A link to the past

van de Camp, Matje

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

A Link to the Past

Constructing Historical Social Networks from Unstructured Data

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
Tilburg University op gezag van de rector magnificus,
prof.dr. E.H.L. Aarts, in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie in
de aula van de Universiteit op

woensdag 2 maart 2016 om 14.15 uur

door

Margaretha Maria van de Camp

geboren op 16 februari 1981 te Tilburg

Promotores:	Prof. dr. A.P.J. van den Bosch Prof. dr. E.O. Postma
Promotiecommissie:	Prof. dr. L. Heerma van Voss Prof. dr. H.J. van den Herik Prof. dr. A. Mehler Prof. dr. M.F. Moens

The research reported in this thesis was funded by the Netherlands Organization for Scientific Research (NWO) in the project Historical Timeline Mining and Extraction (HiTiME), grant number NWO 640.004.803. The HiTiME project is part of the Continuous Access To Cultural Heritage (CATCH) research programme.

SIKS Dissertation Series No. 2016-08
The research reported in this thesis was carried out under auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

TiCC Dissertation Series No. 44

ISBN: 978-94-6203-988-9
Printed by CPI Koninklijke Wöhrmann
Cover design by Matje van de Camp

ACKNOWLEDGEMENTS

I would sincerely like to thank all who were involved with, or contributed to the completion of this thesis. There were times when even I did not think that I would make it to the end, but here we are, and what a journey it has been.

First and foremost, my gratitude goes out to my promotors and supervisors, prof. dr. Antal van den Bosch and prof. dr. Eric Postma. Antal, thank you so much for giving me the opportunity to discover my passion and for your guidance, support and patience throughout. I fondly remember our inspiring meetings and lunch walks. Eric, even though your involvement was not from the start of my PhD, your efforts to get me to the finish line were indispensable. Thank you for all the discussions and pep talks.

I would like to thank all the members of the committee for reading my thesis and providing me with their comments: prof. dr. Jaap van den Herik, prof. dr. Alexander Mehler, prof. dr. Marie-Francine Moens, and prof. dr. Lex Heerma van Voss. Jaap, thank you sincerely for all the invaluable advice you gave me in the years when we were both at TiCC. Alexander, thank you for coming over from Germany for my defense. I enjoyed meeting you in Leipzig and look forward to discussing my methods and analyses with you during and after the ceremony. Marie-Francine, thank you for reading my thesis and for coming to Tilburg for the ceremony. I look forward to meeting you. Lex, thank you for the discussions that we had at IISH when you were part of the HiTiME advisory board. They were a direct inspiration for the work presented in this thesis.

Thank you also in this respect to the rest of the HiTiME advisory board: dr. Dennis Bos, dr. Andrea Scharnhorst, dr. Angelie Sens, and project leader dr. Marien van der Heijden. Marien, thank you for continued enthusiasm for the project and for always receiving me with open arms at IISH.

My time at TiCC was a turbulent one for me, personally, but my colleagues there have always made me feel like part of the family, which is something I really needed and will never forget. Martin, I knew we would get along the first time I heard you comment on the spelling mistakes in someone else's presentation. Thank you for being a mentor and a friend and for always believing in me. I look forward to many years of working together on whatever we can come up with. Ko, Menno, Sander and Bart, thank you for the fun evenings we spent at the university's pub quiz. I will think of foof whenever I drink an Erdinger. Menno and Tanja, thank you for hosting those wonderful "winter barbies" where we all drank and sang and played

pinball. And thank you to Iris (Balemans!), Martha, Steve, Herman, Alain, Maarten, Paai, Nanne, Yevgen and Yu for making my workdays so much more enjoyable.

Good friends are hard to come by, but I am happy to realize that I have found some real gems. Anke Dijkstra and Anouk Gaillard, my beautiful paranymphs, I love and admire you both to no end. Anke, thank you for always providing an alternative viewpoint and helping me to put things in perspective, even from half way around the world. You continue to inspire me and I am so proud to be your friend. Anouk, you are without a doubt the most relaxed person that I have ever met. Thank you for your never-ending patience and all the trips, concerts and afternoons just hanging out. I look forward to many more years of that, while watching your beautiful daughter Aurora growing up. Roy, you are my cousin, but really you are a brother from another mother. I am so grateful that you are always there to listen and understand and for all the fun that we have in between. I hope we never lose that. And last, but definitely not least, a big thank you to Tony, Mili, Michiel and Margot for all the conversations, drinks and good times that we shared over the years. I truly cherish your friendship.

Matje van de Camp
Tilburg, 18 January 2016

TABLE OF CONTENTS

ACKNOWLEDGEMENTS V

TABLE OF CONTENTS VII

1 INTRODUCTION 1
- 1.1. RESEARCH MOTIVATION 2
- 1.2. PROBLEM STATEMENT AND RESEARCH QUESTIONS 3
- 1.3. RESEARCH METHODOLOGY 4
- 1.4. THESIS OUTLINE 5

2 SOCIAL HISTORY 7
- 2.1. SOCIAL HISTORY 8
- 2.2. BWSA 8
- 2.3. CHALLENGES IN SOCIAL HISTORICAL RESEARCH 13
  - 2.3.1. *Accessibility* 13
  - 2.3.2. *Efficiency* 14
  - 2.3.3. *Technophobia* 15
- 2.4. RELATED RESEARCH 15

3 WHAT'S IN A NAME? 19
- 3.1. RECOGNITION AND IDENTIFICATION OF NAMED ENTITIES 20
  - 3.1.1. *Ambiguity in names* 20
  - 3.1.2. *Consequences of misidentification* 21
- 3.2. PREVIOUS RESEARCH INTO NER AND NED 22
  - 3.2.1. *Named Entity Recognition* 23
  - 3.2.2. *Named Entity Disambiguation* 24
- 3.3. BWSA-NERD 25
  - 3.3.1. *Named Entity Recognition* 26
  - 3.3.2. *Named Entity Disambiguation* 30
- 3.4. EXPERIMENTS AND RESULTS 33
  - 3.4.1. *Named Entity Recognition* 33
  - 3.4.2. *Within-document disambiguation* 35
  - 3.4.3. *Cross-document disambiguation* 38
- 3.5. DISCUSSION 39
- 3.6. SOCIAL NETWORK MODEL CONSTRUCTION 42

4 MASTERING TIME 47
- 4.1. RELATED RESEARCH 48
  - 4.1.1. *TimeML* 48
  - 4.1.2. *TempEval* 50
  - 4.1.3. *Dutch temporal analysis* 54
- 4.2. METHOD 55
  - 4.2.1. *TIMEX3* 56
  - 4.2.2. *EVENT* 59

4.2.3. TLINK *62*
4.3. RESULTS 66
4.3.1. TIMEX3 ..... *66*
4.3.2. EVENT *69*
4.3.3. TLINK *69*
4.4. DISCUSSION 74

5 **THE SOCIALIST NETWORK** **81**
5.1. SOCIAL NETWORK ANALYSIS 82
5.1.1. Small-world networks *83*
5.1.2. Scale-free networks *84*
5.1.3. Centrality *85*
5.1.4. Dynamic graphs *88*
5.2. STATISTICAL ANALYSIS 90
5.2.1. Small-worldliness *91*
5.2.2. Growth mechanisms *91*
5.2.3. Centrality ranking correlations *93*
5.3. EVENT ANALYSIS 96
5.4. DISCUSSION 97

6 **DISCUSSION AND CONCLUSIONS** **99**
6.1. ANSWERS TO RESEARCH QUESTIONS 99
6.2. ANSWER TO PROBLEM STATEMENT 101
6.3. THESIS CONTRIBUTIONS 102
6.4. FUTURE RESEARCH 103

REFERENCES **105**

LIST OF TABLES **117**

LIST OF FIGURES **119**

APPENDIX A **121**

APPENDIX B **135**

SUMMARY **143**

CURRICULUM VITAE **147**

LIST OF PUBLICATIONS **149**

SIKS DISSERTATION SERIES **151**

TICC DISSERTATION SERIES **165**

1 INTRODUCTION

Social Networking Services such as Facebook, Twitter, Instagram, and Google+, allow us to communicate with people across the globe with great ease. We are able to find like-minded people anywhere in the world and share ideas with them about anything. The networks that arise from these activities are digitally recorded, creating a multitude of data on human interactions. The availability of such structured data has sparked new interest in the fields of Social Network Extraction and Analysis, especially within the computer science domain. However, social networks are not a new phenomenon. In fact, they have always formed the basis of society as we know it, which grows and evolves through our relationships and interactions with one another. As such, Social Network Analysis (SNA) has a long history as a research methodology within the social sciences (Wasserman & Faust, 1994). It has been applied to answer questions related to decision making processes (Bavelas, 1950; Laumann & Pappi, 1973; Laumann, Marsden, & Galaskiewicz, 1977), diffusion of innovations (Coleman, Katz, & Menzel, 1957; Coleman, Katz, & Menzel, 1966; Rogers E. M., 1979), fraud and corporate interlocking, which occurs when corporate board members serve on boards of multiple corporations at the same time (Levine, 1972; Mizruchi & Schwartz, 1992), social support (Gottlieb, 1981; Wellman & Wortley, 1990), and more. One research area where use of SNA is less prolific is Social History. Still, study of the networks of people and organizations underlying historic events or movements could also lead to new insights for social historians. An example of this new insight is found in the work of (Düring, 2015) who investigates covert support networks that existed for persecuted Jews in Germany during World War II. The networks in this study are constructed from varied sources, including autobiographical accounts and Gestapo interrogation reports, in a large manual undertaking.[1] The amount of effort required to process these, often free text and not digitized, sources to a format suitable for network analysis is one of the reasons why little research has been done on the formation and evaluation of longitudinal, historical social networks. Social historians are also reluctant in adopting methods originating from other fields, especially if these methods are automated, because they trust only their own judgment. Still, computer science, specifically Natural Language Processing (NLP),

---

Parts of this Chapter have previously been published in:
– Van de Camp, M., Van den Bosch, A. (2012). The socialist network. *Decision Support Systems*, 53(4), 761-769.

[1] http://programminghistorian.org/lessons/creating-network-diagrams-from-historical-sources

provides tools that can be utilized to delineate indirect traces of real-world interactions from historical, secondary sources and reconstruct the underlying social networks, making SNA a feasible enterprise for any field using textual sources.

This thesis describes methods for extracting social networks that are implicitly recorded in unstructured data, making them explicit and ordering them on a timeline, to facilitate computational analysis. The methods are applied to a social historical dataset of biographical, free text documents. The resulting network is evaluated through visualisations and comparisons to manual annotations on the same dataset, as well as characteristics of social networks as they are reported in related literature.

1.1. Research Motivation

Considering the art of scientific research, there exist as many opinions on what is "good" research as there are researchers. They each have their own beliefs about what is correct, interesting, scientifically valid, and whether or not such terms even apply. For instance, the idea of interdisciplinary research is generally more contested than monodisciplinary research. This especially seems to be the case for researchers in the social sciences when it comes to the applicability and usability of computer science methods to their field. Although they can all bask in the idea of a super computer that finds them every piece of (textual) information related to their research, presenting it in an organized manner that highlights just the interesting bits that they are after, most are convinced that this is merely a futuristic dream. They are more comfortable relying on their own mind's processing capabilities. In some respects they are right in taking this stance. Most of the tools that are currently available for text analysis perform best on relatively simple tasks such as part of speech tagging or stemming. Tools needed for deeper semantic analysis that would reveal the information relevant to social scientists often produce output that is far too noisy for further use in a scientific context.

That said, there is definitely something to be gained from the combination of the social and computer sciences. The research presented in this thesis is motivated by the belief that the application of simple, straightforward NLP methods to collections of textual data can enrich that data in such a manner that it is reusable within the field that it originated from, providing a benefit over limiting oneself to the use of only domain specific methods of data collection, processing, and presentation. Since we respect and understand the hesitance of social researchers to trust the validity of automatically annotated or generated data, we are motivated to illustrate the advantage of the use of these methods by applying them to an existing collection of texts, and displaying the enriched results in targeted visualizations designed to complement various parts of the research process. Moreover, we aim to show that the specific combination of methods chosen for this task is innovative

and useful for a variety of tasks, providing a powerful new set of tools for textual analysis that could enhance research across multiple fields.

1.2. Problem Statement and Research Questions

Overall, social historians agree that a tool that automatically gathers all the "facts" for them would facilitate their research. We say "facts", since the word itself can stir up quite a dispute in these circles. The nature of the data used in historically inspired research is almost always such that the information contained in it is multi-interpretable. History deals with other people's accounts of what happened, and in the case of secondary sources, to other people. It may seem absurd to try to extract hard facts from such soft data, let alone use them for any meaningful form of analysis. The task of turning unstructured text into structured, quantifiable data is indeed a precarious one. However, it may be assumed that there is at least some consistency across sources, and near to complete consistency within a single source.

The general assumption made in the field of NLP is that, if a dataset is of considerable size and consistent in its contents as well as its style and use of language, recurring patterns will form that can be detected, and sometimes replicated or even predicted, uncovering the underlying structure and meaning. The obvious limitation of this kind of method is the fact that it can only uncover things that are represented in the data enough to form a detectable pattern. Under this assumption, we argue that considering a data source in isolation from its historical real-world context, and processing it using techniques developed for more straightforward tasks and resources, can aid in extracting at least the most obvious facts that hold true within the context of the source under consideration. Furthermore, we postulate that visualization and further computational analysis of the extracted information can be of added value to (social historical) research, if not by inspiring spontaneous findings, then by saving time otherwise spent searching, annotating, or fact checking.

The aim of this thesis is to decrease the reluctance of social historians – and social scientists in general – to use automated methods in their research by proving the suitability of state-of-the-art NLP methods for tasks related to their domain. We focus on social networks as a research tool for Social History and design a method that will simultaneously improve source accessibility and efficiency of information processing for this domain. The main issue to be addressed is summarized by the following problem statement.

> **PS** Can computational methods be used to successfully extract a detailed social network from historical, textual data, enriching the data in such a way that is of added value to social historical research?

The dataset that we use for our experiments consists of biographies written in modern Dutch. We identify two main tasks with regards to the extraction of social network data from our dataset, namely the recognition of named entities, and the

recognition of temporal expressions. This leads us to formulate the following research questions.

RQ 1 To what degree can we reliably recognize and identify named entities in Dutch biographical text using state-of-the-art techniques?

RQ 2 To what degree and level of specificity can we reliably recognize and normalize temporal information in Dutch biographical text using state-of-the-art techniques?

To validate the outcome of our extraction process, we determine some basic characteristics of social network models from literature regarding Social Network Analysis and investigate whether our model shows similar characteristics. To this end we formulate a third research question.

RQ 3 Do social network models constructed with the described method adhere to properties commonly observed in social networks?

To answer our problem statement and research questions, we develop and adapt methods resulting in the following thesis contributions:

1. Evaluation of the current state-of-the-art for Named Entity Recognition on Dutch biographical text;
2. A robust, competitive method for Dutch Named Entity Disambiguation;
3. An accurate method for Dutch Temporal Expression Recognition and Normalization;
4. A method for constructing accurate social network models from unstructured text.
5. Evidence that automated methods, even at the most basic level, can aid in the exertion of Social Historical research.

1.3. Research Methodology

The research methodology used in this thesis consists of five parts: (1) reviewing relevant literature, (2) analysing the findings, (3) selecting the most robust and straightforward methods for each task, (4) adapting and combining the found techniques to test their applicability within the social history domain, and (5) evaluating the results both quantitatively and qualitatively.

First, we conduct a literature study to catalogue the main methods, aspects, and pitfalls within the fields of social network extraction, social network analysis, and relation extraction within the context of both social history and computer science. Literature specifically related to subparts of the research will be reviewed when applicable. Next, we analyse our findings to determine what techniques and methods are most suitable with regards to our purpose and dataset. We then use

these techniques and methods to design a unique set of tools capable of extracting the networks and relations underlying collections of unstructured text, and recombining them in an insightful way.

We test the developed tools by applying them on a collection of historical, secondary sources that describe the actions and whereabouts of a group of several hundreds of people. A quantitative evaluation of the results against manually annotated data is performed at intermediate stages. The specific evaluation metrics used are explained in their respective chapters. Finally, the entire process is evaluated qualitatively by comparing the results to those generally obtained on social networks in related research.

1.4. Thesis Outline

Chapters 1 and 2 form the introduction to this thesis. The motivation, research questions, and methodology are given in Chapter 1. Chapter 2 introduces the field of Social History by providing a description of the chosen dataset and highlighting some of the challenges faced by researchers in the domain. We include a review of related literature in the fields of social network extraction, social network analysis and relation extraction. In Chapter 3 we detail our approach to the recognition and identification of entity names (RQ 1), followed by an explanation of our method for temporal normalisation in Chapter 4 (RQ 2). In Chapter 5 we investigate the validity of the enriched data by interpreting it as a dynamic social network structure and testing this structure for phenomena common to social networks (RQ 3). Finally, we answer our problem statement and conclude the thesis with recommendations for future research in Chapter 6.

2 SOCIAL HISTORY

Social history might be defined negatively as the history of a people with the politics left out.
— George Macaulay Trevelyan (1942)

Social History deals with the thoughts and actions of the common man, and their effects on the historical development of our society. It studies, for instance, under which circumstances certain ideas arise, how they spread through a community, and how they are ultimately combined and transformed into an ideology. These processes are all grounded in interactions between human beings. Networks provide a convenient way to model and study such interactions between entities in general, whether they are people, countries, computers, or even proteins in the human body. Consequently, network analysis has since long been recognized as a valuable asset to many fields in the social, natural, and computer sciences. When the concept of the *social* network first arose in the early 20th century, all annotation and analysis had to be done by hand. Under this constraint, the type of longitudinal data that would most benefit social historians is costly to acquire and therefore not many have ventured into this avenue. In recent years, however, the growing availability of powerful computers and the involvement of computer scientists in the field of social network analysis have opened doors to finally make such large-scale endeavors feasible. Our approach combines methods from the field of Natural Language Processing, or more specifically, Text Mining, into a processing pipeline that extracts those elements from unstructured documents that are needed to construct the social networks underlying the data.

We introduce the Social History domain in Section 2.1, followed by a description of our dataset in Section 2.2. Section 2.3 describes challenges in social historical research that we have identified and aim to mitigate with our approach. We

Parts of this Chapter have previously been published in:
– Van de Camp, M., Van den Bosch, A. (2011). A link to the past: constructing historical social networks. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 61- 69). Portland, Oregon, United States: ACL.
– Van de Camp, M., Van den Bosch, A. (2012). The socialist network. *Decision Support Systems*, 53(4), 761-769.
– Van de Camp, M., & Christiansen, H. (2013). Resolving relative time expressions in Dutch text with Constraint Handling Rules. In D. Duchier, & Y. Parmentier, Constraint Solving and Language Processing (pp. 166-177). Orléans, France: Springer.

conclude the Chapter with an overview of related research, both from the social history domain and the computer science domain, in Section 2.4.

2.1. Social History

In its current form, Social History has existed as a research area since the early 1960s and continues to be a dominant branch of historical research overall. Subfields of social history focus on specific subparts of the population, dividing people by gender, ethnicity, location, social status, profession etc. Labor history, for example, pertains to matters related to the working class, including everything from their personal well being to their organization into worker's unions, political parties, and protest movements. One of the more dominant branches of social history is demographic history, which concerns research into population history based on statistical data, such as population registries and census data.

The analytical methodology of social history most resembles that of the social sciences, approaching matters both from the collective and the individual perspective. Methods used can similarly vary along the spectrum from *quantitative*, where phenomena are examined using logical, quantifiable data and statistical analysis, to *qualitative*, where research relies more on contextually subjective observations. It is worth noting that, even in cases where research is performed purely on statistical data, the interpretation of results is never completely objective. The perspective of the interpreter always influences the interpretation. Therefore social historians tend to prefer the use of primary sources, such as letters, diaries, autobiographies, interviews, newspapers, census data, and even artwork, such as drawings, novels, and plays, to get a full understanding of their chosen research subject. When such sources are not readily available, they fall back to secondary sources, which generally include databases created post-hoc, textbooks, and biographies.

The International Institute of Social History [2] (IISH) in Amsterdam, the Netherlands, hosts an extensive archive of data on global social history and serves as a meeting place for social historians from around the world. On account of its own Dutch origins, IISH's archive includes, among others, a large collection of both primary and secondary sources regarding the worker's movement in the Netherlands. They have graciously provided us access to parts of the collection for the current research. One of the secondary sources included in it is the Biographical Dictionary of Socialism and the Worker's Movement in the Netherlands (henceforth BWSA), which we will use as input for our method.

2.2. BWSA

[2] http://www.socialhistory.org

A biography can be seen as a summary of the most important events in a person's life. It mentions the most relevant people and organizations that the person interacts with, often in a chronological order. It allows us to follow the path that someone walked and to see things more or less from their perspective, although always combined with the interpretation of the author. The BWSA is a collection of 573[3] biographies that describe a relatively coherent group of politicians, artists, thinkers, and the like, who were paramount to the rise of socialism in the Netherlands. Their lives span a period from 1778 to 1998. The biographees interacted with each other both professionally and personally, all of which is summarized in the texts. People from all walks of life and all facets of the political spectrum (as it existed at the time) are represented in the collection, so it provides a very broad view on the Dutch Socialist movement.

The biographies were written by over 200 different authors, which under normal circumstances would result in a high variety of styles and vocabularies. However, the BWSA is under constant review of an editorial board consisting of domain experts. They ensure that every biography in the collection adheres to the same format and that none of the documents contradict any other document content-wise, which makes the BWSA a consistent and high quality source. At the surface, this is directly evident from the structure that is imposed on the biographies, which can be summarized as follows. The introductory paragraph starts with the name of the biographee in the format "LAST NAME, First Names", followed by a short description of their significance to the Socialist movement and their dates and places of birth and death (Figure 2.1). Next, their parents are named, followed by a list of spouses, if any, in chronological order. If the biographee has any known aliases, these are listed at the end of the introduction. Parents and spouses generally do not reoccur unless they also have a biography in the collection. The biographee largely gets referenced by his or her last name, which is sometimes preceded by their initials to clarify their identity. Consequently, most of the person names mentioned in the introduction do not occur anywhere else in the BWSA. The remainder of the biography reports further details relevant to the domain and subject in chronological order from the biographee's childhood through their professional life up to their death and legacy. It may contain as many paragraphs as are needed to complete the story. The length of the biographies therefore tends to vary: the shortest text has 308 tokens, while the longest has 7,188 tokens. The mean length is 1,546 tokens with a standard deviation of 784.

We can quantitatively validate the consistency of a document collection, or *corpus*, by considering its vocabulary growth rate. Figure 2.2 shows the vocabulary growth curve for the entire BWSA, which consists of 888,190 tokens (words and punctuation markers), 49,523 types (uniquely occurring tokens), and 39,433 lemmas (the "dictionary" word forms underlying the word types). The graph shows

[3] All numbers and statistics are based on the BWSA as it was donated to the project in September 2009. New documents have since been added to the original collection.

NIEUWENHUIS, Ferdinand

bekend als Ferdinand Domela Nieuwenhuis, pionier van het socialisme, stichter van het blad *Recht voor Allen*, later sociaal-anarchist, is geboren te Amsterdam op 31 december 1846 en overleden te Hilversum op 18 november 1919. Hij was de zoon van Ferdinand Jacobus Domela Nieuwenhuis, luthers predikant en hoogleraar theologie, en Henriette Frances Berry. Op 24 maart 1870 trad hij in het huwelijk met Johanna Lulofs, met wie hij twee zoons kreeg. Na haar overlijden op 26 maart 1872 hertrouwde hij op 29 oktober 1874 met Johanna Adriana Verhagen, met wie hij twee dochters kreeg. Na haar overlijden op 1 augustus 1877 hertrouwde hij op 21 april 1880 met Johanna Frederika Schingen Hagen, met wie hij een zoon kreeg. Na haar overlijden op 27 februari 1884 hertrouwde hij op 27 mei 1891 met Johanna Egberta Godthelp, met wie hij een dochter en twee zoons kreeg. Bij Koninklijk Besluit van 10 juli 1859 werd de naam door toevoeging gewijzigd in Domela Nieuwenhuis.
Pseudoniemen: Criticus, Ex-Theoloog, Germanus, Philalethes, Dr. Sagittarius.

Figure 2.1 – Introduction of the BWSA biography of Ferdinand Domela Nieuwenhuis. The parts marked in green are included as fields in the BWSA database. Parts marked in blue and orange hold useful information regarding Domela's life (parents, spouses, dates of weddings etc.) and can be extracted using named entity extraction and temporal expression analysis, respectively.

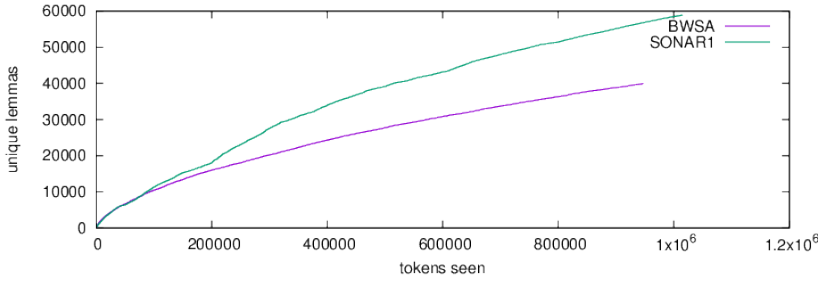


Figure 2.2 – Vocabulary growth curve for the BWSA compared to the SoNaR-1 reference corpus of contemporary Dutch

the cumulative number of unique lemmas as a function of the number of words seen when considering all documents sequentially. Curves such as these generally show a rapid increase at the start after which the growth rate quickly decreases and then seemingly stabilizes (Heaps, 1978). The growth rate will keep decreasing slowly over time, but will never reach zero. In other words, a corpus will never contain the

Figure 2.3 – Alphabetical index of biographies on the BWSA website, available at http://socialhistory.org/bwsa/bios. The middle column lists all available biographies with a short description about the biographee. The right column lists people that were born or passed away on the current day of the year, followed by recent edits made to the collection.

First name	Ferdinand
Last name	Nieuwenhuis
Year of birth	1846
Date of birth	31-12
Year of death	1919
Date of death	18-11
Extra info	(bekend als Ferdinand Domela Nieuwenhuis) pionier van het socialisme, stichter van het blad Recht voor Allen, later sociaal-anarchist

Figure 2.4 – Example record from the BWSA database showing the information for Ferdinand Domela Nieuwenhuis. The record overlaps with the green parts in Figure 2.1. Non-informative fields pertaining to database configuration are not displayed.

entire vocabulary of a language. This phenomenon is easily explained for Dutch corpora, since the language allows for use of closed compounds of arbitrary length[4], meaning that nouns may be compounded into longer words to create new words.

For comparison, we include the vocabulary growth for SoNaR-1[5], which is a million-word reference corpus of contemporary Dutch. When considering both curves, the growth curve of BWSA seems very smooth whereas the SoNaR-1 curve shows many clearly visible bumps. SoNaR-1 contains documents from various genres describing
a multitude of topics. When the text transitions into a new genre or topic, many new words are introduced and the vocabulary growth rate temporarily increases, which is what we see in the graph. The BWSA, on the other hand, is very specific in its domain and storyline, which leads to a smaller and more homogeneous vocabulary. We do, however, notice the same effect in the BWSA growth curve on closer inspection, but at a much smaller scale: the introductory paragraph of each biography also leads to a temporary increase in the growth rate, due to the aggregation of unique names.

IISH hosts the BWSA online.[6] The documents are presented as separate pages, accessible either through an alphabetized index of person names (Figure 2.3) or via keyword search using Boolean operators (AND, OR, NOT). A query result links to the full article in which the search terms are highlighted. Links between documents are minimal: only the first mention of a person who also has a biography in the collection, links to that biography; succeeding mentions and non-BWSA entities are not linked.[7] The links were added manually; doing this for all mentions would definitely be an arduous task. The biographies are accompanied by a database that holds such metadata as a person's full name, dates of birth and death, and a short description of the role they played within the worker's movement (Figure 2.4). These details are used to generate the index on the website. When we compare the database records to the introductory paragraphs (Figure 2.1 and Figure 2.4), we notice that there are many easily detectible pieces of information in the introduction that are missing from the database. Moreover, these pieces and the techniques used to extract them form the exact basis of our method of Social Network Extraction. This simple comparison already demonstrates the potential of straightforward text processing techniques for the social (historical) sciences, since their application to the BWSA will directly allow us to expand the current database, which otherwise would cost many man-hours.

[4] A popular example of a long Dutch compound is the 53-letter word *Kindercarnavalsoptocht-voorbereidingswerkzaamhedenplan*. It translates to: a plan for preparation activities for a children's carnival procession.
[5] http://tst-centrale.org/producten/corpora/sonar-corpus/6-85 (last accessed on July 27, 2015)
[6] http://www.socialhistory.org/bwsa/
[7] The online version of the BWSA has been updated since the start of this thesis, making use of some of the outcomes of the current research, and now contains more linked names per biography.

2.3. Challenges in social historical research

The existence of highly curated collections such as the BWSA is a necessary precondition for successful qualitative research, especially in domains where the interpretation of the data is highly subjective, which is definitely the case for Social History. Without such sources it becomes increasingly difficult to define a common ground and make useful comparisons between different research efforts. However, the mere availability of sources does not provide any guarantees. In fact, we recognize several factors that still inhibit the efficacy of social historical research. We highlight three of such factors that we aim to alleviate with our approach: source accessibility, efficiency of examination, and the reluctance of social historical researchers to us automated methods.

2.3.1. Accessibility

Since it is a secondary, modern source, the BWSA has the benefit of being available in digital form. However, large portions of the data concerning social history remain locked away in paper documents. In recent years, IISH has put much effort into the digitization of its paper archive in order to increase its accessibility and sustainability. Despite these efforts, a number of challenges continue to hinder the archive's accessibility.

The main technique used by IISH to digitize the archive is called Optical Character Recognition (OCR). It entails the conversion of scanned images of typed, printed, or even handwritten text into data that is interpretable by a computer. Since this process is not flawless, it introduces noise into the data, for instance in the form of spelling errors or unknown characters. At the time of writing, most of the interfaces that IISH provides to its archival data allow only straightforward queries that rely on exact string matching. OCR errors greatly inhibit the success rate of these types of searches, since relevant documents containing spelling variations of the query term are not included in the results. Researchers accessing the institute's archive may therefore never find documents that are crucial to the research questions.

Accessibility is further hindered by the fact that different collections within the archive still exist largely disconnected from one another. This is partly due to metadata standards for this field being plentiful and difficult to consolidate. In some cases items are manually classified by attaching descriptive labels from a predefined, ordered set before being added to the archive, which allows for some structured querying over a partially integrated section of the total archive. Intelligent queries across larger subsets would aid researchers in quickly retrieving *all* documents related to their search, and may even increase serendipity by returning results not known to, or not expected by the researcher. However, these queries are as of yet largely impossible. Moreover, the task of assigning labels to documents falls on the shoulders of only a handful of people and, thus, the resulting classification is in all likelihood skewed towards their particular

vocabulary and interpretation. Researchers from other institutions, especially institutions in other countries, might use different terms for the same concepts and, consequently, experience difficulties in locating the data that they are after.

The Social Network Extraction process inevitably includes a normalization phase, where different surface forms of the same name are aggregated and replaced with a uniquely identifiable label (e.g. "John Smith", "J. Smith", and "Mr. Smith" could be replaced with JOHN_SMITH). This operation can theoretically be performed on all of the digitized, textual sources in the archive. On the one hand this facilitates the translation of different search terms to the correct target, and on the other hand it ensures that all documents referencing the same entity are returned for a query.

2.3.2. Efficiency

An in-depth study on any topic in social history currently requires meticulously poring over numerous documents and manually tracking and connecting all the elements in play. As a result social historical projects usually run very long and many historians spend their entire career focused on a single topic. Studies tend to target very specific aspects of broader subjects so as to filter the input data and reduce the time needed to analyze all sources. However, too much specificity and filtering decreases the chance of spontaneous discovery, which is the true catalyst of any scientific domain. It also plays into the hand of a well-known problem that occurs in comparative statistical research, namely Galton's problem (Naroll, 1961; Naroll, 1965).

Galton's problem refers to the statistical phenomenon of *autocorrelation*. In essence it describes the problem of making statistical inferences about a population when the elements in the sample are statistically dependent on a variable that is not accounted for by the model. For example, if two people from the same household are quizzed about the brand of cereal that they eat, their answers are likely to be dependent with respect to the fact that they live together. Therefore, the effectiveness of their answers to the research question "What brand of cereal do people prefer?" is lower than the answers of two individuals from different households would be. As the number of these external dependencies within a dataset increases, the statistical significance of inferences made over the data decreases (Murdock & White, 1969).

Methods developed within the computer science domain allow for the processing of considerably more data in a shorter timespan, which in turn allows the researcher to cast a wider net. More data implies a larger sample. Theoretically, a larger sample size decreases the negative effect of external dependencies on the significance of inferences. More importantly, automation facilitates the creation of persistent, labeled connections within, but also across datasets. With respect to the BWSA and the entities described therein, these connections are in a sense a direct representation of the story behind the data. By considering them as an

interconnected whole, we can break free from the ordered, linear view imposed by the (heavily edited) text. Visualization of the connections in matrices and graphs can further aid in identifying all dependencies between elements before any statistical model is applied to the data. Our method is specifically designed to uncover such relations. By increasing the data throughput and explicating all connections, it will serve to alleviate Galton's problem while simultaneously allowing the social historian to widen his scope.

2.3.3. Technophobia

Although social historians, and researchers in the social sciences in general, do recognize that methods for automated processing provide advantages, at the very least in terms of time saved, they remain reluctant still to actually incorporate them into their normal workflow. Some do not feel confident that they themselves have the skills to use these tools, while others simply do not trust analyses made by a machine. Computers take matters at face value, while the human mind can recognize multiple dimensions that are not visible on the surface. Social historians are mostly interested in these dimensions, as they capture the actual human experience. In these circles, a mere mention of the word "fact" can stir up quite a conversation. When human experience is involved, it quickly becomes difficult, if not impossible, to speak of *hard data*, since everything is subject to interpretation.

We acknowledge the validity of social historians' concerns regarding the accuracy of automated analysis versus manual analysis. Computers will most likely never reach the same level of accuracy as the human brain when it comes to the interpretation of qualitative data. However, we do not consider this a reason to discard them altogether. The type of quantitative analysis and manipulation of huge series of symbols that is needed for text analysis is much faster when automated, even if the results are not 100% accurate. We argue that the correction of the output will take less time than a full manual analysis of the same type and thus that the automated approach will provide an increase in efficiency. Furthermore, automated processes are easily repeated and can be adapted if necessary. This allows the researcher to conveniently view a single dataset from multiple perspectives, or to compare different datasets in the same light.

As argued in this section, automation may solve some major issues plaguing social history as a research domain. Its application does not negate the validity or necessity of manual classification and analysis, as some researchers may fear. Instead they should be seen as a powerful tool to supplement and jumpstart the investigative process.

2.4. Related research

The use of Social Network Analysis as a tool for historical research has been demonstrated by various efforts. For instance, Barkey & Van Rossum (1997)

analyse social unrest among peasants in 17th-century Ottoman villages through a network approach and find that contention, or dissidence, is highest in intermediate villages located between the most central and most isolated villages. Fulminante (2012) apply the same methodology to the locations of early settlements in central Italy, connecting them based on their mutual access to rivers and roads as a way to model the flow of communication and goods through the area. They use the models to study how the geographic location and relative positioning of settlements influence the formation of complex societies. De Benedictis & Tajoli (2011) show the use of the network approach from an economic standpoint in their study of global trade relations and how they evolve over time. Sairio (2008) focus on first-order personal relations within an elite social circle with scholarly ambitions in 18th-century England. The connections between actors in this study are gathered from letters, biographies, and contemporary sources examining the same group of people. These are only a few examples of the many applications of SNA in a historical context. The datasets used in these studies were all constructed manually by the researchers involved, which means that much time was spent on data gathering. This time could have been spent on data analysis, and perhaps have lead to more substantial results, if the data had somehow been harvested automatically.

Text Mining is a research field within the computer science domain that studies methods to convert unstructured, textual data into structured information that is reusable for further analysis. As such, it provides, at least in essence, exactly the tools needed to automate network construction from textual sources. Our research combines methods from several sub fields of Text Mining. The broadest of the fields is Relation Extraction, or more specifically, Social Network Extraction, which we review here. More focussed areas and associated methods are reviewed in their applicable chapters. Relation Extraction is applied in contexts where connections between elements of arbitrary types hold meaning with regards to the overall structure that they form as a whole. Social Network Extraction is the application of Relation Extraction to interactions between human entities. Most of the research in this area focuses on the largest data source that is openly available: the Internet.

A widely used method for determining the relatedness of two entities was first introduced by Kautz, Selman, & Shah (1997) in the context of expert recommendation. They compute the relatedness of two entities by retrieving the number of results returned by a search engine to queries containing one or both names. They then divide the number of web pages mentioning both entities by the number of web pages mentioning either of the entities (Jaccard coefficient). If this measure reaches a certain threshold, the entities are considered to be related to one another. When multiple entities share the same name, keywords may be added to the queries to filter out irrelevant results. The process of determining to which entity an occurrence of a name actually belongs is commonly referred to as Named Entity Disambiguation and will be reviewed in Chapter 3.

Matsuo & Ishizuka (2004) apply Kautz, Selman, & Shah's method to find connections between members of a closed community of researchers. They gather person names from conference attendance lists to create the nodes of the network. The organizational affiliations of each person are added to the queries as a crude form of Named Entity Disambiguation. When a connection is found, the nature of the relation is determined by classifying the contents of websites where both entities are mentioned, based on the occurrence of several manually selected keywords. For instance, occurrence of the words "publication" or "presentation" is considered signs of a co-author relation, whereas "conference", "symposium", or "meeting" implies a conference attendance relation.

A more elaborate approach to network mining is taken by Mika (2005) in his presentation of the *Flink* system. In addition to web co-occurrence counts of person names, the system uses data mined from other — highly structured — sources such as email headers, publication archives, and so-called Friend-Of-A-Friend (FOAF) profiles, which are profiles describing people using a particular machine-readable ontology (Brickley & Miller, 2012). Co-occurrence counts of a name and different personal interests taken from a predefined set are used to determine a person's expertise and to enrich their profile. These profiles are then used to disambiguate the named entities and to find new connections.

Even though search engine counts have become a popular measure of entity relatedness, the counts are not always reliable (Bollegala, Matsuo, & Ishizuka, 2007; Manning, Raghavan, & Schütze, 2008). Higher hit counts oftentimes are mere approximations of the actual web page count. Furthermore, due to indexing, the ever-growing size of the World Wide Web, and the distribution of servers over different locations, the result counts for a given query may vary over time. Bollegala, Matsuo, & Ishizuka (2007) propose to solve this by changing the queries to those that yield fewer results than the approximation threshold. They do so by adding an auxiliary term to the query, which is distributed uniformly throughout the web and independent of the original query term. The result count for the original term is estimated by dividing the result count of the query with the auxiliary term over the probability of the auxiliary term occurring on a web page.

Outside the contexts of the web or scientific research, Social Network Extraction and Network Analysis have also proven to be useful assets to Linguistics and Literary studies. Texts themselves can be interpreted as graphs where words or sentences are connected for instance through co-occurrence or syntactic dependency relations. Syntactic dependency graphs are shown to be scale-free, self-organizing structures, as opposed to randomly formed networks (Masucci & Rodgers, 2006; i Cancho, Mehler, Pustylnikov, & Díaz-Guilera, 2007). A comprehensive review of research into linguistic networks is found in Mehler (2008). In a literary context, Elson, Dames, & McKeown (2010) reconstruct the social networks of characters in classic novels, such as Jane Austen's *Mansfield Park*, by searching the text for quoted speech, which occurs when characters are talking to themselves or each other in the first person. Character names

surrounding the quoted speech are automatically labelled as either the speaking, or spoken to party using a machine learning approach. The number of words spoken determines the weight of a connection from one character to another. An effort that goes beyond this and looks at actual relationship typing is described by Karsdorp, Kestemont, Schöch, & Van den Bosch (2015), who try to detect which characters are romantically involved with one another in French dramatic plays. They approach the problem as a ranking task where, given a character, the system returns a list of other characters in the same play ranked by the likelihood that they are lovers.

More general efforts into Relation Extraction from text have focused on finding recurring patterns and transforming them into triples (RDF). Relation types and labels are then deduced from the most commonly occurring patterns (Ravichandran & Hovy, 2002; Culotta, McCallum, & Betz, 2006). These approaches work well for the induction and verification of straightforwardly verbalized factoids, but they are restricted in their ability to capture much of the subtlety that comes with human interaction. For instance, the details of a romantic relationship are rarely described by one, or even a few stated facts, such as A-LOVES-B, A-IS_MARRIED_TO-B or A-HAS_CHILDREN_ WITH-B.

3 WHAT'S IN A NAME?

It wasn't me.
— Shaggy (2001)

The first step in the creation of a social network model (a *graph*) is the identification of the acting entities. These entities form the agents, or *nodes*, in the graph. Nodes in social networks traditionally represent people, or groups of people, such as families (Padgett & Ansell, 1993), clubs (Galaskiewicz, 1989), or even countries (Wasserman & Faust, 1994). The purpose of the BWSA as a biographical dictionary is to describe how socialist ideas have historically spread through a relatively closed community within the Netherlands. In this scenario, the nodes are logically formed by the biographees. Outside of this initial set, there are also plenty of references to other people, as well as organizations and locations, throughout the data. To fulfil our goal of creating an accurate and complete model of the social network underlying the BWSA, we need to recognize all mentions of all of these entities throughout the text. Moreover, we need to connect each mention to the correct real-world entity. In Natural Language Processing, these tasks are respectively referred to as Named Entity Recognition (NER) and Named Entity Disambiguation (NED).

In this chapter we describe our approach to NER and NED for Dutch biographical texts. Most of the research into these problems is centered on newspaper data because of its abundant availability. In Natural Language Processing, however, good performance on one genre generally does not guarantee good performance on another. We aim to investigate the performance of methods for NER and NED developed on newspaper data when they are applied to biographical data. To increase our chance of success, we restrict ourselves to the use of only proven methods for the recognition, classification, and identification of named entities. We provide a side-by-side comparison of performance on both genres through a series of experiments that serve to answer our first research question:

RQ 1 To what degree can we reliably recognize and identify named entities in Dutch biographical text using state-of-the-art techniques?

We introduce the overall topic of named entities in Section 3.1 and explain what types of ambiguity exist in names and their effects on the identification process. We

provide a review of state-of-the-art research into NER and NED for unstructured text in Section 3.2 and relate it to our goal and dataset. We then select the approach that best fits the BWSA and motivate our choice, followed by a detailed description of our experiments in Section 3.3. In Section 3.4 we present our experiments and the results obtained on the BWSA, and compare these against results obtained on similar datasets. Finally, we end the Chapter in Section 3.5 with a discussion of the results against the backdrop of our goal of constructing a social network.

3.1. Recognition and identification of named entities

In data mining, the term *named entity* is formally used for phrases that refer to one distinct item from a clearly defined set (Grishman & Sundheim, 1996a). For example, named entities may refer to people, organizations, locations, publications, named events, phone numbers, etc. The ability to recognize and identify named entities in text is of fundamental importance to a myriad of tasks within NLP, including, but not limited to, summarization, topic detection and tracking, information retrieval, and relation extraction.

Named entities come in different forms. For instance, a single person may be referred to by his or her first name, surname, both first and surname, or any other format that is valid within the language in question. Besides named entities, he or she might also be referred to by a title or position, a nickname, or simply by a pronoun. The members of the community described by the BWSA are sometimes related or married to one another and therefore they may have similar names. Keeping in mind our goal of creating an accurate model of their combined social network, we must take care to identify each name mentioned in the BWSA and to connect them to the right person.

3.1.1. Ambiguity in names

Names generally carry two types of ambiguity (Wan, Gao, Li, & Ding, 2005): *multi-referent* ambiguity, which exists when one name refers to multiple distinct entities; and *multi-morphic* ambiguity, which exists when one entity is referred to with different names. Multi-referent ambiguity affects precision, while multi-morphic ambiguity affects recall. Both forms of ambiguity occur in the BWSA.

Multi-referent ambiguity

Within the BWSA, the common practice is to refer to the biographee by their (best known) surname throughout their own biography. However, because of existing kinship relations between different people described in the BWSA, two people might share a surname. At the document level, this becomes a factor when, for instance, two people with the same surname are both referenced by only that surname within the same biography. Similarly, at the corpus level, an occurrence of

"Troelstra" in the biography of Pieter Jelles Troelstra likely refers to a different entity than an occurrence of the same name in the biography of his brother, Dirk Jelles Troelstra.

Multi-morphic ambiguity

Among the people described in the biographies are writers, artists, and activists who don't shy away from a pseudonym or two. On top of that, the data that we are dealing with is of a historical nature. As the spelling of everyday language has evolved, so has the spelling of names. The biographies that we are studying were written by a multitude of authors, increasing the risk of spelling variations and, simultaneously, the multi-morphic ambiguity. This mostly comes into play when identifying names of organizations and locations. Fortunately, a biographee's aliases are in most cases explicitly mentioned in the introduction of his or her biography. For our purpose, we can extract these using regular expressions and add them to a list of known names.

A common phenomenon in Dutch surnames, which adds to the ambiguity of a name, is a surname prefix ("tussenvoegsel"): a string of one to three non-capitalized tokens which consists of mostly prepositions and determiners (e.g. "van", "van de", "op de"). These prefixes usually occur at the beginning of a surname, though in some cases they are found in the middle of the surname ("van den Bergh van Eysinga"). They come from a limited set of words that occur frequently in the language and thus many Dutch surnames have the same prefix. As a result, these tokens carry less information in the sense that an overlapping prefix between two surnames does not attribute much to matching them to one another. Titles form another such group of less informative tokens that are not officially part of a person's name, but do help in identifying the referent. They are more informative than prefixes, because overall they are shared between a smaller number of entities, but still contribute to the ambiguity of a name. In the same way, organization names may contain more general terms that denote, for instance, the type of organization, its geographic location, or its ideological background.

3.1.2. Consequences of misidentification

The connections between the nodes in our social network will ultimately be formed by co-occurrences of named entities within a span of text. An incorrectly identified entity can therefore severely change the structure of the network, and thus the outcome of any analysis made on the graph. Consider the example tree structure in Figure 3.1.a, representing a small organization with two departments, where a node represents one employee and all communication flows from top to bottom. Suppose employees E and F have the same first initial and last name, "J. Smith". Without any further contextual knowledge, the two are indistinguishable from one another. If all communication from and to node F is mistakenly identified as going

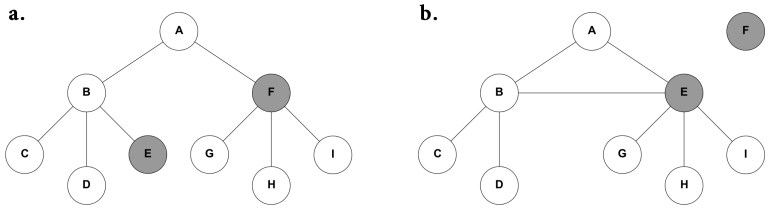


Figure 3.1 - *a.* Hierarchical tree graph representing a straightforward top-to-bottom information flow within a small organization. *b.* Graph representation of the information flow in the same organization if all communication to and from node F is mistakenly attributed to node E.

from and to node E, node F is left completely disconnected from the graph and the overall structure is changed to such an extent that it is no longer a valid tree structure (Figure 3.1.b). By erroneously identifying one as the other, we run the risk of mistaking an influential node for a less influential one, or vice versa, and adversely changing the represented course of history. Luckily, our biographical data lends itself perfectly for conversion to a social network representation, since the 573 documents already give us an equal number of distinct nodes for the network. Each document describes one person's life and mentions other entities with which the biographee interacts. Any of those entities that we can classify as being of a specific type is a suitable candidate for a node in our network.

3.2. Previous research into NER and NED

The task of automatically detecting and identifying named entities in text has received much attention since the mid 1990's (Grishman & Sundheim, 1996a; Nadeau & Sekine, A survey of named entity recognition and classification, 2007). Research in this area covers many domains and languages. Platforms such as the Message Understanding Conference and the Conference on Computational Natural Language Learning have included shared tasks on Named Entity Recognition (MUC-6, MUC-7, CoNLL-2002, and CoNLL-2003), resulting in an increased availability of systems and datasets (Grishman & Sundheim, 1996b; Tjong Kim Sang E. F., 2002; Tjong Kim Sang & De Meulder, 2003). CoNLL-2002 even included a Dutch dataset in their shared task.

Both MUC and the Automatic Content Extraction (ACE) program have included a within-document coreference task, where all references to a single entity within the same document needed to be grouped together (Chinchor & Hirschmann, 1997; Doddington, Mitchell, Przybocki, Ramshaw, Strassel, & Weischedel, 2004). This disambiguation task was not limited to named entities, but also included references in the forms of nouns and pronouns. This problem is commonly referred to as Coreference Resolution. When resolving entities across document boundaries it is called Cross-Document Coreference Resolution (CDC). For our purpose, we restrict ourselves to a review of disambiguation of named entities alone, and do not

consider other word types, such as personal pronouns and descriptive nouns. We realize that, for most types of input data, the inclusion of pronouns in the entity analysis would likely increase the recall of the relationships found in the text. However, the vast majority of the pronouns in the BWSA refer to the biographee of the document that they occur in. Since we consider the presence of the biographee to be implied throughout his biography and connect him to all other entities mentioned anyway, analysis of pronouns does not reveal any additional information for our graph. In fact, it might do more harm than good, since it could also introduce more noise into the output.

3.2.1. Named Entity Recognition

Nadeau & Sekine (2007) provide a comprehensive overview of methods and features commonly used in NER, of which we will discuss only the most relevant points here. The first systems for NER were mostly based on handcrafted rules (Black, Rinaldi, & Mowatt, 1998; Hanisch, Fundel, Mevissen, Zimmer, & Fluck, 2005). One disadvantage of using a rule-based approach is that the system will never recognize any instances that do not match the rules. While it may intuitively seem that the format of names, especially person names, follows a distinct set of rules, it actually appears to be highly dependent on the context, domain, and language (Maynard, Tablan, Ursu, Cunningham, & Wilks, 2001; Minkov, Wang, & Cohen, 2005).

Nowadays, the common approach is to use Machine Learning techniques to solve the task. These fall into three categories: supervised learning (SL), semi-supervised learning (SSL), and unsupervised learning (UL). SL requires a training set of considerable size in which both positive and negative examples of named entities are given, accompanied by features that describe contextual and/or structural aspects of each example. The system determines which of the features are most distinctive for each type of entity and uses this information to identify them in unseen documents. Methods commonly used in SL for NER include, but are not limited to, Hidden Markov modelling (Bikel, Miller, Schwartz, & Weischedel, 1997), Support Vector Machines (Kazama, Makino, Ohta, & Tsujii, 2002; Asahara & Matsumoto, 2003), Maximum Entropy Models (Borthwick, 1999; Chieu & Ng, 2002), and Conditional Random Fields (CRF) (McCallum & Li, 2003; Finkel, Grenager, & Manning, 2005). All of these methods incorporate the Markov property, which implies that decisions about the current state of the system, i.e. the decision whether or not the current word is part of a named entity, depends only on local properties. As noted by Finkel, Grenager, & Manning (2005), this approach inhibits consistency of labels across a dataset. This means that two occurrences of the same name within a single text might be mistakenly classified as different types of entities, because the properties of one occurrence have no effect on the classification of the other. To overcome this issue, they combine a CRF sequence labeller with the approximate inference algorithm Gibbs sampling, instead of the more commonly used Viterbi algorithm (Lafferty, McCallum, & Pereira, 2001;

Geman & Geman, 1984; Forney Jr., 1973). Gibbs sampling allows the encoding of non-local dependencies for sequence models. Finkel, Grenager, & Manning (2005) report F-measures of 92.3, 81.7, and 88.5 for the classes PERSON, ORGANIZATION, and LOCATION, respectively, on the CoNLL-2003 dataset, which are among the highest scores that have been reported on this dataset.

When training data is not available, SSL or UL can be applied. SSL methods require only a few positive examples. The system extracts features from the given examples in context and uses this information to bootstrap the recognition process (Nadeau & Sekine, 2007; Ekbal & Bandyopadhyay, 2008; Buchholz & Van den Bosch, 2000). UL, on the other hand, requires no pre-classified examples, but instead uses the lexical and statistical properties of the input data and compares them to those of other existing lexical resources (e.g. WordNet, thesauri, other corpora) (Alfonseca & Manandhar, 2002; Nadeau, Turney, & Matwin, 2006).

Many NER systems rely on word-level features, such as capitalization or morphology (Bick, 2004). These can be observed on the word itself, and on any of the words surrounding it, or combinations thereof, allowing the feature space to grow and encode more details. Features might also be derived from external resources. For example, *gazetteers* are precompiled lists or dictionaries of known entities by name that can be used to look up strings suspected to be names in an input document (Mikheev, Moens, & Grover, 1999). External corpora can provide useful prior probabilistics, such as the likelihood of a word being capitalized when not at the beginning of a sentence, or the probability of a multi-word unit belonging to a named entity (Ferreira Da Silva, Kozareva, Gabriel, & Lopes, 2004).

3.2.2. Named Entity Disambiguation

A method that is now widely used for named entity disambiguation was first developed by Bagga & Baldwin (1998). For within-document coreference, they make use of the UPenn CAMP system (Baldwin, et al., 1998). Personal names are resolved by matching rule-induced alternate forms of a full name with names found in the text. Organizations and locations are not disambiguated. To solve cross-document coreference, they first create a summary of each entity in each document by extracting all sentences containing a mention of that entity. Next, they calculate similarity among the summaries using the Vector Space Model for Information Retrieval (Salton, 1989) to determine which summaries are about the same entities. They report F-Measures up to 84.6% on a collection of articles from the New York Times with a similarity threshold of 0.1 (Bagga & Baldwin, 1998).

Gooi & Allan (2004) implement a simplified version of Bagga & Baldwin's algorithm for cross-document coreference. While their system achieves equivalent scores on Bagga & Baldwin's corpus, they show that it does not translate well to larger corpora. They also experiment with two other forms of clustering, agglomerative clustering and clustering based on Kullback-Leibler Divergence

(Kullback & Leibler, 1951; Kullback, 1968). They find agglomerative clustering to be the best approach.

Mann & Yarowsky (2003) take an unsupervised approach by generating patterns using Web queries that can extract biographical facts surrounding the entity mention. These rich features are then used to cluster mentions of the same entity. The task of the Web People Search Evaluation Workshops (WePS) is to cluster a set of search results, obtained using an ambiguous person name as the query, in such a way that each cluster of documents describes one unique entity (Artiles, Gonzalo, & Verdejo, 2005; Artiles, Gonzalo, & Sekine, 2007; Artiles, Gonzalo, & Sekine, 2009). Documents can be in multiple clusters, turning this into a *fuzzy*, instead of a *strict* clustering task. Many of the systems that have participated in WePS use an adapted version of the algorithm described by Bagga & Baldwin (1998) (Chen, Lee, & Huang, 2009; Ikeda, Ono, Sato, Yoshida, & Nakagawa, 2009). Other clustering algorithms used to solve the disambiguation task include k-means clustering (Pedersen, Purandare, & Kulkarni, 2005), and more semantically motivated methods such as Pointwise Mutual Information (Bollegala, Honma, Matsuo, & Ishizuka, 2008; Chen, Lee, & Huang, 2009) and Latent Semantic Analysis (Balog, Azzopardi, & De Rijke, 2008).

In an attempt to validate the reliance on the cluster hypothesis in much of the name identification research, Balog et al (2008) compare two different clustering techniques on this task, namely single pass clustering and Probabilistic Latent Semantic Analysis (PLSA). They implement single pass clustering with two types of similarity measures: Naive Bayes and cosine similarity calculated on tf.idf weighted term vectors. Despite the encoding of semantic relatedness between documents in PLSA, single pass clustering with the standard Information Retrieval vector representation is reported to work best and produces state-of-the-art results (Balog, Azzopardi, & De Rijke, 2008).

3.3. BWSA-NERD

Much of the NLP research into named entities focuses on newspaper data, resulting in an increased availability of training data of this genre. As a domain, we consider news to be much broader than a biographical dictionary surrounding a specific theme such as Socialism. While anything may be, and is, reported in the news in greater or lesser detail, biographies in general, and the BWSA in particular, require only the most important events to be described, and only in such detail as is deemed necessary for the telling of the overarching story. Borthwick (1999), among others, note the importance of consistency across training and test data. They report a drop in performance of their system on the MUC7 NER task of almost 7 points when training on airline disaster articles and testing on articles about rocket and missile launches, compared to both training and testing on airline disaster articles. If a mere shift in topic can already have such a detrimental effect on performance, we anticipate that a training set consisting of newspaper data alone

will not suffice for acceptable classification of the entities in the BWSA. Because of its non-specific nature and abundant availability newspaper data might make a useful additional resource for the processing of more specialized datasets, but it would seem to require a combination with data from the target domain. In order to ensure optimal performance of our NERD system on the BWSA, we therefore need to investigate the effects of training on a dataset from one domain and applying it to the other. Unfortunately, there currently are no datasets available that include annotations for NERD for Dutch text that are suited for comparison to the BWSA. For this task we will instead compare our scores to those reported for a comparable English language dataset.

3.3.1. Named Entity Recognition

We use the Stanford NER[8] system for the classification of named entities in biographies. This system is an implementation of the CRF approach described in (Finkel, Grenager, & Manning, 2005), but utilizing the standard Viterbi algorithm instead of Gibbs sampling. F-measures reported for this implementation are 90.4, 80.8, and 88.2 for classes PERSON, ORGANIZATION, and LOCATION, respectively. These scores are slightly lower than those reported for the Gibbs implementation, but still representative of the current state-of-the-art. We also choose the Stanford NER system because of the ease with which it is retrained for a new language or domain, given the presence of a large enough set of annotated examples.

Een	O
gemeentelijke	O
woningstichting	O
,	O
Centraal	B-ORG
Woningbeheer	I-ORG
,	O
door	O
Baart	B-PER
in	O
1923	O
in	O
het	O
leven	O
geroepen	O
...	...

Figure 3.2 - Example of data annotated with named entities in the BIO-format. *B-ORG* means that 'Centraal' is the first token in an organization name; *I-ORG* means that 'Woningbeheer' is the consecutive token in the same name. Since the next token ',' is tagged with *O*, it is not considered as part of the named entity and the full name is, consequently, 'Centraal Woningbeheer'. The next named entity is a one token person name, 'Baart'.

[8] http://nlp.stanford.edu/software/CRF-NER.shtml

		BWSA		BD98		CoNLL-2002	
		training	*test*	*training*	*test*	*training*	*test*
PER	*total*	3,576	1,655	3,569	944	4,716	1,098
	per 1,000	31.9	32.0	21.5	18.7	23.3	15.9
	avg. # occ.	2.55	2.27	1.82	1.80	1.86	1.62
	overlap	21.9 %	38.3 %	6.2 %	17.8 %	8.7 %	24.7 %
ORG	*total*	2,082	961	1,739	536	2,082	882
	per 1,000	18.6	18.6	10.5	10.6	10.3	12.8
	avg. # occ.	2.40	2.05	2.02	1.68	2.62	2.52
	overlap	42.7 %	54.4 %	17.7 %	28.4 %	35.1 %	47.5 %
LOC	*total*	1,391	702	3,869	1,299	3,208	774
	per 1,000	12.4	13.6	23.3	25.7	15.8	11.2
	avg. # occ.	3.57	3.22	2.82	2.38	3.43	2.52
	overlap	65.2 %	74.1 %	45.9 %	59.3 %	49.3 %	56.5 %

Table 3.1 - Descriptive statistics regarding the training and test sets for the NER experiments. For each set the total number of annotated entities is given, their frequency per 1,000 tokens, followed by the average number of occurrences per unique entity, which indicates their global consistency. The overlap measure expresses the percentage of the total number of entity types in the training set that is also included in its accompanying test set, and vice versa.

	PER			*ORG*			*LOC*		
	BWSA	*BD98*	*CoNLL*	*BWSA*	*BD98*	*CoNLL*	*BWSA*	*BD98*	*CoNLL*
BWSA	-	4.4 %	1.3 %	-	7.1 %	3.9 %	-	61.2 %	49.1 %
BD98	3.0 %	-	6.2 %	5.6 %	-	14.9 %	23.6 %	-	23.4 %
CoNLL	0.8 %	4.6 %	-	2.9 %	12.9 %	-	23.0 %	49.8 %	-

Table 3.2 - Percentages of entity type overlap per entity category between all three datasets. The training and test sets have been merged for this purpose.

To test the transferability of one genre to the other, we compare performance between CRF-classifiers trained on newspaper data with those trained on biographical data, and a combination of the two. Though the biographies are written in modern day Dutch, the contents are of a historic nature, which is reflected in the formulation of some of the names. Therefore, comparing the datasets also allows us to test whether modern day data is suitable training data for the recognition of historic names.

We use three datasets in our experiments:

- **BWSA**: A subset of the BWSA biographies, manually annotated with classes person, organization, and location. The training set contains 70 biographies with a total of 112,228 tokens. The test set contains 30 biographies with a total of 51,690 tokens.
- **BD98**: A selection of issues of a Dutch newspaper, *Brabants Dagblad*, from January 1998, annotated with classes person, organization, and location (Buchholz & Van den Bosch, 2000). The training set contains 166,286 tokens; the test set contains 50,564 tokens.

- **CoNLL-2002:** The Dutch dataset as it was compiled for the Named Entity Recognition task at CoNLL-2002 (Tjong Kim Sang E. F., 2002), consisting of four editions of Belgian newspaper *De Morgen* published in the summer of 2000. The dataset is supplied with two test sets, of which we only use one, set b. It contains 68,875 tokens, versus 202,644 tokens in the training set.

We create an additional training set by combining the training sets from all three sources. This set contains 481,158 tokens. Entities are annotated according to the BIO-scheme (Figure 3.2), where a *B-[class]* tag denotes the first token in a named entity and can be followed by one or more *I-[class]* tags marking tokens that are inside the named entity. Tokens that are not part of a named entity are tagged with *O*.

Table 3.1 shows per dataset, for both the training and test sets, how many named entities are annotated of each class in total, their ratio per 1,000 tokens, the average number of occurrences of a unique name, and the percentage of overlap the set has with its counterpart (training versus test). Even though the BWSA sets are much smaller than the other sets, the density of person and organization names in the text is a lot higher judging by their occurrence per 1,000 tokens. This is a clear signal of the differing genres. Considering the average number of occurrences, the entities also show far greater consistency in the BWSA for all classes, except organization names. Here, the CoNLL-2002 set clearly diverges from both the other sets. Since we have not disambiguated these named entities, we cannot conclude with absolute certainty from this that less unique organizations are mentioned in the CoNLL-2002 data, though it seems to be the most plausible explanation. The fact that the CoNLL-2002 data has a Belgian instead of a Dutch origin may also play a part, since Flemish is lexically and grammatically slightly different from Netherlandic Dutch.

When comparing the three named entity classes amongst one another, we see that location names show the highest consistency across all sets. This is logically explained by the fact that there exists a finite number of named locations, which is considerably smaller than the number of distinctly named organizations or persons that exist in the world. It also explains the relatively high overlap for this category in all sets. Location names in the BWSA occur at a lower rate than in the BD98 sets, but the names that do occur have a higher number of average occurrences. This is a consequence of the BWSA describing a specific group in a specific location (i.e. the Netherlands). The general nature of BD98 allows for the occurrence of any location in the world, which is reflected by the high number of unique locations in these sets. The CoNLL-2002 sets show more similarity to the BWSA with regards to the location name density. This is possibly explained by the short period over which the CoNLL-2002 articles were published (4 months, versus 1 month for BD98).

Perhaps the most interesting measure in Table 3.1 is the percentage of entity overlap between the training and test sets. This can be seen as a crude baseline for the performance of a set on itself. Again we see the largest consistency across the

BWSA sets. A higher consistency between a training and test set increases the expected performance of that training set on that test set, but at the same time it decreases its transferability to another, unseen dataset. This is not desirable, since we ultimately aim to apply the trained classifier to the entire BWSA. Combining the BWSA training sets with the other training sets decreases the percentage of entities from the training set that overlap with the test set, while the percentage of entities from the test set that overlap with the training set stays the same. Theoretically, this would increase the expected performance on unseen data because of the presence of more entities in the training data, while expected performance on the test set remains roughly similar. For completeness, we also report the measured overlap between the different datasets (Table 3.2). Judging from these numbers, the BWSA and CoNLL-2002 sets are furthest apart, with the BD98 set in between. We therefore expect this last set to be most suitable as training data for the other two.

We perform three experiments on our chosen datasets. In the first experiment, we cross-validate the three datasets to find out whether a classifier trained on one domain performs well on a dataset of another domain, or even on a different dataset from the same domain. We expect to see a drop in performance when training and testing on different domains. Since both BD98 and CoNLL-2002 are composed of newspaper articles and from roughly the same period, we hypothesize that these sets will translate fairly well to one another. Since the BWSA set does not just differ from the other two sets in genre (biographies versus news), but also in temporal domain (historical versus contemporary), this experiment concurrently allows us to test the temporal transferability of the data for this task. The classifiers have been configured to use the following word-level features:

- the previous, current, and next token (i.e. words or punctuation markers);
- capitalization of the previous, current, and next token;
- bigrams constructed from the previous, current, and next token;
- character *n*-grams with a maximum length of 6 constructed from the current token.

In our second experiment, we train a single classifier on the combined training sets of all three datasets to investigate whether additional training data from a different domain improves performance. In an attempt to further boost performance on the BWSA set, which is our primary goal, we run a third experiment, in which we supply external, domain-dependent information to the classifier. This information comes in the form of a domain-specific gazetteer containing 127,083 person names, 7,053 organization names, and 2,266 location names. The list consists of terms extracted from the archival files of the IISH (Zervanou, Korkontzelos, Van den Bosch, & Ananiadou, 2011), supplemented with all current Dutch municipalities and provinces, and all current countries. Since the gazetteer has been prepared from data specifically comparable to the BWSA, we expect it to boost performance especially on this dataset. The existence of overlap between the different datasets (Table 3.2) implies that performance on BD98 and CoNLL-2002 should also slightly improve, given that some of the overlapping entities occur in the gazetteer.

Stanford NER provides two options for applying the gazetteer in the NER process: *clean*, which means that only names that fully match names on the list are accepted, or *sloppy*, which means that names which partially match with a name on the list are also accepted. We experiment with both settings.

3.3.2. Named Entity Disambiguation

Name disambiguation is similar to Word Sense Disambiguation (WSD) in the sense that both assume that two words that are similar in meaning will appear in a similar context (Miller & Charles, 1991; Balog, Azzopardi, & De Rijke, 2009). Motivated by this hypothesis, a common approach to the task of identifying named entities in unstructured text is to use a clustering technique. However, the main difference between NED and WSD is that in WSD, the number of possible clusters (senses) for a word are limited and usually known a priori. Names are more ambiguous in the sense that a single name may be a valid reference to multiple real-world entities. For instance, several of our biographies are about people with the last name Cohen. When we encounter the string "Cohen" in the biography of someone not named Cohen, it can refer to any of them, or even to another person for whom we have no biography. The same applies to organization names, and in rare cases even to locations. Since we also want to locate and identify these entities that do not have their own document in our dataset, we cannot determine the desired number of clusters beforehand. This rules out k-means clustering, since the method requires the number of clusters to be known in advance. A solution to this problem would be to apply k-means clustering with the number of clusters set to the total number of mentions found in the corpus, and stopping the process when a certain condition, for instance a minimum similarity between cluster members, is met. However, the result depends on which clusters are chosen as the initial seeds and this makes the method too unreliable in comparison to single pass clustering, which will generate the same result in each run.

Our method for named entity disambiguation is closely related to that of Bagga & Baldwin (1998). However, the task solved by their method differs somewhat from ours. We aim to disambiguate every mention of a named entity in every document, while Bagga & Baldwin try to cluster documents as a whole based on the occurrence of one specific personal entity with an ambiguous name ('John Smith'). We use the ambiguity of a name as an aid in resolving its reference by allowing a name to only be matched to a name with less or equal ambiguity. To determine the ambiguity of a name, we compare each of its substrings against predefined lists of general terms (Appendix B) and assign an ambiguity value to each substring based on the entity class and term type. For person names, we check the name for the occurrence of titles and prefixes, which respectively have an assigned ambiguity of 0.25 and 0.5. Organization names are checked for the occurrence of infixes, domain-specific adjectives ('Rooms-Katholiek', 'Internationaal'), and collectives ('Bond', 'Comité'), with respective ambiguity values of 0.5, 0.25, and 0.25. We calculate the ambiguity of the entire name by taking the average of the ambiguity values of its substrings,

weighted by the length of the substring. By including string length in the calculation, we ensure that longer strings are always less ambiguous than shorter ones. Location names are compared to lists of terms denoting different types of locations, such as addresses, buildings, or municipalities (Appendix B). All types carry equal ambiguity, so this classification serves recognition and further processing more than it serves disambiguation.

To disambiguate the entities in the BWSA we apply *strict single-pass agglomerative clustering*. In this approach, every newly encountered entity is compared to all the clusters that have already been identified on a predefined set of properties. If any clusters are found that match above a set threshold, then the entity is added to only the best matching cluster, since a named entity inherently refers to a single entity. If no matches are found, a new cluster is created. Entities are processed in their order of occurrence in the text under the assumption that an ambiguous name is always preceded by a less ambiguous variation of the same name. For instance, the occurrence of the surname 'Troelstra' can only be disambiguated if it has been preceded by an occurrence of the same name combined with an initial or first name, such as 'P.J. Troelstra'. Most of the research into NED focuses strictly on resolution of personal names. We extend this to include names referring to organizations and locations in order to investigate whether the method also performs well on these types of entities. We compare our method to three baselines:

• **OneInOne:** Each entity mention/cluster is placed in its own cluster. This baseline favors precision over recall.

• **AllInOne:** All entity mentions/clusters are added to the same cluster. This baseline favors recall over precision.

• **Match:** All exactly matching entity mentions/clusters are added to the same cluster. This baseline completely ignores the issues of both multi-morphic and multi-referent ambiguity.

The NED process is split into two parts. First, we disambiguate the entities within a single document. Second, we resolve the references across document boundaries. We apply all methods first to a development set consisting of 10 biographies from the BWSA in order to determine the optimal similarity thresholds for both stages. The methods are then applied to a separate test set of 25 biographies with the threshold set to the best performing value to measure performance on unseen data. We will now explain the methodological differences between within- and cross-document disambiguation.

Within-document disambiguation

When disambiguating the entities at the document level, for each name that we encounter, we first check if we have already identified a cluster within the same

document containing an exact match on string level. For entities of type organization that contain an abbreviated name, we also check for any full names that are compatible with the abbreviation. For instance, 'Sociaal Democratische Arbeiders Partij' is compatible with the abbreviation 'SDAP', and 'Partij van de Arbeid' is compatible with 'PvdA'. If multiple matches are found, we add the entity to the cluster containing the closest preceding name with less or equal ambiguity. If no matches are found, we compare the entity's name to every name in each existing cluster, regardless of their ambiguity or position in the text. We anticipate that this will help in identifying any names for which the preceding, less ambiguous form was not recognized by our NER system.

Names are compared on string similarity. We first calculate their distance using the Levenshtein edit distance metric (Levenshtein, 1966), where we count character replacements as having distance 1. The distance is converted to a normalized similarity measure within the 0 to 1 range using the following equation:

$$similarity(n_1, n_2) = \frac{M}{M + Levenshtein(n_1, n_2)}$$

where M denotes the sum of the lengths of the matching substrings of n_1 and n_2. The similarity of an entity to a cluster is calculated by taking the mean of the similarities with the names in the cluster. If this measure is above the set string similarity threshold, the cluster is accepted as a match. Again, if one matching cluster is found, we add the entity to that cluster. If multiple matches are found, we add the entity to the cluster with the highest similarity, containing the closest preceding name with less or equal ambiguity. If no matches are found, the entity is placed in its own cluster.

Cross-document disambiguation

After we have clustered all entities in each document, we move on to identification at the corpus level. We apply the same clustering mechanism as in the within-document setup, but here we consider similarity at the level of the cluster rather than a string. Each cluster from each document ("document-level cluster") is compared to each cluster that has already been identified at the corpus level ("corpus-level cluster"), with the added constraint that the corpus-level cluster may not contain any document-level clusters originating from the same document as the current document-level cluster. This constraint exists to ensure that the cross-document disambiguation process does not override the outcome of the within-document disambiguation. If the least ambiguous name in the document-level cluster has a similarity over a predefined string similarity threshold with the least ambiguous name in the corpus-level cluster, then the corpus-level cluster is considered to be a possible match for the document-level cluster.

In a second step, we compare all possibly matching corpus-level clusters to the document-level cluster based on the context of the entities contained within them.

The context of a cluster consists of all lemmatized nouns, proper nouns, verbs, and adjectives that occur within a 50-word window around all the names of all entities in the cluster. Each lemma is added to the cluster's context vector with a weight equal to its absolute number of occurrences divided by the total number of lemmas in the context. This normalization is applied to cancel out the total context size as a factor in the comparison of document-level clusters and the expectedly larger corpus-level clusters. The final cluster similarity is then measured by taking the cosine similarity between the clusters' context vectors. If this measure is above a set context similarity threshold, then the corpus-level cluster is considered a definite match for the document-level cluster. The document-level cluster is merged into the corpus-level cluster with the highest context similarity. If no matching cluster is found, a new corpus-level cluster is created from the document-level cluster. The thresholds for string and context similarity are first determined on a development set consisting of 10 biographies from the BWSA. The best performing settings are then tested on a test set consisting of 25 biographies.

3.4. Experiments and results

In this Section, we offer the results of our methods for NER and NED on biographical Dutch text, where possible specifically compared to the same methods applied to Dutch newspaper data. We first present the outcomes of our three experiments into NER and its domain transferability. This is followed by our findings in performing within- and cross-document NED on the BWSA.

3.4.1. Named Entity Recognition

Experiment I — Testing transferability

Table 3.3 lists the NER results when we train the Stanford NER system on each of our three data sources' training sets, and test it on their own, and each other's test sets. All presented scores are F1-scores as reported by the CoNLL-2002 evaluation script.[9] Overall, the BWSA-trained classifier tested on the BWSA test set delivers the best results, which is promising for our task. It outperforms even the BD98 and CoNLL-2002 classifiers when applied to their own test set, which is also where the best results for these sets are achieved. Neither of the classifiers trained on newspaper data perform particularly well on the BWSA, or on each other. Out of the two, BD98 seems to be a better fit for the BWSA than CoNLL-2002, which is in line with our expectations based on entity overlap. Still, scores drop around 20 points for all classes with this change of genre. CoNLL-2002 proves only slightly better at classifying BD98 than the BWSA is, but again both score around 20 points lower than the BD98 classifier itself. Vice versa, BD98 performs reasonably well in comparison for person and location name recognition on the CoNLL-2002 data. Organization names turn out to be the most problematic for BD98, both with

[9] http://www.cnts.ua.ac.be/conll2002/ner/bin/conlleval.txt

regards to their recognition within the BD98 test set, as to the ability of the BD98 classifier to recognize entities of this class in the other test sets. In both cases the F1 score never goes beyond 68 points, which is quite low for this class in this task. This is in all likelihood a consequence of the lower density, consistency, and overlap in organization names across this set, as shown in Table 3.1. When the training set contains less examples of a class, the classifier will learn fewer cues for distinguishing instances of that class. Hence, the performance for that class will deteriorate.

Experiment II — Combining forces

The classifier for our second experiment is configured to use the same features as described in Experiment I, but this time it is trained on the combination of all three training sets. The results for our tests on each individual test set are listed in Table 3.3 under 'All'. The results on the CoNLL-2002 test set show a definite improvement over training on just its own training set. Surprisingly, combining the corpora does not help much to improve the scores for the BWSA or BD98 in comparison to the classifiers trained on their own test sets. In fact, when it comes to the recognition of person names, scores drop more than a full point. When we consider this outcome in conjunction with the results from our first experiment, we can conclude that most of the cues needed to successfully recognize entities in the BWSA that exist in the newspaper corpora, also exist in the BWSA itself. Therefore, adding the extra samples to the training data does not provide the classifier with better information for the task. Moreover, some of the cues in the newspaper data actually seem to contradict cues in the BWSA data, causing a mild drop in performance for the PER class.

Experiment III — Adding domain knowledge

For our third experiment, we train two classifiers on each individual training set, plus two classifiers on all combined training data, and we add a gazetteer to each pair. One classifier is configured to accept only clean, exact matches on the gazetteer, while the other also accepts sloppy, partial matches. Results are listed in Table 3.4.

As expected, performance on the BWSA set improves with the addition of a gazetteer, whether it is sloppy or clean, which underlines the substantive compatibility between BWSA and the domain-specific gazetteer. The combined training set delivers slightly better results when compared to the BWSA classifier without gazetteer from Experiment I, but the overall best performance is achieved by the classifier trained on the BWSA alone with sloppy gazetteer, improving scores with around two points for each class. The results on BD98 are improved only for the ORG class, and only when the classifier is trained on the combined training set. This supports our earlier assumption that the BD98 training set by itself does not contain enough examples of organization names to reliably recognize this class.

Test set	Training set	*PER*	*ORG*	*LOC*
BWSA	***BWSA***	89.3	85.7	88.3
	BD98	70.3	67.9	67.5
	CoNLL-2002	68.5	60.7	67.1
	All	88.1	86.0	88.8
BD98	***BWSA***	63.5	49.0	58.6
	BD98	83.4	66.3	84.4
	CoNLL-2002	68.2	50.3	61.9
	All	82.2	67.4	83.6
CoNLL-2002	***BWSA***	59.3	40.9	53.9
	BD98	73.7	50.9	74.9
	CoNLL-2002	83.0	72.2	81.7
	All	85.3	76.0	86.2

Table 3.3 - F1 scores for the different named entity classes trained and tested on all three datasets separately, and combined.

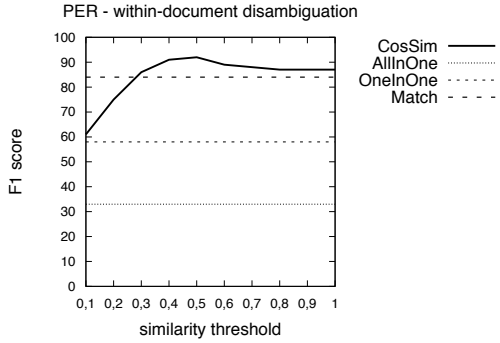
Test set	Training set	Gazetteer	*PER*	*ORG*	*LOC*
BWSA	***BWSA***	*clean*	92.8	87.6	90.9
		sloppy	**93.8**	**88.2**	**90.9**
	All	*clean*	90.5	86.8	90.5
		sloppy	91.0	87.6	90.7
BD98	***BD98***	*clean*	83.8	66.7	85.5
		sloppy	83.3	67.2	85.5
	All	*clean*	81.9	69.5	84.6
		sloppy	80.7	68.8	84.0
CoNLL-2002	***CoNLL-2002***	*clean*	73.9	68.6	79.4
		sloppy	80.2	71.3	81.7
	All	*clean*	87.4	76.7	87.0
		sloppy	88.3	76.4	87.6

Table 3.4 - F1 scores for the different named entity classes trained on the combined training sets and the separate training sets, and tested on each separate test set.

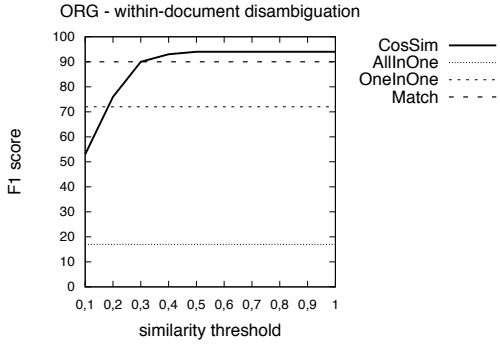
Cues learned from the other training sets and the examples provided by the gazetteer manage to increase performance with three points when applying clean gazetteer matching. The greatest gain is achieved on the CoNLL-2002 set, where scores increase with around five points for each class when we train on all training sets and add a sloppy gazetteer. The extra data contributes most to the improvement of organization and location name recognition, while the gazetteer greatly improves person name recognition.

3.4.2. Within-document disambiguation

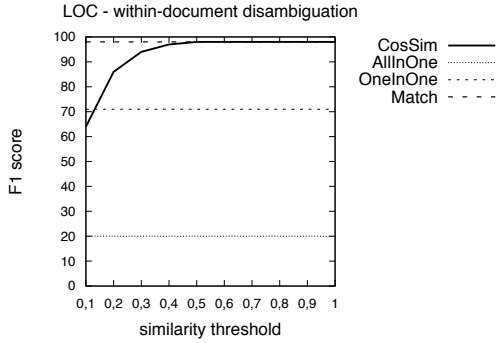
The results for the within-document disambiguation on the development set are displayed in the charts of Figure 3.3. The graphs show, for each of the named entity



BWSA	PER		
set	*algorithm*	*t*	*F1*
dev	CosSim	0.5	92.1
	Match	-	84.5
test	CosSim	0.5	92.0
	Match	-	83.8



BWSA	ORG		
set	*algorithm*	*t*	*F1*
dev	CosSim	0.6	94.8
	Match	-	90.1
test	CosSim	0.6	95.6
	Match	-	90.4



BWSA	LOC		
set	*algorithm*	*t*	*F1*
dev	CosSim	0.7	98.8
	Match	-	98.2
test	CosSim	0.7	99.2
	Match	-	99.0

Figure 3.3 - Left: **Within-document disambiguation results for, respectively, person, organization, and location names averaged over the 10 documents in the BWSA development set. All reported scores are F1 scores. The labels on the x-axis represent the string similarity threshold.** Right: **Results obtained on the development and test sets using the best performing CosSim algorithm and the best performing baseline.**

	PER		ORG		LOC	
	dev	*test*	*dev*	*test*	*dev*	*test*
CosSim	84.9	86.1	87.8	81.6	78.6	75.5
Match	77.2	77.6	82.7	76.3	78.2	76.5

Table 3.5 - Cross-document disambiguation results obtained on the development and test sets using the best performing CosSim algorithm and the best performing baseline. All reported scores are F1 scores.

classes, the F1 score achieved by our algorithm, which we have labeled *CosSim*, compared to our three baselines. In order to determine the optimal threshold to apply to the test set, we test our approach with similarity threshold values ranging from 0.1 to 1, with intervals of 0.1. For all three of the entity classes, we observe an initial increase in performance as the string similarity threshold rises. It reaches its peak at threshold values of 0.5 for person names, 0.6 for organization names, and 0.7 for location names. After a very slight decrease the performance evens out for the highest tested threshold values. The tables accompanying the charts in Figure 3.3 show the performance of our CosSim method on both the development and the test set with the optimal threshold values for each class. For means of comparison, we also provide the equivalent scores for the best performing baseline.

The best performing baseline is *Match*. *CosSim* succeeds at defeating it for all categories, though the performance increase for location names is only minimal. Since the *Match* algorithm does not solve the problem of multi-morphic ambiguity, it fails to cluster different forms of the same name, such as 'Henriette Roland Holst' and 'H. Roland Holst' for a person, and 'Sociaal-Democratische Arbeiderspartij' and 'SDAP' for an organization. For the location category, *CosSim* gives approximately the same performance as the *Match* baseline. Both give near perfect outcomes. We can conclude from this that the use of location names in the BWSA is very consistent. Further inspection of the output shows that the only instances where the *CosSim* algorithm outperforms the baseline are instances where there is a spelling error in the name, which increases multi-morphic ambiguity.

Parts of the BWSA have been digitized through Optical Character Recognition, where a scanned image of the text is converted into a digital text document. Misrecognition of characters has introduced spelling errors into the data. However, the BWSA is actively maintained and when errors are found, these are manually corrected. Therefore, we can assume that the number of errors is kept to a minimum, which is reflected by the small difference in outcomes between *CosSim* and *Match*. We achieve F1-scores of around 92% for person names, 95% for organization names, and 99% for location names. In all cases, precision approaches 100%, while recall varies between 86% and 98% for *CosSim*, and between 73% and 97% for *Match* on the development set. The same is observed in the results obtained on the test set. This implies that both algorithms are able to distinguish different entities from one another. However, *CosSim* is less capable of solving the multi-morphic ambiguity, and *Match* does not solve this problem at all. This results in the generation of more clusters, and thus more nodes in the social network than

actually exist, which is something we need to keep in mind when constructing the graph structure.

3.4.3. Cross-document disambiguation

Table 3.5 displays the results for cross-document disambiguation, where the clusters from different documents are merged when judged to be describing the same entity. In contrast to within-document disambiguation, we observe no change in performance when we vary the context similarity threshold for cross-document disambiguation. This implies that the similarity of cross-document contexts is generally low and almost all information needed for disambiguation is derived from the names themselves. To ensure that we do not miss those cases where the context does supply differentiating data, we do not discard it completely. For the rest of our experiments, we set the threshold to 0.01 for all classes and report only the F1 scores for this value on the development and test set. The tables list the results achieved using the *CosSim* algorithm and the best performing baseline, which again is the *Match* algorithm. For the person class, *CosSim* achieves an F1 score of 84.9 on the development set, with precision and recall at 98.9 and 74.7, respectively. These results are comparable to the results obtained by Bagga & Baldwin on their 'John Smith' corpus with a context threshold of 0.1 (Bagga & Baldwin, 1998). The *Match* algorithm achieves an F1 score of 77.2, with precision at 99.9 and recall at 63.3, on the development set. The high precision scores indicate that we are able to correctly separate the different entities in this category. *CosSim* reaches a recall score approximately 11.5 points higher than *Match*, confirming its superior ability to cluster different references to the same entity based on string and context similarity.

However, performance is nowhere near perfect. In the within-document disambiguation process, we reach an F1 score of around 92 when resolving person names. These errors carry over to the cross-document disambiguation, since we do not allow matching between clusters that come from the same document. This results in an increased number of clusters and a lower recall. On top of that, the *CosSim* algorithm does not succeed in clustering all references to one entity, because in most cases their contexts just aren't similar enough.

For organization names we achieve an F1 score of 87.8, with a precision of 97.0 and a recall of 80.6. *Match* scores around 5 points lower with an F1 score of 82.7, a high precision of 99.8, and a recall of 72.0. A factor that greatly increases the multi-morphic ambiguity in this category is the use of abbreviations of organization names, which happens throughout the BWSA. *CosSim* successfully clusters these with their full name on document level. In some instances, however, an abbreviation is considered to be so well known, that it is used without first mentioning the organization's full name. In other cases, the full name is not recognized by our NER system. When this happens, the abbreviation is the least ambiguous name in the cluster and, thus, it is used to compare the cluster against

the corpus clusters. If the abbreviation overlaps with the abbreviation of another organization (such as 'NV' for 'Nederlandsche Vereeniging voor Spoor- en Tramwegpersoneel', and 'NVV' for 'Nederlands Verbond van Vakverenigingen'), they often exceed the string similarity threshold, and thus are considered a possible match, while they are not. This lowers the precision score for *CosSim*. The wrong pairing of names could also play a factor in *CosSim*'s inability to fully solve multi-morphic ambiguity, by reducing the coherence of the cluster's combined context.

The average F1 score for this location class is 78.6, with precision and recall at 99.5 and 65.3, respectively. The *Match* algorithm achieves approximately the same results, with 78.2, 99.3, and 74.9 for F1 score, precision, and recall. The low F1 score is surprising, considering the high score we achieved on this category for within-document disambiguation. Inspection of the data reveals that some of the errors are caused by wrongly recognized named entities which contain a place name, such as the organization name 'Universiteit van Amsterdam', or the conjugation 'Engelschen' (*English*, old Dutch spelling), which we would classify as *miscellaneous*, a category which is not included in the current setup. In the gold standard, these instances have been given the same identifier as the location name that they refer to: 'Amsterdam' for the first; 'Engeland' for the latter. There are also instances where a location name occurs in the data in multiple languages. The *Match* algorithm cannot cluster these entities because the names are not the same. The *CosSim* algorithm fails to cluster them together because either the strings or their contexts aren't similar enough.

3.5. Discussion

In the experiments outlined in this Chapter, we set out to investigate the compatibility and transferability between newspapers and biographical data for the task of Dutch NER and NED in order to answer our first research question:

> **RQ 1** To what degree can we reliably recognize and identify named entities in Dutch biographical text using state-of-the-art techniques?

As shown by the results of our experiments, NER classifiers trained on newspaper data perform better on biographical material than the other way around. This is in line with our initial assumption that newspaper data is of a more general nature, and thus more suitable as input for domain-specific datasets. However, our results show that the combination of biographical texts and newspapers in a single classifier does not significantly improve performance on either genre. Therefore, the application of generalised input only seems to be a viable solution if no training data from the target domain is available. From the perspective of the BWSA this implies that it is a mostly self-contained dataset, meaning that the entities therein are drawn from a relatively small and closed set, and that they occur in structurally very similar contexts within the dataset. Considering the results in light of the fact that the BWSA sets are much smaller than both newspaper corpora, we can also

conclude that less training data is needed to successfully perform NER on a specialized and curated biographical source, if the training data originates from this same source. This is most likely caused by the high consistency and high density of entities in the BWSA, which in turn is a direct result of the biographical genre in which encounters and other events are presented in a condensed manner.

Inclusion of a gazetteer with examples of domain-specific named entities provides a definite advantage over classifying based on the training data alone. The recognition of person names most benefits from the addition of this extra data. Overall, organization names prove to be the most difficult to recognize, especially when transferring from one domain to another. Here the historic nature of the BWSA undoubtedly plays a part, since organization names tend to be of a rather descriptive nature, explicating the nature and goal of the organization, often in era-dependent words. For instance, during the period described in the BWSA, the ideological background of an organization was of far greater importance than it is nowadays and was often explicitly included in the name, such as in "Bond van *Christen-Democratische* Propagandaclubs" and "Amsterdamse *Sociaal-Democratische* Vrouwenpropagandaclub".

Our scores for NER are comparable to what is reported for state-of-the-art systems on this task. On our BWSA test set we achieve F1 scores as high as 93.8, 88.2, and 90.9 for the recognition of persons, organizations, and locations, respectively. When we apply the best performing classifier to the remainder of the BWSA we have to take into account the possibility that the unseen biographies contain instances that are less compatible with our training set. We therefore add some caution to our prediction of its ultimate performance on the totality of the BWSA. Based on the variation in our results, we feel confident in stating that our BWSA NER classifier can classify named entities in Dutch biographical text from the Social History domain with F1 scores between 90 and 94 for person names, between 86 and 88 for organization names, and between 88 and 91 for location names, which is still compatible with the state-of-the-art.
As shown by our experimental results for NED, multi-morphic ambiguity forms the biggest problem in resolving the entities in the BWSA corpus. Compared to a naïve baseline, our algorithm performs well for both person and organization names, reaching state-of-the-art scores, but it does not contribute much for the location category. This is in part due to the already consistent nature of location names in general. It could be argued that the errors made due to entities misrecognized as locations are not really errors and that they in fact should have been separated in the gold standard. This issue would be resolved with improved NER. However, given the already state-of-the-art performance of our NER classifier, it is unreasonable to assume that this problem can be fully resolved via that avenue. The multi-morphic ambiguity in instances where one location name appears in multiple languages could be resolved using a geocoding service, though we chose not to apply this technique to the current dataset at this time.

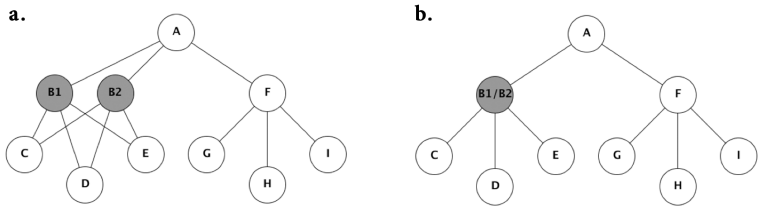


Figure 3.4 - *a.* Hierarchical tree graph representing a straightforward top-to-bottom information flow within a small organization, where one node has been wrongly split into two nodes. *b.* Graph representation of the information flow in the same organization where the two nodes are merged based on common attributes.

The errors made in the person and organization categories are largely attributed to the fact that the contexts of the clusters are too dissimilar. This can be seen as evidence supporting Gooi & Allan's finding that the Bagga & Baldwin algorithm is less suitable for larger corpora (Gooi & Allan, 2004), since the BWSA contains over 900,000 tokens, compared to approximately 170,000 in Bagga & Baldwin's 'John Smith'-corpus. Their corpus is also much more focussed than the BWSA, since it describes 35 distinct people from completely disjoint communities, whereas the BWSA describes several hundred entities whose lives and actions are all intertwined. It is uncertain at this point whether adding more data, whether it be from the BWSA itself or from another domain-specific source, would help in increasing context similarity.

The multi-morphic ambiguity that is left in the data after the complete disambiguation process varies among the entity categories. For person names it comes down to approximately 15%, compared to 18% for organization names, and 25% for location names. These errors will ultimately cause our graph representation to contain too many nodes, resulting in an unnecessarily, and undesirably, high fragmentation of the social network. Although improved NER would definitely also help in improving the results for persons and organizations, this is only part of the problem and, as of yet, there exists no automated service to look up all people or businesses, let alone historic ones. However, there are other possible solutions.

For instance, we could add extra features to the disambiguation process. These features could be gathered from other data sources, such as Wikipedia (Ratinov, Roth, Downey, & Anderson, 2011), or more domain-specific documents from IISH's archive. But they could also be derived from the disambiguation results themselves. When we build the graph structure, we could for instance place a minimum on the number of aggregated occurrences of an entity before accepting it as a node in the network. Entities that fall below this threshold can be analysed separately, comparing their structural properties (e.g. the other nodes that they are connected to) and merging them if these are similar enough.

We illustrate this with an example. Consider the graph in Figure 3.4.a. It is the same graph as is displayed in Figure 1a, representing a hierarchically structured organization, only in this case different references to the same node *B* have been mistakenly separated from one another. Since nodes *B1* and *B2* actually refer to the same person, they show similar characteristics: they both communicate with nodes *A*, *C*, *D*, and *E*. Since there are no other nodes in the network that have these exact properties, we could merge them, by which we restore the tree structure (Figure 3.4.b.). This example shows how the social network extraction process itself can serve to signal possible errors in the data and allows the researcher to decrease noise in the results of the analysis in a focussed manner.

3.6. Social Network Model Construction

Now that we have located and identified all named entities in the data that belong to the categories of people, organizations, and locations, we can build our first social network model. The core of the socialist community roughly consists of the 573 people whose biographies are present in the BWSA. There are also references to person entities outside of this set. However, these entities on average are mentioned only once or a few times in a few biographies, and as such they do not contribute much to the overall community structure. In light of these considerations, we include only the 573 biographees as person nodes in our network models. From a SNA perspective, organizations and locations perform vastly different roles than people do. Compared to an individual person, an organization is a collective. It may seemingly carry out actions, though these are actually performed by the people within the organization. A location performs the same type of role when it is used as a reference to its inhabitants. However, when it occurs in reference to a geographical location, it loses its power to act altogether and can only be acted on by other types of entities. Essentially, each role forms a different layer, or *mode*, in the network, each with different abilities for connectivity. For now, we will focus solely on the person nodes.

The edges in the graph are formed by the actual co-occurrences of the different entities in the dataset. For this purpose we need to define what we mean with "co-occurrence". Considering the BWSA, two entities can co-occur at different levels: they appear in the same biography, in the same paragraph, or in the same sentence. The particular genre of the (short) biography is characterized by its compact description of events, due to its need to condense an entire life span into a few paragraphs. We won't know how long of a period is described by each paragraph or sentence until we perform an analysis of all temporal expressions in the text. However, given that a document describes an entire life, we can assume that a paragraph describes several years, if not decades. If we would calculate our co-occurrences based on paragraphs, we would likely group together many unrelated occurrences, which create unrealistic connections in our graph. Therefore, we base the edges exclusively on within-sentence co-occurrences using the following approach. For each biography we collect all sentences that mention one or more

entities, not including occurrences of the biographee. Entities that occur in a sentence by themselves are connected only to the biographee, while entities that occur together in the same sentence are connected both to the biographee and to each other. Each occurrence increases the weight of the edge by one. We evaluate the quality of our graph by comparing it to a graph constructed directly from the original BWSA HTML pages. As previously mentioned, the BWSA editorial board manually adds hyperlinks between biographies. We can safely assume that the connections resulting from these links are valid, because of the expertise and diligence of the editors. As such, the total network of inter-biography hyperlinks can be seen as a gold standard of connections that need to be present in our graph for it to be an accurate representation of the described community. Besides this gold standard, we also generate a baseline graph model by scanning each biography for exact occurrences of names of BWSA community members and connect these to the person whose biography they occur in. The names are gathered from the accompanying BWSA database. For each person we include their full name (e.g. "Pieter Jelles Troelstra"), their last name (e.g. "Troelstra"), their last name with all initials (e.g. "P.J. Troelstra"), and their last name with the first initial (e.g. "P. Troelstra"). Occurrences of multi-referent last names are not included, since it is impossible to determine the correct referent using this technique.

The gold standard HTML graph model and the person-to-person graph constructed from our disambiguated data (from here on: NED P-P) are displayed in Figure 3.5 a and Figure 3.5 b, respectively. Nodes are sized in proportion to their *degree*, which is equal to the total number of edges attached to the node. The colours express the *modularity class* or community that a node belongs to (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). Modularity classes are calculated based on structural similarities between subparts of the graph, i.e. by grouping together nodes that have many connections in common. The layout of the graph is decided by a force-directed algorithm, which tries to minimize the number of edge crossings, thus grouping together nodes in a way that is aesthetically appealing and easy to interpret. At first glance, the graphs look very similar, both consisting of a large, highly interconnected group (right) and a smaller, less connected group (left). The similarity in shape is an indication that our method is successful at extracting the edges that also exist in the HTML graph. However, the larger node sizes and thicker edges reveal that the NED P-P graph is a lot denser than the gold standard. Furthermore, the communities seem to be distributed slightly differently across both graphs, which most logically implies that our method also finds edges that are not identified in the HTML.

We calculate several statistics to further investigate the (dis)similarities of our graph models, the results of which are listed in Table 3.6. In the HTML graph, we are able to connect 564 out of 573 biographees, with a total of 2,969 edges. The average degree is not much lower than its weighted counterpart. This reflects the limited number of hyperlinks that are added to the BWSA HTML and, consequently, the limited number of occurrences that we find for each connection.

	HTML	Baseline	NED P-P
nodes	573	573	573
nodes (degree > 0)	564	573	567
edges	2,969	4,800	3,350
average degree	10.4	16.8	12.3
average weighted degree	12.1	22.8	26.0
network diameter	7	4	6
communities	8	9	10
edge Precision		47.3	77.5
edge Recall		76.5	92.1
edge F1		58.4	84.2
degree ranking correlation		0.6799	0.9199
betweenness ranking correlation		0.5239	0.8423

Table 3.6 – Statistical comparison of the baseline and NED P-P graph models to the gold standard HTML graph.

The diameter of the graph is 7, which means that from any node in the graph it takes at maximum 7 steps to reach any other node. The baseline graph model is much denser than the HTML graph, with 4,800 edges that connect 100 % of the nodes. This results in a lower network diameter (4) and a higher average degree (16.8). We also see a great increase in the average weighted degree, which implies that more occurrences of the same edges are found using this technique. The NED P-P graph model connects 567 nodes with 3,530 edges. The average degree is closer to that of the HTML graph, while the average weighted degree is higher, as it is in the baseline graph. This means that our method finds more occurrences of edges that also exist in the gold standard than the baseline method, while it also finds less new edges than the baseline method. We confirm this by calculating Precision and Recall of unweighted edges for both graphs. The F1 score for NED P-P reaches 84.2, which is in line with our disambiguation results for the person class. The baseline F1 score is a mere 58.4, which is caused by the large number of new edges in this network (2,530). In comparison, the number of new edges in NED P-P totals 795. We cannot say if the new edges are valid or not from the statistics alone and without the proper domain knowledge. However, when we consider the HTML graph as a gold standard, it is fair to assume that it is a good representation of the structural composition of the actual network, and that any newly added edges should not alter that structural composition too rigorously for the graph to remain an accurate representation. We test structural similarity by calculating Spearman's rank correlation coefficient on the nodes ranked by unweighted degree, as well as *betweenness centrality*. The degree rank correlation measures to what extent the overall importance of nodes in the graph correlates. Betweenness centrality combines importance with (structural) position by incorporating measures taken over adjacent nodes. As such, it is a slightly more precise measure. Both the baseline and NED P-P graphs show significant correlations with the gold standard ($p < 0.00001$). Nevertheless, the correlations are considerably higher for our NED P-P graph, leading us to the conclusion that it is indeed the more accurate of the two.

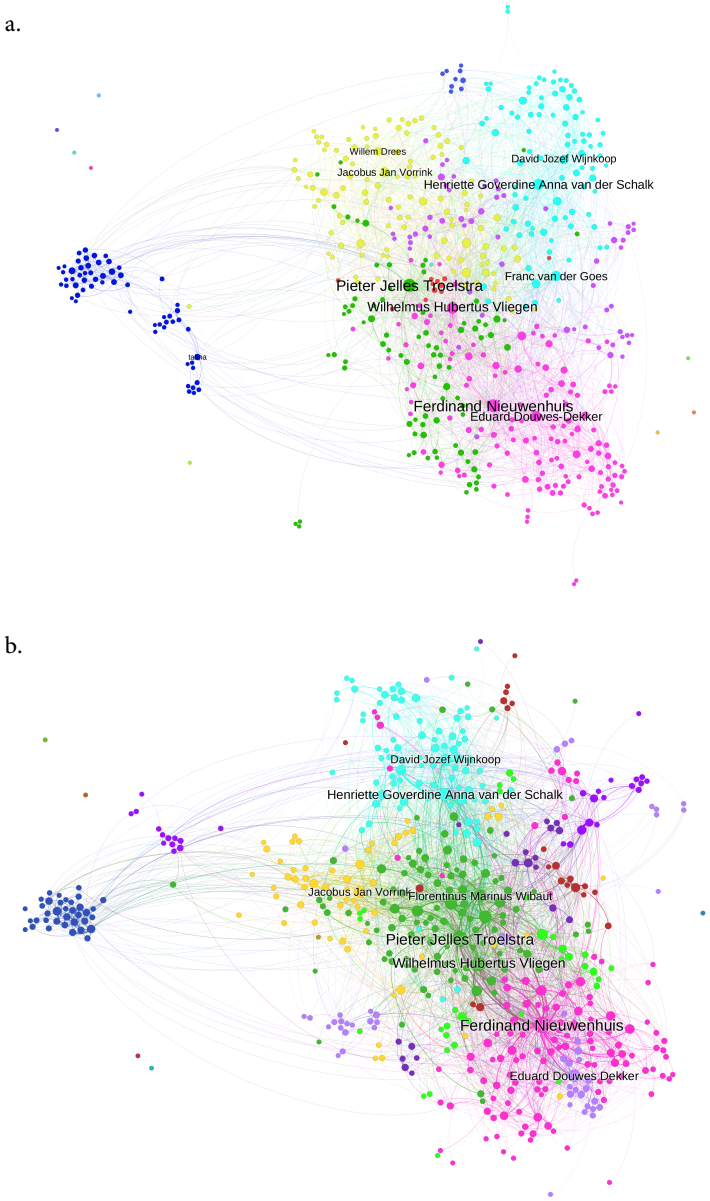


Figure 3.5 – a. Force-directed visualization of the links between the BWSA HTML pages. b. Force-directed visualization of the static social network of BWSA biographees constructed from sentence-level co-occurrences. In both graphs, the nodes are colored by modularity class to distinguish communities, and sized by degree (number of adjacent edges) as a crude measure of influence. Only the most prominent nodes are labeled.

4 MASTERING TIME

Who controls the past controls the future: who controls the present controls the past.
— George Orwell

After the identification of all nodes and edges for our graph we are able to view the accumulated social network of the BWSA as a static structure. This shows us who is connected to whom, but it reveals no information as to the period, duration, or order of the connections, which would help to expose the evolution and flow of ideas through the community. To be able to add this information and transform the network into a dynamic graph, we need to recognize and normalize all temporal expressions and events in the text. When we read a text or listen to a story, our mind processes numerous cues related to the ordering of events in time. These cues may be explicit in their reference, such as "1911" or "May 25th, 2013", or vague, such as "later" or "around the same time". We also process signals that order events relative to one another, as happens in the sentence "We had a beer *after* work". There may be other, more subtle hints related to order or causality which we, humans, have no problem processing and relating to our general knowledge of the world, while a computer will encounter problems deriving the actual timeline of events. Consider the following fragment:

(1) Mary's hair was wet. She went for a swim in the lake.

Knowledge of the fact that water makes things wet helps us to deduce quite easily that the event of Mary going for a swim must have started *before* the event of her hair getting wet. Without this piece of knowledge, one might mistakenly infer that Mary went for a swim *because* her hair was wet, placing the events in reverse order on the timeline. For our current purpose – the construction of a social graph that evolves over time – it is crucial that we place the events and their resulting connections in the correct order, as not to give a misrepresentation of history as it is described in the text.

Parts of this Chapter have previously been published in:
– Van de Camp, M., & Christiansen, H. (2013). Resolving relative time expressions in Dutch text with Constraint Handling Rules. In D. Duchier, & Y. Parmentier, Constraint Solving and Language Processing (pp. 166-177). Orléans, France: Springer.

In this chapter we investigate methods for the automatic recognition, normalization, and ordering of temporal expressions and events in Dutch biographical text to answer our second research question:

> **RQ 2** To what degree and level of specificity can we reliably recognize and normalize temporal information in Dutch biographical text using state-of-the-art techniques?

We discuss related research into temporal analysis in Section 4.1. Our methods for its subtasks are detailed in Section 4.2, followed by our experiments and results in Section 4.3. We conclude the Chapter with a discussion in Section 4.4.

4.1. Related research

Initially, research in Text Mining regarded temporal expressions in the same light as named entities. Their recognition was therefore generally included in the NER task. However, as more advanced systems required more sophisticated temporal analysis, the tasks became separated. Over the last decade or so, research into temporal text analysis has greatly expanded. A major catalyst for this expansion was the definition of the TimeML ISO-standard for temporal annotation (Pustejovsky, et al., 2003), which provides exact rules for the markup of temporal events in documents. It allowed for a more formal approach to temporal expression analysis in which systems can be easily evaluated and compared to one another. To facilitate this process and further research in this area, three iterations of challenges regarding temporal expressions were organized under the flag of SemEval[1], which provides a periodic evaluation of the computational semantic analysis domain in the form of focussed tasks. The temporal task is known as TempEval and has provided a host of baseline systems for the recognition, normalization, and ordering of events and temporal expressions in unstructured text. In this Subsection we introduce the TimeML annotation standard, followed by a brief overview of the TempEval tasks and their best performing systems. It should be noted that most of these systems are implemented only for English language newswire or newspaper text. We therefore provide a separate overview of research specifically related to temporal analysis in Dutch text.

4.1.1. TimeML

Many approaches to automated temporal text analysis involve a machine learning (ML) component, whether it be on its own, or in conjunction with a rule-based approach in a hybrid system. The training and evaluation of ML-based systems usually require some amount of data in which the targeted information has been consistently marked. This ensures that datasets are sufficiently consistent to allow a comparative evaluation, and that the annotations may easily be ported to different

[1] http://aclweb.org/aclwiki/index.php?title=SemEval_Portal

genres or languages. The most widely used scheme for temporal annotation is TimeML. It provides a broad framework to annotate events and temporal expressions in many languages, allowing for straightforward interpretation and evaluation of automatically induced temporal information. TimeML annotation is specifically designed to serve the following purposes (Pustejovsky, et al., 2003):

- anchoring of events to temporal expressions;
- temporal ordering of events, both within the same sentence and across sentence boundaries;
- interpretation of contextually underspecified temporal expressions, such as "two months later", or "yesterday";
- reasoning involving the duration of events.

TimeML uses four different entity classes, each with a specific role and meaning. Complex expressions can be represented through combinations of these base types, which can be summarized as follows:

TIMEX3: denotes a temporal expression. Its main attributes are *type* and *value*, with a possible modifier. Type can be of value TIME, DATE, or DURATION. The value is a normalized value of the temporal reference formatted according to specifications defined within TimeML. The modifier can be used to define imprecise temporal expressions by indicating their position relative to the given value (e.g. "before", "during", "at the end of").

EVENT: used to indicate linguistic expressions that refer to eventualities, which include both happenings and states. Events have three attributes: class, tense, and aspect. Class denotes whether the event refers to an action or state, both with and without intent (e.g. "to have" versus "to want"), plus several other types such as a reporting event (e.g. "say", "claim") or an aspectual event (e.g. "begin", "stop", "continue"). We should note that we make a distinction between these linguistic events and historical events. A historical event refers to a unique happening or process that took place at a certain point in time. A linguistic event is the part of a sentence that refers to the action described by the sentence, which in most cases is the main verb. From here on, we shall use the term *event* to refer to historical events, and *linguistic event* or EVENT to refer to the expressions marked by the EVENT tag.

SIGNAL: indicates words that express how the linguistic events and temporal expressions are related to each other. It is used for temporal prepositions (e.g. "before", "from", "until"), connectives (e.g. "when", "then"), subordinators (e.g. "if", "or"), quantifiers (e.g. "twice", "repeatedly"), and polarity shifting words (e.g. "no", "not", "none"), yet it does not include any attributes to actually define the type.

LINK: expresses one of three types of relations: a logical temporal relation between two linguistic events, or between a linguistic event and a temporal expression

(TLINK); a subordination relation between two linguistic events, or between a linguistic event and a signal (SLINK); an aspectual relation between two linguistic events (ALINK). We will focus only on the TLINK entity in this chapter. The types of relations that can be expressed by a TLINK are drawn from James Allen's interval logic (Allen, 1983). This logic contains 13 possible relations that may exist between two linguistic events, broadly falling into the categories before, during, after, and overlap (Table 4.3).

4.1.2. TempEval

TempEval is a platform for the periodic evaluation of the state-of-the-art in temporal analysis. It is organized as part of the SemEval workshop for Semantic Evaluation in the form of a predefined task to be performed on a supplied dataset. This type of controlled evaluation allows for a fair comparison of the available systems. The main goal of TempEval is to stimulate research into the recognition, normalization, and ordering of temporal expressions and events in text, and to provide a useful framework for the evaluation of such temporal analyses (Verhagen, Gaizauskas, Schilder, Hepple, Katz, & Pustejovsky, 2007; Verhagen, Saurí, Caselli, & Pustejovsky, 2010; UzZaman, Llorens, Allen, Derczynski, Verhagen, & Pustejovsky, 2012). So far, the task has been organized on three occasions (SemEval-2007, SemEval-2010, and SemEval-2013). Even though TempEval is directly inspired by the definition of TimeML, it does not employ the exact TimeML annotation scheme. The most notable differences are as follows:

- for TIMEX3 the type attribute can also be SET, which is applied to temporal expressions that refer to a repeated occurrence, such as "twice a week", or "every Friday". In TimeML this is included through the use of the SIGNAL entity, which in return is not included in the TempEval data;
- two extra attributes are encoded for EVENT: *modality* and *polarity*. Modality reflects the modal verb that governs the event expression. Polarity can either be positive or negative (binary) and refers to the affective polarity of the event. An extra binary attribute *mainevent* denotes whether the current event is the main event in the current sentence;
- regarding relation encoding, TempEval exclusively considers the TLINK category. SLINK and ALINK are not encoded.

In our method, we apply the modified TempEval annotation scheme, with one exception: we do not include TIME or SET as a class for TIMEX3. These classes have little significance in the context of biographies, where the ordering of events is more important than their exact timing. Instead, we add an extra class REFERENCE for the annotation and recognition of temporal expressions that refer to past or future dates.

Tasks

The tasks featured within TempEval have varied over the years, but always centered on the topics of recognition, normalization, and ordering. The full list of tasks specified for TempEval is the following:

A. Recognition and classification of temporal expressions
B. Recognition and classification of events
C. Relating events to temporal expressions in the same sentence
D. Relating events to the document creation time
E. Relating events in consecutive sentences to one another
F. Relating events from the same sentence between which a syntactic subordination relation exists

The first TempEval included only tasks C, D, and E, and provided datasets exclusively in English. The data was preprocessed to split the sentences and each sentence was manually annotated with TIMEX3, EVENT, and TLINK entities. TempEval-2 included all tasks and datasets in six languages, though participants only developed systems for English and Spanish. TempEval-3 consequently offered only datasets in these two languages. The most notable difference between TempEval-3 and its previous instalments is in the setup of the tasks. Participating teams could choose to implement a solution to only one of the tasks, or to solve all of them in a pipeline for complete temporal analysis.

Evaluation

Regarding the evaluation, TempEval and TempEval-2 implemented two approaches: strict and relaxed (Verhagen, Gaizauskas, Schilder, Hepple, Moszkowicz, & Pustejovsky, 2009). TempEval-3 only implemented the strict scoring method. The strict evaluation only regards an assigned label as correct if it is exactly the same as the label in the gold standard. The relaxed method allows for some fuzziness and also counts partial matches as correct, though it weighs the score depending on the overlap with the gold standard answer. For example, if the answer given by the system for a particular TLINK is *before*, while the gold standard answer is *before-or-overlap*, then under the relaxed evaluation method the answer is considered as being correct with a weight of 0.5, while the strict method considers it as false. For both scoring methods, precision and recall are calculated over all answers using the following equations:

$$Precision = R_c / R$$
$$Recall = R_c / K$$

where R_c is the number of correct answers, R is the number of answers in the system's output, and K is the number of answers in the gold standard key. The F1 metric calculated from these values is used to determine the final score for a system on a task (Van Rijsbergen, 1979).

Systems

Many systems have participated in the TempEval tasks using many different approaches. Here, we review a few of the best performing methods (WVALI, HeidelTime, TIPSem, and ClearTK-TimeML) and discuss their suitability for temporal analysis on the biographical genre.

The WVALI system displayed the best performance in the first instalment of the TempEval task (Puscasu, 2007). In this approach, temporal relations between events, between events and temporal expressions, and between events and the document creation time (DCT) are all determined using heuristics. Parameters used include tense and aspectual features for the events, the TimeML types of the constituents, syntactic relations between them, and the presence of temporal signal words. Puscasu (2007) breaks the process down into three steps. First, they derive temporal relations between constituents of a clause from the syntactic dependencies between the constituents and add this information to the parse tree of the clause. In a next step, each pair of clauses from the same sentence is related to one another based on the tenses of their main verbs and the dependency relations between the clauses. Finally, the temporal relation is determined based on propagation of the previously derived temporal relations between the constituents through the parse tree. WVALI's best scores for each of the tasks (C, D, E) are, respectively, 64%, 81%, and 64%.

Before determining the temporal relation between an EVENT-TIMEX3 pair, WVALI applies common-sense reasoning to test whether a certain temporal relation must logically exist between the entities based, among other things, on their relation to the DCT. For instance, they argue that if a TIMEX3 is classified as *before* the DCT and the verb tense of the EVENT is *future*, then the EVENT must be *after* the TIMEX3. While this assumption may generally hold for texts of a reporting genre, such as the newswire text in the TempEval corpus, it cannot be taken for granted when processing other genres of text. Especially in fiction one can make references to timelines that have no logical relation to the DCT whatsoever. In historical genres, such as the BWSA, literal quotes are often included, which cannot be interpreted from the perspective of the static DCT. In fact, the DCT seems completely irrelevant in these contexts, since the time of writing or publishing is of no consequence to the description of the events in the text. Moreover, in the case of digitized historical documents, the *digital* DCT differs from the *actual* DCT, complicating matters even further.

The HeidelTime system first participated in TempEval-2 (Strötgen & Gertz, 2010). It was the best performing system for the detection and normalization of temporal expressions (task A) with a score of 86%. It participated again in TempEval-3 for the same task. This time HeidelTime was the second best system with a score of 90.3%, only 0.02% below the best performing system. HeidelTime uses handcrafted rules, which are defined separately for each type of TIMEX3 entity (TIME, DATE, DURATION, and SET). The rules connect possible temporal expressions (matched

against regular expressions) with a normalization function and a TIMEX3 type label. The regular expressions are constructed using a combination of actual tokens, part of speech tags, and normalized values representing commonly occurring tokens, such as day and month names and numbers possibly representing a time, date, or year. The rules are first applied to the entire document. The normalization functions allow for varying degrees of specificity, which may result in the extraction of underspecified temporal expressions. For example, "mid 90s" is initially normalized to 'XX95-XX-XX', where day, month, and century are left undefined. A single mention of "June" is normalized to 'XXXX-06-XX'. In a next step, HeidelTime applies some basic reasoning to these cases, trying to resolve them by relating the expression to either the DCT, or to the other temporal expressions that occur in the document. If these do not provide sufficient information, the system looks for linguistic cues, such as verb tense, to further determine the value of the expression. Regular expressions are a powerful tool, but limited in their application, since only *exact* matches are processed. This requires the input text to be as clean as possible, without any spelling mistakes or OCR errors. Unfortunately, it is these types of errors that occur most often in historical, digitized data like the BWSA. Therefore, the HeidelTime system seems less appropriate for this genre.

The TIPSem system, which participated in all tasks of TempEval-2 for both English and Spanish, approaches temporal annotation as a supervised sequence labelling task using Conditional Random Fields (CRF) (Llorens, Saquete, & Navarro, 2010). The CRF method requires as input a matrix X of random variables over a sequence of data instances, plus a vector Y containing the categorical labels for the instances in sequence (Lafferty, McCallum, & Pereira, 2001). It then defines a conditional probability $p(Y|x)$ over label sequences Y given a particular observation sequence x, and assigns the label sequence y with the highest conditional probability. This approach seems highly appropriate for temporal analysis tasks, since the extent of linguistic events, temporal expressions, and their relations are dependent on the structural properties of the sentence. For English, TIPSem scored comparable to HeidelTime for task A (85%) and was the best performing system for the tasks of event recognition (B) and identification of event-DCT relations (D) with respective scores of 83% and 82%. It should be noted that the system achieved the lowest scores on English data of all systems participating in task C, relating events to temporal expressions in the same sentence. It scored a mere 54-55%, while all other systems scored between 62% and 65%. For the Spanish version of the same task, however, the system reached a score of 81%. This difference in performance could be attributed to a number of factors, including the quality of the two datasets and linguistic differences between Spanish and English. However, the exact cause of the difference in performance was not investigated by the researchers.

ClearTK-TimeML (Bethard, 2013) was one of the best systems participating in TempEval-3 for all English tasks. It placed first on the combined tasks, fully processing all temporal information from raw text input, yet with a low score of 30.98%. This underlines the complexity of the task at hand. ClearTK approaches the TIMEX3 and EVENT recognition tasks (A and B) both as sequence labelling

tasks, using features such as token, stem, part of speech, and word form for each token and its surrounding 6 tokens. The normalization of TIMEX3 values is handled using an existing rule-based system, TimeN (Llorens, Derczynski, Gaizauskas, & Saquete, 2012). All other tasks are dealt with in a multi-class classification setup, with varying, but straightforward, features. For the relation identification tasks, only EVENT-TIMEX3 pairs are considered where the path of syntactic relations and sub-/superordinations between them matched a predefined regular expression. This ensures a certain level of consistency in both the training and test instances, which expectedly improves a system's precision. Bethard (2013) compares performance between CRF classifiers, Maximum Entropy classifiers (MaxEnt), and Support Vector Machines (SVM). SVM classifiers are found to deliver the most robust performance, while also being easy to train.

4.1.3. Dutch temporal analysis

To get a full understanding of the options and challenges in analysing the temporal expressions in the BWSA, we also need to review research into Dutch temporal expressions. However, automated temporal analysis of Dutch text has not received much attention. Instead, most efforts in this field focus only on purely (cognitive) linguistic and logical aspects of expressing and processing temporal information (Maes & Oversteegen, 1999; Oversteegen, 2005). However, within the context of several large research projects (D-Coi, AMASS++, SoNaR) a linguistically motivated annotation scheme for temporal and geospatial information has been developed, which is named STEx (previously MiniSTEx) (Schuurman, 2008; Schuurman, Hoste, & Monachesi, 2010). STEx has been designed to be applicable to a wide range of genres and languages. Besides temporal expressions, STEx also covers *geospatial* expressions, and *geotemporal* expressions, which combine geospatial and temporal properties (i.e. historic place names). One key aspect of STEx is that it makes use of document metadata beyond the DCT in the annotations. For instance, if the origin of the text is known, the intended audience can be determined, which serves a role in the definition and interpretation of ambiguous references. Compared to TimeML, STEx provides a much broader framework. Expressions such as "winter" or "Christmas" are not covered in TimeML, while they are in STEx. The exact dates of seasons and holidays may differ across regions, but STEx solves this through the connection of the temporal and the geospatial information. The disambiguation of geospatial and geotemporal expressions is solved using a large database of entities. This somewhat hampers the applicability of STEx, since such databases are not always available. Because it has only recently been developed, STEx is not yet in wide use as an annotation scheme and has not been fully evaluated with respect to its applicability in an automated setup. We therefore do not consider STEx to be part of the state-of-the-art and as such do not judge it as a viable solution to our problem.

In a previous effort, we ourselves have attempted to resolve implicit and referential temporal expressions through the application of Constraint Handling Rules (CHR)

implemented in Prolog, which is a logical programming language (Van de Camp & Christiansen, 2013). CHR entails the definition of rules for the extraction and logical reformulation of constraints posed by the data itself, and tries to resolve these constraints into simpler ones that are computationally more comprehensible (Fruhwirth, 1998; Fruhwirth, 2011). For instance, if a biography refers to "the 20s" and the biographee lived from 1785 to 1840, then the temporal expression can logically be inferred to refer to the 1820s in consideration of the constraints posed by the biographee's lifespan. Intuitively, a logical approach to temporal analysis makes sense, if time is viewed (and referred to) as a linear process. However, references to time in text may not adhere to such a chronological order. Furthermore, the references themselves may appear in a multitude of forms, which are not always logically explained. In CHR, each surface form requires its own rule in order to be recognized correctly. The rules that need to be defined for acceptable coverage of the data thus quickly grow to an infeasible number. Moreover, the rules are applied to the data continually, until no further resolution is possible. In this process, the resolution of one expression may provide information for the resolution of another. Any errors that are made early in the process will consequently have a detrimental effect on the final result. In our application of the method, we find that the effort required to define the rules is too great compared to the poor performance of CHR on this particular task.

4.2. Method

Our current goal is to locate all information in the BWSA that pertains to the temporal ordering of the edges in our social network model. To this end, we at least need to recognize and normalize all temporal expressions in the text, as defined by TempEval task A. In an attempt to further enrich the graph, we also aim to extract all linguistic events and relate them to the temporal expressions, which is equivalent to TempEval tasks B and C. This will hopefully allow us to uncover some of the events that are pivotal in the forming, endurance, and dissolution of connections between the biographees of the BWSA. We apply both rule-based and machine learning (ML) methods to solve the tasks at hand. The data for our experiments is gathered by randomly selecting 100 biographies from the entire BWSA. From this we create three random subsets: a training set of 50 biographies (84,835 tokens), a development set of 30 biographies (45,970 tokens), and a test set of 20 biographies (28,136 tokens). The temporal information in these biographies is annotated by two independent human annotators (both male) without any linguistic background, but with good understanding of the task. All disagreements are resolved by a third annotator (female) with expert knowledge of linguistics and temporal analysis. Next, we describe our exact approaches to each of the three temporal entity classes, TIMEX3, EVENT, and TLINK.

4.2.1. TIMEX3

We approach the TIMEX3 recognition task as a straightforward sequence-labelling task according to the BIO scheme (Ramshaw & Marcus, 1995; Tjong Kim Sang & Veenstra, 1999). In this approach the tokens of a sentence are sequentially fed into the classifier, which then assigns one of the following three class labels: B-TIMEX3, I-TIMEX3, and O. These respectively imply that the token is the *beginning* of a TIMEX3 expression, that it is *inside* a TIMEX3, or that it is *outside* a TIMEX3. In our gold standard, 6,501 tokens out of 158,941 are found to be part of a temporal expression (4.1 %). Together they constitute 3,262 temporal expressions. Of these, 2,346 (71.9 %) are of type DATE, 781 (23.9 %) are of type DURATION, and 135 (4.1 %) are of type REFERENCE.

For the classification we use two different types of machine learning algorithms: a MATLAB implementation of Conditional Random Fields[2], and the *k*-nearest neighbours algorithm (*k*-NN) as implemented by MBT (Daelemans, Zavrel, Van den Bosch, & Van der Sloot, 1996). *k*-NN is a *lazy learning* algorithm, which means that no computation or modelling is done on the data until the moment of classification. To classify a new instance, *k*-NN compares it to its *k* most similar instances in the training set and assigns the majority class of this set of neighbours to the instance. MBT implements a slightly altered version of *k*-NN, namely *k*-nearest *distances*, in which the set of neighbours is comprised of all instances that are at one of the *k* shortest distances from the test instance in the feature space. Distance is weighted based on exact feature overlap, a metric that is also suitable for processing of categorical feature values. For each token in a sentence we include the following features: lemma, part of speech tag, named entity type (Chapter 3), and word form. We add the same features for the three preceding and the three following tokens to the feature vector. The feature vectors for the CRF classifier are numerically encoded before being fed into the classifier. The encoding is done through a numeric index of all unique non-numeric features in the entire dataset. We use the same types of classifiers for the TIMEX3 identification task, where all previously recognized temporal expressions are labelled with one of the following type labels: DATE, DURATION, or REFERENCE. For this task we again include lemma, part of speech tag, named entity type and word form for the current token and its six direct neighbours in the feature vector, supplemented with numeric representations of the date components included in the temporal expression. For instance, "November" is converted to 11, and "Monday" is converted to 1. We also run experiments with classifiers that try to detect and identify the TIMEX3 entities in one step. The concatenation of multiple classifiers commonly results in a decreased accuracy, since the errors made by one classifier are fed into the next. We aim to investigate whether the recognition and identification of temporal expressions is best approached as a 1-step, or a 2-step problem.

[2] http://www.cs.ubc.ca/~schmidtm/Software/crfChain.html

In order to be able to compare our ML systems to an intelligent baseline, we implemented a rule set for Dutch in HeidelTime.[3] In a first step, we translated all regular expressions already included in the German and English rule sets to Dutch. In a second step, we converted the normalization rules in the German set to match Dutch syntax. We choose the German set for this, since this language is syntactically more similar to Dutch, compared to English. In a third step, we checked the Dutch normalization rules against the rules for English to make sure all similar cases were covered by both sets. In total, 150 rules are defined (date: 83; duration: 26; set: 19; time: 22).

Regarding the type classification of TIMEX3 entities, we do not consider expressions of types TIME and SET. Biographies describe the flow of major events occurring during a timespan of usually several decades. Expressions referring to specific times are very unlikely to occur in this genre and when they do, the exact time of the event is of negligible consequence to the flow of events surrounding it. Similarly, sets rarely occur in the BWSA. Recurring events that do occur are mostly in the form of named events, such as “Pasen” (“Easter”) and “Kamerverkiezingen” (“parliamentary elections”). Moreover, we observe many occurrences of dates as part of a duration in the BWSA. For instance, when giving an overview of someone's professional career, the authors often choose a construction of the following form:

(2) "Vanaf 1920 was zij als eerste vrouw als waarnemend griffier verbonden aan het Amsterdams Gerechtshof. Van 1924 tot 1928 was Katz bestuurslid van de Nederlandsche Advocaten Vereeniging en de Internationale Vereeniging van Vrouwelijke Advocaten."

("From 1920 onward she was the first woman as Acting Registrar attached to the Amsterdam Court. From 1924 to 1928 Katz was a board member of the Dutch Lawyers Association and the International Association of Women Lawyers.")

According to the TimeML guidelines, "1920" in the first sentence should be annotated as DATE. However, the event described in the sentence is not restricted to a single date or year, but rather starts at the given date and lasts for an as of yet undetermined amount of time. Similarly, "1924" and "1928" should formally both be annotated as DATE, while they clearly indicate the boundaries of a duration. We choose to deviate from the TimeML annotation standard in this respect, and consider these instances as being of type DURATION. The annotators are also asked to mark whether a selected DURATION describes a starting point, an endpoint, or a full duration. The inter-annotator agreement (mutual F-score) for the TIMEX3 type attribute is 94.6 % over all classes, 94.7 % for DATE, 79.7 % for DURATION, and 72.1 % for REFERENCE. The agreement for the DURATION

[3] https://github.com/HeidelTime/heideltime/

subtypes is 86.7 % for starting points, 90.0 % for endpoints, and merely 60.5 % for full durations.

We interpret temporal expressions as describing a period or interval, as is done in James Allen's interval logic (Allen, 1983). He views time in terms of intervals that may or may not overlap, connect, or include one another. We motivate our choice for an interval-based logic with the following example:

(3a) John arrived at 7.12 PM.
(3b) John arrived in the evening.

Sentence 3a considers the event of John's arrival as a momentary happening of which the exact time is given. Sentence 3b is much more fuzzy in its indication of the actual moment that the event occurred, but gives us a timespan within which it happened. If time is treated as a sequence of moments or points, the timespan in sentence 3b could be represented by (one of) the points within the subset of points that fall between 6 PM and 12 AM. However, the event of arriving somewhere in itself can be further decomposed into smaller events. For instance, John might have travelled to his destination by car, in which case the arrival could include the parking of the car, turning off the engine, opening the door, stepping out of the car, closing the door, and so forth. If we would consider 7.12 PM to be one point, we would not be able to further represent its components if needed, resulting in a loss of information. An interval-based representation of time preserves all of the information that is given, and leaves room for the addition of information of a finer granularity if and when it becomes available. In this representation, "7 PM" and "the evening" can be easily compared to one another. With this in mind, we use two attributes to describe the TIMEX3 value: *start* and *end*.

To normalize a TIMEX3 to an interpretable *start* and *end* value, we apply a straightforward rule-based algorithm. For temporal expressions of types DATE and DURATION we match the string of the expression to several regular expressions in order to locate any possible references to a day, month, year, decade, or century. For the DATE type, the extracted values are assigned to the appropriate slot in both the start and end date of the TIMEX3. For instance, the string "24 september 1863" will return the values 24 (day), 11 (month), 3 (year), 6 (decade), and 18 (century) for both the *start* and *end* attributes, which are then combined into "1863-11-24". Missing values are replaced with zeroes, so a single mention of "1863" will be marked as "1863-00-00". Season names and signals such as "het begin van" ("the beginning of") are also parsed and converted to day and month values. For entities of type DURATION the context of the expression is then parsed in an effort to determine whether the expression denotes a starting point, an endpoint, or a full date. For instance, a starting point will most often be preceded by a preposition such as "van(af)" ("from") or "sinds" ("since"), and an endpoint will be preceded by "tot" ("until"). Full durations can be described in numerous ways, for example by a concatenation of two dates with a hyphen, fixed phrases such as "de periode" ("the

period") or "de jaren" ("the years") followed by a temporal expression, or the name of a season. If the TIMEX3 is determined to be a starting point, only the *start* attribute is assigned its extracted date, while the endpoint is set to "0000-00-00"; the reverse is done when the duration is found to be an endpoint.

Temporal expressions of type REFERENCE require a different approach. To be able to assign a value to a reference, we need to determine the distance value and unit described by the expression, and the referent, which is selected from the temporal expressions occurring before the referential expression. The referent is set to the *end* value or, if that is not defined, to the *start* value of the last occurring TIMEX3 that is part of the running text. This implies that temporal expressions which occur between brackets or quotes are not considered as candidates for reference, since they may not adhere to the chronology of the biography. The value of the REFERENCE entity is then calculated by adding or subtracting the distance described by the expression.

4.2.2. EVENT

The recognition and classification of EVENT entities is handled in a rule-based setting and depends largely on the syntactic analysis delivered by Frog (Van den Bosch, Busser, Canisius, & Daelemans, 2007). In fact, we use the verb and noun phrases detected by Frog's chunking module as the baseline for this experiment. For our own method, we depend on Frog's dependency parser (Canisius, Bogers, Van den Bosch, Geertzen, & Tjong Kim Sang, 2006). The rules are defined as follows. From the dependency parse of a sentence, we first select all finite verbs. Next, we locate any verbal complements that are dependent on the selected finite verb. We then check for any reflexive pronouns dependent on the verb phrase and, if the verbal complement is an infinitive, we also check for a dependency relation to the word "te", which can be considered the Dutch equivalent of the English "to be" when it is followed by an infinitive (e.g. "to be standing"). Next, we look for verbal particles and, finally, we check for the occurrence of a direct dependency to or from any of the negation words "niet" ("not"), "niets" ("nothing"), "geen" ("none"), "nooit" and "nimmer" (both: "never"). The complete phrase is then marked as an event with a binary attribute *negated*. As is done in TempEval, we add an attribute indicating whether the event is the current clause's *main event*. Additionally, we add two event type markers based on the main verb in the event phrase. The first type is described by one of the following labels: "alteratie" ("alteration"), "actie" ("action"), "gebeurtenis" ("happening"), and "verloop" ("course"). The label is determined by querying Cornetto (Vossen, et al., 2013), the Dutch equivalent of WordNet, for a chain of hyperonymy relations between the input verb and the four verbs corresponding to the type labels: "wijzigen" ("to alter"), "handelen" ("to act"), "gebeuren" ("to happen"), and "gaan" ("to go on") in that order. The label corresponding to the first verb with which such a path of relations is found is set as the event's *main type*. The *subtype* is set to the first hyponym of the main type verb in the path to the input verb. This allows us to later make generalizations over the

events on three levels: the main type, the subtype, and the main verb, thus synthesizing a hierarchical taxonomic perspective on the data. Further attributes for the event include its *tense*, its *modality*, and its *aspect*, the last two of which are determined by the occurrence of any of the auxiliary verbs listed in Table 4.2. We repeat this process for any remaining infinitives and past participles occurring in the sentence. In total, 3,729 unique verbs occur in the BWSA.

Nouns referring to events are selected using Cornetto. All nouns occurring in the BWSA are compared to the words "ontwikkelingsgang" ("development"), "alteratie" ("change"), "evenement" ("event"), and "handeling" ("action"). These

	Auxiliary verb	**Translation**
modality	kunnen	can
	moeten	must
	mogen	may
	hoeven	should
	weten	know
	willen	want
	lijken	seem
	blijken	turn out
	schijnen	appear
	voorkomen	appear
aspect	beginnen	begin
	stoppen	stop
	ophouden	quit
	blijven	keep (on)
	gaan	go, as in "I'm going fishing"
	komen	come, as in "to come to pass"
	lopen	walk*
	staan	stand*
	liggen	lie*
	zitten	sit*
		* in the sense of "being"

Table 4.1 - Overview of modal and aspectual auxiliary verbs occurring in the BWSA

noun	**translation**	**frequency**
functie	function / position	496
oorlog	war	443
congres	conference	425
strijd	conflict	345
raad	advice	343
leiding	leadership	328
vergadering	meeting	315
dienst	service	313
ontwikkeling	development	280
actie	action	270

Table 4.2 – The ten most frequently occurring event nouns in the BWSA

Rule-based EVENT extraction

In 1861 begon zij zich voor te bereiden op het onderwijzersexamen dat zij wegens langdurige ziekte van 1864 tot 1869 niet heeft afgelegd.

(Translation: "In 1861 she began to prepare for the teachers exam which she did not complete due to prolonged illness from 1864 to 1869.")

finite verbs: In 1861 **begon** zij zich voor te bereiden op het onderwijzersexamen dat zij wegens langdurige ziekte van 1864 tot 1869 niet **heeft** afgelegd.

verbal complements: In 1861 **begon** zij zich voor te $_{vc}$**bereiden** op het onderwijzersexamen dat zij wegens langdurige ziekte van 1864 tot 1869 niet **heeft** $_{vc}$**afgelegd**.

reflexive pronouns: In 1861 **begon** zij $_{obj1/se}$**zich** voor te **bereiden** op het onderwijzersexamen dat zij wegens langdurige ziekte van 1864 tot 1869 niet **heeft afgelegd**.

infinitive completion: In 1861 **begon** zij **zich** voor $_{none}$**te bereiden** op het onderwijzersexamen dat zij wegens langdurige ziekte van 1864 tot 1869 niet **heeft afgelegd**.

verbal particles: In 1861 **begon** zij **zich** $_{svp}$**voor te bereiden** op het onderwijzersexamen dat zij wegens langdurige ziekte van 1864 tot 1869 niet **heeft afgelegd**.

negation: In 1861 **begon** zij **zich voor te bereiden** op het onderwijzersexamen dat zij wegens langdurige ziekte van 1864 tot 1869 $_{mod}$**niet heeft afgelegd**.

noun events: In 1861 **begon** zij **zich voor te bereiden** op het **onderwijzersexamen** dat zij wegens langdurige **ziekte** van 1864 tot 1869 **niet heeft afgelegd**.

extracted EVENTs:

EVENT 1:	*begon zich voor te bereiden* mainevent = 1 aspect = *beginnen* main type = *actie* subtype = *werken* ("to work")	**EVENT 2:**	*heeft niet afgelegd* negated = 1 main type = *actie* subtype = *werken*
EVENT 3:	*onderwijzersexamen* main type = *actie* subtype = *experiment*	**EVENT 4:**	*ziekte* main type = *alteratie* subtype = *proces*

Figure 4.1 - Example of the rule-based EVENT extraction process applied to one sentence of the BWSA

words are not among the most commonly used words in modern Dutch. They are chosen because of their high position in the tree of semantic relations within Cornetto, which implies a great coverage over the target words. If the first sense of the input noun is a hyponym of the first sense of any of these four words in the Cornetto database, then the noun is considered to refer to an event. We filter the nouns to keep only those with 25 or more occurrences in the BWSA, resulting in a list of 266 unique nouns. After manual inspection, 85 nouns are removed that do not refer to an event, which include three of the most common nouns in the

collection: “partij” (“[political] party”), “organisatie” (“organisation”), and “beweging” (“movement”). Although the latter two strictly speaking do refer to events - namely, the events of organising and moving - within the BWSA “organisation” generally refers to a previously mentioned named entity of the same type, while “movement” refers to the overall topic of the collection, the labour movement, or any of its submovements. Table 4.3 list the top ten occurring event nouns.

The success of this approach to event detection is principally dependent on the performance of Frog’s parts-of-speech tagger and dependency parser, and the integrity of the lexical relations in Cornetto. However, the method itself is applicable to text of any genre or domain without any adaptations. We evaluate our approach on two sets of 100 manually annotated sentences, one originating from the BWSA, the other from the BD98 contemporary news corpus previously introduced in Chapter 3. The sentences in the BWSA set are selected based on their expected complexity with respect to the current task, which is measured by their verb to token ratio. The higher the ratio, the higher the expected complexity. The sentences in the training, development, and test sets used for the TIMEX3 tasks are ordered by their complexity in descending order. Sentences that do not contain a TIMEX3 are filtered out. From the remaining sentences, the top 100 is selected for this task. The BD98 sentences are selected at random and may or may not contain a TIMEX3 entity. This is done to test how the method performs on less complex sentences. The BWSA set is randomly split into two equal parts, one of which is used to model the event extraction process. The remainder of the BWSA set, and the BD98 set are used for testing.

4.2.3. TLINK

Since the network based on this analysis is aimed to be a tool for reliable historic investigation, we want to maximize its accuracy. We choose to restrict the temporal relation analysis to only those same-sentence TIMEX3-EVENT pairs between which a dependency relation exists, under the assumption that these explicitly dated events represent the most significant events from a historic perspective. Further analysis of the relations between same-sentence events can be achieved through analysis of the sub- and superordination relations between the tokens of the sentence. However, we choose not to apply such a technique here as to not decrease the reliability of the outcome too much.

The type of the TIMEX3 determines the relation label that is attached to the TLINK entity. The labels are selected from the 13 relations in Allen’s logic, which are described in Table 4.3. The label is determined as follows:

- **DATE:** if an event and a temporal expression referring to a specific date, month, or year are connected through a dependency relation, then the event

most likely occurred during the timespan described by the TIMEX3. The TLINK is therefore labeled with the relation "equal";
DURATION: if the duration includes both a *start* and *end* value, then the event could occur somewhere within this timespan, or it could be ongoing during the entire timespan. We labels these connections with the label "EduringT", which in our interpretation spans Allen's classes "equal", "EstartsT", "EendsT", and "EduringT". Thus the original fuzziness of the expression is retained. When only the *start* value is defined, the event is judged to have started at the given point in time. If the next occurring TIMEX3 is of type DURATION and only has an *end* value, then this date is considered to be the end point of the event. In this case two TLINK entities are added, the first with relation "TstartsE", the second with relation "TendsE". If no endpoint is found, then the event, for the time being, is considered to last for the remainder of the biographee's lifetime. Similarly, if an event is connected to a duration endpoint, then it is checked whether the previous TIMEX3 is the start of a duration that can serve as the event's starting point. If it is not found, the event is considered to have started at the biographee's birth;

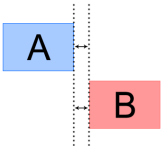
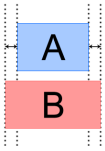
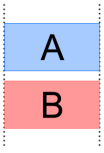
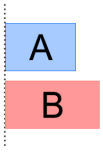
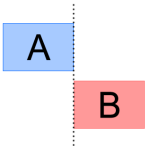
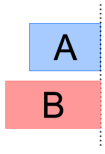
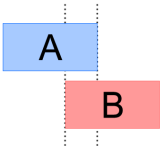
Relation	Illustration	Relation	Illustration
A before B		*A during B*	
A equals B		*A starts B*	
A meets B		*A finishes B*	
A overlaps B			

Table 4.3 - Relations according to Allen's logic

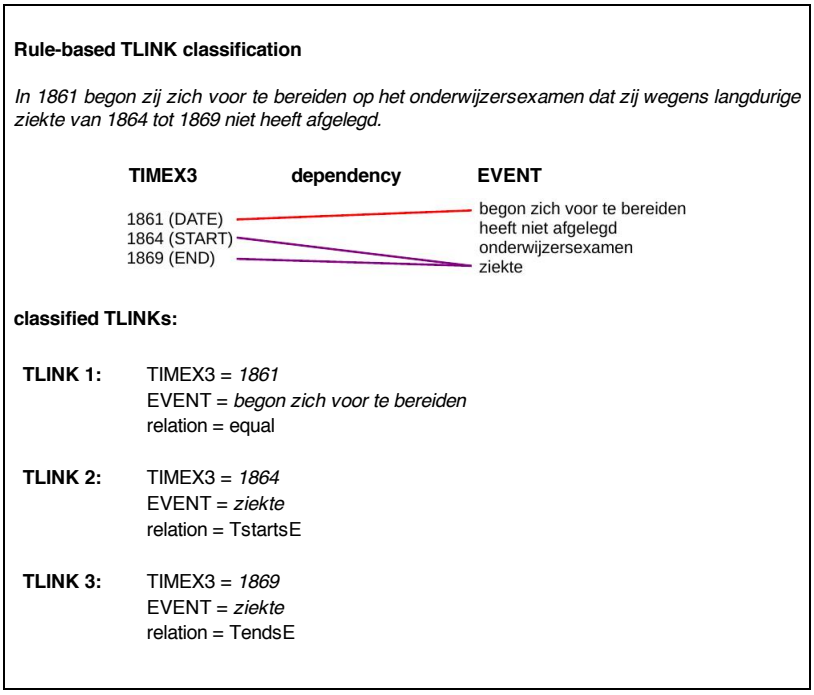


Figure 4.2 - Example of the rule-based TLINK classification process applied to one sentence of the BWSA

- **REFERENCE:** since references are calculated as if they refer to dates rather than durations (though technically, this is a possibility), TLINKs between events and references are also labeled with the relation "equals".

In some cases, an EVENT is dependent on multiple TIMEX3 entities. When this happens, we check the types and values of the two closest TIMEX3 entities and apply the following rules:

- if one of the TIMEX3 entities is of type REFERENCE or contains a DURATION without specified values, then the EVENT is linked to only the non-referential, specified TIMEX3;

- if both TIMEX3 entities are of type DATE, then we add two TLINKs: one between each TIMEX3 and the EVENT, thus creating a repeated event;
- if one of the TIMEX3s is of type DATE, and the other is of type DURATION with only a starting or endpoint, then the DATE value is set as

set	classifier	relaxed F1	strict F1
training	*HeidelTime*	87.7 (± 2.2)	84.6 (± 2.4)
	MBT-k1	91.6 (± 2.3)*	90.2 (± 2.5)*
	CRF	**94.0 (± 1.5)****	**92.5 (± 1.8)****
dev	*HeidelTime*	89.7	87.6
	MBT-k1	91.8	90.4
	MBT-k3	92.7	91.3
	MBT-k5	92.6	91.2
	CRF	**93.9**	**92.5**
test	*HeidelTime*	87.6	85.1
	CRF	**94.7**	**93.5**

Table 4.4 – F1 scores for TIMEX3 recognition. Scores reported for the training set are averages taken over ten 10-fold cross validation experiments. Scores marked with * are significantly better than the lowest scoring system ($p < 0.0005$). Scores marked with ** are significantly better than all other systems ($p < 0.0005$). The scores for the development and test sets are F1 scores for a single run on the entire set. The highest score for each class in each set is displayed in bold. For the test set we only report the scores of the baseline and best performing system.

set	classifier	DATE	DURATION	REFERENCE
training	*HeidelTime*	80.9 (± 4.2)	8.8 (± 6.7)	6.4 (± 7.1)
	MBT-k1	88.3 (± 2.9)*	59.1 (± 10.9)*	73.2 (± 14.6)*
	CRF	**94.0 (± 1.7)****	**78.4 (± 6.8)****	**77.8 (± 15.2)****
dev	*HeidelTime*	82.7	12.6	11.5
	MBT-k1	89.1	55.8	60.2
	MBT-k3	89.0	53.6	67.1
	MBT-k5	88.7	50.3	67.5
	CRF	**95.9**	**83.0**	**67.9**
test	*HeidelTime*	78.9	10.8	0.0
	MBT-k3	89.0	63.2	**64.2**
	CRF	**94.5**	**82.3**	61.2

Table 4.5 – F1 scores for 2-step TIMEX3 identification. Scores reported for the training set are averages taken over ten 10-fold cross validation experiments. Scores marked with * are significantly better than the lowest scoring system ($p < 0.005$). Scores marked with ** are significantly better than all other systems ($p < 0.005$). The scores for the development and test sets are F1 scores for a single run on the entire set. The highest score for each class in each set is displayed in bold.

classifier	evaluation	DATE	DURATION	REFERENCE
HeidelTime	*relaxed F1*	81.8 (± 3.8)	27.7 (± 20.4)	5.7 (± 6.5)
	strict F1	81.5 (± 3.7)*	19.7 (± 19.6)	0.0 (± 0.0)
MBT-k1	*relaxed F1*	82.4 (± 3.7)	66.5 (± 8.7)*	**73.2 (± 13.0)***
	strict F1	79.9 (± 3.5)	53.6 (± 11.1)*	70.3 (± 13.1)*
CRF	*relaxed F1*	**90.3 (± 2.2)****	**80.5 (± 6.3)****	70.4 (± 16.9)*
	strict F1	90.2 (± 2.2)**	76.3 (± 7.4)**	67.6 (± 16.4)*

Table 4.6 – F1 scores for 1-step TIMEX3 recognition and identification training. Scores reported are averages taken over ten 10-fold cross validation experiments on the training set. Scores marked with * are significantly better than the lowest scoring system ($p < 0.005$). Scores marked with ** are significantly better than all other systems ($p < 0.005$). The highest score for each class is displayed in bold.

- the missing value of the DURATION entity and the EVENT is linked to the entire resulting timespan;
- if one of the TIMEX3s is of type DATE, and the other is of type DURATION with both a start and end value specified, then the EVENT is linked only to the DURATION;
- if both TIMEX3 entities are of type DURATION, but only one of them has a start and end value, then the EVENT is linked only to the fully specified TIMEX3;
- if both TIMEX3 entities are of type DURATION, and both have a start and end value, then the EVENT is linked only to the closest fully specified TIMEX3.

4.3. Results

In this Section we detail our experimental results with regards to temporal analysis of the BWSA.

4.3.1. TIMEX3

The results for the TIMEX3 recognition task are displayed in Table 4.4. All reported results are F1 scores. We calculate a *relaxed* and a *strict* score using the same formulas as used in TempEval (Section 4.1.2). The HeidelTime rule-based baseline performs comparable to the English version of HeidelTime, which competed in TempEval 2 and 3. The relative ease with which new rules are implemented in this system combined with its high recognition accuracy underlines the competitive strength of the HeidelTime implementation. For the experiments on the training set, MBT is used with k set to the default value 1. This means that each token receives the class label that is attached to the majority of the instances at only the lowest distance from the input instance. A measure such as this is rather crude and therefore becomes less suitable as the complexity of the task - and thus the complexity of the partition of different classes in the feature space - increases. However, it performs significantly better than HeidelTime, scoring almost 6 points higher in the strict evaluation. The CRF classifier performs significantly better than both the HeidelTime baseline and MBT. Since CRF does not just take into account the current feature vector, but rather processes them in blocks, the structural composition of the context of a TIMEX3 entity plays a bigger role in the classification. This makes CRF more suitable for complex problems such as temporal analysis. We also see that the CRF results are slightly more consistent, judging from the lower standard deviations for this classifier. Next, we run the same classifiers on the 30 biographies in the development set. We use MBT with three different values for k, to investigate whether increasing the search space increases the performance. Increasing the value of k does lead to a better classification, though it still does not reach the same levels of accuracy as the CRF classifier. We therefore only tested the latter on the test set, which results in a score 7 to 8 points over the baseline.

classifier	evaluation	DATE	DURATION	REFERENCE
HeidelTime	*relaxed F1*	82.9	40.5	9.1
	strict F1	82.7	34.8	0.0
MBT-k1	*relaxed F1*	84.7	69.8	57.0
	strict F1	83.8	42.8	50.0
MBT-k3	*relaxed F1*	84.7	69.8	57.0
	strict F1	83.8	42.8	50.0
MBT-k5	*relaxed F1*	83.9	65.7	55.7
	strict F1	83.0	32.3	51.0
CRF	*relaxed F1*	**91.4**	**85.1**	**61.2**
	strict F1	90.9	80.0	56.5

Table 4.7 – F1 scores for 1-step TIMEX3 recognition and identification on the development set. The highest score for each class is displayed in bold.

classifier	evaluation	DATE	DURATION	REFERENCE
HeidelTime	*relaxed F1*	78.8	38.5	0.0
	strict F1	78.7	32.6	0.0
CRF	*relaxed F1*	**90.5**	**86.9**	**63.4**
	strict F1	90.4	82.7	59.3

Table 4.8 – F1 scores for 1-step TIMEX3 recognition and identification on the test set. The highest score for each class is displayed in bold. We only report the scores of the baseline and best performing system.

We run two sets of experiments to determine the type labels for the TIMEX3 entities. In the first setup, the output of the TIMEX3 recognition process is used as input for the type classification (Table 4.5). In the second setup, the input is untagged and the classifier has to perform recognition and identification in a single step (Tables 4.7 to 4.9). What immediately stands out in both setups, is that the HeidelTime baseline scores extremely low on the DURATION and REFERENCE classes. For the DURATION class, this is explained by our different interpretation of what constitutes a duration. HeidelTime conforms to the TimeML guidelines in this respect, and marks a date as DATE even if it implies the start or end of a duration. For the REFERENCE class, HeidelTime also assigns the DATE label as type and stores the reference classification as the value of that TIMEX3. In our setup, these are converted to entities of type REFERENCE with an undefined value before the score is calculated. Still, the score of this category is much lower than it is for the ML-based approaches. This is explained by the low number of REFERENCE entities that actually occur compared to DATE entities. A wrong classification of a single entity will consequently have a great effect on the score. This is also reflected by the comparatively large standard deviations measured on the results of MBT and CRF for these smaller classes.

Overall, CRF delivers the best performance in the 2-step setup. Only for the REFERENCE category does MBT outperform CRF on the test set, though the increase is a mere 3 points. In contrast to timex recognition, increasing the value of *k* decreases the quality of MBT's output on the DATE and DURATION classes, while it again increases performance on the REFERENCE class. In the 1-step classification setup, HeidelTime performs slightly better, especially on the

DURATION class. MBT and CRF also reach a higher score for this class, though only when calculating the relaxed F1 score. The *strict* score is always lower than the score in the 2-step setup. This implies that the extent of the temporal expressions is more often incorrectly recognized, while the attached type label is correct, and indicates that the division of TIMEX3 recognition and identification over separate classifiers is indeed most suitable for this complex task.

The results of the TIMEX3 normalization task are listed in Tables 4.10 and 4.11. The scores produced by HeidelTime are used as a baseline for this experiment. As a consequence of the way in which DURATION and REFERENCE expressions are processed by this system, no exact interval or date can be derived for these entities. As explained before, references are actually marked as DATE and given a value denoting only its referential character, not its date value. Similarly, durations are only labelled with their derived length, but are not linked to the timeline with a start and/or end value. The instances where HeidelTime does give the correct value are instances where no value could be assigned even by the human annotators and, therefore, the start and end dates are set to all zeroes.

Our own rule-based method performs comparable to HeidelTime for the DATE class. In most cases, for a DATE entity the start and end date are the same. The only instances where this is not the case are those where the DATE refers to a year and contains a modifier such as "the beginning of" or "the summer of", in which case the dates will be further defined to match the boundaries of the described period. For regular dates, the end date is always incorrect if the start date is incorrect. Since there are no recorded instances where the end date is correctly defined, while the start date is incorrect, we can conclude that this holds for both the regular dates and the modified years. If we only take into account the predicted year value of the start and end dates, the HeidelTime baseline outperforms our implementation, scoring 2 points higher for the start date and 0.7 points higher for the end value. Comparing the scores on the full dates versus the scores on the year values for both methods, we see that the partial matches from our method always contain a correct year value. This is not the case for the start dates predicted by HeidelTime. Here, half of the partially matching start dates actually contain an incorrect year.

For both the DURATION and REFERENCE classes, our algorithm is more successful in normalizing start dates than it is in defining end dates. Considering the entire date, duration start dates are completely correct in 26.4 % of the cases, versus 18.8 % for end dates. For the REFERENCE class these scores are extremely low: 1.3 % and 0 %, respectively. Looking only at the year of the date, the scores improve tremendously. In this respect, the scores for DURATION are 66.8 % for the start year and 36.0 % for the end year. For REFERENCE the scores are 39.5 % versus 4.1 %. Despite the improvement, these scores imply that the values of the DURATION and REFERENCE classes in most cases will be incorrectly defined. The results will therefore be highly unreliable and unusable in a social historical context.

4.3.2. EVENT

The results for the EVENT extraction process are listed in Table 4.11. We again calculate a *strict* and a *relaxed* F1 measure, this time for each sentence. For the *strict* measure this means that an event is counted as a *true positive* only if it contains all of, and no more than the tokens of the same event in the gold standard, and as a *false negative* otherwise. Events that only occur in the gold standard are also counted as false negative. Events recognized by the classifier that do not occur in the gold standard are counted as *false positives*. In the *relaxed* evaluation, a partly overlapping event is not counted as a false negative, but as a weighted true positive. The weight is calculated by dividing the number of tokens in the overlap by the total number of unique tokens in the gold standard event and the classified event combined. Next, we calculate precision and recall values for each sentence, from which we then calculate the sentence F1 scores. The final F1 scores are calculated by taking the average of the F1 scores of all sentences in the test set. The "correct" column shows the number of sentences in which all recognized events are completely correct compared to the gold standard annotations.

The results show that the verb phrases recognized by Frog's chunker have very low accuracy, which is especially apparent when looking at the number of completely correct sentences. The rule-based method executed on the dependency parse on average scores 25 to 30 points higher. Since we calculate the F1 scores on such small units as sentences and then average these, the variation of the scores is quite high, reflected by the large standard deviations. The performance on the test sets is somewhat lower than performance on the development set. This suggests that some degree of overfitting exists between the development set and the rule-based algorithm. The BWSA sentences were selected based on their expected complexity measured in verb-token ratio, under the assumption that complex sentences are more likely to result in an incorrect dependency parse and thus provide more of a challenge for the event extraction process. The sentences in the BD98 set were chosen at random. In light of our assumption, the scores on the BD98 set are expected to be equal to, or higher than the scores on the BWSA sets, while in fact they are lower. This is partly explained by the fact that the list of noun events used is based on frequencies gathered from the BWSA. Upon inspection it is revealed that the BD98 sentences include much shorter sentences, mere statements even, which do not always contain a verb. It is at these structurally straightforward phrases where the dependency parser actually fails most.

4.3.3. TLINK

During the annotation process, we did not collect information regarding the temporal relations between events and temporal expressions, as to not overwhelm our annotators. As a result, we are unable to perform a quantitative evaluation of the TLINK classification process at this stage. Instead we provide a qualitative evaluation using examples from the output of the process.

			end date			end year		total
			incorrect	partial	correct	incorrect	correct	
DATE	start date	incorrect	0.5 %	0.0 %	0.0 %			0.5 %
		partial	1.6 %	0.4 %	0.0 %			2.0 %
		correct	1.7 %	0.2 %	95.6 %			97.5 %
	start year	incorrect				1.0 %	0.0 %	1.0 %
		correct				2.8 %	96.2 %	99.0 %
	total		3.8 %	0.6 %	95.6 %	3.8 %	96.2 %	
DURATION	start date	incorrect	96.8 %	0.0 %	0.0 %			96.8 %
		partial	0.0 %	0.0 %	0.0 %			0.0 %
		correct	0.0 %	0.0 %	3.2 %			3.2 %
	start year	incorrect				96.8 %	0.0 %	96.8 %
		correct				0.0 %	3.2 %	3.2 %
	total		96.8 %	0.0 %	3.2 %	96.8 %	3.2 %	
REFERENCE	start date	incorrect	93.9 %	0.0 %	0.0 %			93.9 %
		partial	0.0 %	0.0 %	0.0 %			0.0 %
		correct	0.0 %	0.0 %	6.1 %			6.1 %
	start year	incorrect				93.9 %	0.0 %	93.9 %
		correct				0.0 %	6.1 %	6.1 %
	total		93.9 %	0.0 %	6.1 %	93.9 %	6.1 %	

Table 4.9 – Normalization baselines (HeidelTime) per TIMEX3 class

			end date			end year		total
			incorrect	partial	correct	incorrect	correct	
DATE	start date	incorrect	3.0 %	0.0 %	0.0 %			3.0 %
		partial	0.4 %	1.0 %	0.0 %			1.4 %
		correct	0.1 %	0.1 %	95.4 %			95.6 %
	start year	incorrect				3.0 %	0.0 %	3.0 %
		correct				0.5 %	95.5 %	96.0 %
	total		3.5 %	1.1 %	95.4 %	3.5 %	95.5 %	
DURATION	start date	incorrect	15.9 %	14.0 %	3.1 %			33.0 %
		partial	35.6 %	3.7 %	1.3 %			40.6 %
		correct	11.6 %	0.4 %	14.4 %			26.4 %
	start year	incorrect				16.2 %	17.0 %	33.2 %
		correct				47.8 %	19.0 %	66.8 %
	total		63.1 %	18.1 %	18.8 %	64.0 %	36.0 %	
REFERENCE	start date	incorrect	60.2 %	0.3 %	0.0 %			60.5 %
		partial	34.4 %	3.8 %	0.0 %			38.2 %
		correct	1.3 %	0.0 %	0.0 %			1.3 %
	start year	incorrect				60.2 %	0.3 %	60.5 %
		correct				35.7 %	3.8 %	39.5 %
	total		95.9 %	4.1 %	0.0 %	95.9 %	4.1 %	

Table 4.10 – Normalization results per TIMEX3 class

		Chunk-based extraction		Rule-based extraction	
		F1	correct	F1	correct
BWSA-dev	*relaxed*	66.5 (± 24.4)		94.8 (± 9.0)	
	strict	52.9 (± 29.5)	3 / 50	85.7 (± 16.4)	22 / 50
BWSA-test	*relaxed*	66.3 (± 23.4)		90.3 (± 15.7)	
	strict	50.6 (± 31.9)	1 / 50	76.1 (± 26.2)	19 / 50
BD98-test	*relaxed*	54.1 (± 40.7)		88.0 (± 25.8)	
	strict	37.5 (± 40.6)	12 / 100	62.4 (± 41.4)	34 / 100

Table 4.11 – Average F1 scores on the development and test sets for event extraction

Table 4.13 shows some statistics measured on the analysis of the 100 biographies that were annotated for the TIMEX3 extraction, and on the unannotated remainder of the BWSA. The numbers for both partitions are very similar, from which we can conclude that the overall quality of the temporal analysis process is sufficiently consistent throughout the corpus. On average, over 75% of the recognized temporal expressions is linked to an EVENT through a TLINK entity, which is encouraging, since it means that we will be able to capture approximately that amount of the described timeline in our final product: the dynamic social graph. In contrast, only 18% of the EVENT entities are linked to a TIMEX3. This is due to the large number of events compared to temporal expressions, combined with the fact that we limit the TLINK analysis to only those EVENT-TIMEX3 pairs between which a dependency relation exists. The average number of entities to which a TIMEX3 is linked by a TLINK is slightly higher than it is for EVENTs. This implies that there are more sentences in the BWSA that connect multiple eventualities to the same point (or interval) in time, than sentences that connect a single event to multiple timespans (i.e. recurring events). Regarding the TLINKs, the unannotated set contains far less links with the relations "TstartsE" and "TendsE". Further investigation reveals that these mostly get labeled as "EduringT" as a result of an incorrect value attributed to the linked TIMEX3. For instance, a DURATION starting point is required to label a TLINK with relation "TstartsE". If the DURATION is not recognized as a starting point, but rather considered to be a full duration, then the relation is automatically set to "EduringT".

When inspecting the output, we see that sentences where the entities of a related EVENT-TIMEX3 pair occur far apart from each other often lead to an incorrect analysis. This is illustrated by Example 4. In the English translation, the TIMEX3 and EVENT appear right next to each other, while in the Dutch sentence they are separated by the object of the EVENT and a subordinate clause. The complex structure of the sentence is not correctly recognized by the dependency parser, which is why the TLINK analysis fails.

(4) Hij $_{\text{EVENT\_1}}$[was] trots op zijn grootvader van moederskant, professor Wopko Cnoop Koopmans, hoogleraar aan het Doopsgezind Seminarie te Amsterdam, die daar in $_{\text{TIMEX3\_1}}$[1848] een brochure Algemeen Stemregt behoudens maatschappelijke orde, samen met de hoogleraar in de

vaderlandse geschiedenis aan het Atheneum Illustre Hugo Beijerman, EVENT_2[publiceerde].

(He EVENT_1[was] proud of his maternal grandfather, Professor Wopko Cnoop Koopmans, professor at the Baptist Seminary in Amsterdam, who in TIMEX3_1[1848] EVENT_2[published] a brochure Suffrage under condition of social order there, together with the professor of national history at the Athenaeum Illustre, Hugo Beijerman.)
TLINK: none

Example 5 shows a particular instance where a single TIMEX3 is connected to multiple EVENT entities. Here, the combination of TIMEX3_1 and EVENT_1 actually denotes a new temporal expression of the form: *after*-[EVENT-*equal*-TIMEX3]. EVENT_2 is now linked to the recognized TIMEX3_1 of type DATE, while it should formally be linked to the new temporal expression, which would be of type DURATION or, more specifically, the starting point of a DURATION. However, the event of joining a committee is a momentary happening and, as such, should logically be connected to a DATE instead of a DURATION. Taking this into consideration, the resulting analysis can be accepted as is, if the ordering of the events on the timeline is kept the same as the ordering of the EVENTs in the sentence.

(5) Na de EVENT_1[spoorwegstaking] van TIMEX3_1[1903] EVENT_2[kwam] hij als opvolger van de vrije socialist H. Alkema in het Comité van Verweer.

(After the EVENT_1[railway strike] of TIMEX3_1[1903] he EVENT_2[joined] the Committee of Defence as the successor of the free socialist H. Alkema.)

TLINKs: EVENT_1 → TIMEX3_1 : *equal*
EVENT_2 → TIMEX3_1 : *equal*

A similar case is demonstrated by Example 6. Here, the event of *becoming* blind (EVENT_3) is said to end at the moment that the subject passes away, while in fact this pertains to the event of *being* blind. If EVENT_3 is interpreted as a state rather than a happening, then the analysis is correct, though possibly incomplete: the inception of the state most likely occurred "after the war", which is not recognized as a temporal expression by our method.

(6) Na de EVENT_1[oorlog] EVENT_2[leefde] Reyndorp, EVENT_3[blind geworden], nog tot TIMEX3_1[begin 1950].

(After the EVENT_1[war] Reyndorp, EVENT_3[blinded], EVENT_2[lived] until TIMEX3_1[early 1950].)

TLINKS: EVENT_2 → TIMEX3_1 : *TendsE*
EVENT_3 → TIMEX3_1 : *TendsE*

A common error in the part of speech tagging of the text is the wrongful classification of the word "zijn", which can occur as a verb ("to be") or as a pronoun ("his"). In many cases, the pronoun is mistakenly tagged as a verb. As the part of speech tags are used in the syntactical parsing of the sentence, this has a detrimental effect on both the EVENT and TLINK extraction processes, as illustrated by Example 7. Here, the interference caused by the wrongful classification of two occurrences of the pronoun "zijn" leads to three incorrect TLINKs, while the actually targeted EVENT "schreef" does not get linked to any TIMEX3s.

(7) Behalve de duizenden artikelen die in de loop van zijn leven uit zijn pen EVENT_1[vloeiden] EVENT_2[schreef] hij TIMEX3_1[begin jaren twintig] tijdens EVENT_3[zijn] royementsperiode De communist en EVENT_4[zijn] sexueele moraal (Overschie TIMEX3_2[1926]), dat in een kleine oplage met een voorwoord van Roland Holst EVENT_5[verscheen].

(In addition to the thousands of articles that EVENT_1[flowed] from his pen during the course of his life, in TIMEX3_1[the early 20s] during EVENT_3[his] expulsion period he EVENT_2[wrote] The communist and EVENT_4[his] sexual morals (Overschie TIMEX3_2[1926]), which EVENT_5[appeared] in a limited edition with a foreword by Roland Holst.)

TLINKS: EVENT_3 → TIMEX3_1 : *TstartsE*
EVENT_4 → TIMEX3_1 : *TstartsE*
EVENT_4 → TIMEX3_2 : *TendsE*
EVENT_5 → TIMEX3_1 : *TstartsE*

Example 8 shows a sentence where multiple recurring events are detected. The nature of EVENT_1 is so that it can only happen once. The other events might have multiple occurrences, though not as expressed by this sentence. The error made here is that the dependency relations between each EVENT-TIMEX3 pair span the entire sentence, instead of only its separate clauses. It should further be noted that "ziekte" ("illness") is not recognized as an EVENT. It does not occur on our list of noun events, because its frequency of occurrence in the BWSA is below 25.

(8) Na een ernstige ziekte EVENT_1[stierf] Heijermans in TIMEX3_1[juli 1938], vlak voor het EVENT_2[rapport] in TIMEX3_2[augustus] in druk EVENT_3[verscheen].

(After a serious illness Heijermans EVENT_1[died] in TIMEX3_1[July 1938], just before the EVENT_2[report] EVENT_3[appeared] in print in TIMEX3_2[August].)

TLINKS: EVENT_1 → TIMEX3_1 : *equal*
EVENT_1 → TIMEX3_2 : *equal*
EVENT_2 → TIMEX3_1 : *equal*
EVENT_2 → TIMEX3_2 : *equal*
EVENT_3 → TIMEX3_1 : *equal*
EVENT_3 → TIMEX3_2 : *equal*

	Gold	Unannotated
tokens	1589.4	1541.8
TIMEX3, of which:	32.6	31.9
DATE	72.2 %	71.9 %
DURATION	23.6 %	23.2 %
REFERENCE	4.2 %	4.0 %
no label	0.0 %	0.5 %
linked TIMEX3	75.9 %	73.5 %
links per linked TIMEX3	1.21	1.19
EVENT	167.8	160.3
linked EVENT	18.2 %	18.2 %
links per linked EVENT	1.05	1.04
TLINK, of which:	30.2	29.4
equal	76.1 %	79.3 %
EduringT	14.1 %	18.6 %
TstartsE	5.7 %	1.0 %
TendsE	4.2 %	0.5 %

Table 4.12 – Descriptive statistics of the temporal analysis of the BWSA. The left column shows the averages over the gold annotations (100 biographies) for TIMEX3 and EVENT, with automated TLINK extraction. The right column shows the same numbers measured on the fully automated analysis of the 473 remaining, unannotated biographies.

4.4. Discussion

In this chapter we tested the performance of state-of-the-art temporal analysis methods on the BWSA, with the intention of using the results of the analysis to convert our static social network model into a dynamic graph where the edges are anchored to a timeline. We included analysis of linguistic events in the hopes of using the outcome to further classify edges and detect patterns. Our efforts serve to answer our second research question:

> **RQ 2** To what degree and level of specificity can we reliably recognize and normalize temporal information in Dutch biographical text using state-of-the-art techniques?

Until now, the only method available out-of-the-box for automated Dutch temporal analysis was our own implementation of a Dutch rule set in the state-of-the-art HeidelTime system (Section 4.2.1). This system performs recognition, identification, and normalization of TIMEX3 entities, but does not carry out EVENT or TLINK extraction. Our current method replaces the rule-based TIMEX3 analysis of HeidelTime with a hybrid approach, in which recognition and identification are solved using two CRF classifiers, and normalization is dealt with using a straightforward rule-based process. Our system most differs from HeidelTime in its classification of dates that occur as one half of a duration (e.g. "*from July*", "*until the end of summer*"): it identifies these as DURATION with an unspecified start or end date, while HeidelTime, which follows the TimeML guidelines more strictly, classifies them as DATE. We also diverge from TimeML in the normalization of TIMEX3 entities. Instead of a single, momentary value, we

assign a start and end value to each TIMEX3 to denote the interval that it describes. Either value can be left blank if it cannot be derived from the expression, thus retaining all information contained in the text. We believe that these small modifications allow for a more detailed and intuitive analysis of the temporal dimension.

Our system performs significantly better than HeidelTime on the BWSA, reaching strict F1 scores of up to 93.5 for TIMEX3 recognition, and F1 scores between 61.2 and 94.5 for the identification of the different TIMEX3 types. DATE normalization is performed with an accuracy of 95.4 %. These scores are among the highest reported for this task. Normalization of DURATION and REFERENCE entities reaches accuracy levels of only 20 % and 4 %, respectively, on completely correct values, and around 84 % and 40 % on partially correct values. Although these results are much lower than for the DATE class, they are still a great improvement over those obtained with HeidelTime, which can be considered as the current state-of-the-art for Dutch temporal analysis. We can therefore conclude that our approach to temporal expression detection and normalization is actually better than state-of-the-art.

In contrast to HeidelTime, our system also extracts EVENT entities, and links these to any TIMEX3 entities in the same sentence, both in a rule-based setup. The quality of the outcome heavily depends on the result of the TIMEX3 processing and on the dependency parse delivered by Frog. As we have seen, the values assigned to DURATION and REFERENCE entities are in most cases incorrect, which often leads to an erroneous classification of the TLINKs associated with them. Furthermore, sentences containing multiple clauses are often incorrectly parsed by Frog, which causes the incorrect EVENT-TIMEX3 pairs to be linked. Our current efforts are aimed towards the application of automated text analysis in the context of social historical research. In light of this fact, we need to consider the effect of the errors included in our results, before we incorporate any of them into the BWSA graph. One of our stated goals is to *improve efficiency* with regards to semantic analysis of large, interconnected, textual sources from the Social History domain. This improvement should then serve to decrease the *apprehension of automated methods* that generally exists among social historians. Therefore, we need to take care not to frustrate them by including too many unreliable factoids in the graph. Manual post-correction is always an option, but in essence defeats the purpose of this thesis and should be kept to a minimum. In its current state, the analysis of DURATION and REFERENCE entities is not reliable enough for the output to be reapplied in a scientific context. We therefore exclude these from any further applications and analyses of the data. Luckily, most of the TIMEX3s are of the DATE class. These are not only identified with great accuracy, but their very nature intuitively implies that their associated TLINKs are of type *equal*. This intuition is supported by the large percentage of TLINKs with this class (Table 4.12). We should note that the exclusion of the DURATION class does not mean that all TIMEX3s will have their end date equal to their start date. Phrases such as "*the 1860s*" are classified as DATE, but normalized to the start and end date of the entire

period. With this in mind, we can confidently enrich the edges in our social graph with the start and end dates of DATE expressions that are associated with them, and also include any EVENTs attached to the DATE, and thus convert our static graph from Figure 3.5 *b* into a dynamic network that models how connections evolve over time.

In the previous Chapter, we identified all nodes through NER and NED. We then determined the edges by connecting all entities that occurred together in a sentence. For each edge, we now check whether the sentence that it was derived from contains a DATE. If it does, then we add the start and end date to the edge as attributes. Each EVENT that is related to the DATE with type *equal* is also added as an edge attribute. Of course, not every sentence in the BWSA contains a temporal expression. In fact, out of the 68,536 named entity mentions in the total of the BWSA, only 30,754 (44.9 %) appear in a sentence containing a temporal expression. In order to bind the remaining mentions to the timeline and determine whether they fall into the time frame of our graphs, we need to derive a timespan for these sentences as well. For this purpose, we assume that the events detailed in the biographies are mentioned in chronological order, and that a paragraph describes a coherent sequence of events spanning the period between the first and last mentioned temporal expressions. The sentences within a paragraph that do not contain a temporal expression are classified according to their surrounding sentences with temporal expressions. Paragraphs containing no temporal expressions at all are temporally placed in between their preceding and following paragraphs, or restricted to the period between the biographee's birth and death dates if no neighbouring paragraph with a timespan is found.

Acknowledging the life spans of all biographees, the entire time frame described by the BWSA spans from 1778 to 1998. However, the most active period in the data in terms of the number of people alive ranges from approximately 1860 to 1920. Therefore, we cut off the time frame for the BWSA graph to the period between January 1, 1860, and December 31, 1919. The resulting graph (from here on: TIMEX P-P) is displayed in Figure 4.4. Table 4.13 lists the accompanying statistics alongside those obtained on the HTML gold standard and NED P-P graphs from Chapter 3. Technically, the TIMEX P-P graph is a sub graph of the NED P-P graph, which is reflected by the fewer number of edges and connected nodes. In total, there are 69 disconnected nodes in TIMEX P-P. The absence of these nodes has an effect on all statistics. The most notable statistical difference is the number of communities, which increases to 13 in TIMEX P-P. This implies that some of the missing nodes serve to form bridges between communities in the HTML and NED P-P graphs. Even though Recall and the ranking correlations decrease, this does not mean that TIMEX P-P is a less accurate model than the NED P-P graph. It is merely limited to the number of connections that we can reliably connect to the timeline. We have not currently captured any direct edges between the 69 unconnected nodes and any of the other BWSA members that we could reliably anchor in time. However, we do have

connections to other types of entities that

	HTML	NED P-P	TIMEX P-P
nodes	573	573	573
nodes (degree > 0)	564	567	504
edges	2,969	3,350	2,280
average degree	10.4	12.3	8.0
average weighted degree	12.1	26.0	15.2
network diameter	7	6	9
communities	8	10	13
edge Precision		77.5	77.9
edge Recall		92.1	59.8
edge F1		84.2	67.6
degree ranking correlation		0.9199	0.7050
betweenness ranking correlation		0.8423	0.6019

Table 4.13 – Statistical comparison of the NED P-P and TIMEX P-P graph models to the gold standard HTML graph.

potentially show traces of their actions. We might use these edges to secondary nodes to strengthen and/or supplement the connections in the direct person-to-person graph. We conduct a small experiment to investigate whether we can successfully translate mutual connections to organizations, locations, or people from outside the BWSA community into edges that will bring our TIMEX graph model closer to the gold standard HTML graph. To this end, we create a series of adapted graph models in which we consolidate common edges to the same secondary entity to form direct (BWSA) person-to-person edges. Consider the graphs in Figure 4.3 as an example. Whenever we encounter an organization, location, or non-BWSA person node (A) that is directly connected to multiple biographees (1 through 4), we add edges between all pairs of the connected nodes and remove the secondary node from the graph. The weight of the new edges is determined by the minimum of the weights of each pair of nodes to the secondary node. For instance, if edge 1-A has weight 1, while edge 2-A has weight 2, then the weight of the new edge 1-2 will be 1. If an edge already exists between a pair of nodes, then the weight of that edge is increased with the weight of the new edge. We place constraints on the edge formation in regards to the minimum weight required for a consolidated edge to be added to the existing graph. We expect that the inclusion of low weight consolidations will generate a lot of noise, since these represent connections for which we have found only a few pieces of evidence. However, if we reject too many edges we run the risk of excluding just those edges that we are after. We aim to find a compromise between filtering out the noise and increasing the accuracy of the graph. For each of the secondary node classes we therefore construct graphs with minimum consolidated edge weights ranging from 1 to 3. We evaluate each graph with respect to the number of nodes that it connects, paying special attention to the 69 nodes that are disconnected from the TIMEX P-P graph. These nodes are divided into three groups:

- 24 nodes that are *unlikely* to appear in the graph, either because they were not alive between 1860 and 1920, or because they were born after 1899 and therefore are unlikely to show any activity during this period;

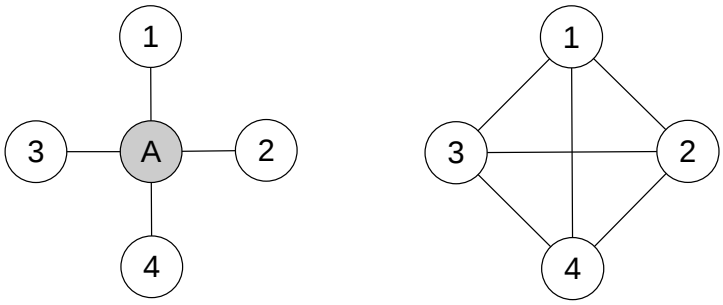


Figure 4.3 – Example of edge consolidation through a common node. Left: graph before consolidation. Right: graph after consolidation of edges through node A.

- 18 nodes that could *possibly* appear in the graph, but might not because of there young age during the selected time frame (i.e. born between 1890 and 1899);
- 27 nodes that are *expected* to have some registered activity during the period, i.e. they were born before 1890 and alive after 1860.

The consolidated graphs are labelled TIMEX P-p-P, TIMEX P-o-P, and TIMEX P-l-P for consolidations over people, organizations, and locations, respectively. Figure 4.5 shows per graph what percentage of nodes from the unlikely, possible, and expected groups are connected to the graph after consolidation with minimum edge weights of 1 to 3. We see that each of the graphs is able to connect some of the expected nodes, though the TIMEX P-o-P and TIMEX P-l-P graphs also include many unlikely nodes at the lower weight levels. Increasing minimum edge weight leads to better node filtering for these graphs. Still, we also have to take into account the number of newly created edges between the already connected nodes. If many edges are added to this part of the graph that do not exist in the gold standard, then the Precision of the graph will go down, resulting in a less accurate graph model. In fact, the number of new edges runs into the thousands for all graphs, with the TIMEX P-l-P graph topping the list with over 76,000 new edges at weight level 1. Given that the HTML gold standard barely contains 3,000 nodes this is a gross over estimation of the connectedness of the graph. These results lead us to the conclusion that it is not feasible at this time to strengthen and supplement the edges in the BWSA graph using connections to secondary nodes.

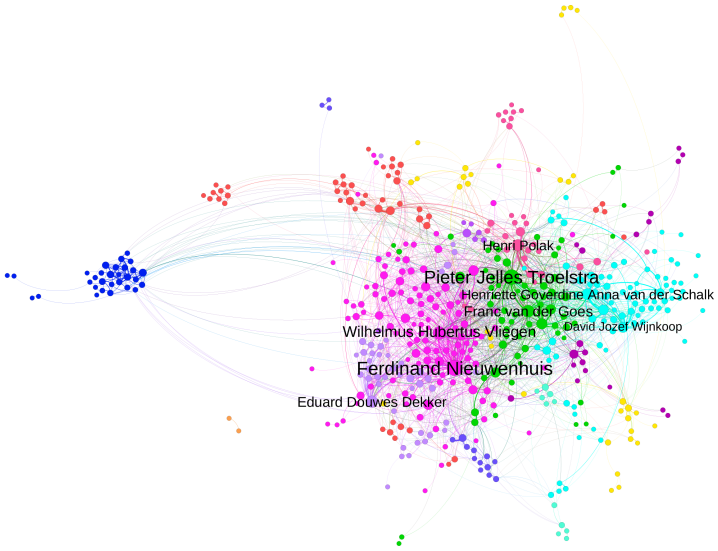


Figure 4.4 – Force-directed visualization of the social network of BWSA biographees between January 1, 1860 and December 31, 1919 constructed from sentence-level co-occurrences. The nodes are coloured by modularity class to distinguish communities, and sized by degree. Only the most prominent nodes are labelled.

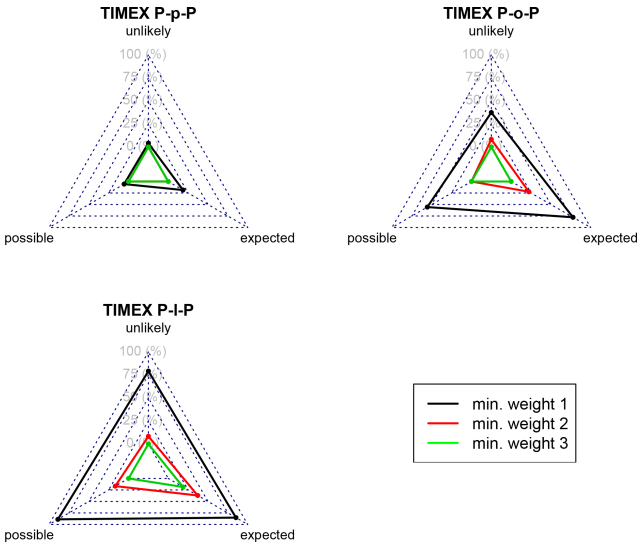


Figure 4.5 – Percentage of unlikely, possible, and expected nodes connected in graphs consolidated over common connections to non-BWSA people (TIMEX P-p-P), organizations (TIMEX P-o-P), and locations (TIMEX P-l-P) with minimum edge weights varying from 1 to 3.

5 THE SOCIALIST NETWORK

Men make their own history, but they do not make it as they please; they do not make it under self-selected circumstances, but under circumstances existing already, given and transmitted from the past.
- Karl Marx, *The Eighteenth Brumaire of Louis Napoleon*, 1852

Static social network models that aggregate nodes and edges over a longer period provide a great tool for the examination of the overall organizational patterns within communities. However, in order to uncover patterns of growth, decline, confluence, or disruption of a network of people, temporal information needs to be included in the model to enable us to study how the graph develops over time. This transforms the *static* network to a *dynamic* one. Dynamic graphs have been used in the past to prove the existence of, for instance, preferential attachment mechanisms, where new nodes are more likely to connect to an already highly connected node in the graph. This, and other mechanisms, as well as structural properties seem to occur in similar fashions over the majority of graph models studied, whether they encode connections between people, computers, or genes. The study of network structure and evolution is called Graph Theory and has been applied to varying fields such as Anthropology, Sociology, Linguistics, Chemistry, and Biology, to name a few (Hage & Harary, 1983; Levine, 1972; Krahmer, Van Erk, & Verleg, 2003; Hansen & Jurs, 1988; Mason & Verwoerd, 2007). So far, however, research into social networks in particular has mostly been limited to the study of static graphs representing relatively small communities. Only since recently have online Social Networking Services, such as Facebook, Twitter, and Instagram, been able to provide us with the data required to study human interaction networks at a larger scale. Nevertheless, this data is suited only for very specific research questions that currently fall outside the scope of most historical domains.

We have developed a method of constructing social networks from free text, in order to provide a solution to the lack of historical, longitudinal data suitable for dynamic social network analysis. Through the application of straightforward NLP methods we are able to construct an accurate graph model of the social network of Dutch socialists using a biographical dictionary as input. So far, we have only considered the network as a static structure. In this chapter, we make use of the fact that the edges in the graph are bound to a timeline, and investigate the evolution of the graph. We aim to validate our method of social network construction by

checking whether our dynamic graph model of the BWSA community adheres to properties commonly found in social networks, thus answering our third research question:

RQ 3 Do social network models constructed with the described method adhere to properties commonly observed in social networks?

The rest of this chapter is structured as follows. In Section 5.1, we provide an introduction into Social Network Analysis and Graph Theory, and identify properties that our graph should comply with as a valid social network model. We also discuss the difference between static, aggregated graphs versus dynamic graphs in terms of their analytical power. In Section 5.2 we investigate whether the BWSA graph model actually adheres to the properties described for social networks. In Section 5.3, we analyse the linguistic events that we have extracted in Chapter 4 in relation to the type of edge that they are attached to. We close the chapter with a discussion of our findings in Section 5.4.

5.1. Social Network Analysis

The fields of Sociology and Anthropology, but also Social History, are generally concerned with the structural composition and procedural changes of communities of people. Individuals are regarded, not in light of their personal characteristics, but in the roles that they play within their social environment. The general assumption underlying this approach is that a connection between two individuals is the product of their combined characteristics and thus that the network of relations provides a mapping of (the consequences of) the communal preferences. Social Network Analysis (SNA) is the method that arose from these fields to study such complex social structures (Wasserman & Faust, 1994). SNA lends much of its terminology and methodology from Graph Theory (West, 2001). Formally, Graph Theory studies the ways in which sets of points can be connected using lines. In the context of SNA the points, or *nodes*, usually represent people, and the lines, or *edges*, represent relations, or *ties*. The edges of a graph can be mutual, or *undirected*, where a connection from A to B implies the same type of connection from B to A, or they can be one-sided, or *directed*, where a connection from A to B does not imply a reciprocal connection from B to A. A complete collection of nodes and edges is referred to as a *graph*. We make a distinction between a *network*, interpreted as a complex structure of relations between a collection of entities, and a graph, interpreted as a formal model of such a network. When studying a graph of a social network, measurements can be taken over the nodes and edges to reveal different aspects of the network. These measurements are divided into *structural variables* measured on edges, and *compositional variables* measured purely on nodes (Wasserman & Faust, 1994). Since the sociological background places more emphasis on the global composition than on the individual characteristics, research in SNA focuses mostly on the structural variables. In this Section, we introduce the most important concepts from Graph Theory as they are applied in SNA, and simultaneously review relevant research into SNA. We start by introducing the

concept of *small-world networks*, a type of network model that is found to be a good representation of real-world social networks. Differences in the growth mechanisms of social networks lead to different types of small-world networks, as we will explain in Subsection 5.1.2. We will discuss their common characteristics so that we can compare them to our own network data. Since we are also interested in locating the most significant nodes in the sense of their impact on the overall structure, we will discuss the structural variable *centrality*, which is the most widely used measure of the importance of nodes within a network. Edges most show their significance over time, where repeated connections provide for stronger ties. The order and evolution of ties provide a handle for studying the *flow* of a graph, i.e. the way in which information, disease, friendship, and so on, spreads through the network. To this end, we discuss *dynamic graphs* and review how centrality is best measured over an evolving network.

5.1.1. Small-world networks

The number of edges connected to a node denotes the node's *degree*. A well-known concept in the domain of social networks is the "six degrees of separation" theorem (Milgram, 1967), which postulates that every person in the world is separated from any other person by maximally six connections in the form of personal acquaintances. It has even found its way into popular culture through the trivia game "six degrees of Kevin Bacon", where players try to connect arbitrary actors to actor Kevin Bacon in the least number of steps through the movies that they have played in.[1] This so-called *small-world* phenomenon was empirically established, most famously, by Milgram in his small-world experiment (Milgram, 1967). In the experiment, a group of randomly selected US residents were asked to pass a folder to a single target person unknown to them and living thousands of miles away. As a rule, participants were only allowed to pass the folder to a person with whom they were personally acquainted, so they had to carefully consider which of their friends and relatives was more likely to know the target person than they were themselves. Milgram found that, on average, it took only 6 steps for the folders to reach the target person, thus confirming the theorem. In a more recent effort, Watts & Strogatz (1998) mathematically validated the concept using Graph Theory. They describe an algorithm that starts with a ring-shaped graph with N nodes. Each node is initially connected to its k nearest neighbours. As a consequence, each node has a degree equal to k. If all nodes in a graph have equal degree, then the graph is said to be *d(egree)-regular*. Watts and Strogatz's algorithm passes over every initial edge in this graph and rewires each with probability p to another node, which is selected from all nodes in the graph with uniform probability. As the rewiring probability p increases, so does the randomness of the graph. They found that as the randomness increases, the *average path length* between nodes in the graph decreases, while *clustering* remains more or less the same. However, as the graph approaches

[1] http://en.wikipedia.org/wiki/Six_Degrees_of_Kevin_Bacon

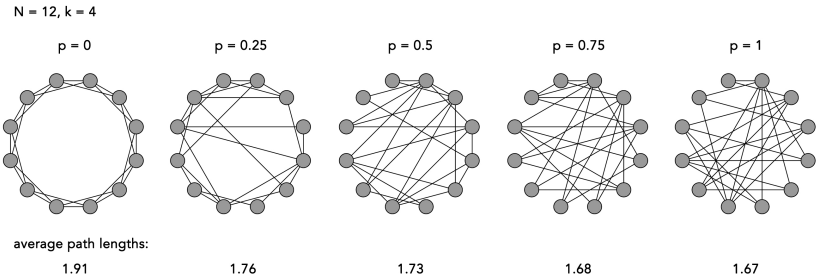


Figure 5.1 – Output of the Watts & Strogatz algorithm on a graph with 12 nodes and k = 4, with p varying from 0 to 1. As p increases, the average path length decreases (Watts & Strogatz, 1998).

complete randomness, clustering disappears and the graph no longer complies with the small-world theorem. Compared to d-regular graphs, small-world models provide a better model of real-world networks. Figure 5.1 displays the output of the Watts & Strogatz algorithm on a graph with 12 nodes and $k = 4$ for probabilities varying from 0 to 1. At the heart of the "six degrees" principle lays the graph theoretical concept of *shortest paths*. As the name suggests, the shortest path between nodes A and B is the route from A to B that traverses the least number of edges. Out of all shortest paths, or *geodesics*, between A and any other node in the graph, the length of the longest of these paths is called the *eccentricity* of node A. The maximum of all node eccentricities is the *diameter* of the graph. It should be noted that in a strict small-world model, a path exists between every pair of nodes, thus creating a graph consisting of a single *component*. In a real-world setting, it is not unthinkable that a situation occurs where some nodes simply are not reachable. This breaks the graph into multiple, unconnected components, the largest of which is referred to as the *giant component*. All of the smaller components together are commonly referred to as the *middle region*, while solitary nodes that do not connect to others are called *singletons*. Keeping track of the nodes in these different regions of a graph can reveal much about the evolution of the graph (Kumar, Novak, & Tomkins, 2010).

5.1.2. Scale-free networks

Even though the average shortest path length is a good measure of a graph's global connectedness, it does not reveal anything about its inner workings. Yet this is the information that is most crucial in many situations. For example, a marketer spreading the word about a new product will want to know who the opinion leaders in a network are. Similarly, to stop an infectious disease from turning into an epidemic, a health officer will want to know which people to vaccinate in order to minimize the spread. In both cases, the task is to find the most influential, or most connected, nodes (Morone & Makse, 2015).

If a graph is truly random, then the probability distribution of the node degrees should follow a Poisson distribution, that is, most nodes should have a degree close to the average. Watts & Strogatz (1998) have shown that small-world models have increasing measures of randomness. However, when studying the degree distribution of a mapping of the World Wide Web, which is also a small-world network, Barabási & Albert (1999) found that it did not resemble a Poisson distribution. Instead, it followed a power law, where most nodes have low degree and a few nodes have very high degree. They labelled these high-degree nodes as *hubs* and hypothesized that many, seemingly random networks would actually display this *scale-free* property. Further research revealed that this hypothesis was correct and that the phenomenon in fact occurs in a wide variety of networks (Braha & Bar-Yam, 2006). A unique characteristic of a scale-free network is that the removal of a random node that is not a hub will only change the size of the graph, not its overall structure. The removal of a hub, however, does have a big impact on the structure and, in extreme cases, can even result in a complete disintegration of the network. A network with the scale-free property also adheres to the small-world property.

To understand how the scale-free property emerges, it is imperative to look at the evolution of a network. The mechanism proposed by Barabási & Albert (1999) to explain why some nodes are much more "popular" than others is called *preferential attachment*. It states that the likelihood of a newly formed edge being attached to a node increases with the degree of that node. In layman's terms: people are more likely to befriend someone who is already friends with a lot of people. There are, however, other mechanisms of attachment that will result in different degree distributions. Amaral, Scala, Barthelemy, & Stanley (2000) compare the degree distributions of several technological, economic, social, and natural networks, and conclude that none of them strictly adhere to the scale-free property. Out of the three social networks that they examine, they find that two of them have Gaussian degree distributions, while the third starts off as a power law, but results in a Gaussian decay of the tail. Based on their findings, Amaral et al. define two additional network types. The first is the *broad-scale* network, which is characterized by a distribution that initially follows a power law, but has a tail with exponential or Gaussian decay. The second is the *single-scale* network, which is characterized by a distribution with completely exponential or Gaussian decay. The cause of these differing distributions comes from possible interruptions of the preferential attachment mechanism: nodes may age past the point where they no longer receive new edges, the cost of adding edges to a node may be too high, or a node may have a limited edge-capacity (Amaral, Scala, Barthelemy, & Stanley, 2000).

5.1.3. Centrality

The popularity or importance of a node is expressed by its *centrality*. Centrality can be calculated on the individual nodes (*point centrality*), or over the entire graph

(*graph centrality*) (Freeman, 1979). We focus our review on point centrality. Point centrality is used to determine the relative importance of a node to the overall graph structure. Generally, four different types are distinguished (Figure 5.2): *degree centrality*, which ranks nodes according to the number of connections they have to other nodes (Shaw, 1954; Nieminen, 1974); *betweenness centrality*, which measures how often a node is included in the shortest path between node pairs (Anthonisse, 1971; Freeman, 1977); *closeness centrality*, which expresses how embedded in the graph a node is by taking the sum of its shortest path lengths to all other nodes in the graph (Bavelas, 1950; Beauchamp, 1965; Rogers D., 1974); and *eigenvector centrality*, which assigns higher ranks to nodes that are either well-connected themselves, or that are connected to other well-connected nodes (Bonacich, 1987; Bonacich, 1991). Each type of centrality expresses a different quality of the node in question and, consequently, serves a different purpose (Freeman, 1979). Degree-based centrality indicates how active a node is within the graph and thus is most useful in contexts where communication activity has focus. Betweenness centrality highlights those nodes that have the highest control over the flow of information through a graph. This can for instance be applied to domains where it is important to manage access to information. Closeness centrality measures how efficiently nodes can contact one another and how much they depend on each other to receive information. Finally, eigenvector centrality can be used to determine which nodes have the most influence over others. Centrality measures lie at the basis of the research into graph *flow*, which is defined as the manner in which quantities (information, power, goods, diseases, etc.) are

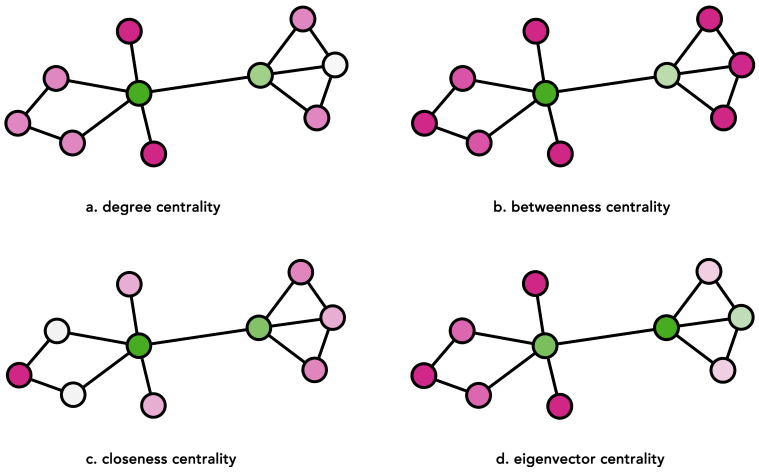


Figure 5.2 – Nodes ranked according to four types of centrality. Green implies higher centrality, while red implies lower centrality.

	Transfer	Serial duplication	Parallel duplication
Shortest paths	package delivery *closeness, betweenness*	mitosis *closeness*	- -
Paths	hand-me-down clothing -	sexually transmitted diseases *eigenvector*	web server *closeness, degree*
Trails	book lending -	gossip -	chain letters *closeness, degree*
Walks	money exchange -	emotional support -	ideology transfer *closeness, degree, eigenvector*

Table 5.1 – Examples of diffusion across networks classified according to their method of transition and their trajectory type. Also listed are the most appropriate measures of centrality. Borrowed and adapted from **(Borgatti, 2005)**.

distributed across a network. However, as Borgatti (2005) indicates, all point-centrality measures make implicit assumptions about the flow that do not hold in many situations. For example, betweenness centrality assumes that all traffic in a graph always follows the shortest routes possible. For this to apply, either the source has to be aware of both the identity and the position of the target (and any nodes in between) in the network, or the possible targets should be restricted to the direct neighbours of the source, resulting in a maximum path length of one. A real-world example of the former is the post office, where packages are sent to clearly identified addresses across a, presumably, optimally configured grid of sorting and distribution centers. An example of the latter would be a revolutionary spreading his ideology with the intention to convince as many as possible. Contrary to a package, belief in an ideology may exist at multiple positions in the graph at the same time. Once it spreads, the source node is still in possession of the ideology and the target node may also start spreading it. Furthermore, a source may "convert" multiple nodes at once, and it might repeat the conversion of the same node at multiple points in time to reinforce that node's conviction. This example highlights aspects of flow that can be classified along two dimensions (Borgatti, 2005): the method of transition, and the type of potential trajectories. Transition is divided into *transfer*, where something exists at only one point at a time; *serial duplication*, where a node can convert only one adjacent node at a time; and *parallel duplication*, where a node can convert multiple adjacent nodes at the same time. Regarding the potential trajectories, there are four types to consider. The first type is the shortest path, which, as discussed, presumes that all traffic follows distance-optimal routes. The other three types are distinguished by their allowance to revisit nodes and to traverse the same edge more than once. When both are not allowed, and the route chosen does not have to be optimal regarding its distance, then the trajectory is said to follow a *path*. When it is allowed to revisit nodes, but not to traverse the same edge twice, then the trajectory forms a *trail*. Lastly, when both nodes and edges can be traversed multiple times, then we speak of a *walk*. Table 5.1 lists examples for each transition-trajectory combination and the centrality measures that comply with it. For the BWSA graph we find betweenness centrality

to be the best fit due to its capacity to capture flow of information in a parallel duplication setup.

5.1.4. Dynamic graphs

The crux of the centrality incompatibility issue stems from the manner in which much of the research into longitudinal social networks so far has modelled graph dynamics. Instead of looking at each point in time separately, most aggregate all momentary snapshots into a single static structure (Ferrer i Cancho, Janssen, & Solé, 2001; Albert & Barabási, 2002; Kossinets & Watts, 2006). This approach poses two major problems. First, centrality measured on this graph is presumed to reflect the overall tendency of the entire network, but, as confirmed by Braha & Bar-Yam (2006), these aggregated measures do not accurately reflect the actual importance of the nodes. Their analysis shows that most of the nodes that have the highest centrality at any given point in time only hold this position of importance for a small fraction of the entire timespan. In fact, the centrality distributions of the majority of the nodes adhere to a power law, where they are unimportant most of the time (and, as a consequence, have low aggregated centrality), but as time goes by, eventually achieve their proverbial "15 minutes of fame".

Second, aggregation over time does not factor in the actual sequence of connections (Lerman, Ghosh, & Hyung Kang, 2010). This problem is illustrated in Figure 5.3. The aggregated graph (Figure 5.3a) shows a path between nodes A and D. However, when taking each separate timeframe into consideration (Figure 5.3b-5.3d), we see that a path starting at node A can never reach node D, since the edge between nodes C and D is only activated at t_0. At that point nothing has yet been transferred from A to C that can be passed along to D. So, even though the graph appears to be completely connected, it actually consists of two, partially overlapping components (A-B-C-E and B-C-D-E), each of which consists of several sub-components bounded to the timeframes.

As a solution to the discrepancy between aggregated node centrality and time-aware node centrality, Lerman, Ghosh, & Hyung Kang (2010) propose a method that incorporates the dynamic evolution of a graph. They model dynamic centrality both as a memoryless process, where the state at t_i depends only on the state at t_{i-1}, and as a process with memory, where the state at t_i depends on all previous states. Using two attenuation factors that describe the probability that a node either initiates a transmission, or passes on a previously received message to another node at t_i, they calculate the amount of information that is expected to be transferred between two arbitrary nodes over a given time interval. Nodes are then ranked according to the amount of information they are able to send to all other nodes over the entire timespan of the network. The same models can also be applied at the node level, where they can be used to determine which node has the most influence over another given node. Although Lerman, Ghosh, & Hyung Kang's method

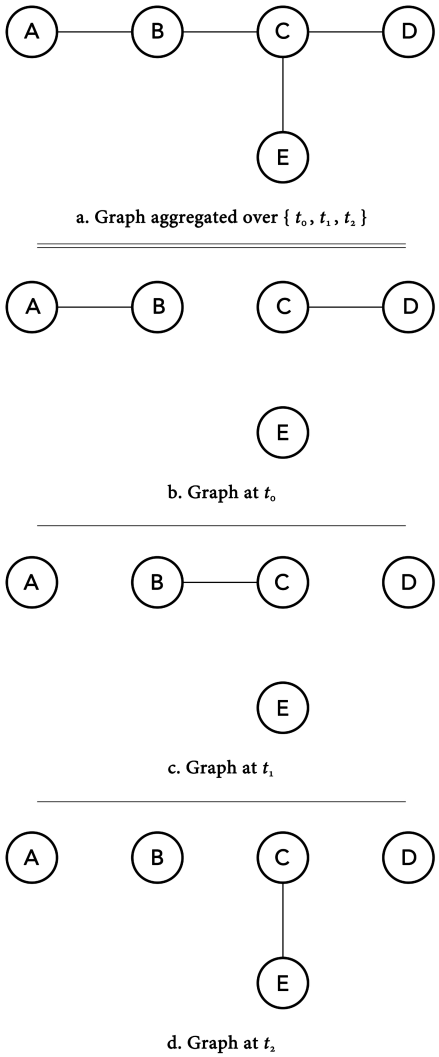


Figure 5.3 – Model of a dynamic graph. Figure a shows the graph aggregated over all time frames; figures b-d show the graphs for each individual time frame.

succeeds in producing seemingly more accurate rankings than the ones resulting from measurements on a static, aggregated graph, the values of the attenuation factors are mere estimations that are not easily validated on actual network data. Moreover, the method does not escape its own criticism of the static centrality measures, since it still produces approximations over aggregated data, though at an adjustable scale. We argue that the importance of a node is an inherently

momentary property, meaning that it depends mostly on the current state of the graph, and perhaps a few of its preceding states. Consider the following: as the centrality of a node increases, so does the probability for it to transmit something to other nodes in the graph. If the transmission involves an important piece of information, or perhaps a connection to another influential node, the node in fact spreads some of its influence to its neighbours, thus decreasing its own importance. This explains the power law centrality distributions found by Braha & Bar-Yam (2006), where nodes are highly important for only a very short period. Once values are being aggregated, no matter how small the interval, they will naturally converge to an average. Different temporal scales of the aggregation merely provide different views on the data. Importance therefore is also a matter of *perspective*. The same applies, for instance, when we review world history. Considering the entire history of the world versus a more or less isolated event such as World War II, there is a great difference in the people that we would list as being the most influential within each timeframe. Where larger timeframes provide more generalized views, smaller timeframes give insight into the details of the dynamics. However, unlike much of the previous research has claimed, one does not disqualify the other. Which perspective applies is simply a matter of the type of query that is posed against the data. The dynamics then involve the changes between consecutive (static) states at lower time scales, rather than aggregations over multiple timeframes.

5.2. Statistical analysis

We have identified several characteristics that are commonly found in social networks. In this Section, we investigate whether our BWSA graph adheres to these properties in order to determine whether it is a valid social network model. All tests in this Section are performed on the static, aggregated person-to-person graph ("BWSA-full"), as well as two sets of consecutive static graphs aggregated over set intervals: one set with an interval of 10 years per graph ("BWSA-decade"), and one set with an interval of 1 year per graph ("BWSA-year"), resulting in a set of 6, and a set of 60 graphs, respectively. To make sure that the patterns that we observe in the BWSA data are not due to some random process, we create a random counterpart to each BWSA set. The random graphs are automatically generated and contain the same number of nodes and edges as their BWSA sibling, though the edges are distributed over the nodes at random. First, we check if our graphs model a small-world network by measuring basic statistics, such as average path lengths and network diameters. Second, we study the growth mechanisms of the network by determining the scale of the degree distributions. Finally, we zoom in on the dynamics of the graphs and test our hypothesis that there is little to no correlation between the node rankings of consecutive graphs in our BWSA-decade and BWSA-year sets. This would confirm our belief that importance is a momentary, rather than a permanent property.

5.2.1. Small-worldliness

Previous studies have shown that small-world networks are characterized by short average path lengths and high clustering. Figure 5.4 shows these measures for each of the BWSA graph sets. Indeed, we see that the average path length is always between 3 and 4. This number is lower even than the 6 degrees commonly reported (Milgram, 1967; Watts & Strogatz, 1998), which is likely due to the small size of the BWSA community. It slightly increases as the size of the temporal interval decreases, which is caused by the overall smaller number of edges in these graphs. The clustering coefficient starts at approximately 0.35 for the BWSA-full graph, and rises to almost 0.5 for the BWSA-year set. In contrast, the random graphs that we generated have varying average path lengths and no clustering, confirming that they are truly random and do not adhere to the small-world theorem. To investigate the difference more closely we look at the composition of the graphs in each of the sets (Table 5.2). Listed are the average number of components in each set, followed by the average number of nodes in the giant component, and the average number of nodes in the middle region. Considering the full graphs, we see that the random graph is fully connected, while the BWSA graph contains 2 nodes that are disconnected from the giant component (but connected to each other). For both the decade- and year-sets, however, the random graphs are much more fragmented than the BWSA graphs, with more than double the number of components and many more nodes in the middle region. Our findings show that the BWSA graph sufficiently differs from a random graph for us to conclude that it itself is not random. Moreover, the short path lengths combined with high clustering lead us to conclude that the graph indeed models a small-world network. Next, we focus on the evolutionary mechanics of the network.

5.2.2. Growth mechanisms

According to the literature, many social networks display the scale-free property, which is an indication of a possible preferential attachment mechanism underlying

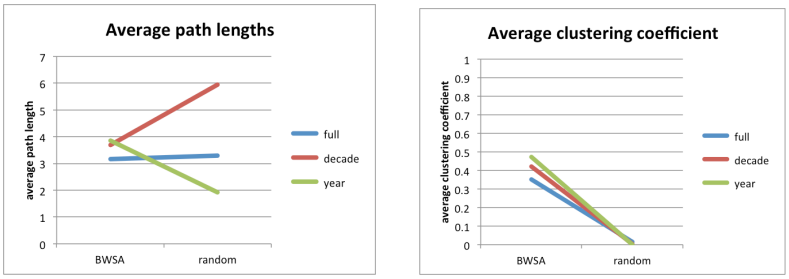


Figure 5.4 – Average path lengths and average clustering coefficients measured on the BWSA graphs versus randomly generated graphs.

	components	giant component size	middle region size
BWSA-full	2	502	2
random-full	1	573	0
BWSA-decade	15	210	34
random-decade	31	326	101
BWSA-year	26	56	68
random-year	72	10	193

Table 5.2 – Composition of full, decade, and year graphs for BWSA and random sets.

	power law	truncated pl	exponential	stretched exp.	log-normal
BWSA-full	1				
random-full			1		
BWSA-decade	3	3			
random-decade		2		1	3
BWSA-year	10	33	8	2	7
random-year		3	13	22	22

Table 5.3 – Distribution of graphs per set over degree distribution types. Power law distributions are an indication of preferential growth in small-world networks.

the evolution of a network. We can determine whether the BWSA adheres to this property by studying the distribution of degrees (number of edges) over the nodes. If we can reliably fit a power law to this distribution, then the graph is considered scale-free. We determine scale for all graphs at each interval size. Power laws are fitted using the Python powerlaw package[2]. The goodness of fit is compared to that of a truncated power law, an exponential, a stretched exponential, and a log-normal distribution. The first indicates a broad scale network, while the latter three are signs of single-scale networks. The scale for each graph is determined by the best fitting model.

As the results in Table 5.3 show, most of the BWSA graphs do indeed have degree distributions that fit a (truncated) power law and, therefore, the evolution of the network is likely controlled by a preferential attachment mechanism. Only for the BWSA-year graphs do we encounter single-scale graphs where the distribution follows a strictly exponential or log-normal scale. The latter is an unlikely finding in a social network that adheres to the small-world principle. Upon inspection we find that the graphs in question are largely disconnected, leading to a degree distribution with mostly zero values. In these cases, the algorithm selects the log-normal scale by default. These distributions do not provide enough data to fit a power law, so their occurrence does not rule out the small-world hypothesis for the BWSA graph. Because not all of the graphs' distributions follow strict power laws, we can conclude that the BWSA network is a broad-scale network. This implies that, if there is indeed a preferential attachment mechanism governing the graph, it is somehow restricted or interrupted. Considering the nature of the data and the long timespan that it describes, this is likely explained by members of the

[2] https://pypi.python.org/pypi/powerlaw

community passing away during the course of the evolutionary process. In these cases, the rate at which the node acquires new connections dramatically decreases and, in most cases, it comes to a complete stop. When connections to deceased community members do occur, the nature of the connection is different from when they were alive. The only opportunity for transfer of knowledge or conviction then lies in primary

There are also some known cases in the data where a person suddenly changes their main ideology and consequently the people that they most associate with (e.g. Ferdinand Domela Nieuwenhuis). Overall, the existence of the growth mechanism ultimately confirms that the BWSA network model is a small-world network. In contrast, we see that the degree distributions of the random graphs rarely fit a power law, confirming that it is truly random and not a small-world network.

5.2.3. Centrality ranking correlations

In Section 5.1.3 we identified betweenness centrality as the appropriate measure for node importance in the context of our graph. The dataset in its essence deals with *ideology transfer*, which is a process involving serial duplication along walks (see Table 3.1). A node with high betweenness centrality occurs along many of the shortest paths in the network. In terms of the BWSA this implies that the person in question has many opportunities to influence the information that travels through the network and, consequently, he or she has the opportunity to control the dominant opinion.

There is much discussion surrounding the suitability of centrality measures for the purpose of ranking nodes by importance, especially in regards to dynamic graphs. The controversy stems mostly from discrepancies found between rankings based on aggregated graphs versus those obtained on smaller slices of the same graph. As discussed, we do not consider this issue to be an argument against centrality measures, but merely cause for a different interpretation. We postulate that rankings taken over graphs of different scales merely provide different perspectives on the same data, and none of the rankings are an invalid representation of that data. We expect high correlations to exist between neighboring graphs of the same locality (i.e. the same temporal scale), which quickly decrease for graphs that are further removed from it in time. Similarly, we expect there to be a medium correlation between graphs of different temporal scales, but describing overlapping periods. To test our hypotheses, we calculate the Spearman rank correlation coefficients between each BWSA-year graph and all of its preceding graphs. Figure 5.5 (top) shows the correlations averaged on the distance between both graphs in years. Figure 5.5 (bottom left) shows the same numbers for the BWSA-decade graphs averaged on the distance in decades. As we expected, graphs at low temporal distances show high correlations. For both graph sets, the correlations decrease to a medium level after approximately 10 years, and dwindle into the lower regions after

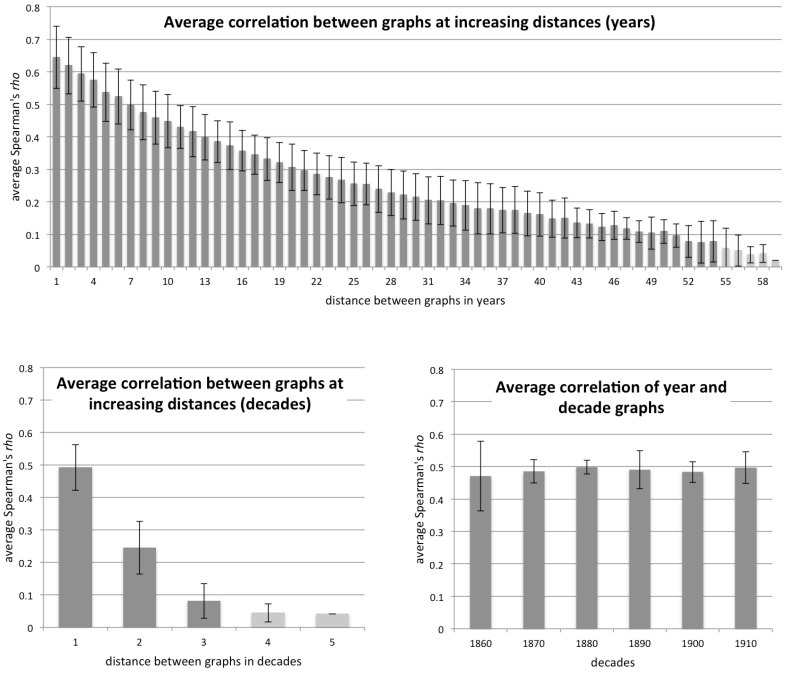


Figure 5.5 – Average Spearman rank correlations of betweenness centrality rankings of nodes. Correlations are calculated between each graph and all of its preceding graphs at the same temporal scale – year or decade – and then averaged over the distance between the current and the preceding graph (top and bottom left). Correlations are also calculated between each BWSA-year graph and the BWSA-decade graph of the corresponding decade, and then averaged over the decade (bottom right). Significant correlations are displayed in dark grey ($p < 0.05$).

20 to 30 years. This confirms our expectation that importance is a momentary property. The decay seems to occur at a faster rate in the BWSA-decade graphs. This is most likely a consequence of the large number of unconnected nodes in each of the BWSA-year graphs. These all have a betweenness centrality of zero and, thus, they all have the same averaged rank in the node rankings. Because they are large in number, there will always be some overlap between graphs and this leads to a slight overestimation of the correlation.

Next, we calculate the correlation coefficients between BWSA-year graphs and the appropriate BWSA-decade graph to test whether node rankings over different temporal scales agree with each other or not. Figure 5.5 (bottom right) shows the BWSA-year correlations averaged per decade. The correlations are all significant at a 95% confidence level. Overall, they are higher than we expected, especially when we take into account the fact that many of the nodes that are unconnected in the BWSA-year graphs are most likely connected in the BWSA-decade graph. This undoubtedly leads to differences in the node rankings that we would expect to see

reflected by lower correlations. It seems that, with regards to the BWSA, temporal scale has little effect on node importance. This indicates the existence of a few highly influential hubs that exude their power over longer periods of time. Further inspection of the data reveals that this is in fact the case. All across the period from 1860 to 1920 we observe high centrality measures for three nodes in particular[3]:

- **Ferdinand Domela Nieuwenhuis** was a controversial key figure in the starting period of the Dutch social movement. He started his career as a Lutheran preacher, but soon parted with the church to become the pioneer of Dutch social democracy. When a growing division between socialists and anarchists divided the socialist party, he turned his back on politics and became an anarchist himself. Through his status as a pioneer and writer, combined with the ideological changes that he experienced on a personal level, Domela exercised influence over a wide variety of people and had the power to instigate real change in the community. Domela occurs in the top 3 of betweenness centrality ranked nodes in 5 out of 6 BWSA-decade graphs and he is the top-ranked node in 31 out of 60 BWSA-year graphs;

- **Eduard Douwes Dekker** was a writer who published under the pseudonym Multatuli. He went to work as a public official in the Dutch East Indies at a young age, where he witnessed the devastating consequences of the colonial system. Back in the Netherlands, he started writing pamphlets, articles, and ultimately novels that would serve to make the injustice known to the public. His book "Max Havelaar" had a great impact at the time it was first published in 1860 and long thereafter. Multatuli is ranked in the top 3 in the BWSA-decade graphs from 1860 up to 1890. In the same period, he is also ranked in the top 3 in 15 of the BWSA-year graphs, where he is the top-ranked node in 8 of them;

- **Pieter Jelles Troelstra** was a socialist politician and lawyer who most famously fought for universal suffrage. He started his parliamentary career around the same time that the then current socialist party was crumbling under disputes between socialists and anarchists (circa 1890). Where Domela made the switch to anarchism, Troelstra remained loyal to the socialist ideology and became one of the twelve founders of the new socialist party, the Social Democratic Workers' Party (SDAP). As the leader of the SDAP, Troelstra enjoyed great respect from his peers and successors alike. He occurs in the top 10 of all BWSA-decade graphs, starting in 1870, and is the top-ranked node from 1900 to 1920. In the BWSA-year graphs, Troelstra is in the top 10 of all but two of the graphs since 1889. He is the top-ranked node in 13 of those graphs.

[3] All centralities and rankings are available at: http://bwsa.taalmonsters.nl

		Person-Person		Person-Organization		Person-Location	
	rank	*Dutch*	*English*	*Dutch*	*English*	*Dutch*	*English*
Top 10	1	treden	tread	aansluiten	join	wonen	live, inhabit
	2	tekenen	draw, sign	worden	become	verhuizen	move, migrate
	3	aangeven	indicate, give	kiezen	choose	vestigen	settle
	4	dragen	carry	oprichten	establish	vertrekken	leave
	5	portret	portrait	toetreden	join	verblijven	stay
	6	sturen	send, guide	raad	council	trouwen	wed
	7	hangen	hang	sluiten	close	keren	turn
	8	begrafenis	funeral	voortkomen	originate	studeren	study
	9	ontmoeten	meet	betrekken	involve	bundelen	combine
	10	krijgen	get	verweer	defence	verblijf	residence
Bottom 10	1655	verbetering	improvement	treden	tread	vormen	to form
	1656	vergadering	meeting	verblijf	residence	opzeggen	terminate
	1657	oprichten	establish	vermelden	mention	opkomen	rise
	1658	keren	turn	bundelen	combine	beschouwen	consider
	1659	kiezen	choose	krijgen	get	sturen	send, guide
	1660	wonen	live, inhabit	verhuizen	move, migrate	dragen	carry
	1661	aansluiten	join	vertalen	translate	worden	become
	1662	vestigen	settle	vestigen	settle	treden	tread
	1663	worden	become	zijn	be	behoren	belong
	1664	verhuizen	move, migrate	wonen	live, inhabit	aansluiten	join

Table 5.4 – The 10 most and least correlated base type events for person-to-person, person-to-organization, and person-to-location edges in the full BWSA graph.

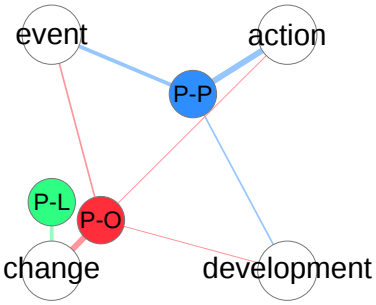


Figure 5.6 – Visualization of the Pearson product-moment correlation coefficients between person-to-person (P-P), person-to-organization (P-O), and person-to-location (P-L) edges and events with main type "event", "action", "development", and "change" in the full BWSA graph. A line indicates a positive correlation. Negative correlations are expressed by the absence of a line. A thicker line implies a stronger correlation, though all correlations displayed are weak (0.002-0.03) and not significant.

5.3. Event analysis

The edges in our BWSA network model are typed in terms of the linguistic events that occur in the sentences from which the edges originate. In the full graph we have connections not just between person nodes, but also between organization-

and location-classed nodes. In this Section, we investigate whether edges between different types of nodes are labelled with events that are semantically reasonable considering the node types. We restrict our study to edges that involve at least one person node. As a result, the edge types that we distinguish are: person-to-person, person-to-organization, and person-to-location. The linguistic events on the edgesare also divided into three classes: main type events, sub type events, and base type events. To determine the relatedness of individual events with each of the edge types, we create a binary vector for both the event and the edge type that indicates for each edge whether it has the event attached to it and whether it is of the selected edge type. The relatedness is then resolved by calculating the Pearson product-moment correlation coefficient between these two vectors. Table 5.4 lists top 10 and bottom 10 of the base type events per edge type ranked on relatedness. It must be noted that the correlation coefficients are extremely low across the board (-0.1 to 0.12). Despite the low correlations, the top 10 rankings do contain linguistic events that seem logical considering the edge type. For instance, the person-to-location edges are mostly related to events that involve housing. There also exists a clear difference in the nature of the events related to person-to-person edges ("carry", "portrait", "funeral") versus person-to-organization edges ("establish", "council"). What is perhaps most interesting, is the fact that many of the events that occur in the top 10 for one edge type, also occur in the bottom 10 of the other edge types. This indicates that the events in question could serve well to classify edges in cases where the node classes, and thus the edge types, are unknown. For the sub type events, we observe similar effects, therefore we do not display those results here. The main type events are limited to four global classes: "event", "action", "development", and "change". In Figure 5.6, we visualize the correlations between the edge types and each of the four main event types in a force-directed graph. Even though the actual correlations are very low (0.002 to 0.03), the differences between them cause the coloured edge type nodes to move in the direction of the main type that they are more related to. The person-to-person edges are more correlated to "event" and "action", while person-to-location edges are only related to "change". Person-to-organization edges have correlations to all four main event types, though it is also most strongly related to "change". The difference we see between strict person edges on the one hand, and edges connected to organizations or locations on the other hand follows our view that they are inherently different types of entities with different roles in the network. The slight correlations between person-to-organization edges and all four main types is also in line with our assumption that organizations, in the form of collectives of people, do have some ability to act as independent entities. Still, in essence any action originates from a person, as is reflected beautifully in Figure 5.6.

5.4. Discussion

In this Chapter, we have identified the basic characteristics of social networks in general and checked if our BWSA network model adheres to these properties. We find that the network agrees with the small-world principle both when considered

as a static, aggregated graph, as well as when split into a series of dynamically evolving graphs at a lower temporal scale. The growth mechanism governing the evolutionary processes is one of preferential attachment. When we decrease the temporal scale to intervals of one year, we observe a broad scale, rather than a scale free network, which indicates a disturbance in the preferential attachment. This is to be expected with a network that spans several decades, since people may die or fall from grace and no longer make any new connections.

We are able to successfully distinguish the most influential people in the network, both locally and globally, by calculating rank correlations on betweenness centrality-ranked nodes. Contrary to what is reported by Braha & Bar-Yam (2006), we discover high correlations between node rankings at different temporal scales. To us this is proof of the tight-knit nature of the BWSA community on the one hand, and the consistency of our graph model on the other hand. We have gathered enough evidence to answer our third research question:

> **RQ 3** Do social network models constructed with the described method adhere to properties commonly observed in social networks?

Overall, our model complies very well with all characteristics of a valid social network. Compared to a randomly generated graph of the same size and with the same number of edges, the BWSA graph demonstrates consistency and structure over randomness. Moreover, the linguistic events attached to the edges make sense semantically in regards to the types of nodes that the edge connects to. The correlations of these events to the specific edge types are very low, which is mostly due to the large number of distinct base type events. Because there are so many, each of them is likely to occur only a few times. There are only four main type events, which occur at a higher rate and are consequently better suited to separate the different edge types. However, the reduction in event types also implies a reduction of the semantic space and makes the edges more difficult to interpret. Sub type events might provide a happy medium. The number of distinct events in this category comes to 310 for the BWSA, which is a completely manageable number.

6 DISCUSSION AND CONCLUSIONS

I discovered that if one looks a little closer at this beautiful world, there are always red ants underneath.
- David Lynch, *Lynch on Lynch*, 1997

In this thesis we set out to develop a method for the automatic extraction of social networks from free text. Specifically, we applied our method to a collection of biographical texts in Dutch from the domain of Social History, with the purpose of providing an innovative way to look at old data. Throughout, our intention has been to prove the validity and use of automated, computational methods when applied to soft sciences. We conclude the thesis in this chapter, which is structured as follows. In Section 6.1 we provide the answers to our three research questions. In Section 6.2 we answer our main problem statement. Section 6.3 details the contributions of this thesis. Finally, we present recommendations for future research in Section 6.4.

6.1. Answers to Research Questions

In this Section, we summarize the answers to the three research questions defined in Chapter 1. The first two research questions naturally arose when we considered the building blocks required to create a social network: in its essence, a network consists of nodes and edges. In Chapter 3, we identified the subjects of the biographies in our dataset to be the most suitable candidates for the nodes in our network, since they are known up front and can therefore be easily validated. We recognized that inclusion of other often-occurring entities, such as organizations and locations can help to strengthen connections and to uncover previously hidden patterns. To gather the actual data for the graph, we therefore needed to locate and identify all references to these entities in the text, which lead us to formulate our first research question:

RQ 1 To what degree can we reliably recognize and identify named entities in Dutch biographical text using state-of-the-art techniques?

To answer RQ 1 we took a two-step approach. First, we performed and evaluated recognition of named entities on our data using a retrained version of an existing

state-of-the-art classifier. When compared to Dutch newspaper data, the experiments show that entities occur at a higher rate in our biographical data. Consequently, less training data seems to be needed to achieve comparable results on this genre. We achieve competitive scores with a classifier trained only on biographical data, though the result is improved when the classifier is trained on a combination of genre-specific (i.e. biographical) data and non-specific (i.e. newspaper) data and supplemented with a domain-specific gazetteer. The final results are at the high end of the spectrum of what is currently achievable regarding automated entity detection and classification. Second, we implemented identification of named entities for Dutch by adapting an existing method for English, as no method existed yet for Dutch. In the identification phase, we did not make use of any external identification services under the assumption that all the information required to disambiguate the entities is included in the source itself.

Ultimately, recognition and identification of person names is most successful, with the end result reaching scores of around 85 %. Organizations and locations score slightly lower (82 % and 75 %, respectively). This means that, depending on entity class, the error margin is expected to be between 15 and 25 %. The identification process in essence serves to cluster entities with similar surface forms and contexts. This clustering itself provides a structured view on the data that can serve to quickly recognize and correct errors. Moreover, most errors occur in the identification of rarely occurring entities, while references to the more prominent community members are easily distinguished. As research into social networks shows, changes at the level of the least important nodes in a graph have little effect on the global structure and composition of the graph. To validate the accuracy of the network's composition we compared it to a similar structure constructed from hyperlinks in our original dataset that were manually added by domain experts. The graphs show comparable statistics and high correlation regarding node rankings, which leads us to the conclusion that the recognition and identification of named entities by our system is sufficiently reliable for the purpose of social network extraction.

Static edges are not well suited for the analysis of flow through a graph, especially when they are aggregated over multiple centuries, as is the case with the full network model of our dataset. Keeping in mind the goal of studying dynamics on the network models produced by our method, we formulated the second research question.

> **RQ 2** To what degree and level of specificity can we reliably recognize and normalize temporal information in Dutch biographical text using state-of-the-art techniques?

Through our experiments regarding analysis of temporal cues we found that dates are most easily recognized and normalized, reaching an accuracy of over 95 %. Other types of temporal expressions, such as durations, are not normalized with enough accuracy to include them in the network construction process. However,

dates cover over 70 % of the temporal expressions recognized in the data, so we can confidently conclude that we are able to capture at least that much of the described timeline. The mere construction of the graph from its parts is insufficient for our purpose of developing a new tool for research in the social sciences. To prove its actual worth, we need to validate the network model, as is described by our third, and final, research question.

> **RQ 3** Do social network models constructed with the described method adhere to properties commonly observed in social networks?

We answer RQ 3 positively, assured by the statistical analysis of the graph as it is provided in Chapter 5. The network model that we have constructed using our method is shown to be compatible with the small-world principle, both when considered as a static aggregated graph, and as a dynamic graph at varying temporal intervals. We can also reliably identify the growth mechanisms that are at work in the graph and prove that they sufficiently differ from the mechanisms in a random network to be considered intentional. Furthermore, we semantically validate the enrichments provided through Temporal Analysis in Chapter 4 by relating the events attached to edges to the classes of their associated nodes. Even though the sparseness of the data prevents us in finding any highly significant correlations, the analysis does show event-node relations that seem to be semantically motivated. This finding underlines the power of the method and its potential for application to other, larger datasets.

6.2. Answer to Problem Statement

The potential of Social Network Analysis as an investigative tool is recognized by many, though the ability to create large-scale datasets suitable for this purpose is generally limited to the more computationally advanced researchers. In recent years, much scientific effort has gone into the crossover between the humanities and computer sciences to allow each field to benefit from achievements made in the other. This thesis is an example of such an effort. Regardless of such endeavors, many humanities researchers and social scientists remain reluctant to adopt purely computational methods into their toolkit, because they see no room for hard analyses on soft data, or because they do not trust the results provided by a machine. Our work attempts to bridge the gap between hard and soft and show the validity of cross-domain research, while simultaneously enabling large-scale SNA on textual data. To this end we formulated the following problem statement.

> **PS** Can computational methods be used to successfully extract a detailed social network from historical, textual data, enriching the data in such a way that is of added value to social historical research?

This problem statement yields two main criteria that our method should meet. First of all, it should produce graphs that accurately model the social network of the

community described by the input data. Second, the information encoded in the graph should be sufficiently detailed to enable sensible semantic analysis of its contents. The first criterion serves to ensure that the output contains as little noise as possible. We do not wish to frustrate the social scientist confronted with the graph by including erroneous data, since this would work against our goal of bringing computational and social sciences closer together. We feel confident in stating that our method succeeds in filtering the information from the noise to an agreeable level. The majority of errors involve entities and temporal expressions that occur less frequent in the data and, as such, the errors do not have a detrimental effect on the final outcome. Moreover, we believe that visualization of the results after processing is a suitable way of validating the quality of the analysis. Domain experts often have a picture in their mind of what their domain looks like, so they will easily recognize elements that contradict their own view and can correct them in the graph. This, combined with the fact that automatic processing of the data is exponentially faster than manual processing, leads us to conclude that our method provides a great increase in efficiency to the social historian, even if he is not directly interested in social networks. For instance, our method can also be applied as a form of document indexing, as it records the locations and contexts of many of the key elements in the data. The graph can then be used to quickly select the (parts of) documents that comply with a certain query. The structural properties of the graph allow queries that are more complex than the Boolean queries currently enabled on the data source. For instance, one could search for all documents that mention people moving to or from Amsterdam, or for all people that interact with organizations that have the word "socialist" in their name. On top of that, our experiments have shown that linguistic events can be sensibly classified according to their nature based on the edges that they are attached to. This allows for different contexts to be defined over the graph (e.g. professional versus personal) that can be used to accurately filter the data to fit very particular research questions. In that, our graph also meets the second criterion, leading us to an overall positive answer to our problem statement.

6.3. Thesis Contributions

We have presented experiments and analyses regarding the automated extraction of social networks from Dutch biographical text with the aim to provide a new and efficient way of gathering data to the field of Social Historical research. From a computational point of view, this requires approaches to Named Entity Recognition and Disambiguation for the identification of nodes, and a method for Temporal Analysis to bind the edges to a timeline. From the perspective of the social historian, the results produced by the process should serve to increase their trust in automated methods of data analysis in general by providing an accurate, clean model that is easily validated. We identify five contributions that we have made through the current thesis.

1. We have examined the suitability of state-of-the-art Named Entity Recognition for Dutch biographical text on the classes "person", "organization", and "location". The results were compared to those obtained on Dutch newspaper data. Our approach is shown to be competitive on both genres, though the entities in the more focused domain of the biographies are more easily recognized using less training data.

2. We have adapted an existing competitive method for Named Entity Disambiguation that was initially developed for English to the Dutch language. The method is tested on the biographical genre and reaches scores that are at par with the state-of-the-art for English and other languages. In doing so, we have provided the first comprehensive study of Named Entity Disambiguation on real-world data for Dutch.

3. We have improved on the current state of the art for Dutch Temporal Analysis by developing a method that incorporates a more accurate normalization of dates, as well as the identification of linguistic events in relation to the timeline.

4. Through statistical analysis of the network model that arises from the detection of entities, temporal expressions, and events, we have shown that such a graph constructed automatically from unstructured input sufficiently agrees with a graph constructed from manual annotations on the same data, and that it complies with properties commonly found in social networks. As such, we have validated that the results obtained through our approach are suitable for further application in a scientific context.

5. The combination of the previous four contributions forms the final product and key contribution of this thesis. Namely, we have described a new investigative tool for the social sciences that can, at least in theory, be applied to any collection of documents, and enables also the less technologically savvy researcher to confidently perform SNA on his own data. As such, we have provided evidence that automated methods can definitely be advantageous to fields such as Social History, where the overall tendency still is to avoid these methods.

6.4. Future Research

Future endeavors based on the current research can be globally divided into two directions: improvement of the subparts of the system to maximize graph accuracy, and application of the method to broader problems. So far, we have tested our method only on the biographical genre within a historical domain. An obvious avenue for future research is the testing of its applicability to datasets of different sizes, genres, and domains. Either way, the accuracy of graphs constructed using our approach principally depends on the outcome of the Named Entity

Recognition, Named Entity Disambiguation, and Temporal Analysis processes. As we have seen, the quality of entity recognition is already high, so we see little room for improvement there, with the exception of extending the types of entities that are recognized by the system. Still, the main concern in this area remains the availability of sufficient amounts of training data that are compatible to the target data, which is a general concern in the application of computational methods and not an issue that can be resolved through further research into named entities. Since we are already able to reliably recognize the vast majority of entity occurrences, focusing on improving name disambiguation seems a more promising avenue. Our dataset was supplemented with a database listing the main entities, which was a great help in the identification of the entities and the evaluation of the graph. However, a truly unstructured dataset does not come with such information. In those cases, the system would greatly benefit from access to an external service for identification of names, such as Wikipedia or GeoNames.

The greatest improvement can be achieved on the analysis of temporal expressions. We were only able to successfully extract edges for dates, while durations and repetitions are currently left out of the ultimate results. The inclusion of repetitions especially would be very interesting from a SNA perspective, since it would allow the study of habitual patterns. Our work shows that, in essence, these types can be recognized by the current system, but they do not occur at a high enough rate to be reliably detected. Therefore, we expect that recognition would most be helped by the inclusion of more examples in the training data, either through general increase of training data or oversampling of just those two classes in the training data. Temporal normalization is currently solved in a rule-based setup, which also contributes to the poor results for durations and repetitions. Again, more training samples could lead to more accurate rules for the normalization of these expressions.

The validity of the graph is currently tested only against itself. Future efforts should include a comparison of multiple graphs constructed from different datasets to fully validate the method's ability to extract accurate social network models. Besides social network construction and document indexing, we also foresee applications of our method in automatic summarization and storyline extraction. Both involve detection and tracking of the most salient pieces of information and logically ordering them against a timeline, which is exactly what our method does. We have focused more on the binding of edges to the timeline and events than on the actual flow of events through the graph. In our analysis, events appear to be suitable for edge classification purposes (or vice versa). Further research should be conducted to test whether chains of temporally consecutive events show the same discriminative power.

REFERENCES

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics , 74* (1), 47.

Alfonseca, E., & Manandhar, S. (2002). An unsupervised method for general named entity recognition and automated concept discovery. *Proceedings of the 1st international conference on general WordNet*, (p. 34). Mysore, India.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM , 26* (11), 832-843.

Alstott, J., Bullmore, E., & Plenz, D. (2014). powerlaw: a Python package for analysis of heavy-tailed distributions. *PLOS ONE , 9* (1).

Amaral, L. A., Scala, A., Barthelemy, M., & Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Sciences , 97* (21), 11149-11152.

Anthonisse, J. (1971). The rush is a directed graph. *Stichting Mathematisch Centrum. Mathematische Besliskunde* (BN 9/71), 1-10.

Artiles, J., Gonzalo, J., & Sekine, S. (2007). The semeval-2007 weps evaluation: Establishing a benchmark for the web people search task. *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 64-69). Association for Computational Linguistics.

Artiles, J., Gonzalo, J., & Sekine, S. (2009). Weps 2 evaluation campaign: overview of the web people search clustering task. *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, (p. 9).

Artiles, J., Gonzalo, J., & Verdejo, F. (2005). A testbed for people searching strategies in the WWW. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 569-570). ACM.

Asahara, M., & Matsumoto, Y. (2003). Japanese named entity extraction with redundant morphological analysis. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. 1*, pp. 8-15. Association for Computational Linguistics.

Bagga, A., & Baldwin, B. (1998). Entity-based cross-document coreferencing using the vector space model. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. 1*, pp. 79-85. Association for Computational Linguistics.

Baldwin, B., Morton, T., Bagga, A., Baldridge, J., Chandraseker, R., Dimitriadis, A., et al. (1998). Description of the UPenn CAMP system as used for coreference. *Proceedings of the Seventh Message Understanding Conference (MUC-7).*

Balog, K., Azzopardi, L., & De Rijke, M. (2008). Personal name resolution of web people search. *WWW2008 Workshop: NLP Challenges in the Information Explosion Era (NLPIX 2008).*

Balog, K., Azzopardi, L., & De Rijke, M. (2009). Resolving person names in web people search. In I. King, & R. Baeza-Yates, *Weaving services and people on the World Wide Web* (pp. 301-323). Springer.

Barabási, A.-L. (2009). Scale-Free Networks: A Decade and Beyond. *Science , 325* (5939), 412-413.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science , 286* (5439), 509-512.

Barkey, K., & Van Rossum, R. (1997). Networks of Contention: Villages and Regional Structure in the Seventeenth-Century Ottoman Empire 1. *American Journal of Sociology , 102* (5), 1345-1382.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *Journal of the acoustical society of America , 22*, 725-730.

Beauchamp, M. (1965). An improved index of centrality. *Behavioral Science , 10* (2), 161-163.

Bethard, S. (2013). ClearTK-TimeML: A minimalist approach to TempEval 2013. *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, 2, pp. 10-14.

Bick, E. (2004). A Named Entity Recognizer for Danish. *The 4th International Conference on Language Resources and Evaluation.* LREC.

Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1997). Nymble: a high-performance learning name-finder. *Proceedings of the fifth conference on Applied natural language processing* (pp. 194-201). Association for Computational Linguistics.

Black, W. J., Rinaldi, F., & Mowatt, D. (1998). FACILE: Description of the NE System Used for MUC-7. *Proceedings of the 7th Message Understanding Conference.*

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* .

Bollegala, D., Honma, T., Matsuo, Y., & Ishizuka, M. (2008). Automatically extracting personal name aliases from the web. In *Advances in Natural Language Processing* (pp. 77-88).

Bollegala, D., Matsuo, Y., & Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. *www* , *7*, 757-766.

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology* , 1170-1182.

Bonacich, P. (1991). Simultaneous group and individual centralities. *Social Networks* , *13* (2), 155-168.

Borgatti, S. (2005). Centrality and network flow. *Social Networks* , *27* (1), 55-71.

Borthwick, A. (1999). *A maximum entropy approach to named entity recognition.* New York University.

Braha, D., & Bar-Yam, Y. (2006). From Centrality to Temporary Fame: Dynamic Centrality in Complex Networks. *Complexity* , *12* (2), 59-63.

Brickley, D., & Miller, L. (2012). FOAF vocabulary specification 0.98.

Buchholz, S., & Van den Bosch, A. (2000). Integrating Seed Names and ngrams for a Named Entity List and Classifier. *LREC.*

Canisius, S., Bogers, T., Van den Bosch, A., Geertzen, J., & Tjong Kim Sang, E. (2006). Dependency parsing by inference over high-recall dependency predictions. *Proceedings of the Tenth Conference on Computational Natural Language Learning* (pp. 176-180). Association for Computational Linguistics.

Chen, Y., Lee, S. Y., & Huang, C.-R. (2009). Polyuhk: A robust information extraction system for web personal names. *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference.*

Chieu, H., & Ng, H. (2002). Named entity recognition: a maximum entropy approach using global information. *Proceedings of the 19th international conference on Computational linguistics. 1*, pp. 1-7. Association for Computational Linguistics.

Chinchor, N., & Hirschmann, L. (1997). MUC-7 coreference task definition, version 3.0. *Proceedings of MUC.*

Christakis, N., & Fowler, J. (2013). Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine , 32* (4), 556-577.

Clauset, A., Shalizi, C., & Newman, M. (2009). Power-law distributions in empirical data. *SIAM Review , 51* (4), 661-703.

Coleman, J., Katz, E., & Menzel, H. (1966). *Medical innovation: A diffusion study.*

Coleman, J., Katz, E., & Menzel, H. (1957). The diffusion of an innovation among physicians. *Sociometry* , 253-270.

Culotta, A., McCallum, A., & Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* (pp. 296-303). Association for Computational Linguistics.

Düring, M. (2015). *Verdeckte soziale Netzwerke im Nationalsozialismus: Die Entstehung und Arbeitsweise von Berliner Hilfsnetzwerken für verfolgte Juden.* De Gruyter Oldenbourg.

Daelemans, W., Zavrel, J., Van den Bosch, A., & Van der Sloot, K. (1996). Mbt: memory-based tagger. *Proceedings of the Fourth Workshop on Very Large Corpora* , 14-27.

De Benedictis, L., & Tajoli, L. (2011). The world trade network. *The World Economy , 34* (8), 1417-1454.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., & Weischedel, R. M. (2004). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. *LREC.*

Ekbal, A., & Bandyopadhyay, S. (2008). Bengali Named Entity Recognition Using Support Vector Machine. *IJCNLP*, (pp. 51-58).

Elson, D. K., Dames, N., & McKeown, K. R. (2010). Extracting social networks from literary fiction. *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 138-147). Association for Computational Linguistics.

Ferreira Da Silva, J. F., Kozareva, Z., Gabriel, J., & Lopes, P. (2004). Cluster Analysis and Classification of Named Entities. *Proceedings of the Conference on Language Resources and Evaluation.* LREC.

Ferrer i Cancho, R., Janssen, C., & Solé, R. (2001). Topology of technology graphs: Small world patterns in electronic circuits. *Physical Review E , 64* (4), 046119.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 363-370). Association for Computational Linguistics.

Forney Jr., G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE , 61* (3), 268-278.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry , 40*, 35-41.

Freeman, L. (1979). Centrality in Social Networks - Conceptual Clarification. *Social Networks , 1* (3), 215-239.

Fruhwirth, T. (2011). *Constraint handling rules.* BoD--Books on Demand.

Fruhwirth, T. (1998). Theory and practice of constraint handling rules. *The Journal of Logic Programming , 37* (1), 95-138.

Fulminante, F. (2012). Social Network Analysis and the Emergence of Central Places. *BABesch , 87*, 1-27.

Galaskiewicz, J. (1989). Interorganizational networks mobilizing action at the metropolitan level. *Networks of power: Organizational actors at the national, corporate, and community levels* , 81-96.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (6), 721-741.

Gooi, C. H., & Allan, J. (2004). *Cross-document coreference on a large scale corpus.* DTIC Document.

Gottlieb, B. H. (1981). *Social networks and social support* (Vol. 4). Sage Publications, Inc.

Grishman, R., & Sundheim, B. (1996a). Design of the MUC-6 evaluation. *Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996* (pp. 413-422). Association for Computational Linguistics.

Grishman, R., & Sundheim, B. (1996b). Message Understanding Conference-6: A Brief History. *COLING. 96*, pp. 466-471. COLING.

Hage, P., & Harary, F. (1983). *Structural models in anthropology.* Cambridge University Press.

Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R., & Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC bioinformatics , 6*.

Hansen, P. J., & Jurs, P. C. (1988). Chemical applications of graph theory. Part I. Fundamentals and topological indices. *Journal of Chemical Education , 65* (7), 574.
Heaps, H. S. (1978). *Information retrieval: Computational and theoretical aspects.* Academic Press, Inc.

Ikeda, M., Ono, S., Sato, I., Yoshida, M., & Nakagawa, H. (2009). Person name disambiguation on the web by two-stage clustering. *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference.*

Karsdorp, F., Kestemont, M., Schöch, C., & Van den Bosch, A. (2015). The Love Equation: Computational Modeling of Romantic Relationships in French Classical Drama. In M. A. Finlayson, B. Miller, A. Lieto, & R. Ronfard (Ed.), *Proceedings of the Workshop on Computational Models of Narrative (CMN'15)*, (pp. 98-107). Atlanta, USA.

Kautz, H., Selman, B., & Shah, M. (1997). Referral Web: combining social networks and collaborative filtering. *Communications of the ACM , 40* (3), 63-65.

Kazama, J., Makino, T., Ohta, Y., & Tsujii, J. (2002). Tuning support vector machines for biomedical named entity recognition. *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain. 3*, pp. 1-8. Association for Computational Linguistics.

Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 137-146). ACM.

Kossinets, G., & Watts, D. (2006). Empirical analysis of an evolving social network. *Science , 311* (5757), 88-90.

Krahmer, E., Van Erk, S., & Verleg, A. (2003). Graph-based generation of referring expressions. *Computational Linguistics , 29* (1), 53-72.

Kullback, S. (1968). *Information theory and statistics.* Courier Corporation.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The annals of mathematical statistics , 22* (1), 79-86.

Kumar, R., Novak, J., & Tomkins, A. (2010). Structure and Evolution of Online Social Networks. In P. Yu, J. Han, & C. Faloutsos, *Link Mining: Models, Algorithms, and Applications* (pp. 337-357). Springer.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning (ICML-2001).* Morgan Kaufmann.

Laumann, E. O., & Pappi, F. U. (1973). New directions in the study of community elites. *American Sociological Review* , 212-230.

Laumann, E. O., Marsden, P. V., & Galaskiewicz, J. (1977). Community-elite influence structures: Extension of a network approach. *American Journal of Sociology* , 594-631.

Lerman, K., Ghosh, R., & Hyung Kang, J. (2010). Centrality Metric for Dynamic Networks. *Proceedings of the Eighth Workshop on Mining and Learning with Graphs* (pp. 70-77). ACM.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady , 10* (8), 707-710.

Levine, J. H. (1972). The sphere of influence. *American Sociological Review* , 14-27.

Llorens, H., Derczynski, L., Gaizauskas, R. J., & Saquete, E. (2012). TIMEN: An Open Temporal Expression Normalisation Resource. *LREC*, (pp. 3044-3051).

Llorens, H., Saquete, E., & Navarro, B. (2010). Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 284-291). Association for Computational Linguistics.

Maes, A., & Oversteegen, L. (1999). Nominal and temporal interpretation in discourse. *Issues in Cognitive Linguistics* , 549-566.

Mann, G. S., & Yarowsky, D. (2003). Unsupervised personal name disambiguation. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. 4*, pp. 33-40. Association for Computational Linguistics.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge university press.

Mason, O., & Verwoerd, M. (2007). Graph theory and networks in biology. *Systems Biology , 1* (2), 89-119.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools , 13* (1), 157-169.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., & Wilks, Y. (2001). Named entity recognition from diverse text types. *Recent Advances in Natural Language Processing 2001 Conference*, (pp. 257-274).

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003. 4*, pp. 188-191. Association for Computational Linguistics.

Mehler, A. (2008). Large text networks as an object of corpus linguistic studies. *Corpus linguistics. An international handbook of the science of language and society.* , 328-382.

Mika, P. (2005). Flink: Semantic web technology for the extraction and analysis of social networks. *Web Semantics: Science, Services and Agents on the World Wide Web , 3* (2), 211-223.

Mikheev, A., Moens, M., & Grover, C. (1999). Named entity recognition without gazetteers. *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics* (pp. 1-8). Association for Computational Linguistics.

Milgram, S. (1967). The Small-world Problem. *Psychology Today , 2* (1), 60-67.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes , 6* (1), 1-28.

Minkov, E., Wang, R. C., & Cohen, W. W. (2005). Extracting personal names from email: Applying named entity recognition to informal text. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 443-450). Association for Computational Linguistics.

Mizruchi, M. S., & Schwartz, M. (1992). *Intercorporate relations: the structural analysis of business* (Vol. 1). Cambridge University Press.

Moreno, J. (1960). *The sociometry reader.* Free Press.

Morone, F., & Makse, H. A. (2015). Influence maximization in complex networks through optimal percolation. *Nature , 524* (7563), 65-68.

Murdock, G. P., & White, D. R. (1969). Standard cross-cultural sample. *Ethnology* , 329-369.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes , 30* (1), 3-26.

Nadeau, D., Turney, P., & Matwin, S. (2006). Unsupervised named-entity recognition: Generating gazetteers and resolving ambiguity. *Advances in Artificial Intelligence: 19th Conference of the Canadian Society for Computational Studies of Intelligence. 4013*, p. 266. Quebec City: Canadian AI.

Naroll, R. (1965). Galton's problem: The logic of cross-cultural analysis. *Social Research* , 428-451.

Naroll, R. (1961). Two solutions to Galton's problem. *Philosophy of Science* , 15-39.

Nieminen, J. (1974). On centrality in a graph. *Scandinavian Journal of Psychology , 15*, 322-336.

Oversteegen, L. (2005). Causality and Tense—Two Temporal Structure Builders. *Journal of semantics , 22* (3), 307-337.

Padgett, J. F., & Ansell, C. K. (1993). Robust Action and the Rise of the Medici, 1400-1434. *American journal of sociology* , 1259-1319.

Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical review letters , 86* (14), 3200.

Pedersen, T., Purandare, A., & Kulkarni, A. (2005). Name discrimination by clustering similar contexts. In *Computational Linguistics and Intelligent Text Processing* (pp. 226-237).

Puscasu, G. (2007). Wvali: Temporal relation identification by syntactico-semantic analysis. *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 484-487). Association for Computational Linguistics.

Pustejovsky, J., Castano, J. M., Ingria, R., Saurí, R., Gaizauskas, R. J., Setzer, A., et al. (2003). TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering , 3*, 28-34.

Ramshaw, L. A., & Marcus, M. P. (1995). Text chunking using transformation-based learning. *arXiv preprint cmp-lg/9505040* .

Ratinov, L., Roth, D., Downey, D., & Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. *Proceedings of the 49th Annual*

Meeting of the Association for Computational Linguistics: Human Language Technologies. 1, pp. 1375-1384. Association for Computational Linguistics.

Ravichandran, D., & Hovy, E. (2002). Learning Surface Text Patterns for a Question Answering System. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 41-47). Philadelphia: Association for Computational Linguistics.

Rogers, D. (1974). Sociometric analysis of interorganizational relations: Application of theory and measurement. *Rural Sociology , 39* (4), 487-503.

Rogers, E. M. (1979). Network analysis of the diffusion of innovations.

Sairio, A. (2008). A social network study of the eighteenth-century Bluestockings: the progressive and preposition stranding in their letters. *Historical Sociolinguistics and Sociohistorical Linguistics , 8*.

Salton, G. (1989). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. *Reading: Addison-Wesley* .

Schuurman, I. (2008). Spatiotemporal annotation using MiniSTEx: How to deal with alternative, foreign, vague and/or obsolete names? *LREC*.

Schuurman, I., Hoste, V., & Monachesi, P. (2010). Interacting Semantic Layers of Annotation in SoNaR, a Reference Corpus of Contemporary Written Dutch. *LREC*.

Shaw, M. (1954). Group structure and the behaviour of individuals in small groups. *Journal of Psychology , 38*, 139-149.

Strötgen, J., & Gertz, M. (2010). HeidelTime: High quality rule-based extraction and normalization of temporal expressions. *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 321-324). Association for Computational Linguistics.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of CoNLL-2002* (pp. 155-158). Association for Computational Linguistics.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* (pp. 142-147). Association for Computational Linguistics.

Tjong Kim Sang, E., & Veenstra, J. (1999). Representing text chunks. *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics* (pp. 173-179). Association for Computational Linguistics.

UzZaman, N., Llorens, H., Allen, J., Derczynski, L., Verhagen, M., & Pustejovsky, J. (2012). Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333* .

Van de Camp, M., & Christiansen, H. (2013). Resolving relative time expressions in Dutch text with Constraint Handling Rules. In D. Duchier, & Y. Parmentier, *Constraint Solving and Language Processing* (pp. 166-177). Orléans, France: Springer.

Van den Bosch, A., Busser, B., Canisius, S., & Daelemans, W. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. *LOT Occasional Series* , 191-206.

Van Rijsbergen, C. J. (1979). *Information Retrieval* (2 ed.). Butterworths.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., & Pustejovsky, J. (2007). Semeval-2007 task 15: Tempeval temporal relation identification. *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 75-80). Association for Computational Linguistics.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., & Pustejovsky, J. (2009). The TempEval challenge: identifying temporal relations in text. *Language Resources and Evaluation , 43* (2), 161-179.

Verhagen, M., Saurí, R., Caselli, T., & Pustejovsky, J. (2010). SemEval-2010 task 13: TempEval-2. *Proceedings of the 5th international workshop on semantic evaluation* (pp. 57-62). Association for Computational Linguistics.

Vossen, P., Maks, I., Segers, R., Van der Vliet, H., Moens, M.-F., Hofmann, K., et al. (2013). Cornetto: a combinatorial lexical semantic database for Dutch. In *Essential Speech and Language Technology for Dutch* (pp. 165-184). Springer.

Wan, X., Gao, J., Li, M., & Ding, B. (2005). Person resolution in person search results: Webhawk. *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 163-170). ACM.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis - Methods and Applications.* Cambridge University Press.

Wasserman, S., & Galaskiewicz, J. (1989). Mimetic Processes within an Interorganizational Field: An Empirical Test. *Administrative Science Quarterly , 34* (3), 454-479.

Watts, D., & Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature , 393* (6684), 440-442.

Wellman, B., & Wortley, S. (1990). Different strokes from different folks: Community ties and social support. *American journal of Sociology* , 558-588.

West, D. (2001). *Introduction to graph theory* (Vol. 2). Upper Saddle River: Prentice hall.

Zervanou, K., Korkontzelos, I., Van den Bosch, A., & Ananiadou, S. (2011). Enrichment and structuring of archival description metadata. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 44-53). Association for Computational Linguistics.

LIST OF TABLES

Table 3.1 - Descriptive statistics regarding the training and test sets for the NER experiments. For each set the total number of annotated entities is given, their frequency per 1,000 tokens, followed by the average number of occurrences per unique entity, which indicates their global consistency. The overlap measure expresses the percentage of the total number of entity types in the training set that is also included in its accompanying test set, and vice versa... 27
Table 3.2 - Percentages of entity type overlap per entity category between all three datasets. The training and test sets have been merged for this purpose. 27
Table 3.3 - F1 scores for the different named entity classes trained and tested on all three datasets separately, and combined. 35
Table 3.4 - F1 scores for the different named entity classes trained on the combined training sets and the separate training sets, and tested on each separate test set.35
Table 3.5 - Cross-document disambiguation results obtained on the development and test sets using the best performing CosSim algorithm and the best performing baseline. All reported scores are F1 scores..... 37
Table 3.6 – Statistical comparison of the baseline and NED P-P graph models to the gold standard HTML graph. 44
Table 4.1 - Overview of modal and aspectual auxiliary verbs occurring in the BWSA60
Table 4.2 - Top ten event nouns in the BWSA 60
Table 4.3 - Relations according to Allen's logic..... 63
Table 4.4 – F1 scores for TIMEX3 recognition. Scores reported for the training set are averages taken over ten 10-fold cross validation experiments. Scores marked with * are significantly better than the lowest scoring system (p < 0.0005). Scores marked with ** are significantly better than all other systems (p < 0.0005). The scores for the development and test sets are F1 scores for a single run on the entire set. The highest score for each class in each set is displayed in bold..... 65
Table 4.5 – F1 scores for 2-step TIMEX3 identification. Scores reported for the training set are averages taken over ten 10-fold cross validation experiments. Scores marked with * are significantly better than the lowest scoring system (p < 0.005). Scores marked with ** are significantly better than all other systems (p < 0.005). The scores for the development and test sets are F1 scores for a single run on the entire set. The highest score for each class in each set is displayed in bold..... 65
Table 4.6 – F1 scores for 1-step TIMEX3 recognition and identification training. Scores reported are averages taken over ten 10-fold cross validation experiments on the training set. Scores marked with * are significantly better than the lowest scoring system (p < 0.005). Scores marked with ** are

significantly better than all other systems ($p < 0.005$). The highest score for each class is displayed in bold. 65
Table 4.7 – F1 scores for 1-step TIMEX3 recognition and identification on the development set. The highest score for each class is displayed in bold. 67
Table 4.8 – F1 scores for 1-step TIMEX3 recognition and identification on the test set. The highest score for each class is displayed in bold. 67
Table 4.9 – Normalization baselines (HeidelTime) per TIMEX3 class 70
Table 4.10 – Normalization results per TIMEX3 class 70
Table 4.11 – Average F1 scores on the development and test sets for event extraction 71
Table 4.12 – Descriptive statistics of the temporal analysis of the BWSA. The left column shows the averages over the gold annotations (100 biographies) for TIMEX3 and EVENT, with automated TLINK extraction. The right column shows the same numbers measured on the fully automated analysis of the 473 remaining, unannotated biographies. 74
Table 4.13 – Statistical comparison of the NED P-P and TIMEX P-P graph models to the gold standard HTML graph. 77
Table 5.1 – Examples of diffusion across networks classified according to their method of transition and their trajectory type. Also listed are the most appropriate measures of centrality. Borrowed and adapted from (Borgatti, 2005). 87
Table 5.2 – Composition of full, decade, and year graphs for BWSA and random sets 92
Table 5.3 – Distribution of graphs per set over degree distribution types. Power law distributions are an indication of preferential growth in small-world networks. 92
Table 5.4 – The 10 most and least correlated base type events for person-to-person, person-to-organization, and person-to-location edges in the full BWSA graph. 96

LIST OF FIGURES

Figure 2.1 – Introduction of the BWSA biography of Ferdinand Domela Nieuwenhuis. The parts marked in green are included as fields in the BWSA database. Parts marked in blue and orange hold useful information regarding Domela's life (parents, spouses, dates of weddings etc.) and can be extracted using named entity extraction and temporal expression analysis, respectively. 10
Figure 2.2 – Vocabulary growth curve for the BWSA compared to the SoNaR-1 reference corpus of contemporary Dutch 10
Figure 2.3 – Alphabetical index of biographies on the BWSA website, available at http://socialhistory.org/bwsa/bios. The middle column lists all available biographies with a short description about the biographee. The right column lists people that were born or passed away on the current day of the year, followed by recent edits made to the collection. 11
Figure 2.4 – Example record from the BWSA database showing the information for Ferdinand Domela Nieuwenhuis. The record overlaps with the green parts in Figure 2.1. Non-informative fields pertaining to database configuration are not displayed. 11
Figure 3.1 - *a.* Hierarchical tree graph representing a straightforward top-to-bottom information flow within a small organization. *b.* Graph representation of the information flow in the same organization if all communication to and from node F is mistakenly attributed to node E..... 22
Figure 3.2 - Example of data annotated with named entities in the BIO-format. *B-ORG* means that 'Centraal' is the first token in an organization name; *I-ORG* means that 'Woningbeheer' is the consecutive token in the same name. Since the next token ',' is tagged with *O*, it is not considered as part of the named entity and the full name is, consequently, 'Centraal Woningbeheer'. The next named entity is a one token person name, 'Baart'. 26
Figure 3.3 - Left: Within-document disambiguation results for, respectively, person, organization, and location names averaged over the 10 documents in the BWSA development set. All reported scores are F1 scores. The labels on the x-axis represent the string similarity threshold. Right: Results obtained on the development and test sets using the best performing CosSim algorithm and the best performing baseline. 36
Figure 3.4 - *a.* Hierarchical tree graph representing a straightforward top-to-bottom information flow within a small organization, where one node has been wrongly split into two nodes. *b.* Graph representation of the information flow in the same organization where the two nodes are merged based on common attributes. 41

Figure 3.5 – a. Force-directed visualization of the links between the BWSA HTML pages. b. Force-directed visualization of the static social network of BWSA biographees constructed from sentence-level co-occurrences. In both graphs, the nodes are coloured by modularity class to distinguish communities, and sized by degree (number of adjacent edges) as a crude measure of influence. Only the most prominent nodes are labelled.....45
Figure 4.1 - Example of the rule-based EVENT extraction process applied to one sentence of the BWSA.....61
Figure 4.2 - Example of the rule-based TLINK classification process applied to one sentence of the BWSA.....64
Figure 4.3 – Example of edge consolidation through a common node. Left: graph before consolidation. Right: graph after consolidation of edges through node A.....78
Figure 4.4 – Force-directed visualization of the social network of BWSA biographees between January 1, 1860 and December 31, 1919 constructed from sentence-level co-occurrences. The nodes are coloured by modularity class to distinguish communities, and sized by degree. Only the most prominent nodes are labelled.....79
Figure 4.5 – Percentage of unlikely, possible, and expected nodes connected in graphs consolidated over common connections to non-BWSA people (TIMEX P-p-P), organizations (TIMEX P-o-P), and locations (TIMEX P-l-P) with minimum edge weights varying from 1 to 3.....79
Figure 5.1 – Output of the Watts & Strogatz algorithm on a graph with 12 nodes and k = 4, with p varying from 0 to 1. As p increases, the average path length decreases (Watts & Strogatz, 1998).....84
Figure 5.2 – Nodes ranked according to four types of centrality. Green implies higher centrality, while red implies lower centrality.....86
Figure 5.3 – Model of a dynamic graph. Figure a shows the graph aggregated over all time frames; figures b-d show the graphs for each individual time frame.....89
Figure 5.4 – Average path lengths and average clustering coefficients measured on the BWSA graphs versus randomly generated graphs.....91
Figure 5.5 – Average Spearman rank correlations of betweenness centrality rankings of nodes. Correlations are calculated between each graph and all of its preceding graphs at the same temporal scale – year or decade – and then averaged over the distance between the current and the preceding graph (top and bottom left). Correlations are also calculated between each BWSA-year graph and the BWSA-decade graph of the corresponding decade, and then averaged over the decade (bottom right). Significant correlations are displayed in dark grey ($p < 0.05$).....94
Figure 5.6 – Visualization of the Pearson product-moment correlation coefficients between person-to-person (P-P), person-to-organization (P-O), and person-to-location (P-L) edges and events with main type "event", "action", "development", and "change" in the full BWSA graph. A line indicates a positive correlation. Negative correlations are expressed by the absence of a line. A thicker line implies a stronger correlation, though all correlations displayed are weak (0.002-0.03) and not significant.....96

APPENDIX A

Alphabetical list of all biographies included in the version of the BWSA used for this research.

Biography identifier	Name	Lifespan
aalberse	Petrus Josephus Mattheus Aalberse	1871 - 1948
adama-van-scheltema	Carel Steven Adama van Scheltema	1877 - 1924
albarda	Johan Willem Albarda	1877 - 1957
alma	Petrus Alma	1886 - 1969
amelink	Herman Amelink	1881 - 1957
andreae	Wabina Andreae	1874 - 1966
andriessen	Wilhelmus Johannes Andriessen	1894 - 1978
ankersmit	Johan Frederik Ankersmit	1871 - 1942
ankersmit-g	Gerharda Johanna Helena Ankersmit	1869 - 1944
ansing	Willem Ansing	1837 - 1900
arbeid	Isaäc Arbeid	1882 - 1944
ariens	Alphonse Marie Auguste Joseph Ariëns	1860 - 1928
baan	Leendert de Baan	1880 - 1929
baars	Asser Baars	1892 - 1944
baart-l	Lucretia Jacoba Baart	1850 - 1932
baart-m	Maria Elize Baart	1854 - 1879
baart-s	Servatius Protatius Baart	1871 - 1950
baas	Gerrit Baas Kzn.	1884 - 1978
bahlman	Ignatius Bernardus Maria Bahlmann	1852 - 1934
bakker	Sybe Kornelis Bakker	1875 - 1918
bakker-j	Jacobus Bakker	1875 - 1950
bakker-p	Pieter Oege Bakker	1897 - 1960
banning	Willem Banning	1888 - 1971
bartels	Derk Bartels	1883 - 1937
bax	Willem Bax	1836 - 1918
baye	Hermanus Franciscus Baye	1852 - 1894
beekman	Christiaan Hendrik Beekman	1887 - 1964
bella	Simon de la Bella	1889 - 1942
bennink	Gerrit Bennink	1858 - 1927
berdenis	Mathilde Berdenis van Berlekom	1862 - 1952
berger	Johannes Adriaan Berger	1895 - 1961
bergh	George van den Bergh	1890 - 1966
bergh-van-eysinga	Henri Wilhelm Philippus Elize van den Bergh van Eysinga	1868 - 1920
bergmeijer	Jan Andries Bergmeijer	1854 - 1941

bergsma	Pieter Bergsma	1882 - 1946
berlage	Hendrik Petrus Berlage	1856 - 1934
besuijen	Karel Paulus Willem Besuijen	1880 - 1916
beuzemaker	Nicolaas Beuzemaker	1902 - 1944
beversluis	Martinus Beversluis	1894 - 1966
bleekrode	Meijer Bleekrode	1896 - 1943
boekhorst	Johannes Petrus Antonius te Boekhorst	1862 - 1944
boekman	Emanuel Boekman	1889 - 1940
boer-h	Harmen de Boer	1887 - 1941
boer-k	Klaas de Boer	1883 - 1945
boers	Benjamin Boers	1871 - 1952
bokkel	Jan Gerhard ten Bokkel	1856 - 1932
bokma-de-boer	Sjoukje Maria Diderika Bokma de Boer	1860 - 1939
bonger	Willem Adriaan Bonger	1876 - 1940
borgesius	Hendrik Borgesius	1847 - 1917
bos-d	Dirk Bos	1862 - 1916
bos-k	Karel Antonie Bos	1846 - 1899
bosch	Johannes van den Bosch	1780 - 1844
bosch-kemper	jonkvrouw Jeltje de Bosch Kemper	1836 - 1916
boshart	Maurits Boshart	1905 - 1964
bosman	Fokke Bosman	1893 - 1971
bot	Lambertus Johannes Bot	1897 - 1988
bouwman	Bertus Bouwman	1882 - 1955
bouwmeester	Maria Catharina Bouwmeester	1885 - 1956
braambeek	Hendrik Jan van Braambeek	1880 - 1960
brandsteder	Jacob Andries Brandsteder	1887 - 1986
brautigam	Johan Brautigam	1878 - 1962
brink	Johannes Antonius Hendrikus van den Brink	1865 - 1933
brinkhuis	Johannes Brinkhuis	1860 - 1938
broeksz	Johannes Bartholomeus Broeksz	1906 - 1980
brok	Klaas Cornelis Brok	1873 - 1944
brommert	Johannes Brommert	1891 - 1975
brouwer	Hendrik Brouwer	1870 - 1946
bruens	Hendrikus Johannes Bruens	1865 - 1924
bruijn	Adrianus Cornelis de Bruijn	1887 - 1968
bruin	Pieter de Bruin	1879 - 1957
bruins-a	Angenita Engelina Johanna Bruins	1874 - 1957
bruins-j	Jan Anthonie Bruins	1872 - 1947
bruinsma	Vitus Jacobus Bruinsma	1850 - 1916
bult	Franciscus Xaverius Wilhelmus Bult	1866 - 1925
burink	Gerrit van Burink	1891 - 1961
buskes	Johannes Jacobus Buskes Jr.	1899 - 1980
bymholt	Berend Bymholt	1864 - 1947
ceton	Jan Cornelis Ceton	1875 - 1943
clercq	Daniël de Clercq	1854 - 1931
cohen-f	Frederika Sophia Cohen	1903 - 1943

cohen-j	Jozef Alexander Cohen	1864 - 1961
cohen-l	Levie Cohen	1864 - 1933
collem	Abraham Eliazer van Collem	1858 - 1933
coltof	Bernhard Coltof	1889 - 1940
coltof-s	Samuel Wolf Coltof	1854 - 1932
conjong	Johannis Adrianus Conjong	1895 - 1965
constandse	Anton Levien Constandse	1899 - 1985
corduwener	Gerardus Antonius Corduwener	1882 - 1940
cornelissen	Christianus Gerardus Cornelissen	1864 - 1942
coronel	Samuel Senior Coronel	1827 - 1892
cramer	Charles Guillaume Cramer	1879 - 1976
cremer	Jacobus Johannes Cremer	1827 - 1880
croll	Cornelis Croll	1857 - 1895
dam	Aron van Dam	1881 - 1942
danz	Peter Danz	1876 - 1969
det	Eldert Johannes van Det	1863 - 1948
diemer	Hendrik Diemer	1879 - 1966
dieters	Jan Dieters	1901 - 1943
dijkman	Albert Pieter Dijkman	1894 - 1953
dijkstra	Johannes Dijkstra	1854 - 1917
dok	Jan van Dok	1865 - 1956
dolleman	Willem Frederik Dolleman	1894 - 1942
douwes-dekker	Eduard Douwes Dekker	1820 - 1887
drees	Willem Drees	1886 - 1988
drion	Franciscus Johannes Wilhelmus Drion	1874 - 1948
drop	Willem Drop	1880 - 1939
duijs	Jan Eliza Wilhelm Duijs	1877 - 1941
eck	Dirk Antonie van Eck	1867 - 1948
eeden	Frederik Willem van Eeden	1860 - 1932
effendi	Roestam Effendi	1903 - 1979
eichelsheim	Henri Johannes Jacobus Eichelsheim	1865 - 1933
emmenes	Adrianus van Emmenes	1857 - 1906
engelbert	Adrien Jean Eliza Engelbert van Bevervoorde tot Oldemeule	1819 - 1851
engels	Arnoldus Hendrikus Johannes Engels	1869 - 1940
engels-j	Jacobus Alphonsus Engels	1896 - 1982
erkel	Gerrit van Erkel	1860 - 1937
faber	Jan Lambertus Faber	1875 - 1958
feringa	Frederik Feringa	1840 - 1905
fimmen	Eduard Carl Fimmen	1881 - 1942
fleischacker	Eliza Carolina Ferdinanda Fleischacker	1822 - 1904
fles	Levie Fles	1871 - 1940
fortuijn	Jan Antoon Fortuijn	1855 - 1940
frowein	Pieter Coenraad Frederik Frowein	1854 - 1917
funkekupper	Albert Johann Funke Küpper	1894 - 1934
fuykschot	Frans Pieter Fuykschot	1896 - 1961

gebing	Friedrich Wilhelm Gebing	1848 - 1923
geesink	Coenraad Albertus Jacobus Geesink	1828 - 1883
gelderen	Jacob van Gelderen	1891 - 1940
gerhard	Adrien Henri Gerhard	1858 - 1948
gerhard-h	Hendrik Gerhard	1829 - 1886
gerritsen	Carel Victor Gerritsen	1850 - 1905
giesen	Josephus Maria Petrus Jacobus Giesen	1881 - 1932
giezen	Antje Giezen	1858 - 1898
giezen-w	Wilhelmus Giezen	1860 - 1944
goedhart	Frans Johannes Goedhart	1904 - 1990
goes	Franc van der Goes	1859 - 1939
gool	Andries Johannes Jacobus van Gool	1882 - 1919
goot	Willemien Hendrika van der Goot	1897 - 1989
gorter	Herman Gorter	1864 - 1927
gotze	Johannes George Götze	1861 - 1940
goudsmit	Isaäc Goudsmit	1881 - 1966
goulooze	Daniël Goulooze	1901 - 1965
graaff	Willem Cornelis de Graaff	1847 - 1902
groeneweg	Susanna Groeneweg	1875 - 1940
groeningen	August Pieter van Groeningen	1866 - 1894
gronvald	August Ferdinand Gronvald	1879 - 1967
groot	Saul de Groot	1899 - 1986
groustra	Harmannus Groustra	1861 - 1944
gruyter	Jan de Gruyter	1859 - 1932
guit	Leonardus Franciscus Guit	1884 - 1937
gunst	Frans Christiaan Günst	1823 - 1885
haan	Jacob Israël de Haan	1881 - 1924
haas	Gerardus Horreüs de Haas	1879 - 1943
haas-j	Johan de Haas	1897 - 1945
haazevoet	Petrus Johannes Josephus Haazevoet	1876 - 1954
hagoort	Roelf Hagoort	1899 - 1965
hahn	Albert Pieter Hahn	1877 - 1918
haighton	Elise Adelaïde Haighton	1841 - 1911
harms	Antonie Harms	1867 - 1937
hartman	Evert Hendrik Hartman	1811 - 1873
hartog	Henri Benjamin Hartog	1869 - 1904
hartogh-heijs	Hermanus Hartogh Heijs	1841 - 1891
haver	Theodore Petronella Bernardine Haver	1856 - 1912
havers	Willem Havers	1865 - 1946
heeg	Tonnis van der Heeg	1886 - 1958
hegeraat	Herman Jacobus Hendrikus Conraad Hegeraat	1873 - 1919
heide	Albertinus van der Heide	1872 - 1953
heijenbrock	Johann Coenraad Hermann Heijenbrock	1871 - 1948
heijermans	Ida Sarah Heijermans	1861 - 1943
heijermans-c	Catharine Mariam Heijermans	1859 - 1937
heijermans-h	Herman Heijermans	1864 - 1924

heijermans-l	Louis Heijermans	1873 - 1938
heijkoop	Arie Wouter Heijkoop	1883 - 1929
heinen	Marie Heinen	1881 - 1948
heldt	Bernardus Hermanus Heldt	1841 - 1914
hell	Johannes Gerardus Diederik van Hell	1889 - 1952
helsdingen	Willem Pieter Gerardus Helsdingen	1850 - 1921
hermans-h	Hendrik Gerard Maria Hermans	1874 - 1949
hermans-l	Louis Maximiliaan Hermans	1861 - 1943
hiemstra	Pieter Feddes Hiemstra	1878 - 1953
hilgenga	Jan Hilgenga	1883 - 1968
hinte	Nico van Hinte	1869 - 1932
hobbel	Jan Hobbel	1857 - 1931
hoeven	Willem van der Hoeven	1879 - 1956
hof	Jan Hof	1870 - 1953
hogendorp	Anna van Hogendorp	1841 - 1915
hogerhuis	Keimpe Hogerhuis	1859 - 1919
	Marten Hogerhuis	1869 - 1936
	Wybren Hogerhuis	1863 - 1948
hoitsema	Maria Wilhelmina Hendrika Hoitsema	1847 - 1934
hommes	Jan Poppes Hommes	1843 - 1916
hondt	Luberta de Hondt	1885 - 1959
hoogcarspel	Jan Hoogcarspel	1888 - 1975
hoogland	Pieter Hoogland	1877 - 1958
hout	Isaac Salomon van der Hout	1843 - 1918
houten-h	Hendrik van Houten	1892 - 1952
houten-s	Samuel van Houten	1837 - 1930
houven	Gijsbert van der Houven	1883 - 1963
hoving	Jan Hoving	1877 - 1939
hovy	Willem Hovy	1840 - 1915
hudig	Dirk Hudig	1872 - 1934
hugenholtz	Frederik Willem Nicolaas Hugenholtz	1868 - 1924
huig	Rika Huig	1906 - 1967
huisman	Hendrik Hendicus Huisman	1821 - 1873
huizinga	Johannes Huizinga	1867 - 1946
huygens	Cornélie Lydie Huygens	1848 - 1902
ijzerman	Arie Willem IJzerman	1879 - 1956
ivens	George Henri Anton Ivens	1898 - 1989
jacobs	Aletta Henriëtte Jacobs	1854 - 1929
jansonius	Jan Gerbrandus Jansonius	1870 - 1957
janssen-g	Gerardus Lambertus Janssen	1859 - 1932
janssen-p	Peter Arnoldus Janssen	1835 - 1916
jentink	Geertruida Christina Jentink	1852 - 1918
jong-a	Aaltje de Jong	1885 - 1943
jong-a-a	Albert Andries de Jong	1891 - 1970
jong-a-m	Adrianus Michiel de Jong	1888 - 1943
jong-a-r	Année Rinzes de Jong	1883 - 1970

jonge	Willem Caspar de Jonge	1866 - 1925
junghuhn	Franz Wilhelm Junghuhn	1809 - 1864
jungius	Hendrika Maria Aleida Jungius	1864 - 1908
kadt	Jacques de Kadt	1897 - 1988
kalma	Jacobus Kalma	1870 - 1943
kaspers	Hendrik Ebo Kaspers	1869 - 1953
kater	Klaas Kater	1833 - 1916
katz	Cornelia Frida Katz	1885 - 1963
katz-s	Samuël Katz	1845 - 1890
kaulbach	Anna Maria Kaulbach	1869 - 1960
keesing	Isidore Keesing	1876 - 1943
kemper	Jeronimo Kemper	1808 - 1876
kenther	Harm Jan Kenther	1870 - 1944
keppler	Arie Keppler	1876 - 1941
ketner	Cornelis Hendrik Ketner	1876 - 1959
kieft	Johan van de Kieft	1884 - 1970
kies	Paul Charles Joseph Kiès	1895 - 1968
kitsz	Cornelis Kitsz	1873 - 1955
klaren	Uilke Jans Klaren	1852 - 1947
kleefstra-a	Anna Catharina Kleefstra	1884 - 1977
kleefstra-g	Geertje Kleefstra	1874 - ?
kleerekoper	Asser Benjamin Kleerekoper	1880 - 1943
knappert	Emilie Charlotte Knappert	1860 - 1952
knuttel	Johannis Adrianus Nelinus Knuttel	1878 - 1965
koch	Daniël Marcellus George Koch	1881 - 1960
koe	Anne de Koe	1866 - 1941
koejemans	Anthoon Johan Koejemans	1903 - 1982
kol	Hendrikus Hubertus van Kol	1852 - 1925
kolthek	Harm Kolthek Jr.	1872 - 1946
kom	Cornelis Gerhard Anton de Kom	1898 - 1945
koo	Johannes de Koo	1841 - 1909
kooijman	Pieter Adrianus Kooijman	1891 - 1975
kooistra	Ietje Kooistra	1861 - 1923
korteweg	Bastiaan Pieter Korteweg	1849 - 1879
kramers	Martina Gezina Kramers	1863 - 1934
krelage	Jan Antoon Krelage	1885 - 1957
krijthe	Hendricus Cornelius Josephus Krijthe	1825 - 1902
krop	Hildebrand Lucien Krop	1884 - 1970
kruithof	Klaas Kruithof	1875 - 1956
kruseman	Wilhelmina Jacoba Pauline Rudolphine Krüseman	1839 - 1922
kruyt	John William Kruyt	1877 - 1943
kuijkhof	Johannes Gerardus van Kuijkhof	1864 - 1921
kuijper	Abraham Kuijper	1837 - 1920
kuiken	Johannes Kuiken Jzn.	1860 - 1936
kuiper	Cornelis Johannes Kuiper	1875 - 1951
kuiper-h	Hendrikus Jacobus Kuiper	1897 - 1985

kulk	Willem van der Kulk	1891 - 1971
kupers	Evert Kupers	1885 - 1965
kuyper	Rudolph Karel Herman Kuyper	1874 - 1934
laan-j	Jan ter Laan	1872 - 1956
laan-k	Kornelis ter Laan	1871 - 1963
laar	Hendrik van Laar	1898 - 1955
lacet	Carolina Lacet	1856 - 1920
ladenius	Gaatske Adriana Ladenius	1883 - 1953
lalleman	Gerrit Bernardus Lalleman	1820 - 1901
lammers	Casper Antonius Franciscus Lammers	1885 - 1966
lange-d	Daniël de Lange	1878 - 1948
lange-j	Jakob Leendert Adriaan de Lange	1869 - 1940
langeraad	Krijn Adriaan van Langeraad	1865 - 1943
lansink	Bernardus Lansink jr.	1884 - 1945
last	Josephus Carel Franciscus Last	1898 - 1972
lebeau	Joris Johannes Christiaan Lebeau	1878 - 1945
leeuw	Alexander Salomon de Leeuw	1899 - 1942
leeuwen	Frederik van Leeuwen	1905 - 1968
leguit	Paulus Leguit	1856 - 1937
lende	Coendert van der Lende	1887 - 1964
lensing	Wilhelmina Elisabeth Lensing	1847 - 1925
leusink	Andries Leusink	1884 - 1973
levenbach	Marius Gustaaf Levenbach	1896 - 1981
levita	Adolf Samson de Levita	1868 - 1934
liebers	Bernhard Bruno Ferdinand Liebers	1836 - 1900
lieme	Nehemia de Lieme	1882 - 1940
ligt	Bartholomeus de Ligt	1883 - 1938
ligthart	Gerard Jan Ligthart	1859 - 1916
lindeijer	George Frederik Lindeijer	1878 - 1943
lodewijk	Johan Jacob Lodewijk	1871 - 1942
loerakker	Anthonius Josephus Loerakker	1873 - 1950
lokerse	Neeltje Lokerse	1868 - 1954
loopuit	Joseph Loopuit	1864 - 1923
lubbe	Marinus van der Lubbe	1909 - 1934
luitink	Petrus Jacobus Luitink	1841 - 1871
luitjes	Tjerk Luitjes	1867 - 1946
luremans	Cornelis Johannes Luremans	1866 - 1936
maenen	Josephus Hubertus Maenen	1888 - 1972
mannoury	Gerrit Mannoury	1867 - 1956
mansholt-d	Derk Roelfs Mansholt	1842 - 1921
mansholt-l	Lambertus Helprig Mansholt	1875 - 1945
manus	Rosette Susanna Manus	1881 - 1943
marken	Jacob Cornelis van Marken	1845 - 1906
masereeuw	Harmanus Masereeuw	1861 - 1941
matthijsen	Jan Willem Matthijsen	1879 - 1949
mazirel	Laura Carola Mazirel	1907 - 1974

meeter	Eillert Meeter	1818 - 1862
meij	Henriette Rosina Dorothea van der Meij	1850 - 1945
meijer-r	Rudolf Carel Meijer	1826 - 1904
meijer-w	Willem Hendrik Meijer	1877 - 1951
melchers	Gerrit Willem Melchers	1869 - 1952
meliefste	Pieter Meliefste	1901 - 1985
mendels	Maurits Mendels	1868 - 1944
menist	Abraham Menist	1896 - 1942
mensing	Maria Anna Catharina Mensing	1854 - 1933
mercier	Helena Mercier	1839 - 1910
methofer	Johannes Cornelis Hendrik Philippus Methöfer	1863 - 1933
meurs	Frans van Meurs	1889 - 1973
michels	Andreas Wilhelmus Michels	1880 - 1961
michon	Christiaan Peter Michon	1843 - 1899
middelhuis	Johannes Antonius Middelhuis	1902 - 1978
mierop	Dirk Lodewijk Willem van Mierop	1876 - 1930
miranda	Salomon Rodrigues de Miranda	1875 - 1942
mok	Salomon Mok	1902 - 1948
mol	Hendrik Mol	1880 - 1959
moltmaker	Petrus Moltmaker	1882 - 1941
mooij	Johannes Mooij	1888 - 1977
muller	Hendrik Clemens Muller	1855 - 1927
munster	Gijsbert Jasper van Munster	1883 - 1945
muysken	Geertruida Agneta Muysken	1855 - 1920
naber	Johanna Wilhelmina Antoinette Naber	1859 - 1941
nabrink	Gerard Nabrink	1903 - 1993
nanninga	Gerrit Nanninga	1858 - 1933
nathans	Nathan Nathans	1883 - 1937
nauta	Jakob Nauta	1886 - 1949
nawijn	Tjepke Nawijn	1862 - 1939
nederhorst	Gerard Marinus Nederhorst	1907 - 1979
nieuwenhuis-a	Adrianus Jacobus Nieuwenhuis	1820 - 1880
nieuwenhuis-f	Ferdinand Nieuwenhuis	1846 - 1919
nieuwenhuis-j	Johannes Adam Nieuwenhuis	1856 - 1939
nivard	Franciscus Lambertus Deodatus Nivard	1879 - 1945
nobel	Otto Willem de Nobel	1867 - 1950
nolens	Willem Hubert Nolens	1860 - 1931
nolting	Pieter Nolting	1852 - 1923
noordhoff	Franciscus Siebren Noordhoff	1882 - 1970
odinot	Maria Elisabeth Odinot	1908 - 1998
oldenbroek	Jacobus Hendrik Oldenbroek	1897 - 1970
ommeren	Bartholomeus van Ommeren	1859 - 1907
onsman	Inte Onsman	1872 - 1929
oogen	Johannes Jacobus van Oogen	1867 - 1926
ortt	jonkheer Felix Louis Ortt	1866 - 1959
ossedorp	Frans Lodewijk Ossedorp	1863 - 1941

oudegeest	Jan Oudegeest	1870 - 1950
oudens	Franciscus Petrus Oudens	1855 - 1920
palar	Lambertus Nicodemus Palar	1900 - 1981
pannekoek	Antonie Pannekoek	1873 - 1960
paris	Johannes Hubertus Paris	1873 - 1939
pas	Wilhelmus Martinus van de Pas	1901 - 1960
passtoors	Willem Caspar Joseph Passtoors	1856 - 1916
pekelharing	Baltus Hendrik Pekelharing	1841 - 1922
penning	Paulus Jacobus Penning	1843 - 1904
perk	Christina Elizabeth Perk	1833 - 1906
pieck	Henri Christiaan Pieck	1895 - 1972
pieters-c	Caspar Hubertus Pieters	1892 - 1964
pieters-e	Elise Hubertine Pieters	1897 - 1976
pieters-g	Gerardus Hubertus Pieters	1860 - 1947
poels	Henricus Poels	1868 - 1948
poesiat	Bart Poesiat	1831 - 1898
polak-a	Anna Sophia Polak	1874 - 1943
polak-e	Eliazer Polak	1880 - 1962
polak-h	Henri Polak	1868 - 1943
polak-kerdijk	Arnoldus Polak Kerdijk	1846 - 1905
porreij	Jacoba Maria Petronella Porreij	1851 - 1930
posthumus	Nicolaas Wilhelmus Posthumus	1880 - 1960
postma	Thomas Postma	1824 - 1906
postma-j	Jan Postma	1895 - 1944
potharst	Johannes Theodorus Potharst	1841 - ?
pothuis	Samuel Pothuis	1873 - 1937
poutsma	Hessel Poutsma	1866 - 1933
prijes	Sientje Prijes	1876 - 1933
quack	Hendrick Peter Godfried Quack	1834 - 1917
querido	Israël Querido	1872 - 1932
raay	Cornelis Johannes van Raay	1859 - 1893
rauwerda	Anne Rauwerda	1859 - 1945
ravesteyn	Willem van Ravesteyn	1876 - 1970
redele	Eduard Charles Philippus Redelé	1866 - 1913
reens	Abraham Mozes Reens	1870 - 1930
rees	Jacob van Rees	1854 - 1928
reinalda	Marius Antoon Reinalda	1888 - 1965
reve	Gerardus Johannes Marinus van het Reve	1892 - 1975
reyndorp	Bernard Daniël Guillaume Charles Reyndorp	1870 - 1950
ribbius-peletier	Anna Elisabeth Ribbius Peletier	1891 - 1989
rijen	Antonius Theodorus van Rijen	1878 - 1946
rijk	Ester van Rijk	1853 - 1937
rijnders	Gerhard Rijnders	1876 - 1950
rijzewijk	Joannes van Rijzewijk	1880 - 1939
ris	Klaas Ris	1821 - 1902
rodrigues	Flora Rodrigues	1893 - 1996

roestenberg	Cornelis Roestenberg	1877 - 1955
roland-holst	Richard Nicolaüs Roland Holst	1868 - 1938
romein	Jan Marius Romein	1893 - 1962
rommerts	Hoeke Rommerts	1834 - 1890
rommerts-o	Obbe Rommerts	1835 - 1902
roode	Justus Johannes de Roode	1866 - 1945
roorda	Gerrit Roorda	1890 - 1977
roorda-van-eysinga	Sikko Ernest Willem Roorda van Eysinga	1825 - 1887
rosa	Andries de Rosa	1869 - 1943
rot-a	Adriaan Rot	1861 - 1927
rot-j	Jacob Rot	1854 - 1938
rot-jan	Jan Rot	1892 - 1982
rot-t	Thomas de Rot	1840 - 1915
royaards	Godfried Johan Royaards	1842 - 1904
rugge	Eltjo Rugge	1872 - 1950
ruijs-van-beerenbroek	jonkheer Charles Joseph Maria Ruijs van Beerenbroek	1873 - 1936
ruijter	Pieter Cornelis de Ruijter	1855 - 1889
ruppert	Marinus Ruppert	1911 - 1992
ruppert-j	Johan Stephaan Ruppert jr.	1877 - 1934
rutgers-j	Johannes Rutgers	1850 - 1924
rutgers-s	Sebald Justinus Rutgers	1879 - 1961
sajet	Benedictus Herschel Sajet	1887 - 1986
samson	Izak Samson	1872 - 1928
sannes	Goswijn Willem Sannes	1875 - 1930
savornin-lohman	jonkvrouw Catharina Anna Maria de Savornin Lohman	1868 - 1930
schaepman	Herman Johan Aloysius Maria Schaepman	1844 - 1903
schaft	Jannetje Johanna Schaft	1920 - 1945
schaik	Johannes Gerardus van Schaik	1871 - 1956
schalk	Henriette Goverdine Anna van der Schalk	1869 - 1952
schaper	Johan Hendrik Andries Schaper	1868 - 1934
scheepers	Johannes Theodorus Scheepers	1838 - 1882
scheps	Johannes Hermanus Scheps	1900 - 1993
schermerhorn	Dirk Schermerhorn	1900 - 1937
schermerhorn-n	Nicolaas Jacob Cornelis Schermerhorn	1866 - 1956
schermerhorn-w	Willem Schermerhorn	1894 - 1977
schilperoort	Anna Barbera Schilperoort	1778 - 1853
schmidt	Frerich Ulfert Schmidt	1870 - 1939
schmidt-p	Petrus Johannes Schmidt	1896 - 1952
schotting	Louis Schotting	1870 - 1957
schroder	Petrus Hendrikus Antonius Schröder	1836 - 1914
schurer	Fedde Schurer	1898 - 1968
schutte	Johannes Antonius Schutte	1882 - 1945
seegers	Leendert Seegers	1891 - 1970
seggelen	Josephus Adriaan van Seggelen	1883 - 1952
serrarens	Petrus Josephus Servatius Serrarens	1888 - 1963

sikkel	Johannes Cornelis Sikkel	1855 - 1920
simonis	Pieter Simonis	1898 - 1983
sinoo	Eimert Sinoo	1880 - 1971
slotemaker	Jan Rudolph Slotemaker	1869 - 1941
sluijser	Meijer Sluijser	1901 - 1973
smeenk	Christaan Smeenk	1880 - 1964
smit	Jan Martinus Smit	1852 - 1942
smit-jr	Gerrit Johan Adam Smit Jr.	1879 - 1934
smit-m	Mattijs Willem Smit	1866 - 1916
smit-w	Wilhelmina Carolina Benjamina Smit	1872 - 1951
smits	Hendrik Smits	1865 - 1937
sneevliet	Hendricus Josephus Franciscus Marie Sneevliet	1883 - 1942
soep	Abraham Soep	1874 - 1958
spiekman	Hendrik Spiekman	1874 - 1917
staal	Karl Rigard van Staal	1889 - 1961
staalman	Andries Popke Staalman	1858 - 1938
stad	Pieter van der Stad Jbz.	1850 - 1905
stam	Jan Cornelis Stam	1884 - 1943
stap	Jan Abrahams Stap	1859 - 1908
stapelkamp	Antoon Stapelkamp	1886 - 1960
staveren	David van Staveren	1881 - 1966
steinmetz	Willem Steinmetz	1891 - 1968
stellingwerf	Oebele Stellingwerf	1847 - 1897
stenhuis	Roelof Stenhuis	1885 - 1963
sternheim	Andries Sternheim	1890 - 1944
sterringa-g	Geert Sterringa	1876 - 1944
sterringa-j	Jan Sterringa	1870 - 1951
stienstra	Tjeerd Jans Stienstra	1859 - 1935
stins	Hermanus Josephus Stins	1877 - 1932
stoffel	Jan Stoffel	1851 - 1921
stokman	Jacobus Gerardus Stokman	1903 - 1970
stokvis	Jozef Emanuel Stokvis	1875 - 1951
stoop	Theodoor Stoop	1861 - 1933
stork	Dirk Willem Stork	1855 - 1928
struik	Thomas Antonie Struik	1897 - 1945
suchtelen	Nicolaas Johannes van Suchtelen	1878 - 1949
sunito	Raden Mas Djojowirono Sunito	1912 - 1979
suurhoff	Jacobus Gerardus Suurhoff	1905 - 1967
tak	Pieter Lodewijk Tak	1848 - 1907
talma	Aritius Sybrandus Talma	1864 - 1919
tas	Salomon Tas	1905 - 1976
temme	Hendrik Lodewijk Temme	1857 - 1917
tempel	Jan van den Tempel	1877 - 1955
tendeloo	Nancy Sophia Cornélie Tendeloo	1897 - 1956
tenthoff	Johannes Theodorus Tenthoff	1847 - 1916
terwey	Jan Pieter Terwey	1883 - 1965

thijsse	Jacobus Pieter Thijsse	1865 - 1945
thijssen	Theodorus Johannes Thijssen	1879 - 1943
tiggers	Petrus Johannes Tiggers	1891 - 1968
tilanus	Liede Tilanus	1871 - 1953
tinbergen	Jan Tinbergen	1903 - 1994
treub	Marie Willem Frederik Treub	1858 - 1931
troelstra-d	Dirk Jelles Troelstra	1870 - 1902
troelstra-h	Hendrika Troelstra	1867 - 1944
troelstra-p	Pieter Jelles Troelstra	1860 - 1930
tusveld	Johan Tusveld	1865 - 1902
tuuk	Titia Klasina Elisabeth van der Tuuk	1854 - 1939
urban	Francois Ernst Lodewijk Urban	1843 - 1904
vader	Gerard Albert Vader	1865 - 1940
valkhoff	Johan Valkhoff	1897 - 1975
veen-s	Sjoerd Si(e)brens van Veen	1828 - 1897
veen-y	Ybele Geert van der Veen Hzn.	1884 - 1940
veer	Johannes Koenraad van der Veer	1869 - 1928
vegt	Helmig Jan van der Vegt	1864 - 1944
velde	Cornelis Antonius van der Velde	1860 - 1932
veldman	Joeke Veldman	1889 - 1971
ven	Ernestus Philippus Hubertus van der Ven	1846 - 1913
veraart	Joannes Antonius Veraart	1886 - 1955
verdorst	Pieter Marinus Verdorst	1858 - 1944
verschoor	Anna Helena Margaretha Verschoor	1895 - 1978
versluis	Willem Gerardus Versluis	1892 - 1972
verstegen	Alexander Gustaaf Adolph Verstegen	1870 - 1936
visscher	Johannes Visscher	1872 - 1945
visser	Louis Leonardus Hendrikus de Visser	1878 - 1945
vleeschdrager	Lion Vleeschdrager	1898 - 1958
vleggeert	Johannis Cornelis Vleggeert	1899 - 1970
vletter	Jacob de Vletter	1818 - 1872
vliegen	Wilhelmus Hubertus Vliegen	1862 - 1947
vlies	Anke van der Vlies	1873 - 1939
vliet	Pieter van Vliet jr.	1858 - 1941
voo	Goose Wijnant van der Voo	1806 - 1902
voogd	Petrus Voogd	1873 - 1939
vorrink	Jacobus Jan Vorrink	1891 - 1955
vorst	Hendrikus Johannes van Vorst	1867 - 1927
vos-g	Grietje Vos	1891 - 1985
vos-h	Hein Vos	1903 - 1972
vos-m	Maria Wilhelmina Vos	1897 - 1994
vos-p	Petrus Josephus Wilhelmus de Vos	1805 - 1866
vos-r	Roosje Vos	1860 - 1932
voskuil	Klaas Voskuil	1895 - 1975
vries	Nathan Albert de Vries	1878 - 1924
vrijburg	Willem Vrijburg	1850 - 1925

wacht	Jan Wacht	1885 - 1967
waerden	Theodorus van der Waerden	1876 - 1940
wal	Feike Obbes van der Wal	1873 - 1937
werkhoven	Cornelis Werkhoven	1887 - 1928
werthweijn	Petrus Werthweijn	1822 - 1900
wertwijn	Jan Willem Wertwijn	1839 - 1899
wessels	Adriaan Cornelis Wessels	1867 - 1960
wiardi-beckman	Herman Bernard Wiardi Beckman	1904 - 1945
wibaut	Florentinus Marinus Wibaut	1859 - 1936
wichmann	Clara Gertrud Wichmann	1885 - 1922
wiedijk	Pieter Wiedijk	1867 - 1938
wiessing	Henri Pierre Leonard Wiessing	1878 - 1961
wijhe	Marie Cornelis van Wijhe	1881 - 1953
wijk	Jan Hendrik van Wijk	1907 - 1981
wijk-j	Johannes van der Wijk	1848 - 1913
wijnkoop	David Jozef Wijnkoop	1876 - 1941
willekes-macdonald	Ina Elisa Willekes MacDonald	1886 - 1979
willemse	Wijbrecht Willemse	1897 - 1984
wink	Pieter Marinus Wink	1875 - 1924
wit	Anna Augusta Henriette de Wit	1864 - 1939
witt-hamer	Michiel Jacobus de Witt Hamer	1843 - 1925
wittert	Everardus Bonifacius François Frederik baron Wittert	1875 - 1959
wolbers	Julien Wolbers	1819 - 1889
wolff	Salomon de Wolff	1878 - 1960
wollring	Hendrik Herman Wollring	1869 - 1939
wormser	Johan Adam Wormser jr.	1845 - 1916
woudenberg	Cornelis Woudenberg	1883 - 1954
woudenberg-h	Hendrik Jan Woudenberg	1891 - 1967
zadelhoff	Johannes Hendrikus Franciscus van Zadelhoff	1868 - 1946
zandstra	Lourens Zandstra	1848 - 1923
zee	Daniël van der Zee	1880 - 1969
zeeuw	Arie Bastiaan de Zeeuw	1881 - 1967
zinderen-bakker	Rindert van Zinderen Bakker	1845 - 1927
zomeren	Johannes Nicolaas van Zomeren	1848 - 1902
zutphen	Johannes Andries van Zutphen	1863 - 1958
zwaag	Geert Lourens van der Zwaag	1858 - 1923
zwertbroek	Gerrit Jan Zwertbroek	1893 - 1977

APPENDIX B

Term lists for Named Entity Disambiguation.

ORGANIZATION – ADJECTIVES:

Chr., Christelijk, Christelijke
Coöp., Coöperatieve, Coop., Cooperatieve
Alg., Algemene, Algemeene, Algemeen
A'damsche, Amsterdamse, Amsterdamsche
Int., Internationale, Internationaal
Kath., Katholiek, Katholieke
Kon., Koninklijk, Koninklijke
Nat., Nationale, Nationaal
Ned., Nederlandsche, Nederlandse
R'damsche, Rotterdamse, Rotterdamsche
R.K., Roomsch Katholieke, Rooms Katholieke, Roomsch Katholiek, Rooms Katholiek, Roomsch, Rooms, Katholieke, Katholiek
Soc., Sociaal, Sociale, Socialistisch, Socialistische
Dem., Demokraat, Demokraten, Demokratisch, Demokratische, Democraat, Democraten, Democratisch, Democratische
Centraal, Centrale

ORGANIZATION – COLLECTIVES:

Vennootschap, vennootschap, Vennootschappen, vennootschappen
Maatschap, maatschap, Maatschappij, maatschappij, Mij, Mij., Maatschappen, maatschappen, Maatschappijen, maatschappijen
Vereniging, vereniging, Vereeniging, vereeniging, Ver., Ver., Verenigingen, verenigingen, Vereenigingen, vereenigingen
Coöperatie, coöperatie, Cooperatie, cooperatie, Coöp., coöp., Coop., coop., Coöperaties, coöperaties
Stichting, stichting, Stichtingen, stichtingen
Commissie, commissie, Comm., comm., Commissies, commissies
Comité, comité, Comite, comite, Comités, comités, Comites, comites
Genootschap, genootschap, Genootschappen, genootschappen
Bureau, bureau, Bureaus, bureaus
Partij, partij, Partijen, partijen
Club, club, Clubs, clubs
Bestuur, bestuur, Besturen, besturen
Universiteit, universiteit, Universiteiten, universiteiten
School, school, Scholen, scholen
Academie, academie, Akademie, akademie, Academies, academies, Akademies, akademies
Beweging, beweging, Bewegingen, bewegingen
Verbond, verbond, Verbonden, verbonden

Centrale, centrale, Centrales, centrales
Fonds, fonds, Fondsen, fondsen
Gemeenschap, gemeenschap, Gemeenschappen, gemeenschappen
Federatie, federatie, Federaties, federaties
Onderneming, onderneming, Ondernemingen, ondernemingen
Secretariaat, secretariaat, Secretariaten, secretariaten
Instituut, instituut, Inst., Instituten, instituten
Dispuut, dispuut, Disputen, disputen
Groep, groep, Groepen, groepen
Unie, unie, Unies, unies
Kamer, kamer, Kamers, kamers
Bond, bond, Bonden, bonden
Raad, raad, Raden, raden
Bedrijf, bedrijf, Bedrijven, bedrijven
Kolonie, kolonie, Kolonies, kolonies, Koloniën, koloniën, Kolonien, kolonien

ORGANIZATION – INFIXES:

De, de
Den, den
Der, der
Te, te
Ter, ter
Van, van
Voor, voor
Om, om
En, en
In, in
Door, door
Naar, naar
Tot, tot
Het, het, 't, t, 'T, T
Bij, bij
Met, met
Een, een

Person – surname prefixes:

A, a
Aan, aan
Aan de, aan de
Aan den, aan den
Aan der, aan der
Aan het, aan het, Aan t, aan t, Aan ‘t, aan ‘t
Af, af
Al, al
Am, am
Am de, am de
Auf, auf
Auf dem, auf dem
Auf den, auf den
Auf der, auf der
Auf ter, auf ter
Aus, aus
Aus ‘m, aus ‘m, Aus m, aus m
Aus dem, aus dem
Aus den, aus den
Aus der, aus der
Bij, bij
Bij de, bij de
Bij den, bij den
Bij het, bij het, Bij t, bij t, Bij ‘t, bij ‘t
Bin, bin
Boven d, boven d, Boven d’, boven d’
D, d, D’, d’
Da, da
Dal, dal
Dal’, dal’, Dalla, dalla
Das, das
De, de
De die, de die
De die le, de die le
De l, de l, De l’, de l’
De la, de la
De las, de las
De le, de le
De van der, de van der
Deca, deca
Degli, degli
Dei, dei
Del, del
Della, della
Den, den
Der, der
Des, des
Di, di
Die le, die le
Do, do
Don, don
Dos, dos
Du, du
El, el
Het, het, T, t, ‘T, ‘t
I, i
Im, im
In, in
In de, in de
In den, in den
In der, in der
In het, in het, In t, in t, In ‘t, in ‘t
L, l, L’, l’
La, la
Las, las
Le, le
Les, les
Lo, lo
Los, los
Of, of
Onder, onder
Onder de, onder de
Onder den, onder den
Onder ‘t, onder ‘t, Onder het, onder het, Onder t, onder t
Op, op
Op de, op de
Op den, op den
Op der, op der
Op gen, op gen
Op het, op het, Op t, op t, Op ‘t, op ‘t
Op ten, op ten
Over, over
Over ‘t, over ‘t, Over het, over het
Over de, over de
Over den, over den
Over t, over t
S, s, S’, s’, ‘S, ‘s
Te, te
Ten, ten
Ter, ter
Tho, tho
Thoe, thoe
Thor, thor
To, to
Toe, toe
Tot, tot
Uijt, uijt
Uijt ‘t, uijt ‘t
Uijt de, uijt de
Uijt den, uijt den
Uijt te de, uijt te de
Uijt ten, uijt ten
Uit, uit
Uit de, uit de
Uit den, uit den
Uit ‘t, uit ‘t, Uit het, uit het, Uit t, uit t
Uit te de, uit te de

Uit ten, uit ten
Unter, unter
Van, van
Van de, van De, van de
Van de l, van de l, Van de l', van de l'
Van Den, Van den, van den
Van Der, Van der, van der
Van gen, van gen
Van het, van het, Van 't, van 't, Van t, van t
Van ter, van ter
Van la, van la
Van van de, van van de
Ver, ver
Vom, vom
Von, von
Von 't, von 't, Von t, von t
Von dem, von dem
Von den, von den
Von der, von der
Voor, voor
Voor 't, voor 't
Voor de, voor de
Voor den, voor den
Voor in 't, voor in 't, Voor in t, voor in t
Vor, vor
Vor der
vor der
Zu, zu
Zum, zum
Zur, zur

Person – titles:

aartsbisschop
abdis
abt
admiraal
advocaat
advocaat-generaal
ambassadeur
ambtenaar
apotheker
arts
attaché, attache
baron
barones
bisschop
broeder
burgemeester
commandeur
commissaris
consul
consul-generaal
curator
deken
directeur
directeur-generaal
doctor, dr.
doctorandus, drs.
dokter, dr.
frater
gemeenteraadslid
generaal
generaal-majoor
gezant
gouverneur
graaf
gravin
griffier
grootofficier
heer, hr.
hoofdinspecteur
hoofdofficier
hoogleraar
ingenieur, ir., ing.
inspecteur
inspecteur-generaal
jonkheer
jonkvrouw
kamerlid
kandidaat-notaris
kantonrechter
kanunnik
kapelaan
kapitein
kapitein-luitenant
kardinaal
kolonel
koning
koningin
kroonprins
kroonprinses
luitenant
luitenant-admiraal
luitenant-generaal
luitenant-kolonel
majesteit
majoor
meester, mr.
meneer
mevrouw
minister
monseigneur
notaris
officier
opperrabbijn
oud-minister
pastoor

pater
predikant
prins
prinses
procureur-generaal
professor
rabbijn
rechter
rechter-commissaris
rector
referendaris
registeraccountant, ra, ra., r.a.
religieuze
ridder
ritmeester
schoolopziener
secretaris-generaal
staatssecretaris
substituut-officier
tandarts
vice-admiraal
vice-consul
wethouder
zuster

Location – address:

straatweg
steenweg
dwarsweg
zijweg
heerweg
kiezelweg
weg
kiezel
heerbaan
voetbaan
baan
kassei, kalsijde
dwarsstraat
heerstraat
steenstraat
straatje
straat
paadje
voetpad
pad
steeg
wegel
ringlaan
laan
ring
singel
lei

vest
dreef
drift
boulevard
passage
promenade
gracht
graaf
rei
rui
dam
zeedijk
dijk
kade
kaai
wal
brug
heul
sas
sluis
til
poort
voord, voorde, voort
wad
gaard, gaarde
hof
marktplein
markt
plein
plaats
plantsoen
park
rotonde
square
aard
bies, biest, bist
brink
tuin
dries
eng
meent
berg, burg
voord, voorde, voort
borg, burg, burcht
erf
geleeg
heem, hiem
heerd, heert
hoeve, hoef
hof
hofstee, hofstad
kluis
munster
sluis
sas

werf
winning
akker
beemd, bemd, bamd
boomgaard, bogaard, bogerd
braak
broek
bunder, bonder, boender
dagmaat
dal, del
duin
geer
hei, heide
heuvel
hil, hul, hulle
hoek
hof
terp
tuin
veen, ven, vin
veld
wei, weide
wijngaard, wingerd
aard
bies, biest, bist
bocht
blok, blook
bleuk
beluik, bilk
brink
donk
dries
eng
gaard, gaarde
goor
ham, haam, hem
horst
kamp

Location – administrative:

land
rijk
staat
vrijstaat
republiek
keizerrijk
koninkrijk
hertogdom
graafschap
gouw
provincie
baronie
kasselrij
proosdij
stad
bisdom
dorp
gemeente
continent
wereld
wijk

Location – cultural:

museum
gallerie
amphitheater
theater
concert hal
cinema, bioscoop, bios
opera
symphonie
droom

Location – educational:

universiteit
hogeschool
middelbare school
kleuterschool
college
klaslokaal
gymnasium
school
bibliotheek

Location – professional:

fire station
politiebureau
benzinepomp
postkantoor
congrescentrum
beurs
opslagplaats
magazijn
warenhuis
broeikas
rechtbank
koffiehuis
supermarkt
hypermarkt

markt
world trade centre, world trade center, WTC
gemeentehuis
olieboorplatform
boorplatform
werkplaats
winkel
nachtclub
club
energiecentrale
centrale
bank
bar
pub
forum
hotel
motel
kantoor
eetcafe
restaurant
cafe
wolkenkrabber
silo
stal
bloemenzaak
bakkerij
graanschuur
graanzolder
consulaat
ambassade
parlement
brouwerij
fabriek
gieterij
mijn
raffinaderij

Location – religious:

kerk
basiliek
kathedraal
domkerk
dom
kapel
martyrium
moskee
imambargah
monikenklooster
nonnenklooster
klooster
abdij
mithraeum
pyramide
altaar
synagoge
tempel
pagoda
sticht

Location – residential:

bejaardentehuis
verzorgingstehuis
weeshuis
tehuis
appartement
asiel
condominium
duplex
huis
villa
bungalow
landhuis
riddergoed
boerderij
boerenwoning

Location – transport:

vliegveld
terminal
station
parkeerplaats
parkeergarage

Location – water:

oceaan
stuwmeer
stuwdam
dam
meer
zwembad
bad
haven
strand
atol
kanaal
fjord
loch
oase
polder

poel
rif
reservoir
rivier
kust
bron, born
beek
straat
stroom
ij
ee
moeras
mangrove
delta
baai
ie
diep
gracht
graaf
greppel
grub
heul, hool
kreek
laak
loop
plas
vaart
vijver, wijer, wouwer
waal, weel, wiel
water
zeeëngte
zee
zijl
zoe, zouw, zoei
sluis
ven, venne

schuur
schuurtje
hospitaal
ziekenhuis
kliniek
hut
stadion
arena
stal
triomfboog
boog
gym
kelder
loods
barak
keet
stulp
put
bunker
kasteel
citadel
toren
molen, meulen
gevangenis
paleis
brug

Location – other:

boothuis
badhuis
hooizolder
hooischuur
watermolen
wal
fortificatie
fort
carport
garage
hangar
silo
aqueduct
hal

SUMMARY

Social networks have a long history in the social sciences as a means to study interactions between members of different types of communities. In the absence of computers to record the data, the network models for these studies were all constructed manually. Since this is an expensive process in terms of the time and expertise needed, the models that resulted from these efforts are limited in size and complexity. Advances made since in computer science, coupled with the increasing availability of real-world data from sources such as Facebook and Twitter, has reignited interest in social network models, or *graphs*, from a computational perspective. Automated methods and powerful computers have made it feasible to conduct more complex experiments at larger scales very cost-effectively. Furthermore, developments in Natural Language Processing provide just the right tools to extract the building blocks of social networks, even from sources that were not originally created for this purpose, like free text documents.

The ability to automatically construct graphs from unstructured sources adds Social Network Analysis as a viable instrument to a host of new fields. One of those fields is Social History, where it can for instance be applied to study the propagation of an ideology, or the mechanics of support networks in historical contexts. Unfortunately, there still exists some apprehension among social historians, and humanities researchers in general, to trust the results of automated methods. They commonly fear that their own subjective interpretation will be compromised by the objective decisions, and inevitable mistakes, of a computer. Our goal in this thesis is to mitigate such trepidations by developing an automated method for social network extraction on an existing source from the Social History domain, and proving its validity by comparing the outcome against existing social network models and expert annotations on the original data. To this end, we formulate the following problem statement:

> *Can computational methods be used to successfully extract a detailed social network from historical, textual data, enriching the data in such a way that is of added value to social historical research?*

We identify two main tasks in our problem statement, namely the construction of the social network model from the input data, and the validation of the resulting graph. Social networks consist of *actors*, represented by *nodes*, and their *relationships*, represented by *links* or *edges*. Actors are usually people, or groups of people, such as organizations or locations (i.e. their population). Relationships may represent any type of interaction or state between any pair of actors, for instance

friendship, cooperation, hate, or mere awareness. In order to ensure that a graph constructed with our method is sufficiently dense, i.e. that it contains enough nodes and edges to perform meaningful calculations on it, we require the input to contain a high concentration of terms that refer to distinct (groups of) people, or *named entities*. Biographies generally fit the bill because of their condensed description of events. Our chosen dataset, therefore, is a biographical dictionary, which we introduce in Chapter 2 together with the field of Social History and its challenges. We conclude Chapter 2 with an overview of related research into network and relation extraction, both from a social scientific, and a computational point of view.

The network construction process can be logically divided into the extraction of nodes and the extraction of links. In Chapter 3, we apply Named Entity Recognition and Disambiguation to gather the nodes of our social network, determined by the people, organizations, and locations that are mentioned in the biographical dictionary. We are able to identify these entities with great accuracy. Next, we develop a new state-of-the-art method for the cross-document disambiguation of names in Dutch text, which performs comparable to similar systems developed for English. At the end of the chapter, we have a first look at a static graph constructed from sentence-based co-occurrences of person entities and find a definite equivalence to existing manual annotations on the biographies.

Part of our goal is to enable the study of evolutionary patterns within the graph. We therefore require the links to be bounded to a timeline. In Chapter 4, we develop a hybrid method for the recognition, identification, and normalization of temporal expressions in Dutch text. Recognition and identification are dealt with using a machine-learning component, while normalization is solved in a rule-based setup. This method is shown to perform significantly better than the current state of the art, which is a completely rule-based approach. On top of the improved performance, our method also adds analysis of linguistic events. Together with the temporal expressions, these events are used to further specify the links in our graph: the start and end dates of the temporal expression determines the time and duration of the relationship, while the events describe its context. Again, we conclude the chapter with a brief look at a static graph from sentence-based co-occurrences of person entities, this time restricted to sentences describing a specific timeframe. Since the resulting graph is rather sparse, we attempt to increase its connectivity by collapsing mutual connections to other people, organizations, or locations, but find that none of these entity types are suitable for this purpose in the current setup.

At the start of Chapter 5, we introduce some key concepts in Social Network Analysis and describe the characteristics that social network models generally adhere to according to the literature. Next, we test our own model to see if it complies with these characteristics. We analyze the statistics of graph at different temporal intervals (decade and year) and find that it is a small world network where growth is governed by a slightly distorted preferential attachment mechanism. We also check the correlations of node centrally rankings between

different sub graphs and conclude that importance is a momentary rather than permanent attribute. We conclude the chapter with a brief analysis of the linguistic events from Chapter 4 that are attached to the links in the graphs with respect to the types of nodes (i.e. person, organization, location) that the link connects to. Although the correlations are all insignificant, ordering them from high to low does reveal some patterns that could be semantically motivated at both ends of the scale.

We conclude the thesis in Chapter 6 with a discussion of our results and answers to our problem statement and research questions.

CURRICULUM VITAE

Matje van de Camp was born in Tilburg, the Netherlands, on the 16th of February, 1981. After completing her secondary education (VWO) at the Willem II College in Tilburg, she obtained her propedeutic diploma in Philosophy from the Vrije Universiteit, Amsterdam, in 2000. In 2001 she made the switch to St. Joost Academy of Art in Breda, and in 2004 moved to studying Business Communication and Digital Media at Tilburg University in 2004. She completed her Bachelor's degree in 2007, followed by a Master's degree in Human Aspects of Information Technology in 2008. Her Master thesis addresses issues regarding quality assessment of large corpora.

Matje van de Camp next worked as a functional designer at ABS LBS, a software company in the laundry industry. In 2009, she joined the Induction of Linguistic Knowledge (ILK) research group at Tilburg University as a PhD student for the HiTiME project under the supervision of prof. Antal van den Bosch. His move to Radboud University in Nijmegen in 2011 was counterbalanced by expanding the supervision team by prof. Eric Postma. The HiTiME project's main goal was to chart the evolution of the Dutch Social Movement using text analytics and visual representations. The project was funded by NWO under the Continuous Access To Cultural Heritage programme, and conducted in cooperation with the International Institute of Social History in Amsterdam. Her research within HiTiME focussed on the extraction of social networks from biographical texts.

In 2013, she founded the company De Taalmonsters, with which she develops intelligent search applications for people in the fields of Linguistics, Literary studies, Journalism, and beyond.

LIST OF PUBLICATIONS

Journal

- Van de Camp, M., Van den Bosch, A. (2012). The socialist network. *Decision Support Systems*, 53(4), 761-769.

Conference & Workshop

- Van de Camp, M., Van den Bosch, A. (2011). A link to the past: constructing historical social networks. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 61-69). Portland, Oregon, United States: ACL.

- Van de Camp, M., & Christiansen, H. (2013). Resolving relative time expressions in Dutch text with Constraint Handling Rules. In D. Duchier, & Y. Parmentier, *Constraint Solving and Language Processing* (pp. 166-177). Orléans, France: Springer.

Technical Report

- Van den Bosch, A., Zervanou, K., Van de Camp, M., Van den Hoven, M., Hunt, S., Van der Heijden, M. (2010). Baseline measurement CATCH-HiTime, version 1.0. Tilburg University.

Award

- Van de Camp, M. (2012). *Leveraging historic social networks using present-day data*. Winning submission for Leipzip eHumanities Innovation Award 2012. Leipzig University.

- Van de Camp, M. (2013, February 8). *Leveraging historic social networks using present-day data* [Video file]. Retrieved from https://www.youtube.com/watch?feature=player_embedded&v=dp_odkdC3qM

SIKS DISSERTATION SERIES

2009

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

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
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

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

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 Muehl (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 Gurkok (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

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 B˚rner (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 LÈon Planken (TUD)
Algorithms for Simple Temporal Reasoning
2013-43 Marc Bron (UVA)
Exploration and Contextualization through Interaction and Concepts

2014

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 Vinicius 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 Cock (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 (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 Ooijen (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

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 Bˆsch (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 AndrÈ 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 LoÔs VanhÈe (UU)
Using Culture and Values to Support Flexible Coordination
2015-21 Sibren Fetter (OUN)
Using Peer-Support to Expand and Stabilize Online Learning
2015-22 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 column-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 (OUN)
Real-time Multimodal Emotion Recognition in E-Learning
2015-31 Yakup Koç (TUD)
On Robustness of Power Grids
2015-32 Jerome Gard (UL)
Corporate Venture Management in SMEs
2015-33 Frederik Schadd (UM)
Ontology Mapping with Auxiliary Resources
2015-34 Victor de Graaff (UT)
Geosocial Recommender Systems
2015-35 Junchao Xu (TUD)
Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

2016

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

TiCC Dissertation Series

1. Pashiera Barkhuysen. *Audiovisual Prosody in Interaction.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 3 October 2008.
2. Ben Torben-Nielsen. *Dendritic Morphology: Function Shapes Structure.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotor: K.P. Tuyls. Tilburg, 3 December 2008.
3. Hans Stol. *A Framework for Evidence-based Policy Making Using IT.* Promotor: H.J. van den Herik. Tilburg, 21 January 2009.
4. Jeroen Geertzen. *Dialogue Act Recognition and Prediction.* Promotor: H. Bunt. Co-promotor: J.M.B. Terken. Tilburg, 11 February 2009.
5. Sander Canisius. *Structured Prediction for Natural Language Processing.* Promotores: A.P.J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.
6. Fritz Reul. *New Architectures in Computer Chess.* Promotor: H.J. van den Herik. Co-promotor: J.W.H.M. Uiterwijk. Tilburg, 17 June 2009.
7. Laurens van der Maaten. *Feature Extraction from Visual Data.* Promotores: E.O. Postma, H.J. van den Herik. Co-promotor: A.G. Lange. Tilburg, 23 June 2009 (cum laude).
8. Stephan Raaijmakers. *Multinomial Language Learning.* Promotores: W. Daelemans, A.P.J. van den Bosch. Tilburg, 1 December 2009.
9. Igor Berezhnoy. *Digital Analysis of Paintings.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 7 December 2009.
10. Toine Bogers. *Recommender Systems for Social Bookmarking.* Promotor: A.P.J. van den Bosch. Tilburg, 8 December 2009.
11. Sander Bakkes. *Rapid Adaptation of Video Game AI.* Promotor: H.J. van den Herik. Co-promotor: P. Spronck. Tilburg, 3 March 2010.
12. Maria Mos. *Complex Lexical Items.* Promotor: A.P.J. van den Bosch. Co-promotores: A. Vermeer, A. Backus. Tilburg, 12 May 2010 (in collaboration with the Department of Language and Culture Studies).
13. Marieke van Erp. *Accessing Natural History. Discoveries in data cleaning, structuring, and retrieval.* Promotor: A.P.J. van den Bosch. Co-promotor: P.K. Lendvai. Tilburg, 30 June 2010.
14. Edwin Commandeur. *Implicit Causality and Implicit Consequentiality in Language Comprehension.* Promotores: L.G.M. Noordman, W. Vonk. Co-promotor: R. Cozijn. Tilburg, 30 June 2010.
15. Bart Bogaert. *Cloud Content Contention.* Promotores: H.J. van den Herik, E.O. Postma. Tilburg, 30 March 2011.
16. Xiaoyu Mao. *Airport under Control.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotores: N. Roos, A. Salden. Tilburg, 25 May 2011.
17. Olga Petukhova. *Multidimensional Dialogue Modelling.* Promotor: H. Bunt. Tilburg, 1 September 2011.
18. Lisette Mol. *Language in the Hands.* Promotores: E.J. Krahmer, A.A. Maes, M.G.J. Swerts. Tilburg, 7 November 2011 (cum laude).
19. Herman Stehouwer. *Statistical Language Models for Alternative Sequence Selection.* Promotores: A.P.J. van den Bosch, H.J. van den Herik. Co-promotor: M.M. van Zaanen. Tilburg, 7 December 2011.
20. Terry Kakeeto-Aelen. *Relationship Marketing for SMEs in Uganda.* Promotores: J. Chr. van Dalen, H.J. van den Herik. Co-promotor: B.A. Van de Walle. Tilburg, 1 February 2012.
21. Suleman Shahid. *Fun & Face: Exploring non-verbal expressions of emotion during playful interactions.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 25 May 2012.

22. Thijs Vis. *Intelligence, Politie en Veiligheidsdienst: Verenigbare Grootheden?* Promotores: T.A. de Roos, H.J. van den Herik, A.C.M. Spapens. Tilburg, 6 June 2012 (in collaboration with the Tilburg School of Law).
23. Nancy Pascall. *Engendering Technology Empowering Women.* Promotores: H.J. van den Herik, M. Diocaretz. Tilburg, 19 November 2012.
24. Agus Gunawan. *Information Access for SMEs in Indonesia*. Promotor: H.J. van den Herik. Co-promotores: M. Wahdan, B.A. Van de Walle. Tilburg, 19 December 2012.
25. Giel van Lankveld. *Quantifying Individual Player Differences.* Promotores: H.J. van den Herik, A.R. Arntz. Co-promotor: P. Spronck. Tilburg, 27 February 2013.
26. Sander Wubben. *Text-to-text Generation Using Monolingual Machine Translation.* Promotores: E.J. Krahmer, A.P.J. van den Bosch, H. Bunt. Tilburg, 5 June 2013.
27. Jeroen Janssens. *Outlier Selection and One-Class Classification*. Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 11 June 2013.
28. Martijn Balsters. *Expression and Perception of Emotions: The Case of Depression, Sadness and Fear*. Promotores: E.J. Krahmer, M.G.J. Swerts, A.J.J.M. Vingerhoets. Tilburg, 25 June 2013.
29. Lisanne van Weelden. *Metaphor in Good Shape.* Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 28 June 2013.
30. Ruud Koolen. *"Need I say More? On Overspecification in Definite Reference."* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 20 September 2013.
31. J. Douglas Mastin. *Exploring Infant Engagement. Language Socialization and Vocabulary. Development: A Study of Rural and Urban Communities in Mozambique.* Promotor: A.A. Maes. Co-promotor: P.A. Vogt. Tilburg, 11 October 2013.
32. Philip C. Jackson. Jr. *Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language.* Promotores: H.C. Bunt, W.P.M. Daelemans. Tilburg, 22 April 2014.
33. Jorrig Vogels. *Referential choices in language production: The Role of Accessibility.* Promotores: A.A. Maes, E.J. Krahmer. Tilburg, 23 April 2014.
34. Peter de Kock. *Anticipating Criminal Behaviour.* Promotores: H.J. van den Herik, J.C. Scholtes. Co-promotor: P. Spronck. Tilburg, 10 September 2014.
35. Constantijn Kaland. *Prosodic marking of semantic contrasts: do speakers adapt to addressees?* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 1 October 2014.
36. Jasmina Marić. *Web Communities, Immigration and Social Capital.* Promotor: H.J. van den Herik. Co-promotores: R. Cozijn, M. Spotti. Tilburg, 18 November 2014.
37. Pauline Meesters. *Intelligent Blauw*. Promotores: H.J. van den Herik, T.A. de Roos. Tilburg, 1 December 2014.
38. Mandy Visser. *Better use your head. How people learn to signal emotions in social contexts.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 10 June 2015.
39. Sterling Hutchinson. How symbolic and embodied representations work in concert. Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2015.
40. Marieke Hoetjes. *Talking hands. Reference in speech, gesture and sign.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 7 October 2015
41. Elisabeth Lubinga. *Stop HIV. Start talking? The effects of rhetorical figures in health messages on conversations among South African adolescents*. Promotores: A.A. Maes, C.J.M. Jansen. Tilburg, 16 October 2015.
42. Janet Bagorogoza. *Knowledge Management and High Performance. The Uganda Financial Institutions Models for HPO.* Promotores: H.J. van den Herik, B. van der Walle, Tilburg, 24 November 2015.
43. Hans Westerbeek. *Visual realism: Exploring effects on memory, language production, comprehension, and preference.* Promotores: A.A. Maes, M.G.J. Swerts. Co-promotor: M.A.A. van Amelsvoort. Tilburg, 10 Februari 2016.
44. Matje van de Camp. *A Link to the Past: Constructing Historical Social Networks from Unstructured Data.* Promotores: A.P.J. van den Bosch, E.O. Postma. Tilburg, 2 maart 2016.