Consuming America

Text Mining the Dutch Infatuation with the United States

Melvin Wevers

November 6, 2017

Digital Humanities Group - KNAW Humanities Cluster

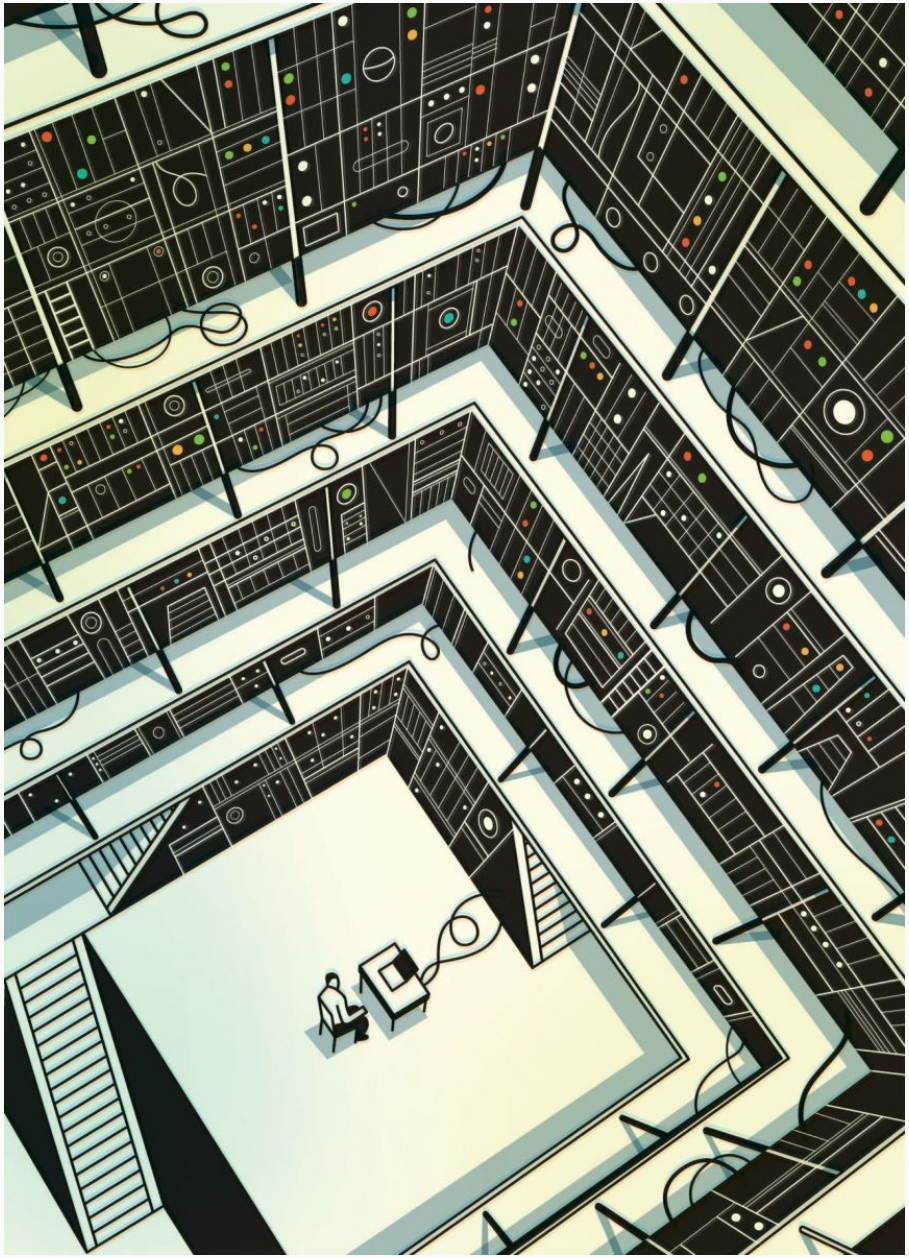
Table of contents

Introduction

Project goal

"computational methods to analyze the role of reference cultures in debates about social issues and collective identities, looking specifically at the emergence of the United States in public discourse in the Netherlands from the end of the nineteenth century to the end of the Cold War."

Digitized Newspapers as a Historical Source



- Newspapers as proxy for public discourse
- Periodical source that represents "broad, multiform collection of opinions and attitudes." (Van Vree, 1989)
- Millions of articles and advertisements

Case Study: Cigarettes

- Consumer Goods
- Product-country image, (skegaard, 1998)
- Product that defined America (Brandt, 2009)
- How did newspaper discourse on cigarettes reflect America's role as a reference culture?

Figure 1: *Limburger Koerier*, April 13, 1938

Approach

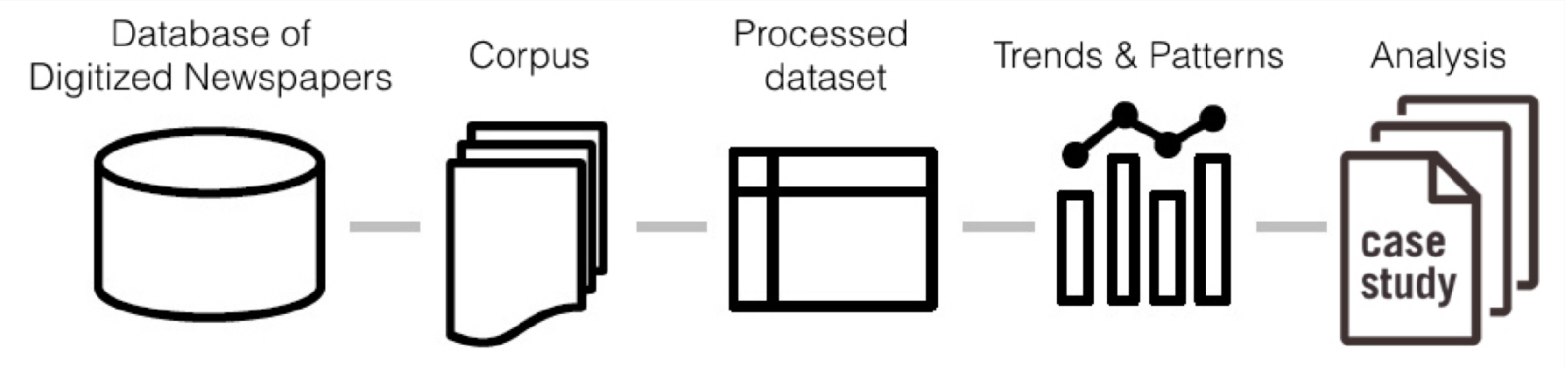


Figure 2: workflow for working with digitized newspapers

Understanding the Dataset

What's in the data?

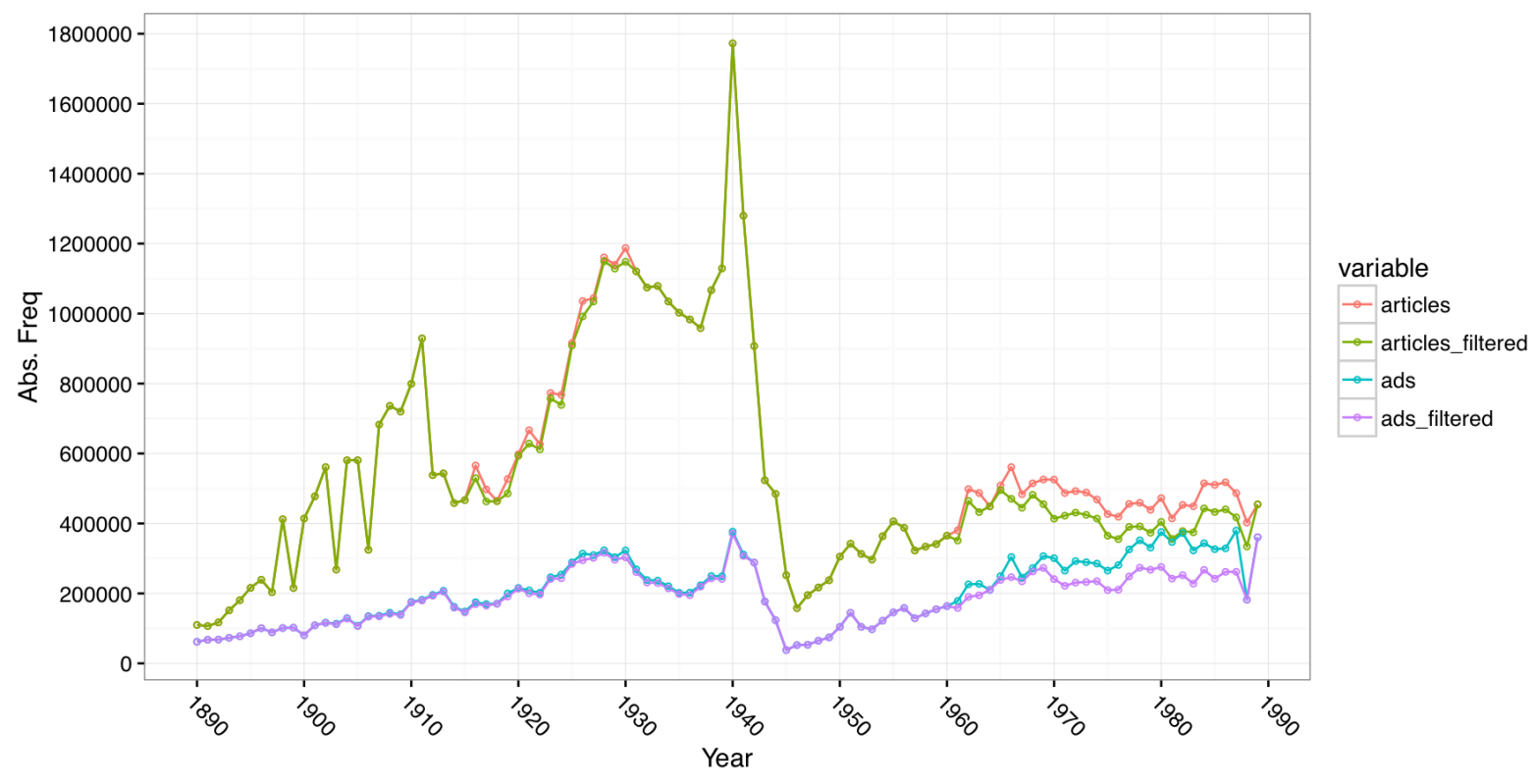


Figure 3: 52.5 million articles — 18.7 million advertisements

Distribution of regional and national newspapers

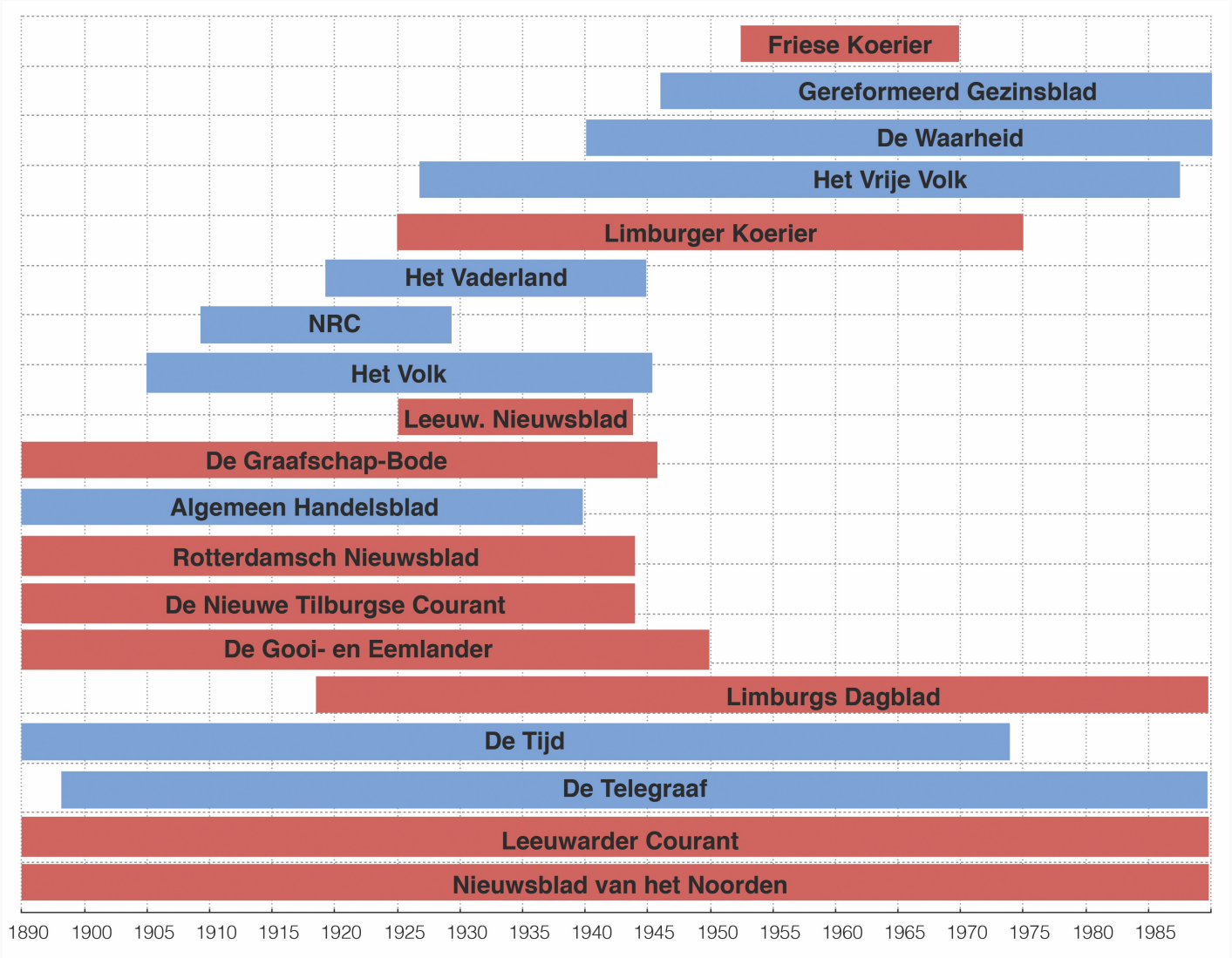


Figure 4: Temporal distribution of newspapers

Corpus Construction

Making a corpus

- Metadata filtering (document type, date, newspaper title, etc.)
- Search query
- Advanced search operators
  - Boolean (AND, OR, NOT
  - Proximity ("roken amerika" 10 -"roken amerika" 1)
  - Fuzzy Matching (+Verenigde 2 +Staten 2)
  - Regular Expressions (vere{1,2}nigde staten—ameri[k—c]a*)
- **Query**: Articles with references to the United States *and* cigarettes published between 1890 and 1990 in national and regional newspapers
- **Result**: Dataset (CSV or JSON)

Pre-processing a corpus

- Text cleaning
  - Tokenization
  - Stop word removal
  - Remove digits, punctuation, etc.
  - Remove duplicate articles
  - Filter out frequent and infrequent words
- Feature Extraction
  - N-gram counts
  - TF-IDF matrix (Term Frequency * Inverse Document Frequency
  - Topic modeling

Extract Trends and Patterns

Methods to Extract Trends and Patterns

- Document distribution
- N-gram analysis
- Co-occurrence networks
- Topic modeling

Document Distribution

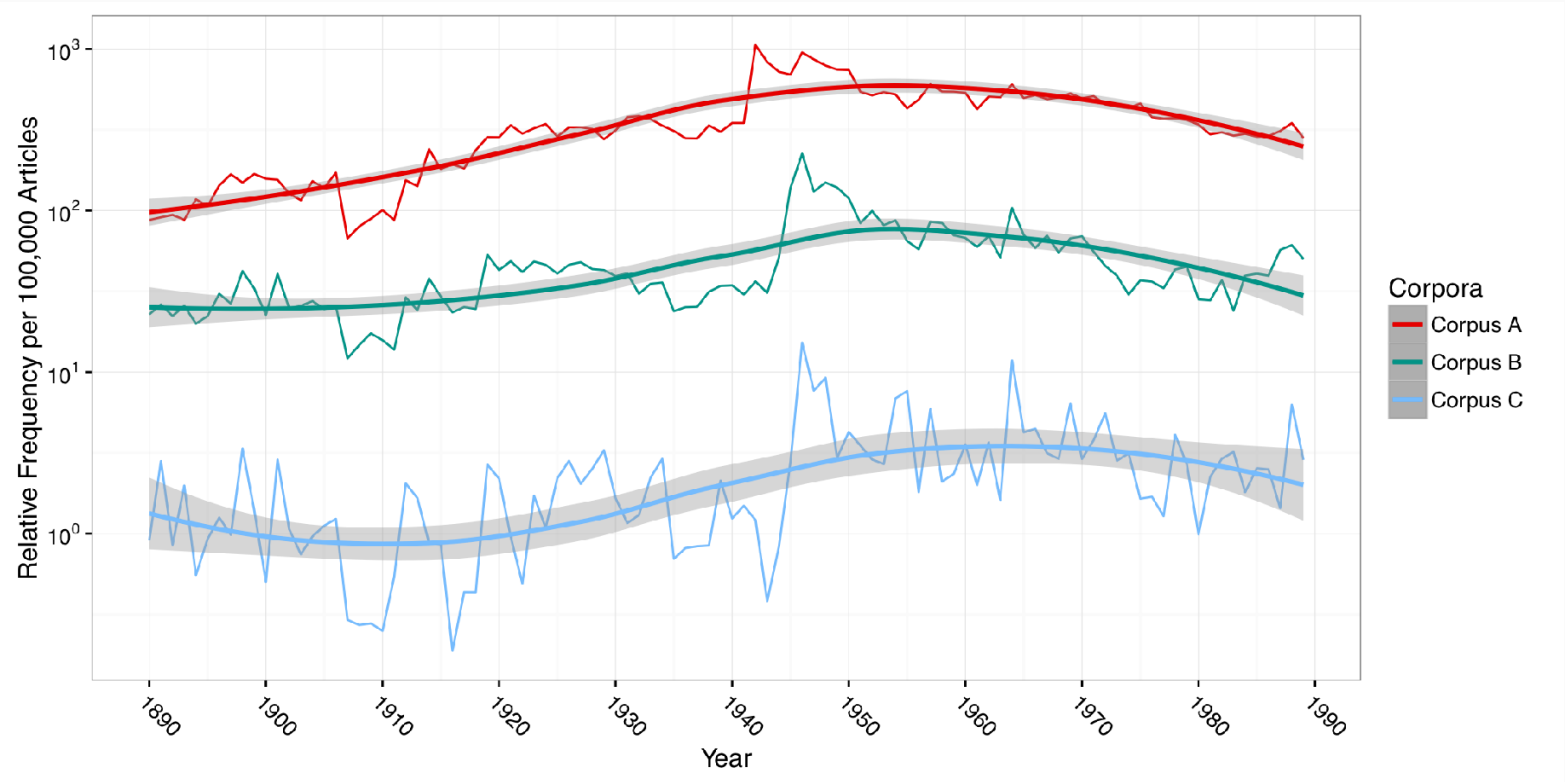


Figure 5: Relative frequency of articles on cigarettes (n = 185,004), cigarettes and the United States (n = 22,565), and the latter two in close proximity (n = 1,133)

Closer Examination of Peaks

Figure 6: Word Cloud 1954-1957

Figure 7: Word Cloud 1964

N-gram Analysis: American Products

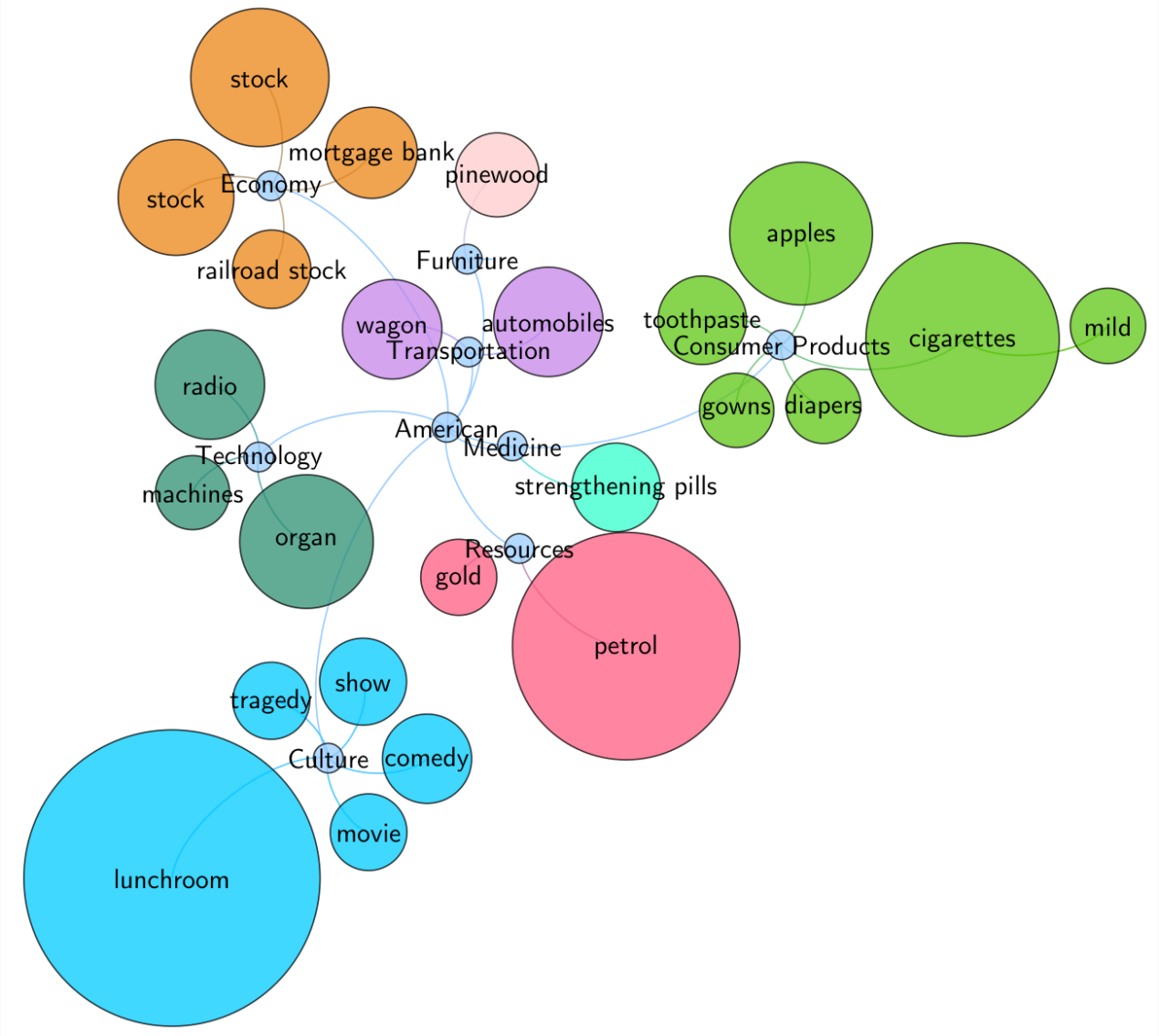


Figure 8: Bigrams with 'American' as advertisements 1930-1939

N-gram Analysis: The Nationality of Cigarettes

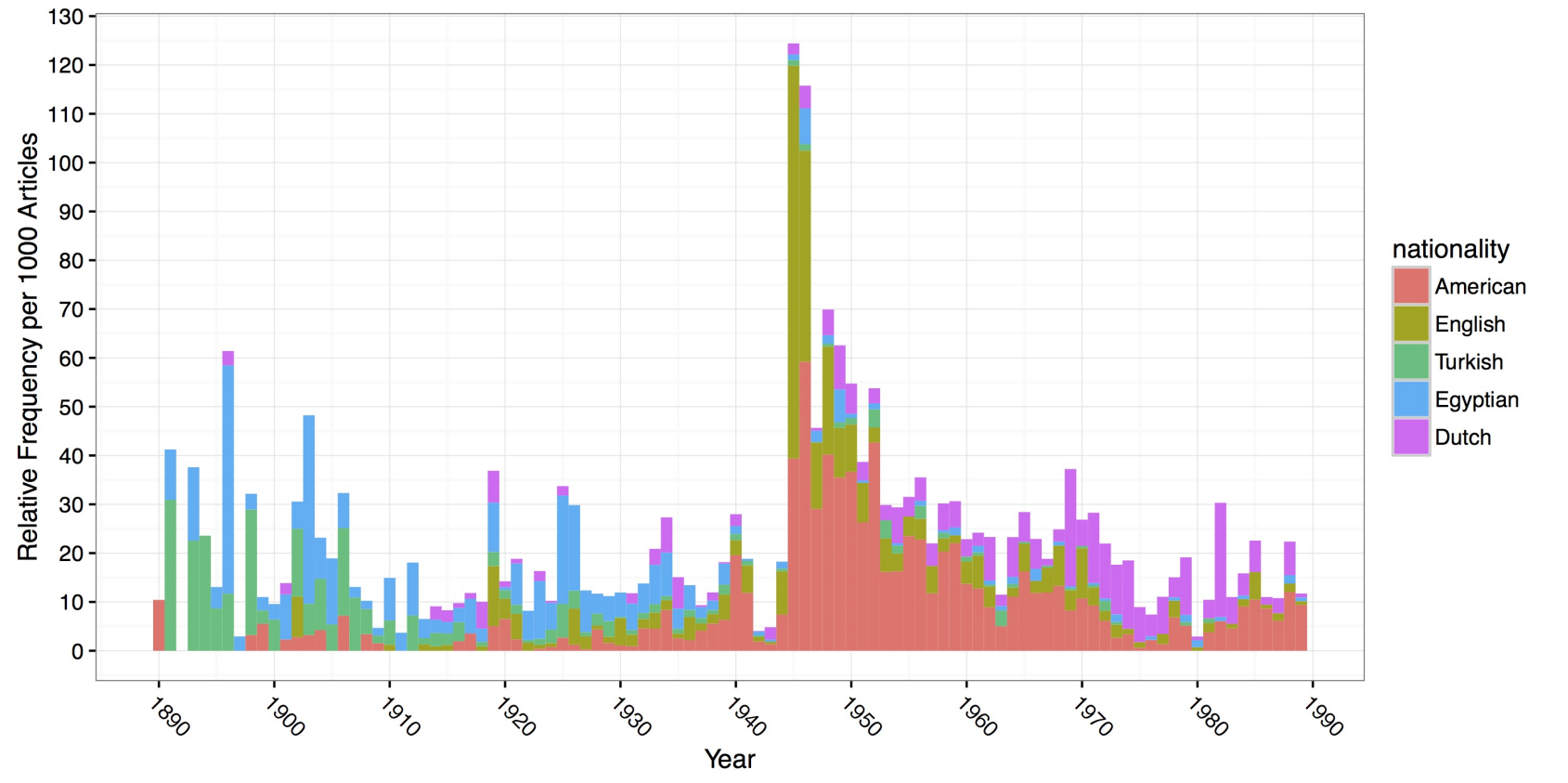


Figure 9: Nationalities associated with cigarettes in articles

Co-occurence Networks

- Count co-occurring words within specific span
- Mutual Information: $I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \left(\frac{P(x,y)}{P(x)\,P(y)} \right)$
- Compare between different types of cigarettes and periods
- Visualize using Gephi

Co-occurence Networks

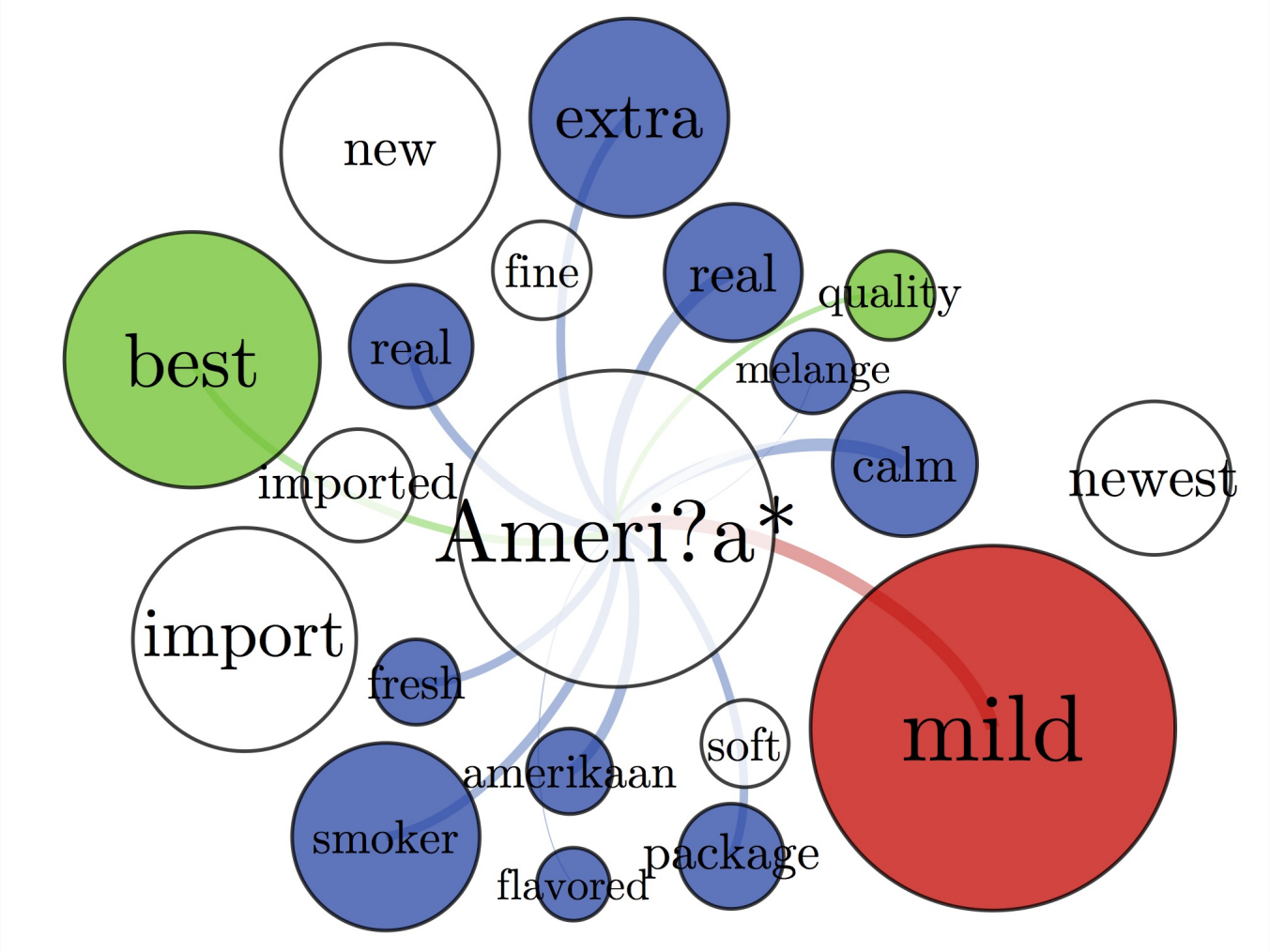


Figure 10: Features associated with American cigarettes, 1919-1939

Co-occurence Networks

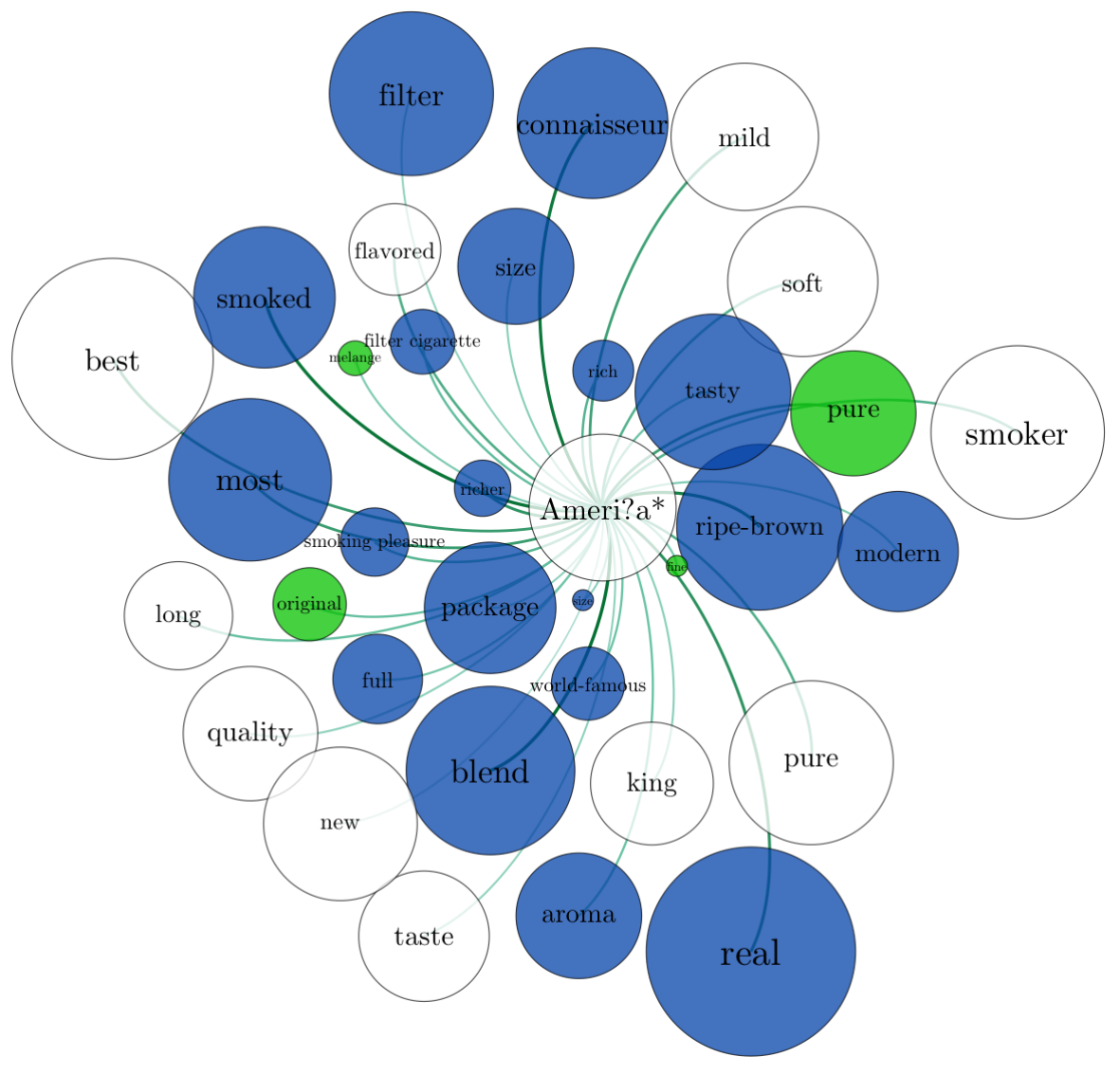


Figure 11: Features associated with American cigarettes, 1945-1970

Topic Modeling

- Probabilistic model to uncover hidden structure of set of documents (Blei, 2002)
- A topic is a cluster of words that often occur together
- Topic modeling offers a simple way to browse, search, and summarize large volumes of text

Topic Modeling Discourse on Cigarettes

H3	Highest Prob: roken, longkanker, sigaretten, rokers, rapport, amerikaanse, tussen FREX: longkanker, rokers, kanker, meinsma, sigarettenrokers, roken, ziekten Lift: kankerregistratie, sterfgevallen, verwekken, longkanker, meinsma, sterftecijfer, sigarettenrokers	Lung cancer / report / meinsma	health
H11	Highest Prob: jaar, sigaretten, procent, miljoen, amerikaanse, roken, miljard FREX: miljard, procent, waarschuwing, januari, miljoen, ministerie, verbruik Lift: wetsontwerp, sigarenwinkeliers, sigarettenreclame, shag, sigaartjes, tabaksverbruik, sigarettenverbruik	Cigarette consumption in the United States / Ban on advertising.	Culture / Politics / Health
H18	Highest Prob: sigaret, tabak, sigaretten, nederlandse, merken, merk, nederland FREX: teer, merken, tabak, merk, nicotine, produkt, filter Lift: typen, consumentenbond, nicotinegehalte, consument, teer, importeur, turmac	Introduction of filter cigarettes in the Netherlands. Dutch debates on health risks of smoking	Health

Figure 12: Topic model (n = 2,508; k = 20) of articles on cigarettes, 1960-1969

Data-Driven Analysis

Data-Driven Analysis

- Contrasting output from several computational techniques
- Trace continuity and change in discourse
- Query expansion
- Guide further explorations of the archive

Iterate between close and distant reading

- Read the actual sources
- Refine query and corpus
- Generate hypotheses
- Combine insights to construct historical narrative

Take-Home Messages

Some pointers....

- Get to know your data
- Computation is not an replacement but an addition to traditional hermeneutics and heuristics
- Don't rely on one monolithic tool
- Some basic programming skills can come in handy

Questions?

melvinwevers@icloud.com

`http://www.melvinwevers.nl`