The Mediascape of Dutch Chroniclers (1500-1850)

Labeling Media Mentions in Early Modern Chronicles Using CRF

Alie Lassche & Roser Morante

Leiden University, VU Amsterdam

DH Benelux. June 3, 2021

Outline I

1. Chronicles and the early modern mediascape

2. Corpus and annotation task

3. Machine learning media mentions

4. Discussion

Outline

1. Chronicles and the early modern mediascape
2. Corpus and annotation task
3. Machine learning media mentions
4. Discussion

Chronicles and the early modern mediascape

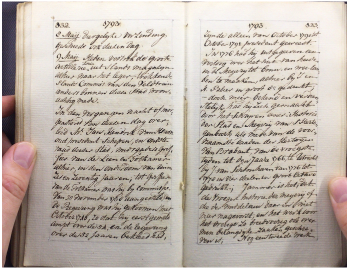
Chronicling Novelty. New knowledge in the Netherlands, 1500-1850

- Project managers: Judith Pollmann (Leiden University) & Erika Kuijpers (VU Amsterdam)
- About: Circulation and evaluation of new knowledge, ideas and technologies among a non-specialist public
- www.chroniclingnovelty.com

Changing mediascapes and the collection of knowledge

- Explore how the use of computational methods allow me to get more insight in the media early modern chroniclers used and the information they received

Chronicles and the early modern mediascape



332. 1793

8 Maij. Diergelyke Verzending geschiede ook deezen Dag.

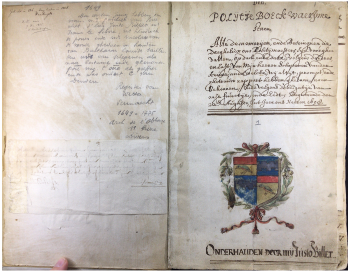
9 Maij. Heden vertrok de groote artillerie train uit Slands magazyn alhier naar het leger, trekkende Slands Commis van den Veldtrain anders binnen deeze Stad woon-achtig mede.

In den vergangen nacht of mor-genstond van deezen dag over-leed Mr. Jan Hendrik van Heum oud president Schepen, en Oudste raad deezer Stad, mitsgaders Grif-fier van de Leen en Tolkamer alhier, in den Ouderdom van Ruim Zes en Zeventig Jaaren; tot Griffier van de Tolkamer was hij by Commissie van 25 November 1738 aangesteld, in de Regeering was hij gekoomen met October 1716, zo dat hij eerstgemeld ampt over de 54, en de Regeering over de 52 Jaaren bekleed had,

1793. 333.

Zynde alleen van October 1790 tot October 1791 president geweest.

In 1776 had hij uitgegeven een Vertoog over het nut van heide in de Meyery tot bouw- en wei-lan-den te maaken, alhier by J. en A. Palier in groot 8vo. gedrukt; – doch meer bekend en verdien-stelijk had hij zich gemaakt door het schrijven eener Historie der Stad en Meyerij van 's Herto-genbosch als mede van de voor-naamste daaden der Hertogen van Brabant, van de vroegste tyden tot den Jaare 1766, te Utrecht by J. van Schoonhoven van 1776 tot 1778 in Vier deelen in groot Octavo gedrukt; Jammer is het, dat de Vroeger historie der Meyerij of die der middeleeuw daar in niet meer nagevorst, en het werk voor het overige zo breedvoerig ook over min belangrijke zaaken geschre-ven is; Nog een tweede werk



Gebracht in 186., doen binden in 1856. Van Duyse

Bill. univ. 9 Dec 1865 V [illegible]

1649

Den eersten Juny hebben den voorn. m.r pachter van Peer-vliet, d'heer Justo Billiet m.r Frans le febvre, m.r hendrick de pours ende m.r Nicolaes van de voorde ghedaen in handen van Balthasar Cruwde baillieu den eedt van schepenen als naer costume ende ghenomen pote my t'oere als greffier ende was onderst. C. van Dendere

Register van Wetten Rennaecht 1649 - 1775

Arch. de l'Abbaye St Pierre Divers

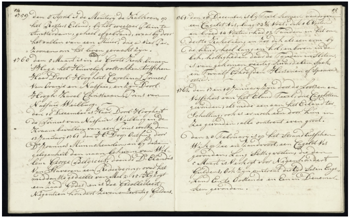
Den,

POLYTIE BOECK waerinne staen,

Alle de memoryen, ende Besoingnen die Daeghelicx ons Polytymeesters syn voorghe-vallen, op dach, ende date, volgens decrees en last van Myn heeren Schepenen vande Keure, ende welcke, wy altyt, prompt ende ghetrauw rapport hebben ghedaen, naer behooren; ende volgens d'obligatye van onse functye, ende Eedt; Beginnende den 26.e Augusto, Int Jaer ons Heeren 1659.

1

ONDERHAUDEN DOOR MIJ JUSTO BILLET.



14.

1759. den 8 April is de Moutery de Kalkoen op het Rotters Eiland, by het weesper Plein te Amsterdam, geheel afgebrand, waarby door het vallen van een Muur, drie a vier Per-zoonen om het leven geraakt zyn.

1760 den 5 Maart is in de Groote Kerk binnen 's Hage het Huwelyk voltrokken, tusschen Haar Doorl. Hoogheit. Carolina Princes van Oranje en Nassau, en Zyn Doorl. Hoogh. Karel Christiaan Prins van Nassau Weilburg.

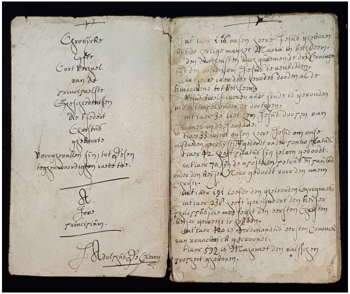
den 18 December is Haar Doorl. Hoogheit de Princes van Nassau Weilburg in de Kraam bevallen van een Prins welke den 13 January 1761 den H. Doop ontfing door Do. Joannes Munnekemolen, en by deze gelegenheit den naam bekwam van Wil-lem George Belgicus, doende Dr. Clandius van Staveren een Redevoering, over het middenste gedeelte van Mal. 2: 15: Hy Zogt een Zaad Godes en is doe Gecollecteert, Negentien hondert Zeven en Zeventig Guldens.

15

1761. den 18 December, is by Texel komen aandryven een Cagelot Vis, lang 53½ Voet, dik 16 Voeten en breed 14 Voeten, had 52 Tanden in het on-derste Kakebeen, en zoo dik als een arm by de hand, heel lang en hol, en boven in de bek, holligheden daar de Tanden in sloten, is van gekomen veertig Swaardeken Spek, en Twaalf Oxhoofden Hersenen of Sperma-cetie.

1762 den 17 en 18 January zyn door de Loodzen en Visschers van het Eiland Texel, drie Cagelotten gevonden, als mede een aan het Eiland ter Schelling, ook is er noch een door hun in Zee gevonden, alle omtrent even groot.

den 20 February is op het Strand tusschen Wijk op Zee en Zandvoort, een Cagelot Vis gevonden, Lang Sestig voeten, die den 5 Maart is Verkogt voor Negenhondert Guldens, ook zyn omtrent die tyd Ses in Enge-land en te Oostende en Een in Denemar-ken gevonden.



Chronycke
Ofte
Cort Verhael
van de
principaelste
Gheschiedenissen
die sedert
Christus
geboorte
voorgevallen syn, tot desen
tegenwoordigen iaere toe.

Et
Jesu
principium.

J. Adolphus de Cluey.

1

Int iaer 1.16 onsen Heere Jesus gheboren uyt de heylige maeght Maria in Bethleem.

den derthiensten dach quamen de drij Coningen uyt den oosten om Jesus te aenbidden.

int selve iaer dede Herodes dooden al de kinderkens tot Bethleem.

Jesus twelf iaeren audt synde is gevonden in den tempel onder de doctoren.

int iaer 30 liet hem Jesus doopen van Joannes in de Jordaen.

tiaer 33 wiert onsen Heere Jesus om onse misdaden gecruyst ende gedoodt onder Pontio Pilatus.

tiaer 42 heeft Pilatus syn selven gedoodt.

int tiaer 70 syn de apostelen Petrus en Paulus onder den Keyser Nero gedoodt voor den naem Christi.

int tiaer 191 leefde den geleerden Origines.

int tiaer 238 heeft geregneert den Keyser Philippus die men seght den eersten Christen Keyser geweest te hebben.

int tiaer 420 is Ferdinandus eersten Coninck van Vranckrijck gecroondt.

tiaer 592 is Magomet den valschen propheet geboren.

Chronicles and the early modern mediascape

- The whole of sources of information available to the early modern chronicler
- Two elements:
    - **The media that are mentioned by the authors of the chronicle**
    - The information that the authors are reporting on, which has been obtained from the media

Outline

1 Chronicles and the early modern mediascape

2 Corpus and annotation task

3 Machine learning media mentions

4 Discussion

Corpus

Corpus characteristics

- 350 manuscripts, 70,000 pages
- Written in the Dutch language in the Low Countries between 1500 and 1850
- Transcribed and annotated with Transkribus (HTR) and the help of volunteers (VeleHanden)

Three labels

- `receiver`, `source`, `perception` (`oral/heard`, `written/read`, `seen`, `else`)

This morning `<source>`mayor Vorsterman`</source>` came `<perception: oral/heard>` telling `</perception>` `<receiver>`us`</receiver>` that because of the disease, no one was allowed to be buried in the church

Annotation task

Inter Annotator Agreement shows difficulty of the task

	F-score 1	**F-score 2**
all labels	0.589	0.742
source	0.208	0.764
receiver	0.777	0.619
perception	0.707	0.727

Table: Inter Annotator Agreement in the media annotation task.

Used data in first experiments: 12 volume chronicle by Jozef Van Walleghem about Bruges (1779 - 1800)

chronicle	**p (% labeled)**	**n sources**	**n receivers**	**n perceptions**
Van Walleghem	1165 (17%)	519	272	510

Table: Characterics of the data.

Outline

1. Chronicles and the early modern mediascape

2. Corpus and annotation task

3. Machine learning media mentions

4. Discussion

Machine learning media mentions

- Classifier should assign a label to every token
  - `O`, `source-B`, `source-I`, `receiver-B`, `receiver-I`, `perception-B`, or `perception-I`
- Word vector model was trained on the data, using `fastText`
- 100 dimensional vector were used as 100 features for the classifier
- Model was trained using Conditional Random Fields (CRF) in `sklearn-crfsuite`
  - Able to deal with sequential data implicitly
  - Useful when working with sequence labeling tasks
  - Calculates features during training
- Information about the previous and next word as features (onsets, offsets, digits, vectors)
- Train set and test set were manually split (70% training and 30% testing of every volume)

Machine learning media mentions

	precision	recall	f1-score	support
perception-B	0.643	0.552	0.594	134
receiver-B	0.828	0.578	0.681	83
source-B	0.513	0.488	0.500	121
source-I	0.670	0.448	0.537	145
perception-I	0.500	0.143	0.222	7
receiver-I	0.125	0.111	0.118	9

Table: F-scores per label.

Preliminary results

label n	label n+1	weight
O	O	4.163518
receiver-I	receiver-I	4.029864
receiver-B	receiver-I	3.232955
source-I	source-I	2.888268
perception-I	perception-I	2.374591
source-B	source-I	1.727776

Table: Top likely transitions.

weight	label	feature
1.988054	source-I	+1:word.isdigit()
1.744118	perception-B	word[-5:]: lesen
1.472060	perception-B	word[-5:]: ndigt
1.466354	O	bias
1.145304	source-B	-1:word[-5:]: nde

Table: Top positive state features.

Preliminary results

Three main types of errors

- The token with which the label starts: which token gets the suffix `-B`?
  - **de** Gendsche Gazette
  - **alderverschrikkelijkste** berichten
- Lonely tokens: model is better in predicting when there is a sequence of labeled tokens
  - decreet
  - gerugt
- Consistency of the model
  - The word *men* often appears as `receiver`, but not always
  - Model labels media mentions that are overlooked by the annotators
  - Annotator confused `source` with `receiver`, but model assigned the correct labels

Outline

1. Chronicles and the early modern mediascape
2. Corpus and annotation task
3. Machine learning media mentions
4. Discussion

Discussion

- **Removal of punctuation**: our data consisted of one long sentence
  - Problematic when trying to optimize the model
  - Early modern chroniclers were very inconsistent in using punctuation
- **Lowering all characters**: mentions of oral sources or newspaper could not be recognized by its capital
  - Early modern chroniclers were very inconsistent in using capitals
- **Cluster word vectors**
- **Add a lexicon**: using a lexicon with frequently used words might improve the labeling of lonely tokens
- **Goal**: train a model that can be used to label media mentions in the whole corpus, and that can be used as a method to facilitate the close reading

Thanks for your attention!