

# TEXT ANALYSIS

---

Christopher Gandrud

SG1022, City University London

- What is text analysis and why use it?
- Human coding
- Automated coding
- Pitfalls

## DEFINING TEXT ANALYSIS

---

“When we perform textual analysis on a text, we make an educated guess at some of the most likely interpretations that might be made of that text.” (McKee 2003, 1)

“Content Analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter [e.g. videos, audio]) to the contexts of their use.” (Krippendorff 2013, 24)

“Content Analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter [e.g. videos, audio]) to the contexts of their use.”

**Replicable:** different researchers, independent of each other should get the same results when applying the same technique.

E.g. independent researchers come to the understanding of the text using the same method.

Replicable results are more reliable.

“Content Analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter [e.g. videos, audio]) to the contexts of their use.”

Valid: research is open to careful scrutiny and your claims can be upheld given independently available evidence.

“Content Analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter [e.g. videos, audio]) to the contexts of their use.”

Texts: something that is produced by someone to have meaning for someone else.

E.g. newspaper articles, treaties, transcripts, tweets, maps, advertisements, press releases, movies, party manifestos.

In this course we focus exclusively on texts composed of words.



1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.
2. Texts do not have single meanings.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.
2. Texts do not have single meanings.
3. Meanings invoked by texts need not be shared.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.
2. Texts do not have single meanings.
3. Meanings invoked by texts need not be shared.
4. Contents refer to something other than themselves.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.
2. Texts do not have single meanings.
3. Meanings invoked by texts need not be shared.
4. Contents refer to something other than themselves.
5. Texts have meanings relative to particular contexts.

1. Texts have no objective–reader-independent qualities. Meaning (data) arises from someone reading the text, often expecting other's understanding.
2. Texts do not have single meanings.
3. Meanings invoked by texts need not be shared.
4. Contents refer to something other than themselves.
5. Texts have meanings relative to particular contexts.
6. Content analysts infer answers to particular research questions from their texts. Their inferences are merely more systematic, explicitly informed, and verifiable ...than what ordinary readers do.

## WHY USE TEXT ANALYSIS?

---

You *use* and *contribute* to text analysis every day.





kanye west is



kanye west is **dead**  
kanye west is **jesus**  
kanye west is **from**  
kanye west is **worth**



the labour party is



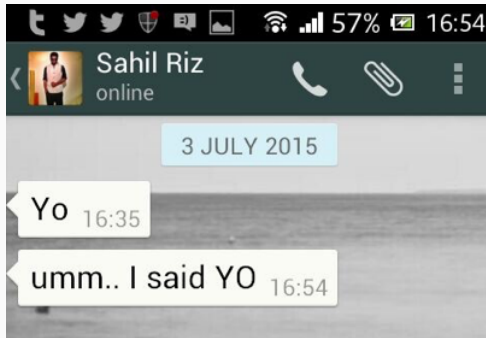
the labour party is **now a threat**

the labour party is **heading for a showdown on trident**

the labour party is **finished**

the labour party is **a socialist party and proud of it**

(Some of you) are building a data set that will be used  
for text analysis **right now**.



Source: <http://www.buzzfeed.com/shayanroy/blocking-you-now.obE7eXDgAP>

People are creating increasingly more (machine accessible) texts.

Massive new source of data for social science analysis.

We may have research questions where we conducted a survey with an open-ended question.

We need some systematic way to understand these texts and make comparisons across survey respondents.

We may have research questions where we want to interview a group of people that are hard to access, but who produce many texts.

For example, in an ideal world we may want to survey world leaders for their preferences to handling Syrian refugees. We may want to see how these preferences change over time.

World leaders don't given many interviews (especially not multiple interviews on the same topic), but they—often filtered through a press office—do create many texts.



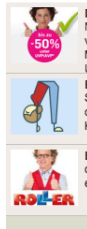


11. Januar 2016, 20:22 Uhr Ermittlungen zu den Übergriffen in Köln

## Kölner Polizei: Vor allem Marokkaner fallen auf

- Die Ermittler haben bisher 19 Tatverdächtige identifiziert.
- Neun Verdächtige halten sich illegal in Deutschland auf, zehn Personen sind Asylsuchende.
- Von den 19 Verdächtigen kommen 14 aus nordafrikanischen Ländern, vor allem aus Marokko.
- Den Statistiken der Kölner Ermittler zufolge werden 40 Prozent der nordafrikanischen Zuwanderer innerhalb eines Jahres straffällig.

ANZEIGE



D  
N  
vi  
M  
U  
  
B  
S  
di  
K  
  
B  
di  
ei

Source: <http://www.sueddeutsche.de/panorama/ermittlungen-zu-den-uebergriffen-in-koeln-vor-allem-marokkaner-fallen-auf-1.2814336>

January 11, 2016, 20:22 Investigations on the attacks in Cologne

## **Cologne police: Especially Moroccans to fall**

- Investigators have identified so far 19 suspects.
- Nine suspects keep illegal in Germany, ten people are asylum seekers.
- 14 Of the 19 suspects come from North African countries, mainly from Morocco.
- Statistics of Cologne investigators According to 40 percent of North African immigrants become delinquent within a year.

Source: <http://www.sueddeutsche.de/panorama/ermittlungen-zu-den-uebergreifen-in-koeln-vor-allem-marokkaner-fallen-auf-1.2814336>

via Google Translate

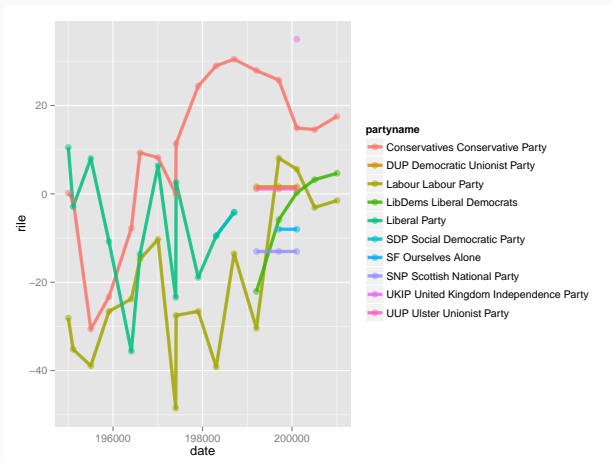
A word cloud visualization of a German Chancellor's press release from January 2015. The words are arranged in a circular pattern, with the most prominent words in the center. The words are color-coded: orange, blue, purple, yellow, and grey. The central words are 'minister' (grey), 'federal' (grey), and 'refugees' (yellow). Other prominent words include 'crimes' (purple), 'germany' (purple), 'laws' (purple), 'offender' (purple), 'justice' (blue), 'maas' (blue), 'interior' (blue), 'sexual' (blue), 'heiko' (blue), 'criminals' (purple), 'government' (purple), 'chancellor' (orange), 'police' (orange), 'she' (orange), 'good' (orange), 'answers' (orange), 'country' (orange), 'offenders' (orange), 'integration' (orange), 'swift' (orange), 'residence' (orange), 'violence' (orange), 'television' (blue), 'german' (blue), 'attacks' (blue), 'deport' (orange), 'steps' (orange), 'leave' (orange), 'account' (blue), 'called' (orange), 'law' (orange), 'state' (blue), 'place' (blue), 'declared' (orange), 'commit' (orange), 'protection' (orange), 'maizièrè' (orange), 'response' (orange), 'legislation' (orange), and 'thomas' (orange).

response legislation thomas  
protection maizièrè minister  
declared place  
commit state  
leave account  
called law  
crimes  
germany  
deport attacks  
steps german  
television  
refugees  
violence residence swift  
maas chancellor police  
interior she  
sexual heiko  
criminals government  
laws  
offender justice  
answers country  
good  
offenders  
integration

We may have research questions about units that are not able to be surveyed, but which produce texts.

E.g. International organisations, political parties, neighbourhood groups.

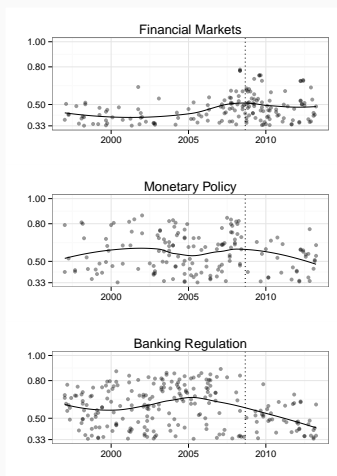
## Left-Right Position of UK Parties Based on their Party Manifestos



We may have research questions about how actors communicate to achieve goals.

For example, what topics do monetary policy bureaucrats talk about more when there is a financial crisis?

# TOPICS OF US FEDERAL RESERVE GOVERNOR SPEECHES



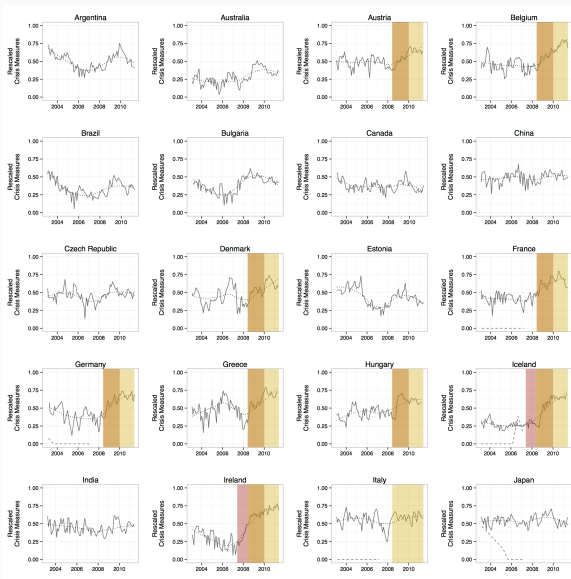
Source: Young and Gandrud (2016)

We may have research questions about widely held beliefs across time, where a survey would be too costly to run.

For example, if we wanted to study monthly perceptions of financial market stress across 180 countries.



# REAL-TIME PERCEPTIONS OF FINANCIAL MARKET STRESS



Generally, text analysis results in data that is:

- Nominal: e.g. the main topic of a text.
- Continuous: e.g. scale (negative to positive, left-right), proportion of a document dedicated to a specific word or words.

## HUMAN AND MACHINE CODING

---

You can analyse texts either by relying exclusively on human coders or primarily rely on machine-assistance.

Note: you should never exclusively rely on machine coding. At a minimum, you need to check the validity of your machine codes. Do they make sense?

Machine coding has the advantage of being much more efficient for large numbers of texts + more easily reproducible.