language texts. Lui et al. [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 et al., 2014, Peng et al., 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 et al., 2014, Solorio and Liu, 2008a,b], Turkish-Dutch [Nguyen and Doğruöz, 2013], Mandarin-English [Adel et al., 2013, Peng et al., 2014], and French-English [Jurgens et al., 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 Subsubsection 2.3.4.1. 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 at 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 the level of sentences is also challenging. For example, Zaidan and Callison-Burch [2014] 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 at the word level.

2.5.2 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

et al., 2014, Kim et al., 2014, Papalexakis et al., 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, 2014, 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 et al., 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 for 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 et al., 2013] or by incorporating syntactic inversion constraints [Li and Fung, 2012]. Peng et al. [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.

2.5.3 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 et al. [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 et al. [2015a] 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 et al. [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.

2.6 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 chapter, 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.

2.6.1 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 inter-

est 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 2.3 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 (Subsection 2.4.3). Others employ measures such as average word and sentence length (e.g., in Section 2.3). 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 at many other levels. Some computational studies have focused on phonological [Eisenstein, 2013b, Jain et al., 2012, Jørgensen et al., 2015] and syntactic [Doyle, 2014, Gianfortoni et al., 2011, Johannsen et al., 2015, Wiersma et al., 2011] 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 2.3). 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.

2.6.2 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 this chapter, 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 2.3) 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 et al., 2011] and explored the effectiveness of domain adaptation approaches [Nguyen et al., 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 et al. [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 (Subsection 2.3.2), age (Subsection 2.3.3) and language variety (Subsection 2.3.4) 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.

2.6.3 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 et al. [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 et al., 2015]). One of the approaches to improve the performance of tools has been to normalize the texts, but as Eisenstein [2013a] 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.

2.7 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 chapter, 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.

3

Computational Folkloristics

3.1 Introduction

In the second part of this dissertation, the studies focus on the domain of folk narratives. Folk narratives (e.g., legends, fairy tales, jokes) are part of folklore and are a significant part of a society's culture. Traditionally, folk narratives have been transmitted from generation to generation through oral communication, but nowadays many of such narratives are also transmitted through written means. Variations of a story appear as the story is transmitted over time. For example, a character might be added or a location might be changed, and such changes are often influenced by the needs of society. Folk narratives are therefore an important source to study diachronic and synchronic variability, shifts in beliefs and moral values, and individual and collective identities [Meder, 2010]. The increasing digitization of folk narratives [Abello et al., 2012, La Barre and Tilley, 2012, Meder, 2010] and advances in the fields of computational linguistics and information retrieval enable the discovery of new patterns and could help speed up the digitization by automatically enriching the narratives with metadata, such as the genre or keywords. In this dissertation, variation in folk narratives is explored by focusing on computational models to analyze folk narrative similarity.

This chapter provides background on the research presented in this dissertation. Section 3.2 describes relevant concepts from folk narrative research. Section 3.3 then continues with a discussion of related work. Section 3.4 discusses the Dutch Folktale Database, which served as the data source for experiments in this dissertation. Finally, this chapter concludes with a short summary and outlook on presented research in this dissertation (Section 3.5).

3.2 Folktales Background

Folklore has been studied using different approaches. One of the main approaches has been the historic-geographic method, which has its roots in Finland and was developed by Julius and Kaarle Krohn. The historic-geographic method assumes that

1	A car driver picks up a hitchhiker. They talk about spiritual topics in life. Suddenly the hitchhiker vanishes. The driver tells the story to the police. They tell him that they have heard the story earlier that day as well.
2	A guy bikes through the park at night. He encounters a girl covered in blood. He brings her to the police, but during the trip she suddenly disappears. She resembles a murdered girl.
3	A car driver picks up a hitchhiker and borrows her his sweater. When he stops by to pick up the sweater, he discovers she passed away due to a car accident a while ago. He finds his sweater on her grave.
4	A car driver picks up a girl wearing a white dress. He accidently spills red wine on her dress. He brings her home, and the next day he finds out she died a year ago. When the police open her grave, they find the white dress with the red wine spot.

Table 3.1: Variants of *The Vanishing Hitchhiker* (BRUN 01000), "A ghostly or heavenly hitchhiker that vanishes from a vehicle, sometimes after giving warning or prophecy".

variants are derived from a single source [Goldberg, 1984]. By collecting and analyzing multiple variants of the same story, the geographical and temporal trajectory of the story can be studied and the original form can be identified. In the words of Uther [2009, p. 19]: "*variation [..] is essential for understanding the narrative's age, the process of its transmission, and its place in tradition*". For example, there may be different variants for the identity of a character in a particular tale (e.g., king vs. emperor). By analyzing the distribution of such variants over time and geographically, the original form of a story may be recovered [Thompson, 1951].

Tale types. *Tale types* – also called story types – are a fundamental concept that traditionally originated in the historic-geographic approach within the study of folklore. According to Thompsom, "*A type is a traditional tale that has an independent existence, it may be told as a complete narrative and does not depend for its meaning on any other tale*" [Thompson, 1951, p. 415]. Stories with the same type are usually 'similar' in terms of plot, motifs or themes. Furthermore, they are assumed to have the same genetic origin [Thompson, 1951]. To illustrate the concept of a tale type, we use *the Vanishing Hitchhiker*, a well-known urban legend found in the Brunvand index. In the Dutch Folktale Database, there are 19 variants of this particular story (March 2016). A selection of the variants of this story is presented in Table 3.1. The characters (the hitchhiker as well as the driver), can be male or female. Sometimes the clothing of the hitchhiker is described in specific details (e.g., wearing white clothes or a red coat). The particular vehicle also varies (e.g., car, motor, bicycle, horse and carriage). The location can be unspecified, or set in a specific place (e.g., city or park). The person that disappears is sometimes described as someone who was murdered, or as an angel. In some variants the clothes of the hitchhiker are found on a grave. Thus, many story elements can be varied.

As another example, the tale type *Little Red Riding Hood* (ATU 333) is about a young girl who visits her grandmother, but then is eaten by a wolf disguised as her grandmother. In some variants the girl manages to escape from the wolf, and in others the story is transformed into a joke. Variants of Little Red Riding Hood have been the subject of various studies (e.g., Tehrani [2013] and Zipes [1993]).

Folktale narrative researchers have developed *type indexes* to classify and organize stories according to their tale types. For example, many narratives are written in language varieties that are not well known to most researchers. By classifying the narratives with tale types and annotating them with various other metadata, such narratives can still be included in the analyses. Many type indexes have been proposed (see discussions by Uther [1996, 2009]), some tailored to certain narrative genres or geographical locations. The best-known type index is the Aarne-Thompson-Uther (ATU) type index [Uther, 2004] that covers many fairy tales, but also legends, jokes and other folktale genres. The initial version was published by Antti Aarne in 1910 [Aarne, 1910] and the most recent version was published in 2004 [Uther, 2004]. An example tale type in the ATU index is *Little Red Riding Hood* (ATU 333). Examples of other type indexes are the Type Index of Urban Legends proposed by Brunvand [2012] and the SINSAG [Sinninghe, 1943] catalogue which focuses specifically on legends.

While tale types have been useful to organize narratives, they also suffer from limitations [Dundes, 1997, Uther, 2009]. First, tale types provide a simplified, binary view of similarity (stories either belong or not belong to the same tale type), while as some critics note, folktales are fluid. The operationalization of story similarity this way, as argued by critics, ignores aspects such as the social context and stylistic elements [Goldberg, 1984]. Being able to distinguish tale types using computational methods has been seen as providing support for the validity of the use of tale types. For example, Tehrani [2013] demonstrate that using phylogenetic methods it is possible to distinguish ATU 333 and ATU 123 as distinct types. In this dissertation, we also use computational methods to take a closer look at the concept of tale types. Second, tale types do not always group narratives at the same level of specificity. Some tale types are very specific, grouping narratives that share a common plot (e.g., *Little Red Riding Hood*, ATU 333), while others group narratives that share a common structure (e.g., repetition). The broadest category of tale types only share a common theme (e.g., *Jokes about devout women*, ATU 1855). Third, many type indexes are heavily biased towards a geographic region. For example, the original type index developed by Antti Aarne was based on Finnish, Danish and German folktales. Materials from other regions were used for further revisions of the index, but the geographic bias remains a limitation in many of these type indexes.

Motifs. Motifs are the smallest elements "*in a tale having a power to persist in tradition*" [Thompson, 1951, p. 415]. Each tale type entry in the ATU index is accompanied with a description that is annotated with *motifs* from the Thompson's Motif Index (TMI, [Thompson, 1955-1958]). The number of motifs to describe a tale type ranges from one to many [Thompson, 1951]. Motifs play a key role in the classification of stories into tale types. However, the tale type descriptions with annotated motifs in the ATU do not capture the variation in the motifs that are present in the individual stories [Haase, 2008]. The second edition of the TMI was published in the 1950s and contains over 45,000 motifs. Example motifs are *Choice between evils* (J210) and *Transformation to person of different social class* (D20). Like the ATU classification, TMI has also received criticism. For example, a study by Karsdorp et al.

[2012a] found that the description level of motifs is too specific for modeling tale types as sequences of recurring motifs.

Subgenres. Folktales belong to various subgenres. The most frequent subgenres in the Dutch Folktale Database are [Nguyen et al., 2012]:

- *Fairy tales* are set in an unspecified time and place, with often a happy ending and magical elements.
- *Legends* are situated in a known place and recent past, with human characters but also supernatural elements such as witches.
- *Urban legends* take place in modern times and are claimed to have happened. They are often about hazardous or embarrassing situations.
- *Jokes* are stories told to entertain, frequently ending with a punch line.

Narratives belonging to the same tale type do not necessarily occur in only one genre. For example, besides as a fairy tale, the *Little Red Riding Hood* tale type also has variants in which the story is told as a joke. In the Dutch Folktale Database, some of these instances revolve around a funny dialogue between Little Red Riding Hood and the Big Bad Wolf, the two main characters of the story.

3.3 Related Work

In this section we discuss prior work on using computational approaches to support the processing and analysis of folktale data. Traditionally, studies of folklore are confined to small, well-defined collections and 'close reading' is applied to study the data. However, as Abello et al. [2012] argue, the sizes of current corpora prevent close-reading of even a fraction of the corpora. They label the emerging field that uses computational approaches for the study of folklore as *computational folkloristics*. Abello *et al*. propose a network representation of folklore data to overcome the limitation of classification schemes and to combine 'distant reading' with 'close reading'. La Barre and Tilley [2012] studied information-seeking and information-use practices of folktale searchers. They note that contextual information and domain-specific features are desirable in search systems and that integrating tale types and motifs into search mechanisms (e.g., for query expansion) is a promising research direction.

Metadata generation. Metadata facilitates the exploration and analysis of folktale data and therefore the automatic extraction of metadata has been pursued in various studies to annotate large corpora. One direction has focused on automatically extracting the characters in folktales. As Karsdorp et al. [2012b] point out, traditional NER systems are not easily transferred to the domain of folktales. For example, inanimate entities may be characters as well, such as the fleeing pancake in the story of *The Fleeing Pancake*. Declerck et al. [2012] identified characters based on extracting indefinite and definite nominal phrases and matching them with an ontology. However, the approach was evaluated on only one narrative. Karsdorp et al. [2012b] identified characters by searching on direct and indirect speech, which they see as

indicators of intentionality. The characters were ranked based on their dispersion pattern in the narrative. Genre classification in the folktale domain has also been explored. Guerini and Strapparava [2016] developed a system to automatically distinguish between urban legends, news and fairy tales. They built on the idea that urban legends should share some properties with both news (who, where, when) and fairy tales (emotional). Nguyen et al. [2012] developed a classifier to distinguish between nine different genres. Character n-grams were found to be effective, however the performance on some genres, such as songs, was low due to the low number of training examples in their data.

Motif identification and analysis. Thompson's Motif Index contains over 45,000 motifs and thus manually assigning them to stories is a daunting task. Therefore, the automatic identification of such motifs is a promising research direction to scale-up motif analyses and to improve the accessibility of folktale data. Voigt et al. [1999] aimed to automatically identify motifs in Hungarian texts. They employed principal component factor analysis, where the identified principal components were used as motifs. However, interpreting the principal components may be difficult. Karsdorp and Van den Bosch [2013] explored the use of Labeled LDA and a simple retrieval model to automatically assign motifs. They found that both models performed fairly well. Ofek et al. [2013] explored classifying sequences of motifs into categories (e.g., magic tales, jokes) of tale types of the ATU. Motif sequences were represented at different levels of abstraction.

Easy accessibility of a source like TMI is essential for its usefulness and success in research. The TMI consists of six volumes, with the last volume being an index to facilitate searching in the TMI. However, as Karsdorp et al. [2015a] point out, manually searching through the TMI using the index is still problematic in many cases. Recently, NLP and IR methods have been explored to enrich the TMI and improve its accessibility. Declerck and Lendvai [2011] used lexical, syntactic and semantic analysis to annotate the TMI index. Much of their system was based on manually crafted rules. For example, when searching for a human, the system would now also return results with e.g., woman, old man, etc. Similarly, Karsdorp *et al*. developed MOMFER [Karsdorp et al., 2015a], a search engine for the TMI. By making use of tools such as WordNet, MOMFER enables searching for semantic concepts. They demonstrate the usability of their tool using several case studies.

Story similarity. Measuring the similarity between folktales is an interesting task because of the inherent variation of this data. Lestari and Manurung [2015] measured the similarity between folktales based on structural similarity between their plot graphs, which they extracted using a dependency parser and a coreference resolution tool. The approach was tested on a small corpus of folktales (24) that were grouped into five categories based on the ATU index. They found that combining their approach with a bag of words approach performed best. Measuring story similarity is an essential component in clustering systems. Lobo and de Matos [2010] explored clustering a set of fairy tales from Project Gutenberg using a combination of Latent Semantic Mapping and item-based top-n recommendation algorithm. Grundkiewicz and

Gralinski [2011] explored different clustering algorithms and the use of text normalisation for clustering urban legends. Abello et al. [2012] created links between stories based on various approaches, including shared keywords and a domain-specific ontology, and for the final clustering various similarity measures were explored. However, evaluating the quality of the produced clusterings remains problematic. For example, the produced clustering is often compared against an existing classification (e.g., Grundkiewicz and Gralinski [2011]) or based on an informal demonstration through a use case (e.g., Abello et al. [2012]). The presented studies in this dissertation consider variation in stories by focusing on story similarity. In Chapter 9 automatic identification of tale types using a learning-to-rank framework is explored and in Chapter 10 human perception of folktale similarity is investigated using a crowdsourcing experiment.

Sentiment analysis. Folktales – in particular fairy tales– have also been of interest for the study of emotions in text. Mohammad [2011] found that fairy tales have a wider distribution of emotion word densities (the expected number of emotion words in a certain window of words) than novels. Alm and Sproat [2005] found that fairy tales often start neutral, but end with a happy emotion.

3.4 The Dutch Folktale Database

With the increasing interest in digital humanities, many cultural heritage collections have been digitized. One such digitized resource is the Dutch Folktale Database (DFD), which is the main resource used in Part III of this dissertation, which focuses on variation in folk narratives. The DFD (screenshot in Figure 3.1) is a large collection of Dutch folktales and available online at http://www.verhalenbank.nl. The database contains over 45,000 folk narratives (Aug, 2016) from a variety of subgenres, such as fairy tales, urban legends and jokes [Meder, 2010]. The narratives have been collected through various methods, ranging from fieldwork to scraping social media, and the narratives have been manually annotated with metadata such as a summary, keywords, language, tale type and named entities. When performing research based on data from the DFD, folktale researchers typically use an online search interface to make the relevant selection of narratives for their research. Searches are typically aimed at groups of narratives from the same subgenre or collector, or with the same main character [Trieschnigg et al., 2013a].

Manually annotating the folktale data is time consuming and as a result there is a large backlog of narratives waiting to be annotated to facilitate accessibility. In addition, folk narrative research using the DFD has traditionally been limited to small sets of narratives, because the size of the collection prevents manually analyzing all the narratives. In the context of the FACT (Folktales as Classifiable Texts) and Tunes and Tales projects, the use of computational tools to analyze and process these folk narratives has been explored, such as the automatic classification of genre [Nguyen et al., 2012], automatic extraction of keywords [Trieschnigg et al., 2013b], animacy detection [Karsdorp et al., 2015b] and motif detection [Karsdorp and Van den Bosch, 2013]. The majority of this research has been summarized in Meder et al. [2016].

The DFD has been at the forefront of digitization in the folklore community, but more recently the availability of other digitized folk narrative data has also increased, for example, with a large collection of online Danish folklore[1] and efforts in annotating French folktales using NLP methods [Garcia-Fernandez et al., 2014]. Thus, future research in this area could also consider other folktale databases.

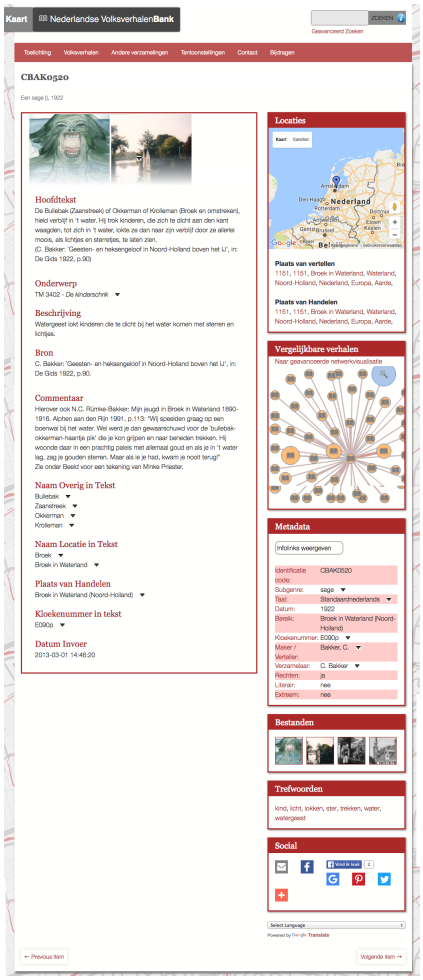


Figure 3.1: Screenshot of the Dutch Folktale Database, August 2016

3.5 Conclusion

Analyzing folk narratives using computational approaches to study culture fits in the emerging trend of treating text as cultural data. The transmission of stories over time and across space gives rise to variation in folk narrative data. This variation, and its implications for measuring and perceiving the similarity between folk narratives, is studied in this dissertation using computational approaches. In this chapter, the main relevant concepts from folk narrative research were explained and relevant research was discussed. Then, the Dutch Folktale Database was introduced, which will be the source of Chapters 9 and 10 in this dissertation.

[1] http://etkspace.scandinavian.ucla.edu/danishfolklore

Part II

Computational Sociolinguistics

Language varies across social groups (e.g., gender and age) and geographically (e.g., the north versus the south of the Netherlands). There is also variation within the language of a single speaker. Certain aspects of a person's identity may be shown more or less explicitly in language use, depending on their culture, the recipient of their utterance, the topic of the conversation, etc. From a sociolinguistic perspective, language is a resource which can be drawn on to study different aspects of a person's social identity [Holmes and Meyerhoff, 2003]. From a computational linguistics perspective, language variation offers the possibility to automatically infer aspects of a person's identity from their language use (see Section 2.3 for an overview).

This part of the dissertation first focuses on predicting the gender and age of authors based on their texts. Earlier research in sociolinguistics regarded male and female, and age as static variables (e.g., Labov [1966] and Trudgill [1974]). However, researchers within sociolinguistics now view them primarily as social variables. Concepts such as gender and age are shaped differently depending on an individual's experiences and personality, as well as the society and culture a person is part of [Eckert, 1997, Holmes and Meyerhoff, 2003]. The operationalizations of gender and age in latent attribute prediction studies are revisited in this dissertation.

Next, computational methods are studied that could be used to support the processing and analysis of large-scale text corpora for sociolinguistic research. Regional dialect studies are often based on a small set of linguistic variables. A study is presented that investigates statistical methods that could support the selection of such variables. A study is also carried out to analyze the use of two regional (or, minority) languages in the Netherlands on Twitter. The use of multiple languages in a single text introduces challenges for NLP tools. An automatic language identification system is therefore developed to identify languages at a fine-grained level in texts to help processing and analyzing texts written by multilinguals.

This part of the dissertation is organized as follows. In **Chapter 4**, a study is described on automatic age prediction of Twitter users based on their tweets. The chapter also discusses the TweetGenie demo, which resulted from this research.

In **Chapter 5**, a reflection is provided on the tasks of gender and age prediction based on data collected using the TweetGenie demo. By analyzing data in which demo visitors tried to guess the gender and age of Twitter users themselves, the operationalizations of gender and age in latent attribute tasks are questioned.

In **Chapter 6**, different approaches are compared to test whether the geographical variation of a linguistic variable is significant. Furthermore, an approach is presented based on non-parametric statistics that overcomes limitations of existing approaches.

In **Chapter 7**, work is presented on identifying Dutch and Turkish in posts from a large online forum for Turkish-Dutch speakers living in the Netherlands. Compared to previous work on automatic language identification that traditionally has focused on documents (see Section 2.5), this chapter explores language identification at the fine-grained level of words.

In **Chapter 8**, automatic language identification is used to study the influence of the (target) audience on the use of minority languages by Twitter users in two Dutch provinces.

4

A Study of Language and Age in Twitter

This chapter is based on:

D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, "'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, 2013 [Nguyen et al., 2013a]

and on: D. Nguyen, D. Trieschnigg, and T. Meder, "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, 2014 [Nguyen et al., 2014b]

4.1 Introduction

A person's language use reveals much about their social identity. Early sociolinguistic studies only had access to relatively small datasets (e.g., a couple of hundred persons), due to time and practical constraints on the collection of data. With the rise of social media such as Twitter, new resources have emerged that can complement these analyses. Twitter offers the opportunity to gather large amounts of informal language from many individuals. However, the Twitter population might be biased and only little is known about the studied persons. To overcome this, we carried out a large annotation effort to annotate the gender and age of Twitter users. While gender is one of the most studied variables within the emerging field of computational sociolinguistics, the relation between age and language has only recently become a topic of interest.

In this chapter we present work on automatically predicting people's age, which can offer new insights into the relation between language use and age. An automatic age prediction system could also be used to improve targeting of advertisements and to support fine-grained analyses of trends on the web. So far, age prediction has primarily been approached by classifying persons into age categories. We revisit this approach by being the first to approach age prediction from three different angles: classifying users into *age categories* (20-, 20-40, 40+), predicting their *exact age*, and

classifying users by their *life stage* (secondary school student, college student, employee). We compare the performance of an automatic system with that of humans on these tasks. Next, to allow a more fine-grained analysis, we use the exact ages of Twitter users and analyze how language use changes with age.

Specifically, we make the following contributions: 1) We present a characterization of Dutch Twitter users as a result of a fine-grained annotation effort; 2) we explore different ways of approaching age prediction (age categories, life stages and exact age); 3) we find that an automatic system has better performance than humans on the task of inferring age from tweets; 4) we analyze variables that change with age, and find that most changes occur at younger ages; 5) we develop a public demo of the research, called TweetGenie, and use the demo to evaluate the age prediction system in the wild.

We start with discussing related work (Section 4.2) and our dataset (Section 4.3). Next, we discuss our experiments on age prediction (Section 4.4). We then continue with a more fine-grained analysis of variables that change with age (Section 4.5). The next section describes the public demo and presents an evaluation of the system using the collected data (Section 4.6). We conclude with a summary (Section 4.7).

4.2 Related Work

Eckert [1997] distinguishes between chronological age (number of years since birth), biological age (physical maturity) and social age (based on life events). Studies about language and age usually consider chronological age and apply an *etic* approach, grouping speakers based on age spans (e.g., Barbieri [2008], Labov [1966], Trudgill [1974]). But speakers can have a very different position in society than their chronological age indicates. Therefore, it might be reasonable to apply an *emic* approach, grouping speakers according to 'shared experiences of time', such as school as a shared experience for teenagers [Eckert, 1997].

So far, automatic age prediction has mostly been approached as a two-class or three-class classification problem based on age spans with for example boundaries at 30 or 40 years (e.g., Garera and Yarowsky [2009], Goswami et al. [2009], Rao et al. [2010]), thus corresponding to an etic approach. However, as choosing boundaries still remains problematic, several researchers have looked more closely into this issue. For example, Rosenthal and McKeown [2011] experimented with varying the binary split for creating age categories. In contrast, Nguyen et al. [2011] approached age prediction as a regression problem, eliminating the need to create age categories. In our work, we will experiment with age prediction as a regression problem, as a classification problem based on age categories and explore an emic approach, by classifying persons according to their life stages.

Both content features and stylistic features (such as part-of-speech and the amount of slang words) have been found to be useful for predicting the age of users [Argamon et al., 2007, Goswami et al., 2009, Nguyen et al., 2011]. Pennebaker and Stone [2003] found that as people get older, they tend to use more positive and fewer negative words, focus more on the future and less on the past and make fewer self-references. Not much research has been done yet on investigating the relationship

between gender and age from a computational perspective. Argamon et al. [2007] found that certain linguistic features that increase with age, also increase more with males. Nguyen et al. [2011] incorporated gender using a binary variable, only allowing a simple interaction between gender and age. Many others have ignored the effect of gender when predicting the age of users.

Experiments on automatic classification of users according to latent attributes such as gender and age have been done on a wide range of resources, including telephone conversations [Garera and Yarowsky, 2009], blogs [Sarawgi et al., 2011], forum posts [Nguyen et al., 2011] and scientific articles [Bergsma et al., 2012b, Sarawgi et al., 2011]. Recently, Twitter has started to attract interest by researchers as a resource to study automatic identification of user attributes, such as ethnicity [Pennacchiotti and Popescu, 2011, Rao et al., 2011], gender [Bamman et al., 2014b, Burger et al., 2011, Fink et al., 2012, Rao et al., 2010, 2011], geographical location [Eisenstein et al., 2010] and age [Rao et al., 2010].

4.3 Data

In this section we describe a large annotation effort we carried out to annotate Dutch Twitter users. Based on the results we present a characterization of Dutch Twitter users.

4.3.1 Selecting and Crawling Users

Twitter users can indicate information such as their name, location, website and short biography in their profile. However, gender and age are not explicit fields in Twitter profiles[1]. As a result, other researchers working on identification of such attributes have resorted to a variety of approaches to construct a corpus, ranging from focused crawling to using lists with common names.

Rao et al. [2010] constructed a corpus by *focused* crawling. To collect users they used a crawl with seeds by looking for profiles that had *baby boomers, junior, freshman* etc. in their description. However, this leads to a potential bias by starting with users that explicitly indicate their age identity in their profile. Burger et al. [2011] sampled users from the Twitter stream and used links to blogging sites, indicated in their profile, to find the gender. Therefore, their set of users was restricted to users having blogs and willing to link them using Twitter. Some approaches used lists of male and female names, for example obtained using Facebook [Fink et al., 2012] or from the US social security department [Bamman et al., 2014b, Zamal et al., 2012].

Our goal was to select a set of users as randomly as possible, and not biasing user selection by searching on well-known stereotypical behavior or relying on links to explicit sources (see Subsection 2.3.1 for a discussion of labeling methods). This did create the need for a large annotation effort, and resulted in a smaller user sample. Using the Twitter API we collected tweets that contained the word *het*, which can be used as a definite article or pronoun in Dutch. This allowed us to restrict our tweets to Dutch as much as possible, and limit the risk of biasing the collection somehow. Dur-

[1] In 2015, an optional birthday field was added to Twitter profiles.

ing a one-week period in August 2012 we sampled users according to this method. Of these users, we randomly selected a set for annotation. We then collected all followers and followees of these users and randomly selected additional users from this set. We only included accounts with less than 5000 followers, to limit the inclusion of celebrities and organizations. For all users, we initially downloaded their last 1000 tweets. Then new tweets from these users were collected from September to December 2012.

	Het		**Followe(e/r)s**	
Annotated	1842	(76%)	1343	(43%)
Not enough tweets	15	(0.6%)	129	(4%)
Not a person	221	(9%)	441	(14%)
Not public	264	(11%)	719	(23%)
Not Dutch	51	(2%)	468	(15%)
Other	46	(2%)	17	(0.5%)
Total	2439		3117	

Table 4.1: Reasons why accounts were discarded/kept by sampling method.

4.3.2 Dutch Twitter Users

In this section we analyze the effect of our sampling procedure, and present a characterization of Dutch Twitter users in our corpus. We employed two students to perform the annotations. Annotations were done by analyzing a user's profile, tweets, and additional external resources (like Facebook or LinkedIn) if available. In this chapter, we only focus on the annotations that are relevant to this study.

Effect of Sampling Method
The annotators were instructed to only annotate the users that met the following requirements:

- The account should be publicly accessible.
- The account should represent an actual person (e.g., not an organization).
- The account should have 'sufficient' tweets (at least 10).
- The account should have Dutch tweets (note that this does not eliminate multilingual accounts).

We separated the reasons why accounts were discarded by the two sampling methods (*het* and followers/followees) that were used (the first requirement in the list that was not satisfied was marked). The results are reported in Table 4.1. We observe that the proportion of actual annotated users is much higher for the users obtained using the query *het*. The users obtained by sampling from the followers and followees included more non-Dutch accounts, as well as accounts that did not represent persons. In addition, there was also a group of people who had protected their account between the time of sampling and the time of annotation. In total, 3,185 users were annotated.

Gender
The biological gender was annotated for 3,166 persons (for some accounts, the annotators could not identify the gender). The gender ratio was almost equal, with 49.5% of the persons being female. However, as we will see later, the ratio depends on age. The annotation of the gender was mostly determined based on the profile photo or a person's name, but sometimes also their tweets or profile description.

Mislove et al. [2011] analyzed the US Twitter population using data from 2006-2009. Using popular female and male names they were able to estimate the gender of 64% of the people, finding a highly biased gender ratio with 72% being male. A more recent study [Beevolve.com] however found that 53% were women, based on information such as name and profile.

Age
Because we expected most Twitter users to be young, the following three categories were used: 20-, 20-40, 40+. The age category was annotated for 3,110 accounts. The results separated by gender are shown in Table 4.2[2]. There are more females in the young age group, while there are more men in the older age groups. The same observation was made in statistics reported by [Beevolve.com].

	20-	**20-40**	**40+**
M	796	488	265
F	1078	316	157

Table 4.2: Age and gender

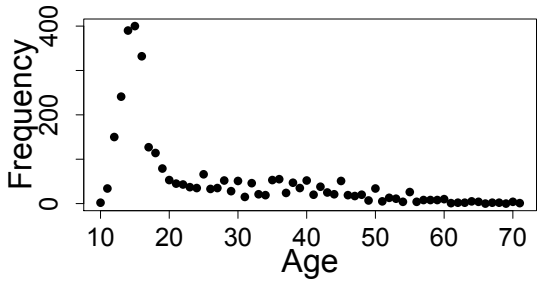


Figure 4.1: Frequencies per age

We also asked our annotators to annotate the exact age. Sometimes it was possible to get an almost exact estimate, for example by using LinkedIn profiles, exact age mentions in the profile, tweets, or mentioning which grade the person was in. However, since this was not always the case, annotators also indicated a margin (0, 2, 5 or 10 years) of how sure they were. Figure 4.1 shows a graph with the frequencies per year of age. Table 4.3 reports the frequencies of the indicated margins. In our data, we find that the margin for young users is low, and that for older users the margin is much higher.

As discussed earlier in this chapter, it may be more natural to distinguish users according to their *life stage* instead of a fixed age category. Life stages can be approached from different dimensions. In this chapter, we use life stages based on the occupation of people, by distinguishing between students, employed, retired etc. The results are displayed in Table 4.4. Unfortunately, the decision to annotate this was done while the annotation process was already underway; therefore the accounts of some users were not available anymore (either removed or protected).

[2]Note that this table only takes persons into account for whom both age and gender were annotated.

Age estimation margin	Frequency
0	703
2	1292
5	918
10	173

Table 4.3: Age annotation margins

Life Stage	Frequency
Secondary school student	1352
College student	316
Employee	1021
Retired	5
Other	15
Unknown	132
Not accessible	344

Table 4.4: Life stage frequencies

We find that the most common life stages are associated with clear age boundaries, although the boundaries are not the same as for the age categories. We find the following age spans in which 90% of the persons fall: secondary school students (12–16 yrs), college students (16–24 yrs), employees (24–52 yrs). However, note that with the life stage approach, people may be assigned to a different group than the group that most resembles their age, if this group matches their life stage better. We have plotted the overlap between life stage and age categories in Figure 4.2.

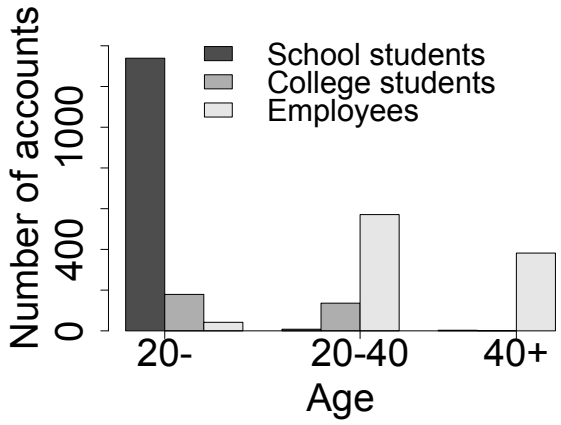


Figure 4.2: Overlap life stage and age categories

4.3.3 Inter-annotator Agreement

We employed two students to perform the annotations. 84 accounts were annotated by both. Inter-annotator agreement was measured using Cohen's kappa. Generally, a value above 0.7 is considered acceptable. We found the following kappa values: gender (1.0), age category (0.83) and life stage (0.70). For the actual age, the mean absolute difference was 1.59 years.

4.4 Age Prediction

4.4.1 Goal

In this section we compare the different ways of approaching age, by testing how feasible age prediction is using simple features based only on the text of tweets. We will automatically predict the following:

- *Age category*: 20-, 20-40, 40+
- *Age*: continuous variable
- *Life stage*: secondary school student[3], college student, employee

For the life stage, we only use categories for which we had a sufficient number of persons. Note that classifying age according to age category and life stage are multiclass classification problems, while treating age as a continuous variable results in a regression problem. In addition, we compare our systems with the performance of humans on this task.

4.4.2 Evaluation

We will evaluate the performance of our classification methods (to predict the age category and life stage) using the F_1 measure. We will report both the macro and micro averages. The regression problem (predicting age as a continuous variable) will be evaluated using the Pearson's correlation coefficient, mean absolute error (MAE) and accuracy, where a prediction was counted as correct if it fell within the margin as specified by the annotators.

4.4.3 Dataset

We restricted our dataset to users who had at least 20 tweets and for whom the gender, age category and exact age were annotated (Table 4.5). For each user we sampled up to 200 tweets. We divided the dataset into a train and test set. Each set contains an equal number of males and females, and the same age distribution (according to the annotated age categories) across gender categories. This limits the risks of the model learning features that for example are more associated with a particular gender, due to that gender occurring more in the particular age category. Parameter tuning and development of the features were done using cross-validation on the training set.

4.4.4 Learning Algorithm

We use linear models, specifically logistic and linear regression, for our tasks. Given an input vector $\mathbf{x} \in \mathbb{R}^m$, $x_1, \ldots, x_m$ represent features (also called independent variables or predictors). In the case of classification with two classes, e.g., $y \in \{-1, 1\}$, the model estimates a conditional distribution $P(y|\mathbf{x}, \beta) = 1/(1+exp(-y(\beta_0+\mathbf{x}^\top\boldsymbol{\beta})))$, where β_0 and $\boldsymbol{\beta}$ are the parameters to estimate. We use a one-versus-all method to

[3]In Dutch this is translated to *scholier*, which includes all students up to and including high school, there is no direct translation in English.

	Train		Test	
	M	**F**	**M**	**F**
20-	602	602	186	186
20-40	231	231	73	73
40+	118	118	37	37
Total	1902		592	

Table 4.5: Dataset statistics

handle multiclass classification. In the case of regression, we find a prediction $\hat{y} \in \mathbb{R}$ for the exact age of a person $y \in \mathbb{R}$ using a linear regression model: $\hat{y} = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$. In order to prevent overfitting we use Ridge (also called L_2) regularization. We make use of the liblinear [Fan et al., 2008] and scikit-learn [Pedregosa et al., 2011] libraries.

4.4.5 Preprocessing & Features

Tokenization is done using the tool by O'Connor et al. [2010]. All user mentions (e.g., *@user*) are replaced by a common token. Because preliminary experiments showed that a unigram system already performs very well, we only use unigrams to keep the approach simple. We keep words that occur at least 10 times in the training documents. In the next section, we will look at more informed features and how they change as people are older.

4.4.6 Results

In this section we present the results of the three age prediction tasks. The results can be found in Tables 4.6 and 4.7. We find that a simple system using only unigram features can already achieve high performance, with micro F_1 scores of above 0.86 for the classification approaches and a MAE of less than 4 years for the regression approach. We also experimented with applying a log transformation of the exact age for the regression task. The predicted values were converted back when calculating the metrics. We find that the MAE and accuracy both improve. In the rest of this section, when referring to the regression run, we refer to the run without a log transformation.

Run	F_1 **macro**	F_1 **micro**
Age categories	0.7670	0.8632
Life stages	0.6785	0.8628

Table 4.6: Results classification

Run	ρ	**MAE**	**Accuracy**
Age regression	0.8845	3.8812	0.4730
Age regression - log	0.8733	3.6172	0.5709

Table 4.7: Results age regression

A scatterplot of the actual age versus the predicted age can be found in Figure 4.3. Figure 4.4 shows the errors per actual age. We find that starting from older ages (around 40-50) the system almost always underpredicts the age. This could have several reasons. It may be that the language changes less as people get older (we show evidence for this in the next section), another plausible reason is that we have very little training data in the older age ranges.

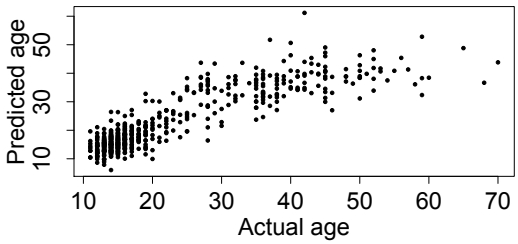


Figure 4.3: Scatterplot age

The most important features for old and young persons are presented in Tables 4.8 and 4.9. We find both content features and stylistic features to be important. For example, content words like *school*, *son*, and *daughter* already reveal much about a person's age. Younger persons talk more about themselves (*I*), and use more chat language such as *haha, xd*, while older people use more conventional words indicating support or wishing well (e.g., *wish, enjoy, thanks, take care*).

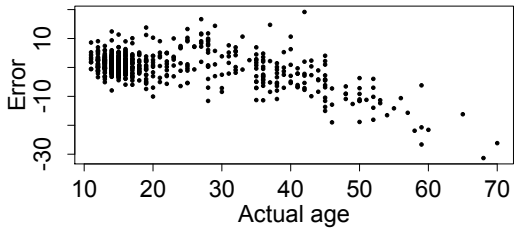


Figure 4.4: Scatterplot absolute error

For the age categories we redid the classification using only persons for whom the life stage was known to allow better comparison between the two classification tasks. We found that people in the 40+ class are often misclassified as belonging to the 20-40 class, and college students are often classified as secondary school students. The precision and recall for the individual classes are listed in Tables 4.10 and 4.11. The performances are comparable. The micro average for life stages is slightly better (0.86 vs. 0.85), the macro average is worse (0.68 vs. 0.75) as the metric is heavily affected by the bad performance on the *students* class. Although life stages are better motivated from a sociolinguistics viewpoint [Eckert, 1997], it is not yet clear which classes are the most suitable. In our corpus, almost all persons were either secondary school students or employees. If a more fine-grained distinction is necessary (for example for personalization), it is still a question which categories should be used.

Dutch	English	Weight
school	school	-0.081
ik	I	-0.073
:)	:)	-0.071
werkgroep	work group	-0.069
stages	internships	-0.069
oke	okay	-0.067
xd	xd	-0.066
ben	am	-0.066
haha	haha	-0.064
als	if	-0.064

Table 4.8: Top features for younger people (regression)

Dutch	English	Weight
verdomd	damn	0.119
dochter	daugther	0.112
wens	wish	0.112
zoon	son	0.111
mooie	beautiful	0.111
geniet	enjoy	0.110
dank	thanks	0.108
goedemorgen	good morning	0.107
evalueren	evaluate	0.105
sterkte	take care	0.102

Table 4.9: Top features for older people (regression)

	Precision	Recall
20-	0.9297	0.9775
20 - 40	0.6739	0.7561
40+	0.8158	0.4493

Table 4.10: Results per class: Age categories

	Precision	Recall
Sec. school student	0.8758	0.9853
College student	0.6667	0.1250
Employee	0.8541	0.8977

Table 4.11: Results per class: Life stages

Train	Test	Age categories		Regression			Life stages	
		Macro F_1	Micro F_1	ρ	MAE	Accuracy	Macro F_1	Micro F_1
All	F	0.7778	0.8750	0.9101	3.4220	0.5135	0.7038	0.8765
	M	0.7563	0.8514	0.8625	4.3405	0.4324	0.6538	0.8500
Male	F	0.6861	0.8277	0.8784	3.9617	0.5135	0.6151	0.8642
	M	0.7027	0.8311	0.8431	4.5017	0.4459	0.6116	0.8346
Female	F	0.7281	0.8581	0.8965	3.5586	0.5270	0.6438	0.8560
	M	0.6373	0.8041	0.8195	5.2099	0.3682	0.6829	0.8538

Table 4.12: Effect of gender

Treating age prediction as a regression problem eliminates the need to choose boundaries. The main drawback is that annotating the exact age of users requires more effort than annotating the life stage or an age category. However, as mentioned before, our annotators showed that reliable annotations are possible (on average less than 2 years difference).

In summary, we believe that both classifying users according to their life stage and treating age prediction as a regression problem are promising approaches. Both approaches complement each other. Age prediction as a regression problem relies on chronological age, while life stages are built on shared experiences between people. Depending on the practical application, knowing the chronological age or life stage might be more informative. For example, groups based on life stage might be more useful for marketing purposes, while the chronological age might be more informative when targeting medical information.

Effect of gender. In Table 4.12 we have separated the performance according to gender. We also experimented with training on data of only one gender, and reported the performance separated by gender. Across the three tasks the performance for females is better than the performance for males. We also find that across the three tasks, the performance for females is better when trained on only females, compared to the performance of males, when trained on only males.

One of the explanations could be that females write slightly more than men (average #tokens: 2,235 versus 2,130), although the differences between the means are small and there is no significant difference in the number of tweets per person (note that we sampled up to 200 tweets per person).

Another explanation can be found in sociolinguistic studies. It has been pointed out that females assert their identity more through language than males [Eckert, 1989, Labov, 1990]. Hence, they might use all kinds of in-group vocabulary more often, thereby marking their affiliation with a certain group. Men's vocabulary, on the contrary, is more homogenous across the in-groups [Eckert, 2000]. Consistent with this, Ling [2005, p. 348] found that females "*seem to have a broader register when using SMS*". Due to this, it might be easier to determine the age of women. However, neither Eckert [1989] nor Labov [1990] looked at age specifically, and the studied people were also not comparable (e.g,. Eckert [1989] only studied young people, and social media settings have not been explored much yet).

4.4.7 Error Analysis

As reported in the previous section, not for all cases the correct age was predicted. This is of course not surprising. People do not only constitute their identity on the basis of their age, but they combine various variables in order to express their selves. For example, a person is not only a teenager, but also a female, a high school student, a piano player, etc. [Eckert, 2008]. Depending on what a person wants to express at a particular moment and towards a particular person, certain aspects of his/her identity may be more emphasized, making age prediction even more complicated. To illustrate this, we will discuss two Twitter users for whom the age was incorrectly predicted.

Case study 1
The first person is a 24-year old student, who the system estimated to be a 17-year old secondary school student. The top 10 most frequent words for this user are *@USER, RT, •, Ik (I), ≪, G, :D, Hahaha, tmi,* and *jij (you)*. The use of special characters like a dot (•) and the 'much less than' sign (≪) is characteristic for younger Twitter users, who separate statements in their tweets employing these characters. *I* is one of the words being the most predictive of younger people as was presented in the feature analysis (see Table 4.8) and the other words like *hahaha, you* etc. are also highly associated with younger persons in our corpus. As we can see, this person employs these words with such a high frequency that he can easily be mistaken for a secondary school student under 20. Examples containing salient words are the ones below:

@USER kommmdan nurd
@USER comeonthen nurd [nerd]

Hahaahhahaha kkijk rtl gemist holland in da hood, bigga huiltt ik ga stukkk
Hahaahhahaha [I am] wwatching rtl gemist[4] *holland in da hood*[5]*, bigga is cryingg it's killinggg me*

RT @USER: Ook nog eens rennen voor me bus #KutDag • Ik heb weekend :)
RT @USER: Had to run for my bus too #StupidDay • I have weekend :)

In addition to the words mentioned above, *me (my)*, and *heb (have)* appear, which are indicative for younger persons in our corpus, as well. Next to the fact that this person employs words rather associated with teenagers on Twitter, we can also derive what kind of identity is constituted here. In the tweets, unconventional punctuation, emoticons, ellipsis, in-group vocabulary (*nurd*), and alphabetical lengthening (*stukkk*) are used to create an informal, unconventional style particularly addressing an in-group. It can be concluded that this person does not appear to stress his identity as an adult, but finds other aspects of his identity more important to emphasize. These aspects, however, are expressed with features employed most frequently by younger persons in our corpus, resulting in a wrong age prediction for this person.

Case study 2
The second person is a 19-year old student. However, the system predicted him as being a 33-year old employee. The top 10 most frequent words for this user are *@USER, CDA, RT, Ik (I), VVD, SGP, PvdA, D66, bij (at)* and *Groenlinks*. It becomes clear that this person tweets about politics a lot, with Dutch political parties (*CDA, VVD, SGP, D66, Groenlinks*) being six out of his ten most frequent words. Tweets that are characteristic for this user and that relate to some of his most salient words are, for example:

@USER Woensdagochtend 15 augustus start het landelijke CDA met haar regiotour op Goeree-Overflakkee i.s.m. @USER.
@USER On Wednesday morning, the 15th of August the national CDA starts with its tour through the region in Goeree-Overflakkee in collaboration with @USER

RT @USER: Vanmiddag met @USER gezellig bij @USER een wijntje gedaan en naar de Emmaüskerk #Middelharnis geweest. Mooie dag zo!
RT @USER: Had fun this afternoon had wine at @USER with @USER and went to the Emmaüschurch #Middelharnis. Beautiful day!

Almost all of his tweets are (like the first example) about politics, so we can assume this user wants to stress his identity as a person interested in politics, or even

[4]Website where people can watch tv shows online.
[5]Dutch reality show.

as a politician on Twitter. Certainly, this is a more common topic for users older than a 19-year old. Proof for this is the fact that words such as *ministers, elections, voter* etc. are highly ranked features associated with older people in the regression model. In addition, the person uses more prepositions, conventional punctuation, formal abbreviations and for example mentions *wine* which is also rather associated with older people in our corpus. Moreover, *beautiful* is one of the top ten features predictive of older people. Thus, not only the main topic of his tweets (politics) is associated more with older people, but he also represents himself as a grown-up person in his other tweets by using which what we perceive as rather conservative vocabulary and punctuation.

Thus, the discussed cases show that people can emphasize other aspects of identity than age. This can result in a deviation from style and content from their peers, thereby making the automatic prediction of age more difficult.

4.4.8 Manual Prediction

In this section we compare the performance of our systems with the performance of humans on the task of inferring age *only* from tweets. A group of 17 people (including males and females, old and young, active and non-active Twitter users) estimated the gender, life stage, exact age and age categories for a random subset of the Twitter users in the test set. Each person was assigned a different set of about 20 Twitter users. For each Twitter user, a text file was provided containing the same text as used in our automatic prediction experiments. The participants received no additional information such as the name, profile information etc. They could decide themselves how carefully they would read the text, as long as they could make a serious and informed prediction. On average, it took about 60-90 min to do the task. In total there are 337 users for whom we both have manual and automatic predictions. The results can be found in Tables 4.13 and 4.14.

Run	F_1 **macro**	F_1 **micro**
Age categories		
Manual	0.619	0.752
Automatic	0.751	0.858
Life stages		
Manual	0.658	0.778
Automatic	0.634	0.853

Table 4.13: Results classification - manual vs. automatic

Run	ρ	**MAE**	**Acc.**
Manual	0.784	4.875	0.552
Automatic	0.879	4.073	0.466

Table 4.14: Results age regression - manual vs. automatic

Using McNemar's Test we find that the automatic system is significantly better in classifying according to age categories (χ^2 = 18.01, df=1, $p < 0.01$) and life stages (χ^2 = 9.76, df=1, $p < 0.01$). The automatic system is also significantly better in predicting the exact age when comparing the MAEs (paired t-test, t(336) = 2.79, $p < 0.01$). In addition, for each metric and task we calculated which fraction of the persons performed *equal or better than* the automatic system. This ranged from 0.24

(age cat., all metrics) to 0.41 (life stages, micro F_1) and 0.47 (life stages, macro F_1), to 0.29 (exact age, MAEs) and 0.82 (exact age, accuracy).

In addition we find the following. First, humans achieve a better accuracy for the regression task. The accuracy is based on margins as indicated by the annotators. Humans were often closer at the younger ages, where the indicated margins were also very low and a slightly off prediction would not be counted as correct. Second, humans have trouble predicting the ages of older people as well. The correlation between the MAEs and exact ages are 0.58 for humans and 0.60 for the automatic system. Third, humans are better in classifying people into life stages than in age categories.

To conclude, we find that an automatic system is capable of achieving *better* performance than humans, and being much faster (on average, taking less than a second compared to 60-90 minutes to predict the age of 20 users).

4.5 Analysis of Age-Related Linguistic Variables

By analyzing the importance of features in an automatic prediction system, only general effects can be seen (i.e., this feature is highly predictive for old versus young). However, to allow for a more detailed analysis, we now use the exact ages of Twitter users to track how variables change with age.

4.5.1 Variables

We explore variables that capture style as well as content.

Style

The following style variables capture stylistic aspects that a person is aware of and explicitly chooses to use:

- *Capitalized words*, for example *HAHA* and LOL. The words need to be at least 2 characters long.
- *Alphabetical lengthening*, for example *niiiiiice* instead of *nice*. Matching against dictionaries was found to be too noisy. Therefore, this is implemented as the proportion of words that have a sequence of the same three characters in the word. The words should also contain more than one unique character (e.g., tokens such as *www* are not included) and contain only letters.
- *Intensifiers*, which enhance the emotional meaning of words (e.g., in English, words like *so, really* and *awful*)[6].

The following variables capture stylistic aspects that a person usually is not aware of:

- *LIWC-prepositions*, the proportion of prepositions such as *for, by* and *on*. The wordlist was obtained from the Dutch LIWC [Zijlstra et al., 2005][7] and contains 48 words.

[6] *Zo, heel, helemaal, hele, super, so, kut, kapot, fucking, vet, zeer, kanker, zoo, dik, facking, kei, tering, bijzonder, zooo, klote, keihard, rot, zoooo.*

[7] Obtained from Zijlstra in October 2012.

- *Word length*, the average word length. Only tokens starting with a letter are taken into account, so hashtags and user mentions are ignored. Urls are also ignored.
- *Tweet length*, the average tweet length.

References
Pennebaker and Stone [2003] found that as people get older, they make fewer self-references. We adapt the categories for the Dutch LIWC [Zijlstra et al., 2005] to use on Twitter data by including alphabetical lengthening, slang, and English pronouns (since Dutch people often tweet in English as well).

- *I*, such as *I, me, mine, ik, m'n, ikke*.
- *You*, such as *you, u, je, jij*.
- *We*, such as *we, our, ons, onszelf, wij*.
- *Other*, such as *him, they, hij, haar*.

Conversation

- *Replies*, proportion of tweets that are a reply or mention a user (and are not a retweet).

Sharing

- *Retweets*, proportion of tweets that are a retweet.
- *Links*, proportion of tweets that contain a link.
- *Hashtags*, proportion of tweets that contain a hashtag.

4.5.2 Analysis

We calculate the Pearson's correlation coefficients between the variables and the actual age using the same data from the age prediction experiments (train and test together), and report the results separated by gender in Table 4.15.

We find that younger people use more explicit stylistic modifications such as alphabetical lengthening and capitalization of words. Older people tend to use more complex language, with longer tweets, longer words and more prepositions. Older people also have a higher usage of links and hashtags, which can be associated with information sharing and impression management. The usage of pronouns is one of the variables most studied in relation with age. Consistent with Pennebaker and Stone [2003] and Barbieri [2008] we find that younger people use more first-person (e.g., *I*) and second person singular (e.g., *you*) pronouns. These are often seen as indicating interpersonal involvement. In line with the findings of Barbieri [2008], we also find that older people more often use first-person plurals (e.g., *we*).

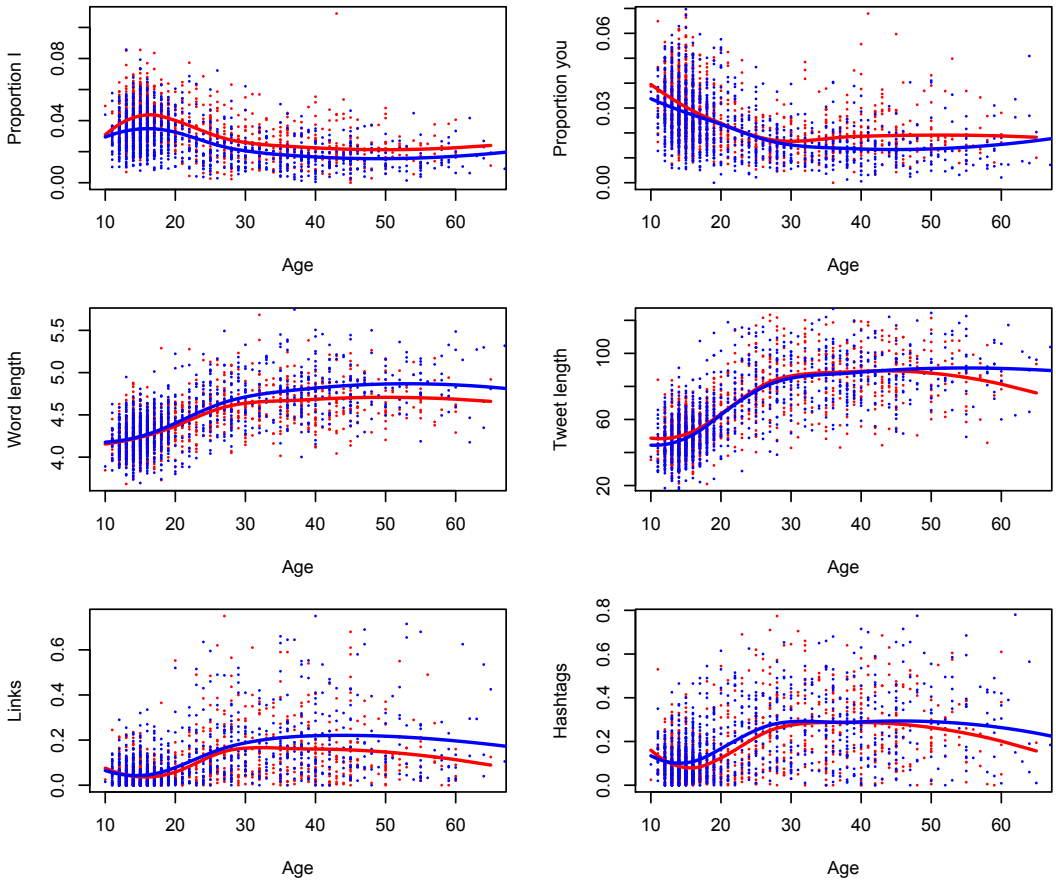


Figure 4.5: Plots of variables as they change with age. Blue: males, Red: females

In Figure 4.5 we have plotted a selection of the variables as they change with age, separated by gender. We also show the fitted LOESS curves [Cleveland et al., 1992]. One should keep in mind that we have less data in the extremes of the age ranges. We find strong changes in the younger ages; however after an age of around 30 most variables show little change. What little sociolinguistics research there is on this issue has looked mostly at individual features. Their results suggest that the differences between age groups above age 35 tend to become smaller [Barbieri, 2008]. Such trends have been observed with stance [Barbieri, 2008] and tag questions [Tottie and Hoffmann, 2006]. Related to this, it has been shown that adults tend to be more conservative in their language, which could also explain the observed trends. This has been attributed to the pressure of using standard language in the workplace in order to be taken seriously and get or retain a job [Eckert, 1997].

One should keep in mind, however, that we have studied people with different ages, and we did not perform a longitudinal study that looked at changes within persons as they became older. Therefore the observed patterns may not indicate actual change within persons, but could be a reflection of changes between different generations (see also Subsubsection 2.3.3.3).

Variable	Females ρ	Males ρ
Style		
Capitalized words	-0.281**	-0.453**
Alph. lengthening	-0.416**	-0.324**
Intensifiers	-0.308**	-0.381**
LIWC-prepositions	0.577**	0.486**
Word length	0.630**	0.660**
Tweet length	0.703**	0.706**
References		
I	-0.518**	-0.481**
You	-0.417**	-0.464**
We	0.312**	0.266**
Other	-0.072	-0.148**
Conversation		
Replies	0.304**	0.026
Sharing		
Retweets	-0.101*	-0.099*
Links	0.428**	0.481**
Hashtags	0.502**	0.462**

Table 4.15: Analysis of variables. For both genders n = 1247. Bonferroni correction was applied to p-values. $^{*}p \leq 0.01$ $^{**}p \leq 0.001$

Reflecting on the age prediction task and the analysis presented in this section, we make the following observations. First, for some variables there is almost no difference between males and females (e.g., tweet length), while for some other variables one of the genders consistently uses that variable more (e.g., the first singular pronouns for females, links for men). In our prediction experiments, we also observed differences in the prediction performance between genders. We also found differences in the gender distribution across age categories on Twitter. Therefore, we conclude that researchers interested in the relation between language use and age should not ignore the gender variable.

Second, in the automatic prediction of exact age we found that as people get older the system almost always underpredicts the age. When studying how language changes over time, we find that most change occurs in the younger ages, while at the older ages most variables barely change. This could be an explanation of why it is harder to predict the correct age of older people (for both humans and the automatic system). This also suggests that researchers wanting to improve an automatic age prediction system should focus on improving prediction for older persons, and thus identifying variables that show more change at older ages.

4.6 Evaluation in the Wild

To evaluate the age prediction system in the wild, we developed TweetGenie[8], a website that allows visitors to enter public Dutch Twitter accounts. We evaluate the system in two ways: 1) using the feedback from users; and 2) using manual annotation. We first describe the demo and then discuss the results.

4.6.1 TweetGenie

Description. TweetGenie predicts the gender and age of the users behind the entered accounts based on the 200 most recent tweets. For this demo, age was modeled as a continuous variable as described earlier in this chapter. The gender prediction was carried out using logistic regression. Due to press attention from various media outlets, we were able to attract a large number of visitors. After a visitor enters a public Twitter account, a results page is shown (see Figure 4.6 for a screenshot). The results page shows the predicted age (in years), the gender, and a gender 'score' indicating how strong the prediction was (based on $\mathbf{x}^\top \boldsymbol{\beta}$ with $\mathbf{x}$ being the features and $\boldsymbol{\beta}$ the estimated parameters). In addition, an option is available to share their results page on Twitter. To collect data and improve the system, users are encouraged to provide feedback on the predictions. On the page with the automatic prediction (Figure 4.6), users have the option to enter the correct age and confirm whether the gender prediction was correct. An overview of the components is shown in Figure 4.8. The first webserver hosts the frontend. A second webserver is used to retrieve the data from Twitter and perform the predictions. A MySQL database is used to keep track of the progress of each prediction, and to store logs and feedback received by users.

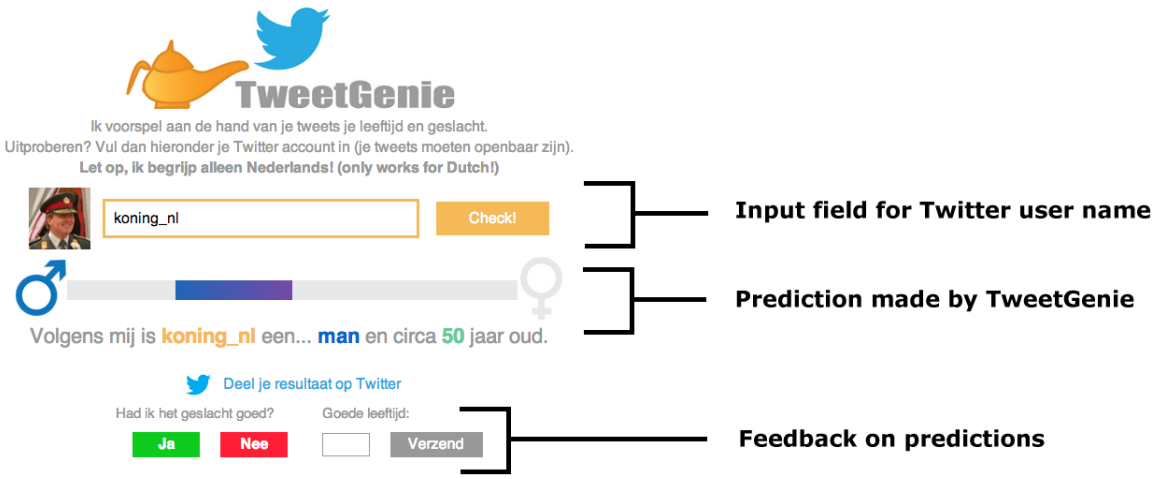


Figure 4.6: Screenshot prediction interface

Public response. TweetGenie was launched on May 13, 2013 at around 11.30 AM. To reach a large audience, a press statement was released and messages were posted on social media networks. We now analyze the data in the first week after the launch.

[8] www.tweetgenie.nl

Figure 4.7 shows the number of entered Twitter users and the number of tweets mentioning TweetGenie in the first week after the launch. The number of tweets and the number of users entered follow similar trends. We observe a high peak in the beginning, but it also rapidly decreases over time. The system was asked to make a prediction 87,818 times and 9,291 tweets were posted with the word 'TweetGenie'. 1,931 of these tweets were created using the tweet sharing function of TweetGenie. The observed sentiment was mostly positive. If TweetGenie made an incorrect prediction, most people joked about it (e.g., "**grin* I just became 13 years younger without plastic chirurgy #tweetgenie*").

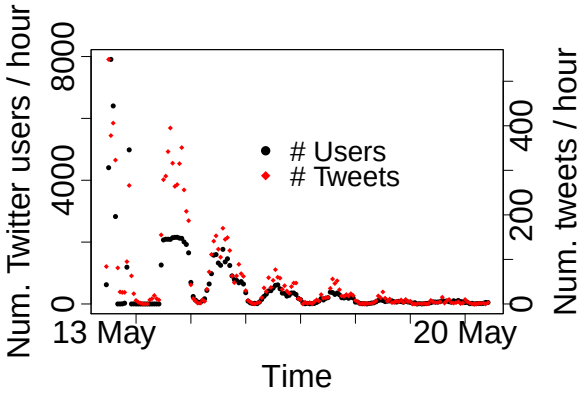


Figure 4.7: Entered users and tweets per hour

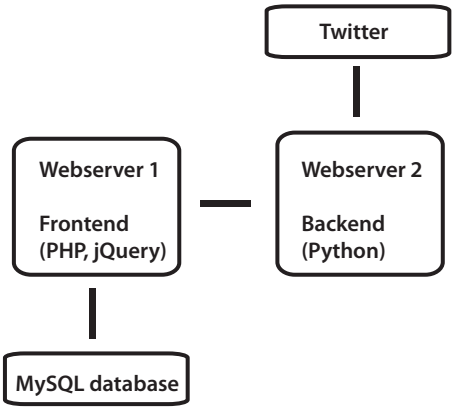


Figure 4.8: Architecture

4.6.2 Evaluation

We now discuss the evaluation of the system based on data collected using TweetGenie.

Evaluation based on user feedback. Visitors were encouraged to give feedback on the predictions of TweetGenie. In the first week, we received feedback on the gender of 16,563 users and on the age of 17,034 users. We randomly sampled 150 Twitter users for which we received feedback on both the gender and age. We checked the feedback of these users by visiting their Twitter profiles. If the feedback seemed plausible based on the profile, we assumed the feedback was correct (i.e., we did not visit any other social media profiles to find the exact age). The results are shown in Table 4.16. We find that 90% of the feedback appears to be correct. Only a small fraction (4%) of the feedback was incorrect, this could be deliberate or due to sloppiness. The remaining feedback was on Twitter accounts of non-Dutch users (e.g., English, German, French), or accounts that did not represent a person (e.g., a sports team, animal, multiple persons).

We calculate the performance based on the 135 users for who we received correct feedback. The results show that the users who gave feedback are *not* representative of the general Dutch Twitter population as estimated in Section 4.3. The users are older than average (the age distribution is shown in Figure 4.9). There are more older

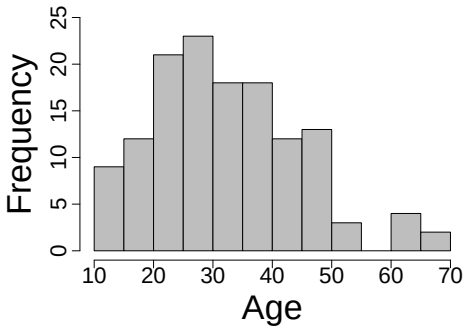


Figure 4.9: Age distribution feedback

Feedback	Freq.	Perc.
Correct	135	90%
Incorrect	6	4%
Not a Dutch account	5	3.33%
Not a person	4	2.67%

Table 4.16: Statistics feedback reliability

males, and more younger females using Twitter in the Netherlands (Section 4.3), and as a consequence the number of males (60.7%) is higher than the number of females (39.3%).

Based on this dataset, we find that the accuracy of the gender predictions was 94%. The Mean Absolute Error (MAE) for the age predictions is 6.1 years, which is higher than reported earlier in this chapter (Section 4.4.6). However, this can be explained by the observation that relatively many older Twitter users give feedback, and as discussed, automatic age predictions for older Twitter users tend to be less accurate.

Evaluation based on manual annotation. We also evaluated the system by manually annotating 50 users that were randomly sampled from the entered users in the logs. We did not include accounts that were not Dutch or did not represent individual persons. If feedback was available for a Twitter user, we used the provided feedback (after a manual check). Otherwise, we manually annotated the gender and age using all available information (e.g., social media profiles, websites).

The gender was correctly predicted for 82% of the users, which is lower than measured in the evaluation based on the user feedback. The Mean Absolute Error (MAE) is 6.18 years, which is in line with the observed MAE based on the user feedback.

Our analyses confirm that users for who feedback was available are *not* representative of all users who were entered in the system. Of the sampled 50 entered users, the fraction of males and females is almost equal (52% and 48%) compared to 60.7% and 39.3% according to the evaluation based on the user feedback. The number of users who were less than 20 years old (15) is similar to the number of users in the range of > 20 and ≤ 30 years (17), while in the previous analysis the fraction of users below 20 years is smaller. Thus, less feedback was received for younger Twitter users. In line with the analysis based on user feedback, we also find that relatively many older Twitter users were entered into TweetGenie compared to the more representative set of Dutch Twitter users that was collected to train the age prediction model.

4.7 Conclusion

We presented a study on the relation between the age of Twitter users and their language use. A dataset was constructed by means of a fine-grained annotation effort of more than 3,000 Dutch Twitter users. We studied age prediction based only on tweets. Next, we presented a detailed analysis of variables as they change with age.

We approached age prediction in different ways: predicting the age category, life stage, and the actual age. Our system was capable of predicting the exact age within a margin of 4 years. Compared with humans, the automatic system performed better and was much faster than humans. For future research, we believe that life stages or exact ages are more meaningful than dividing users based on age groups. In addition, gender should not be ignored as we showed that how age is displayed in language is also strongly influenced by the gender of the person.

We also found that most changes occur when people are young, and that after around 30 years the studied variables show little change. This may also explain why it is more difficult to predict the age of older people (for both humans and the automatic system).

Our models were based *only* on the tweets of the user. This has as a practical advantage that the data is easy to collect, and thus the models can easily be applied to new Twitter users. However, a deeper investigation into the relation between language use and age should also take factors such as the social network and the direct conversation partners of the tweeters into account.

The age prediction model was tested in the wild by developing a public demo called TweetGenie. The demo was an excellent opportunity to gather more data. Compared to the evaluation on the dataset that was used to develop the model, the performance numbers achieved by TweetGenie were lower. Users who were entered into the system were not representative of the Dutch Twitter population: relatively more older Twitter users were entered in the system, leading to more errors in the automatic age prediction.

5

On Gender and Age Prediction: Lessons from a Crowdsourcing Experiment

This chapter is based on D. Nguyen, D. Trieschnigg, A.S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F. de Jong, "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, 2014 [Nguyen et al., 2014a]

5.1 Introduction

In line with scholars from a variety of disciplines, including the social sciences and philosophy, sociolinguists consider age and gender as social and fluid variables [Eckert, 2012]. Gender and age are shaped depending on the societal context, the culture of the speakers involved in a conversation, the individual experiences and the multitude of social roles: a female teenager might also be a high school student, a piano player, a swimmer, etc. [Eckert, 2008].

Speakers use language as a resource to construct their identity [Bucholtz and Hall, 2005]. For example, a person's gender identity is constructed through language by using linguistic features associated with male or female speech. These features gain social meaning in a cultural and societal context. On Twitter, users construct their identity through interacting with other users [Marwick and boyd, 2011]. Depending on the context, they may emphasize specific aspects of their identity, which leads to linguistic variation both within and between speakers. We illustrate this with the following three tweets:

Tweet 1: *I'm walking on sunshine <3 #and don't you feel good*
Tweet 2: *lalaloveya <3*
Tweet 3: *@USER loveyou ;D*

In these tweets, we find linguistic markers usually associated with females (e.g., a heart represented as *<3*). Indeed, 77% of the 181 players guessed that a female

wrote these tweets in our online game. However, this is a 16-year old biological male[1], whose Twitter account reveals that he mostly engages with female friends. Therefore, he may have accommodated his style to them [Danescu-Niculescu-Mizil et al., 2011] and as a result he employs linguistic markers associated with the opposite biological sex.

Most of the NLP research focusing on predicting gender and age has approached these variables as *biological* and *static*, rather than *social* and *fluid*. For example, current approaches use supervised machine learning models trained on tweets from males and females. However, the resulting stereotypical models are ineffective for Twitter users who tweet differently from what is to be expected from their biological sex. As explained above, language use is based on social gender and age identity, and not on biological sex and chronological age. In other words, treating gender and age as fixed biological variables in analyzing language use is too simplistic. By comparing the *biological sex and chronological age* of Twitter users with how they are perceived by the crowd (as an indication of socially constructed identities), we shed light on the *difficulty* of predicting gender and age from language use and draw attention to the *inherent limitations* of current approaches.

As has been demonstrated in several studies, the crowd can be used for experimentation (e.g., Munro et al. [2010]). Our study illustrates the value of the crowd for the study of human behavior, in particular for the experimental study of the social dimension of language use. To collect data, we created an online game (an example of *gamification* [Deterding et al., 2011]) in which thousands of players (the crowd) guessed the biological sex and chronological age of Twitter users based on only the users' tweets. While variance between annotators has traditionally been treated as noise, more recently variation is being treated as a *signal* rather than noise [Aroyo and Welty, 2013]. For example, Makatchev and Simmons [2011] analyze how English utterances are perceived differently across language communities.

This chapter follows this trend, treating variation as meaningful information. We assume that the crowd's perception (based on the distribution of the players' guesses) is an indication of to what extent Twitter users emphasize their gender and age identity in their tweets. For example, when a large proportion of the players guess the same gender for a particular user, the user is assumed to employ linguistic markers that the crowd associates with gender-specific speech (e.g., iconic hearts used by females). Our contributions are as follows:

- We demonstrate the use of gamification to study sociolinguistic research problems (Section 5.3).
- We study the difficulty of predicting an author's gender (Section 5.4) and age (Section 5.5) from text alone by analyzing prediction performance by the crowd. We relate our results to sociolinguistic theories and show that approaching gender and age as fixed biological variables is too simplistic.
- Based on our findings, we reflect on current approaches to predicting age and gender from text, and draw attention to the limitations of these approaches (Section 5.6).

[1] Estimated based on his Twitter profile and tweets.

5.2 Related Work

Gender. Within sociolinguistics, studies on gender and language have a long history [Eckert and McConnell-Ginet, 2013]. More recently, the NLP community has become increasingly interested in this topic. Most of the work aims at predicting the gender of authors based on their text, thereby focusing more on prediction performance than sociolinguistic insights.

A variety of datasets have been used, including Twitter [Bamman et al., 2014b, Bergsma and Van Durme, 2013, Burger et al., 2011, Fink et al., 2012, Rao et al., 2010], blogs [Mukherjee and Liu, 2010, Schler et al., 2006], telephone conversations [Garera and Yarowsky, 2009], YouTube [Filippova, 2012] and chats in social networks [Peersman et al., 2011]. Females tend to use more pronouns, emoticons, emotion words, and blog words (*lol, omg*, etc.), while males tend to use more numbers, technology words, and links [Bamman et al., 2014b, Nguyen et al., 2013a, Rao et al., 2010]. These differences have also been exploited to improve sentiment classification [Volkova et al., 2013] and cyberbullying detection [Dadvar et al., 2012].

To the best of our knowledge, the study by Bamman et al. [2014b] is the only computational study that approaches gender as a social variable. By clustering Twitter users based on their tweets, they show that multiple gendered styles exist. Unlike their study, we use the crowd and focus on implications for gender and age prediction.

Age. Eckert [1997] makes a distinction between chronological (number of years since birth), biological (physical maturity), and social age (based on life events). Most of the studies on language and age focus on chronological age. However, speakers with the same chronological age can have very different positions in society, resulting in variation in language use. Computational studies on language use and age usually focus on automatic (chronological) age prediction. This has typically been modeled as a classification problem, although this approach often suffers from ad hoc and dataset dependent age boundaries [Rosenthal and McKeown, 2011]. In contrast, recent works also explored predicting age as a continuous variable and predicting lifestages [Nguyen et al., 2011, 2013a].

Similar to studies on gender prediction, a variety of resources have been used for age prediction, including Twitter [Nguyen et al., 2013a, Rao et al., 2010], blogs [Goswami et al., 2009, Rosenthal and McKeown, 2011], chats in social networks [Peersman et al., 2011] and telephone conversations [Garera and Yarowsky, 2009]. Younger people use more alphabetical lengthening, more capitalization of words, shorter words and sentences, more self-references, more slang words, and more Internet acronyms [Barbieri, 2008, Goswami et al., 2009, Nguyen et al., 2013a, Pennebaker and Stone, 2003, Rao et al., 2010, Rosenthal and McKeown, 2011].

Perception experiments. In perception experiments by sociolinguists, non-linguists have been asked to identify social characteristics of speakers based on their speech [Clopper, 2013]. Perception experiments have also been used to map regional variation from the viewpoint of non-linguists [Preston, 2015]. The use of crowdsourcing to obtain perception data is a relatively unexplored direction.

5.3 Data

To study how people perceive the gender and age identity of Twitter users based on their tweets, we created an online game. Players were asked to guess the gender and age of Twitter users from tweets. The game was part of a website (TweetGenie, www.tweetgenie.nl) that also hosted an automatic system that predicts the gender and age of Twitter users based on their tweets [Nguyen et al., 2014b]. To attract players, a link to the game was displayed on the page with the results of the automatic prediction, and visitors were challenged to test if they were better than the automatic system (TweetGenie).

5.3.1 Twitter Data

We sampled Dutch Twitter users in the fall of 2012. We employed external annotators to annotate the biological sex and chronological age (in years) using all information available through tweets, the Twitter profile and external social media profiles such as Facebook and LinkedIn. In total over 3,000 Twitter users were annotated. For more details regarding the collection of the dataset we refer to Nguyen et al. [2013a] (Chapter 4).

We divided the data into train and test sets. 200 Twitter users were randomly selected from the test set to be included in the online game (statistics are shown in Table 5.1). Named entities were manually anonymized to conceal the user's identity. Names in tweets were replaced by 'similar' names (e.g., a first name common in a certain region in the Netherlands was replaced with another common name in that region). This was done without knowing the actual gender and age of the Twitter users. Links were replaced with a general [LINK] token and user mentions with @USER.

Gender and age	**F, <20**	**M, <20**	**F, [20-40)**	**M, [20-40)**	**F, ≥40**	**M, ≥40**
Frequency	61	60	24	23	17	15

Table 5.1: Statistics Twitter users in our game

5.3.2 Online Game

Game setup. The interface of the game is shown in Figure 5.1. Players guessed the biological sex (male or female) and age (years) of a Twitter user based on only the tweets. Note that the ground truth labels were collected using *all* available information (e.g., profile picture, name, etc.). For each user, {20, 25, 30, 35, 40} tweets were randomly selected. For a particular Twitter user, the same tweets were displayed to all players. Twitter users were randomly selected to be displayed to the players.

To include an entertainment element, players received feedback after each guess. They were shown the correct age and gender, the age and gender guessed by the computer, and the average guessed age and gender distribution by the other players. In addition, a score was shown of the player versus the computer.

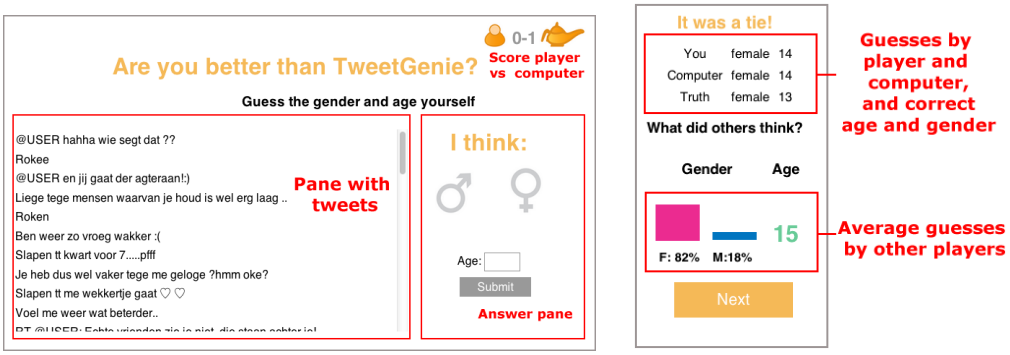


Figure 5.1: Screenshot of the game. Text is translated into English (originally in Dutch). Left shows the interface when the user needs to make a guess. Right shows the feedback interface.

5

Collection. In May 2013, the game was launched. Media attention resulted in a large number of visitors [Nguyen et al., 2014b] (Chapter 4). We use the data collected from May 13, 2013 to August 21, 2013, resulting in a total of 46,903 manual guesses. Players tweeted positively about the game, such as "*@USER Do you know what is really addictive? 'Are you better than Tweetgenie' ...*" and "*@USER Their game is quite fun!*" (tweets translated to English).

We filter sessions that do not seem to contain genuine guesses: when the entered age is 80 years or above, or 8 or below. These thresholds were based on manual inspection, and chosen because it is unlikely that the shown tweets are from users of such ages. For each guess, we registered a session ID and an IP address. A new session started after 2 hours of inactivity. To study player performance more robustly, we excluded multiple sessions of the same player. After three or more guesses had been made in a session, all next sessions from the same IP address were discarded.

Statistics. Statistics of the data are shown in Table 5.2. Figure 5.2 shows the distribution of the number of guesses per session. The longest sessions consisted of 18 guesses. Some of our analyses require multiple guesses per player. In that case, we only include players having made at least 7 guesses.

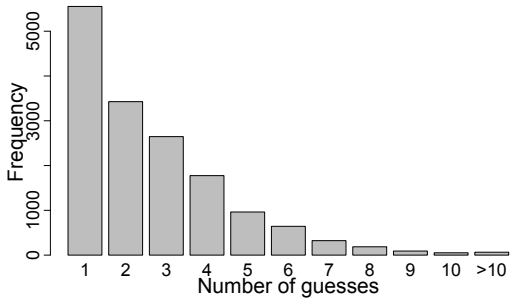
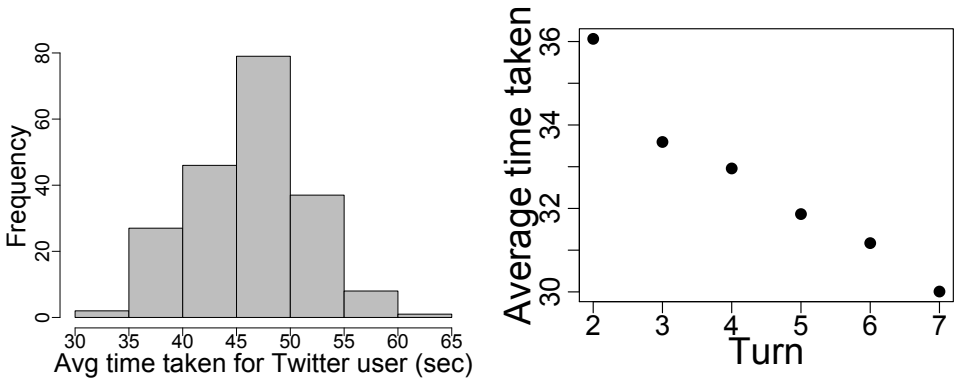


Figure 5.2: Number of guesses per session

# guesses	41,989
# sessions	15,724
Avg. time (sec) per guess	46
Avg. # guesses / session	2.67

Table 5.2: Statistics online game (after cleaning)



(a) Average time taken for Twitter users **(b)** Average time taken per turn

Figure 5.3: Time taken in game

We calculate the time taken for a guess by taking the time difference between two guesses (therefore, no time for the first guess in each session could be measured). For each Twitter user, we calculate the average time that was taken to guess the gender and age of the user. (Figure 5.3a). There is a significant correlation (Pearson's r = 0.291, $p < 0.001$) between the average time the players took to evaluate the tweets of a Twitter user and the number of displayed tweets.

There is also a significant correlation between the average time taken for a user and the entropy over gender guesses (Pearson's r = 0.410, $p < 0.001$), and the average time taken for a user and the standard deviation of the age guesses (Pearson's r = 0.408, $p < 0.001$). Thus, on average, players spent more time on Twitter users for whom it was more difficult to estimate gender and age.

We observe that as the game progresses, players tend to take less time to make a guess. This is shown in Figure 5.3b, which shows the average time taken for a turn (restricted to players with at least 7 guesses). There was no significant correlation between time spent on a guess and the performance of players and we did not find trends of performance increase or decrease as players progressed in the game.

5.3.3 Automatic Prediction

Besides studying human performance, we also compare the predictions of humans with those of an automatic system. We split the data into train and test sets using the same splits as used by Nguyen et al. [2013a] (Chapter 4). We train a logistic regression model to predict gender (male or female), and a linear regression model to predict the age (in years) of a person.

More specifically, given an input vector $\mathbf{x} \in \mathbb{R}^m$, $x_1, \ldots, x_m$ represent features. In the case of gender classification (e.g., $y \in \{-1, 1\}$), the model estimates a conditional distribution $P(y|\mathbf{x}, \beta) = 1/(1 + exp(-y(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta})))$, where β_0 and $\boldsymbol{\beta}$ are the parameters to estimate. Age is treated as a regression problem, and we find a prediction $\hat{y} \in \mathbb{R}$ for the exact age of a person $y \in \mathbb{R}$ using a linear regression model: $\hat{y} = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$. We use Ridge (also called L_2) regularization to prevent overfitting.

We make use of the liblinear [Fan et al., 2008] and scikit-learn [Pedregosa et al., 2011] libraries. We only use unigram features, since they have proven to be very effective for gender [Bamman et al., 2014b, Peersman et al., 2011] and age [Nguyen et al., 2013a] prediction. Parameters were tuned using cross-validation on the training set.

5.4 Gender

Most of the computational work on language and gender focuses on gender classification, treating gender as fixed and classifying speakers into females and males. However, this assumes that gender *is* fixed and is something people have, instead of something people *do* [Butler, 1990].

In this section, we first analyze the *task difficulty* by studying crowd performance on inferring gender from tweets. We observe a relatively large group of Twitter users who employ language that the crowd associates with the *opposite* biological sex. This, then, raises questions about the upper bound that a prediction system based on only text can achieve.

Next, we place Twitter users on a *gender continuum* based on the guesses of the players and show that treating gender as a binary variable is too simplistic. While historically gender has been treated as binary, researchers in fields such as sociology [Lorber, 1996] and sociolinguistics [Bergvall et al., 1996, Holmes and Meyerhoff, 2003] find this view too limited. Instead, we assume the simplest extension beyond a binary variable: a one-dimensional gender continuum (or scale) [Bergvall et al., 1996]. For example, Bergvall [1999] talks about a "*continuum of humans' gendered practices*". While these previous studies were based on qualitative analyses, we take a quantitative approach using the crowd.

5.4.1 Task Difficulty

Majority vote. We study crowd performance using a system based on the *majority* of the players' guesses. Majority voting has proven to be a strong baseline to aggregate votes (e.g., in crowdsourcing systems [Le et al., 2010, Snow et al., 2008]). On average, we have 210 guesses per Twitter user, providing substantial evidence per Twitter user. A system based on majority votes achieves an accuracy of 84% (Table 5.3a shows a confusion matrix). Table 5.3b shows a confusion matrix of the majority predictions versus the automatic system. We find that the biological sex was predicted incorrectly by both the majority vote system and the automatic system for 21 out of the 200 Twitter users (10.5%, not in the tables).

Automatic classification systems on English tweets achieve similar performances as our majority vote system (e.g., Bergsma and Van Durme [2013] report an accuracy of 87%, Bamman et al. [2014b] 88%). More significantly, the results suggest that 10.5% (automatic + majority) to 16% (majority) of the Dutch Twitter users do not employ language that the crowd associates with their biological sex. As said, this raises the question of whether we can expect much higher performances by computational systems based on only language use.

Crowd	Biological sex: Male	Biological sex: Female
Male	82	16
Female	16	86

(a) Crowd (majority)

Automatic	Crowd: Male	Crowd: Female
Male	68	22
Female	30	80

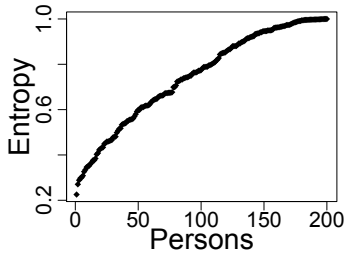
(b) Automatic vs. crowd

Table 5.3: Confusion matrices crowd prediction

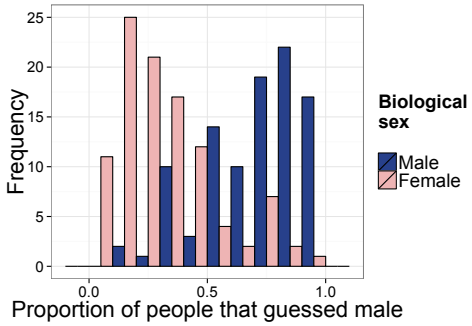
Individual players versus an automatic system. When considering players with 7 or more guesses, the average accuracy for a player is 0.71. Our automatic system achieves an accuracy of 0.69. The small number of tweets per Twitter user in our data (20-40) makes it more difficult to automatically predict gender.

5

Entropy. We characterize the difficulty of inferring a user's gender by calculating the entropy for each Twitter user based on the gender guesses (Figure 5.4a). We find that the difficulty varies widely across users, and that there are no distinct groups of 'easy' and 'difficult' users. However, we do observe an interaction effect between the entropy of the gender guesses and the ages of the Twitter users. At an aggregate level, we find no significant trend. Analyzing females and males separately, we observe a significant trend with females (Pearson's $r = 0.270$, $p < 0.01$), suggesting that older female Twitter users tend to emphasize other aspects than their gender in tweets (as perceived by the crowd).



(a) Entropy over gender guesses



(b) A histogram of all Twitter users and the proportion of players who guessed the users were male. For example, there are 25 female users for which 10 - 20% of the players guessed they were male.

Figure 5.4: Gender prediction

5.4.2 Binarizing Gender, a Good Approach?

Using data collected through the online game we *quantitatively* put speakers on a gender continuum based on how their tweets are perceived by the crowd. For each Twitter user, we calculate the proportion of players who guessed the users were male and female (Figure 5.4b). We can make the following observations.

First, the guesses by the players are based on their expectations about what kind of behavior and language is used by males and females. The plot shows that for some users, almost all players guessed the same gender, indicating that these expectations are quite strong and that there are stylistic markers and topics that the crowd strongly associates with males or females.

Second, if treating gender as a binary variable is reasonable, we would expect to see two distinct groups. However, we observe quite an overlap between the biological males and females. There are 1) users who conform to what is expected based on their biological sex; 2) users who deviate from what is expected; and 3) users whose tweets do not emphasize a gender identity or whose tweets have large variation using language associated with both genders. We investigated whether this is related to their use of Twitter (professional, personal, or both), but the number of Twitter users in our dataset who used Twitter professionally was small and not sufficient to draw conclusions.

We now illustrate our findings using examples. The first example is a 15-year old biological female for whom the crowd guessed most strongly that she is female (96% of n=220). Three tweets from her are shown below. She uses language typically associated with females, talking about spending time with her girlfriends and the use of stylistic markers such as hearts and alphabetical lengthening. Thus, she conforms strongly to what the crowd expects from her biological sex.

Tweet 4: *Gezellig bij Emily en Charlotte.*
Translation: *Having fun with Emily and Charlotte.*

Tweet 5: *Hiiiiii schatjesss!*
Translation: *Hiiiiii cutiesss!*

Tweet 6: ♥ *@USER*

Below are two tweets from a 40-year old biological female who does not employ linguistic markers strongly associated with males or females. Therefore, only 46% of the crowd (n=200) was able to guess that she is female.

Tweet 7: *Ik viel op mijn bek. En het kabinet ook. Geinig toch? #Catshuis*
Translation: *I went flat on my face. And the cabinet as well. Funny right? #Catshuis*

Tweet 8: *Jeemig. Ik kan het bijna niet volgen allemaal.*
Translation: *Jeez. I almost can't follow it all.*

Twitter users vary in how much they emphasize their gender in their tweets. As a result, the difficulty of inferring gender from tweets varies across persons, and treating gender as a binary variable ignores much of the interesting variation within and between persons.

Automatic system. We now analyze whether an automatic system is capable of capturing the position of Twitter users on the gender continuum (as perceived by the crowd). We calculate the correlation between the proportion of male guesses (i.e., the position on the gender continuum) and the scores of the logistic regression classifier: $\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$. While the training data was binary (users were labeled as male or female), a reasonable Spearman correlation of $\rho = 0.584$ ($p < 0.001$) was obtained between the classifier score and the score based on the crowd's perception. We did not observe a significant relation between the score of the classifier (corresponding to the confidence of the gender prediction) and age.

5.5 Age

We start with an analysis of task difficulty, by studying crowd performance on inferring age from tweets. Next, we show that it is particularly hard to accurately infer the chronological age of older Twitter users from tweets.

5.5.1 Task Difficulty

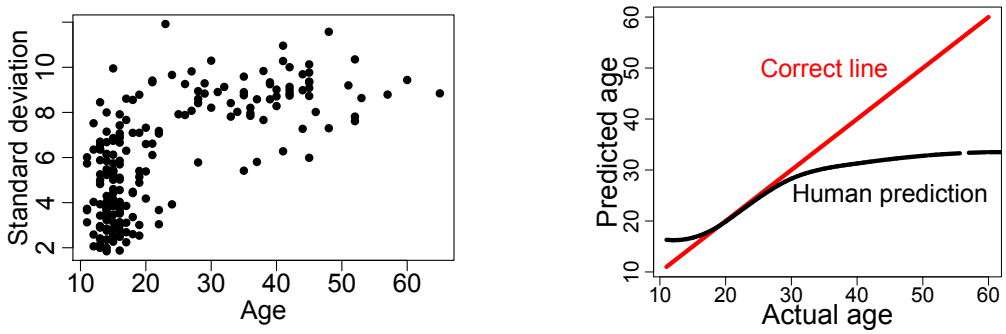
The crowd's average guesses. As with a system based on majority vote for gender prediction, we test the performance of a system that predicts the ages of Twitter users based on the average of all guesses. We find that such a system achieves a Mean Absolute Error (MAE) of 4.844 years and a Pearson's correlation of 0.866. Although the correlation is high, the absolute errors are quite large. We find that the crowd has difficulty predicting the ages of older Twitter users. There is a positive correlation (Pearson's ρ = 0.789) between the absolute errors and the actual age of Twitter users. There is a negative correlation between the errors (predicted - actual age) and the actual age of Twitter users (Pearson's ρ = -0.872).

We calculate the standard deviation over all the age guesses for a user (Figure 5.5a) to measure the difficulty of inferring a user's age. There is a positive correlation between age and standard deviation of the guesses ($\rho = 0.691$), which indicates that players have more difficulty in guessing the ages of older Twitter users.

Individual players versus an automatic system. To estimate the performance of individual players, we restrict our attention to players with at least 7 guesses. We find that individual players are, on average, 5.754 years off. A linear regression system achieves a MAE of 6.149 years and a Pearson correlation of 0.812. The small number of tweets in our data (20-40) increases the difficulty of the task for automatic systems.

5.5.2 Inferring the Age of Older Twitter Users

Figure 5.5b shows the average player predictions with the actual age of the Twitter users. The red line is the 'perfect' line, i.e., the line when the predictions would match the exact age. Black represents a fitted LOESS curve [Cleveland et al., 1992] based on the human predictions. We find that the players tend to overpredict the age of younger Twitter users, but even more strikingly, on average they consistently



(a) Standard deviation and actual age

(b) Average age prediction by humans

Figure 5.5: Age prediction

underpredict the age of older Twitter users. The prediction errors already start at the end of the 20s, and the gap between actual and predicted age increases with age.

This could be explained by sociolinguistic studies that have found that people between 30 and 55 years use standard forms the most, because they experience the maximum societal pressure in the workplace to conform [Holmes, 2013]. On Twitter, this has been observed as well: Nguyen et al. [2013a] found fewer linguistic differences between older age groups than between younger age groups. This makes it difficult for the crowd to accurately estimate the ages of older Twitter users. Younger people and retired people use more non-standard forms [Holmes, 2013]. Unfortunately, our dataset does not contain enough retired users to analyze whether this trend is also present on Twitter.

5.6 Discussion

We now discuss the implications of our findings for research on automatically predicting the gender and age of authors from their texts.

Age and gender as *social* variables. Most computational research has treated gender and age as fixed, biological variables. The dominant approach is to use supervised machine learning methods to generalize across a large number of examples (e.g., texts written by females and males). While the learned models so far are effective at predicting age and gender of *most* people, they learn stereotypical behavior and therefore provide a simplistic view.

First, by using the crowd we have shown that Twitter users emphasize their gender and age in varying degrees and in different ways, so that for example, treating gender as a binary variable is too simplistic [Butler, 1990, Eckert and McConnell-Ginet, 2013]. Many users do not employ the stereotypical language associated with their biological sex, making models that take a static view of gender ineffective for such users. More detailed error analyses of the prediction systems will increase understanding of the reasons for incorrect predictions, and shed light on the relation between language use and social variables.

Second, models that assume static variables will not be able to model the interesting variation [Eisenstein, 2013a]. Models that build on recent developments in sociolinguistics will be more meaningful and will also have the potential to contribute to new sociolinguistic insights. For example, modeling what influences speakers to show more or less of their identity through language, or jointly modeling variation between and within speakers, are in our opinion interesting research directions. The ever-increasing amounts of social media data offer opportunities to explore these research directions.

Sampling. We have shown that the difficulty of tasks such as gender and age prediction varies across persons. Therefore, creating datasets for such tasks requires maximum attention. For example, when a dataset is biased towards people who show a strong gender identity (e.g., by sampling followers of accounts highly associated with males or females, such as sororities [Rao et al., 2010]), the results obtained on such a set may not be representative of a more random set (as observed when classifying political affiliation [Cohen and Ruths, 2013]).

Task difficulty. Our study also raises the question of what level of performance can be obtained for tasks such as predicting gender and age from only language use. Since we often form an impression based on someone's writing, crowd performance is a good indicator of the task difficulty. While the crowd performance does not need to be the upper bound, it does indicate that it is difficult to predict gender and age of a large number of Twitter users.

When taking the majority label, only 84% of the users were correctly classified according to their biological sex. This suggests that about 16% of the Dutch Twitter users do not use language that the *crowd* associates with their biological sex.

We also found that it is hard to accurately estimate the ages of older Twitter users, and we related this to sociolinguistics studies that found less linguistic differences in older age groups due to societal pressure in the workplace.

Limitations. A limitation of our work is that we focused on language variation *between* persons, and not on variation *within* persons. However, speakers vary their language depending on the context and their conversation partners (e.g., accommodation effects were found in social media [Danescu-Niculescu-Mizil et al., 2011]). For example, we assigned Twitter users an overall 'score' by placing them on a gender continuum, ignoring the variation we find within users.

Crowdsourcing as a tool to understand NLP tasks. Most research on crowdsourcing within the NLP community has focused on how the crowd can be used to obtain fast and large amounts of annotations. This study is an example of how the crowd can be used to obtain a deeper understanding of an NLP task. We expect that other tasks where disagreement between annotators is meaningful (i.e., it is not only due to noise), could potentially benefit from crowdsourcing experiments as well.

5.7 Conclusion

In this chapter, we demonstrated the successful use of the crowd to study the relation between language use and social variables. In particular, we took a closer look at inferring gender and age from language using data collected through an online game. We showed that treating gender and age as fixed variables ignores the variety of ways people construct their identity through language.

Approaching age and gender as *social* variables will allow for richer analyses and more robust systems. It has implications ranging from how datasets are created to how results are interpreted. We expect that our findings also apply to other social variables, such as ethnicity and status. Instead of only focusing on performance improvement, we encourage NLP researchers to also focus on what we can *learn* about the relation between language use and social variables using computational methods.

6

A Kernel Independence Test for Geographical Language Variation

This chapter is based on D. Nguyen and J. Eisenstein, "A Kernel Independence Test for Geographical Language Variation" to appear in Computational Linguistics and presented at New Ways of Analyzing Variation 44 (NWAV44), 2015.

6.1 Introduction

Figure 6.1 shows the geographical location of 1000 Twitter posts containing the word *hella*, an intensifier used in expressions like *I got hella studying to do* and *my eyes got hella big* [Eisenstein et al., 2014]. While the word appears in major population centers throughout the United States, the map suggests that it enjoys a particularly high level of popularity on the west coast, in the area around San Francisco. But does this represent a real geographical difference in American English, or is it the result of chance fluctuation in a finite dataset?

Regional variation of language has been extensively studied in sociolinguistics and dialectology [Chambers and Trudgill, 1998, Grieve et al., 2011, 2013, Lee and Kretzschmar, 1993, Nerbonne and Kretzschmar Jr, 2013, Szmrecsanyi, 2012]. A common approach involves mapping the geographic distribution of a linguistic variable (e.g., the choice of *soda*, *pop*, or *coke* to refer to a soft drink) and identifying boundaries between regions based on the data. The identification of linguistic variables that exhibit regional variation is therefore the first step in many studies of regional dialects. Traditionally, this step has been based on the manual judgment of the researcher; depending on the quality of the researcher's intuitions, the most interesting or important variables might be missed.

The increasing amount of data available to study dialectal variation suggests a turn towards data-driven alternatives for variable selection. For example, researchers can mine social media data such as Twitter [Doyle, 2014, Eisenstein et al., 2010, Huang et al., 2016] or product reviews [Hovy et al., 2015] to identify and test thousands of dialectal variables. Despite the large scale of available data, the well-known "long tail" phenomenon of language ensures that there will be many potential variables with low

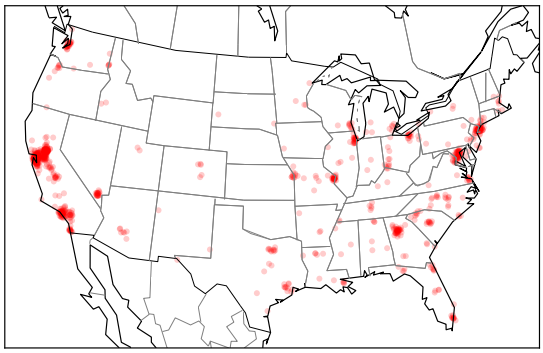


Figure 6.1: 1000 geolocated tweets containing the word *hella*

counts. A statistical metric for comparing the strength of geographical associations across potential linguistic variables would allow linguists to determine whether finite geographical samples — such as the one shown in Figure 6.1 — reveal a statistically meaningful association.

The use of statistical methods to analyze spatial dependence has been only lightly studied in sociolinguistics and dialectology. Existing approaches employ classical statistics such as Moran's I (e.g., Grieve et al. [2011]), join count analysis (e.g., Lee and Kretzschmar [1993]) and the Mantel Test (e.g., Scherrer [2012]); we review these statistics in Section 6.2. These approaches suffer from a common problem: each type of test can capture only a specific parametric form of spatial linguistic variation. As a result, these tests can incorrectly fail to reject the null hypothesis if the nature of the geo-linguistic dependence does not match the underlying assumptions of the test.

To address these limitations, we propose a new test statistic that builds on a rich and growing literature on kernel embeddings for non-parametric statistics [Shawe-Taylor and Cristianini, 2004]. In these methods, probability distributions, such as the distribution over geographical locations for each linguistic variable, are embedded in a Reproducing Kernel Hilbert space (RKHS). Specifically, we employ the Hilbert-Schmidt Independence Criterion (HSIC; Gretton et al. [2005a]). Due to its ability to compare arbitrarily high-order moments of probability distributions, HSIC can be used to compare arbitrary probability measures, by computing kernel functions on finite samples. Unlike prior approaches, HSIC is statistically consistent: in the limit of a sufficient amount of data, it will correctly determine whether the distribution of a linguistic feature is geographically dependent. As a further convenience, because it is built on kernel similarity functions, HSIC can be applied with equal ease to any type of linguistic data, as long as an appropriate kernel function can be constructed.

To validate this approach, we compare it against three alternative spatial statistics: Moran's I, the Mantel test, and join count analysis. For a controlled comparison, we use synthetic data to simulate different types of regional variation, and different types of linguistic data. This allows us to measure the capability of each approach to detect true geo-linguistic associations, and to avoid Type I errors even in noisy data. Next, we apply these approaches to three real linguistic datasets: a corpus of Dutch tweets, a Dutch syntactic atlas and letters to the editor in North American newspapers.

To summarize, the contributions of this article are:

- We show how the Hilbert-Schmidt Independence Criterion can be applied to linguistic data. HSIC is a non-parametric test statistic, which can handle both frequency and categorical data, requires no discretization of geographic data, and is capable of detecting arbitrary geo-linguistic dependencies (Section 6.3).
- We use synthetic data to compare the power and calibration of HSIC against three alternatives: Moran's I, the Mantel Test, and join count analysis (Section 6.4).
- We apply these methods to analyze dialectal variation in three empirical datasets, in both English and Dutch, across a variety of registers (Section 6.5).

6.2 Related Work

This section describes prior work on global methods for quantifying the degree of spatial dependence in a geotagged corpus.[1] While other global spatial statistics exist, we focus on the following three methods because they have been used in previous work on dialect analysis: Moran's I [Grieve et al., 2011], join count analysis [Lee and Kretzschmar, 1993] and the Mantel test [Scherrer, 2012].

We define a consistent notation across methods. Let x_i represent a scalar linguistic observation for unit $i \in \{1 \dots n\}$ (typically, the presence or frequency of a linguistic variable), and let y_i represent a corresponding geolocation. For convenience, we define d_{ij} as the spatial distance between y_i and y_j. Suppose we have n observations, so that the data $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. Our goal is to test the strength of association between X and Y, against the null hypothesis that there is no association.

6.2.1 Moran's I

Grieve et al. [2011] introduced the use of Moran's I [Cliff and Ord, 1981, Moran, 1950] in the study of dialectal variation. To define the statistic, let $W = \{w_{ij}\}_{i,j \in \{1 \dots n\}}$ represent a *spatial weighting* matrix, such that larger values of w_{ij} indicate greater proximity, and $w_{ii} = 0$. In their application of Moran's I to a corpus of newspaper letters-to-the-editor, Grieve *et al.* define W as,

$$w_{ij} = \begin{cases} 1, & d_{ij} < \tau, i \neq j \\ 0, & d_{ij} \geq \tau, \text{ or } i = j \end{cases} \tag{6.1}$$

[1] Global methods test for dependence over the entire dataset. In some cases, there will be local dependence in a few "hot spots", even when global dependence is not detected, and local autocorrelation statistics have been proposed to capture such dependences [Anselin, 1995]. For example, Grieve [2016] uses the Getis-Ord G_i statistic [Getis and Ord, 1992] in his analysis of regional American English. Local statistics are particularly useful as an exploratory tool, but Grieve argues that the associated p-values are difficult to interpret due to the issue of multiple comparisons. We therefore focus on global tests in this chapter. The adaptation of the proposed HSIC statistic into a local measure of dependence is an intriguing topic for future work.

where τ is some critical threshold [Grieve et al., 2011]. When the spatial weighting matrix is defined in this way, Moran's I can be seen as a statistic that quantifies whether observations x_i and x_j are more similar when $w_{ij} = 1$ than when $w_{ij} = 0$.[2]

Moran's I is based on a hypothesized autoregressive process $X = \rho W X + \epsilon$, where X is a vector of the linguistic observations $x_1, \ldots x_n$, and ϵ is a vector of uncorrelated noise. Since X and W are given, the estimation problem is to find ρ so as to minimize the magnitude of ϵ. To take a probabilistic interpretation, it is typical to assume that ϵ consists of independent and identically-distributed (IID) normal random variables with zero mean [Ord, 1975]. Under the null hypothesis of no spatial dependence between the observations in X, we would have $\rho = 0$. Note, however, that we may fail to reject the possibility that $\rho = 0$ even in the presence of spatial dependence, if the form of this dependence is not monotonic or nonlinear in W.

Because ρ is difficult to estimate exactly [Ord, 1975], Moran's I is used as an approximation. It is computed as,

$$I = \frac{n}{\sum_i^n (x_i - \overline{x})^2} \frac{\sum_i^n \sum_j^n w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_i^n \sum_j^n w_{ij}}, \tag{6.2}$$

where $\overline{x} = \frac{1}{n}\sum_i x_i$. The ratio on the left is the inverse of the variance of X; the ratio on the right corresponds to the covariance between points i and j that are spatially similar. Thus, the statistic rescales a spatially-reweighted covariance (the ratio on the right of Equation 6.2) by the overall variance (the ratio on the left of Equation 6.2), giving an estimate of the overall spatial dependence of X. A compact alternative notation is to rewrite the statistic in terms of the vector of *residuals* $R = \{r_i\}_{i \in 1 \ldots n}$, where the residual $r_i = x_i - \overline{x}$. This yields the form $I = \frac{R^\top W R}{R^\top R}$, with $R^\top$ indicating the transpose of the column vector R, and with W assumed to be normalized so that $\sum_{i,j} w_{ij} = n$. Moran's I values often lie between -1 and 1, but the exact range depends on the weight matrix W, and is theoretically unbounded [de Jong et al., 1984].

In hypothesis testing, our goal is to determine the p-value representing the likelihood that a value of Moran's I at least as extreme as the observed value would arise by chance under the null hypothesis. The expected value of Moran's I in the case of no spatial dependence is $-\frac{1}{n-1}$. Grieve *et al.* compute p-values from a closed-form approximation of the variance under the null hypothesis of total randomization. A non-parametric alternative is to perform a *permutation test*, calculating the empirical p-value by comparing the observed test statistic against the values that arise across multiple random permutations of the original data.

In either case, Moran's I does not test the null hypothesis of no statistical dependence between the linguistic features X and the geo-coordinates Y. Rather, it tests whether the estimated value of ρ would be likely to arise if there were no such dependence. But if the nature of the geo-linguistic dependence defies the assumptions of the statistic, then we risk incorrectly failing to reject the null hypothesis, a type II

[2]The matrix W can be defined in other ways. We can define a continuous-valued version of W by setting $w_{ij} = \exp(-\gamma d_{ij})$, with d_{ij} equal to the geographical distance between units i and j. Alternatively, we could define a topological spatial weighting matrix by setting $w_{ij} = 1$ when j is one of the k nearest neighbors of i [Getis and Aldstadt, 2010].

error. Put another way, there are forms of strong spatial dependence for which $\rho = 0$, such as non-monotonic spatial relationships. This risk can be somewhat ameliorated by careful choice of the spatial weighting matrix W, which could in theory account for non-linear or even non-monotonic dependencies. However an exhaustive search for some W that obtains a low p-value would clearly be an invalid hypothesis test, and so W must be fixed *before* any test is performed. In some cases, the researcher may bring substantive insights to the determination of W, and so the flexibility of Moran's I in this sense could be regarded as a positive feature. But there is little theoretical guidance, and a poor selection of W will result in inflated type II error rates.

From a practical standpoint, Moran's I is applicable to only some types of linguistic data. In the study of dialect, X typically represents the frequency or presence of some linguistic variable, such as the use of *soda* versus *pop*. We are unaware of applications of Moran's I to variables with more than two possibilities (e.g., *soda*, *pop*, *coke*). One possible solution, proposed by one of the reviewers, would be to perform multiple tests, with each alternant pitted against all the others. But it is not clear how the p-values from these multiple tests should be combined. For example, selecting the minimum p-value across the alternants would mean that the null hypothesis would be more likely to be rejected for variables with more alternants; averaging the p-values across alternants would have the opposite problem.

6.2.2 Join Count Analysis

If the linguistic data X consists of discrete observations, a simple approach is the use of *join count analysis*. For each pair of points (i, j), we compute $w_{ij}\delta(x_i = x_j)$, where $\delta(x_i = x_j)$ returns a value of 1 if x_i and x_j are identical, and 0 otherwise. As in Moran's I, w_{ij} is an element of a spatial weighting matrix, which could be binary or continuous. The global sum of the counts is computed as,

$$\text{num-agree} = \sum_i^n \sum_j^n w_{ij}(x_i x_j + (1 - x_i)(1 - x_j)) \tag{6.3}$$

$$= X^\top W X + (1 - X)^\top W (1 - X), \tag{6.4}$$

with $X^\top$ indicating the transpose of the column vector X. Note the similarity to the numerator of Moran's I, which can be written as $R^\top W R$. The number of agreements can be compared with its expectation under the null hypothesis, yielding a hypothesis test for global autocorrelation [Cliff and Ord, 1981].

Join count analysis has been applied to the study of dialect by Lee and Kretzschmar [1993], who take each linguistic observation $x_i \in \{0, 1\}$ to be a binary variable indicating the presence or absence of a dialect feature. They then build a binary spatial weighting matrix by performing a Delaunay triangulation over the geolocations of participants in dialect interviews, with $w_{ij} = 1$ if the edge (i, j) appears in the Delaunay triangulation. A nice property of Delaunay triangulation is that points tend to be connected to their closest neighbors, regardless of how distant or near those neighbors are: in high-density regions, the edges will tend to be short, while in low-density regions, the edges will be long. The method is therefore arguably more suitable to

data in which the density of observations is highly variable — for example, between densely-populated cities and sparse-populated hinterlands.

Because join count statistics are based on counts of agreements, this form of analysis requires that each x_i is a categorical variable — possibly non-binary — rather than a frequency. In this sense, it is the complement of Moran's I, which can be applied to frequencies, but not to non-binary variables. Thus, join count analysis is best suited to cases where observations correspond to individual utterances (e.g., Twitter data, dialect interviews), rather than cases where observations correspond to longer texts (e.g., newspaper corpora).

6.2.3 The Mantel Test

The Mantel test can in principle be used to measure the dependence between any two arbitrary signals. In this test, we compute *distances* for each pair of linguistic variables, $d_x(x_i, x_j)$, and each pair of spatial locations, $d_y(y_i, y_j)$, forming a pair of distance matrices D_x and D_y. We then estimate the element-wise correlation (usually, the Pearson correlation) between these two matrices. Scherrer [2012] uses the Mantel test to correlate linguistic distance with geographical distance, and Gooskens and Heeringa [2006] correlate perceptual distance with linguistic distance. The Mantel test has also been applied to non-human dialect analysis, revealing regional differences in the call structures of Amazonian parrots by computing a linguistic distance matrix D_x directly from spectral measurements [Wright, 1996].

Because it is built around distance functions, the Mantel test is applicable to binary, categorical, and frequency data — any kind of data for which a distance function can be constructed. For spatial locations, a typical choice is to compute the distance matrix based on the Euclidean distance between each pair of points. For binary or categorical linguistic data, the entries of the linguistic distances matrix can be set to 0 if $x_i = x_j$, and 1 otherwise. For linguistic frequency data, we use the absolute difference between the frequency values.

The role of hypothesis testing in the Mantel test is to determine the likelihood that the observed test statistic — in this case, the correlation between the distance matrices D_x and D_y — could have arisen by chance under the null hypothesis. In the ideal case of perfect correlation, twice as much geographical distance should imply twice as much as linguistic distance. But this situation is highly unlikely to obtain in practice. In fact, as one of the reviewers noted, such a perfect correlation *cannot* arise from any linguistic data involving a single variable: even if a linguistic variable obeys a perfect dialect continuum (e.g., varying in frequency from east to west), the distances in the orthogonal north-south direction would diminish the resulting correlation. In realistic settings in which the geo-linguistic dependence is obscured by noise, this can dramatically diminish the power of the test. Note that even non-linear transformations of the distance metric would not correct this issue. The key problem, as identified by Legendre et al. [2015], is that the Mantel test is not designed to test for independence between X and Y, but rather, the correlation between *distances* on X and Y. When distances are the primary units of analysis — as, for example, in the work of Gooskens and Heeringa [2006] — the test is applicable. But for the task of determining whether a specific linguistic variable is geographically dependent, the

test is incorrectly applied; as we show in Section 6.4, this results in inflated Type II error rates.

6.2.4 Other Related Work

Several computational studies attempt to characterize linguistic differences across geographical regions, although in general these studies do not perform hypothesis testing on geographical dependence. In general, these studies rely on aggregating geotagged social media content into geographical bins. Some studies rely on politically defined units such as nations and states [Hovy et al., 2015]; however, *isoglosses* (the geographical boundaries between linguistic features) need not align with politically-defined geographical units [Nerbonne and Kretzschmar Jr, 2013]. Other approaches rely on automatically defined geographical units, induced by computational methods such as geodesic grids [Wing and Baldridge, 2011], KD-trees [Roller et al., 2012], Gaussians [Eisenstein et al., 2010], and mixtures of Gaussians [Hong et al., 2012]. While these approaches offer insight about the nature of geographical language variation, they do not provide test statistics that allow us to quantify the geographical dependence of various linguistic features.

As described in the next section, our approach is based on reproducing kernel Hilbert spaces, which enable us to non-parametrically compare probability distributions. Another way in which kernel methods can be applied to spatial analysis is in Gaussian Processes, which are often used to represent spatial data [Cressie, 1988, Ecker and Gelfand, 1997]. Specifically, we can define a kernel over space, so that a response variable is distributed as a Gaussian with covariance defined by the kernel function. For example, it might be possible to model the popularity of linguistic features as a Gaussian Process, using the spatial covariance kernel to make smooth predictions at unknown locations. Our approach in this chapter is different, as we are interested in hypothesis testing, rather than modeling and prediction. Another difference is that we apply kernels to both the geographical and linguistic data sources, while a Gaussian Process approach would make the parametric assumption that the linguistic signal is Gaussian distributed with covariance defined by the spatial covariance kernel.

6.3 Hilbert-Schmidt Independence Criterion (HSIC)

Moran's I, join count analysis, and the Mantel test share an important drawback: they do not directly test the independence of language and geography, but rather, they test for autocorrelation between X and Y, or between distances on these variables. Moran's I tests whether the parameter of a linear autoregressive model is nonzero; the Mantel test is performed on the correlations between pairwise distances; join count statistics enable tests of whether spatially adjacent units tend to have the same linguistic features. In each case, rejection of the null hypothesis implies dependence between the geographical and linguistic signals. However, each test can incorrectly fail to reject the null hypothesis if its assumptions are violated, even if given an arbitrarily large amount of data.

We propose an alternative approach: directly test for the independence of geographical and linguistic variables, $P_{XY} \stackrel{?}{=} P_X P_Y$. Our approach, which is based on the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2005a, 2008], makes no parametric assumptions about the form of these distributions. The proposed test is *consistent*, in the sense that it will always reach the right decision, if provided enough data [Fukumizu et al., 2007].[3]

To test independence for arbitrary distributions P_{XY}, P_X, and P_Y, HSIC employs the framework of Reproducing Kernel Hilbert spaces (RKHS). This framework will be familiar to the computational linguistics community through its application to support vector machines [Collins and Duffy, 2001, Lodhi et al., 2002], where *kernel similarity functions* between pairs of instances are used to induce classifiers with highly non-linear decision boundaries. HSIC is a kernelized independence test, and it offers an analogous advantage: by computing kernel similarity functions on pairs of observations, it is possible to implicitly compare probability distributions across high-order moments, enabling non-parametric independence tests that are statistically consistent. An additional advantage of the RKHS framework is that it can be applied to arbitrary linguistic data — including dichotomous, polytomous, continuous, and vector-valued observations — as long as an appropriate kernel similarity function can be defined.

6.3.1 Comparing Probability Distributions

The *maximum mean discrepancy* (MMD) is a non-parametric statistic that compares two arbitrary probability distributions. In the HSIC test, this statistic is used to compare the joint distribution P_{XY} with the product of marginal distributions $P_X P_Y$. The MMD is defined as,

$$\text{MMD}(P, Q) = \sup_f (\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]), \tag{6.5}$$

where we take the supremum f over a set of possible functions, and compute the difference in the expected values under the distributions P and Q. Clearly, if $P = Q$, then $\text{MMD} = 0$, but the challenge is to compute MMD for arbitrary P and Q, based only on finite samples from these distributions.

To explain how to do this, we introduce some concepts from RKHS. Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+$ denote a kernel function, mapping from pairs of observations (x_i, x_j) to non-negative reals. A classical example is the radial basis function (RBF) kernel on vectors, where $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||_2^2\right)$. Many other kernel functions are possible; the conditions for valid kernels are enumerated in Subsection 6.3.3.

For any instance x, the kernel function k defines a corresponding *feature map* $k(\cdot, x) : \mathcal{X} \mapsto \mathbb{R}_+$, which is simply the function that arises by fixing one of the arguments of the kernel function to the value x. The "reproducing" property of RKHS implies an identity between kernel functions and inner products of feature maps:

$$k(x_i, x_j) = \langle k(\cdot, x_i), k(\cdot, x_j) \rangle. \tag{6.6}$$

[3] HSIC was introduced as a test of independence by Gretton et al. [2008]. In this chapter, we present the first application to computational linguistics.

Thus, even though the feature map may be an arbitrarily complex function of x, we can compute inner products of feature maps by computing the kernel similarity function over the associated instances. The MMD can be expressed in terms of such inner products, and thus, can be computed in terms of kernel similarity functions.

For a probability measure P, the *mean element* of P is defined as the expected feature map, $\mu_P = \mathbb{E}_P k(\cdot, x)$. The MMD can then be computed in terms of kernel functions of the mean elements,

$$\text{MMD}^2(P, Q) = \langle \mu_P, \mu_P \rangle + \langle \mu_Q, \mu_Q \rangle - 2\langle \mu_P, \mu_Q \rangle. \tag{6.7}$$

If $\mu_P = \mu_Q$, then the MMD is zero. The key observation is that $\mu_P = \mu_Q$ if and only if $P = Q$, so long as an appropriate kernel similarity function is chosen [Fukumizu et al., 2007]; see Subsection 6.3.3 for more on the choice of kernel functions.

Each of the inner products in Equation 6.7 corresponds to an expectation that can be estimated empirically from finite samples $\{x_1, x_2, \ldots, x_m\}$ and $\{y_1, y_2, \ldots, y_n\}$,

$$\text{MMD}^2(P, Q) = \mathbb{E}_{x,x' \sim P} k(x, x') + \mathbb{E}_{y,y' \sim Q} k(y, y') - 2\mathbb{E}_{x \sim P, y \sim Q} k(x, y) \tag{6.8}$$

$$\widehat{\text{MMD}^2}(P, Q) = \frac{1}{m^2} \sum_{i,j}^{m} k(x_i, x_j) + \frac{1}{n^2} \sum_{i,j}^{n} k(y_i, y_n) - \frac{2}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j). \tag{6.9}$$

The full derivation is provided by Gretton et al. [2008]. Having shown how to estimate a statistic on whether two probability measures are identical, we now use this statistic to test for independence.

6.3.2 Derivation of HSIC

To construct an independence test over random variables X and Y, we test the MMD between the joint distribution P_{XY} and the product of marginals $P_X P_Y$. In this setting, each observation i corresponds to a pair (x_i, y_i), so we require a kernel function on paired observations, $k((x_i, y_i), (x_j, y_j))$. We define this as a *product kernel*,

$$k((x_i, y_i), (x_j, y_j)) = k_{\mathcal{X}}(x_i, x_j) k_{\mathcal{Y}}(y_i, y_j), \tag{6.10}$$

where $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are kernels for the linguistic and geographic observations respectively.

Using the product kernel, we can define mean embeddings for the distributions P_{XY} and $P_X P_Y$, enabling the application of the MMD estimator from Equation 6.9. The Hilbert-Schmidt Independence Criterion (HSIC) is precisely this application of maximum mean discrepancy to compare the joint distribution against the product of marginal distributions.

Concretely, let us define the *Gram matrix* K_x so that $(K_x)_{i,j} = k_{\mathcal{X}}(x_i, x_j)$ for all pairs i, j in the sample. Analogously, $(K_y)_{i,j} = k_{\mathcal{Y}}(y_i, y_j)$. Then the HSIC can be estimated from a finite sample of m observations as

$$\widehat{\text{HSIC}} = \frac{1}{n^2} \sum_{i,j}^{m} (K_x)_{i,j} (K_y)_{i,j} + \frac{1}{n^4} \sum_{i,j,q,r}^{m} (K_x)_{i,j} (K_y)_{q,r} - \frac{2}{n^3} \sum_{i,j,q}^{m} (K_x)_{i,j} (K_y)_{i,q} \tag{6.11}$$

$$= \frac{\mathbf{tr} K_{\mathcal{X}} H K_{\mathcal{Y}} H}{n^2}, \tag{6.12}$$

where $\mathbf{tr}$ indicates the matrix trace and H is a centering matrix, $H = \mathbb{I}_m - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. With this definition of H, we have,

$$(K_\mathcal{X} H)_{ij} = k_\mathcal{X}(x_i, x_j) - \frac{1}{n}\sum_{j'} k_\mathcal{X}(x_i, x_{j'}) \tag{6.13}$$

$$(K_\mathcal{Y} H)_{ij} = k_\mathcal{Y}(y_i, y_j) - \frac{1}{n}\sum_{j'} k_\mathcal{Y}(y_i, y_{j'}). \tag{6.14}$$

These two terms can be seen as mean-centered Gram matrices. By computing the trace of their matrix product, we obtain a cross-covariance between the Gram matrices. This trace is directly proportional to the maximum mean discrepancy between P_{XY} and $P_X P_Y$. If X and Y are dependent, then large values of $k_\mathcal{X}(x_i, x_j)$ will imply large values of $k_\mathcal{Y}(y_i, y_j)$ — similar geography implies similar language — and so the cross-covariance will be greater than zero. If X and Y are independent, then large values of $k_\mathcal{X}(x_i, x_j)$ are not any more likely to correspond to large values of $k_\mathcal{Y}(y_i, y_j)$, and so the expectation of this cross-covariance will be zero.

6

6.3.3 Kernel Functions

To apply HSIC to the problem of detecting geo-linguistic dependence, we must define the kernel functions $k_\mathcal{X}$ and $k_\mathcal{Y}$. In the RKHS framework, valid kernel functions must be symmetric and positive semi-definite. To ensure consistency of the kernel-based estimator for MMD, the kernel must also be *characteristic*, meaning that it induces an injective mapping between probability measures and their corresponding mean elements [Fukumizu et al., 2007]: each probability measure P must correspond to a single unique mean element μ_P. Muandet *et al.* elaborate these and other properties of several well known kernels [Muandet et al., 2016, Table 3.1].

For the spatial kernel $k_\mathcal{Y}$, we employ a Gaussian radial basis function (RBF), which is a widely used choice for vector data. Specifically, we define $k_\mathcal{Y}(y_i, y_j; \gamma) = \exp(-\gamma d_{i,j}^2)$, where d_{ij}^2 is the squared Euclidean distance between y_i and y_j, and γ is a parameter of the kernel function. We also employ the RBF in $k_\mathcal{X}$ when the linguistic observations take on continuous values, such as frequencies or phonetic data. The RBF kernel is symmetric, positive semi-definite, and characteristic.[4]

The parameter γ corresponds to the "length-scale" of the kernel. Intuitively, as this parameter increases, the kernel similarity drops off more quickly with distance. In this chapter, we follow the popular heuristic of setting γ to the median of the data $(y_1, \ldots y_n)$, as proposed by Gretton et al. [2005b]. We empirically test the sensitivity of HSIC to this parameter in Section 6.4. More recent work offers optimization-based approaches for setting this parameter [Gretton et al., 2012], but we do not consider this possibility here.

[4]Flaxman recently proposed a "kernelized Mantel test", in which correlations are taken between kernel similarities rather than distances [Flaxman, 2015]. The resulting test statistic is similar, but not identical to HSIC. Specifically, while HSIC centers the kernel matrix against the local mean kernel similarities for each point, the kernelized Mantel test centers against the global mean kernel similarity. This makes the test more sensitive to distant outliers. We implemented the kernelized Mantel test, and found its performance to be similar to the classical Mantel test, with lower statistical power than HSIC. Flaxman made similar observations in his analysis of the spatiotemporal distribution of crime events.

Linguistic data is often binary or categorical. In this case, we use a Delta kernel (also sometimes called a Dirac kernel). This kernel is simply defined as $k_{\mathcal{X}}(x_i, x_j) = 1$ if $x_i = x_j$ and 0 otherwise. The Delta kernel has been used successfully in combination with HSIC for high-dimensional feature selection [Song et al., 2012, Yamada et al., 2014], and is symmetric, positive semi-definite, and characteristic for discrete data.

6.3.4 Scalability

The size of each Gram matrix is the square of the number of observations. For large datasets, this will be too expensive to compute. Following Gretton et al. [2005a], we employ a low-rank approximation to each Gram matrix, using the incomplete Cholesky decomposition [Bach and Jordan, 2002]. Specifically, we approximate the symmetric matrices $K_{\mathcal{X}}$ and $K_{\mathcal{Y}}$ as low-rank products, $K_{\mathcal{X}} \approx AA^T$ and $K_{\mathcal{Y}} \approx BB^T$, where $A \in \mathbb{R}^{n \times r_A}$ and $B \in \mathbb{R}^{n \times r_B}$. The approximation quality is determined by the parameters r_A and r_B, which are set to ensure that the magnitudes of the residuals $K - AA^T$ and $L - BB^T$ are below a predefined threshold. HSIC may then be approximated as:

$$\widehat{\text{HSIC}} = \frac{\mathbf{tr} K_{\mathcal{X}} H K_{\mathcal{Y}} H}{n^2}, \tag{6.15}$$

$$\approx \frac{\mathbf{tr}(AA^T)H(BB^T)H}{n^2}, \tag{6.16}$$

$$= \frac{\mathbf{tr}(B^T(HA))(B^T(HA))^T}{n^2} \tag{6.17}$$

where the matrix product HA can be computed without explicitly forming the $n \times n$ matrix H, due to its simple structure. Alternative methods for scaling the computation of HSIC are discussed in a recent note by Zhang et al. [2016].

6.4 Synthetic Data

Real linguistic datasets lack ground truth about which features are geographically distinct, making it impossible to use such data to quantitatively evaluate the proposed approaches. We therefore use synthetic data to compare the power and sensitivity of the various approaches described above. Our main goals are: (1) to calibrate the p-values produced by each approach in the event that the null hypothesis is true, using completely randomized data; (2) to test the power of each approach to capture spatial dependence, particularly under conditions in which the spatial dependence is obscured by noise.

We compare HSIC with specific instantiations of the methods described in Section 6.2, focusing on previous published applications of these methods to dialect analysis. Specifically, we consider the following methods:

Moran's I. We follow Grieve et al. [2011], using a binary spatial weighting matrix with a distance threshold τ, usually set to the median of the distances between points in the dataset. This method is not applicable to categorical data.

Join counts. We follow the approach of Lee and Kretzschmar [1993], who define a binary spatial weighting matrix from a Delaunay triangulation, and then compute join counts for linked pairs of observations. This method is not applicable to frequency data.

Mantel test. We use Euclidean distance for the geographical distance matrix. For continuous linguistic data, we also use Euclidean distance; for discrete data, we use a delta function.

For all approaches, a one-tailed significance test is appropriate, since in nearly all conceivable dialectological scenarios we are testing only for the possibility that geographically proximate units are *more* similar than they would be under the null hypothesis. For some methods, it is possible to calculate a p-value from the test statistic using a closed form estimate of the variance. However, for consistency, we employ a permutation approach to characterize the null distribution over the test statistic values. We permute the linguistic data x, breaking any link between geography and the language data, and then compute the test statistics for many such permutations.

6.4.1 Data Generation

To ensure the verisimilitude of our synthetic data, we target the scenario of geo-tagged tweets in the Netherlands. For each municipality i, we stochastically determine the number and location of the tweets as follows:

Number of data points. For each municipality, the number of tweets n_i is chosen to be proportional to the population, as estimated by Statistics Netherlands (CBS). Specifically, we draw $\tilde{n}_i \sim \text{Poisson}(\mu_{obs} * \text{population}_i)$ and then set $n_i = \tilde{n}_i + 1$, ensuring that each municipality has at least one data point. The parameter μ_{obs} controls the frequency of the linguistic variable. For example, a common orthographic variable (e.g., "g-deletion") might have a high value of μ_{obs}, while a rare lexical variable (e.g., *soda* versus *pop*) might have a much lower value. Note that μ_{obs} is shared across all municipalities.

Locations. Next, for each tweet t, we determine the location y_t by sampling without replacement from the set of real tweet locations in municipality i (the dataset is described in Subsection 6.5.3). This ensures that the distribution of geo-locations in the synthetic data matches the real geographical distribution of tweets, rather than drawing from a parametric distribution which may not match the complexity of true geographical population distributions. Each location is represented as a latitude and longitude pair.

For each variable, each municipality is assigned a frequency vector θ_i, indicating the relative frequency of each variable form: e.g., 70% *soda*, 30% *pop*. We discuss methods for setting θ_i below, which enable the simulation of a range of dialectal phenomena.

We simulate both counts data and frequency data. In counts data — such as geotagged tweets — the data points in each instance in municipality i are drawn from a binomial or multinomial distribution with parameter θ_i. In frequency data,

(a) Angle: 0 degrees (east to west)

(b) Angle: 120 degrees

Figure 6.2: Synthetic frequency data with dialect continua in two different angles

we observe only the relatively frequency of each variable form for each municipality. In this case, we draw the frequency from a Dirichlet distribution with expected value equal to θ_i, drawing $\phi_t \sim \text{Dirichlet}(s\theta_i)$, where the scale parameter s controls the variance within each municipality.

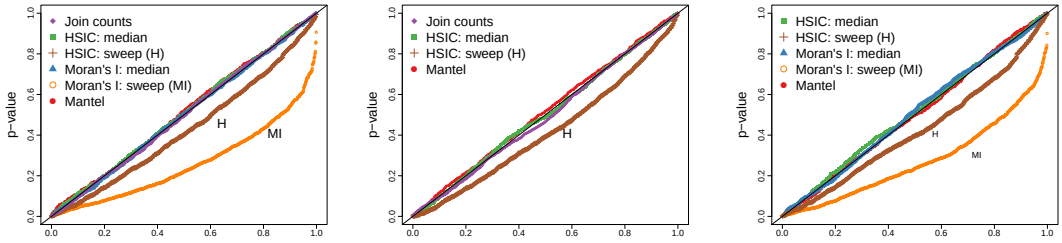
6.4.2 Calibration

Our first use of synthetic data is to examine the p-values obtained from each method when the null hypothesis is true — that is, when there is no geographical variation in the data. The p-value corresponds to the likelihood of seeing a test statistic at least as extreme as the observed value, under the null hypothesis. Thus, if we repeatedly generate data under the null hypothesis, a well-calibrated test will return a distribution of p-values that is uniform in the interval $[0, 1]$: for example, we expect to observe $p < .05$ in exactly 5% of cases, corresponding to the allowed rate of Type I errors (incorrect rejection of the null hypothesis) at the threshold $\alpha = 0.05$.

To measure the calibration of each of the proposed tests, we generate 1,000 random datasets using the procedure described above, and then compute the p-values under each test. In these random datasets, the relative frequency parameters θ_i are the same for all municipalities, which is the null hypothesis of complete randomization. To generate the binary and categorical data, we use $\mu_{obs} = 10^{-5}$, meaning that the expected number of observations is one per hundred thousand individuals in the municipality or province; for comparison, this corresponds roughly to the tweet frequency of the lengthened spelling *hellla* in the 2009-2012 Twitter dataset gathered by Eisenstein et al. [2014].

To visualize the calibration of each test, we use quantile-quantile (Q-Q) plots, comparing the obtained p-values with a uniform distribution. A well-calibrated test should give a diagonal line from the origin to $(1, 1)$. Figure 6.3 shows the Q-Q plots obtained from each method on each relevant type of data (recall that not all methods can be applied to all types of data, as described in the previous section).

HSIC and Moran's I each have tuning parameters that control the behavior of the test: the kernel bandwidth in HSIC and the distance cutoff in Moran's I. A simple heuristic is to use the median Euclidian distance $\overline{d}$: in Moran's I, we use $\overline{d}$ as the distance threshold for constructing the neighborhood matrix W; in HSIC, we use $\frac{1}{\overline{d}^2}$ as the kernel bandwidth parameter. Figure 6.3 shows that by basing these parameters on



(a) Binary data **(b)** Categorical data **(c)** Frequency data

Figure 6.3: Quantile-quantile plots comparing the distribution of the obtained p-values with a uniform distribution. The y-axis is the p-value returned by the tests. The x-axis shows the corresponding quantile for a uniform distribution on the range [0,1]. The approaches that optimize the parameters, i.e., the cutoff for Moran's I (MI) and the bandwidth for HSIC (H), lead to a skewed distribution of p-values.

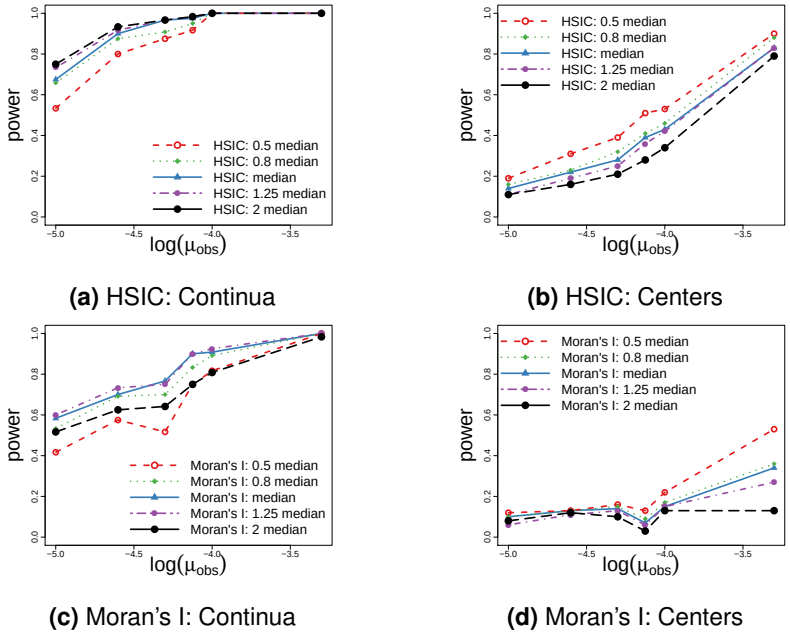
6

the median distance between pairs of points, we get well-calibrated results. However, some prior work takes an alternative approach, sweeping over parameter values to obtain the most significant results [Grieve et al., 2011]. In our experiments we sweep across the distance cutoff for Moran's I, and the bandwidth for the spatial distances in HSIC. This distorts the calibration, particularly for Moran's I, meaning that the resulting p-values are not reliable. This is most severe for Moran's I at the municipality level, reaching type I error rates of 11.7% (binary data) and 14.3% (frequency data) when the significance threshold α is set to 5%. Given that such parameter sweeps are explicitly designed to maximize the number of positive test results — and not the overall calibration of the test — this is unsurprising. We therefore avoid parameter sweeps in the remainder of this article, and rely instead on median distance as a simple heuristic alternative.

6.4.3 Power

Next, we consider synthetic data in which there is geographical variation by construction. We assess the *power* of each approach by computing the fraction of simulations for which the approaches correctly rejected the null hypothesis of no spatial dependence, given a significance threshold of $\alpha = 0.05$. We again use the Netherlands as the stage for all simulations, and consider two types of geographical variation.

Dialect continua. We generate data such that the frequency of a linguistic variant increases linearly through space, as in a dialect continuum [Heeringa and Nerbonne, 2001]. In most of the synthetic data experiments below, we average across a range of angles, from $0°$ to $357°$ with step sizes of $3°$, yielding 120 distinct angles in total. Each angle aligns differently with the population distribution of the Netherlands, so we also assess sensitivity of each method to the angle itself. Figure 6.2 shows two synthetic datasets with dialect continua in different angles.



(a) HSIC: Continua

(b) HSIC: Centers

(c) Moran's I: Continua

(d) Moran's I: Centers

Figure 6.4: Power across different parameter settings. Higher values indicate a greater likelihood of correctly rejecting the null hypothesis.

Geographical centers. Second, we consider a setting in which variation is based on one or more geographical *centers*. In this setting, all cities within some specified range of the center (or one of the centers) have some maximal frequency value θ_i; in other cities, this value decreases as distance from the nearest center grows. This corresponds to the dialectological scenario in which a variable form is centered on one specific city, as in, say, the association of the word *hella* with the San Francisco metropolitan area. We average across twenty five possible centers: the capitals of each of the twelve provinces of the Netherlands; the national capital of the Netherlands (Amsterdam); the *two* most populous cities in each of the twelve provinces. For each setting, we randomly generate synthetic data four times, resulting in a total of 100 synthetic datasets for this condition.

Parameter settings. We use these data generation scenarios to test the sensitivity of HSIC and Moran's I to their hyperparameters, by varying the kernel bandwidths in HSIC (Figures 6.4a and 6.4b) and the distance threshold in Moran's I (Figures 6.4c and 6.4d). The sensitivity of HSIC to the bandwidth value decreases as the number of data points increases (as governed by μ_{obs}), especially in the case of dialect continua. The sensitivity of Moran's I to the distance cutoff value decreases with the amount of data in the case of dialect continua, but in the case of center-based variation, Moran's I becomes *more* sensitive to this parameter as there is more data. For both methods, the same trends regarding the best performing parameters can be observed. In the case of dialect continua, larger cutoffs and bandwidths perform best, but in the case of variation based on centers, smaller cutoffs and bandwidths lead to higher power. Overall, there is no single best parameter setting, but the median heuristics perform reasonably well for both types of variation.

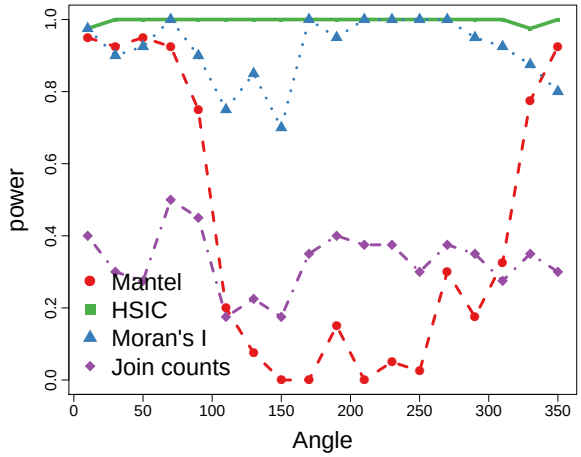
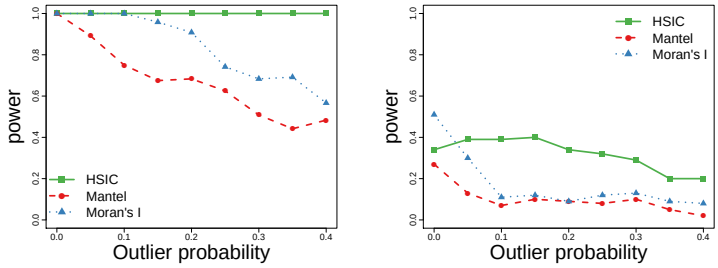


Figure 6.5: Relationship between the statistical power of each test and the angle of the dialect continuum across the Netherlands.

Direction of dialect continua. We simulate dialect continua by varying the frequency of linguistic variables linearly through space. Due to the heterogeneity of population density, different spatial angles will have very different properties: for example, one choice of angle would imply a continuum cutting through several major cities, while another choice might imply a rural-urban distinction. Figure 6.5 shows the power of the methods on binary data (there are two variant forms, and each instance contains exactly one of them), in which we vary the angle of the continuum. HSIC is insensitive to the angle of variation, demonstrating the advantage of this kernel nonparametric method. Moran's I is relatively robust, while join count analysis performs poorly across the entire range of settings. The Mantel test is remarkably sensitive to the angle of variation, attaining nearly zero power for some scenarios of dialect continua. This is caused by the complex interaction between the underlying linguistic phenomenon and the east-west variation of the population density of the Netherlands. For example, when the dialect continuum is simulated at an angle of 105 degrees, the south east of the Netherlands has a higher usage of the variable, but this is only a very small region due to the shape of the country. The Mantel test apparently has great difficulty in detecting geographical variation in such cases.

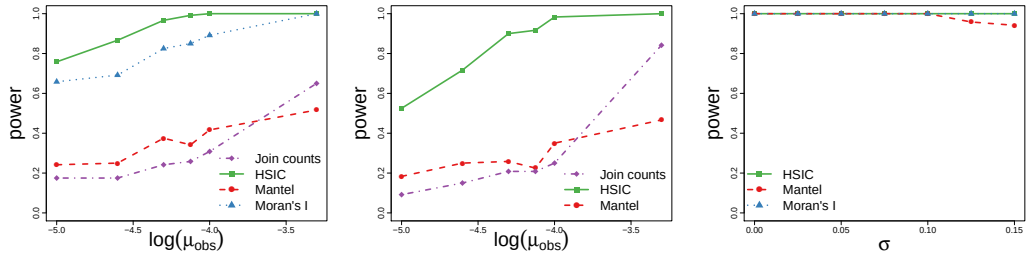
Outliers. In the frequency-based synthetic data, each instance uses each variable form with some continuous frequency — this is based on the scenario of letters-to-the-editors of regional newspapers, as explored in prior work [Grieve et al., 2011]. We test the robustness of each approach by introducing *outliers*: randomly selected data points whose variable frequencies are replaced at random with extreme values of either 0 or 1. As shown in Figure 6.6, HSIC is the most robust against outliers, while the performance of the Mantel test is the most affected by outliers (recall that join count analysis applies only to discrete observations, so it cannot be compared on this measure).



(a) Dialect continua **(b)** Centers

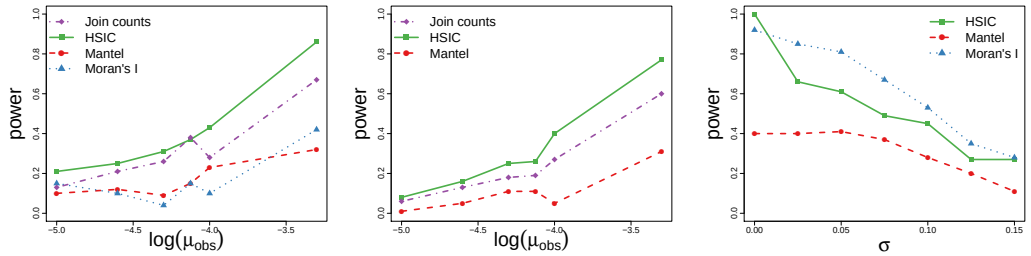
Figure 6.6: Results on synthetic frequency data ($\sigma = 0.1$) with outliers

Overall. We now compare the methods by averaging across various settings simulating dialect continua (Figure 6.7) and variation based on centers (Figure 6.8). To generate the categorical data, we vary μ_{obs} in our experiments, with a higher μ_{obs} resulting in more tweets and consequently less variation at the municipality level. As expected, the power of the approaches increases as μ_{obs} increases in the experiments on the categorical data, and the power of the approaches decreases as σ increases in the experiments on the frequency data. The experiments on the binary and categorical data show the same trends: HSIC performs the best across all settings. Join count analysis does well when the variation is based on centers, and Moran's I does best for dialect continua. Moran's I performs best on the frequency data, especially in the case of variation based on centers.



(a) Binary data **(b)** Categorical data (3) **(c)** Frequency data

Figure 6.7: Dialect continua



(a) Binary data **(b)** Categorical data (3) **(c)** Frequency data

Figure 6.8: Centers

6.4.4 Summary

In this section, we evaluated each statistical test for geographical language variation on a battery of synthetic data. HSIC and the Mantel test are the only approaches applicable to all data types (binary, categorical and frequency data). Overall, HSIC is more effective than the Mantel test, which is much more sensitive to the specifics of the synthetic data scenario, such as the angle of the dialect continuum. HSIC is robust against outliers, and performs particularly well when the number of data points increases. Join count analysis is suitable for capturing non-linear variation, but its power is low compared to other approaches in the analysis of dialect continua. Conversely, when the linguistic data is binary, Moran's I performs well on dialect continua, but its power is low in situations of variation based on centers. In our experiments on frequency data, Moran's I performs well in both scenarios.

6.5 Empirical Data

We now assess the spatial dependence of linguistic variables on three real linguistic datasets: letters to the editor (English), syntactic atlas of the Dutch dialects, and Dutch geotagged tweets. In each dataset, we compute statistical significance for the geo-linguistic dependence of multiple linguistic variables. To adjust the significance thresholds for multiple comparisons, we use the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] to bound the overall false discovery rate (FDR).

6.5.1 Letters to the Editor

In their application of Moran's I to English dialects in the United States, Grieve et al. [2011] compile a corpus of letters-to-the-editors of newspapers to measure the presence of dialect variables in text. To compute the frequency of the lexical variables, letters are aggregated to core-based statistical areas (CBSA), which are defined by the United States to capture the geographical region around an urban core. The frequency of 40 manually selected lexical variables is computed for each of 206 cities.

We use the Mantel test, HSIC, and Moran's I to assess the spatial dependence of variables in this dataset. Join count analysis was excluded, because it is not suitable for frequency data. We verified our implementation of Moran's I by following the approach taken by Grieve *et al.*: we computed Moran's I for cutoffs in the range of 200 to 1000 miles and selected the cutoff that yielded the lowest p-value. The obtained cutoffs and test statistics closely followed the values reported in the analysis by Grieve *et al.*, with slight deviations possibly due to our use of a permutation test rather than a closed-form approximation to compute the p-values.

After adjusting the p-values using the false discovery rate (FDR) procedure, a 500-mile cutoff results in three significant linguistic variables.[5] However, recall that the approach of selecting parameters by maximizing the number of positive test results tends to produce a large number of Type I errors. When setting the distance cutoff to the median distance between data points, none of the linguistic variables were

[5] Grieve *et al.* report five significant variables. In our analysis, there are two variables with FDR-adjusted p-values of 0.0559.

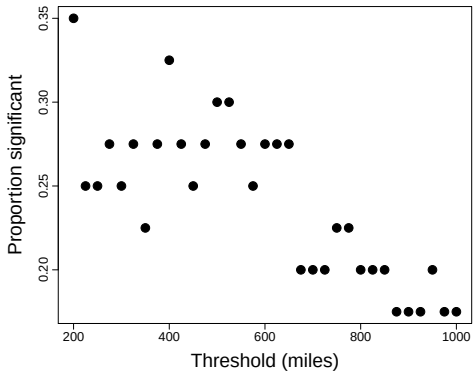


Figure 6.9: The proportion of variables detected to be significant (p<.05) by Moran's I by varying the distance cutoff (without adjusting for multiple comparisons).

found to have a significant geographical association. Similarly, HSIC and the Mantel test also found no significant associations after adjusting for multiple comparisons. Figure 6.9 shows the proportion of significant variables according to Moran's I based on different thresholds. The numbers vary considerably depending on the threshold. The figure also suggests that the median distance (921 miles) may not be a suitable threshold for this dataset.

6.5.2 Syntactic Atlas of the Dutch Dialects (SAND)

SAND [Barbiers et al., 2005, 2009] is an online electronic atlas that maps syntactic variation of Dutch varieties in the Netherlands, Belgium, and France.[6] The data was collected between the years of 2000 and 2005. SAND has been used to measure the distances between dialects, and to discover dialect regions [Spruit, 2006, Tjong Kim Sang, 2015].

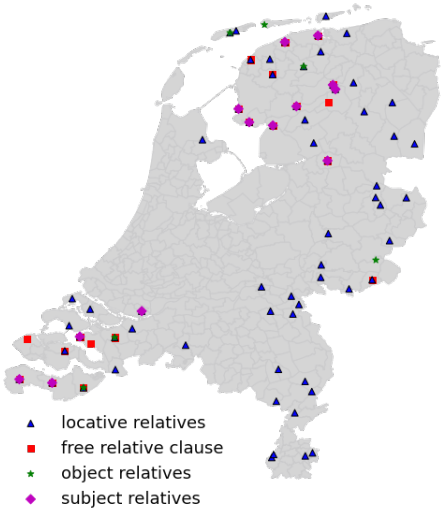
In our experiments, we consider only locations within the Netherlands (157 locations). The number of variants per linguistic variable ranges from one (due to our restriction to the Netherlands) to eleven. We do not include Moran's I in our experiments, since it is not applicable to linguistic variables with more than two variants. We apply the remaining methods to all linguistic variables with twenty or more data points and at least two variants, resulting in a total of 143 variables.

Table 6.10a lists the 10 variables with the highest HSIC values. Statistical significance at a level of $\alpha = 0.05$ is detected for 65.0% of the linguistic variables using HSIC, 78.3% when using join count analysis, and 52.4% when using the Mantel test. The three methods agree on 99 out of the 143 variables, and HSIC and join count analysis agree on 118 variables. From manual inspection, it seems that the non-linearity of the geographical patterns may have caused difficulties for the Mantel test. For example, Figure 6.10b is an example of a variable where HSIC and join count analysis both had an adjusted p-value $< .05$, but the Mantel test did not detect a significant pattern.

[6] http://www.meertens.knaw.nl/sand/

Map id	Description
1:84b	Free relative, complementizer following relative pronoun
1:84a	Short subject and object relative, complementizer following relative pronoun
1:80b	ONE pronominalisation
1:33a	Complementizer agreement 3 plural
1:76a	Reflexive pronouns; synthesis
2:36b	Form of the participle of the modal verb willen 'want'
1:29a	Complementizer agreement 1 plural
1:69a	Correlation weak reflexive pronouns
2:30b	Interruption of the verbal cluster; synthesis I
2:61a	Forms for iemand 'somebody'

(a) Highest ranked variables by HSIC. All methods had an adjusted p-value $< .05$ for all variables, except variable 2:30b (Mantel: $p = .118$.



(b) SAND map 16b, book 1. Finite complementizer(s) following relative pronoun (N=112). HSIC: p=.024; Mantel: p=.506; Join counts: p=.002.

Figure 6.10: Results SAND

6.5.3 Twitter

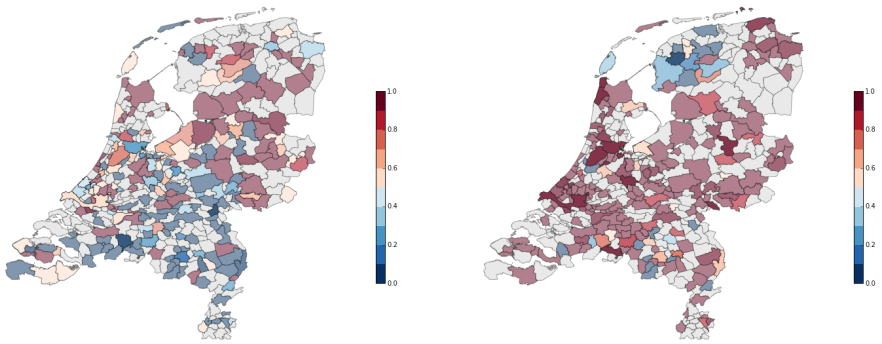
Our Twitter dataset consists of 4,039,786 geotagged tweets from the Netherlands, written between January 1, 2015 and October 31, 2015. We manually selected a set of linguistic variables (Table 6.1), covering examples of lexical variation (e.g., two different words for referring to french fries), phonological variation (e.g., t-deletion), and syntactic variation (e.g., *heb gedaan* ('have done') vs. *gedaan heb* ('done have')). We are not aware of any previous work on dialectal variation in the Netherlands that uses spatial dependency testing on Twitter data. The number of tweets per municipality varies dramatically, and for the less frequent linguistic variables there are no tweets at all in some municipalities. In our computation of Moran's I, we only include municipalities with at least one tweet.

Table 6.1 shows the output of each statistical test for this data. Some of these linguistic variables exhibit strong spatial variation, and are identified as statistically significant by all approaches. An example is the different ways of referring to french fries (*friet* versus *patat*, Figure 6.11a), where the figure shows a striking difference between the south and the north of the Netherlands. Another example is Figure 6.11b, which shows two different ways of saying 'for a little while' (*efkes* versus *eventjes*). The less common form, *efkes* is mostly used in Friesland, a province in the north of the Netherlands.

Examples of linguistic variables where the approaches disagree are shown in Figure 6.12. The first case (Figure 6.12a) is an example of lexical variation, with two different ways of saying *bye* in the Netherlands. A commonly used form is *doei*, while *houdoe* is known to be specific to North-Brabant, a Dutch province in the south of the Netherlands. HSIC and join count analysis both detect a significant pattern, but

Linguistic variables	Description	N	Moran's I	HSIC	Mantel	Join counts
Friet / patat	french fries	842	0.0004	0.0002	0.0003	0.0003
Proficiat / gefeliciteerd	congratulations	14,474	0.0004	0.0002	0.0080	0.0003
Iedereen / een ieder	everyone	13,009	0.8542	0.0002	0.8769	0.0432
Doei / aju	bye	4,427	0.7163	0.0050	0.2570	0.3868
Efkes / eventjes	for a little while	969	0.0036	0.0002	0.0003	0.0003
Naar huis / naar huus	to home	3, 942	0.8542	0.1090	0.1245	0.9426
Niet meer / nie meer	not anymore	11,596	0.0793	0.0002	0.5590	0.0329
Of niet / of nie	or not	1,882	0.8357	0.1010	0.4191	0.9426
-oa- / -ao-	e.g., *jao* versus *joa*	754	0.0004	0.0002	0.0003	0.0003
Even weer / weer even	for a little while again	921	0.0004	0.0002	0.0003	0.0003
Have + participle	e.g., *heb gedaan* ('have done') vs. *gedaan heb* ('done have')	1,122	0.8587	0.2849	0.6668	0.0255
Be + participle	e.g., *ben geweest* ('have been') vs. *geweest ben* ('been have')	1,597	0.0793	0.2849	0.7862	0.0051
Spijkerbroek / jeans	jeans	1,170	0.7796	0.0002	0.0080	0.0003
Doei/ houdoe	bye	4,491	0.5016	0.0002	0.6668	0.0047
Bellen / telefoneren	to call by telephone	4,689	0.2730	0.0003	0.9781	0.5941

Table 6.1: Twitter results. The p-values were calculated using 10,000 permutations and corrected for multiple comparisons.



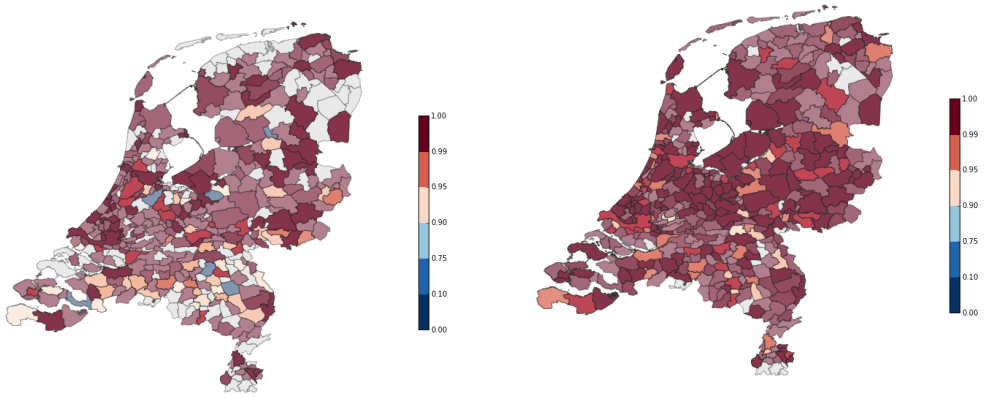
(a) French fries (*friet* versus *patat*), N=844

(b) For a little while (*efkes* versus *eventjes*), N=970

Figure 6.11: Highly significant linguistic variables on Twitter. Grey indicates areas with no data points. The intensity indicates the number of data points.

Moran's I and the Mantel test do not. The trend is less strong than in the previous examples, but the figure does suggest a higher usage of *houdoe* in the south of the Netherlands.

Another example is t-deletion for a specific phrase (*niet meer* versus *nie meer*), as shown in Figure 6.12b. Previous dialect research has found that geography is the most important external factor for t-deletion in the Netherlands, with contact zones, such as the Rivers region in the Netherlands (at the intersection of the dialects of the southern province of North-Brabant, the south-west province of Zuid-Holland and the Veluwe region), having high frequencies of t-deletion [Goeman, 1999]. Both HSIC and join count analysis report an FDR-adjusted $p < .05$, while for Moran's I, the geographical association does not reach the threshold of significance.



(a) Bye (*doei* versus *houdoe*), N=4,491

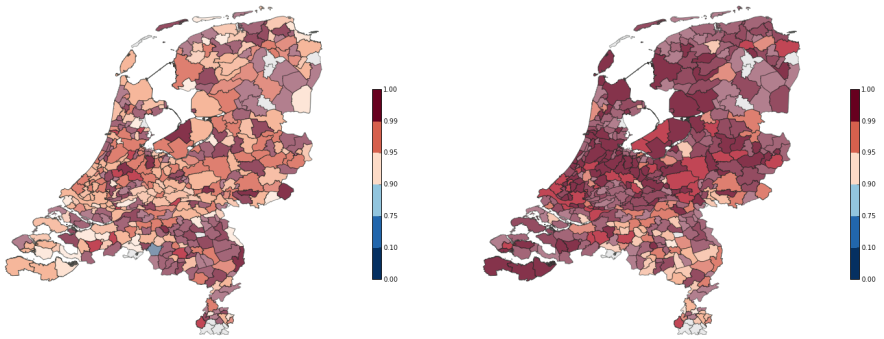
(b) t-deletion (*niet meer* vs. *nie meer*), N=11,596

Figure 6.12: Linguistic variables on Twitter where tests disagreed

We also present preliminary results on using HSIC as an exploratory tool on the same Twitter corpus. To focus on active users who are most likely tweeting on a personal basis, we exclude users with 1,000 or more followers and users who have fewer than 50 tweets, resulting in 8,333 users. We exclude infrequent words (used by fewer than 100 users) and very frequent words (used by 1000 users or more), resulting in a total of 5,183 candidate linguistic variables. We represent the usage of a word by each author as a binary variable, and use HSIC to compute the level of spatial dependence for each word.

The top 10 words with the highest HSIC scores are *groningen* (city), *zwolle* (city), *eindhoven* (city) *arnhem* (city), *breda* (city), *enschede* (city), *nijmegen* (city), *leiden* (city), *twente* (region) and *delft* (city). While these words do not reflect dialectal variation as it is normally construed, we expect their distribution to be heavily influenced by geography. Manual inspection revealed that many English words (e.g., *his, very*) have high geographical dependence. English speakers are more likely to visit tourist and commercial centers, so it is unsurprising that these words should show a strong geographical association. The top-ranked non-topical word is *proficiat*, occurring at rank 34 according to HSIC. *Proficiat* had previously been identified as a candidate dialect variable, and was included in our analysis in Table 6.1; this replication of prior dialectological knowledge validates the usage of HSIC as an exploratory tool. Less well known are *joh* (an interjection) and *dadelijk* ('immediately'/'just a second'), which are ranked respectively at #60 and #71 by HSIC. The geographical distributions of these words are shown in Figures 6.13a and 6.13b; both seem to distinguish the southern part of the Netherlands from the rest of the country. The identification of these words speaks to the potential of HSIC to guide the study of dialect by revealing geographically-associated terms.[7]

[7]The top 10 words with the highest Moran's I scores are similar: *groningen, eindhoven, friesland* (province), *leeuwarden* (city), *zwolle*, *proficiat*, *drachten* (city), *carnaval* (a festival), *brabant* (province), *enschede* (city).



(a) Usage of *joh* **(b)** Usage of *dadelijk*

Figure 6.13: Linguistic features on Twitter

6.6 Conclusion

We have reviewed four methods for quantifying the spatial dependence of linguistic variables: Moran's I, which is perhaps the best-known in sociolinguistics and dialectology; join count analysis; the Mantel test; and the Hilbert-Schmidt Independence Criterion (HSIC), which we introduce to linguistics in this chapter. Of these methods, only HSIC is *consistent*, meaning that it converges to an accurate measure of the statistical dependence between X and Y in the limit of sufficient data. In contrast, the other approaches are based on parametric models. When the assumptions of these models are violated, the power to detect significant geo-linguistic associations is diminished. All three of these methods can be modified to account for various geographical distributions: for example, the spatial weighting matrix employed in Moran's I and join count analysis can be constructed as a non-linear or non-monotonic function of distance [Getis and Aldstadt, 2010], the distances in the Mantel test can be censored at some maximum value [Legendre et al., 2015], and so on. However, such modifications require the user to have strong prior expectations of the form of the geolinguistic dependence, and open the door to p-value hacking through iterative "improvements" to the test. In contrast, HSIC can be applied directly to any geotagged corpus, with minimal tuning. By representing the underlying probability distributions in a Hilbert space, HSIC implicitly makes a comparison across high-order moments of the distributions, thus recovering evidence of probabilistic dependence without parametric assumptions.

These theoretical advantages are borne out in an analysis of synthetic data. We consider a range of realistic scenarios, finding that the power of Moran's I, the Mantel test, and join count analysis depends on the nature of the geographical variation (e.g., dialect continua versus centers), and in some cases, even on the direction of variation. Overall, we find that HSIC, while not the most powerful test in every scenario, offers the broadest applicability and the least potential for catastrophic failure of any of the proposed approaches.

We then showed how to apply these tests to a diverse range of real datasets: frequency observations in letters to the editor, binary observations in a dialect atlas,

and binary observations in social media. We find that previous results on newspaper data were dependent on the procedure of selecting the geographical distance cutoff to maximize the number of positive test results; using all other test procedures, the significance of these results disappears. On the dialect atlas, we find that the fraction of statistically significant variables ranges from 55.2% to 78.3% depending on the statistical approach. On the social media data, we obtain largely similar results from the four different tests, but HSIC detects the largest number of significant associations, identifying cases in which geography and population density were closely intertwined.

To conclude, we believe that kernel embeddings of probability measures offer a powerful new approach for corpus analysis. In this chapter, we have focused on measuring geographical dependence, which can be used to test and discover new dialectal linguistic variables. But the underlying mathematical ideas may find application in other domains, such as tracking change over time [Popescu and Strapparava, 2014, Štajner and Mitkov, 2011], or between groups of authors [Koppel et al., 2002]. Of particular interest for future research is the use of structured kernels, such as tree kernels [Collins and Duffy, 2001] or n-gram kernels [Lampos et al., 2015], which could test for structured linguistic phenomena such as variation in syntax [Johannsen et al., 2015] or phonological change [Bouchard et al., 2007].

7

Word-Level Language Identification

This chapter is based on D. Nguyen and A.S. Doğruöz, "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, 2013 [Nguyen and Doğruöz, 2013]

7.1 Introduction

Automatic language identification is frequently the first step in processing mixed-language texts. In addition, it supports large-scale analyses of language choice patterns in multilingual communication. In this chapter, we identify Dutch (NL) and Turkish (TR) at the word level in a large online forum for Turkish-Dutch speakers living in the Netherlands. The users in the forum frequently switch languages within posts, for example:

<TR> *Sariyi ver* </TR>
<NL> *Wel mooi doelpunt* </NL>
Translation:<TR> *Give me the yellow one* </TR>
<NL> *Rather nice goal* </NL>

So far, language identification has mostly been modeled as a document classification problem. Most approaches rely on character or byte n-grams, by comparing n-gram profiles [Cavnar and Trenkle, 1994], or using various machine learning classifiers. While McNamee [2005] argues that language identification is a solved problem, classification at a more fine-grained level (instead of document level) remains a challenge [Hughes et al., 2006]. Furthermore, language identification is more difficult for short texts [Baldwin and Lui, 2010, Vatanen et al., 2010], such as queries and tweets [Bergsma et al., 2012a, Carter et al., 2013, Ceylan and Kim, 2009]. Tagging individual words (without context) has been done using dictionaries, affix statistics and classifiers using character n-grams [Gottron and Lipka, 2010, Hammarström, 2007]. Although Yamaguchi and Tanaka-Ishii [2012] segmented text by language, their data was artificially created by randomly sampling and concatenating text segments (40-160 characters) from monolingual texts. Therefore, the language switches do not

reflect realistic switches as they occur in natural texts. Most related to ours is the work by King and Abney [2013] who labeled languages of words in multilingual web pages, but evaluated the task only using word-level accuracy. A more detailed discussion of related work is provided in Subsection 2.5.2.

This chapter makes the following contributions: 1) We explore two new ways to evaluate the task for analyzing multilingual communication and show that only using word accuracy gives a limited view; 2) We are the first to apply this task on a conversational and larger dataset; 3) We show that features using the context improve the performance; 4) We present a new public dataset to support research on language identification. In the rest of the chapter, we first describe our dataset (Section 7.2). Secondly, we present our experiments (Section 7.3) and discuss the results (Section 7.4). We finally conclude with a summary and suggestions for future work (Section 7.5).

7.2 Data

Our data[1] comes from one of the largest online communities in The Netherlands for Turkish-Dutch speakers. All posts from May 2006 until October 2012 were crawled. Although Dutch and Turkish dominate the forum, English fixed phrases (e.g., *no comment, come on*) are also occasionally observed. Users switch between languages within and across posts. Examples 1 and 2 illustrate switches between Dutch and Turkish within the same post. Example 1 is a switch at the sentence level, example 2 is a switch at the word level.

Example 1:
<NL>Mijn dag kan niet stuk :) </NL> <TR> Cok guzel bir haber aldim </TR>
Translation: <NL> This made my day:) </NL><TR> I received good news </TR>

Example 2:
<TR>kahvalti</TR><NL>met vriendinnen by my thuis </NL>
Translation: <TR>breakfast </TR><NL> with my girlfriends at my home </NL>

The data is highly informal with misspellings, lengthening of characters (e.g., *hotttt*), replacement of Turkish characters (e.g., *kahvalti* instead of *kahvaltı*) and spelling variations (e.g., *tankyu* instead of *thank you*). Dutch and Turkish sometimes share common spellings (e.g., *ben* is *am* in Dutch and *I* in Turkish), making this a challenging task.

Annotation. For this research, we classify words as either Turkish or Dutch. Since Dutch and English are typologically more similar to each other than Turkish, the English phrases (less than 1%) are classified as Dutch. Posts were randomly sampled and annotated by a native Turkish speaker who is also fluent in Dutch. A native Dutch speaker annotated a random set of 100 posts (Cohen's kappa = 0.98). The following tokens were ignored for language identification:

[1] Available at http://www.dongnguyen.nl/data-langid-emnlp2013.html.

- Smileys (as part of the forum markup, as well as textual smileys such as *:)*).
- Numeric tokens and punctuation.
- Forum tags (e.g., *[u]* to underline text).
- Links, images, embedded videos, etc.
- Turkish and Dutch first names and place names.[2]
- Usernames when indicated with special forum markup.
- Chat words, such as *hahaha*, *ooooh* and *lol* recognized using regular expressions.

Posts for which all tokens are ignored, are not included in the corpus.

Statistics. The dataset was randomly divided into a training, development and test set. The statistics are listed in Table 7.1. The statistics show that Dutch is the majority language, although the difference between Turkish and Dutch is not large. We also find that the documents (i.e., posts) are short, with on average 18 tokens per document. The data represents realistic texts found in online multilingual communication. Compared to previously used datasets [King and Abney, 2013, Yamaguchi and Tanaka-Ishii, 2012], the data is noisier and the documents are much shorter.

	#NL tokens	#TR tokens	#Posts (BL%)
Train	14,900 (54%)	12,737 (46%)	1,603 (15%)
Dev	8,590 (51%)	8,140 (49%)	728 (19%)
Test	5,895 (53%)	5,293 (47%)	735 (17%)

Table 7.1: Number of tokens and posts for Dutch (NL) and Turkish (TR), including % of bilingual (BL) posts

7.3 Experimental Setup

7.3.1 Training Corpora

We used the following corpora to extract dictionaries and language models.

- *GenCor*: Turkish web pages [Sak et al., 2008].
- *NLCOW2012*: Dutch web pages [Schäfer and Bildhauer, 2012].
- *Blog authorship corpus*: English blogs [Schler et al., 2006].

Each corpus was chunked into large segments which were then selected randomly until 5M tokens were obtained for each language. We tokenized the text and kept the punctuation.

[2]Based on online name lists and Wikipedia pages.

7.3.2 Baselines

As baselines, we use *langid.py*[3] [Lui and Baldwin, 2012] and van Noord's *TextCat* implementation[4] of the algorithm by Cavnar and Trenkle [1994]. TextCat is based on the comparison of n-gram profiles and langid.py on Naive Bayes with n-gram features. For both baselines, words were entered individually to each program. Words for which no language could be determined were assigned to Dutch. These models were developed to identify the languages of the documents instead of words and we did not retrain them. Therefore, these models are not expected to perform well on this task.

7.3.3 Models

We start with models that assign languages based on only the current word. Next, we explore models and features that can exploit the context (the other words in the post). Words with the highest probability for English were assigned to Dutch for evaluation.

Dictionary lookup (`DICT`)
We extract dictionaries with word frequencies from the training corpora. This approach looks up the words in the dictionaries and chooses the language for which the word has the highest probability. If the word does not occur in the dictionaries, Dutch is chosen as the language.

Language model (`LM`)
We build a character n-gram language model for each language (max. n-gram length is 5). We use Witten-Bell smoothing and include word boundaries for calculating the probabilities. The language with the highest likelihood of the word is chosen.

Dictionary + Language model (`DICT+LM`)
We first use the dictionary lookup approach (**`DICT`**). If the word does not occur in the dictionaries, a decision is made using the language models (**`LM`**).

Logistic Regression (`LR`)
We use a logistic regression model that incorporates context with the following features:

- *Individual word*: Label assigned by the **`DICT+LM`** model.
- *Context*: The results of the **`LM`** model based on previous + current token, and current token + next token (e.g., the sequence "*ben thuis*" (*am home*) as a whole if *ben* is the current token). This gives the language model more context for estimation. We compare the use of the assigned labels (**`LAB`**) with the use of the log probability values (**`PROB`**) as feature values.

[3] https://github.com/saffsd/langid.py
[4] http://www.let.rug.nl/∼vannoord/TextCat/

	Word classification					Fraction				Post class.	
	TR		NL				MAE				
Run	P	R	P	R	Acc.	ρ	All	Mono.	BL	F_1	Acc.
Textcat	.872	.647	.743	.915	.788	.739	.251	.264	.188	.386	.396
LangIDPy	.954	.387	.641	**.983**	.701	.615	.364	.371	.333	.413	.475
DICT	.955	.733	.802	.969	.858	.827	.196	.200	.175	.511	.531
LM	.950	.930	.938	.956	.944	.926	.074	.076	.065	.699	.703
DICT+LM	.951	.934	.942	.957	.946	.943	.067	.067	**.063**	.711	.717
LR+LAB	.965	.952	.958	.969	.961	.917	.066	.066	.068	.791	.808
LR+PROB	.956	.976	.978	.959	.967	.945	.048	.044	.064	.826	.849
CRF+BASE	**.973**	.974	.977	.976	.975	.940	.043	.027	.119	**.858**	**.898**
CRF+LAB	.964	.977	.979	.967	.972	.933	.046	.033	.111	.855	.891
CRF+PROB	.970	**.980**	**.982**	.973	**.976**	**.946**	**.039**	**.025**	.103	.853	.895

Table 7.2: Results of language identification experiments.

Conditional Random Fields (`CRF`)
We treat the task as a sequence labeling problem and experiment with linear-chain Conditional Random Fields [Lafferty et al., 2001] in three settings:

- *Individual word*: A CRF with only the tags assigned by the **`DICT+LM`** to the individual tokens as a feature (**`BASE`**).
- *Context*: CRFs using the **`LAB`** or **`PROB`** as additional features (same features as in the logistic regression model) to capture additional context.

7.3.4 Implementation

Language identification was not performed for texts within quotes. To handle the alphabetical lengthening (e.g., *lollll*), words are normalized by trimming same character sequences of three characters or more. We use the Lingpipe[5] and Scikit-learn [Pedregosa et al., 2011] toolkits for our experiments.

7.3.5 Evaluation

The assigned labels can be used for computational analysis of multilingual data in different ways. For example, these labels can be used to analyze language preferences in multilingual communication or the direction of the switches (from Turkish to Dutch or the other way around). Therefore, we evaluate the methods from different perspectives. The evaluation at word and post levels is done with the following metrics:

- **Word classification.** Precision (P), recall (R) and accuracy. Although this is the most straightforward approach to evaluate the task, it ignores the document boundaries.

[5] http://alias-i.com/lingpipe/

- **Fraction of language in a post.** Pearson's correlation (ρ) and Mean Absolute Error (MAE) of proportion of Turkish in a post. This evaluates the measured proportion of languages in a post when the actual tags for individual words are not needed. For example, such information is useful for analyzing the language preferences of users in the online forum. Besides reporting the MAE over all posts, we also separate the performance over monolingual and bilingual posts (BL).

- **Post classification.** Durham [2003] analyzed the switch between languages in terms of the amount of monolingual and bilingual posts. Our posts are classified as NL, TR or bilingual (BL) if all words are tagged in the particular language or both. We report F_1 and accuracy.

7.4 Results

The results are presented in Table 7.2. Significance tests were done by comparing the results of the word and post classification measures using McNemar's test, and comparing the MAEs using paired t-tests. All runs were significantly different from each other based on these tests ($p < .05$), except the MAEs of the **DICT+LM** and **LR+LAB** runs and the MAEs and post classification metrics between the CRFs runs.

The difficulty of the task is illustrated by examining the coverage of the tokens by the dictionaries. 24.6% of the tokens (dev + test set) appear in both dictionaries, 31.1% only in the Turkish dictionary, 30.5% only in the Dutch dictionary and 13.9% in none of the dictionaries.

The baselines do not perform well. This confirms that language identification at the word level needs different approaches than identification at the document level. Using language models result in a better performance than dictionaries. They can handle unseen words and are more robust against variation in spelling. The combination of language models and dictionaries is more effective than the individual models. The results improve when context was added using a logistic regression model, especially with the probability values as feature values.

CRFs improve the results but the improvement on the correlation and MAE is less. More specifically, CRFs improve the performance on monolingual posts, especially when a single word is tagged in the wrong language. However, when the influence of the context is too high, CRFs reduce the performance in bilingual posts. This is also illustrated with the results of the post classification. The **LR+PROB** run has a high recall (0.905), but a low precision (0.559) for bilingual posts, while the **CRF+PROB** approach has a low recall (0.611) and a high precision (0.828).

The fraction of Dutch and Turkish in posts varies widely, providing additional challenges to the use of CRFs for this task. Classifying posts first as monolingual/bilingual and tagging individual words afterwards for bilingual posts might improve the performance. The evaluation metrics highlight different aspects of the task whereas word-level accuracy gives a limited view. We suggest using multiple metrics to evaluate this task for future research.

Dictionaries versus language models. The results reported in Table 7.2 were obtained by sampling 5M tokens of each language. To study the effect of the number of tokens on the performance of the **DICT** and **LM** runs, we vary the amount of data. The performance of both methods increases consistently with more data (Figure 7.1). We also find that language models achieve good performance with only a limited amount of data, and consistently outperform the approach using dictionaries. This is probably due to the highly informal and noisy nature of our data.

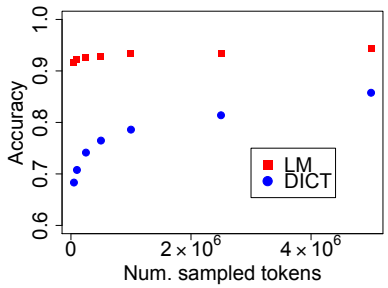


Figure 7.1: Effect of sampling size

Post classification. We experimented with classifying posts into TR, NL and bilingual posts using the results of the word-level language identification (Table 7.2: post classification). Posts were classified as a particular language if all words were tagged as belonging to that language, and bilingual otherwise. Runs using CRFs achieved the best performance.

We now experiment with allowing a margin (e.g., a margin of 0.10 classifies posts as TR if at least 90% of the words are classified as TR). Allowing a small margin already increases the results of simpler approaches (such as the **LR-PROB** run, Table 7.3) by making it more robust against errors. However, allowing a margin reduces the performance of the CRF runs.

Margin	0.0	0.05	0.10	0.15	0.20
Accuracy	0.849	0.873	0.876	0.878	0.865

Table 7.3: Effect of margin on post classification (**LR-PROB** run)

Error analysis. The manual analysis of the results revealed three main challenges: 1) Our data is highly informal with many spelling variations (e.g., *moimoimoi*, *gooooooool-llll*) and noise (e.g., *asdfghjfgshahaha*); 2) Words sharing spelling in Dutch and Turkish are difficult to identify especially when there is no context available (e.g., a post with only one word). These words are annotated based on their context. For example, the word *super* in *"Seyma, super*" is annotated as Turkish since *Seyma* is also a Turkish word; 3) Named entity recognition is necessary to improve the performance of the system and decrease the noise in evaluation. Based on precompiled lists, our system ignores named entities. However, some names still remain undetected (e.g., usernames).

7.5 Conclusion

We presented experiments on identifying the language of individual words in multilingual conversational data. Our results reveal that language models are more robust than dictionaries and adding context improves the performance. We evaluate our methods from different perspectives based on how language identification at the word level can be used to analyze multilingual data. The highly informal spelling in online environments and the occurrences of named entities pose challenges. Future work could focus on cases with more than two languages, and languages that are typologically less distinct from each other or dialects [Trieschnigg et al., 2012].

8

Audience and the Use of Minority Languages on Twitter

This chapter is based on D. Nguyen, D. Trieschnigg and L. Cornips, "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, 2015 [Nguyen et al., 2015a]

8.1 Introduction

Over 10% of the Twitter users tweet in more than one language [Hale, 2014]. Also within a single language, there is much geographical variation [Eisenstein et al., 2010]. Every user has his or her own linguistic repertoire which the individual user can draw linguistic elements or codes (language varieties) from. We illustrate this with two tweets from an international fashion model from Friesland (a Dutch province) who can draw on English, Dutch and Frisian:

> We just touched down in London town #vsfashionshow

> @USER SKATSJE!!! Lekker genietsje fan heit en mem en Fryslan!!
> ik mis jim
> *Translation: @USER CUTIE!!! Enjoy mom and dad in Friesland!!*
> *i miss you*

She mostly tweets in English, possibly to maximize her audience [Androutsopoulos, 2014] and to create an international image. For example, the first tweet is written in English. Using *#vsfashionshow* the tweet becomes part of a public stream about a fashion show, increasing her audience even more. The second tweet is a response to a tweet from her sister and is written in Frisian, a minority language spoken in Friesland. Through tweeting in Frisian, she is constructing their shared localness or 'Frisianess'.

Which language and linguistic elements users select from their linguistic repertoire depends on various factors, including the audience, the topic, the perspective, whether the user wants to mark something as humorous/serious, etc. [Androutsopoulos, 2013a]. In this chapter, we focus on the influence of *audiences* on whether a minority language is used on Twitter. A speaker's style is influenced by the audience [Bell, 1984], and in that sense, social media, and especially Twitter is interesting: multiple audiences (e.g., friends, colleagues) are collapsed into a single context [Marwick and boyd, 2011]. Users with public profiles on Twitter have potentially a limitless audience, but they often imagine an audience when writing tweets and may target tweets to different audiences [Marwick and boyd, 2011].

We study Twitter users in two provinces in the Netherlands, where besides Dutch a minority language is spoken. Frisian (spoken in Friesland) is recognized as an official language in the Netherlands. Limburgish (spoken in Limburg) – a group of what people call dialects – has also received minor recognition by the Netherlands, a signatory of the 1992 European Charter for Regional Languages or Languages of Minorities. There is a positive attitude towards both minority languages, but their use has declined [Cornips, 2013, Riemersma et al., 2001].

In this chapter, we analyze the *language choices* of users at the tweet level, focusing on when users tweet in a minority language. An automatic language identification tool is used to classify tweets according to their language. We distinguish between two types of tweets: tweets that are a response to another tweet, and 'independent' tweets. We first focus on independent tweets, analyzing tweets with direct addressees (where the targeted audience may be reduced) and tweets with hashtags (where the audience may be expanded).

We then study language choices for tweets that are responses to other tweets by extracting conversations. Speakers may often *code-switch*, i.e., use multiple languages in a single speech exchange. In this study, we focus on code-switching at the tweet level. Following Androutsopoulos [2013a], we take a restrictive view on what is considered a speech exchange and confine our attention to code-switching within Twitter conversations.

Our contributions can be summarized as follows:

- We show that Twitter users accommodate to their audiences by studying the influence of direct addressees and hashtag streams on language choice.
- We study code-switching patterns within Twitter conversations and find that characteristics of the conversation partner as well as previous language choices in the conversation influence language choice.

8.2 Related Work

Our study builds on two different lines of work. First, we draw from the frameworks of audience design [Bell, 1984] and communication accommodation theory [Giles et al., 1991], and in particular recent studies that have applied these frameworks to social media settings. On Facebook, users maximize or partition their audience (when starting posts) or align or disalign (when responding) using their language

choices [Androutsopoulos, 2014]. A small-scale study on Twitter revealed that bilingual Welsh/English users more often tweet in Welsh to a user who is also bilingual, and in English when posting a tweet that is not directed to particular users [Johnson, 2013].

Second, we follow recent large-scale quantitative studies of language choice and code-switching based on automatic language identification [Eleta and Golbeck, 2014, Hale, 2014, Jurgens et al., 2014, Kim et al., 2014]. Traditional sociolinguistic studies rely on qualitative analyses (cf. Androutsopoulos [2014]) or quantitative analyses using questionnaires or manual coding (see Androutsopoulos [2013a]). While revealing valuable insights, these studies have been limited to small sets of speakers. Larger datasets that are automatically tagged by language can complement such studies. So far, large-scale studies have mostly focused on the networks of multilingual users, finding that multilingual users connect users who only tweet in one language [Eleta and Golbeck, 2014, Hale, 2014, Kim et al., 2014]. In these studies, users were represented by a language label or the language distribution of their tweets, thus focusing on language choice at the *user level*. In comparison, this study focuses on language choice at the *tweet level*.

8.3 Data

Twitter users from the Dutch provinces Friesland and Limburg were collected by starting with seed users and expanding using followers and followees. The seed users were manually identified users and users with a geotagged tweet from within these provinces (streaming API: January 2013 - July 2014). Users were mapped to locations (city, province, country) based on their provided profile location. For each user we collected the most recent 200 tweets.

An automatic language identifier was used to label the tweets. A training set of over 38k tweets was manually compiled with tweets labeled as English, Dutch, Limburgish or Frisian. Tweets containing multiple languages were labeled according to the predominant language. A logistic regression classifier obtained a cross-validation accuracy of 98%. Because performance was lower on very short tweets, tweets with less than 4 tokens were not labeled by the classifier. Manual rules were constructed to label a subset of the very short tweets. Similar to other studies on language choice (e.g., Kim et al. [2014]), we applied a threshold to determine whether a user uses a minority language on Twitter. We only retained users with at least 7.5%[1] of their tweets marked as containing Frisian or Limburgish, resulting in 2,069 users from Friesland and 2,761 users from Limburg.

We extract conversations based on information from the Twitter API, which provides the identifier of the original tweet in case of a reply. We excluded conversations with tweets from only one user (users can reply to themselves) and conversations for which the first tweet was a response to a missing tweet. We extracted 3,916 conversations, containing a total of 10,434 tweets. Most conversations were of length 2 (mean: 2.664, max 23).

[1]The threshold was based on data analysis, retaining approximately 23% of the users.

8.4 Language Choice

In this section, we focus on tweets that are not a response to another tweet and are not a retweet. We analyze tweets with an explicit mention of another user. In such cases the targeted audience is often shifted towards the addressed user. We also study tweets with a hashtag, which causes a possible expansion of the audience as they are included in public hashtag streams. These differences in audiences are reflected in statistics normalized by user: When users mention a specific user, they are more likely to employ a minority language than when they use hashtags (e.g., users from Limburg use Limburgish in 33.8% of their tweets with a user mention vs. only in 28.6% of their tweets with a hashtag).

Addressee. We first study the influence of addressees on language choice. We restrict our analysis to tweets that start with a user mention (@user). Such tweets are often directed towards the addressed user, in comparison to just tagging a user. For each user, we sampled up to two tweets.

We aim to analyze if addressees influence whether a minority language is used, while controlling for a user's tendency to use a minority language. We use logistic regression, which allows analyzing which factors explain the language choice. We fit a model with the dependent variable being the language choice, modeled as a binary variable (minority language or not). Independent variables are the use of minority language by the addressee, measured as the proportion of the last 100 tweets (before the tweet of interest) containing a minority language, and a binary variable indicating whether the addressee is from the same province. We collected additional data for addressees who were not in our dataset. However, for some we were not able to obtain data and these were excluded from the analysis. The results (Table 8.1) indicate that Twitter users are more likely to use a minority language, when addressing a user who often uses the minority language. From manual inspection we do observe that users not always accommodate to their addressee. For instance, sometimes even celebrities or international companies are addressed in a minority language.

	Coefficient	**Std. Error**
Intercept	-2.010***	0.149
Use of minority lang. by user u	2.685***	0.299
Use of minority lang. by user a	3.221***	0.293
Same province	0.160	0.149

Table 8.1: Logistic regression model of influence of addressee *a* on language choice of user *u*; n = 1272; $^{***}p < .001$

Hashtags. Hashtags have become a common practice on Twitter and they are often included to join public discussions [Huang et al., 2010]. We study the influence of the audiences of these public discussions on the language choice for tweets with hashtags.

For example, one of the most popular Dutch hashtag streams on Twitter is *#dtv* or *#durftevragen* ('dare to ask'). In these streams, Twitter users post questions on various topics, ranging from questions about software, opinions about news, to looking for a certain service. These streams have local variants (albeit less popular), such as for Limburgish *#durftevraoge* and *#durftevroage* and for Frisian *#doartefreechjen* and *#doartefreegjen*. Reaching the right audience is key here, since users are looking for an answer to their question. In our dataset, tweets using the local Limburgish and Frisian hashtag variants are all written in a minority language, whereas 84.6% of the tweets using the Dutch hashtag variants are written in Dutch.

Not all hashtags are added to join public discussions, for example some indicate a feeling (e.g., *#sad*) [Jurgens et al., 2014]. We therefore confine our analysis to hashtags referring to named entities. These are interesting, because they tend to be used to link to a public discussion and most of them do not imply a language choice on their own. We excluded hashtags that had local variants, such as names referring to cities. We manually annotated a random subset of the tweets (at most 1 tweet per user). In addition, we annotated whether the hashtag was referring to a local (e.g., local music festival) or (inter)national entity (e.g., show on national television). Of the annotated hashtags, 44.2% were marked as named entities and 51.7% of these referred to a local entity.

Similar to our previous analysis, we fit a logistic regression model with the language choice being the dependent variable. On Twitter the exact audience is unknown, but users use cues from the environment to imagine their audience [Marwick and boyd, 2011]. In a similar way, we use the last 100 tweets with the same hashtag written *before* the tweet of interest as cues. The tweets were collected using the search on the Twitter website, which allows searching in historical tweets. For each hashtag instance, we calculated the proportion of tweets in the stream containing a minority language. The audience may also consist of lurkers, but their language choices cannot be analyzed.

Table 8.2 shows the results. When many tweets in the hashtag stream contain a minority language, it is more likely that a user will use a minority language as well (even after controlling for the use of minority language by the user and whether the hashtag refers to a local or national entity).

	Coefficient	**Std. Error**
Intercept	-3.718***	0.453
Use of minority lang. by user	4.984***	0.819
Use of minority lang. in stream	6.489***	1.352
Hashtag about local entity	0.513	0.435

Table 8.2: Logistic regression model of influence of hashtag on language choice; n = 236; $^{***}p < .001$

8.5 Code-Switching in Twitter Conversations

In this section, we study the language choices of Twitter users when participating in conversations. Multilingual users often switch language during a conversation (i.e., *code-switching*). Initial tweets are frequently targeted towards a broader audience, but during a conversation the audience often shifts towards the direct conversation partner(s). Speakers accommodate to each other during conversations [Giles et al., 1991], which has also been observed on Twitter [Danescu-Niculescu-Mizil et al., 2011]. While the previous section focused on independent tweets, this section focuses on conversations, and thus an additional factor that influences language choice are the previous language choices made within the conversation.

Influence of previous tweet. We calculate the probability of a language choice for a tweet ($lang_i$) given the language of the tweet the user is responding to ($lang_{i-1}$), i.e., $P(lang_i|lang_{i-1})$ shown in Figure 8.1. Most of the time users align their language choice with the language of the tweet they are responding to (i.e., the self loop probabilities for Dutch and the minority languages are all above 0.5), and this trend is particularly strong when responding to tweets written in a minority language. However, this is not the case for English, which may be explained by the fact that English is most often used emphatically, for example by only inserting *nice* or *thanks*, and thus it is less expected that the conversations continue in English. We also find that users from the Limburg province more often tweet in their minority language than users from Friesland.

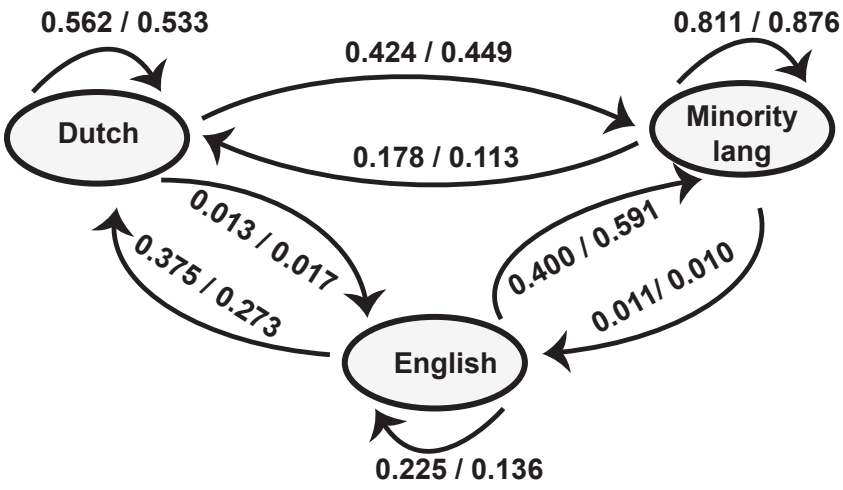


Figure 8.1: Switching behavior. The probabilities are reported for both provinces using [Friesland]/[Limburg]

In our next analysis, we also take into account the previous use of minority language by the users. Only tweets at the second position in a conversation were included in the analysis, to eliminate effects of other language choices. For each user, we sampled at most two tweets. Similar to the previous analyses, we fit a logistic regression model (Table 8.3) with as the dependent variable the language choice

($lang_i$). As independent variables, we include the use of minority language by both users as well as the language of the previous tweet. Location information was not included, since both users are from the same province. The results indicate that while the use of the minority language by the conversation partner is significant, the language of the previous tweet has a larger influence on the language choice.

	Coefficient	Std. Error
Intercept	-1.005***	0.112
Use of min. lang. by user of tweet$_i$	2.053***	0.241
Use of min. lang. by user of tweet$_{i-1}$	0.773**	0.248
Tweet$_{i-1}$ in minority language	1.478 ***	0.132

Table 8.3: Logistic regression model for language choice in conversations; n = 1863; $^{***}p < .001$, $^{**}p < .01$

Language choice over time. Figure 8.2 shows the language distribution by position within a conversation. The analysis is based on all conversations, but we note that the same trends are observed when only including longer conversations. Most of the initial tweets are written in Dutch, possibly to maximize the audience [Androutsopoulos, 2014]. However, as conversations progress, it becomes more unlikely for a tweet to be written in Dutch. Once a switch has been made to a minority language, users tend to continue in that minority language (see also Figure 8.1).

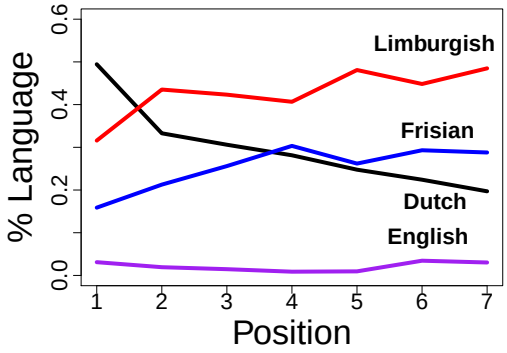


Figure 8.2: Language distribution by position in conversation

By replying in a different language, a user may be trying to negotiate the language choice [Androutsopoulos, 2014]. We expect that once a base language has been established, the probability of switching language decreases. For each $tweet_i$, we find the longest consecutive sequence ending at the previous tweet ($tweet_{i-1}$) written in $lang_{i-1}$ as an indication of the extent of negotiation going on. As expected, there is a significant, negative correlation between the lengths of these sequences and whether a switch occurs (Pearson's r=-.150, p $<$ .0001). The position in a conversation and whether a switch occurs correlate only slightly (Pearson's r=-.058, p $<$ .001) and controlling for this did not lead to notable changes in the trend.

8.6 Conclusion

In this chapter, we studied the use of minority languages on Twitter across various settings. Our findings indicate that users tend to adapt their language choice to their audiences. When users address other Twitter users, the minority language is more likely to be used when the addressed users often make use of the minority language as well. In Twitter conversations, the language choices of users are also influenced by the language of the tweet they are responding to. Furthermore, while many tweets are written in Dutch to reach a broader audience, users often switch to the minority language during a conversation.

Part III

Computational Folkloristics

So far, this dissertation considered variation in text from a sociolinguistics perspective. In this part of the dissertation, variation in text is studied from the perspective of folkloristics, and more specifically, the focus is on folk narratives. Examples of folk narratives are *Little Red Riding Hood*, *Cinderella* and the urban legend about the *Vanishing Hitchhiker*. Different versions of a story appear due to oral and written transmission over time. For example, below are two versions of the same story as found in the Dutch Folktale Database, where the specific pet (dog or cat) and the ending are different from each other.

> **Story STSAG974**: A woman kills her dog by putting him in the microwave to dry. She successfully sues the manufacturer, because the instructions did not mention not to put pets in the microwave.

> **Story STSAG376**: A woman puts her cat in the microwave, she then successfully claims compensation from the manufacturer. Since then, the manufacturer warns against such practices.

The increasing digitization of folk narratives enables the use of computational approaches to help understand and model variation in folk narratives. Like in the context of sociolinguistics, variation may lead to change over time, and thus analyzing the variation in folk narratives could shed more light on how narratives develop over time and across space. Furthermore, modeling the variation could lead to improved tools for the automatic enrichment of meta data, which can be used for organizing and analyzing folk narrative data.

In **Chapter 9**, we discuss automatic identification of tale types, a concept frequently used by folktale researchers to group the different variants of a particular story. In **Chapter 10**, we then perform crowdsourcing experiments to study how experts and non-experts perceive narrative similarity, and to what extent their judgements correspond to the concept of tale types.

9

Automatic Identification of Tale Types

This chapter is based on D. Nguyen, D. Trieschnigg, M. Theune "Folktale Classification using Learning to Rank", In Proceedings of the 35th European Conference on IR Research (ECIR 2013), pages 195-206, Moscow, Russia, 2013 [Nguyen et al., 2013b]

9.1 Introduction

In this chapter we present work on automatically determining the tale types of narratives. Our work is guided by the *type indexes* that folktale narrative researchers have developed to classify and to organize stories according to *tale types*. In our experiments, we limit our focus to two internationally recognized tale type indexes. The first is the frequently used Aarne-Thompson-Uther (ATU) type index [Uther, 2004] that covers many fairy tales, but also legends, jokes and other folktale genres. The second is the Type Index of Urban Legends proposed by Brunvand [2012].

The goal, then, of our work is to automatically determine the tale types of stories. In particular, we cast this as a ranking problem, where the goal is to assign the highest rank to the most applicable tale types. This serves multiple purposes. First, with the increasing digitization of folktales [Abello et al., 2012, La Barre and Tilley, 2012, Meder, 2010], there is a need to (semi-)automate the identification of tale types. Second, such a system could help discover new relationships between stories. And as discussed later in this chapter, this task is related to problems such as detection of text reuse, plagiarism and paraphrase detection.

The goal of this work is to be able to determine the correct tale type of a given folktale. This work adds a novel viewpoint on text similarity. Text similarity can be defined on many levels on the similarity spectrum [Metzler et al., 2005], with document identity on the one end, and topical similarity on the other end. Text reuse [Bendersky and Croft, 2009, Metzler et al., 2005], that includes addition, rewriting or removal of text, is viewed as lying in the middle of the spectrum. Story similarity, as we view it, bears many similarities to text reuse. Stories with the same tale type can be seen to have originated from a common template/model. However, the similarity goes beyond lexical or topical similarity, in the sense that it is based on events, motifs

(narrative elements) and participants of the narrative. Stories are regarded as being of the same type if they match on a more abstract level than just the lexical words (for example locations do not have to match literally), in contrast to text reuse.

We start with a discussion of related work (Section 9.2). Then the dataset is presented (Section 9.3). Next, we describe the experimental setup (Section 9.4) and discuss the results (Section 9.5). We conclude with a summary (Section 9.6).

9.2 Related Work

Fisseni and Löwe [2012] investigated how people perceive story similarity using a user study. Story variations were created by applying character substitutions and varying the order (e.g., reversed temporal order) and style. They found that people focus mostly on motifs, linguistic features and content, and less on the structure of a story when deciding whether two stories are the same.

Friedland and Allan [2008] studied the identification of similar jokes. They framed it as a ranking problem like we do here. Their approach used a bag of words model and abstraction of words using manually constructed word lists. In our work we aim to develop an approach that does not rely on manually constructed lists. Their tale types were identified heuristically and not motivated by an existing type classification system, while we use existing classification systems used by folktale researchers and consider multiple genres.

As mentioned in the introduction, our problem is similar, but not identical to problems such as identification of text reuse, paraphrasing and plagiarism detection. Clough et al. [2002] defined multiple levels of text reuse (wholly derived, partially derived and non derived) and experimented with n-gram overlap, greedy string tiling and sentence alignment using a news corpus. Metzler et al. [2005] looked at text reuse on the sentence and document level.

Our problem is also related to the TDT story link detection task [Allan, 2002], that involved determining whether two stories are discussing the same event (e.g., the *Oklahoma City bombing* topic) in the news domain. Most approaches to the story link detection task relied on text similarity. For example the cosine similarity and the clarity metric have been found to be very effective [Allan et al., 2000, Lavrenko et al., 2002]. In addition, many approaches focused on matching named entities. However, in our problem, stories do not need to match on exact details such as named entities.

Paraphrase detection [Androutsopoulos and Malakasiotis, 2010] involves detecting texts that convey the same information. Used methods include textual similarity measures, as well as using the structure, for example by matching dependency trees. Research in this area has mostly focused on phrases and sentences. The texts in our dataset (as described later) are much longer.

Our problem also shares aspects with plagiarism detection (e.g., Clough [2003]), in particular when certain parts of text are paraphrased. For example Nawab et al. [2012] experiment with query expansion using sources such as WordNet and a paraphrase lexicon to measure text similarity on a semantic level. However, some aspects of plagiarism are not applicable to our problem. This holds in particular for cues that identify inconsistencies in text (such as style and vocabulary).

9.3 Tale Type Indexes

Our dataset is derived from the Dutch Folktale Database. We only consider stories that are written in standard Dutch (the collection also contains many narratives in historical Dutch, Frisian and Dutch dialects). In this chapter we restrict our focus to the two type indexes mentioned in the introduction, the ATU index [Uther, 2004] and the Type Index of Urban Legends [Brunvand, 2012]. We created two datasets based on these type indexes. For each type index, we only keep the tale types that occur at least two times in our dataset. The frequencies of the tale types are plotted in Figures 9.1 and 9.2. Many tale types only occur a couple of times in the database, whereas a few tale types have many instances.

9.3.1 Aarne-Thompson-Uther (ATU)

Our first type index is the Aarne-Thompson-Uther classification (ATU) [Uther, 2004]. Examples of specific tale types are *Red Riding Hood (ATU 0333)* and *The Race between Hare and Tortoise (ATU 0275A)*. The index contains tale types hierarchically organized into categories (e.g., *Fairy Tales* and *Religious Tales*). We discard stories belonging to the *Anecdotes and Jokes* category (types 1200-1999), since the tale types in this category are very different in nature from the rest of the stories[1]. The average number of words per story is 489 words.

9.3.2 Brunvand

Our second type index is proposed by Brunvand [2012] and is a classification of urban legends. Examples of tale types are *The Microwaved Pet (BRUN 02000)*, *The Kidney Heist (BRUN 06305)* and *The Vanishing Hitchhiker (BRUN 01000)*. The stories have on average 158 words.

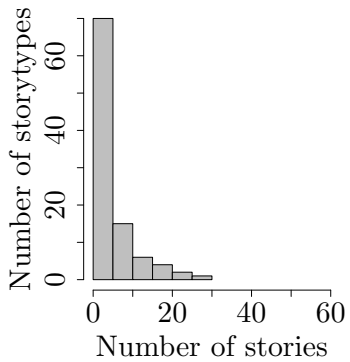


Figure 9.1: ATU tale type frequencies

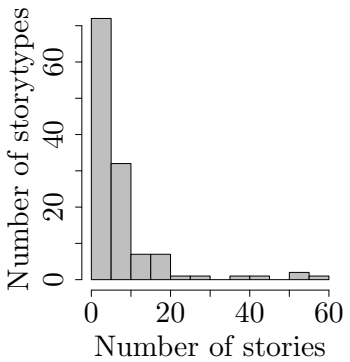


Figure 9.2: Brunvand tale type frequencies

[1] As was suggested by a folktale researcher. Tale types in the *Anecdotes and Jokes* category are mostly based on thematic similarity, while others are based on plot.

9.4 Experimental Setup

In this section we describe our general experimental setup as well as the specific features used.

9.4.1 General Setup

Goal. We cast the problem of determining the correct tale type as a *ranking* problem. Given a story, the system should return a ranking of the tale types for that story. We chose a ranking approach, since there are many tale types, and most of them only have a few instances in our dataset. In addition, in an actual application new tale types could be added over time when new folktales are identified. A ranking of tale types is also useful when used in a semi-automatic system: annotators are presented with the list and can choose the correct one. In addition, a ranking of tale types can easily be converted into a classification, for example by just taking the top ranked tale type as the predicted label.

Evaluation. We will evaluate our approach by the *Mean Reciprocal Rank (MRR)*. We use a rank cutoff, by only considering documents in the top 10. We will also evaluate using the *accuracy*, simulating a classification setting. The highest ranked label is then taken as the predicted class.

We use Terrier [Ounis et al., 2006] as our retrieval component, sofia-ml [Sculley, 2009] as our learning-to-rank toolkit and the Frog tool [Van den Bosch et al., 2007] to obtain POS tags (CGN[2] tagset) and dependency tags.

9.4.2 Baselines

We explore the following baselines, all ranked using BM25:

- *Big document model*. For each tale type, we create a big document with the text of all stories of that particular tale type. We then issue a query, containing the text of our input document, on these big documents. The result is a ranking of tale types. This is similar to the big document models used in Distributed Information Retrieval (e.g., Callan et al. [1995] and Si et al. [2002]), with stories as documents and tale types as collections.

- *Small document model*. For a given story, we issue a query with the text of the story on an index with individual stories. A ranking is returned by ordering tale types based on the individual stories that are ranked (duplicates not taken into account). When taking the top ranked label as the class, this is the same as a Nearest Neighbour classifier (k=1).

Experiments showed that the small document approach was more effective than the big document approach (as discussed in the results section). We therefore aim to improve this baseline in our further experiments.

[2]Corpus Gesproken Nederlands (Spoken Dutch Corpus), http://lands.let.kun.nl/cgn/ehome.htm.

9.4.3 Learning to Rank

Compared to traditional information retrieval methods, learning to rank [Liu, 2011] allows researchers to easily add features to their ranking method. We use the sofia-ml toolkit [Sculley, 2009] with the SGD-SVM learning algorithm and $\lambda = 0.1$. Using learning to rank, we aim to improve the *small document approach* by incorporating a variety of features. Our proposed method contains the following steps.

- Retrieve an initial set of candidate stories using BM25.
- Apply learning to rank to rerank the top 50 candidates.
- Create a final ranked list of tale types, by taking the corresponding labels of the ranked stories and removing duplicates.

9.4.4 Features

We now describe the features that are used in our learning-to-rank setting. All features are normalized within a query. We explore features based on lexical similarity, features that match on a more abstract semantic level, and features that reflect the big document baseline.

I Information Retrieval measures (IR)
These features indicate the score of the query on the text using the BM25 model. We experiment with three types of queries, resulting in three features: fulltext (*BM25 - Full text*), only nouns (*BM25 - nouns*) and only verbs (*BM25 - verbs*). Note that ranking only on the first feature *BM25 - Full text* results in our small document baseline system.

II Lexical Similarity (LS)
These features represent the similarity of the two texts measured using Jaccard and TF-IDF similarity, and calculated on the following token types: unigram, bigrams, character n-grams (2-5), chunks, named entities, time and locations. Location and time words were extracted using Cornetto [Vossen et al., 2007], a lexical semantic database for Dutch, if they were a hyponym of *location* or *timeunit:noun*. The motivation for using these features is that locations (e.g., *house, living room, church*) and time (e.g., *day, September, college year*) can play important roles in the plot of a story.

III Similarity to all stories of the candidate's tale type (bigdoc)
This is a feature that resembles the *big document model* as was used in the baseline. This measures the similarity of the input story to the tale type of the candidate by taking all stories of that tale type into account. As feature we use the retrieval score of the big document of the tale type of our candidate story. Again, we experiment with three types of queries, resulting in three features: fulltext (*Bigdoc - BM25 - Full text*), only nouns (*Bigdoc - BM25 - nouns*) and only verbs (*Bigdoc - BM25 - verbs*).

IV Subject Verb Object (SVO) triplets
Events are central to the identity of a story. We aim to capture these using *verb(subject, object)* (SVO) triplets, such as *lives(princess, castle*) or partial triplets such as *disappear(driver,)*. Recently, triplets have been explored to distinguish between stories and non-stories [Ceran et al., 2012]. Triplets are much sparser than just words; we therefore explore allowing partial matches, and abstraction of verbs to a higher semantic level using VerbNet [Kipper-Schuler, 2005].

Triplet extraction. For each extracted verb, the system tries to find a matching subject or object by traversing the dependency graph (obtained using the Frog parser) and matching on the relation *su* for the subject, or *obj1*, *obj2* for the object. Only certain POS tags such as nouns, pronouns and named entities are taken into account. Manual inspection showed that the triplets are very noisy, often because of errors by the Frog parser. Each word is replaced by its lemma as given by the Frog parser.

Features. To overcome sparsity, we also use features that allow partial matches. For each abstraction level and similarity measure, we create four features representing *Exact* overlap, *Subject-Verb(SV)* overlap, *Object-Verb(OV)* overlap and *Subject-Object(SO)* overlap. We use the Jaccard and TF-IDF similarity.

Abstraction. Abstraction of triplets reduces the sparsity of the features, and allows stories to match on a more abstract level. We experiment with no abstraction, and abstracting the verbs. Abstraction of verbs is done using VerbNet [Kipper-Schuler, 2005], an English verb lexicon, that groups verbs into 270 general classes. Using relations between Cornetto and Wordnet, a mapping is made between verbs in a story and English verbs. For example, the following Dutch verbs are mapped to the 'consider-29.9' class in VerbNet: *achten (esteem), bevinden (find), inzien (realise), menen (think/believe), veronderstellen (presume), kennen (know), wanen (falsely believe), denken (think)*. With verbs, we also experiment with partial matches, but do not add a feature that measures the overlap between subject and object, since these have not been changed.

Reduction of sparsity. To illustrate the reduction of sparsity using the methods described, the number of unique elements are shown in Table 9.1. We find that when allowing partial matches, the number of unique elements decreases a lot (from over 10,000 to 6,000-7,000). When verbs are abstracted, the counts decrease even more. This is partly caused by verbs that were discarded because VerbNet or Cornetto did not cover them.

Abstraction	Exact	Subject-Object	Subject-Verb	Object-Verb
None (Original)	10,260	6,325	6,416	6,925
Verb	8,924	NA	4,505	5,588

Table 9.1: Number of unique elements - Brunvand (index)

9.5 Results

9.5.1 Data

For each type index (ATU and Brunvand) we created a dataset. First, we divided the documents into two sets. The query set contains the stories (documents) for which we need to find the tale types. The index set contains the stories that need to be ranked. The corresponding labels of these stories can then be used to predict a tale type for the query.

Only tale types were kept that had at least 2 stories in the folktale database. Then, for each tale type one document was assigned to our index set, and one document was assigned to our query set (train/dev/test). The rest of the documents for that particular tale type were assigned randomly to either the index or the query set, until the query set was the desired size (e.g., 150 with ATU). The query set then was randomly divided into a train, development and test set while ensuring the desired sizes. Statistics are listed in Tables 9.2 and 9.3.

	Index	**Train**	**Dev**	**Test**
Nr. documents	400	75	25	50
Nr. tale types	98	59	24	43

Table 9.2: ATU dataset statistics

	Index	**Train**	**Dev**	**Test**
Nr. documents	687	175	50	75
Nr. tale types	125	92	40	50

Table 9.3: Brunvand dataset statistics

9.5.2 Baselines

The results for our baseline methods as described in Subsection 9.4.2 can be found in Tables 9.4 and 9.5. We find that for both datasets, the smalldoc baseline performs better, although the difference is much larger for the ATU dataset.

	MRR	**Accuracy**
Smalldoc	0.7779	0.72
Bigdoc	0.4423	0.36

Table 9.4: Baseline results - ATU

	MRR	**Accuracy**
Smalldoc	0.6430	0.56
Bigdoc	0.6411	0.56

Table 9.5: Baseline results - Brunvand

For our reranking approach, we rerank the top 50 stories obtained using the smalldoc approach. For the ATU, we find that the correct tale type is in the top 50 results for 49 out of 50 stories. For the Brunvand index, the correct tale type is present in

71 out of 75 stories. This gives an upper bound on the reranking performance and confirms that only reranking the top 50 stories is sufficient for almost all queries.

9.5.3 Feature Analysis

We evaluate the effectiveness of the feature types by adding them to the baseline model (smalldoc). The results can be found in Tables 9.6 and 9.7.

	MRR	Accuracy
Baseline	0.7779	0.72
+ Bigdoc	0.8367	0.78
+ IR	0.8049	0.76
+ LS	0.7921	0.72
+ Triplets	0.8016	0.72
All	0.8569	0.82

Table 9.6: Feature analysis - ATU

	MRR	Accuracy
Baseline	0.6430	0.56
+ Bigdoc	0.7933	0.72
+ IR	0.7247	0.61
+ LS	0.6810	0.60
+ Triplets	0.6600	0.59
All	0.8132	0.76

Table 9.7: Feature analysis - Brunvand

The performance gains are high compared to the baseline system. The smalldoc baseline had a higher performance on the ATU index, but when including all features the results on the Brunvand index approaches that of ATU.

We also observe that all feature types improve performance. For both datasets the big document features are highly effective. Note that the big document features capture a different type of evidence than the other features. The big document features take similarity to all stories of a particular tale type into account, while the other features reflect the similarity between a pair of documents (the input document and the candidate).

Triplets improve performance, but not by much. We analyze the performance of the triplets in more detail by varying the features based on abstraction level and matches as shown in Tables 9.8 and 9.9. For both datasets, allowing partial matches when not using any abstraction improves the MRR. However, with ATU the accuracy decreases slightly. Abstraction using verbs does not perform well. When adding both feature types (no abstraction and verb abstraction) the performance does increase.

The performance of the triplets is suboptimal for several reasons. First, manual inspection showed that mistakes of the parser caused triplets to be missed or extracted incorrectly. In addition, we rely on general purpose semantic lexicons such as VerbNet and Cornetto. The coverage of such general lexicons might not be sufficient for specific domains such as folktales.

The most important features (i.e., the features with the highest weight) are shown in Tables 9.10 and 9.11. We observe that the models learned for ATU and Brunvand have the same features in the top 3. Important features are the big document features and lexical similarity (unigrams, TF-IDF). The fact that they share so many features indicates that the ATU and Brunvand indexes are very similar in how tale types were defined, and that the same types of evidence are important for finding the correct tale types.

Abstr.	Matching	MRR	Acc.
No	Exact	0.7762	0.72
No	Exact, partial	0.7902	0.70
Verb	Exact, partial	0.7475	0.68
No, Verb	Exact, partial	0.8016	0.72

Table 9.8: Triplet analysis - ATU

Abstr.	Matching	MRR	Acc.
No	Exact	0.6422	0.56
No	Exact, partial	0.6556	0.57
Verb	Exact, partial	0.6419	0.56
No, Verb	Exact, partial	0.6600	0.59

Table 9.9: Triplet analysis - Brunvand

Feature	Weight
Bigdoc: BM25 - nouns	0.179
Bigdoc: BM25 - full text	0.158
LS: unigrams - TFIDF	0.109
Bigdoc: BM25 - verbs	0.069
Triplets: SO match, Jaccard, no abstraction	0.063

Table 9.10: Top 5 features - ATU

Feature	Weight
Bigdoc: BM25 - full text	0.209
Bigdoc: BM25 - nouns	0.204
LS: unigrams - TFIDF	0.065
IR: BM25 - nouns	0.062
Bigdoc: BM25 - verbs	0.051

Table 9.11: Top 5 features - Brunvand

Overall, we believe that the results are very encouraging; a system using all features obtains a high MRR (above 0.8), making this a promising approach to use in a setting where annotators of new stories are presented with a ranked list of possible tale types. However, one should keep in mind that we still need to investigate the performance of the approach for other type indexes and texts written in dialects and historical variations.

9.5.4 Error Analysis

We manually analyzed stories that had a low reciprocal rank using the run with all features. With both the Brunvand index and the ATU index, errors occurred because the system found similar stories that matched on the writing style instead of the actual plot. This happened mostly with stories that had a distinguishing style (for example because they were told by the same narrator in a particular setting), and even more when the input story was very short (often with stories of the Brunvand index) or if the correct tale type had only a few instances. Thus, if not much content was available to match on plot, our system sometimes incorrectly judged stories to be similar due to style.

With the ATU index, we also observed errors where the system judged stories to be similar because they matched on content words, and not on the actual plot. They might share words related to the location of the story (e.g., *the woods*) or the characters (e.g., *father, son*). This happened in particular with very long stories.

In general, challenging stories were stories with very distinguishing writing styles, and stories with extreme lengths (either very short or very long). Future work should focus on improving performance for these types of stories.

9.6 Conclusion

This chapter presented a study on classifying stories according to their *tale types*, a concept used by folktale researchers to organize folktales. Two type indexes were used as the basis of our experiments: the Aarne-Thompson-Uther (ATU) type index [Uther, 2004] and the Type Index of Urban Legends [Brunvand, 2012].

We framed the problem as a ranking problem, where the goal was to rank tale types for a given story. We employed a nearest neighbours approach, by ranking individual stories based on their similarity with the given story, and taking the corresponding label as the predicted class. High performance gains were achieved using learning to rank, with features inspired by approaches from distributed information retrieval and features that compare subject-verb-object triplets.

The problem of classifying stories according to their tale type presents a new angle on text similarity, and we believe further research on this could also provide new insights into related problems like text reuse, paraphrase detection, story link detection and others. The developed methods could also be useful for classification and organization of other types of narrative data, such as literary fiction, and data reflecting oral transmission, such as interviews [De Jong et al., 2008].

The results were very encouraging, however for such a system to be useful to folktale researchers, stories written in a dialect or historical language variety should be considered as well. In addition, other tale type indexes should also be covered.

10

Perception of Narrative Similarity

This chapter is based on D. Nguyen, D. Trieschnigg and M. Theune, "Using Crowdsourcing to Investigate Perception of Narrative Similarity", In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pages 321-330, Shanghai, China [Nguyen et al., 2014c]

10.1 Introduction

Measuring the similarity between documents is essential in many applications. For example, clustering systems are inherently dependent on the used similarity measure. However, for many tasks it is unclear what an appropriate similarity measure should be. Multiple dimensions might play a role (e.g., topic, genre), and different users might not agree on which dimensions are important. So far, most research on text similarity has focused on topical or semantic similarity, thereby ignoring dimensions that might be important from a user's perspective. Research investigating *how humans perceive similarity* between documents has been scarce so far.

Understanding how humans perceive similarity is useful in many situations. It could guide the development of similarity metrics to correspond better with human perception. Clustering systems could benefit by knowing along which dimensions documents should be clustered. And, it could aid in the creation of more suitable datasets [Bär et al., 2011] and evaluation metrics [Kim et al., 2013].

In this chapter, we study perception of similarity in the domain of folk narratives (such as fairy tales and urban legends). While the previous chapter focused on automatic identification of tale types, we now study to what extent perception of narrative similarity actually corresponds to tale types.

With the increasing digitization of folk narratives [Abello et al., 2012, La Barre and Tilley, 2012, Meder, 2010], there is a need for better search and clustering systems [Grundkiewicz and Gralinski, 2011]. However, so far little is known about *how humans perceive narrative similarity*. For example, take the following two narratives (summaries are shown):

Narrative 1 Some men sat around a fire. Nine cats came to sit near the fire, and the men got nervous. One of the men threw fire at the cats with a fire shovel. The next day, nine women in the village lay in bed with burned buttocks.

Narrative 2 Every afternoon a large black cat came to sit by the fire in the kitchen. The people knew about a witch in the neighborhood. One afternoon the cat came again. The woman threw a pan with hot oil at the cat's neck. The next day, the neighbor wore a white scarf, she had burned her neck.

The characters in both narratives are witches, humans and cats. Although the exact events are different, both narratives share a story line: The cats are actually witches, who are recognized by their wounds in their human form. Some people might even recognize that these narratives served a common purpose: demonstrating that witches are real.

Folktale researchers could recognize these narratives as belonging to the same tale type (titled '*Witch hurt as animal; woman turns out to be wounded the next day*', SINSAG 0640). The above example illustrates that similarity between narratives can be based on various dimensions (e.g., characters, plot, theme/purpose, tale types). The goal of our study is to shed light on how narrative similarity is perceived. Which dimensions do people consider when judging narrative similarity, and do non-experts pay attention to different dimensions than experts?

Empirical studies on narrative similarity have only been done on a small scale so far (e.g., Fisseni and Löwe [2012], Kypridemou and Michael [2013]). This study is the first large-scale empirical study on narrative similarity. We collect data from a large number of non-experts using crowdsourcing by asking them to rate similarity between narrative pairs. Crowdsourcing enables rapid and low-cost data collection and has been used in various tasks, including studying the similarity of multimedia files [Vliegendhart et al., 2012], music [Lee, 2010, Urbano et al., 2010] and documents [Zengin and Carterette, 2013]. Data on how experts judge narrative similarity was collected in two ways: 1) By asking experts directly to rate similarity, in the same way as data obtained from non-experts. 2) By using the tale types that the narratives are (manually) classified with. For example, narratives belonging to the same tale type are expected to receive higher similarity judgements.

To summarize, our contributions are as follows:

- We show how crowdsourcing can be used to collect data for studying perception of similarity (Section 10.3).
- We identify the dimensions that play a role in perception of narrative similarity (Section 10.4).
- We show that non-experts and experts have a different perception of narrative similarity and that tale types do not fully correspond with non-expert perception of narrative similarity (Section 10.4).
- We show that automatic methods correspond reasonably well with judgements of the crowd (Section 10.5).

10.2 Related Work

In this section we discuss related work on empirical studies of similarity, narrative similarity, and crowdsourcing.

Empirical studies of similarity. Our study follows recent research on the human perception of similarity in various domains, for example images [Nathalie et al., 2002], style of paintings [Kovashka and Lease, 2010], text [Bär et al., 2011], multimedia files [Vliegendhart et al., 2012], music [Lee, 2010, Urbano et al., 2010] and videos [Cherubini et al., 2009]. Some of these studies also investigated which dimensions play a role in human judgement of similarity, for example judgements of multimedia files [Vliegendhart et al., 2012] and preference judgements of search result lists [Kim et al., 2013]. The influence of structure, style, and content on text similarity have been studied by Bär et al. [2011]. Compared to these previous studies, we collect and compare judgements from both experts and non-experts.

Narrative similarity. Narratives have traditionally been studied by focusing on their plot structure. This is also reflected in research focusing on narrative similarity [Michael, 2012]. Scholars have typically approached this by developing formal systems to represent and find analogies between plot structures of narratives. The approaches rely on in-depth annotations of story structures by humans and as a result have stayed either theoretical [Michael, 2012] or have only been tested on small amounts of data (e.g., 1 narrative pair [Fay, 2012], or 26 Aesop fables [Elson, 2012]).

A different line of work involves a more computational approach but uses shallower features. For example, automatic classification of folk narratives [Nguyen et al., 2013b] or jokes [Friedland and Allan, 2008]. These methods focus on lexical similarity and do not study which dimensions play a role in perception of similarity. In addition, their ground truth labels provide only a binary view of similarity.

Two recent studies investigated perception of narrative similarity, but on a very small scale (16 narrative pairs [Kypridemou and Michael, 2013], variations of two stories [Fisseni and Löwe, 2012]). Their results have suggested that non-experts also focus on dimensions other than structural similarity [Fisseni and Löwe, 2012], and that humans are more likely to rate narratives as similar if they have a common summary [Kypridemou and Michael, 2013].

Crowdsourcing. Crowdsourcing enables the collection of large amounts of data with low costs using platforms such as Amazon Mechanical Turk and Crowdflower. We target Dutch workers in our study, who are fast and of high quality [Pavlick et al., 2014]. Recent studies have explored how the crowd can be used to infer taxonomies [Eckert et al., 2010] and clusterings from data [Gomes et al., 2011]. However, such approaches need judgements for each item. An alternative approach is to use the crowd to learn a similarity metric, which can then be applied on large, growing collections (e.g., Yi et al. [2012]). Our study follows the latter line of thought, by aiming to obtain insight into perceived similarity and develop automatic methods to measure similarity.

10.3 Data

In this section we describe the collection of the narrative similarity judgements.

10.3.1 Preprocessing

We selected a subset of the narratives from the Dutch Folktale Database. We restricted the set to narratives that were easily readable (based on writing style and length) and had all the required metadata to support our analyses. More specifically, we only kept narratives with the following requirements: 1) Written in Standard Dutch [Trieschnigg et al., 2012]; 2) With an annotated tale type, genre and collector; 3) Of intermediate length (between 10 and 250 tokens).

10.3.2 Task Design

To collect data, we designed a human intelligence task (HIT). The task was given to both experts and non-experts. Data from non-experts was collected by posting the HITs on Crowdflower, a crowdsourcing platform. We asked workers to judge the similarity between pairs of stories. We provided as few instructions as possible (e.g., by *not* mentioning terms like plot), so that workers were not influenced by us to pay attention to certain dimensions. Small pilot experiments were carried out while developing the design of the HIT. Each HIT consisted of 6 pairs of narratives (5 pairs + 1 pair with gold labels), and several survey questions. Each HIT was initially judged by 3 workers. We collected additional judgements for narrative pairs with large standard deviations of the judgements, such that all HITs received 3 to 5 judgements. We paid 40 US dollar cents for each HIT.

Survey. To study the influence of characteristics of people on how they perceive narrative similarity, we included several survey questions:

- Gender (male/female)
- Age (in years)
- Location (one of the Dutch provinces, or other)
- Highest completed education[1]
- How often do you read a book?[2]
- What kind of books do you read?[3]
- How often do you watch a movie?[2]

[1] No education, Elementary school, Pre-vocational secondary education, Senior general secondary education/pre-university secondary education, Secondary vocational education, Higher professional education, University education.

[2] Daily, several times a week, several times a month, never.

[3] Fiction, non-fiction, both.

Similarity judgements. Workers were presented with pairs of narratives for which they were asked to rate the similarity on a scale from 1 (no similarity) to 5 ((almost) the same). A similar scale was used in related studies [Lee et al., 2005, Zengin and Carterette, 2013]. Workers were also asked to provide a short motivation for their rating in a free text field for each narrative pair. The order of the displayed narrative pairs within a HIT was randomized.

Gold labels. To improve the detection of spammers, we manually created 12 narrative pairs with 'gold labels'. Workers who provided ratings deviating from these labels were identified as potential spammers. We created pairs with high similarity by copying an existing story and making small edits in spelling, punctuation, word order, etc. For such pairs, we expected a similarity judgement of 4 or 5. We also selected pairs with very low similarity by manually selecting stories that had nothing in common (e.g., plots and characters are completely different). For such pairs, we expected a similarity judgement of 1 or 2. We did not inform workers about their performance on the pairs with gold labels.

10.3.3 Pair Selection

Selecting pairs at random would generate many pairs with little similarity. Therefore, we control the selection of pairs as follows:

1. Similarity between narratives classified with the *same* tale type and *same* genre.
2. Similarity between narratives classified with the *same* tale type but *different* genre.
3. Similarity between narratives with the *same* genre, but *different* tale types.

Under conditions 1 and 2, the narratives are the same based on their tale types. We include pairs by varying the lexical similarity of these pairs based on cosine similarity. A threshold (based on data analysis, see below) was calculated to distinguish between low, mid and high similarity. We include an equal number of pairs from each bin.

Under condition 3, we only include pairs that have a high cosine similarity. We assume that pairs with low or mid similarity are less interesting, since they have little lexical similarity and are also not similar based on their tale types.

Thresholds. We first group all narratives by tale type. For each tale type, we randomly select a pair of narratives and calculate the cosine similarity. Based on the samples, we take their 33% and 67% boundaries to define the thresholds to distinguish between low, mid and high cosine similarity.

Same tale type, different genre. We select pairs of narratives that are classified under the *same* tale type but under different genres. We first generate candidate pairs:

```
For each tale type:
    Group all narratives by genre
    If #genres > 1:
        Sample pairs across genre (up to 3 per bin)
```

The final selection is made by sampling from all the candidate pairs, given the desired distribution for the cosine similarity bins and the number of pairs to include.

Same tale type, same genre. We study similarity between pairs belonging to the same genre and tale type, but with varying levels of cosine similarity.

```
For each genre:
    For each tale type:
        Sample up to 3 pairs per cosine bin
```

The final selection is made by sampling from the candidate pairs ensuring an equal distribution across cosine similarity bins, given a desired genre distribution and the total number of pairs needed.

Same genre, different tale types. We also select pairs belonging to different tale types but with a high cosine similarity. We create candidate pairs as follows:

```
For each genre:
    For each tale type:
        Select up to 3 pairs with high cosine similarity and with
        one of the narratives belonging to this tale type.
```

The final selection is made by sampling from the candidate pairs given a desired genre distribution.

10.3.4 Groups

The designed HITs were given to two different groups: crowdworkers and folk narrative researchers.

Crowdworkers. We posted the tasks on CrowdFlower and targeted workers from the Netherlands. The jobs ran between April 4, 2014 and April 27, 2014. We launched the jobs in several batches, to prevent workers from doing the task too many times and to ban spammers in between. Potential spammers were identified by the following criteria:

- Inconsistent demographics. Most workers completed multiple HITs. We assumed workers with inconsistent demographics information to be spammers.
- Time spent on judgement. Workers who spent less than 3 minutes on a HIT (based on data analysis).
- Gold labels. Workers whose judgements did not match the gold labels.
- Motivation. We manually inspected the answers on the motivation questions. Spammers answered with random characters, by copying parts of the narratives or by always answering with the same sentence.

We manually checked if workers identified using these criteria were spammers. Such workers were excluded from the dataset and blocked for all next HITs. We collected in total 923 HITs (150 workers). 619 HITs (80 workers) were kept after

filtering spammers. Figure 10.1 shows the average times spent on a HIT for workers (median: 677.5 seconds).

Folktale researchers. We also asked three senior folktale researchers (all with a researcher/lecturer position) to do the same task. We selected 40 narrative pairs, ensuring that we included at least 2 pairs from each bin according to our sampling method described above. HITs were the same as presented to the crowdworkers, but without the pairs with gold labels (thus resulting in 5 narrative pairs per HIT).

10.3.5 Statistics

The statistics of the collected data are shown in Table 10.1.

Statistic	Crowdworkers	Experts
# unique narrative pairs	1002	40
# completed HITs	619	24
# persons	80	3

Table 10.1: Dataset statistics

10.4 Analysis

In this section, we analyze the collected data. We start with studying the demographics of the workers and then continue with an analysis of their similarity judgements.

10.4.1 Workers

Demographics. Workers are mostly men (66%), but are relatively spread across different ages and education levels. The workers are spread throughout the Netherlands, but most workers come from the west of the Netherlands (where the population density is higher as well).

Reading and watching movies. Table 10.2 summarizes the users' reading and movie-watching behavior. Most people read both fiction and non-fiction (52, 65%), and some read only fiction (20, 25%). A small fraction only reads non-fiction (8, 10%). We code the education responses, and movies and reading behavior by converting each category to an integer. We find that the education level is highly correlated with frequencies of reading a book (Spearman's $\rho = .424$, $p < 0.001$). The education level is negatively correlated with the frequency of watching movies (Spearman's $\rho = -.229$, $p < 0.05$). Watching movies and reading books is not correlated (Spearman's $\rho = -.086$, not significant).

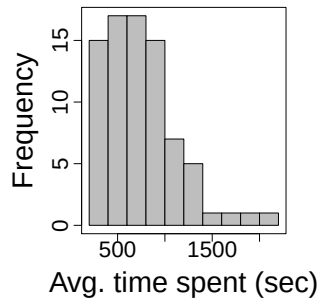


Figure 10.1: Average time spent on task

How often	B	M
Never	11	0
Couple of times a month	44	39
Multiple times a week	18	37
Daily	7	4

Table 10.2: Frequencies of reading books (B) and watching movies (M)

10.4.2 Understandability Ratings

Workers also indicated how well they understood the pair of narratives on a scale from 1 (not understandable) to 5 (well understandable) (Figure 10.2). Manual inspection of pairs with lower ratings, revealed that crowdworkers had difficulty with language use that was less standard (e.g., dialects, slang, uncommon words), unconventional style and structure. Narratives from the more modern genres (i.e., urban legends and jokes) are understood better than narratives from the older genres (i.e., legends and fairy tales), see Table 10.3.

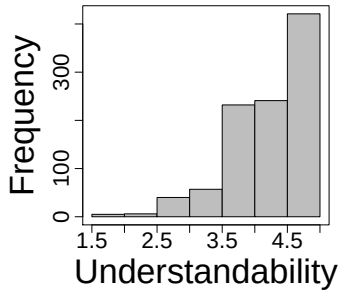


Figure 10.2: Understandability ratings

Genre	U
Urban legends	4.47
Jokes	4.33
Legends	4.12
Fairy tales	4.06

Table 10.3: Understandability (U) and genre

For each worker, we calculate the worker's understandability bias, by calculating the average difference between the worker's score and the average of the scores. We find that higher educated workers tend to give *lower* understandability scores (Spearman's ρ = -.249, $p < 0.05$). While this may seem counterintuitive, they also vary more in their understandability ratings (Spearman's ρ = .278, $p < 0.05$). We found no significant correlations with reading or movie watching behavior. In the remainder of this chapter, we only keep narrative pairs that received an average understandability rating of 3.5 or higher (removing 104 pairs).

10.4.3 Narrative Similarity Ratings

We first analyze the agreement between the judgements. Next, we study the similarity judgements for different conditions. Finally, we study the similarity dimensions by analyzing the free-text motivations.

10.4.3.1 Agreement

Crowdworkers. We first analyze the agreement between crowdworkers. We have in total 80 workers and 898 pairs. For each pair, we have 3 to 5 judgements. Figure 10.3 shows a histogram of the standard deviations of the judgements for each narrative pair. We find that 85% of the pairs have a standard deviation less than 1. We also calculate a user bias (Figure 10.4). For each worker, it is the average over the differences between a judgement made by a worker and the overall mean of the judgements for each narrative pair. Only 17.5% of the workers are on average more than 0.5 points off the mean.

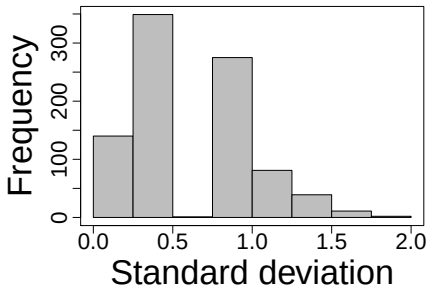


Figure 10.3: Std per narrative

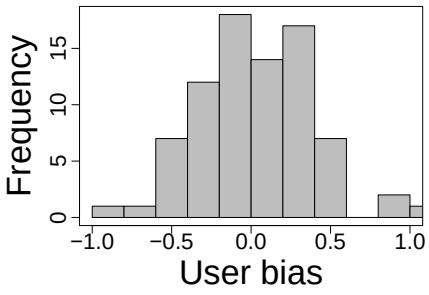


Figure 10.4: User bias

We calculate several agreement metrics (see Table 10.4). Following Lee et al. [2005], we calculate a measure of inter-rater correlation. For each narrative pair, we select at random a judgement and correlate it with the average of the other similarity judgements. We also calculate a pairwise agreement. For each narrative pair, we check the agreement between all pairs of workers. The reported pairwise agreement is the number of pairs workers agreed on divided by the total number of pairs, and is similar to the value reported in Zengin and Carterette [2013]. We also find that mapping the scores to a lower number of categories (1-2, 3, 4-5) leads to a higher pairwise agreement of 0.517.

Metric	Crowd	Experts
Spearman correlation	.556	.778
Pearson correlation	.572	.796
Pairwise agreement	.335	.423

Table 10.4: Agreement crowd and experts

We also analyzed the influence of demographics on agreement. We find that people who read books daily (3) or never (0) tend to agree more within their group (pairwise agreement of 0.433 and 0.444, see Table 10.5). We also tested excluding groups with lower reading frequencies. Only including workers who often read (≥ 2) leads to higher agreement (0.374) than including all workers. No clear trends were observed with watching movies or education.

	Reading frequency			
Criteria	**0**	**1**	**2**	**3**
$\geq$	.335	0.333	0.374	0.444
$=$	.433	0.310	0.340	0.444

Table 10.5: Pairwise agreement and reading frequency

Experts. We calculated the agreement between experts in the same way as with the crowd. Table 10.4 shows the calculated metrics. We find that experts achieve higher inter-annotator agreement than the crowd, probably because their reasoning involves tale types and they agree more on which dimensions are important (see also the next section). We also study to what extent individual experts and the average of the experts correspond with the crowd judgements (Table 10.6). Averaging the expert judgements leads to a higher correlation with the crowd judgements.

	E1	**E2**	**E3**	**Avg. expert**
Crowd	.654	.683	.633	.744

Table 10.6: Spearman correlation of individual experts (E1-3) and average expert with the crowd

10.4.3.2 Analysis

The average similarity for each condition (see Subsection 10.3.3) is shown in Tables 10.7 and 10.8.

Tale types. We first investigate how tale types correspond with judgements by the crowd. We would expect narratives belonging to the same tale type to receive higher ratings than narratives belonging to different tale types.

Narrative pairs (with high cosine similarity) with the same tale type indeed receive higher ratings than pairs (with high cosine similarity) with different tale types (Table 10.7). Figure 10.5 shows a histogram of the similarity ratings for narrative pairs belonging to the *same* tale type. If tale types would correspond strongly with perceived similarity by non-experts, we would see a skewed distribution with most of the ratings being a 4 or 5. Instead, most of the ratings are in the middle and the perceived similarities of the narratives belonging to the same tale type vary widely. Figure 10.6 shows a histogram of the similarity ratings for narrative pairs belonging to *different* tale types. Here, we do see a skewed distribution, with most scores being low (e.g., 1 or 2).

Thus, although narratives belonging to the same tale type tend to be perceived as more similar than narratives belonging to different tale types, tale types do not explain all of the observed variation in similarity judgements by non-experts. This suggests that tale types ignore dimensions that non-experts do find important.

Figures 10.7 and 10.8 show the histograms of expert judgements. The figures reflect that experts use tale types in their research and give more extreme scores than non-experts. Pairs of narratives belonging to the same tale type are mostly rated with

	Urban legends		Jokes		Legends		Fairy tales	
Same tale type, same genre								
Low cosine	2.900	(0.109)	2.119	(0.160)	2.503	(0.133)	2.343	(0.191)
Mid cosine	3.375	(0.134)	2.743	(0.139)	2.793	(0.112)	3.150	(0.268)
High cosine	3.972	(0.089)	3.550	(0.172)	3.536	(0.173)	3.806	(0.194)
Different tale type, same genre								
High cosine	2.095	(0.072)	2.174	(0.070)	2.346	(0.092)	2.106	(0.119)
Same tale type, different genre								
Low cosine	2.226 (0.094)							
Mid cosine	2.721 (0.110)							
High cosine	3.504 (0.121)							

Table 10.7: Mean and standard errors of similarity scores per condition

	All	
Same tale type, same genre		
Low cosine	2.501	(0.077)
Mid cosine	3.008	(0.078)
High cosine	3.719	(0.078)
Different tale type, same genre		
High cosine	2.181	(0.042)

Table 10.8: Mean and standard errors of similarity scores per condition, for all genres

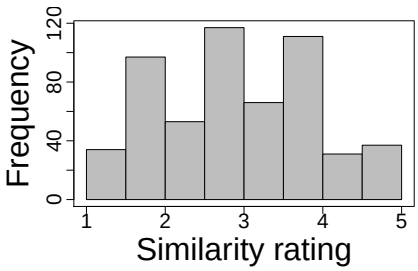


Figure 10.5: Crowd same tale type

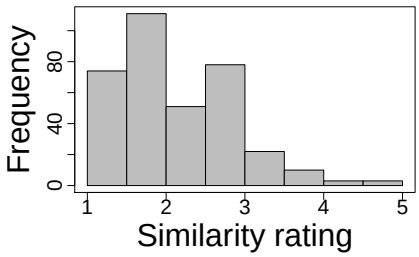


Figure 10.6: Crowd different tale type

a 5, narratives belonging to different tale types often with a 1 or 2. However, we do observe variation indicating that other aspects influence their judgements as well. For example, based on their feedback, we find that experts tend to rate pairs of narratives from broad tale types (e.g., based on theme) lower than tale types defined based on plots.

Genres. Table 10.7 also shows the scores for narratives with the same tale type but classified under *different genres*. We find that narrative pairs with different genres tend to receive a lower similarity judgement than pairs belonging to the same genre. In the next section we study the influence of genre using the provided free-text motivations.

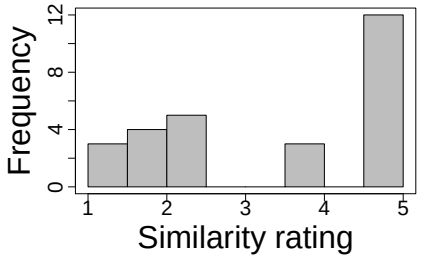


Figure 10.7: Experts same tale type

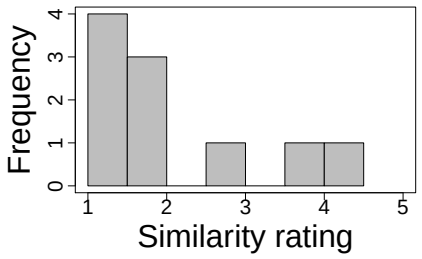


Figure 10.8: Experts different tale type

Cosine similarity. In Table 10.7 we observe that within each genre, a higher cosine bin results in a higher average similarity judgement. In a later section, we experiment how the similarity judgements correspond with various supervised and unsupervised similarity metrics.

10.4.4 Dimensions of Narrative Similarity

For each similarity judgement, we also asked for a free-text answer with a motivation. In this section, we study the importance of different similarity dimensions (e.g., plot, characters) based on these motivations.

Crowdworkers. Motivations given for the narrative pair in the introduction are shown in Table 10.10. The first worker only mentions a similarity in the characters, the other workers also see a similarity in the plot. Note that other dimensions could have (unconsciously) influenced the workers as well, but they did not mention them.

We randomly selected 192 narrative pairs and included all motivations for these pairs (total: 589). The most frequent dimensions were identified after annotating subsets of the data. Each motivation was then manually annotated (Table 10.9) by one coder. A second coder annotated a subset of 64 narratives. Cohen's κ ranged from moderate (e.g., plot: $\kappa = 0.59$, style: $\kappa = 0.66$, theme: $\kappa = 0.59$) to high (e.g., genre: $\kappa = 0.80$, characters: $\kappa = 0.88$, number of details: $\kappa = 1.00$).

The characters, plot, genre and theme were mentioned the most. However, a variety of other dimensions were mentioned as well (e.g., style, number of details). No explanation was given in 18% of the motivations.

For each dimension (except 'none' and 'other'), we annotated whether a *difference* and/or *similarity* was mentioned, e.g. "*plots are different*" (Table 10.9). When workers mentioned the characters, plot or theme, they tended to focus on the similarities between the narratives. However, when they referred to the amount of detail, workers only stated differences.

For most dimensions, whether they are mentioned is influenced by the presented narrative pair. For example, the probability of a random motivation mentioning theme is 0.28. However, knowing that another worker has mentioned theme for the same pair, the probability goes up to 0.51.

To study the importance of these dimensions, we fitted an Ordinary Least Squares model (OLS) with the given score as the dependent variable (Table 10.11). We find that plot, genre and theme are the most important. Characters are not significant after

Dimension	Crowd M	Crowd Sim	Crowd Diff	Crowd P	Experts M	Experts Sim	Experts Diff	Experts P
Characters: The characters or important objects in a narrative (e.g., a princess, a ring)	.43	.88	.17	.80	.51	.74	.44	1.00
Plot: The sequence of events in a narrative	.37	.67	.49	.76	.54	.53	.62	1.00
Genre: For example "*both narratives are jokes*"	.21	.82	.18	.58	.14	.69	.31	1.00
Theme: The central topic/moral (e.g., paranormal events)	.28	.86	.15	.71	.36	.88	.13	1.00
Setting: Where the story is set. This can be more general (e.g., a castle) or a geographic location (e.g., Paris)	.04	.39	.61	.20	.01	1.00	.00	.33
Style: E.g., punctuation, word choice, formal language	.08	.36	.64	.39	.03	.67	.33	.67
Number of details: Length or number of details	.02	.00	1.00	.10	.05	.00	1.00	1.00
Recount facts: E.g., "*narrative 1 could be true*"	.01	.63	.63	.08	.00	-	-	.00
Structure: E.g., repetition of events	.03	.60	.47	.16	.08	.56	.44	1.00
Tale types: E.g., "*both are of the same tale type*"	.00	-	-	.00	.46	.59	.43	1.00
Motifs: Elementary building blocks of narratives	.00	-	-	.00	.06	.43	.71	.67
Other: Such as the narrator, origin of the stories, etc.	.03	-	-	.23	.05	-	-	.67
None: E.g., "*they are not the same*"	.18	-	-	.51	.13	-	-	.67

Table 10.9: Dimensions of narrative similarity. For each group (crowd: 80 persons, 589 motivations, experts: 3 persons, 111 motivations), the table reports the fraction of motivations (M) or persons (P) mentioning a dimension, and for each dimension, the fractions that mentioned similarities (Sim) or differences (Diff).

Not much except they are about a cat *Given score: 2. Dimensions: Characters*
Both narratives are about witches and black cats. Furthermore in both stories the cat gets injured and as a result the woman is also injured. The narratives look very much like each other, but the content differs. Therefore I give it 4 out of 5. *Given score: 4. Dimensions: Characters, Plot*
easy to read, both narratives are about cats who are actually witches who sit at the fire and are thrashed there *Given score: 4. Dimensions: Style, Characters, Plot*

Table 10.10: Translated motivations by crowdworkers

	B	SE
Intercept	2.47***	0.12
Characters.sim	0.03	0.11
Characters.diff	0.04	0.18
Plot.sim	0.99***	0.12
Plot.diff	-0.44***	0.12
Genre.sim	0.27*	0.14
Genre.diff	-0.57*	0.27
Theme.sim	0.21·	0.12
Theme.diff	-0.63**	0.23
Settings.sim	0.46	0.36
Settings.diff	0.14	0.29
Style.sim	0.17	0.29
Style.diff	0.79***	0.21
Num details.diff	1.19**	0.36
Recount facts.sim	0.42	0.49
Recount facts.diff	-0.24	0.49
Structure.sim	-0.31	0.38
Structure.diff	-0.03	0.41
None	-0.90***	0.16
Adjusted $R^2 = 0.292$		

Table 10.11: OLS model (weights and standard errors).
*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; · $p < 0.1$

including the other dimensions. Maybe surprisingly, mentioning differences between style and number of details receives a *positive* weight. From manual inspection, we find that when narratives are already very similar on other dimensions, workers tend to mention these more superficial differences.

We also analyzed the correlation between characteristics of workers (education, frequency of watching movies/reading books). For most of the dimensions, we did not observe a relation with the characteristics of workers. People who read more books more often mention the theme of a narrative ($\rho = .223$, $p < 0.05$). We also found that people who watch more movies more often pay attention to whether narratives differ in number of details or length ($\rho = .210$, $p < 0.1$).

Experts. Motivations given by the experts for the narrative pair in the introduction are shown in Table 10.12. Statistics based on manual annotation are shown in Table 10.9. The dimensions 'tale types' and 'motifs' are used in folk narrative research. Motifs are small elementary building blocks of plots of narratives (e.g., "*disease caused by witchcraft*"). As expected, motifs and tale types were only mentioned by the experts. Tale types were mentioned in many of the motivations (46%). Other dimensions important to experts are the plot, characters and theme of the narratives. Style, whether true facts are recounted, and setting are not important to experts.

Both are the same: the narratives must demonstrate that witches are real. *Given score: 5. Dimensions: Theme, Characters*
Strong similarity in content, I doubt between box 4 and 5: 1 and 2 share the traditional element of a witch changing into a cat, getting hurt, and being recognized in her human form through the wound. Given score: 4. Dimensions: Characters, Plot
Clearly two narratives of the same type: Hexentier verwundet: Frau zeigt am folgenden Tag Malzeichen. Whether it is with multiple cats, or one, it doesn't matter. Moral: night cats are metamorphosed witches, and you don't want them near you. *Given score: 5. Dimensions: Tale type, Theme, Characters*

Table 10.12: Translated motivations by experts

10.5 Estimating Narrative Similarity

In this section we present preliminary experiments on how well unsupervised and supervised methods correspond with the crowd judgements. Studies on document similarity in other domains found low to moderate correlations between automatic measures and human judgements. For example, a correlation of less than 0.2 was observed using cosine similarity [Zengin and Carterette, 2013] and between 0.5-0.6 using different binary, count-based and LSA-based measures [Lee et al., 2005]. To our knowledge, we are the first to perform such experiments on narrative similarity.

10.5.1 Goal and Evaluation

For each narrative pair, we take the mean of the received similarity judgements by the crowdworkers. We experiment with two different setups: 1) Classification, where the goal is to classify the pairs into *low* (≤ 3) and *high* (>3) similarity. The performance is reported using the F-score. 2) Regression, where the goal is to predict the mean of the received judgements. We evaluate the performance using the Spearman correlation and Mean Squared Error (MSE).

10.5.2 Dataset Construction

We randomly divided the dataset into a training and test set. Feature development and parameter tuning was done using cross-validation on the training set. Like in the previous sections, we excluded the narrative pairs that received a low score for understandability. Statistics of the dataset are shown in Table 10.13. The documents were parsed using the Frog parser [Van den Bosch et al., 2007] and a stop word list of 76 frequent Dutch words was used.

Set	# Pairs	Mean	Low	High
Train	498	2.674	344 (69.08%)	154 (30.92%)
Test	400	2.683	271 (67.75%)	129 (32.25%)

Table 10.13: Statistics dataset

10.5.3 Method

We experiment with both unsupervised similarity metrics (e.g., cosine similarity) and supervised machine learning models. We use linear regression and logistic regression with Ridge (L2) regularization to prevent overfitting.

10.5.4 Features

We evaluate a variety of features, most of them based on the dimensions we identified in the previous section. In addition, we explore features based on manually annotated metadata. First, we study the effectiveness of features that only measure lexical similarity. We experiment with different metrics (cosine similarity and Jaccard index) and representations (e.g., words versus character n-grams).

We also extract features from the narratives to approximate elements such as the plot, characters and theme in narratives. Plot elements are approximated by extracting subject-verb pairs. They are extracted by searching on subject (*'su'*) and verb complement (*'vc'*) relations from the Frog parser. Each 'plot element' is a character + root of a verb (e.g., 'lawyer_answer' or 'girl_disappear'). We extract the characters of a narrative by searching on subject (*'su'*) relations from the Frog parser. Only tokens classified as nouns, pronouns, or as 'special' are included. Unfortunately, the narratives are noisy because they come from a variety of sources, and therefore the Frog parser sometimes missed relations or incorrectly extracted them.

Themes are extracted using LDA [Blei et al., 2003]. We train a model on the training documents with 20 topics using the Gensim library [Řehůřek and Sojka, 2010]. We measure the similarity between the topic distributions using the Jensen-Shannon divergence.

Crowdworkers also pay attention to style. We therefore experiment with features that capture stylistic similarities based on statistics such as the length of words and sentences, and similarities in POS structures. Our analyses also revealed that differences in the amount of detail in narratives play a role. We use the difference in length of the narratives to approximate this dimension.

We also study the usefulness of manually annotated metadata. They also capture dimensions identified in the previous section, such as whether the narratives have the same genre (mentioned by the crowd and experts), or tale type (only mentioned by experts). In addition, we study whether manually annotated keywords and named entities are useful. Table 10.14 provides an overview of the used features.

	Lexical
1	Cosine similarity
2	Jaccard index
	Story Elements
3	Plot
4	Theme (LDA)
5	Characters
	Stylistic
6	Absolute difference between average word length
7	Absolute difference between average sentence length
8	1-3 n-gram POS patterns (Jaccard)
	Other
9	Absolute length difference
	Metadata (manual annotation)
10	Same tale type (boolean)
11	Keywords (Jaccard)
12	Same genre (boolean)
13	Named entities (Jaccard)

Table 10.14: Overview features

10.5.5 Results

We first study the individual features. Next, we study the performance achieved by combining them.

Individual features. We first evaluate the individual features in the regression setup. We report the Spearman correlations and MSEs (Table 10.15). For the lexical features, we experimented with using the cosine similarity and Jaccard index. We also experimented with using word unigrams, word unigrams + bigrams, or character n-grams (of lengths 2-5). We find that using n-grams consistently achieves a better performance. In addition, the Jaccard index performs better than the cosine similarity. We find that the stylistic features (POS patterns, word and sentence length) only obtain a low correlation. The features that aim to capture the story elements (e.g., theme) perform moderately. The features based on manually annotated metadata perform well, in particular the features based on tale types and keywords.

Combination of features. We now combine the features using supervised machine learning models. We evaluate them in regression and classification tasks (Table 10.16). We find that a reasonable performance is obtained using only the lexical features. Although the story elements features alone (plot, characters, theme) obtained a moderate performance, they do not help improve on the performance using the lexical features. We suspect this has several reasons. First, the story elements features are directly derived from the text as well and therefore highly correlated with the lexical features. For example, we find a Spearman correlation of .468 between the characters feature and the Jaccard n-grams feature. In addition, manual inspection shows that the extracted story elements are noisy, and thus the extraction of the features itself can be improved.

Metric	ρ	MSE
Lexical		
Cosine - Unigrams	0.182	0.925
Jaccard - Unigrams	0.374	0.816
Cosine - Bigrams	0.206	0.918
Jaccard - Bigrams	0.383	0.865
Jaccard - n-grams	0.418	0.817
Cosine - n-grams	0.357	0.813
Story Elements		
Theme (LDA)	0.122	0.968
Characters	0.155	0.953
Plot	0.168	0.937
Stylistic		
Difference word length	0.076	0.981
Difference sentence length	0.073	0.980
POS n-grams	0.121	0.950
Other		
Length difference	0.079	0.975
Metadata		
Tale type	0.336	0.873
Keywords	0.481	0.797
Genre	0.142	0.984
Named entities	0.184	0.946

Table 10.15: Individual features

Metric	ρ	MSE	F-score
Categories			
Lexical	0.431	0.759	0.590
Story elements	0.181	0.922	0.455
Stylistic	0.124	0.949	0.408
Metadata	0.494	0.746	0.614
Lexical + Category			
Lexical + story elements	0.435	0.761	0.590
Lexical + stylistic	0.491	0.715	0.611
Lexical + metadata	0.569	0.614	0.652
All			
Automatic (Lexical + story elem. + stylistic + other)	0.494	0.715	0.600
Automatic + metadata (Lexical + story elem. + stylistic + other + metadata)	**0.592**	**0.598**	**0.657**

Table 10.16: Feature combinations

The metadata alone are already very effective. However, one should keep in mind that for new narratives no metadata will be available. Using only lexical + stylistic features a good performance is achieved. Adding the remaining features does not lead to improvements. However, the best performance is obtained using both the automatically extracted features and the metadata.

Although the obtained correlation is moderate (.592), we should keep in mind that it is a difficult task. For example, when we randomly selected a judgement for each narrative pair and correlated that with the average of the remaining judgements, a Spearman correlation of .556 was obtained (see the section on agreement analysis).

10.6 Discussion and Implications

We analyzed the relationship between tale types and human perception of similarity. While most narrative pairs from different tale types are indeed perceived as not similar, within a tale type there may be much variation. Dimensions such as genre and style that do not play a role in the definition of tale types, do play a role in perception of similarity. This suggests that a more nuanced view of narrative similarity is desired.

Our results highlighted that non-experts and experts differ in how they judge narrative similarity. Therefore, how similarity between narratives is estimated should depend on the intended users and goal of the application. We also found that non-experts vary in which dimensions they consider. Therefore, efforts to personalize systems that deal with narrative similarity could be an interesting direction of research. In addition, to help users understand the output of an automatic system, explicit explanations of how narratives are related would be useful as well.

Our study has limitations. First, free-text motivations were used to study the importance of dimensions. Users only mentioned dimensions they considered relevant, but (unconsciously) they may have also been influenced by other dimensions. Second, the mentioned dimensions and provided ratings may also have been influenced by the previous pairs a user has seen. We randomized pairs within a HIT to reduce possible effects of displaying order. However, further research is needed to study the influence of sampling and displaying order on the user judgements. Third, to enable a large-scale experiment, we included a large number of narratives from the Dutch Folktale Database. While we posed several restrictions to the final set to improve readability and also asked workers to indicate whether they understood the narratives, unclear or noisy narratives may have led to noise in the obtained judgements and mistakes in the automatic extraction of the features in the prediction experiments. Fourth, our experiments were performed on one specific dataset. Although we expect that our experimental setup can be used in other domains as well, other datasets (for example, movie reviews) should be used to verify this.

10.7 Conclusion

This chapter presented a study on how humans perceive narrative similarity. A better understanding of narrative similarity is a first step towards better clustering and retrieval systems dealing with narrative collections. Data was collected by asking

crowdworkers and folktale experts to rate the similarity between narrative pairs. We analyzed the provided similarity scores as well as their provided motivations. Our results showed that non-experts pay attention to more dimensions than experts, and that tale types only give a limited view of narrative similarity.

Many of the identified dimensions can currently only be approximated in a shallow way using automatic methods. Further work is needed on automatically extracting dimensions such as style, structure, plot, etc. of narratives to improve the automatic estimation of narrative similarity. The findings of this chapter can be used to develop better clustering systems for narratives. While this chapter focused on a particular domain (narratives), we expect that the setup of the experiment and the types of data analyses performed can also be used to shed light on how similarity is perceived in other domains.

Part IV

Discussion and Conclusion

11

Discussion

This chapter reflects on the work carried out in this dissertation by focusing on two challenges: ethical concerns and biases in the data.

11.1 Ethics

The rise of big data, and consequently the introduction of new types of data sources and research directions, has raised numerous complex ethical questions [Bolander and Locher, 2014, boyd and Crawford, 2012]. In this section, three issues are discussed that are particularly relevant to the research in this dissertation.

The most prominent ethical issue regarding the use of social media data for research purposes concerns the visibility of such data, i.e., whether such data should be considered *private* or *public*. Early research using social media data approached privacy in a dichotomous manner, for example as illustrated in the introduction of a volume on Computer Mediated Communication by Herring [1996] with "*The editorial policy [..] makes a distinction between restricted- and open-access electronic fora, the former of which are considered private, while the latter are public*". In more recent years, however, researchers have acknowledged that privacy is much more fluid and complex. As mentioned by boyd and Crawford [2012], "*Just because content is publicly accessible does not mean that it was meant to be consumed by just anyone*". For example, conversations on private matters can be posted in public, and social norms regarding what should be considered public information may be very different from the explicit visibility settings of the data. While for more traditional data collection methods (e.g., lab experiments) practices such as 'informed consent' and oversight by institutional review boards are used to protect the privacy of participants, to date few guidelines exists for the use of social media data. Recognizing the importance and complexity of this issue in social media research is thus only a first step towards better protection of the privacy of social media users.

In this dissertation, only *publicly* available data was used. While it is acknowledged that such data could contain discussions on private matters, the analyses presented focused on global patterns, rather than the patterns of an *individual*. When

individual cases were selected for illustration, texts from public figures were chosen or the examples were chosen in such a way that the social media users themselves were not easily identifiable.

Ethical concerns also surround research that focuses on studying differences between social groups, for example, differences in language use between males and females. As discussed by Talbot [2008], gender research risks '*unintentional reproduction*' of gender stereotypes, for example by focusing on differences between males and females and ignoring similarities between genders. Such risks also accompany latent attribute prediction research, especially when features are highlighted that are predictive of males and females. For example, when the TweetGenie demo was launched, one Twitter user tweeted (translated from Dutch) "*The most characteristic words for men and women are very stereotypical :([LINK] #tweetgenie*", and another Twitter user described TweetGenie with "*TweetGenie: how stereotypical are your tweets? [LINK]*". These cases illustrate the need for carefulness when presenting the results of such research, and in particular when the results are disseminated to the public. More broadly, researchers should take an active role in informing the public debate on the impact of machine learning systems in society. In the specific chapters, examples from individuals were discussed who deviated from gender or age stereotypes to emphasize that not all individuals conform to the stereotypical patterns that are usually learned by prediction models. We conducted a study [Nguyen et al., 2014a] (Chapter 5) to problematize the operationalizations of gender and age in these prediction tasks.

Another concern is the use of latent attribute prediction in user profiling applications. Automatic recognition of latent attributes, such as gender and age, could benefit many applications, such as supporting more fine-grained analyses of trends, and this is indeed often used to motivate research on latent attribute prediction. However, this could also lead to (un)intentional discrimination. For example, while incorporating (inferred) gender as a feature could potentially improve many prediction tasks, it could also trigger gender discrimination. An example of a discriminative classifier is one that would predict a different label (e.g., job success, or repaying a loan) if the only difference in the feature values would be the gender of the person. Factors such as labeling bias, sampling bias and data incompleteness increase the risk for discrimination [Calders and Žliobaitė, 2013]. For example, if historically women are treated unfairly, the labels used for training could encode such biases and the trained model would only reflect, or perhaps even reinforce the biases. Fortunately, labels such as gender are also valuable for the various methods that have been devised to assess and mitigate discrimination in classifiers [Calders and Žliobaitė, 2013].

Thus, while the rise of big data has positively impacted both academia and industry, still little is known about ethical issues and few guidelines exist. Various factors in the data collection and modeling process could lead to unintentional discriminative classifiers, and thus care should be taken in how such classifiers are used in applications and in the interpretation of the results. Furthermore, as researchers are becoming increasingly interested in topics such as computationally modeling the relation between language and variables such as ethnicity and economic status, the ethical concerns will only increase.

11.2 Biases in Data

The selection of data sources may have introduced biases in the collected data. This section reflects on the main types of potential biases and their possible implications on the findings.

11.2.1 Language Selection

With the exception of Chapter 6, this dissertation focuses on textual data generated in the Netherlands. As a result, the language of the datasets used is mainly Dutch, although minority languages in the Netherlands (Limburgish and Frisian) and Turkish (in the context of code-switching) have also been considered. The approaches for the prediction tasks (e.g., age prediction (Chapter 4), language identification (Chapter 7) and tale type classification (Chapter 9)) did not make use of many language-specific resources and could therefore be adapted relatively easily to other languages (e.g., by using the English Wordnet instead of Cornetto). However, as Bender [2009] points out, even the effectiveness of linguistically naive models like n-gram models depends on the typological properties of the languages. Thus, future work should also consider other languages, preferably from other language families.

Compared to the prediction systems, the analyses presented in this dissertation are potentially more affected by the focus on the Dutch language, in particular in terms of generalizability. The study on language choice in Twitter (Chapter 8) might be less affected by the specific choice of languages studied. The study focused on broad patterns regarding language choice, focusing on the influence of the audience on the choice for using a majority language (Dutch) versus a minority language (Limburgish and Frisian). However, future research is needed to test whether similar patterns hold for other majority/minority language pairs. The study that potentially is affected the most is the study on the relation between age and language in Twitter (Chapter 4). For example, some of the highest ranked words in the prediction models are likely to be very specific to the Dutch language. On the other hand, some of the other patterns we found, such as the use of pronouns, matched some of the findings in previous research on other languages (e.g., Barbieri [2008] and Pennebaker and Stone [2003]).

11.2.2 Crowdsourcing

Several chapters in this dissertation made use of the *crowd* to collect data, i.e., data from a large group of (online) people. These people can be volunteers (Chapters 4 and 5) or be paid for their work (Chapter 10). The crowd population is *self-selected*, and thus can exhibit certain biases, which, depending on the type of study, could affect the results to some degree.

Chapter 4 used feedback from visitors of the TweetGenie demo to evaluate the age prediction system in the wild. Almost all entered accounts were from non-public figures, and therefore it is likely that the feedback was mostly provided by the holders of the accounts themselves. TweetGenie was particularly popular among the older Twitter users, and as a result, the Twitter users for whom feedback was received tended to be older than the general Twitter population in the Netherlands.

Chapter 5 used input from players of the game that was part of the TweetGenie demo to reflect on the operationalizations of gender and age in latent attribute prediction tasks. No demographic information was obtained from the players, but how language use is perceived is likely to depend on someone's background. Thus, more information about the players would be desirable, such as their demographics and their social media usage. Flekova et al. [2015] indeed found differences between age and gender groups regarding accuracy and confidence in annotating gender and age based on tweets in a crowdsourcing experiment. Despite the limitations of the used data, a recent paper confirms much of the findings from Chapter 5. Flekova et al. [2016] analyzed human perception of gender and age of Twitter users. Different from Chapter 5, they focused on the English language and collected data through Amazon Mechanical Turk. Similar to our study, annotations from multiple workers were collected for each Twitter user. While a majority-based system for gender prediction achieved an accuracy of 84% in our study, in the study by Flekova et al. the obtained accuracy was 85.8%, thus matching our findings closely. Furthermore, in their study older Twitter users were also are often perceived to be younger (Figure 1 in Flekova et al. [2016] matches closely with Figure 5.5b in this dissertation).

Chapter 10 collected labels from CrowdFlower workers to study the perception of narrative similarity. It is well known that the population of crowdworkers is heavily biased towards certain groups, e.g., while Amazon Mechanical Turk workers were initially moderate-income US workers, in later years the population shifted towards young, well-educated Indian workers [Ross et al., 2010]. In this study, crowdworkers were asked to judge Dutch texts, thus constraining the potential workforce heavily. Demographic information of the workers was obtained and while they were mostly men, there was a relatively good spread across ages, education levels and regions in the Netherlands. Only a few patterns were found that suggested differences in similarity perception related to the background of the workers (Subsection 10.4.4). For example, workers who read more books more often mentioned the theme of a narrative. However, the effects were small and therefore it can be assumed that the findings are not affected much by the composition of the crowdworkers.

11.2.3 Social Media Population Bias

A large part of this dissertation uses social media data. However, the user populations of social media platforms tend to be very different from the general population. For example, the data in Chapter 4 shows that the Dutch Twitter population (in 2013) is more biased towards younger Twitter users. Furthermore, the results suggest differences in the gender distribution across age groups. In Chapter 6 on geographical variation, tweets were used that were geotagged. However, Pavalanathan and Eisenstein [2015a] found that geotagged tweets are more often written by women and younger people, compared to users for whom the location was estimated based on their profile. The focus on social media, and the specific sampling used to collect the data, could thus introduce biases that have an influence on the observed language use, as well as the composition of the demographic and social groups.

When the goal is to make inferences about offline behavior (such as election outcomes or emigration patterns) based on data collected from social media, the biases in social media normally would need to be corrected for. When information about the target population is available (e.g., demographics obtained through census data), the biases can be estimated and corrected for [Wang et al., 2015, Zagheni and Weber, 2015]. For example, Wang et al. [2015] partition the population into fine-grained cells and then aggregate the estimates of each cell by weighting each cell by its relative proportion in the target population. When such information is not available, Zagheni and Weber [2015] suggest focusing on relative changes over time. In this dissertation, however, the focus was *not* on making inferences about offline behavior. Therefore, no corrections were applied to account for the biases, but it should still be kept in mind that the results may not generalize across other social media platforms.

11.2.4 Social Media Platform Selection Bias

The presented studies were based on data collected from Twitter (Chapters 4, 5, 6 and 8), online forums (Chapter 7), a dialect atlas (Chapter 6), letters to the editor (Chapter 6) and folk narratives from the Dutch Folktale Database (Chapters 9 and 10). However, each of the analyses and experiments was focused on a *single* data source and no comparison was made across the data sources.

The focus on individual data sources, and in particular Twitter, is something that this dissertation has in common with many social media studies. The dominance of Twitter is described by Tufekci [2014] as Twitter being the '*model organism*' for social media studies. In biology studies, model organisms refer to species that are extensively studied, with the expectation that findings will also provide insights into other species. However, focusing on one species or social media platform introduces biases related to the observed mechanisms. Not only may the composition of demographic and social groups differ across social media platforms, the design of the platforms may also stimulate different kinds of behavior, potentially leading to different patterns of language use and behavior.

The design of social media platforms affects user behavior at all levels. For example, it may influence how social media users shape their online identities, and with different focuses (e.g., self-expression versus professional self-promotion) norms for self-presentation may vary considerably across platforms [Van Dijck, 2013]. Networks structures differ across social media platforms [Yang and Counts, 2010], interpretation of emojis differs across platforms because of differences in rendering [Miller et al., 2016], tweets are limited to 140 characters, and so on.

Within computational linguistics, only a few studies have so far focused on differences in language use between social media platforms. Baldwin et al. [2013] focused on the 'noisiness' of data sources by analyzing the proportion of grammatical sentences and Hu et al. [2013] found that language use in Twitter was more formal than often expected. Thus, it is clear that there are differences in social behavior and language use across social media platforms and future work should consider to use data from multiple social media platforms.

12

Conclusion

In this dissertation, variation in written language was modeled and analyzed using computational approaches. In this chapter, the main findings are summarized based on the research questions presented in the introduction. The chapter then continues with an outline of future research directions, followed by some final remarks.

12.1 Findings

This section summarizes the main findings for the research questions introduced in Section 1.4.

Computational Sociolinguistics. The emerging area of '*Computational Sociolinguistics*' integrates aspects of sociolinguistics and computer science in studying the relation between language and society from a computational perspective. Chapter 2 surveys this area and demonstrates the potential for synergy between computational linguistics and sociolinguistics.

The largest progress is likely to occur when researchers from both fields join forces to fully leverage the knowledge and experience in both fields. However, in order for real synergy to come out of this, an understanding of the commonalities and differences between the involved research fields is needed. Section 2.2 reflected on the methods employed in the two fields. Besides a strong qualitative branch, sociolinguistics also has a large quantitative branch (variationist sociolinguistics), which therefore shares the most commonalities with computational linguistics. However, the two fields differ in terms of data collection methods, research goals (e.g., focusing on prediction versus explanation) and consequently, also the approaches used to validate the models.

Throughout Chapter 2 as well as in the subsequent chapters in this dissertation, examples were provided of how both fields can complement each other. The list below is not complete, but serves as an illustration of the potential of interaction between computational linguistics and sociolinguistics.

- Sociolinguistics serves as a source of reflection for methods employed in CL. For example, the operationalizations of gender, age and languages as distinct categories were questioned in Chapters 2 and 5.
- Insights from sociolinguistics explain errors made by automatic prediction models, for example why predicting the ages of older Twitter users is more difficult than predicting the ages of younger Twitter users (Chapter 4).
- Methods from CL facilitate scaling up sociolinguistic analyses. For example, the use of statistical methods can help to test whether a certain linguistic feature exhibits a significant geographical variation (Chapter 6) and automatic language identification can scale up analyses of code-switching (Chapter 8).
- Insights from sociolinguistics motivate new research problems for CL researchers, such as language identification at the word level instead of longer segments such as sentences or documents (Chapter 7).

I now return to the research questions relating to computational sociolinguistics, and provide an answer based on the findings presented in this dissertation.

> **RQ1**. *To what extent can the age of Twitter users be predicted based on their tweets?* (Chapter 4)

Automatically predicting variables such as age has been of increasing interest in the CL community. Although previous research on automatic age prediction primarily focused on age categories, Chapter 4 explored modeling age from three different angles: life stages, age categories and age as a continuous variable. Overall, the performance of the automatic systems was high. For example, when predicting age as a continuous variable, the Mean Absolute Error was less than 4 years. However, the predictions were notably less accurate on older Twitter users. An analysis revealed why this was the case: the differences in language use were smaller for older Twitter users in our data. Especially for Twitter users older than 30 years, the data suggests that it is difficult to make an accurate age prediction based on text alone. A possible reason for this pattern is that adults tend to be more conservative in their language use, due to workplace pressures and in order to be taken seriously [Eckert, 1997].

A demo was developed called TweetGenie (Section 4.6) to bring the research to the general public. TweetGenie predicts the gender and age of Dutch Twitter users with public accounts based on tweets. The system builds on the research presented in Chapter 4 and a similar system was adopted for gender prediction. To evaluate the system in the wild, we asked users of the demo to provide feedback on the predictions. The feedback revealed that the persons for whom Twitter account names were entered into the demo were older than the general Dutch Twitter population. Thus, there was a mismatch between the users for whom TweetGenie was asked to provide a prediction, and the user distribution the system was trained on. Predictions from TweetGenie were shared with researchers and students from various institutions, including Statistics Netherlands (CBS), the Institute for Dutch Lexicology, Stanford University, Radboud University and Utrecht University.

RQ2. *What are limitations and consequences of the typical operationalizations of gender and age in latent attribute studies?* (Chapter 5)

The results presented in Chapter 4 raised the need for a closer look at the prediction tasks. Therefore, a reflection on the operationalizations of gender and age was provided in Chapter 5. The TweetGenie demo hosted an online game, in which players were asked to guess the gender and age of Twitter users based on their tweets. By analyzing the guesses of the players, we showed that the perceived gender and age of Twitter users based on their language use often did not correspond with their biological sex and their chronological age. Our analysis questioned the operationalization of gender and age in latent attribute prediction tasks. For example, in the case of gender classification, operationalizing gender as a binary variable is a drastic simplification and even raises the question of whether a perfect performance can be attained on this task when the prediction is based on language use alone. Furthermore, similar to the automatic system, humans tended to underpredict the ages of older Twitter users, an observation that was also made in a more recent study by Flekova et al. [2016].

RQ3. *What is a suitable method to test for geographical language variation?* (Chapter 6)

Identifying linguistic variables that exhibit geographical variation is a key step in many dialect studies. This dissertation explored three approaches that have been used in previous work on dialect analysis: Moran's I, join count analysis and the Mantel test. However, these approaches make assumptions that may not hold in practice and thus may fail when these assumptions are violated. Chapter 6 proposes the use of the Hilbert-Schmidt Independence Criterion (HSIC) [Gretton et al., 2008], which is based on non-parametric statistics and overcomes these limitations. This study was the first to apply HSIC to linguistic data. The test is *consistent*, i.e., given enough data it will reach the right decision. HSIC builds on kernel methods from machine learning and makes no parametric assumptions about the relation between geography and linguistic variables. Furthermore, HSIC can be applied to all types of linguistic data. An extensive comparison was carried out between the different approaches to test for geographical variation. No single approach dominated across all settings but overall, the proposed approach HSIC was the most robust.

RQ4. *How can automatic language identification be performed at the word level?* (Chapter 7)

The increasing amount of social media data enables studying code-switching patterns in a multitude of social situations. Manually annotating the data by language is time-consuming. However, mixed-language texts pose challenges for most of the developed NLP tools, which usually assume that texts are written in a single language. Furthermore, traditional language identification methods focus on document-level identification and are therefore not suitable for analyzing more fine-grained patterns, such as code-switching within a single social media message. Chapter 7 explored automatic language identification approaches to help process and analyze mixed-language texts. In particular, we focused on Dutch-Turkish posts from an online forum.

We explored various approaches for automatic language identification at the word level. Furthermore, we argued for additional evaluation metrics rather than only raw accuracy (e.g., evaluation based on the derived labels as mono- and bilingual posts). We showed that by exploiting the context of words, the prediction performance could be improved. However in some cases, the context would incorrectly push the labels towards more coherence and therefore lead to the incorrect classification of a text as being monolingual. This study also highlighted annotation challenges. For example, for some tokens the language could not be determined (e.g., named entities, smileys).

We were among the first to focus on word-level language identification, and more broadly, the use of computational approaches for analyses of code-switching patterns. Recently, however, both topics have been receiving increasing attention in the CL research community, for example, with shared tasks on language identification at the word level at the First (EMNLP 2014 [Solorio et al., 2014]) and the Second (EMNLP 2016) Workshops on Computational Approaches to Code Switching. Furthermore, in 2016 there is a Special Session on Code-Switching at INTERSPEECH 2016. We also made the annotated data available. Das and Gambäck [2014] applied their language identification system on our data and report a 1.0% increase in accuracy. Furthermore, Gambäck and Das [2016] proposed new metrics for measuring the complexity of code-switching in corpora. Compared to several other datasets, our dataset contained the most switching between utterances.

> **RQ5**. *How does the target audience influence the language choice of social media users?* (Chapter 8)

Which language a user is using online is influenced by many factors, including the topic, emotion and audience. Chapter 8 presented a quantitative analysis of how audiences influence the use of minority languages on Twitter. The analysis focused on Twitter users located in two Dutch provinces, Limburg and Friesland. Using an automatic language identification tool, we provided a quantitative analysis of the use of Frisian and Limburgish, the minority languages spoken in Friesland and Limburg. On Twitter, the exact audience is unknown, but some tweets provide an indication of the target audience. For example, including hashtags often leads to an expansion of the audience, and including user mentions often results in a shift in the target audience towards the addressed user. By analyzing tweets that belonged to these two cases, we found quantitative evidence that the audiences of tweets influence the used languages. Furthermore, whereas most tweets were posted in Dutch, during conversations, users would often switch to the minority language.

Computational Folkloristics. After focusing in Part II on variation from a sociolinguistic perspective, Part III focuses on variation from a folkloristics perspective. Different variants of a story may appear due to oral and written transmission over time. For example, events may be added or left out, or characters might change. These variations are frequently studied by folklorists, since such variations might reveal how stories have developed over time. Furthermore, the type of variations (e.g., the removal of religious elements) may reveal more about the cultural and social context of a story. This dissertation addressed the following two research questions:

RQ6. *Can the tale types of folk narratives be automatically predicted?* (Chapter 9)

Folk narrative researchers use tale types to organize the different variants of a tale. For example, the different variants of *Little Red Riding Hood* are classified as ATU 333. Chapter 9 presented a learning-to-rank approach to automatically identify the tale types of folk narratives. Experiments were carried out with both the ATU and Brunvand type indexes using data from the Dutch Folktale Database (DFD). The performance on both type indexes was high. In particular, features inspired by distributed information retrieval approaches were found to be effective. An error analysis highlighted that some of the errors were caused by the system matching on stylistic cues rather than content. Furthermore, the inconsistent granularity of the tale types increased the difficulty of the task. The learning-to-rank approach was chosen so that a ranking of tale types can be presented to the user in a semi-automatic annotation process. If a fully automatic process is desired, an automatic classification can be obtained by taking the top-ranked tale type. Based on this research, an automatic tale type identifier is now integrated into the DFD to support metadata generation.

RQ7. *How is folk narrative similarity perceived by experts and non-experts?* (Chapter 10)

Whereas Chapter 9 approached folk narrative similarity from the concept of tale types, it was still unclear to what extent the concept of tale types matched human perception of folk narrative similarity. In Chapter 10 a closer look was taken at folk narrative similarity using a crowdsourcing experiment to collect similarity judgements of folktale researchers and crowdworkers. Both groups assigned a low similarity to narratives from different tale types. However, for narrative pairs that were classified with the same tale type, there was still much variation in the similarity judgements. In particular, non-experts also considered dimensions that were not important for folktale researchers, such as style and details such as locations. This study showed that similarity of folk narratives is a complex concept. Moreover, the results suggest that for an exploratory system, an adaptable similarity metric based on individual user preferences is more preferable than a global similarity metric.

To summarize, variation from both a sociolinguistics and folkloristics perspective was studied in this dissertation. Several chapters revolve around prediction tasks, such as gender and age prediction, automatic language identification, and the automatic identification of tale types. The dissertation also contains several chapters that were primarily focused on analyses, such as the studies on perception of narrative similarity, perception of the identities of Twitter users based on their tweets, and the influence of audience on language choices in Twitter. For the prediction tasks, a variety of supervised machine learning methods were used. Linear and logistic regression were used in Chapters 4, 7 and 10. Conditional random fields were used in Chapter 7 to take context into account, and a learning-to-rank approach was taken in Chapter 9 to be able to return a ranking. The study on geographical variation (Chapter 6) was based on non-parametric statistics, to be able to detect arbitrary dependencies between spatial and linguistic signals.

12.2 Future Work

In this dissertation, variation in text was studied and modeled using computational approaches from two different perspectives: sociolinguistics and folkloristics. In this section, four future research directions are discussed that are likely to advance research in both areas.

12.2.1 Beyond Lexical Variation

The presented studies on variation from a sociolinguistic perspective were primarily focused on *lexical* variation, and in particular variation at the word level. For example, the prediction models for age and gender in Chapters 4 and 5 were based on word unigrams, because preliminary experiments and related literature had shown that word unigrams are effective for these tasks. Chapter 7 on language identification focused on word-level language identification. The approaches used in Chapter 6 for analyzing geographical variation are applicable to *any* type of linguistic data, but the main focus of the experiments was lexical variation, especially when the tool was applied in an exploratory setting. Similarly, the studies on variation in folk narrative data (Part III), based on data from the Dutch Folktale Database, made use of lexical features to compute the similarity between narratives. Although some semantic features were explored, e.g., based on characters or named entities, error analyses revealed that these features were not very effective, which was perhaps caused by mistakes in the extraction process.

Other types of variation have been featured in this dissertation as well. For example, orthographic variation has been considered in the study on language and age in Chapter 4, syntactic variation using data from the Dutch Syntactic Atlas in Chapter 6 and variation in the language used (e.g., a minority language or Dutch) in Chapter 8. However, the majority of the work has focused on lexical variation, and this is in line with the dominant focus of research in the area of computational sociolinguistics so far.

In Nguyen and Cornips [2016] we built on the work presented in Chapter 7 on automatic language identification by considering language identification at the morpheme level. In an exploratory setting, words were identified that were composed of subunits associated with different languages. For example, in *oetverkocht* 'sold out', *oet* 'out' is used that is associated with Limburgish, while *verkocht* 'sold' is associated with Dutch. The system was able to identify interesting, creative uses of code-switching within words on Twitter. However, the results also highlighted that preprocessing steps (e.g., removing named entities) are important for this task. Other researchers have also started to consider other layers of variation with recent studies on syntactic variation (e.g., Doyle [2014], Johannsen et al. [2015]) and semantic variation (e.g., Bamman et al. [2014a]) in social media. In the domain of folk narratives, studies could focus on analyzing variation in, say, characters, location, narrative structure and motifs.

Models that would explicitly capture variation at other layers besides the lexical layer could potentially lead to a better task performance by being able to account for variation that is not captured in lexical features. For example, when modeling

variation in social media from a sociolinguistics perspective, variation in syntax, morphology, etc. could be modeled and analyzed. As sociolinguists are generally more interested in variation at these other layers, such research would help bring sociolinguists and computational linguists closer and potentially create new opportunities for collaboration. However, when other types of variation, such as syntactic variation, are considered, more fine-grained extraction and parsing methods are probably needed, such as dependency and part-of-speech taggers [Johannsen et al., 2015], which may not be available for lesser-studied language varieties.

12.2.2 Diachronic Change

The studies in this dissertation focused on *synchronic* variation, i.e., variation at a specific point in time. Although the folktales in the Dutch Folktale Database were not necessarily from the same time period, the analyses did not focus on change over time. However, synchronic variation often is the starting point for *diachronic change*. Moreover, synchronic variation is often studied to *analyze* diachronic change within sociolinguistics through apparent time studies [Meyerhoff, 2006]. However, studying linguistic change using computational approaches has only been explored in a few studies so far. For example, using social media data (e.g., Danescu-Niculescu-Mizil et al. [2013b], Eisenstein et al. [2014]), and historical corpora (e.g., Moscoso del Prado Martin and Brendel [2016]). Similarly, analyzing large-scale diachronic corpora of folk narratives using text mining approaches is currently an unexplored area, while traditionally the study of folklore has tended to focus on diachronic rather than synchronic aspects. For example, the historic-geographic approach within the study of folklore compares variation in folk narratives to study the origin and development of stories [Goldberg, 1984]. Thus, a larger focus on diachronic change in computational studies would also be a step forward in closing the gap with relevant areas such as sociolinguistics and folk narrative research.

Large-scale studies of change over time do need to deal with several challenges. The availability of suitable datasets is thus far limited. For example, social media datasets rarely cover time periods longer than ten years (e.g., the Twitter dataset used by Eisenstein et al. [2014] covered 2009-2012, Danescu-Niculescu-Mizil et al. [2013b] covered 2001-2011, the datasets used by Kershaw et al. [2016] covered less than a year). Furthermore, to properly interpret findings on such datasets it is important to understand the dynamics and possible biases in such datasets. For example, the Google books corpus has been frequently used to study cultural and linguistic change, but Pechenick et al. [2015] point out that findings based on this data can be influenced by biases such as the overrepresentation of scientific literature in the data. Another issue could be the uneven distribution of data across time periods.

However, the amount of digital data will only keep increasing and thus the availability of suitable datasets to study diachronic change on a larger scale is also expected to increase. In particular, the fine-grained contextual data introduces unique possibilities to study synchronic and diachronic change jointly. For example, the data could potentially be used to study questions such as how variation at a specific point in time (e.g., across social groups or geographically) influences trajectories of change over time.

12.2.3 Interpretable Models

In the social sciences and the humanities, statistical models are mostly used for explanation, while in computer science the focus tends to be on prediction. Which goal is leading has a large influence on the selection and validation of models [Shmueli, 2010]. Each of the approaches has a specific value: explanation plays a large role in theory building and testing, while predictive modeling is valued for its use in practical applications. Research in areas such as machine learning and computational linguistics has had an almost exclusive focus on prediction tasks, measuring the quality of the models using metrics such as precision, recall, F-score, accuracy, and so on, which Wagstaf [2012] refers to as the *'hyper-focus on abstract metrics'*. Interpretability of models often comes second to prediction performance [Breiman, 2001]. As a result, the value of such models in explaining the phenomenon of interest has tended to be small. Improving the interpretability of the learned models would increase the role of such models in theory building. Furthermore, when specific actions need to be taken based on the output of a system, the system needs to be *accountable* and it is essential for the users to *understand* the predictions a system has made. Interpretable models can also help to reduce the risk of the model making predictions based on confounding factors (and thus, potentially leading to discriminative models, see Section 11.1), which humans could correct or identify if the learned model is interpretable.

With the emergence of machine learning areas such as deep learning, a call for more interpretable models has been voiced by Manning [2015] and in 2016 several workshops have been organized on this topic[1]. Furthermore, the right for users to obtain an explanation of predictions made using computational methods in certain situations is even codified in EU regulations [Goodman and Flaxman, 2016], illustrating the increased recognition of the importance of this topic. Interestingly, within the social sciences, an opposite trend can be observed, as some now argue for a larger focus on prediction [Yarkoni and Westfall, 2016].

Attention for this topic is fairly recent, and research addressing the development and evaluation of interpretable machine learning models within computational linguistics has so far been sparse. Research on topic modeling is one of the exceptions where interpretability of the models has received much attention, both in training and evaluating the models [Paul, 2016]. However, evaluating the interpretability of models is not trivial. For example, model size alone is not a good indicator of interpretability [Freitas, 2014]. A promising research direction involves integrating interpretation criteria into the learning process of the models, for example by incorporating coherence in learning the models parameters [Paul, 2016] or adding monotonicity constraints provided by the users (e.g., a constraint that when the price of a product increases, it becomes less likely that a consumer will buy the product) [Freitas, 2014]. Another research direction is to develop models that generate explanations for the predictions to users [Ribeiro et al., 2016].

In this dissertation, the need for interpretable models was highlighted in the crowdsourcing experiment in Chapter 10. In this study, users varied in the dimensions they considered important when assessing similarity of folk narratives and thus

[1] 2016 Workshop on Human Interpretability in Machine Learning at ICML, Human Centered Machine Learning at CHI and Human Centered Data Science at CSCW.

a system that is able to adapt to the preferences of users would be desirable. Furthermore, being able to explain why a certain clustering has been made would also help users interpret the output, especially when the outcome is unexpected. The models employed in this dissertation vary in their interpretability. For example, in Chapters 4 and 7 logistic regression was used for the tasks of age prediction and language identification. Logistic regression outputs a probability, which could be treated as an indication of the confidence in the prediction. These confidence values could then help interpreting the results and could possibly be used in subsequent analyses, e.g., by giving less confident cases a lower weight. However, the logistic regression models in this dissertation still use thousands of features. While the top ranked features provide a glimpse into the evidence that the model typically uses, for many end users, these models still remain a black box. The recent interest in interpretable machine learning models is a welcome development that could increase the impact of computational approaches in the social sciences and humanities.

12.2.4 Data Source Comparison

As discussed in Subsection 11.2.4, focusing on a single data source (e.g., a particular social media platform or folk narratives from a specific collector) has the risk of introducing various kinds of biases in the data, thereby potentially limiting the generalizability of the results. Thus, how different sources compare to each other is an important aspect to be considered in future research. For example, are geographical patterns observed in Twitter data also present in other social media platforms such as Instagram or blogs? Or, how does the medium influence the way folk narratives are told and transmitted? A challenge when comparing data sources is that it can be difficult to identify the factors that cause the observed differences (e.g., is a difference in language use caused by different user populations, or is it caused by the characteristics of the specific medium?). Datasets containing the same set of users across different platforms (e.g., the same users across Twitter, Instagram and Foursquare [Han Veiga and Eickhoff, 2016]) are thus a promising source for these studies.

12.3 Concluding Remarks

We live in exciting times. The massive amounts of textual data that are becoming increasingly available enable studying social and cultural phenomena in ways that were not possible before. Furthermore, big social and cultural data are promising sources for reflection on computational approaches to analyzing text. In this dissertation, variation in text was analyzed using computational approaches. I hope to have conveyed that variation in text carries social meaning and that computationally analyzing and modeling this variation can help answer questions about social and cultural phenomena. Many of the presented studies benefited from the interdisciplinary collaborations with social scientists, anthropologists, and sociolinguists during my PhD. I am excited by the increasing interaction between computer scientists and researchers from the humanities and the social sciences, and I am looking forward to see how this will push the disciplines involved forward in the coming years.

Bibliography

A. Aarne. *Verzeichnis der märchentypen*. Number 3 in Folklore Fellows Communications. Helsinki: Academia Scientarium Fennica, 1910.

J. Abello, P. Broadwell, and T. R. Tangherlini. Computational folkloristics. *Communications of the ACM*, 55(7):60–70, 2012.

S. W. Aboelela, E. Larson, S. Bakken, O. Carrasquillo, A. Formicola, S. A. Glied, J. Haas, and K. M. Gebbie. Defining interdisciplinary research: Conclusions from a critical review of the literature. *Health Services Research*, 42(1p1):329–346, 2007.

H. Adel, N. T. Vu, and T. Schultz. 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, 2013.

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

J. Allan. *Topic detection and tracking. Introduction to topic detection and tracking*, pages 1–16. Kluwer Academic Publishers, Norwell, MA, USA, 2002. ISBN 0-7923-7664-1.

J. Allan, V. Lavrenko, D. Malin, and R. Swan. Detections, bounds, and timelines: UMass and TDT-3. In *Proceedings of Topic Detection and Tracking Workshop (TDT-3)*, 2000.

C. O. Alm and R. Sproat. *Affective Computing and Intelligent Interaction*, chapter Emotional Sequencing and Development in Fairy Tales, pages 668–674. Springer Berlin Heidelberg, 2005.

I. Androutsopoulos and P. Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38(1):135–187, 2010.

J. Androutsopoulos. Pragmatics of computer-mediated communication. In S. Herring, D. Stein, and T. Virtanen, editors, *Code-switching in computer-mediated communication*, pages 667–694. De Gruyter Mouton, 2013a.

J. Androutsopoulos. Online data collection. In C. Mallinson, B. Childs, and G. V. Herk, editors, *Data Collection in Sociolinguistics: Methods and Applications*, pages 236–249. Routledge, 2013b.

J. Androutsopoulos. Languaging when contexts collapse: Audience design in social networking. *Discourse, Context & Media*, 4–5(0):62 – 73, 2014.

L. Anselin. Local indicators of spatial association–LISA. *Geographical analysis*, 27(2):93–115, 1995.

E. M. Ardehaly and A. Culotta. 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, 2015.

S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346, 2003.

S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9), 2007.

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

L. Aroyo and C. Welty. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Proceedings of WebSci'13*, 2013.

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

J. L. Austin. *How to do things with words*. Oxford University Press, 1975.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *The Journal of Machine Learning Research*, 3:1–48, 2002.

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

J. Bak, S. Kim, and A. Oh. 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, 2012.

T. Baldwin and M. Lui. 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, 2010.

T. Baldwin, P. Cook, M. Lui, A. MacKinlay, and L. Wang. 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, 2013.

D. Bamman and N. A. Smith. Contextualized sarcasm detection on Twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, pages 574–577, Oxford, UK, 2015.

D. Bamman, C. Dyer, and N. A. Smith. 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, 2014a.

D. Bamman, J. Eisenstein, and T. Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014b.

D. Bär, T. Zesch, and I. Gurevych. A reflective view on text similarity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 515–520, Hissar, Bulgaria, 2011.

F. Barbieri. Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics*, 12(1):58–88, 2008.

S. Barbiers, H. Bennis, G. D. Vogelaer, M. Devos, M. van der Ham, I. Haslinger, M. van Koppen, J. V. Craenenbroeck, and V. V. den Heede. *Syntactic Atlas of the Dutch Dialects: Volume I*. Amsterdam University Press, 2005.

S. Barbiers, J. van der Auwera, H. Bennis, E. Boef, G. D. Vogelaer, and M. van der Ham. *Syntactic Atlas of the Dutch Dialects: Volume II*. Amsterdam University Press, 2009.

Beevolve.com. An exhaustive study of Twitter users across the world. http://www.beevolve.com/twitter-statistics/. Last accessed: Jan 2013. URL `http://www.beevolve.com/twitter-statistics/`.

T. S. Behrend, D. J. Sharek, A. W. Meade, and E. N. Wiebe. The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43(3):800–813, 2011.

A. Bell. Language style as audience design. *Language in Society*, 13(2):145–204, 1984.

A. Bell. *The guidebook to sociolinguistics*. John Wiley & Sons, 2013.

E. M. Bender. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece, 2009.

E. M. Bender, J. T. Morgan, M. Oxley, M. Zachry, B. Hutchinson, A. Marin, B. Zhang, and M. Ostendorf. 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, 2011.

M. Bendersky and W. B. Croft. Finding text reuse on the web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 262–271, Barcelona, Spain, 2009.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

S. Bergsma and B. Van Durme. 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, 2013.

S. Bergsma, P. McNamee, M. Bagdouri, C. Fink, and T. Wilson. Language identification for creating language-specific Twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, pages 65–74, Montréal, Canada, 2012a.

S. Bergsma, M. Post, and D. Yarowsky. 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, 2012b.

V. L. Bergvall. Toward a comprehensive theory of language and gender. *Language in society*, 28(02):273–293, 1999.

V. L. Bergvall, J. M. Bing, and A. F. Freed. *Rethinking Language and Gender Research: Theory and Practice*. Routledge, 1996.

R. M. Bhatt and A. Bolonyai. Code-switching and the optimal grammar of bilingual language use. *Bilingualism: Language and Cognition*, 14(04):522–546, 2011.

J.-I. Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez. 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, 2013.

O. Biran, S. Rosenthal, J. Andreas, K. McKeown, and O. Rambow. Detecting influencers in written online conversations. In *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, pages 37–45, Montréal, Canada, 2012.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

B. Bolander and M. A. Locher. Doing sociolinguistic research on computer-mediated data: A review of four methodological issues. *Discourse, Context & Media*, 3(0):14 – 26, 2014.

A. Bouchard, P. Liang, T. Griffiths, and D. Klein. A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 887–896, Prague, Czech Republic, 2007.

C. Boulis and M. Ostendorf. 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, 2005.

D. Boyd. *Special Issue: Oral History in the Digital Age*, volume 40. Oral History Review, 2013.

d. boyd and K. Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679, 2012.

D. B. Bracewell, M. Tomlinson, and H. Wang. 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, 2012.

L. J. Bracken and E. A. Oughton. 'What do you mean?' The importance of language in developing interdisciplinary research. *Transactions of the Institute of British Geographers*, 31(3):371–382, 2006.

P. Bramsen, M. Escobar-Molano, A. Patel, and R. Alonso. 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, 2011.

C. A. Bravo and L. Hoffman-Goetz. Tweeting about prostate and testicular cancers: Do Twitter conversations and the 2013 Movember Canada campaign objectives align? *Journal of Cancer Education*, 31(2):236–243, 2016.

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

W. Bright. Notes. *Language in Society*, 26(03):469–470, 1997.

S. Brody and N. Diakopoulos. 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, 2011.

P. Brown and S. C. Levinson. *Politeness: Some universals in language usage*, volume 4 of *Studies in Interactional Sociolinguistics*. Cambridge University Press, 1987.

J. H. Brunvand. A type index of urban legends. *Encyclopedia of Urban Legends. Updated and expanded edition*, pages 741–765, 2012.

M. Bucholtz and K. Hall. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614, 2005.

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

J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309, Edinburgh, Scotland, UK, 2011.

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

T. Calders and I. Žliobaitė. Why unbiased computational processes can lead to discriminative decision procedures. In B. Custers, T. Calders, B. Schermer, and T. Zarsky, editors, *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, pages 43–57. Springer Berlin Heidelberg, 2013.

J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–28, Seattle, USA, 1995.

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

S. Carter, W. Weerkamp, and M. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47 (1):195–215, 2013.

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

W. B. Cavnar and J. M. Trenkle. 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, 1994.

B. Ceran, R. Karad, A. Mandvekar, S. R. Corman, and H. Davulcu. A semantic triplet based story classifier. In *The 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 573 – 580, Istanbul, Turkey, 2012.

H. Ceylan and Y. Kim. Language identification of search engine queries. 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 1066–1074, Suntec, Singapore, 2009.

J. K. Chambers and P. Trudgill. *Dialectology*. Cambridge University Press, 1998.

M. Cherubini, R. de Oliveira, and N. Oliver. Understanding near-duplicate videos: a user-centric approach. In *Proceedings of the 17th ACM international conference on Multimedia (MM '09)*, pages 35–44, Beijing, China, 2009.

J. Cheshire. *Handbook of Language Variation and Change*, chapter Sex and gender in variationist research, pages 423–443. Oxford, UK: Blackwell, 2002.

B. C. K. Choi and A. W. P. Pak. 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, 2006.

M. Ciot, M. Sonderegger, and D. Ruths. 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, 2013.

W. S. Cleveland, E. Grosse, and W. M. Shyu. Local regression models. *Statistical models in S*, pages 309–376, 1992.

A. D. Cliff and J. K. Ord. *Spatial processes: models & applications*, volume 44. Pion London, 1981.

C. G. Clopper. Experiments. In C. Mallinson, B. Childs, and G. V. Herk, editors, *Data Collection in Sociolinguistics: Methods and Applications*, pages 151–161. Routledge, 2013.

P. Clough. Old and new challenges in automatic plagiarism detection. *National Plagiarism Advisory Service*, 2003.

P. Clough, R. Gaizauskas, S. S. L. Piao, and Y. Wilks. METER: MEasuring TExt Reuse. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 152–159, Philadelphia, USA, 2002.

R. Cohen and D. Ruths. 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, 2013.

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

M. Collins and N. Duffy. Convolution kernels for natural language. In *Advances in Neural Information Processing Systems 14*, pages 625–632, Vancouver, British Columbia, Canada, 2001.

M. Corney, O. de Vel, A. Anderson, and G. Mohay. 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, 2002.

L. Cornips. Recent developments in the Limburg dialect region. In F. Hinskens and J. Taeldeman, editors, *Language and Place. An International Handbook of Linguistic Variation*. De Gruyter Mouton, 2013.

R. Cotterell, A. Renduchintala, N. Saphra, and C. Callison-Burch. An Algerian Arabic-French code-switched corpus. In *LREC-2014 Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*, pages 34–37, Reykjavik, Iceland, 2014.

N. Cressie. Spatial prediction and ordinary kriging. *Mathematical geology*, 20(4):405–421, 1988.

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

W. Daelemans. 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, 2013.

C. Danescu-Niculescu-Mizil and L. Lee. 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, 2011.

C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais. 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, 2011.

C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. 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, 2012.

C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts. 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, 2013a.

C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. 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, 2013b.

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

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

A. Das and B. Gambäck. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing (ICON-2014)*, pages 169–178, Goa, India, 2014.

A. De Fina, D. Schiffrin, and M. Bamberg, editors. *Discourse and Identity*. Cambridge University Press, 2006.

F. De Jong, D. W. Oard, W. Heeren, and R. Ordelman. Access to recorded interviews: A research agenda. *ACM Journal on Computing and Cultural Heritage (JOCCH)*, 1(1):1–27, 2008.

F. De Jong, A. van Hessen, T. Petrovic, and S. Scagliola. 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, 2014.

P. de Jong, C. Sprenger, and F. van Veen. On extreme values of Moran's I and Geary's c. *Geographical Analysis*, 16(1):17–24, 1984.

T. Declerck and P. Lendvai. Linguistic and semantic representation of the Thompson's motif-index of folk-literature. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL 2011)*, pages 151–158, Berlin, Germany, 2011.

T. Declerck, N. Koleva, and H.-U. Krieger. Ontology-based incremental annotation of characters in folktales. In *Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 30–34, Avignon, France, 2012.

S. Deterding, D. Dixon, R. Khaled, and L. Nacke. From game design elements to gamefulness: Defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15, Tampere, Finland, 2011.

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

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

J. Diesner and K. M. Carley. 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, 2005.

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

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

A. S. Doğruöz and P. Nakov. 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, 2014.

G. Doyle. 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, 2014.

A. Dundes. The motif-index and the tale type index: A critique. *Journal of Folklore Research*, 34(3):195–202, 1997.

M. Durham. Language choice on a Swiss mailing list. *Journal of Computer-Mediated Communication*, 9(1), 2003.

C. Dürscheid and E. Stark. Sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In C. Thurlow and K. Mroczek, editors, *Digital Discourse. Language in the New Media*. Oxford: Oxford University Press, 2011.

N. Dwi Prasetyo, C. Hauff, D. Nguyen, T. van den Broek, and D. Hiemstra. On the impact of Twitter-based health campaigns: A cross-country analysis of Movember. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 55–63, Lisbon, Portugal, 2015.

M. D. Ecker and A. E. Gelfand. Bayesian variogram modeling for an isotropic spatial process. *Journal of Agricultural, Biological, and Environmental Statistics*, 2(4):347–369, 1997.

K. Eckert, M. Niepert, C. Niemann, C. Buckner, C. Allen, and H. Stuckenschmidt. Crowdsourcing the assembly of concept hierarchies. In *Proceedings of the 10th annual joint conference on Digital libraries (JCDL 2010)*, pages 139–148, Surfer's Paradise, Australia, 2010.

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

P. Eckert. Age as a sociolinguistic variable. In F. Coulmas, editor, *The handbook of sociolinguistics*, pages 151–167. Blackwell Publishers, 1997.

P. Eckert. *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Wiley-Blackwell, 2000.

P. Eckert. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476, 2008.

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

P. Eckert and S. McConnell-Ginet. *Language and gender*. Cambridge University Press, 2013.

J. Eisenstein. 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, 2013a.

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

J. Eisenstein. Written dialect variation in online social media. In C. Boberg, J. Nerbonne, and D. Watt, editors, *Handbook of Dialectology*. Wiley, 2015.

J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. 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, 2010.

J. Eisenstein, N. A. Smith, and E. P. Xing. 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, 2011.

J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. Diffusion of lexical change in social media. *PLoS ONE*, 9(11):e113114, 11 2014.

I. Eleta and J. Golbeck. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41:424 – 432, 2014.

H. Elfardy and M. Diab. 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)*, pages 371–378, Istanbul, Turkey, 2012a.

H. Elfardy and M. Diab. Token level identification of linguistic code switching. In *Proceedings of COLING 2012: Posters*, pages 287–296, Mumbai, India, 2012b.

H. Elfardy and M. Diab. 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, 2013.

D. K. Elson. Detecting story analogies from annotations of time, action and agency. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative*, pages 89–97, Istanbul, Turkey, 2012.

F. Willekens et al. Computational humanities. Report of the computational humanities programme committee of the KNAW. Technical report, Amsterdam: Royal Netherlands Academy of Arts and Sciences (KNAW), 2010.

N. Fairclough. *Language and power*. London: Longman, 1989.

R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin. LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

M. Fay. Story comparison via simultaneous matching and alignment. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative*, pages 98–102, Istanbul, Turkey, 2012.

O. Ferschke, I. Gurevych, and Y. Chebotar. 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, 2012.

K. Filippova. 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, 2012.

C. Fink, J. Kopecky, and M. Morawski. 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, 2012.

B. Fisseni and B. Löwe. Which dimensions of narrative are relevant for human judgments of story equivalence? In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative*, pages 112–116, Istanbul, Turkey, 2012.

S. R. Flaxman. *Machine Learning in Space and Time*. PhD thesis, Carnegie Mellon University, 2015.

L. Flekova, D. Preoţiuc-Pietro, J. Carpenter, S. Giorgi, and L. Ungar. Analyzing crowdsourced assessment of user traits through Twitter posts. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

L. Flekova, J. Carpenter, S. Giorgi, L. Ungar, and D. Preoţiuc-Pietro. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 843–854, Berlin, Germany, 2016.

A. A. Freitas. Comprehensible classification models: A position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.

L. Friedland and J. Allan. Joke retrieval: recognizing the same joke told differently. In *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM '08)*, pages 883–892, Napa Valley, California, 2008.

K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496, Vancouver, B.C., Canada, 2007.

B. Gambäck and A. Das. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1850–1855, Portorož, Slovenia, 2016.

A. Garcia-Fernandez, A.-L. Ligozat, and A. Vilnat. Construction and annotation of a French folkstale corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2430–2435, Reykjavik, Iceland, 2014.

P. Gardner-Chloros and M. Edwards. Assumptions behind grammatical approaches to code-switching: When the blueprint is a red herring. *Transactions of the Philological Society*, 102 (1):103–129, 2004.

N. Garera and D. Yarowsky. 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, 2009.

M. Garley and J. Hockenmaier. 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, 2012.

J. P. Gee. *An Introduction to Discourse Analysis: Theory and Method*. New York: Routledge, third edition, 2011.

A. Getis and J. Aldstadt. Constructing the spatial weights matrix using a local statistic. In *Perspectives on spatial data analysis*, pages 147–163. Springer, 2010.

A. Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3):189–206, 1992.

P. Gianfortoni, D. Adamson, and C. P. Rosé. 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, 2011.

E. Gilbert. 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, 2012.

E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*, pages 211–220, Boston, Massachusetts, USA, 2009.

H. Giles and N. Coupland. *Language: Contexts and consequences*. Mapping Social Psychology Series. Brooks/Cole Publishing Company, 1991.

H. Giles, D. M. Taylor, and R. Bourhis. Towards a theory of interpersonal accommodation through language: some Canadian data. *Language in Society*, 2(2):177–192, 1973.

H. Giles, N. Coupland, and J. Coupland. Accommodation theory: Communication, context, and consequence. In H. Giles, J. Coupland, and N. Coupland, editors, *Contexts of Accommodation*, pages 1–68. Cambridge University Press, 1991.

C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering causal structure: Artificial intelligence, philosophy of science, and statistical modeling*. Academic Press, 1987.

T. Goeman. *T-deletie in Nederlandse dialecten; kwantitatieve analyse van structurele, ruimtelijke en temporele variatie*. PhD thesis, Vrije Universiteit Amsterdam, 1999.

C. Goldberg. The historic-geographic method: Past and future. *Journal of Folklore Research*, 21(1):1–18, 1984.

S. A. Golder and M. W. Macy. Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology*, 40(1):129–152, 2014.

R. G. Gomes, P. Welinder, A. Krause, and P. Perona. Crowdclustering. In *Advances in Neural Information Processing Systems 24*, pages 558–566, Granada, Spain, 2011.

A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19, 2010.

B. Goodman and S. Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, New York, NY, USA, 2016.

C. Gooskens and W. Heeringa. The relative contribution of pronunciational, lexical, and prosodic differences to the perceived distances between Norwegian dialects. *Literary and Linguistic Computing*, 21(4):477–492, 2006.

S. Goswami, S. Sarkar, and M. Rustagi. Stylometric analysis of bloggers' age and gender. In *Proceedings of the Third International ICWSM Conference*, pages 214–217, San Jose, California, 2009.

T. Gottron and N. Lipka. A comparison of language identification approaches on short, query-

style texts. In *Proceedings of the 32nd European conference on Advances in Information Retrieval (ECIR 2010)*, pages 611–614, Milton Keynes, UK, 2010.

S. Gouws, D. Metzler, C. Cai, and E. Hovy. 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, 2011.

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

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

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory*, volume 3734 of *Lecture Notes in Computer Science*, pages 63–77. Springer Berlin Heidelberg, 2005a.

A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005b.

A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf, and A. J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592, Vancouver, B.C., Canada, 2008.

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems 25*, pages 1205–1213, Lake Tahoe, Nevada, USA, 2012.

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

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

J. Grieve. *Regional Variation in Written American English*. Cambridge University Press, 2016.

J. Grieve, D. Speelman, and D. Geeraerts. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23(02):193–221, 2011.

J. Grieve, D. Speelman, and D. Geeraerts. A multivariate spatial analysis of vowel formants in American English. *Journal of Linguistic Geography*, 1:31–51, 2013.

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

R. Grundkiewicz and F. Gralinski. How to distinguish a kidney theft from a death car? Experiments in clustering urban-legend texts. In *Proceedings of the RANLP 2011 Workshop on Information Extraction and Knowledge Acquisition*, pages 29–36, Hissar, Bulgaria, 2011.

M. Guerini and C. Strapparava. Why do urban legends go viral? *Information Processing & Management*, 52(1):163–172, 2016.

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

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

G. R. Guy. The cognitive coherence of sociolects: How do speakers handle multiple sociolin-

guistic variables? *Journal of Pragmatics*, 52:63 – 71, 2013.

G. Gweon, M. Jain, J. McDonough, B. Raj, and C. P. Rosé. 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, 2013.

D. Haase, editor. *The Greenwood Encyclopedia of Folktales and Fairy Tales*, volume 1-3. Westport, CT: Greenwood Publishing, 2008.

S. A. Hale. Global connectivity and multilinguals in the Twitter network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, pages 833–842, Toronto, Canada, 2014.

H. Hammarström. A fine-grained model for language identification. In *Proceedings of iNEWS-07 Workshop at SIGIR 2007*, pages 14–20, Amsterdam, The Netherlands, 2007.

B. Han, P. Cook, and T. Baldwin. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012: Technical Papers*, pages 1045–1062, Mumbai, India, 2012.

M. Han Veiga and C. Eickhoff. A cross-platform collection of social network profiles. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 2016)*, pages 665–668, Pisa, Italy, 2016.

A. Hassan, A. Abu-Jbara, and D. Radev. 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, 2012.

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

W. Heeringa and J. Nerbonne. *Language and Space. An International Handbook of Linguistic Variation, Volume III: Dutch*, chapter Dialectometry, pages 624–646. Berlin and New York: Walter de Gruyter, 2013.

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

L. Hemphill and J. Otterbacher. 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, 2012.

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

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

S. C. Herring. A faceted classification scheme for computer-mediated discourse. *Language@Internet*, 4, 2007.

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

T. Heskes, K. Albers, and B. Kappen. Approximate inference and constrained optimization.

In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 313–320, Acapulco, Mexico, 2002.

T. Hey, S. Tansley, and K. Tolle, editors. *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research, 2009.

F. Heylighen and J.-M. Dewaele. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3):293–340, 2002.

V. Hinnenkamp. Deutsch, Doyc or Doitsch? Chatters as languagers–The case of a German–Turkish chat room. *International Journal of Multilingualism*, 5(3):253–275, 2008.

L. Hinrichs. *Codeswitching on the Web: English and Jamaican Creole in E-mail Communication*. John Benjamins Publishing Company, 2006.

D. I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

J. Holmes. *Women, men and politeness*. Routledge, 1995.

J. Holmes. *An introduction to sociolinguistics*. Routledge, 4th edition, 2013.

J. Holmes and M. Meyerhoff, editors. *The handbook of language and gender*. Wiley-Blackwell, 2003.

O. R. Holsti. *Content analysis for the social sciences and humanities*. Addison-Wesley, 1969.

L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st international conference on World Wide Web (WWW '12)*, pages 769–778, Lyon, France, 2012.

D. Hovy. 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, 2015.

D. Hovy and A. Søgaard. 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, 2015.

D. Hovy, A. Johannsen, and A. Søgaard. 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, 2015.

Y. Hu, K. Talamadupula, and S. Kambhampati. 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, 2013.

F. Huang. 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, 2015.

F. Huang and A. Yates. 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, 2014.

J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in Twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia (HT '10)*, pages 173–

178, Toronto, Canada, 2010.

Y. Huang, D. Guo, A. Kasakoff, and J. Grieve. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255, 2016.

D. Huffaker. Dimensions of leadership and social influence in online communities. *Human Communication Research*, 36(4):593–617, 2010.

D. Huffaker, J. Jorgensen, F. Iacobelli, P. Tepper, and J. Cassell. 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, 2006.

B. Hughes, T. Baldwin, S. Bird, J. Nicholson, and A. Mackinlay. 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, 2006.

K. Hyland. *Disciplinary Discourses: Social Interactions in Academic Writing*. The University of Michigan Press, 2004.

Institute of Medicine and National Academy of Sciences and National Academy of Engineering. *Facilitating Interdisciplinary Research*. The National Academies Press, 2004.

M. Jain, J. McDonough, G. Gweon, B. Raj, and C. P. Rosé. 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, 2012.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI meeting corpus. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, pages 364–367, Hong Kong, 2003.

A. Johannsen, D. Hovy, and A. Søgaard. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China, 2015.

I. Johnson. Audience design and communication accommodation theory: Use of Twitter by Welsh-English biliterates. In E. H. G. Jones and E. Uribe-Jongbloed, editors, *Social Media and Minority Languages: Convergence and the Creative Industries*, pages 99–118. Multilingual Matters, 2013.

B. Johnstone. *Discourse Analysis*. Blackwell publishing, second edition, 2007.

S. Jones, R. Cotterill, N. Dewdney, K. Muir, and A. Joinson. 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, 2014.

V. Joosen. *Wit als sneeuw, zwart als inkt*. Lannoo, 2012.

A. K. Jørgensen, D. Hovy, and A. Søgaard. 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, 2015.

A. K. Joshi. 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, 1982.

M. Joshi, W. W. Cohen, M. Dredze, and C. P. Rosé. 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, 2012.

M. Joshi, M. Dredze, W. W. Cohen, and C. P. Rosé. What's in a domain? Multi-domain learning for multi-attribute data. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 685–690, Atlanta, Georgia, 2013.

D. Jurafsky, R. Ranganath, and D. McFarland. 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, 2009.

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

A. Karlqvist. Going beyond disciplines. *Policy Sciences*, 32(4):379–383, 1999.

F. Karsdorp and A. Van den Bosch. Identifying motifs in folktales using topic models. In *Proceedings of BENELEARN 2013*, Nijmegen, The Netherlands, 2013.

F. Karsdorp and A. Van den Bosch. The structure and evolution of story networks. *Royal Society Open Science*, 3(6), 2016.

F. Karsdorp, P. Kranenburg, T. Meder, D. Trieschnigg, and A. Van den Bosch. In search of an appropriate abstraction level for motif annotations. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative*, pages 22–26, Istanbul, Turkey, 2012a.

F. Karsdorp, P. van Kranenburg, T. Meder, and A. Van den Bosch. Casting a spell: Identification and ranking of actors in folktales. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in het Humanities (ACRH-2)*, pages 39–50, Lisbon, Portugal, 2012b.

F. Karsdorp, M. Van der Meulen, T. Meder, and A. Van den Bosch. MOMFER: a search engine of Thompson's motif-index of folk literature. *Folklore*, 126(1):37–52, 2015a.

F. Karsdorp, M. van der Meulen, T. Meder, and A. Van den Bosch. Animacy detection in stories. In *Proceedings of the Workshop on Computational Models of Narrative (CMN'15)*, pages 82–97, Atlanta, Georgia, USA, 2015b.

D. Kershaw, M. Rowe, and P. Stacey. 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, 2016.

J. Kim, G. Kazai, and I. Zitouni. Relevance dimensions in preference-based IR evaluation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13)*, pages 913–916, Dublin, Ireland, 2013.

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

B. King and S. Abney. 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, 2013.

B. King, D. Radev, and S. Abney. 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, 2014.

K. Kipper-Schuler. *VerbNet: a broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005.

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

B. Klimt and Y. Yang. 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, 2004.

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

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

F. Kooti, H. Yang, M. Cha, K. Gummadi, and W. Mason. 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, 2012.

M. Koppel, S. Argamon, and A. R. Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002.

A. Kovashka and M. Lease. Human and machine detection of stylistic similarity in art. In *Proceedings of CrowdConf 2010*, San Francisco, CA, 2010.

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

V. Krishnan and J. Eisenstein. "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, 2015.

E. Kypridemou and L. Michael. Narrative similarity as common summary. In *Proceedings of the Workshop on Computational Models of Narrative 2013*, pages 129–146, Hamburg, Germany, 2013.

K. A. La Barre and C. L. Tilley. The elusive tale: leveraging the study of information seeking and knowledge organization to improve access to and discovery of folktales. *Journal of the American Society for Information Science and Technology*, 63(4):687–701, 2012.

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

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

W. Labov. The intersection of sex and social class in the course of linguistic change. *Language variation and change*, 2(2):205–254, 1990.

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

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

J. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models

for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*, pages 282–289, Williamstown, MA, USA, 2001.

V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox. Assessing the impact of a health intervention via user-generated internet content. *Data Mining and Knowledge Discovery*, 29(5):1434–1457, 2015.

V. Lavrenko, J. Allan, E. DeGuzman, D. LaFlamme, V. Pollard, and S. Thomas. Relevance models for topic detection and tracking. In *Proceedings of the second international conference on Human Language Technology Research (HLT '02)*, pages 115–121, San Diego, California, USA, 2002.

D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Life in the network: the coming age of computational social science. *Science*, 323(5915): 721–723, 2009.

J. Le, A. Edmonds, V. Hester, and L. Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 17–20, Geneva, Switzerland, 2010.

D. Y. Lee. 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, 2001.

J. Lee and W. A. Kretzschmar. Spatial analysis of linguistic data with GIS functions. *International Journal of Geographical Information Science*, 7(6):541–560, 1993.

J. H. Lee. Crowdsourcing music similarity judgments using Mechanical Turk. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 183–188, Utrecht, Netherlands, 2010.

M. D. Lee, B. Pincombe, and M. B. Welsh. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259, Stresa, Italy, 2005.

A. Leemann, M.-J. Kolly, R. Purves, D. Britain, and E. Glaser. Crowdsourcing language change with smartphone applications. *PLoS ONE*, 11(1), 2016.

P. Legendre, M.-J. Fortin, and D. Borcard. Should the Mantel test be used in spatial analysis? *Methods in Ecology and Evolution*, 6(11):1239–1247, 2015.

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

V. A. Lestari and R. Manurung. Measuring the structural and conceptual similarity of folktales using plot graphs. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 25–33, Beijing, China, 2015.

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

R. Levitan, A. Gravano, and J. Hirschberg. 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, 2011.

Y. Li and P. Fung. Code-switch language model with inversion constraints for mixed language

speech recognition. In *Proceedings of COLING 2012: Technical Papers*, pages 1671–1680, Mumbai, India, 2012.

L. Liao, J. Jiang, Y. Ding, H. Huang, and E.-P. Lim. Lifetime lexical variation in social media. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 1643–1649, Québec City, Québec, Canada, 2014.

R. Ling. The sociolinguistics of SMS: An analysis of SMS use by a random sample of Norwegians. *Mobile Communications*, pages 335–349, 2005.

W. Ling, G. Xiang, C. Dyer, A. Black, and I. Trancoso. 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, 2013.

T.-Y. Liu. *Learning to Rank for Information Retrieval*. Foundations and Trends in Information Retrieval. Springer, 2011.

P. V. Lobo and D. M. de Matos. Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 1472–1475, Valletta, Malta, 2010.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444, 2002.

J. Lorber. Beyond the binaries: Depolarizing the categories of sex, sexuality, and gender*. *Sociological Inquiry*, 66(2):143–160, 1996.

M. Lui and T. Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 25–30, Jeju, Republic of Korea, 2012.

M. Lui, J. H. Lau, and T. Baldwin. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2(1):27–40, 2014.

M. Makatchev and R. Simmons. 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, 2011.

C. Mallinson, B. Childs, and G. Van Herk, editors. *Data Collection in Sociolinguistics: Methods and Applications*. Routledge, 2013.

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

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

L. Manovich. Cultural analytics: Analysis and visualization of large cultural data sets. Technical report, Software Studies Initiative @ Calit2/UCSD, 2007.

L. Manovich. The science of culture? Social computing, digital humanities and cultural analytics. *Journal of Cultural Analytics*, 2016.

J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, 2011.

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

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

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

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

A. McAfee and E. Brynjolfsson. Big data. *Harvard Business Review*, 90(10):60–68, 2012.

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

T. Meder. From a Dutch Folktale Database towards an International Folktale Database. *Fabula*, 51(1-2):6–22, 2010.

T. Meder, D. Nguyen, and R. Gravel. The apocalypse on Twitter. *Digital Scholarship in the Humanities*, 2015.

T. Meder, F. Karsdorp, D. Nguyen, M. Theune, D. Trieschnigg, and I. E. C. Muiser. Automatic enrichment and classification of folktales in the Dutch folktale database. *The Journal of American Folklore*, 129(511):78–96, 2016.

D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel. Similarity measures for tracking information flow. In *Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM '05)*, pages 517–524, Bremen, Germany, 2005.

M. Meyerhoff. *Introducing Sociolinguistics*, chapter Real time and apparent time. Routledge, 2006.

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

L. Michael. Similarity of narratives. In *Proceedings of the LREC 2012 Workshop on Computational Models of Narrative*, pages 103–111, Istanbul, Turkey, 2012.

L. Michael and J. Otterbacher. 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, 2014.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop*, 2013.

H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht. "Blissfully happy" or "ready to fight": Varying interpretations of emoji. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, pages 259–268, Cologne, Germany, 2016.

J. Milroy and L. Milroy. Belfast: change and variation in an urban vernacular. In P. Trudgill, editor, *Sociolinguistic patterns in British English*, pages 19–36. London: Edward Arnold, 1978.

J. Milroy and L. Milroy. Linguistic change, social network and speaker innovation. *Journal of linguistics*, 21(2):339–384, 1985.

L. Milroy and M. Gordon. *Sociolinguistics: Method and Interpretation*. Wiley-Blackwell, 2nd edition, 2003.

A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. 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, 2011.

D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE*, 8(4): e61981, 2013.

S. Mohammad. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA, 2011.

P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1-2):17–23, 1950.

R. A. Morrow and D. D. Brown. *Contemporary Social Theory: Critical Theory and Methodology*, chapter Deconstructing the Conventional Discourse of Methodology: Quantitative versus Qualitative Methods. SAGE Publications, 1994.

F. Moscoso del Prado Martin and C. Brendel. Case and cause in Icelandic: Reconstructing causal networks of cascaded language changes. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2421–2430, Berlin, Germany, 2016.

K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyonds. *arXiv preprint arXiv:1605.09522*, 2016.

H. Mubarak and K. Darwish. 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, 2014.

A. Mukherjee and B. Liu. Improving gender classification of blog authors. In *Proceedings of the 2010 conference on Empirical Methods in Natural Language Processing*, pages 207–217, Cambridge, MA, 2010.

R. Munro, S. Bethard, V. Kuperman, V. T. Lai, R. Melnick, C. Potts, T. Schnoebelen, and H. Tily. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130, Los Angeles, California, 2010.

C. Myers-Scotton. *Social motivations for codeswitching: Evidence from Africa*. Oxford: Clarendon, 1995.

C. Myers-Scotton. *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford: Oxford University Press, 2002.

G. Nathalie, L. B. Hervé, H. Jeanny, and G.-D. Anne. Towards the introduction of human perception in a natural scene classification system. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing*, pages 385 – 394, Martigny, Switzerland, 2002.

R. M. A. Nawab, M. Stevenson, and P. Clough. Retrieving candidate plagiarised documents using query expansion. In *Proceedings of the 34th European Conference on IR Research (ECIR 2012)*, pages 207–218, Barcelona, Spain, 2012.

J. Nerbonne. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198, 2009.

J. Nerbonne and W. A. Kretzschmar Jr. Dialectometry++. *Literary and Linguistic Computing*, 28(1):2–12, 2013.

J. Nerbonne and M. Wieling. Statistics for aggregate variationist analyses. In C. Boberg, J. Nerbonne, and D. Watt, editors, *Handbook of Dialectology*. Boston: Wiley, 2015.

D. Nguyen and L. Cornips. Automatic detection of intra-word code-switching. In *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 82–86, Berlin, Germany, 2016.

D. Nguyen and A. S. Doğruöz. 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, 2013.

D. Nguyen and C. P. Rosé. 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, 2011.

D. Nguyen, N. A. Smith, and C. P. Rosé. 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, 2011.

D. Nguyen, D. Trieschnigg, T. Meder, and M. Theune. Automatic classification of folk narrative genres. In *Proceedings of the Workshop on Language Technology for Historical Text(s) at KONVENS 2012*, pages 378–382, Vienna, Austria, 2012.

D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. "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, 2013a.

D. Nguyen, D. Trieschnigg, and M. Theune. Folktale classification using learning to rank. In *Proceedings of the 35th European Conference on IR Research (ECIR 2013)*, pages 195–206, Moscow, Russia, 2013b.

D. Nguyen, D. Trieschnigg, A. S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F. de Jong. 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, 2014a.

D. Nguyen, D. Trieschnigg, and T. Meder. 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, 2014b.

D. Nguyen, D. Trieschnigg, and M. Theune. Using crowdsourcing to investigate perception of narrative similarity. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 321–330, Shanghai, China, 2014c.

D. Nguyen, D. Trieschnigg, and L. Cornips. 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, 2015a.

D. Nguyen, T. van den Broek, C. Hauff, D. Hiemstra, and M. Ehrenhard. #Supportthecause: Identifying motivations to participate in online health campaigns. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2570–2576, Lisbon, Portugal, 2015b.

D. Nguyen, A. S. Doğruöz, C. P. Rosé, and F. de Jong. Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593, 2016.

V.-A. Nguyen, J. Boyd-Graber, P. Resnik, D. A. Cai, J. E. Midberry, and Y. Wang. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95(3):381–421, 2014d.

K. G. Niederhoffer and J. W. Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360, 2002.

M. Nissani. Ten cheers for interdisciplinarity: The case for interdisciplinary knowledge and research. *The Social Science Journal*, 34(2):201–216, 1997.

B. Noble and R. Fernández. 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, 2015.

S. Nowson and J. Oberlander. 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, 2006.

S. Nowson, J. Oberlander, and A. J. Gill. Weblogs, genres and individual differences. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1666–1671, Stresa, Italy, 2005.

B. O'Connor, M. Krieger, and D. Ahn. TweetMotif: exploratory search and topic summarization for Twitter. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 384–385, Washington, DC, 2010.

N. Ofek, S. Darányi, and L. Rokach. Linking motif sequences with tale types by machine learning. In *Proceedings of the Workshop on Computational Models of Narrative 2013*, pages 166–182, Hamburg, Germany, 2013.

K. Ord. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349):120–126, 1975.

J. Otterbacher. 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, 2010.

I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, Seattle, Washington, USA, 2006.

J. C. Paolillo. Language variation on Internet Relay Chat: A social network approach. *Journal of sociolinguistics*, 5(2):180–213, 2001.

E. E. Papalexakis, D. Nguyen, and A. S. Doğruöz. 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, 2014.

M. J. Paul. Interpretable machine learning: lessons from topic modeling. In *Proceedings of the CHI Workshop on Human-Centered Machine Learning*, San Jose, CA, USA, 2016.

U. Pavalanathan and J. Eisenstein. 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, 2015a.

U. Pavalanathan and J. Eisenstein. Audience-modulated variation in online social media. *American Speech*, 90(2):187–213, 2015b.

E. Pavlick, M. Post, A. Irvine, D. Kachaev, and C. Callison-Burch. The language demographics

of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2:79–92, 2014.

E. A. Pechenick, C. M. Danforth, and P. S. Dodds. Characterizing the Google books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*, 10(10): e0137041, 2015.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

C. Peersman, W. Daelemans, and L. V. Vaerenbergh. 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, 2011.

Y. Peirsman, D. Geeraerts, and D. Speelman. The automatic identification of lexical variation between language varieties. *Natural Language Engineering*, 16(04):469–491, 2010.

N. Peng, Y. Wang, and M. Dredze. 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, 2014.

M. Pennacchiotti and A. Popescu. 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, 2011.

J. W. Pennebaker and L. D. Stone. Words of wisdom: language use over the life span. *Journal of personality and social psychology*, 85(2):291–301, 2003.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic Inquiry and Word Count: LIWC 2001. *Mahwah, NJ: Lawrence Erlbaum*, 2001.

K. Peterson, M. Hohensee, and F. Xia. 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, 2011.

J. Pfeffer and G. R. Salancik. *The External Control of Organizations: A Resource Dependence Perspective*. New York: Harper & Row, 1978.

M. Piergallini, A. S. Doğruöz, P. Gadde, D. Adamson, and C. P. Rosé. 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, 2014.

M. Piotrowski. Natural language processing for historical texts. *Synthesis Lectures on Human Language Technologies*, 2012.

O. Popescu and C. Strapparava. Time corpora: Epochs, opinions and changes. *Knowledge-Based Systems*, 69:3–13, 2014.

S. Poplack, D. Sankoff, and C. Miller. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics*, 26(1):47–104, 1988.

T. Postmes, R. Spears, and M. Lea. The formation of group norms in computer-mediated communication. *Human Communication Research*, 26(3):341–371, 2000.

V. Prabhakaran and O. Rambow. 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, 2013.

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

V. Prabhakaran, O. Rambow, and M. Diab. 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, 2012b.

V. Prabhakaran, A. John, and D. D. Seligmann. 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, 2013.

V. Prabhakaran, A. Arora, and O. Rambow. 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, 2014a.

V. Prabhakaran, E. E. Reid, and O. Rambow. 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, 2014b.

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

D. Preoţiuc-Pietro, V. Lampos, and N. Aletras. 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, 2015a.

D. Preoţiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras. Studying user income through language, behaviour and affect in social media. *PLoS ONE*, 10(9):e0138717, 2015b.

D. R. Preston. Perceptual dialectology. In C. Boberg, J. Nerbonne, and D. Watt, editors, *Handbook of Dialectology*. Wiley, 2015.

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

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

S. Rabe-Hesketh and A. Skrondal. *Multilevel and Longitudinal Modeling Using Stata*. Stata Press, 2012.

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

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

F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. Overview of the author profiling

task at PAN 2013. In *CLEF 2013 Evaluation Labs and Workshop Working Notes Papers*, Valencia, Spain, 2013.

F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, and W. Daelemans. Overview of the 2nd author profiling task at PAN 2014. In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, Sheffield, United Kingdom, 2014.

D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. 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, Canada, 2010.

D. Rao, M. J. Paul, C. Fink, D. Yarowsky, T. Oates, and G. Coppersmith. 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, 2011.

R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks.*, pages 46–50, Valletta, Malta, 2010.

B. T. Ribeiro. Footing, positioning, voice. Are we talking about the same things? In A. D. Fina, D. Schiffrin, and M. Bamberg, editors, *Discourse and Identity*, pages 48–82. Cambridge University Press, 2006.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016.

K. Richards. *Language and Professional Identity: Aspects of Collaborative Interaction*. Palgrave Macmillan, 2006.

A. Riemersma, D. Gorter, and J. Ytsma. Frisian in the Netherlands. In *The other languages of Europe: Demographic, Sociolinguistic and Educational Perspectives*, pages 103–118. Clevedon: Multilingual Matters, 2001.

S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1500–1510, Jeju Island, Korea, 2012.

S. Romaine. Bilingualism (2nd edition). *Malden, MA: Blackwell Publishers*, 1995.

C. P. Rosé. *International Handbook of the Learning Sciences*, chapter Learning analytics in the Learning Sciences. Taylor & Francis, in press.

C. P. Rosé and A. Tovares. *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, 2015.

C. P. Rosé, Y.-C. Wang, Y. Cui, J. Arguello, K. Stegmann, A. Weinberger, and F. Fischer. 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, 2008.

S. Rosenthal and K. McKeown. 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, 2011.

J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: Shifting demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872, Atlanta, GA, USA, 2010.

H. Sak, T. Güngör, and M. Saraçlar. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL 2008)*, pages 417–427, Gothenburg, Sweden, 2008.

G. Sankoff. Age: Apparent time and real time. *Encyclopedia of Language and Linguistics. Oxford, UK: Elsevier*, 2006.

M. Sap, G. Park, J. Eichstaedt, M. Kern, D. Stillwell, M. Kosinski, L. Ungar, and A. H. Schwartz. 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, 2014.

R. Sarawgi, K. Gajulapalli, and Y. Choi. 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, 2011.

R. Schäfer and F. Bildhauer. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 486–493, Istanbul, Turkey, 2012.

E. A. Schegloff. *Sequence Organization in Interaction: A Primer in Conversation Analysis*, volume 1. Cambridge University Press, 2007.

Y. Scherrer. Recovering dialect geography from an unaligned comparable corpus. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 63–71, Avignon, France, 2012.

Y. Scherrer and O. Rambow. 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, 2010.

J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker. 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, 2006.

G. Schneider, J. Dowdall, and F. Rinaldi. 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, 2004.

B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. The INTERSPEECH 2010 paralinguistic challenge. In *Proceedings of INTERSPEECH*, pages 2794–2797, Makuhari, Chiba, Japan, 2010.

H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9): e73791, 2013.

L. E. Scissors, A. J. Gill, K. Geraghty, and D. Gergle. 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, 2009.

D. Sculley. Large scale learning to rank. In *NIPS 2009 Workshop on Advances in Ranking*, 2009.

J. R. Searle. *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press, 1969.

J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

G. Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

L. Si, R. Jin, J. Callan, and P. Ogilvie. A language modeling framework for resource selection and results merging. In *Proceedings of the eleventh international conference on Information and knowledge management (CIKM '02)*, pages 391–397, McLean, Virginia, USA, 2002.

S. Singh. A pilot study on gender differences in conversational speech on lexical richness measures. *Literary and Linguistic Computing*, 16(3):251–264, 2001.

J. R. Sinninghe. *Katalog der niederländischen Märchen-, Ursprungssagen-, Sagen-und Legendenvarianten*. Number 132 in FF Communications. Helsinki Suomalainen Tiedeakademia, 1943.

L. Sloan, J. Morgan, P. Burnap, and M. Williams. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PLoS ONE*, 10(3):e0115545, 2015.

D. A. Snow and L. Anderson. Identity work among the homeless: The verbal construction and avowal of personal identities. *American Journal of Sociology*, 92(6):1336–1371, 1987.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. 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, 2008.

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

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

T. Solorio and Y. Liu. 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, 2008b.

T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Ghoneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang, and P. Fung. 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, 2014.

B. Sommerdijk, E. Sanders, and A. Van den Bosch. Can tweets predict TV ratings? In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 2016.

L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(1):1393–1434, 2012.

M. R. Spruit. Measuring syntactic variation in Dutch dialects. *Literary and Linguistic Computing*, 21(4):493–506, 2006.

S. Štajner and R. Mitkov. Diachronic stylistic changes in British and American varieties of 20th

century written English language. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage at RANLP*, pages 78–85, Hissar, Bulgaria, 2011.

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

W. Stoop and A. Van den Bosch. 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, 2014.

T. Strzalkowski, S. Shaikh, T. Liu, G. A. Broadwell, J. Stromer-Galley, S. Taylor, U. Boz, V. Ravishankar, and X. Ren. Modeling leadership and influence in multi-party online discourse. In *Proceedings of COLING 2012*, pages 2535–2552, Mumbai, India, 2012.

S. Swayamdipta and O. Rambow. 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, 2012.

B. Szmrecsanyi. *Grammatical variation in British English dialects: a study in corpus-based dialectometry*. Cambridge University Press, 2012.

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

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

M. Talbot. 20 gender stereotypes: Reproduction and challenge. In J. Holmes and M. Meyerhoff, editors, *The handbook of language and gender*. John Wiley & Sons, 2008.

J. Tam and C. H. Martell. Age detection in chat. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC '09)*, pages 33–39, Berkeley, CA, 2009.

T. R. Tangherlini. Big folklore: A special issue on computational folkloristics. *The Journal of American Folklore*, 129(511):5–13, 2016.

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

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

J. J. Tehrani. The phylogeny of Little Red Riding Hood. *PLoS ONE*, 8(11):e78871, 2013.

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

S. Thompson. *The folktale*. Dryden Press, 1951.

S. Thompson. *Motif-Index of Folk-Literature: A Classification of Narrative Elements in Folktales, Ballads, Myths, Fables, Mediaeval Romances, Exempla, Fabliaux, Jest-books and Local Legends*. Bloomington, Indiana University Press, 1955-1958.

E. Tjong Kim Sang. Discovering dialect regions in syntactic dialect data. In *Workshop European Dialect Syntax VIII - Edisyn 2015*. Zurich, Switserland, 2015.

G. Tottie and S. Hoffmann. Tag questions in British and American English. *Journal of English Linguistics*, 34(4):283–311, 2006.

D. Trieschnigg, D. Hiemstra, M. Theune, F. Jong, and T. Meder. 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, 2012.

D. Trieschnigg, D. Nguyen, and T. Meder. In search of Cinderella: A transaction log analysis of folktale searchers. In *Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage (ENRICH 2013)*, Dublin, Ireland, 2013a.

D. Trieschnigg, D. Nguyen, and M. Theune. Learning to extract folktale keywords. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 65–73, Sofia, Bulgaria, 2013b.

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

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

K. P. Truong, G. J. Westerhof, S. M. A. Lamers, and F. de Jong. 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, 2014.

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

Z. Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, pages 505–514, Oxford, UK, 2014.

J. Urbano, J. Morato, M. Marrero, and D. Martín. Crowdsourcing preference judgments for evaluation of music similarity tasks. In *Proceedings of the SIGIR workshop on crowdsourcing for search evaluation*, pages 9–16, Geneva, Switzerland, 2010.

H. J. Uther. Type- and motif-indices 1980-1995: An inventory. *Asian Folklore Studies*, 55(2): 299–317, 1996.

H. J. Uther. *The Types of International Folktales: A Classification and Bibliography Based on the System of Antti Aarne and Stith Thompson. Vols 1-3.* Suomalainen Tiedeakatemia, Helsinki, 2004.

H. J. Uther. Classifying tales: Remarks to indexes and systems of ordering. *Folks Art - Croatian Journal Of Ethnology and Folklore Research*, 46(1):15–32, 2009.

A. Van den Bosch, B. Busser, S. Canisius, and W. Daelemans. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, pages 99–114, Leuven, Belgium, 2007.

J. Van Dijck. 'You have one identity': performing the self on Facebook and LinkedIn. *Media, Culture & Society*, 35(2):199–215, 2013.

B. Van Durme. 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, 2012.

T. Vatanen, J. J. Väyrynen, and S. Virpioja. Language identification of short text segments with n-gram models. In *Proceedings of LREC 2010*, pages 3423–3430, Valletta, Malta, 2010.

R. Vliegendhart, M. Larson, and J. A. Pouwelse. Discovering user perceptions of semantic similarity in near-duplicate multimedia files. In *Proceedings of the First International Workshop on Crowdsourcing Web Search (CrowdSearch 2012)*, pages 54–58, Lyon, France, 2012.

V. Voigt, M. Preminger, L. Ládi, and S. Darányi. Automated motif identification in folklore text

corpora. *Folklore*, 12, 1999.

S. Volkova, T. Wilson, and D. Yarowsky. 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, 2013.

C. Voss, S. Tratz, J. Laoudi, and D. Briesch. 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, 2014.

P. Vossen, K. Hofmann, M. Rijke, E. Tjong, K. Sang, and K. Deschacht. The Cornetto database: Architecture and user-scenarios. In *DIR 2007*, Leuven, Belgium, 2007.

Y. Vyas, S. Gella, J. Sharma, K. Bali, and M. Choudhury. 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, 2014.

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

K. L. Wagstaf. Machine learning that matters. In *Proceedings of the 29th International Conference on Machine Learning*, pages 529–536, Edinburgh, Scotland, UK, 2012.

W. Wang, D. Rothschild, S. Goel, and A. Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980 – 991, 2015.

Y. Wang, D. Reitter, and J. Yen. 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, 2014.

R. Wardhaugh. *An Introduction to Sociolinguistics*. Wiley-Blackwell, 6th edition, 2011.

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

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

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

M. Wen, D. Yang, and C. P. Rosé. 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, 2014a.

M. Wen, D. Yang, and C. P. Rosé. 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, 2014b.

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

M. Wieling and J. Nerbonne. 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, 2010.

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

M. Wieling, J. Bloem, K. Mignella, M. Timmermeister, and J. Nerbonne. Measuring foreign accent strength in English: Validating Levenshtein distance as a measure. *Language Dynamics and Change*, 4(2):253–269, 2014.

W. Wiersma, J. Nerbonne, and T. Lauttamus. Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing*, 26(1):107–124, 2011.

B. Wing and J. Baldridge. 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, 2011.

S. Wintner. 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, 2002.

T. F. Wright. Regional dialects in the contact call of a parrot. *Proceedings of the Royal Society of London B: Biological Sciences*, 263(1372):867–872, 1996.

M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama. High-dimensional feature selection by feature-wise kernelized lasso. *Neural computation*, 26(1):185–207, 2014.

H. Yamaguchi and K. Tanaka-Ishii. 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, 2012.

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

D. Yang, M. Wen, and C. P. Rosé. 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, 2015.

J. Yang and S. Counts. Comparing information diffusion structure in weblogs and microblogs. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 351 – 354, Washington, DC, 2010.

T. Yarkoni and J. Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning. *In submission*, 2016.

J. Yi, R. Jin, A. K. Jain, S. Jain, and T. Yang. Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning. In *Advances in Neural Information Processing Systems 25*, pages 1772–1780, Lake Tahoe, Nevada, USA, 2012.

E. Zagheni and I. Weber. Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25, 2015.

O. F. Zaidan and C. Callison-Burch. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202, 2014.

F. A. Zamal, W. Liu, and D. Ruths. 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, 2012.

M. Zampieri, L. Tan, N. Ljubešić, and J. Tiedemann. 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, 2014.

M. Zampieri, L. Tan, N. Ljubešić, J. Tiedemann, and P. Nakov. 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, 2015.

M. Zengin and B. Carterette. User judgements of document similarity. In *Proceedings of the SIGIR 2013 Workshop on Modeling User Behavior for Information Retrieval Evaluation (MUBE 2013)*, Dublin, Ireland, 2013.

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

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *arXiv preprint arXiv:1606.07892*, 2016.

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

J. Zipes, editor. *The Trials & Tribulations of Little Red Riding Hood*. Psychology Press, 1993.

Publications

This dissertation is based on the following publications:

- **D. Nguyen** and A. S. Doğruöz, *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, 2013.
  Data: http://www.dongnguyen.nl/data-langid-emnlp2013.html
- **D. Nguyen**, R. Gravel, D. Trieschnigg, and T. Meder, *"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, 2013.
- **D. Nguyen**, D. Trieschnigg, and M. Theune, *Folktale classification using learning to rank*, In Proceedings of the 35th European Conference on IR Research (ECIR 2013), pages 195-206, Moscow, Russia, 2013.
- **D. Nguyen**, D. Trieschnigg, A. S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F. de Jong, *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, 2014.
- **D. Nguyen**, D. Trieschnigg and M. Theune, *Using crowdsourcing to investigate perception of narrative similarity*, In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 321-330, Shanghai, China, 2014.
- **D. Nguyen**, D. Trieschnigg and T. Meder, *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, 2014.
- **D. Nguyen**, D. Trieschnigg, and L. Cornips, *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, 2015.
- **D. Nguyen**, A. S. Doğruöz, C. P. Rosé, and F. de Jong, *Computational sociolinguistics: A survey*, In Computational Linguistics, 42(3), pages 537-593, 2016.
- **D. Nguyen** and J. Eisenstein, *A kernel independence test for geographical language variation*, To appear in Computational Linguistics. Presented at NWAV 44, 2015.

This dissertation was also influenced by the following work carried out by the author during the PhD program:

- **D. Nguyen**, T. Demeester, D. Trieschnigg, D. Hiemstra, *Federated search in the wild*, In Proceedings of the 21st ACM international Conference on Information and Knowledge Management, pages 1874-1878, Maui, Hawaii, USA, 2012.
- **D. Nguyen**, D. Trieschnigg, T. Meder and M. Theune, *Automatic classification of folk narrative genres*, In Proceedings of the First International Workshop on Language Technology for Historical Text(s), pages 378-382, Vienna, Austria, 2012.
- **D. Nguyen**, T. van den Broek, C. Hauff, D. Hiemstra and M. Ehrenhard, *#SupportTheCause: Identifying motivations to participate in online health campaigns*, In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2570-2576, Lisbon, Portugal, 2015.
- **D. Nguyen** and Leonie Cornips, *Automatic detection of intra-word code-switching*, In Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 82–86, Berlin, Germany, 2016.
- T. Meder, **D. Nguyen**, R. Gravel, *The apocalypse on Twitter*, Digital Scholarship in the Humanities, 2015.
- N. Dwi Prasetyo, C. Hauff, **D. Nguyen**, T. van den Broek, and D. Hiemstra, *On the impact of Twitter-based health campaigns: A cross-country analysis of Movember*. In Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis, pages 55-63, Lisbon, Portugal, 2015.
- T. Meder, F. Karsdorp, **D. Nguyen**, M. Theune, D. Trieschnigg, I. Muiser, *Automatic enrichment and classification of folktales in the Dutch folktale database*, Journal of American Folklore, Volume 129, Number 511, pages 78-96, 2016.

Furthermore, earlier work by the author [Nguyen and Rosé, 2011, Nguyen et al., 2011] inspired parts of this dissertation.

Summary

The world is becoming increasingly digital. We create content by posting on social media, we track our movements using mobile apps, we read books on e-readers, and so on. Massive digital datasets, such as social media data, are a promising source to study social and cultural phenomena. They provide the opportunity to study language use and behavior in a variety of social situations on a large scale and often with the availability of detailed contextual information. Thus, such datasets could potentially have a large impact on research in the social sciences and the humanities. However, to fully leverage their potential, new computational approaches are needed. This dissertation explores computational approaches to text analysis for studying cultural and social phenomena and focuses on two emerging areas: computational sociolinguistics and computational folkloristics. Both areas share the recognition that variation in text is often meaningful and may provide insights into social and cultural phenomena. This dissertation focuses on computational approaches to analyze and model this type of variation.

The first part of this dissertation focuses on computational sociolinguistics. So far, the dominant approach for text analysis using computational approaches has been based on the limited view of language as a means to convey information. However, language also serves an important social role and variation is inherent to its social nature. In parallel with the rise of social media and the increasing interest in using large textual datasets for studying social phenomena, the area of computational sociolinguistics is emerging, in which computational approaches are used to study the relation between language and society. Variation in language use is often socially significant and plays a central role in (computational) sociolinguistics. An example of variation is that social media users may use multiple language varieties in their social media messages (e.g., English, Dutch and a regional dialect). The social context (e.g., audience, goal) often influences which language variety is used.

The work reported in this dissertation starts with studies on language variation according to the gender, age and location of authors. First, by building on the insight that language use in texts may be related to the identity of their authors, the tasks of gender and age prediction were explored. An automatic system predicting age based on tweets was developed. A demo, called TweetGenie, resulted from this research and the data collected using TweetGenie was used to evaluate the age prediction system 'in the wild'. Furthermore, the collected data was used as a source to reflect on the tasks of automatic gender and age prediction from text and, in particular, highlighted limitations of representing gender and age as static variables in computational approaches. The next study focused on geographical variation. A key step in many dialect studies is selecting a set of variables that exhibit geographical variation (e.g., two different ways of referring to the same concept). This dissertation provided an extensive comparison of existing methods to test for

geographical linguistic variation and proposed the use of a non-parametric approach that overcomes important limitations of existing methods. The work on computational sociolinguistics concluded with studies that focused on online multilingual communication. In both spoken and written communication, multilingual speakers may use multiple languages in a single conversation (a phenomenon also known as code-switching). To support the processing and analysis of code-switching in texts, automatic identification of languages at a word level was explored. Furthermore, by using automatic language identification a study was carried out to analyze the influence of audiences on the language variety used in Twitter.

The second part of this dissertation focuses on the emerging area of computational folkloristics, a field in which computational approaches are leveraged to study folklore. Variation in folk narratives emerges due to oral and written transmission of the stories across time and space. Modifications to stories are often motivated by social reasons and thus variation in folk narratives is often meaningful. The work conducted in this dissertation focused on variation in the folk narratives from the Dutch Folktale Database. A study was set up to explore supervised machine learning approaches to automatically identify similar folk narratives (e.g., different variants of *Little Red Riding Hood*) according to the concept of tale types, which are used in folk narrative research to organize and analyze folk narratives. Secondly, the perception of narrative similarity was investigated by asking experts and non-experts (crowdworkers) to rate the similarity of folk narratives. The study showed some limitations of the concept of tale types and highlighted the need to adapt to individual user differences when measuring folk narrative similarity.

As the work presented here demonstrates, the combination of large social and cultural datasets, text analysis and computational approaches has a huge potential for impact on the social sciences and the humanities, but many challenges remain to be addressed. First, many large datasets (e.g., obtained from sources such as social media) contain various properties that are very useful for research purposes (e.g., rich contextual information), but they have often been generated and collected in less controlled settings than the datasets that linguists and social scientists traditionally tend to work with. As a result, such datasets tend to exhibit known and unknown biases. These biases may have an influence on the composition of the demographic groups, language use, and online behavior. Second, various challenges revolve around ethical concerns, in terms of privacy (e.g., data collection and storage), but also in terms of applications (e.g., classifiers that encode biases in society). Third, before computational models are likely to be adopted by researchers outside the field of computer science, the interpretability of the models requires more attention in order to support exploratory analyses and theory development. And finally, while this dissertation focused on variation at a specific point in time, the increasing availability of longitudinal datasets provides exciting opportunities to study change over time.

Samenvatting

De wereld wordt steeds digitaler. We genereren data door het plaatsen van berichten op sociale media, we houden onze bewegingen bij met behulp van mobiele applicaties, we lezen boeken op e-readers, enzovoorts. Gigantische digitale collecties, zoals data afkomstig van sociale media, zijn een veelbelovende bron om sociale en culturele fenomenen te bestuderen. Taalgebruik en gedrag kan nu op een grote schaal en in een diversiteit van sociale situaties bestudeerd worden. Deze nieuwe bronnen hebben dus de potentie om een grote impact te hebben op onderzoek in de geestes- en sociale wetenschappen. Nieuwe computationele methoden zijn echter nodig om optimaal gebruik te maken van deze bronnen. Dit proefschrift onderzoekt computationele methoden voor het analyseren van teksten om sociale en culturele fenomenen te bestuderen en richt zich op twee gebieden die in opkomst zijn: computationele sociolinguïstiek en computationele volkskunde. Beide gebieden delen het idee dat variatie in tekst vaak betekenisvol is en inzicht kan geven in sociale en culturele fenomenen. Dit proefschrift richt zich op computationele methoden om dit type variatie te analyseren en te modeleren.

Het eerste deel van dit proefschrift richt zich op de computationele sociolinguïstiek. De meeste computationele methoden voor tekstanalyse zijn tot nu toe gebaseerd op de beperkende gedachte dat taal vooral gebruikt wordt om informatie over te brengen. Taal speelt echter ook een belangrijke sociale rol en variatie is inherent aan het sociale karakter van taal. Parallel aan de opkomst van sociale media en het toenemende gebruik van grote tekstcollecties om sociale fenomenen te bestuderen, is de computationele sociolinguïstiek in opkomst. In dit onderzoeksgebied worden computationele methoden gebruikt om de relatie tussen taal en samenleving te bestuderen. Variatie in taalgebruik heeft vaak een sociale betekenis en speelt een belangrijke rol in de (computationele) sociolinguïstiek. Een voorbeeld van variatie is dat gebruikers op sociale media meerdere talen kunnen gebruiken (bijv. Engels, Nederlands en een lokaal dialect). De sociale context (zoals het publiek en het doel) beïnvloedt vaak welke taal gebruikt wordt.

Dit proefschrift begint met onderzoek waarin taalvariatie met betrekking tot het geslacht, de leeftijd en de locatie van auteurs centraal staat. Eerst werd het automatisch bepalen van geslacht en leeftijd op basis van taalgebruik onderzocht. Op basis van dit onderzoek werd een demo, TweetGenie, ontwikkeld. De daarmee verzamelde data werd gebruikt om beperkingen van computationele tekstanalysemethoden aan het licht te brengen die gericht zijn op de sociale variabelen geslacht en leeftijd. De volgende studie richtte zich op geografische variatie. Een belangrijk onderdeel in veel dialectonderzoek is het selecteren van variabelen die geografische variatie bevatten (bijv. twee verschillende manieren om naar hetzelfde concept te verwijzen). Dit proefschrift bevat een uitgebreide vergelijking van bestaande methoden om te testen of een variabele geografische variatie bevat. Bovendien wordt het gebruik

van een niet-parametrische methode voorgesteld die de beperkingen van bestaande methoden oplost. Dit deel van het proefschrift sluit af met onderzoek gericht op meertalige communicatie in sociale media. Meertalige sprekers maken vaak gebruik van meerdere talen in één gesprek (vaak wordt dit aangeduid als 'codewisselingen'). Om teksten met codewisselingen te kunnen verwerken en analyseren, onderzoekt dit proefschrift automatische taalidentificatie op het niveau van individuele woorden. Bovendien is er een studie uitgevoerd naar de invloed van het beoogde publiek op de taalkeuze in Twitter.

In het tweede deel van dit proefschrift staat de computationele volkskunde centraal, een opkomend onderzoeksgebied waarin computationele methoden gebruikt worden om volkskunde te beoefenen. Variatie in volksverhalen ontstaat doordat verhalen overgebracht worden in mondelinge en geschreven vorm. Veranderingen ontstaan vaak om sociale redenen en variatie in volksverhalen is daarom vaak betekenisvol. Het onderzoek in dit proefschrift richt zich op variatie in volksverhalen in de Nederlandse Volksverhalenbank. Gesuperviseerde leeralgoritmen werden gebruikt om automatisch soortgelijke volksverhalen te identificeren (zoals verschillende versies van *Roodkapje*) op basis van verhaaltypen, die gebruikt worden in volksverhalenonderzoek om verhalen te organiseren en te analyseren. Vervolgens werd onderzocht hoe experts en niet-experts (door middel van crowdsourcing) de gelijkenis tussen volksverhalen waarnemen. Het onderzoek toonde enkele zwakheden aan van het concept van verhaaltypen en liet zien dat het belangrijk is om de manier waarop gelijkenissen tussen verhalen gemeten worden aan te passen aan de gebruiker van het systeem.

Zoals het hier beschreven onderzoek laat zien, kan de combinatie van grote sociale en culturele datacollecties, tekstanalyse en computationele methoden een grote impact hebben op de geestes- en sociale wetenschappen. Maar er zijn ook nog veel uitdagingen. Ten eerste bevatten veel grote datacollecties (bijv. op basis van sociale media) weliswaar informatie die nuttig is voor onderzoek (zoals rijke contextuele informatie), maar deze data is vaak gegenereerd en verzameld in minder gecontroleerde omgevingen dan de datacollecties waar onderzoekers van de geestes- en sociale wetenschappen normaalgesproken mee werken. Als gevolg daarvan bevatten deze datacollecties onzuiverheden die niet altijd bekend zijn. Ten tweede zijn er veel ethische uitdagingen, zoals privacyaspecten (bijv. datacollectie en -opslag) en mogelijke toepassingen van de ontwikkelde systemen (bijv. algoritmen die ongelijkheden in de samenleving versterken). Ten derde, voordat de ontwikkelde methoden gebruikt zullen worden door onderzoekers buiten de informatica, is het belangrijk dat de transparantie van de modellen meer aandacht krijgt om exploratief onderzoek en theorieontwikkeling te ondersteunen. Tot slot, dit proefschrift richt zich op variatie in een bepaalde tijdsperiode, maar de toenemende beschikbaarheid van datacollecties die een langere tijdsperiode beslaan biedt de mogelijkheid om veranderingen over tijd te onderzoeken.

SIKS Dissertatiereeks

2009-01 Rasa Jurgelenaite (RUN) *Symmetric Causal Independence Models*

2009-02 Willem Robert van Hage (VU) *Evaluating Ontology-Alignment Techniques*

2009-03 Hans Stol (UvT) *A Framework for Evidence-based Policy Making Using IT*

2009-04 Josephine Nabukenya (RUN) *Improving the Quality of Organisational Policy Making using Collaboration Engineering*

2009-05 Sietse Overbeek (RUN) *Bridging Supply and Demand for Knowledge Intensive Tasks – Based on Knowledge, Cognition, and Quality*

2009-06 Muhammad Subianto (UU) *Understanding Classification*

2009-07 Ronald Poppe (UT) *Discriminative Vision-Based Recovery and Recognition of Human Motion*

2009-08 Volker Nannen (VU) *Evolutionary Agent-Based Policy Analysis in Dynamic Environments*

2009-09 Benjamin Kanagwa (RUN) *Design, Discovery and Construction of Service-oriented Systems*

2009-10 Jan Wielemaker (UVA) *Logic programming for knowledge-intensive interactive applications*

2009-11 Alexander Boer (UVA) *Legal Theory, Sources of Law & the Semantic Web*

2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin) *Operating Guidelines for Services*

2009-13 Steven de Jong (UM) *Fairness in Multi-Agent Systems*

2009-14 Maksym Korotkiy (VU) *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)*

2009-15 Rinke Hoekstra (UVA) *Ontology Representation – Design Patterns and Ontologies that Make Sense*

2009-16 Fritz Reul (UvT) *New Architectures in Computer Chess*

2009-17 Laurens van der Maaten (UvT) *Feature Extraction from Visual Data*

2009-18 Fabian Groffen (CWI) *Armada, An Evolving Database System*

2009-19 Valentin Robu (CWI) *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*

2009-20 Bob van der Vecht (UU) *Adjustable Autonomy: Controling Influences on Decision Making*

2009-21 Stijn Vanderlooy (UM) *Ranking and Reliable Classification*

2009-22 Pavel Serdyukov (UT) *Search For Expertise: Going beyond direct evidence*

2009-23 Peter Hofgesang (VU) *Modelling Web Usage in a Changing Environment*

2009-24 Annerieke Heuvelink (VUA) *Cognitive Models for Training Simulations*

2009-25 Alex van Ballegooij (CWI) *RAM: Array Database Management through Relational Mapping*

2009-26 Fernando Koch (UU) *An Agent-Based Model for the Development of Intelligent Mobile Services*

2009-27 Christian Glahn (OU) *Contextual Support of social Engagement and Reflection on the Web*

2009-28 Sander Evers (UT) *Sensor Data Management with Probabilistic Models*

2009-29 Stanislav Pokraev (UT) *Model-Driven Semantic Integration of Service-Oriented Applications*

2009-30 Marcin Zukowski (CWI) *Balancing vectorized query execution with bandwidth-optimized storage*

2009-31 Sofiya Katrenko (UVA) *A Closer Look at Learning Relations from Text*

2009-32 Rik Farenhorst (VU) and Remco de Boer (VU) *Architectural Knowledge Management: Supporting Architects and Auditors*

2009-33 Khiet Truong (UT) *How Does Real Affect Affect Affect Recognition In Speech?*

2009-34 Inge van de Weerd (UU) *Advancing in Software Product Management: An Incremental Method Engineering Approach*

2009-35 Wouter Koelewijn (UL) *Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling*

2009-36 Marco Kalz (OUN) *Placement Support for Learners in Learning Networks*

2009-37 Hendrik Drachsler (OUN) *Navigation Support for Learners in Informal Learning Networks*

2009-38 Riina Vuorikari (OU) *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context*

2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin) *Service Substitution – A Behavioral Approach Based on Petri Nets*

2009-40 Stephan Raaijmakers (UvT) *Multinomial Language Learning: Investigations into the Geometry of Language*

2009-41 Igor Berezhnyy (UvT) *Digital Analysis of Paintings*

2009-42 Toine Bogers (UvT) *Recommender Systems for Social Bookmarking*

2009-43 Virginia Nunes Leal Franqueira (UT) *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients*

2009-44 Roberto Santana Tapia (UT) *Assessing Business-IT Alignment in Networked Organizations*

2009-45 Jilles Vreeken (UU) *Making Pattern Mining Useful*

2009-46 Loredana Afanasiev (UvA) *Querying XML: Benchmarks and Recursion*

2010-01 Matthijs van Leeuwen (UU) *Patterns that Matter*

2010-02 Ingo Wassink (UT) *Work flows in Life Science*

2010-03 Joost Geurts (CWI) *A Document Engineering Model and Processing Framework for Multimedia documents*

2010-04 Olga Kulyk (UT) *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments*

2010-05 Claudia Hauff (UT) *Predicting the Effectiveness of Queries and Retrieval Systems*

2010-06 Sander Bakkes (UvT) *Rapid Adaptation of Video Game AI*

2010-07 Wim Fikkert (UT) *Gesture interaction at a Distance*

2010-08 Krzysztof Siewicz (UL) *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments*

2010-09 Hugo Kielman (UL) *A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging*

2010-10 Rebecca Ong (UL) *Mobile Communication and Protection of Children*

2010-11 Adriaan Ter Mors (TUD) *The world according to MARP: Multi-Agent Route Planning*

2010-12 Susan van den Braak (UU) *Sensemaking software for crime analysis*

2010-13 Gianluigi Folino (RUN) *High Performance Data Mining using Bio-inspired techniques*

2010-14 Sander van Splunter (VU) *Automated Web Service Reconfiguration*

2010-15 Lianne Bodenstaff (UT) *Managing Dependency Relations in Inter-Organizational Models*

2010-16 Sicco Verwer (TUD) *Efficient Identification of Timed Automata, theory and practice*

2010-17 Spyros Kotoulas (VU) *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications*

2010-18 Charlotte Gerritsen (VU) *Caught in the Act: Investigating Crime by Agent-Based Simulation*

2010-19 Henriette Cramer (UvA) *People's Responses to Autonomous and Adaptive Systems*

2010-20 Ivo Swartjes (UT) *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*

2010-21 Harold van Heerde (UT) *Privacy-aware data management by means of data degradation*

2010-22 Michiel Hildebrand (CWI) *End-user Support for Access to Heterogeneous Linked Data*

2010-23 Bas Steunebrink (UU) *The Logical Structure of Emotions*

2010-24 Dmytro Tykhonov (TUD) *Designing Generic and Efficient Negotiation Strategies*

2010-25 Zulfiqar Ali Memon (VU) *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective*

2010-26 Ying Zhang (CWI) *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines*

2010-27 Marten Voulon (UL) *Automatisch contracteren*

2010-28 Arne Koopman (UU) *Characteristic Relational Patterns*

2010-29 Stratos Idreos (CWI) *Database Cracking: Towards Auto-tuning Database Kernels*

2010-30 Marieke van Erp (UvT) *Accessing Natural History – Discoveries in data cleaning, structuring, and retrieval*

2010-31 Victor de Boer (UVA) *Ontology Enrichment from Heterogeneous Sources on the Web*

2010-32 Marcel Hiel (UvT) *An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems*

2010-33 Robin Aly (UT) *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval*

2010-34 Teduh Dirgahayu (UT) *Interaction Design in Service Compositions*

2010-35 Dolf Trieschnigg (UT) *Proof of Concept: Concept-based Biomedical Information Retrieval*

2010-36 Jose Janssen (OU) *Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification*

2010-37 Niels Lohmann (TUE) *Correctness of services and their composition*

2010-38 Dirk Fahland (TUE) *From Scenarios to components*

2010-39 Ghazanfar Farooq Siddiqui (VU) *Integrative modeling of emotions in virtual agents*

2010-40 Mark van Assem (VU) *Converting and Integrating Vocabularies for the Semantic Web*

2010-41 Guillaume Chaslot (UM) *Monte-Carlo Tree Search*

2010-42 Sybren de Kinderen (VU) *Needs-driven service bundling in a multi-supplier setting – the computational e3-service approach*

2010-43 Peter van Kranenburg (UU) *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*

2010-44 Pieter Bellekens (TUE) *An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain*

2010-45 Vasilios Andrikopoulos (UvT) *A theory and model for the evolution of software services*

2010-46 Vincent Pijpers (VU) *e3alignment: Exploring Inter-Organizational Business-ICT Alignment*

2010-47 Chen Li (UT) *Mining Process Model Variants: Challenges, Techniques, Examples*

2010-48 *Withdrawn*

2010-49 Jahn-Takeshi Saito (UM) *Solving difficult game positions*

2010-50 Bouke Huurnink (UVA) *Search in Audiovisual Broadcast Archives*

2010-51 Alia Khairia Amin (CWI) *Understanding and supporting information seeking tasks in multiple sources*

2010-52 Peter-Paul van Maanen (VU) *Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention*

2010-53 Edgar Meij (UVA) *Combining Concepts and Language Models for Information Access*

2011-01 Botond Cseke (RUN) *Variational Algorithms for Bayesian Inference in Latent Gaussian Models*

2011-02 Nick Tinnemeier (UU) *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language*

2011-03 Jan Martijn van der Werf (TUE) *Compositional Design and Verification of Component-Based Information Systems*

2011-04 Hado van Hasselt (UU) *Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms*

2011-05 Base van der Raadt (VU) *Enterprise Architecture Coming of Age – Increasing the Performance of an Emerging Discipline.*

2011-06 Yiwen Wang (TUE) *Semantically-Enhanced Recommendations in Cultural Heritage*

2011-07 Yujia Cao (UT) *Multimodal Information Presentation for High Load Human Computer Interaction*

2011-08 Nieske Vergunst (UU) *BDI-based Generation of Robust Task-Oriented Dialogues*

2011-09 Tim de Jong (OU) *Contextualised Mobile Media for Learning*

2011-10 Bart Bogaert (UvT) *Cloud Content Contention*

2011-11 Dhaval Vyas (UT) *Designing for Awareness: An Experience-focused HCI Perspective*

2011-12 Carmen Bratosin (TUE) *Grid Architecture for Distributed Process Mining*

2011-13 Xiaoyu Mao (UvT) *Airport under Control. Multiagent Scheduling for Airport Ground Handling*

2011-14 Milan Lovric (EUR) *Behavioral Finance and Agent-Based Artificial Markets*

2011-15 Marijn Koolen (UvA) *The Meaning of Structure: the Value of Link Evidence for Information Retrieval*

2011-16 Maarten Schadd (UM) *Selective Search in Games of Different Complexity*

2011-17 Jiyin He (UVA) *Exploring Topic Structure: Coherence, Diversity and Relatedness*

2011-18 Mark Ponsen (UM) *Strategic Decision-Making in complex games*

2011-19 Ellen Rusman (OU) *The Mind's Eye on Personal Profiles*

2011-20 Qing Gu (VU) *Guiding service-oriented software engineering – A view-based approach*

2011-21 Linda Terlouw (TUD) *Modularization and Specification of Service-Oriented Systems*

2011-22 Junte Zhang (UVA) *System Evaluation of Archival Description and Access*

2011-23 Wouter Weerkamp (UVA) *Finding People and their Utterances in Social Media*

2011-24 Herwin van Welbergen (UT) *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*

2011-25 Syed Waqar ul Qounain Jaffry (VU) *Analysis and Validation of Models for Trust Dynamics*

2011-26 Matthijs Aart Pontier (VU) *Virtual Agents for Human Communication – Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots*

2011-27 Aniel Bhulai (VU) *Dynamic website optimization through autonomous management of design patterns*

2011-28 Rianne Kaptein (UVA) *Effective Focused Retrieval by Exploiting Query Context and Document Structure*

2011-29 Faisal Kamiran (TUE) *Discrimination-aware Classification*

2011-30 Egon van den Broek (UT) *Affective Signal Processing (ASP): Unraveling the mystery of emotions*

2011-31 Ludo Waltman (EUR) *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*

2011-32 Nees-Jan van Eck (EUR) *Methodological Advances in Bibliometric Mapping of Science*

2011-33 Tom van der Weide (UU) *Arguing to Motivate Decisions*

2011-34 Paolo Turrini (UU) *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*

2011-35 Maaike Harbers (UU) *Explaining Agent Behavior in Virtual Training*

2011-36 Erik van der Spek (UU) *Experiments in serious game design: a cognitive approach*

2011-37 Adriana Burlutiu (RUN) *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference*

2011-38 Nyree Lemmens (UM) *Bee-inspired Distributed Optimization*

2011-39 Joost Westra (UU) *Organizing Adaptation using Agents in Serious Games*

2011-40 Viktor Clerc (VU) *Architectural Knowledge Management in Global Software Development*

2011-41 Luan Ibraimi (UT) *Cryptographically Enforced Distributed Data Access Control*

2011-42 Michal Sindlar (UU) *Explaining Behavior through Mental State Attribution*

2011-43 Henk van der Schuur (UU) *Process Improvement through Software Operation Knowledge*

2011-44 Boris Reuderink (UT) *Robust Brain-Computer Interfaces*

2011-45 Herman Stehouwer (UvT) *Statistical Language Models for Alternative Sequence Selection*

2011-46 Beibei Hu (TUD) *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work*

2011-47 Azizi Bin Ab Aziz(VU) *Exploring Computational Models for Intelligent Support of Persons with Depression*

2011-48 Mark Ter Maat (UT) *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent*

2011-49 Andreea Niculescu (UT) *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality*

2012-01 Terry Kakeeto (UvT) *Relationship Marketing for SMEs in Uganda*

2012-02 Muhammad Umair (VU) *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models*

2012-03 Adam Vanya (VU) *Supporting Architecture Evolution by Mining Software Repositories*

2012-04 Jurriaan Souer (UU) *Development of Content Management System-based Web Applications*

2012-05 Marijn Plomp (UU) *Maturing Interorganisational Information Systems*

2012-06 Wolfgang Reinhardt (OU) *Awareness Support for Knowledge Workers in Research Networks*

2012-07 Rianne van Lambalgen (VU) *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions*

2012-08 Gerben de Vries (UVA) *Kernel Methods for Vessel Trajectories*

2012-09 Ricardo Neisse (UT) *Trust and Privacy Management Support for Context-Aware Service Platforms*

2012-10 David Smits (TUE) *Towards a Generic Distributed Adaptive Hypermedia Environment*

2012-11 J.C.B. Rantham Prabhakara (TUE) *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics*

2012-12 Kees van der Sluijs (TUE) *Model Driven Design and Data Integration in Semantic Web Information Systems*

2012-13 Suleman Shahid (UvT) *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions*

2012-14 Evgeny Knutov (TUE) *Generic Adaptation Framework for Unifying Adaptive Web-based Systems*

2012-15 Natalie van der Wal (VU) *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.*

2012-16 Fiemke Both (VU) *Helping people by understanding them – Ambient Agents supporting task execution and depression treatment*

2012-17 Amal Elgammal (UvT) *Towards a Comprehensive Framework for Business Process Compliance*

2012-18 Eltjo Poort (VU) *Improving Solution Architecting Practices*

2012-19 Helen Schonenberg (TUE) *What's Next? Operational Support for Business Process Execution*

2012-20 Ali Bahramisharif (RUN) *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing*

2012-21 Roberto Cornacchia (TUD) *Querying Sparse Matrices for Information Retrieval*

2012-22 Thijs Vis (UvT) *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?*

2012-23 Christian Mühl (UT) *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction*

2012-24 Laurens van der Werff (UT) *Evaluation of Noisy Transcripts for Spoken Document Retrieval*

2012-25 Silja Eckartz (UT) *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application*

2012-26 Emile de Maat (UVA) *Making Sense of Legal Text*

2012-27 Hayrettin Gürkök (UT) *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games*

2012-28 Nancy Pascall (UvT) *Engendering Technology Empowering Women*

2012-29 Almer Tigelaar (UT) *Peer-to-Peer Information Retrieval*

2012-30 Alina Pommeranz (TUD) *Designing Human-Centered Systems for Reflective Decision Making*

2012-31 Emily Bagarukayo (RUN) *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure*

2012-32 Wietske Visser (TUD) *Qualitative multi-criteria preference representation and reasoning*

2012-33 Rory Sie (OUN) *Coalitions in Cooperation Networks (COCOON)*

2012-34 Pavol Jancura (RUN) *Evolutionary analysis in PPI networks and applications*

2012-35 Evert Haasdijk (VU) *Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics*

2012-36 Denis Ssebugwawo (RUN) *Analysis and Evaluation of Collaborative Modeling Processes*

2012-37 Agnes Nakakawa (RUN) *A Collaboration Process for Enterprise Architecture Creation*

2012-38 Selmar Smit (VU) *Parameter Tuning and Scientific Testing in Evolutionary Algorithms*

2012-39 Hassan Fatemi (UT) *Risk-aware design of value and coordination networks*

2012-40 Agus Gunawan (UvT) *Information Access for SMEs in Indonesia*

2012-41 Sebastian Kelle (OU) *Game Design Patterns for Learning*

2012-42 Dominique Verpoorten (OU) *Reflection Amplifiers in self-regulated Learning*

2012-43 *Withdrawn*

2012-44 Anna Tordai (VU) *On Combining Alignment Techniques*

2012-45 Benedikt Kratz (UvT) *A Model and Language for Business-aware Transactions*

2012-46 Simon Carter (UVA) *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation*

2012-47 Manos Tsagkias (UVA) *Mining Social Media: Tracking Content and Predicting Behavior*

2012-48 Jorn Bakker (TUE) *Handling Abrupt Changes in Evolving Time-series Data*

2012-49 Michael Kaisers (UM) *Learning against Learning – Evolutionary dynamics of reinforcement learning algorithms in strategic interactions*

2012-50 Steven van Kervel (TUD) *Ontologogy driven Enterprise Information Systems Engineering*

2012-51 Jeroen de Jong (TUD) *Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching*

2013-01 Viorel Milea (EUR) *News Analytics for Financial Decision Support*

2013-02 Erietta Liarou (CWI) *MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing*

2013-03 Szymon Klarman (VU) *Reasoning with Contexts in Description Logics*

2013-04 Chetan Yadati (TUD) *Coordinating autonomous planning and scheduling*

2013-05 Dulce Pumareja (UT) *Groupware Requirements Evolutions Patterns*

2013-06 Romulo Goncalves (CWI) *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience*

2013-07 Giel van Lankveld (UvT) *Quantifying Individual Player Differences*

2013-08 Robbert-Jan Merk (VU) *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators*

2013-09 Fabio Gori (RUN) *Metagenomic Data Analysis: Computational Methods and Applications*

2013-10 Jeewanie Jayasinghe Arachchige (UvT) *A Unified Modeling Framework for Service Design.*

2013-11 Evangelos Pournaras (TUD) *Multi-level Reconfigurable Self-organization in Overlay Services*

2013-12 Marian Razavian (VU) *Knowledge-driven Migration to Services*

2013-13 Mohammad Safiri (UT) *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly*

2013-14 Jafar Tanha (UVA) *Ensemble Approaches to Semi-Supervised Learning Learning*

2013-15 Daniel Hennes (UM) *Multiagent Learning – Dynamic Games and Applications*

2013-16 Eric Kok (UU) *Exploring the practical benefits of argumentation in multi-agent deliberation*

2013-17 Koen Kok (VU) *The PowerMatcher: Smart Coordination for the Smart Electricity Grid*

2013-18 Jeroen Janssens (UvT) *Outlier Selection and One-Class Classification*

2013-19 Renze Steenhuizen (TUD) *Coordinated Multi-Agent Planning and Scheduling*

2013-20 Katja Hofmann (UvA) *Fast and Reliable Online Learning to Rank for Information Retrieval*

2013-21 Sander Wubben (UvT) *Text-to-text generation by monolingual machine translation*

2013-22 Tom Claassen (RUN) *Causal Discovery and Logic*

2013-23 Patricio de Alencar Silva (UvT) *Value Activity Monitoring*

2013-24 Haitham Bou Ammar (UM) *Automated Transfer in Reinforcement Learning*

2013-25 Agnieszka Anna Latoszek-Berendsen (UM) *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System*

2013-26 Alireza Zarghami (UT) *Architectural Support for Dynamic Homecare Service Provisioning*

2013-27 Mohammad Huq (UT) *Inference-based Framework Managing Data Provenance*

2013-28 Frans van der Sluis (UT) *When Complexity becomes Interesting: An Inquiry into the Information eXperience*

2013-29 Iwan de Kok (UT) *Listening Heads*

2013-30 Joyce Nakatumba (TUE) *Resource-Aware Business Process Management: Analysis and Support*

2013-31 Dinh Khoa Nguyen (UvT) *Blueprint Model and Language for Engineering Cloud Applications*

2013-32 Kamakshi Rajagopal (OUN) *Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development*

2013-33 Qi Gao (TUD) *User Modeling and Personalization in the Microblogging Sphere*

2013-34 Kien Tjin-Kam-Jet (UT) *Distributed Deep Web Search*

2013-35 Abdallah El Ali (UvA) *Minimal Mobile Human Computer Interaction*

2013-36 Than Lam Hoang (TUe) *Pattern Mining in Data Streams*

2013-37 Dirk Börner (OUN) *Ambient Learning Displays*

2013-38 Eelco den Heijer (VU) *Autonomous Evolutionary Art*

2013-39 Joop de Jong (TUD) *A Method for Enterprise Ontology based Design of Enterprise Information Systems*

2013-40 Pim Nijssen (UM) *Monte-Carlo Tree Search for Multi-Player Games*

2013-41 Jochem Liem (UVA) *Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning*

2013-42 Léon Planken (TUD) *Algorithms for Simple Temporal Reasoning*

2013-43 Marc Bron (UVA) *Exploration and Contextualization through Interaction and Concepts*

2014-01 Nicola Barile (UU) *Studies in Learning Monotone Models from Data*

2014-02 Fiona Tuliyano (RUN) *Combining System Dynamics with a Domain Modeling Method*

2014-03 Sergio Raul Duarte Torres (UT) *Information Retrieval for Children: Search Behavior and Solutions*

2014-04 Hanna Jochmann-Mannak (UT) *Websites for children: search strategies and interface design – Three studies on children's search performance and evaluation*

2014-05 Jurriaan van Reijsen (UU) *Knowledge Perspectives on Advancing Dynamic Capability*

2014-06 Damian Tamburri (VU) *Supporting Networked Software Development*

2014-07 Arya Adriansyah (TUE) *Aligning Observed and Modeled Behavior*

2014-08 Samur Araujo (TUD) *Data Integration over Distributed and Heterogeneous Data Endpoints*

2014-09 Philip Jackson (UvT) *Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language*

2014-10 Ivan Salvador Razo Zapata (VU) *Service Value Networks*

2014-11 Janneke van der Zwaan (TUD) *An Empathic Virtual Buddy for Social Support*

2014-12 Willem van Willigen (VU) *Look Ma, No Hands: Aspects of Autonomous Vehicle Control*

2014-13 Arlette van Wissen (VU) *Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains*

2014-14 Yangyang Shi (TUD) *Language Models With Meta-information*

2014-15 Natalya Mogles (VU) *Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare*

2014-16 Krystyna Milian (VU) *Supporting trial recruitment and design by automatically interpreting eligibility criteria*

2014-17 Kathrin Dentler (VU) *Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability*

2014-18 Mattijs Ghijsen (VU) *Methods and Models for the Design and Study of Dynamic Agent Organizations*

2014-19 Vincius Ramos (TUE) *Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support*

2014-20 Mena Habib (UT) *Named Entity Extraction and Disambiguation for Informal Text: The Missing Link*

2014-21 Kassidy Clark (TUD) *Negotiation and Monitoring in Open Environments*

2014-22 Marieke Peeters (UU) *Personalized Educational Games – Developing agent-supported scenario-based training*

2014-23 Eleftherios Sidirourgos (UvA/CWI) *Space Efficient Indexes for the Big Data Era*

2014-24 Davide Ceolin (VU) *Trusting Semi-structured Web Data*

2014-25 Martijn Lappenschaar (RUN) *New network models for the analysis of disease interaction*

2014-26 Tim Baarslag (TUD) *What to Bid and When to Stop*

2014-27 Rui Jorge Almeida (EUR) *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*

2014-28 Anna Chmielowiec (VU) *Decentralized k-Clique Matching*

2014-29 Jaap Kabbedijk (UU) *Variability in Multi-Tenant Enterprise Software*

2014-30 Peter de Kock Berenschot (UvT) *Anticipating Criminal Behaviour*

2014-31 Leo van Moergestel (UU) *Agent Technology in Agile Multiparallel Manufacturing and Product Support*

2014-32 Naser Ayat (UVA) *On Entity Resolution in Probabilistic Data*

2014-33 Tesfa Tegegne Asfaw (RUN) *Service Discovery in eHealth*

2014-34 Christina Manteli (VU) *The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems*

2014-35 Joost van Oijen (UU) *Cognitive Agents in Virtual Worlds: A Middleware Design Approach*

2014-36 Joos Buijs (TUE) *Flexible Evolutionary Algorithms for Mining Structured Process Models*

2014-37 Maral Dadvar (UT) *Experts and Machines United Against Cyberbullying*

2014-38 Danny Plass-Oude Bos (UT) *Making brain-computer interfaces better: improving usability through post-processing.*

2014-39 Jasmina Maric (UvT) *Web Communities, Immigration, and Social Capital*

2014-40 Walter Omona (RUN) *A Framework for Knowledge Management Using ICT in Higher Education*

2014-41 Frederic Hogenboom (EUR) *Automated Detection of Financial Events in News Text*

2014-42 Carsten Eijckhof (CWI/TUD) *Contextual Multidimensional Relevance Models*

2014-43 Kevin Vlaanderen (UU) *Supporting Process Improvement using Method Increments*

2014-44 Paulien Meesters (UvT) *Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.*

2014-45 Birgit Schmitz (OUN) *Mobile Games for Learning: A Pattern-Based Approach*

2014-46 Ke Tao (TUD) *Social Web Data Analytics: Relevance, Redundancy, Diversity*

2014-47 Shangsong Liang (UVA) *Fusion and Diversification in Information Retrieval*

2015-01 Niels Netten (UvA) *Machine Learning for Relevance of Information in Crisis Response*

2015-02 Faiza Bukhsh (UvT) *Smart auditing: Innovative Compliance Checking in Customs Controls*

2015-03 Twan van Laarhoven (RUN) *Machine learning for network data*

2015-04 Howard Spoelstra (OUN) *Collaborations in Open Learning Environments*

2015-05 Christoph Bösch (UT) *Cryptographically Enforced Search Pattern Hiding*

2015-06 Farideh Heidari (TUD) *Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes*

2015-07 Maria-Hendrike Peetz (UvA) *Time-Aware Online Reputation Analysis*

2015-08 Jie Jiang (TUD) *Organizational Compliance: An agent-based model for designing and evaluating organizational interactions*

2015-09 Randy Klaassen (UT) *HCI Perspectives on Behavior Change Support Systems*

2015-10 Henry Hermans (OUN) *OpenU: design of an integrated system to support lifelong learning*

2015-11 Yongming Luo (TUE) *Designing algorithms for big graph datasets: A study of computing bisimulation and joins*

2015-12 Julie M. Birkholz (VU) *Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks*

2015-13 Giuseppe Procaccianti (VU) *Energy-Efficient Software*

2015-14 Bart van Straalen (UT) *A cognitive approach to modeling bad news conversations*

2015-15 Klaas Andries de Graaf (VU) *Ontology-based Software Architecture Documentation*

2015-16 Changyun Wei (UT) *Cognitive Coordination for Cooperative Multi-Robot Teamwork*

2015-17 André van Cleeff (UT) *Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs*

2015-18 Holger Pirk (CWI) *Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories*

2015-19 Bernardo Tabuenca (OUN) *Ubiquitous Technology for Lifelong Learners*

2015-20 Loïs Vanhée (UU) *Using Culture and Values to Support Flexible Coordination*

2015-21 Sibren Fetter (OUN) *Using Peer-Support to Expand and Stabilize Online Learning*

015-22 Zhemin Zhu (UT) *Co-occurrence Rate Networks*

2015-23 Luit Gazendam (VU) *Cataloguer Support in Cultural Heritage*

2015-24 Richard Berendsen (UVA) *Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation*

2015-25 Steven Woudenberg (UU) *Bayesian Tools for Early Disease Detection*

2015-26 Alexander Hogenboom (EUR) *Sentiment Analysis of Text Guided by Semantics and Structure*

2015-27 Sándor Héman (CWI) *Updating compressed colomn stores*

2015-28 Janet Bagorogoza (TiU) *Knowledge Management and High Performance: The Uganda financial institutions model for HPO*

2015-29 Hendrik Baier (UM) *Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains*

2015-30 Kiavash Bahreini (OU) *Real-time Multimodal Emotion Recognition in E-Learning*

2015-31 Yakup Koç (TUD) *On the robustness of Power Grids*

2015-32 Jerome Gard (UL) *Corporate Venture Management in SMEs*

2015-33 Frederik Schadd (TUD) *Ontology Mapping with Auxiliary Resources*

2015-34 Victor de Graaf (UT) *Gesocial Recommender Systems*

2015-35 Jungxao Xu (TUD) *Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction*

2016-01 Syed Saiden Abbas (RUN) *Recognition of Shapes by Humans and Machines*

2016-02 Michiel Christiaan Meulendijk (UU) *Optimizing medication reviews through decision support: prescribing a better pill to swallow*

2016-03 Maya Sappelli (RUN) *Knowledge Work in Context: User Centered Knowledge Worker Support*

2016-04 Laurens Rietveld (VU) *Publishing and Consuming Linked Data*

2016-05 Evgeny Sherkhonov (UVA) *Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers*

2016-06 Michel Wilson (TUD) *Robust scheduling in an uncertain environment*

2016-07 Jeroen de Man (VU) *Measuring and modeling negative emotions for virtual training*

2016-08 Matje van de Camp (TiU) *A Link to the Past: Constructing Historical Social Networks from Unstructured Data*

2016-09 Archana Nottamkandath (VU) *Trusting Crowdsourced Information on Cultural Artefacts*

2016-10 George Karafotias (VU) *Parameter Control for Evolutionary Algorithms*

2016-11 Anne Schuth (UVA) *Search Engines that Learn from Their Users*

2016-12 Max Knobbout (UU) *Logics for Modelling and Verifying Normative Multi-Agent Systems*

2016-13 Nana Baah Gyan (VU) *The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach*

2016-14 Ravi Khadka (UU) *Revisiting Legacy Software System Modernization*

2016-15 Steffen Michels (RUN) *Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments*

2016-16 Guangliang Li (UVA) *Socially Intelligent Autonomous Agents that Learn from Human Reward*

2016-17 Berend Weel (VU) *Towards Embodied Evolution of Robot Organisms*

2016-18 Albert Meroño Peñuela (VU) *Refining Statistical Data on the Web*

2016-19 Julia Efremova (Tu/e) *Mining Social Structures from Genealogical Data*

2016-20 Daan Odijk (UVA) *Context & Semantics in News & Web Search*

2016-21 Alejandro Moreno Célleri (UT) *From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground*

2016-22 Grace Lewis (VU) *Software Architecture Strategies for Cyber-Foraging Systems*

2016-23 Fei Cai (UVA) *Query Auto Completion in Information Retrieval*

2016-24 Brend Wanders (UT) *Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach*

2016-25 Julia Kiseleva (TU/e) *Using Contextual Information to Understand Searching and Browsing Behavior*

2016-26 Dilhan Thilakarathne (VU) *In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains*

2016-27 Wen Li (TUD) *Understanding Geo-spatial Information on Social Media*

2016-28 Mingxin Zhang (TUD) *Large-scale Agent-based Social Simulation - A study on epidemic prediction and control*

2016-29 Nicolas Höning (TUD) *Peak reduction in decentralised electricity systems -Markets and prices for flexible planning*

2016-30 Ruud Mattheij (UvT) *The Eyes Have It*

2016-31 Mohammad Khelghati (UT) *Deep web content monitoring*

2016-32 Eelco Vriezekolk (UT) *Assessing Telecommunication Service Availability Risks for Crisis Organisations*

2016-33 Peter Bloem (UVA) *Single Sample Statistics, exercises in learning from just one example*

2016-34 Dennis Schunselaar (TUE) *Configurable Process Trees: Elicitation, Analysis, and Enactment*

2016-35 Zhaochun Ren (UVA) *Monitoring Social Media: Summarization, Classification and Recommendation*

2016-36 Daphne Karreman (UT) *Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies*

2016-37 Giovanni Sileno (UvA) *Aligning Law and Action - a conceptual and computational inquiry*

2016-38 Andrea Minuto (UT) *MATERIALS THAT MATTER - Smart Materials meet Art & Interaction Design*

2016-39 Merijn Bruijnes (UT) *Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect*

2016-40 Christian Detweiler (TUD) *Accounting for Values in Design*

2016-41 Thomas King (TUD) *Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance*

2016-42 Spyros Martzoukos (UVA) *Combinatorial and Compositional Aspects of Bilingual Aligned Corpora*

2016-43 Saskia Koldijk (RUN) *Context-Aware Support for Stress Self-Management: From Theory to Practice*

2016-44 Thibault Sellam (UVA) *Automatic Assistants for Database Exploration*

2016-45 Bram van de Laar (UT) *Experiencing Brain-Computer Interface Control*

2016-46 Jorge Gallego Perez (UT) *Robots to Make you Happy*

2016-47 Christina Weber (UL) *Real-time foresight - Preparedness for dynamic innovation networks*

2016-48 Tanja Buttler (TUD) *Collecting Lessons Learned*

2016-49 Gleb Polevoy (TUD) *Participation and Interaction in Projects. A Game-Theoretic Analysis*

2016-50 Yan Wang (UVT) *The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains*

2017-01 Jan-Jaap Oerlemans (UL) *Investigating Cybercrime*

2017-02 Sjoerd Timmer (UU) *Designing and Understanding Forensic Bayesian Networks using Argumentation*

2017-03 Daniël Harold Telgen (UU) *Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines*

2017-04 Mrunal Gawade (CWI) *MULTI-CORE PARALLELISM IN A COLUMN-STORE*

2017-05 Mahdieh Shadi (UVA) *Collaboration Behavior*

2017-06 Damir Vandic (EUR) *Intelligent Information Systems for Web Product Search*

2017-07 Roel Bertens (UU) *Insight in Information: from Abstract to Anomaly*

2017-08 Rob Konijn (VU) *Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery*

Massive digital datasets, such as social media data, are a promising source to study social and cultural phenomena. They provide the opportunity to study language use and behavior in a variety of social situations on a large scale and often with the availability of detailed contextual information. However, to fully leverage their potential for research in the social sciences and the humanities, new computational approaches are needed.

This dissertation explores computational approaches to text analysis for studying cultural and social phenomena and focuses on two emerging areas: computational sociolinguistics and computational folkloristics. Both areas share the recognition that variation in text is often meaningful and may provide insights into social and cultural phenomena. This dissertation develops computational approaches to analyze and model variation in text.