# ExPoSe

## Examples of Exploring the Dutch Hansards

Maarten Marx

ExPoSe Kickoff

2013-10-23

# Content

What can you do with 200 years of Hansards on your PC?

- Trendspotting in historical collections
  - Counting phrases (ngrams)
- Summarization and characterization
  - Natural language processing
  - Statistical language models
- Turn Hansards 90 degrees
  - Not event- or document-centric, but actor-centric information

# Before we start: A quick look at the Dutch Hansards

- PDF: Human readable
- XML: Human and machine readable
- In ExPoSe we turn human readable documents into machine (and human) readable documents.

# Trendspotting in historical collections

- Culturonomics, Google books, ngrams [Michel et.al. 2011]]

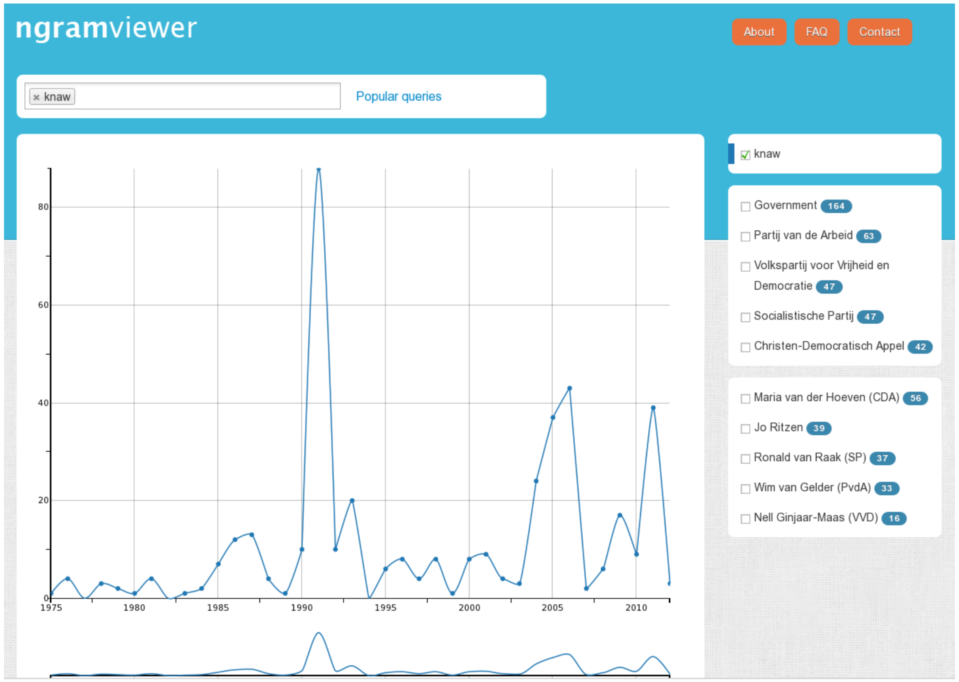
- Plot relative volume per year of phrases of up to 5 words.

- Massive amounts of data

```
    Ngram size |   Types         |    Tokens
-----------|----------------|------------------------
KBngram1   |   49.514.842   |     18.437.979.846
KBngram2   |   39.156.451   |     11.821.165.297
KBngram3   |   65.169.507   |      5.808.214.106
KBngram4   |   47.955.071   |      2.386.522.277
KBngram5   |   46.222.852   |      1.056.997.790    +
-----------|----------------|------------------------
Total      |  248.018.723   |     39.510.879.316
```

# Trendspotting in historical collections

- Provide direct access to underlying data from trend graph
- Dispectu’s Political and KB ngram viewer
- http://ngram.politicalmashup.nl/ (doe normaal, doe eens normaal)



☑ knaw

□ Government 164

□ Partij van de Arbeid 63

□ Volkspartij voor Vrijheid en Democratie 47

□ Socialistische Partij 47

□ Christen-Democratisch Appel 42

□ Maria van der Hoeven (CDA) 56

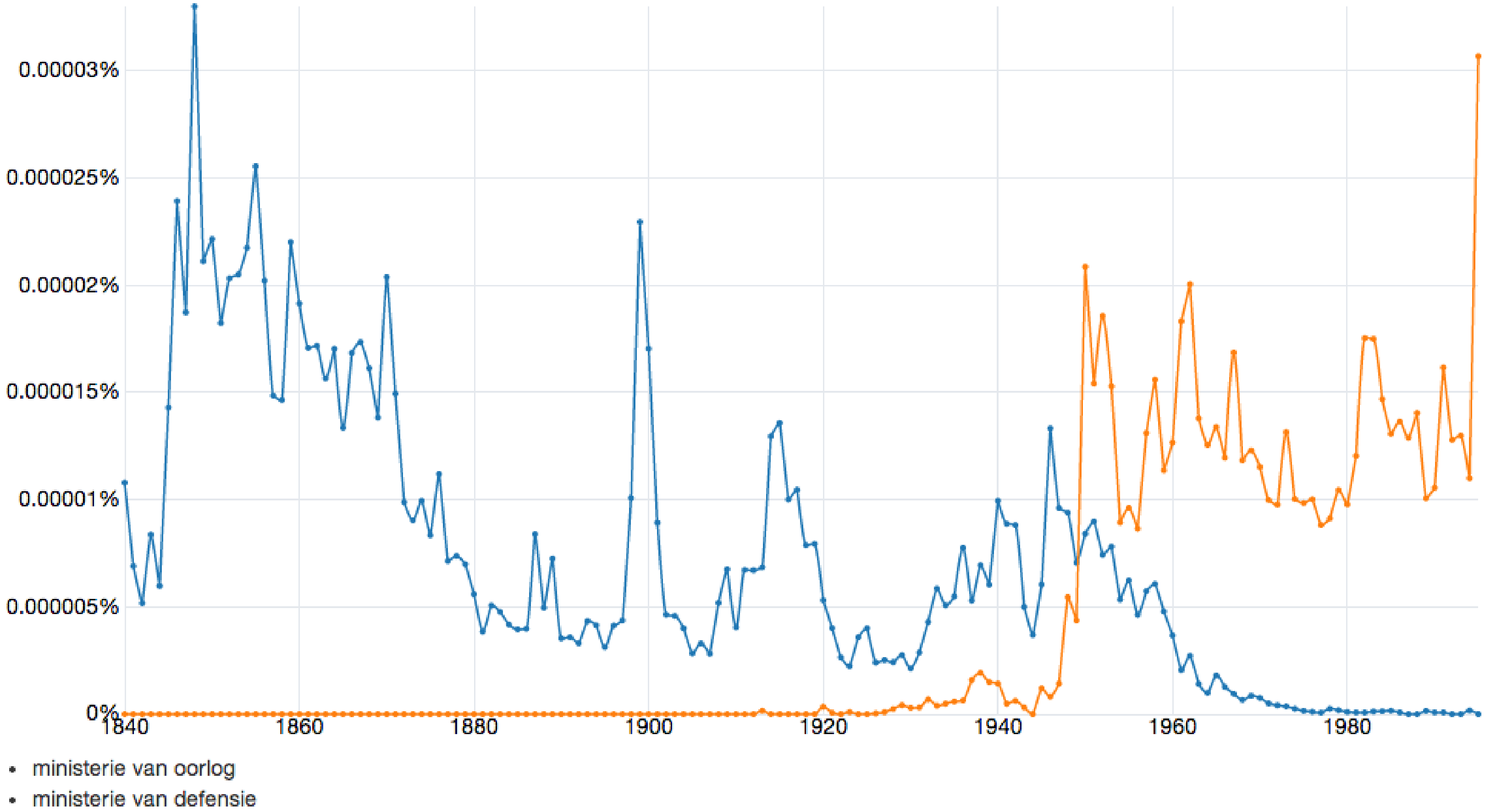
□ Jo Ritzen 39

□ Ronald van Raak (SP) 37

□ Wim van Gelder (PvdA) 33

□ Nell Ginjaar-Maas (VVD) 16

# Trendspotting in historical collections



Tellingen op de kranten collectie bij de KB.

http://kbkranten.politicalmashup.nl

# Content

- ~~Trendspotting in historical collections~~
- Summarization and characterization
  - Natural language processing
  - Statistical language models
- Turn Hansards 90 degrees
  - Not event- or document-centric, but actor-centric information

# Summarization and characterization

- Use extensive structure of Hansards to create language models of
  - individual persons
  - parties
  - government vs opposition
  - left vs right
  - MP vs MG

voortkomen
verenigen
plaatsvinden
schetsen
groeien
lukken
functioneren
bezien
liggen
voeren
beantwoorden
zoeken
menen
versterken
richten
zorgen
wijzen
refereren

plegen
plaatsen
verzoeken
meewerken
detineren
wisselen
delen
opleggen
denken
spreken
verdenken
informeren
wijzen
verblijven
sluiten
constateren
veroordelen

# Summarization and characterization

- Natural language processing and clever statistics leads to more accurate models
  - part of speech tagging (“woordsoortbenoeming”)
  - lemmatization
  - recognizing and linking of named entities
  - corpus comparison techniques

- `http://data.politicalmashup.nl/politiekinzicht/`

- Compare a generalist to a specialist
  - Generalist
  - Specialist

# Content

- ~~Trendspotting in historical collections~~
- ~~Summarization and characterization~~
- Turn Hansards 90 degrees
  - Not event- or document-centric, but actor-centric information

# Turn Hansards 90 degrees

Presentation of information depends on the amount of structure

- Unstructured: presentation is determined by physical (file) format
  - Full text search which returns complete documents (typical search engine)

# Turn Hansards 90 degrees

Presentation of information depends on the amount of structure

- Unstructured: presentation is determined by physical (file) format
  - Full text search which returns complete documents (typical search engine)
- Metadata at document level: faceted search
  - Refine text query with additional fields
  - Show distribution of hits over the facets and their values

# Turn Hansards 90 degrees

Presentation of information depends on the amount of structure

- Unstructured: presentation is determined by physical (file) format
  - Full text search which returns complete documents (typical search engine)

- Metadata at document level: faceted search
  - Refine text query with additional fields
  - Show distribution of hits over the facets and their values

- Metadata inside the documents (semi-structured texts):
  - Exploration can be made easier and faster
  - `search.politicalmashup.nl/?q="doe+eens+normaal"`

# From document- to actor-centric information in Hansards

- Trendsetter: `TheyWorkForYou.com`. View the Hansards from the perspective of your own MP: what does she say, what does she do, what does she ask?

# From document- to actor-centric information in Hansards

- Trendsetter: `TheyWorkForYou.com`. View the Hansards from the perspective of your own MP: what does she say, what does she do, what does she ask?

- Hansard turned 90 degrees:
  - From documents mentioning actors (and their speeches)
  - To actors linking to (speeches in) documents

# From document- to actor-centric information in Hansards

- Trendsetter: `TheyWorkForYou.com`. View the Hansards from the perspective of your own MP: what does she say, what does she do, what does she ask?

- Hansard turned 90 degrees:
  - From documents mentioning actors (and their speeches)
  - To actors linking to (speeches in) documents

- More than just words per person.
  - attention
  - support

# Attention

Giving and receiving attention is an indicator of power.

- We can measure attention by looking at interruptions of speakers in Parliament.

# Attention

Giving and receiving attention is an indicator of power.

- We can measure attention by looking at interruptions of speakers in Parliament.

- Since the 70's these are explicitly marked in the Hansards (using a "□" to indicate a new speaker on the central lectern).

# Attention

Giving and receiving attention is an indicator of power.

- We can measure attention by looking at interruptions of speakers in Parliament.
- Since the 70's these are explicitly marked in the Hansards (using a "□" to indicate a new speaker on the central lectern).
- Compare the interruptions of and by 3 partyleaders in one year.
  - Quite impossible member of government
  - Possible member of government
  - Virtual member of government
  - Member of government

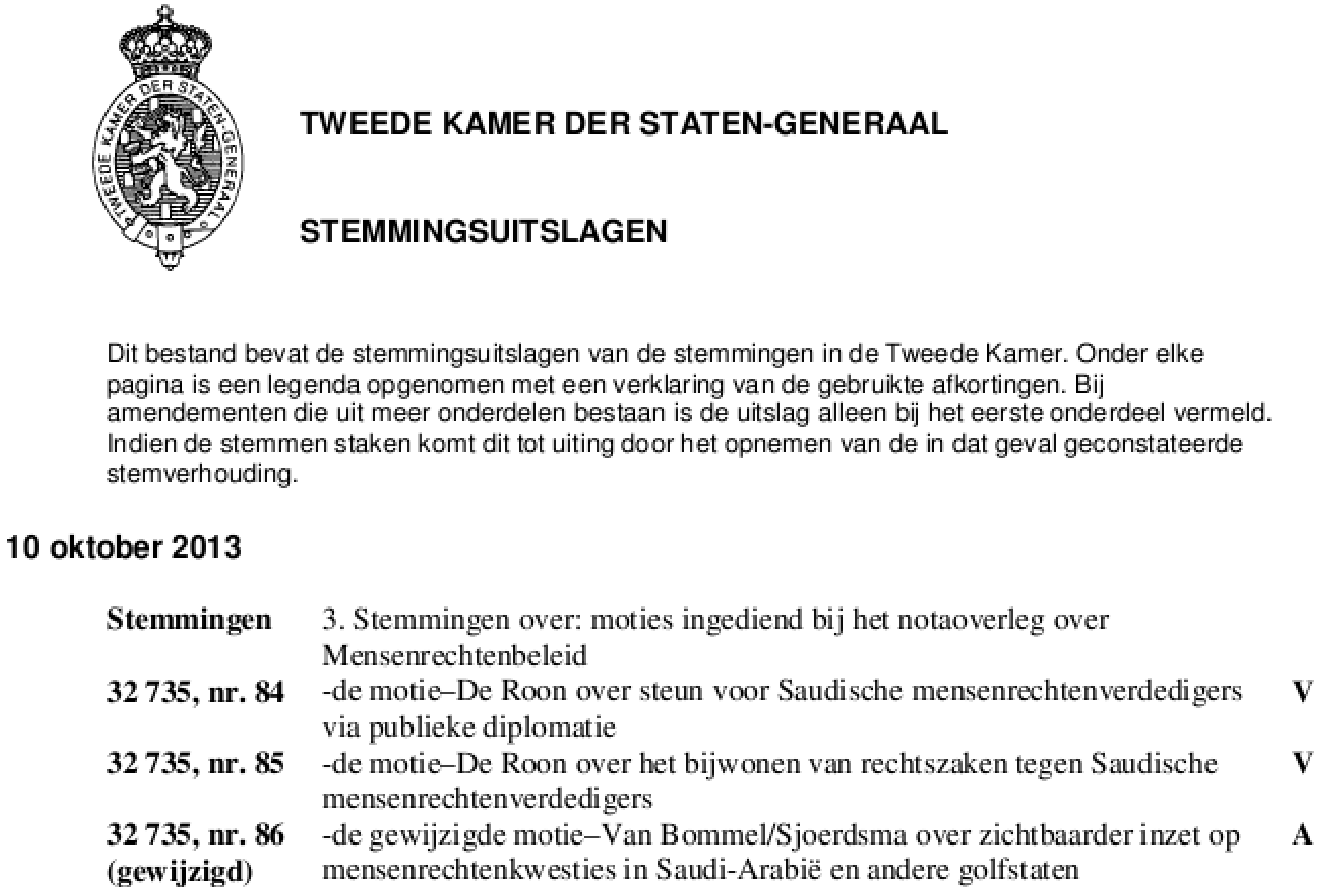
# Hansard with rich internal metadata: example votes

Turn spoken text into data

- Indicate actors and roles
- Create links to actors and mentioned documents
- Example: vote
  - Outcome as text
  - Outcome as data

# Votes: document centric

- Lists of vote-events are published as follows:



TWEEDE KAMER DER STATEN-GENERAAL

STEMMINGSUITSLAGEN

Dit bestand bevat de stemmingsuitslagen van de stemmingen in de Tweede Kamer. Onder elke pagina is een legenda opgenomen met een verklaring van de gebruikte afkortingen. Bij amendementen die uit meer onderdelen bestaan is de uitslag alleen bij het eerste onderdeel vermeld. Indien de stemmen staken komt dit tot uiting door het opnemen van de in dat geval geconstateerde stemverhouding.

**10 oktober 2013**

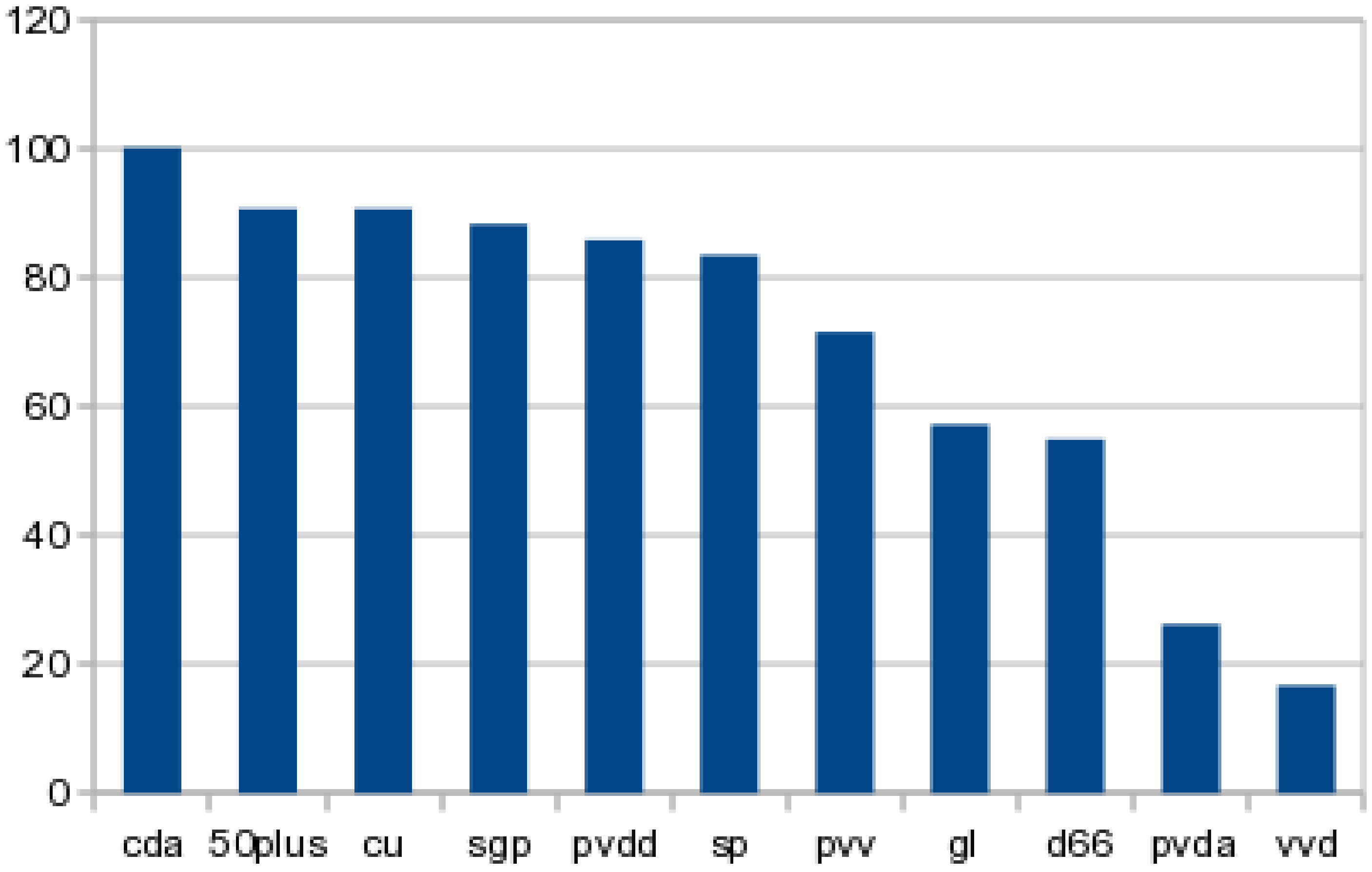
| | | |
|---|---|---|
| **Stemmingen** | 3. Stemmingen over: moties ingediend bij het notaoverleg over Mensenrechtenbeleid | |
| **32 735, nr. 84** | -de motie–De Roon over steun voor Saudische mensenrechtenverdedigers via publieke diplomatie | **V** |
| **32 735, nr. 85** | -de motie–De Roon over het bijwonen van rechtszaken tegen Saudische mensenrechtenverdedigers | **V** |
| **32 735, nr. 86 (gewijzigd)** | -de gewijzigde motie–Van Bommel/Sjoerdsma over zichtbaarder inzet op mensenrechtenkwesties in Saudi-Arabië en andere golfstaten | **A** |

# Votes: actor centric

- group all vote events by their initiators
- show support for actor from all other actors
- For each party, how often supported that party proposals by X?

# Support for Omtzigt

Omtzigt (CDA), 2012-2013, 41 initiated votes



# Conclusion

Exploiting available structure in political documents yields value.

- More informative search results
- Insight by aggregation
- Showing information from relevant perspectives (not only documents but also events, actors, temporal).

# To come in ExPoSE:

- More and older material from all ExPoSe partners
- More links to and from Hansards
- Other actors featured in the Hansards
  - Persons
  - Companies
  - Organisations
  - Quangos
- . . . recognized, disambiguated and appropriately linked
- Network analysis