Development and annotation of a newspaper corpus as part of a doctoral thesis on text structure and cohesion in news items from the 17th and 18th centuries

Katrin Goldschmidt, Universität Bonn

10 March 2017

Content

1. Historical Newspapers
   - Characteristics of historical newspapers
   - Objectives of doctoral thesis
2. A Historical Newspaper Corpus
   - Corpus Development
   - ANNIS
3. Segmentation
   - Typographical Segmentation
   - Functional Segmentation

Historical Newspapers in the 17th and 18th century

Textual structure

- collections of correspondences published by an editor
- single correspondences inform about different news stories
- 3 levels: issue > correspondence > news item(s)

But news stories were written as continuous text...

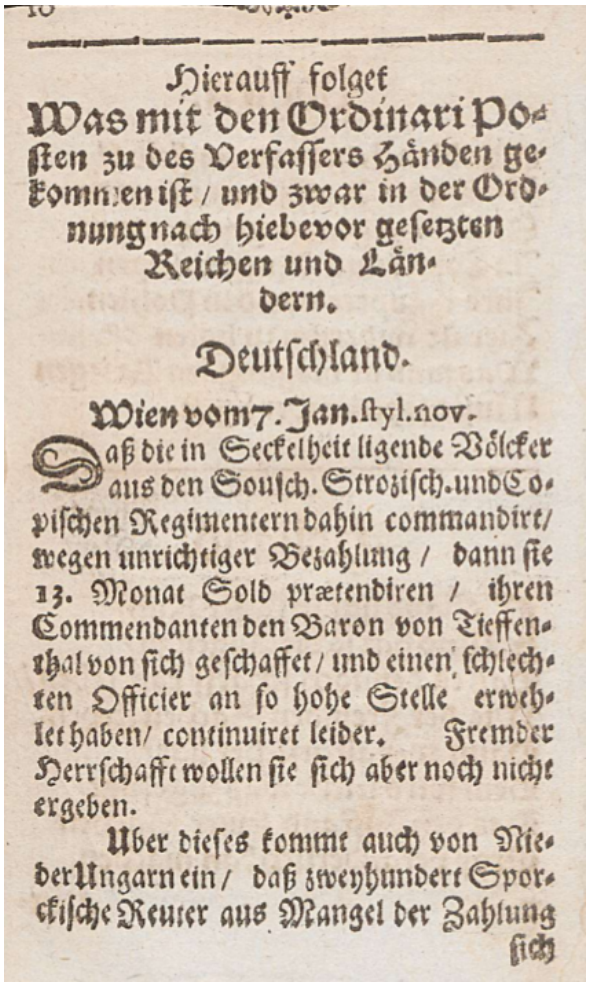
Historical Newspapers in the 17th and 18th century

Textual structure

- collections of correspondences published by an editor
- single correspondences inform about different news stories
- 3 levels: issue > correspondence > news item(s)

But news stories were written as continuous text...

Historical news items in the 17th and 18th centuries



Hierauff folget
Was mit den Ordinari Po-
sten zu des Verfassers Händen ge-
kommen ist / und zwar in der Ord-
nung nach hiebevor gesetzten
Reichen und Län-
dern.

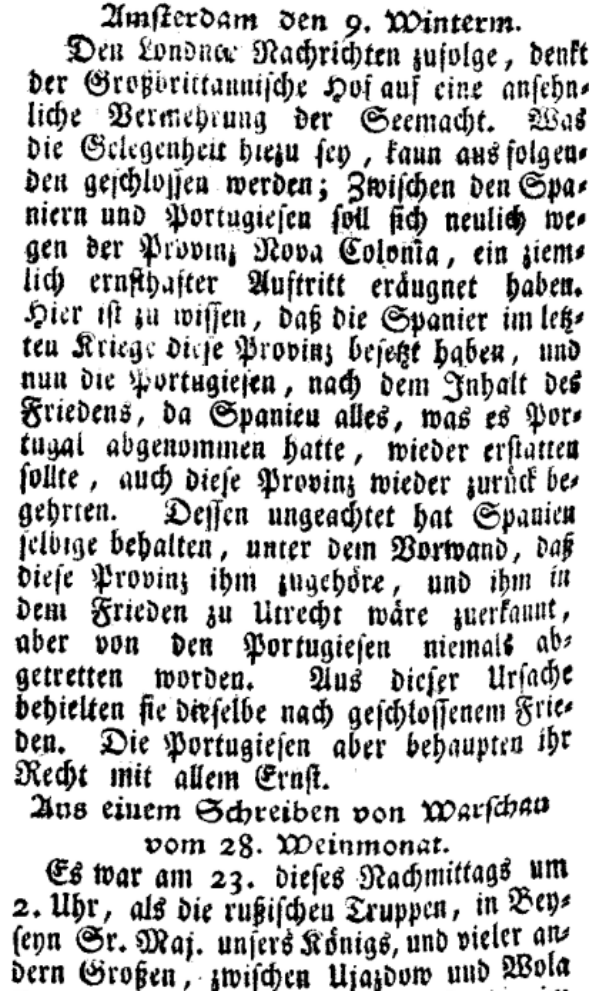
Deutschland.

Wien vom 7. Jan. styl. nov.

Daß die in Seckelheit ligende Völcker aus den Sousch. Strozisch. und Co-
pischen Regimentern dahin commandirt/ wegen unrichtiger Bezahlung / dann sie 13. Monat Sold prætendiren / ihren Commendanten den Baron von Tieffen-
thal von sich geschaffet / und einen schlech-
ten Officier an so hohe Stelle erweh-
let haben/ continuiret leider. Fremder Herrschafft wollen sie sich aber noch nicht ergeben.

Uber dieses komme auch von Nie-
der Ungarn ein / daß zweyhundert Spor-
ckische Reuter aus Mangel der Zahlung sich

Nordischer Mercurius 1664



Amsterdam den 9. Winterm.

Den Londner Nachrichten zufolge, denkt der Großbrittannische Hof auf eine ansehn-
liche Vermehrung der Seemacht. Was die Gelegenheit hiezu sey, kann aus folgen-
den geschlossen werden; Zwischen den Spa-
niern und Portugiesen soll sich neulich we-
gen der Provinz Nova Colonia, ein ziem-
lich ernsthafter Auftritt eräugnet haben. Hier ist zu wissen, daß die Spanier im letz-
ten Kriege diese Provinz besetzt haben, und nun die Portugiesen, nach dem Inhalt des Friedens, da Spanien alles, was es Por-
tugal abgenommen hatte, wieder erstatten sollte, auch diese Provinz wieder zurück be-
gehrten. Dessen ungeachtet hat Spanien selbige behalten, unter dem Vorwand, daß diese Provinz ihm zugehöre, und ihm in dem Frieden zu Utrecht wäre zuerkannt, aber von den Portugiesen niemals ab-
getretten worden. Aus dieser Ursache behielten sie dieselbe nach geschlossenem Frie-
den. Die Portugiesen aber behaupten ihr Recht mit allem Ernst.

Aus einem Schreiben von Warschau vom 28. Weinmonat.

Es war am 23. dieses Nachmittags um 2. Uhr, als die rußischen Truppen, in Bey-
seyn Sr. Maj. unsers Königs, und vieler an-
dern Großen, zwischen Ujazdow und Wola

Wienerisches Diarium 1767

Objectives

Textual Structure

- typographical segmentation of text items
- super- and macrostructural elements as title, headlines, comments, citations

News Structure

- publicistic features as persons, locations, time references
- event-related vs. source-related entities

Coherence Structure

- entities as phrases vs. parts of phrases
- direct and indirect relations within and between news items

Objectives

Textual Structure

- typographical segmentation of text items
- super- and macrostructural elements as title, headlines, comments, citations

News Structure

- publicistic features as persons, locations, time references
- event-related vs. source-related entities

Coherence Structure

- entities as phrases vs. parts of phrases
- direct and indirect relations within and between news items

Objectives

Textual Structure

- typographical segmentation of text items
- super- and macrostructural elements as title, headlines, comments, citations

News Structure

- publicistic features as persons, locations, time references
- event-related vs. source-related entities

Coherence Structure

- entities as phrases vs. parts of phrases
- direct and indirect relations within and between news items

Content

From transcription to annotation

Texts

- 7 german newspaper issues (1609-1767)
- from Berlin, Hamburg (2x), Salzburg, Straßburg, Vienna (2x)
- segmentation of correspondences into news items by 3 raters
  => main corpus
- additional corpus: all January issues of Nordischer Mercurius 1667 with syntactical annotation (Mercurius corpus)

Corpus size

- token: ca. 30.000 [main corpus: 17.784]
- sentences: ca. 1.700 [main corpus: 670]
- news items: ca. 460 [main corpus: 245]
- entities: ca. 5700 [main corpus: 3.384]

From transcription to annotation

Texts

- 7 german newspaper issues (1609-1767)
- from Berlin, Hamburg (2x), Salzburg, Straßburg, Vienna (2x)
- segmentation of correspondences into news items by 3 raters => main corpus
- additional corpus: all January issues of Nordischer Mercurius 1667 with syntactical annotation (Mercurius corpus)

Corpus size

- token: ca. 30.000 [main corpus: 17.784]
- sentences: ca. 1.700 [main corpus: 670]
- news items: ca. 460 [main corpus: 245]
- entities: ca. 5700 [main corpus: 3.384]

From transcription to annotation

Texts

- 7 german newspaper issues (1609-1767)
- from Berlin, Hamburg (2x), Salzburg, Straßburg, Vienna (2x)
- segmentation of correspondences into news items by 3 raters => main corpus
- additional corpus: all January issues of Nordischer Mercurius 1667 with syntactical annotation (Mercurius corpus)

Corpus size

- token: ca. 30.000 [main corpus: 17.784]
- sentences: ca. 1.700 [main corpus: 670]
- news items: ca. 460 [main corpus: 245]
- entities: ca. 5700 [main corpus: 3.384]

Facsimile and Transcription

Ao. 1741. No. XXXIV.

Dienstag den 21. Merz.

Berlinische Nachrichten von Staats- und gelehrten Sachen.

Berlin, vom 21. Merz.

In verwichenen Sonntage brach das Prinz Wilhelmische Regiment von hier nach Schlesien auf. Seine Königl. Hoheit führten dasselbe in hoher Person auf und ein jeder bewunderte die Vollkommenheit dieses Regiments.

In wenig Tagen werden Se. Königl. Hoheit selbst nachfolgen.

London, vom 9. Merz.

Am 4. dieses Monats erhielt der Hof Briefe von dem Ritter Ogle, welche am 4. Jenner geschrieben waren, mit der Nachricht, daß sich der Ritter am 30. December des vorigen Jahres mit seiner Flotte zu St. Domingo vor Anker gelegt hätte. Die Kriegsschiffe Rippon und York wären zu ihm gestossen, allein am 12. November 66. Meilen vom Cap Lezard durch einen Sturm wieder von ihm getrennet worden. Er füget noch hinzu, daß die Schiffe auf dem Fall einer Trennung, sich auf der Insel zu St. Christoph wieder zu finden abgeredet hätten, weswegen er sie auch daselbst anzutreffen hoffete. Diesen Briefen zu Folge hat der Ritter am 5. dieses Monats seinen Weg nach Jamaika fortgesetzet. Der Lord Catheard, commandirender General unserer Truppen in Westindien, ist am 19. Decembr., von einem heftigen Durchfalle angegriffen worden, und am 31. als am andern Tage nach der Ankunft des Ritters Ogle zu St. Domingo, verstorben. So schmerzhaft als seine Krankheit gewesen, so bezeigte er sich als ein wahrhafter Engländer und starb eben so großmüthig als er gelebt hatte. Alles was man von Misvergnügen bey ihm fand, rührte nur daher, daß er ohne seinen Landesleuten die

Wir.

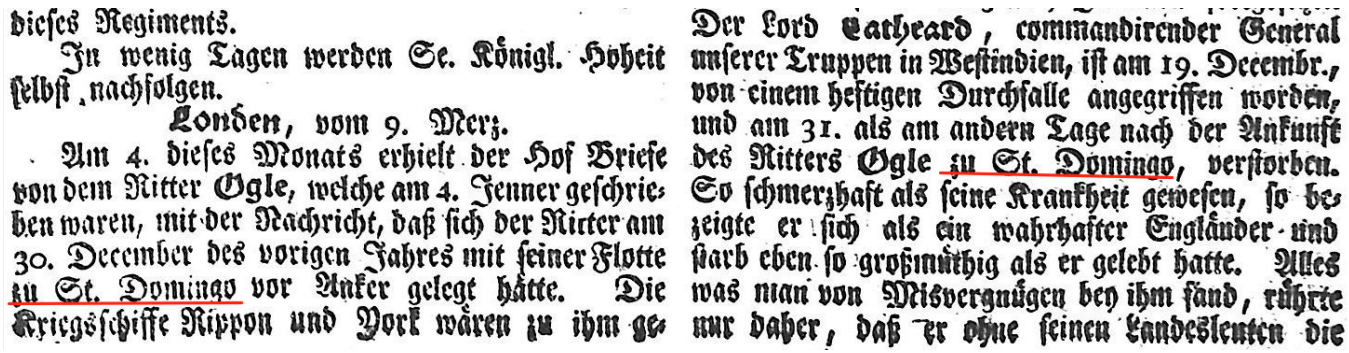
```
% Titel: Berlinische Nachrichten von Staats= und gelehrten Sachen
% Erscheinungsort: Berlin
% Herausgeber: Haude & Spener
% Ausgabe: Di, 34
% Gedruckt: 21-03-1741
% Transkription: IDS-Zeitungskorpus, zur Verf. gestellt durch Inst
% Satzsegmentierung und Preprocessing: Katrin Goldschmidt, Univers
% letzteKorrektur: 16-05-2014
% nicht aufgenommene Originalseiten bzw Textstellen: +K Rubriken '


_Seite1

Ao. 1741. +K Titelkupfer "WAHRHEIT.VND.FREIHEIT." @K No. XXXIV .
Dienstag +K Titelkupfer "WAHRHEIT.VND.FREIHEIT." @K den 21. Merz
+K danach durchgehende Linie @K
&spbr& Berlinische Nachrichten &spbr&
&spbr& von &spbr&
&spbr& Staats= und gelehrten Sachen . &spbr&
+K danach durchgehende Linie @K
+K Beginn linke Spalte @K
&spbr& Berlin , vom 21. Merz . &spbr&
&spbr& +K Schmuckinitiale Adler hält Majuskel A @K *{A*6}m verwicl
Prinz Wilhelmische Regiment von
hier nach Schlesien auf . &spbr& Seine
Königl. Hoheit führten dasselbe in
hoher Person auf und ein jeder
bew<u>nderte die Vollkommenheit
dieses Regiments . &spbr&
&spbr& Jn wenig Tagen werden Se. Königl. Hoheit
selbst nachfolgen . &spbr&
&spbr& Londen , vom 9. Merz . &spbr&
&spbr& Am 4. dieses Monats erhielt der Hof Briefe
von dem Ritter Ogle , welche am 4. Jenner geschrie=
ben waren , mit der Nachricht , daß sich der Ritter am
30. December des vorigen Jahres mit seiner Flotte
zu St. Domingo vor Anker gelegt hätte . &spbr& Die
Kriegsschiffe Rippon und York wären zu ihm ge=
+K Beginn rechte Spalte @K
stossen , allein am 12. November 66. Meilen vom
Cap Lezard durch einen Sturm wieder von ihm ge=
trennet worden . Er füget noch hinzu , daß die Schiffe
auf dem Fall einer Trennung , sich auf der Jnsel zu
St. Christoph wieder zu finden abgeredet hätten ,
weswegen er sie auch daselbst anzutreffen hoffete .
Diesen Briefen zu Folge hat der Ritter am 5. die=
```

Berlinische Nachrichten 1741

Annotation



dieſes Regiments.

In wenig Tagen werden Se. Königl. Hoheit ſelbſt nachfolgen.

London, vom 9. Merz.

Am 4. dieſes Monats erhielt der Hof Briefe von dem Ritter Ogle, welche am 4. Jenner geſchrieben waren, mit der Nachricht, daß ſich der Ritter am 30. December des vorigen Jahres mit ſeiner Flotte zu St. Domingo vor Anker gelegt hätte. Die Kriegsſchiffe Rippon und York wären zu ihm ge-

Der Lord Catheard, commandirender General unſerer Truppen in Weſtindien, iſt am 19. Decembr., von einem heftigen Durchfalle angegriffen worden, und am 31. als am andern Tage nach der Ankunft des Ritters Ogle zu St. Domingo, verſtorben. So ſchmerzhaft als ſeine Krankheit geweſen, ſo bezeigte er ſich als ein wahrhafter Engländer und ſtarb eben ſo großmüthig als er gelebt hatte. Alles was man von Misvergnügen bey ihm fand, rührte nur daher, daß er ohne ſeinen Landesleuten die

tok	tok_ID	pos	text	entity_p	coref_target_p	coref_entity_p	indref_target	indref_entity	indref_reftype	typosegm	is
seiner	107	PPOSAT					sour_Inf	80	poss		
Flotte	108	NN									
zu	109	APPR		loc_H							
St.	110	NE									
Domingo	111	NE									
vor	112	APPR									
Anker	113	NN	bn_b								
.(.)	196	$.	bn_e								
Der	197	ART	bn_b	who_H							
Lord	198	NN									
Catheard	199	NE									
ist	208	VAFIN									
zu	233	APPR		loc_H	loc_H_bn	109					
St.	234	NE									
Domingo	235	NE									
,(,)	236	$,									

Annotation levels

Textual Structure

- **token**
- pos
- **text**
- makrostr
- **typosegm**
- issue
- correspondence
- origin
- meta

News Structure

- **entity_p (phrase)**
- entity_pp (part of phrase)
- how_why
- episode

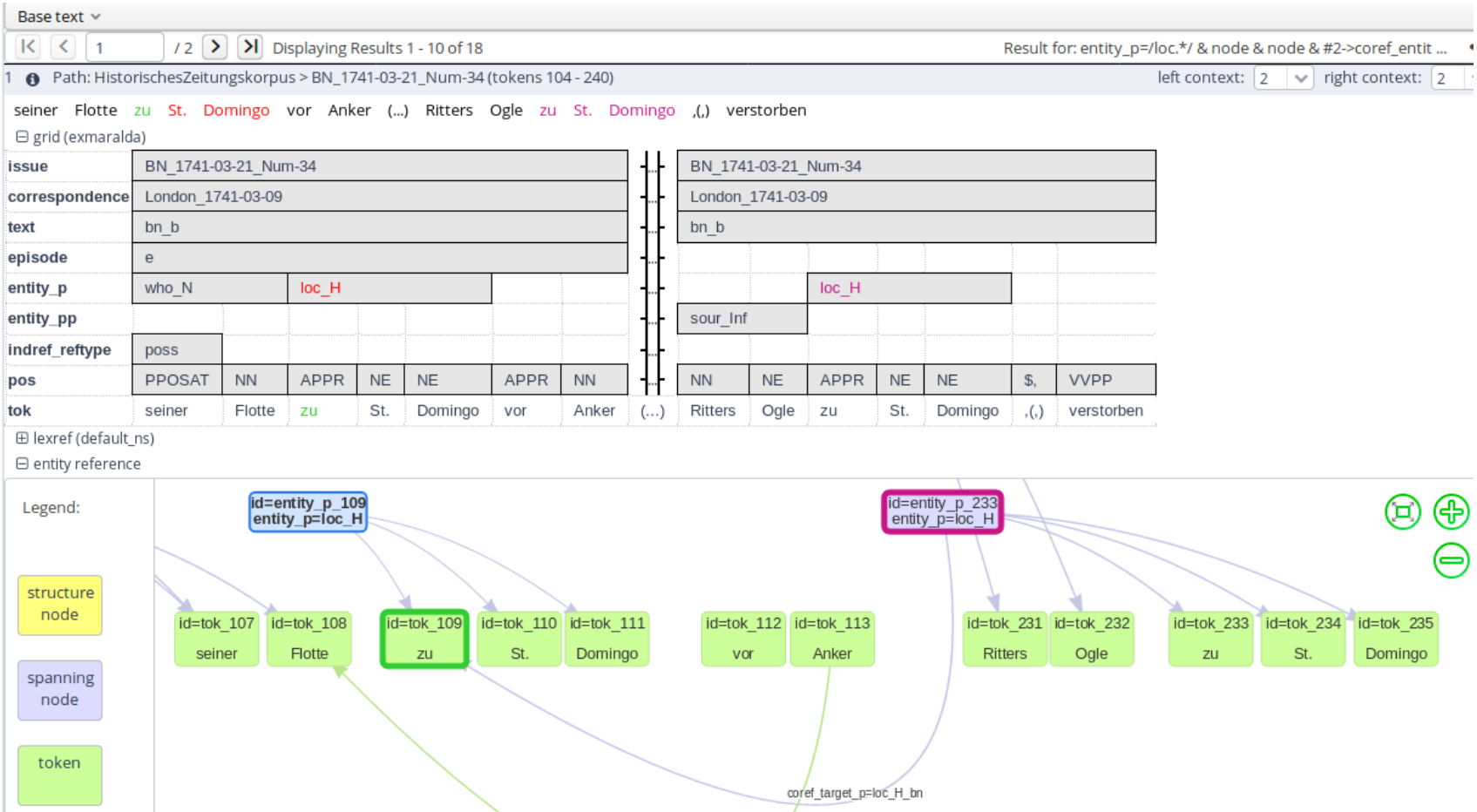
- indref_reftype (func, part, poss)
- time_reftype (before, during, after)

Coherence Relations

- **coref_target_p**
- coref_target_pp
- indref_target
- lexref_target
- ref_episode_target
- ref_howwhy_target

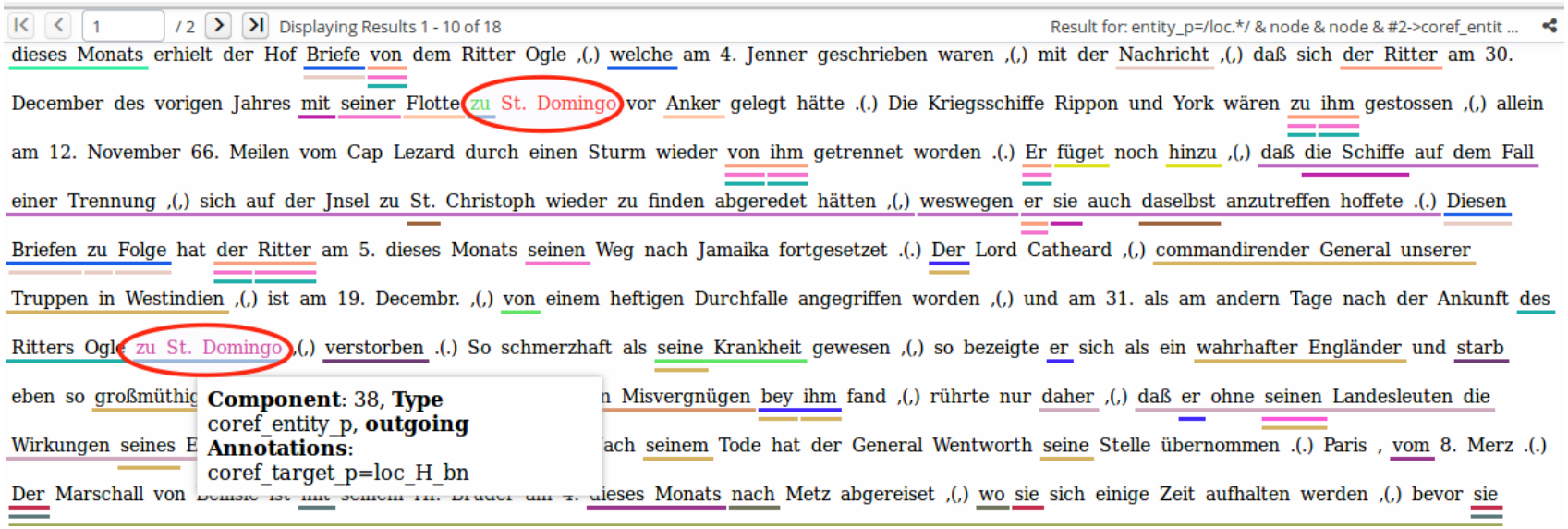
HistZeitKorp in local ANNIS

entity reference view



HistZeitKorp in local ANNIS

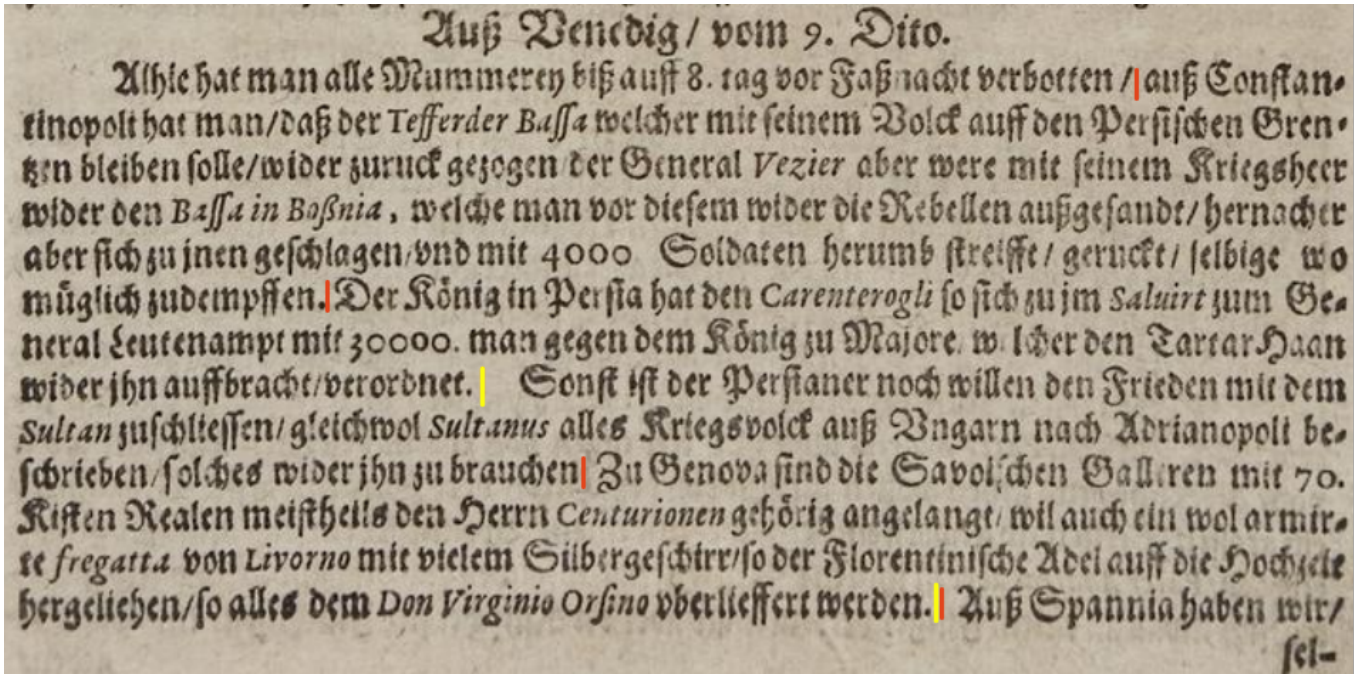
discourse view



Content

Segmentation and the role of typographical means

- following Lefèvre (2013) shifts between paragraphes are marked by the representation type [.+space(+UpperCase)]
- but: no quantitative analysis

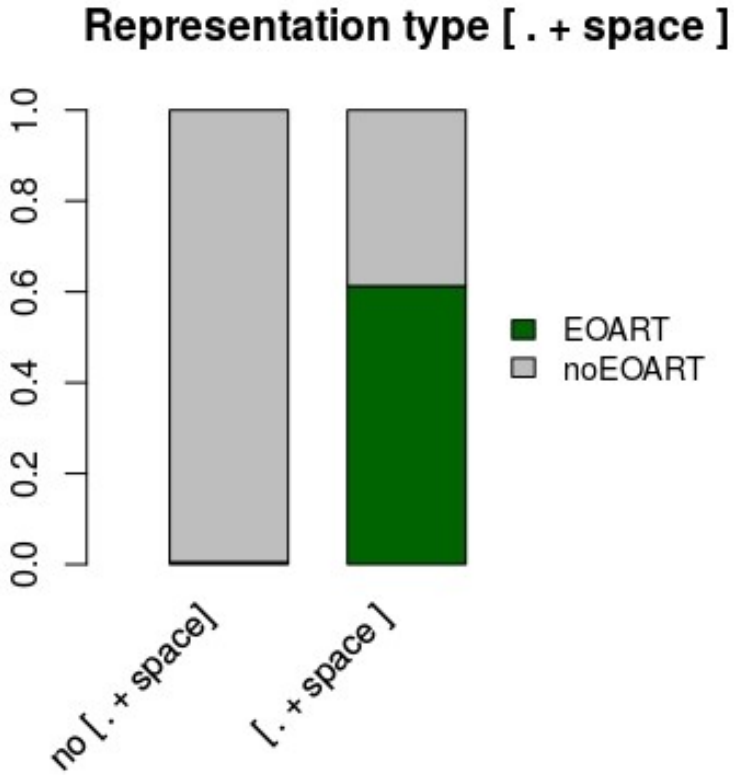


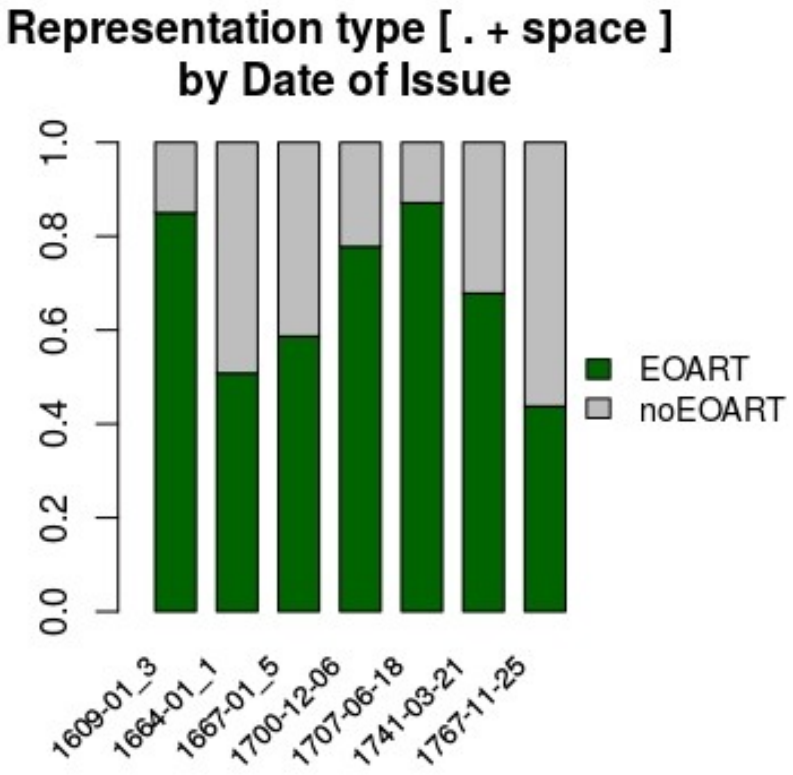
Auß Venedig / vom 9. Dito.

Alhie hat man alle Mummerey biß auff 8. tag vor Faßnacht verbotten / auß Conſtantinopolt hat man / daß der *Tefferder Baſſa* welcher mit ſeinem Volck auff den Perſiſchen Grentzen bleiben ſolle / wider zuruck gezogen der General *Vezier* aber were mit ſeinem Kriegsheer wider den *Baſſa in Boßnia*, welche man vor dieſem wider die Rebellen außgeſandt / hernacher aber ſich zu jnen geſchlagen / vnd mit 4000 Soldaten herumb ſtreiffe / gerucket / ſelbige wo müglich zudempffen. Der König in Perſia hat den *Carenterogli* ſo ſich zu jm *Saluirt* zum General Leutenampt mit 30000. man gegen dem König zu Majore / welcher den Tartar Haan wider jhn auffbracht / verordnet. Sonſt iſt der Perſianer noch willen den Frieden mit dem *Sultan* zuſchlieſſen / gleichwol *Sultanus* alles Kriegsvolck auß Vngarn nach Adrianopoli beſchrieben / ſolches wider jhn zu brauchen. Zu Genova ſind die Savoiſchen Gallren mit 70. Kiſten Realen meiſtheils den Herrn *Centurionen* gehörig angelangt / wil auch ein wol armirte *fregatta* von *Livorno* mit vielem Silbergeſchirr / ſo der Florentiniſche Adel auff die Hochzeit hergeliehen / ſo alles dem *Don Virginio Orſino* vberlieffert werden. Auß Spannia haben wir / ſel-

Relation 1609 [boundaries: Lefèvre (2013); HistNewsKorp]

Typographical segmentation is not that easy...





Segmentation and the role of functional means

- following Schröder (1995) shifts between news items are marked by change of theme, person, location and time [Relation 1609, Aviso 1609]
- over 75% of the news items are introduced by entities that respond to journalistic wh-questions
  (like *who, where, when ...?*)

A first approach: "How many news items begin with entities?"

- ANNIS search for entities as phrases at the direct beginning of a news item: text=/bn_b/ & entity_p & #2 _l_ #1
- 167 matches

Segmentation and the role of functional means

- following Schröder (1995) shifts between news items are marked by change of theme, person, location and time [Relation 1609, Aviso 1609]
- over 75% of the news items are introduced by entities that respond to journalistic wh-questions
  (like *who, where, when ...?*)

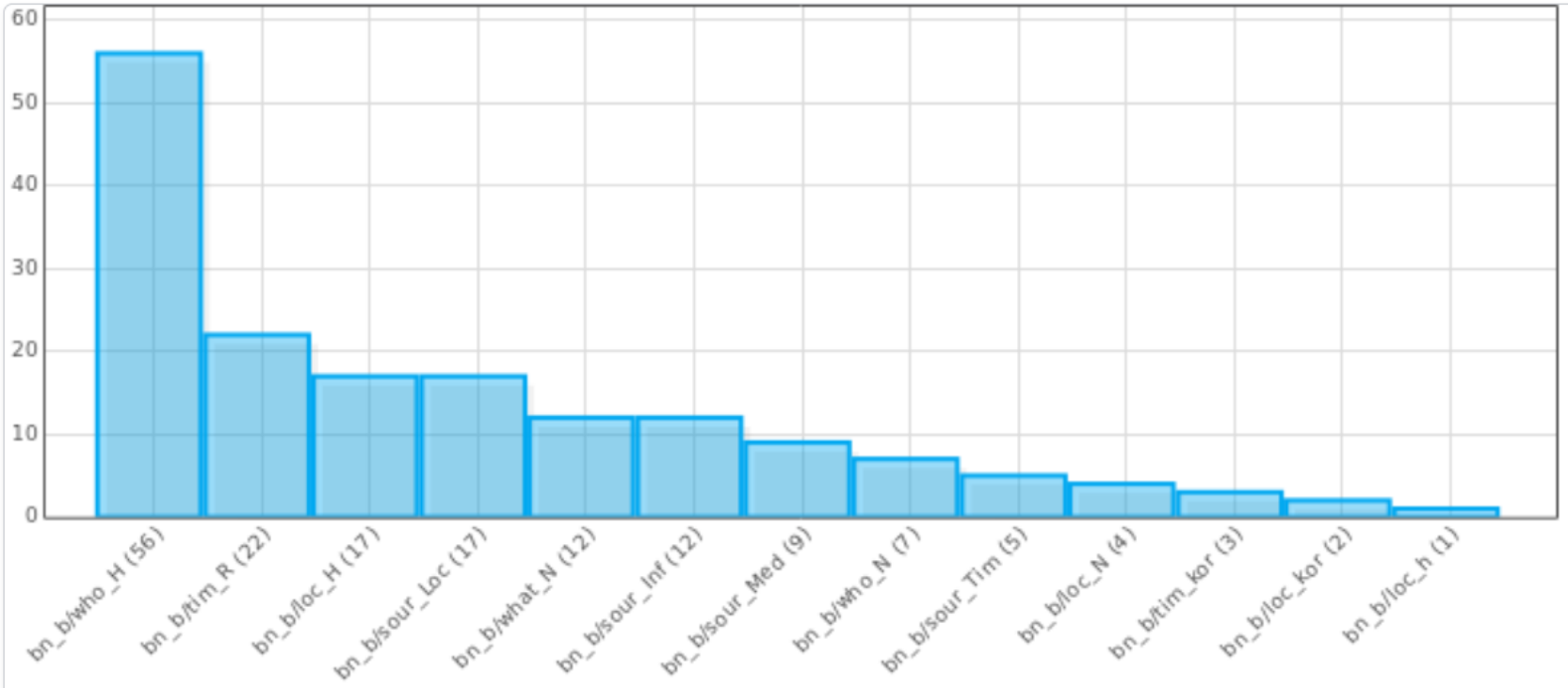
A first approach: "How many news items begin with entities?"

- ANNIS search for entities as phrases at the direct beginning of a news item: text=/bn_b/ & entity_p & #2 _l_ #1
- 167 matches

Help/Examples | Query Result × | Frequency Analysis ×

New Analysis

linear scale
$\log_{10}$ scale



Also functional segmentation is not that easy...

- we see that 64% of the news items begin with an entity that responds to the forecited journalistic wh-questions:

	Schröder1995: Relation 1609	HistZeitKorp	
PubInfs	**percentage**	**percentage**	**N**
per_loc_tim_them	75%	64%	158
other	8%	4%	9
none	17%	32%	78
total	100%	100%	**245**

A second approach: "How many news items begin with entities that refer to entities within former news items?"

- ANNIS search: 11 matches ($11 \div 158 \approx 7.0\%$)

Also functional segmentation is not that easy...

- we see that 64% of the news items begin with an entity that responds to the forecited journalistic wh-questions:

	Schröder1995: Relation 1609	HistZeitKorp	
PubInfs	**percentage**	**percentage**	**N**
per_loc_tim_them	75%	64%	158
other	8%	4%	9
none	17%	32%	78
total	100%	100%	**245**

A second approach: "How many news items begin with entities that refer to entities within former news items?"

- ANNIS search: 11 matches ($11 \div 158 \approx 7.0\%$)

Conclusions

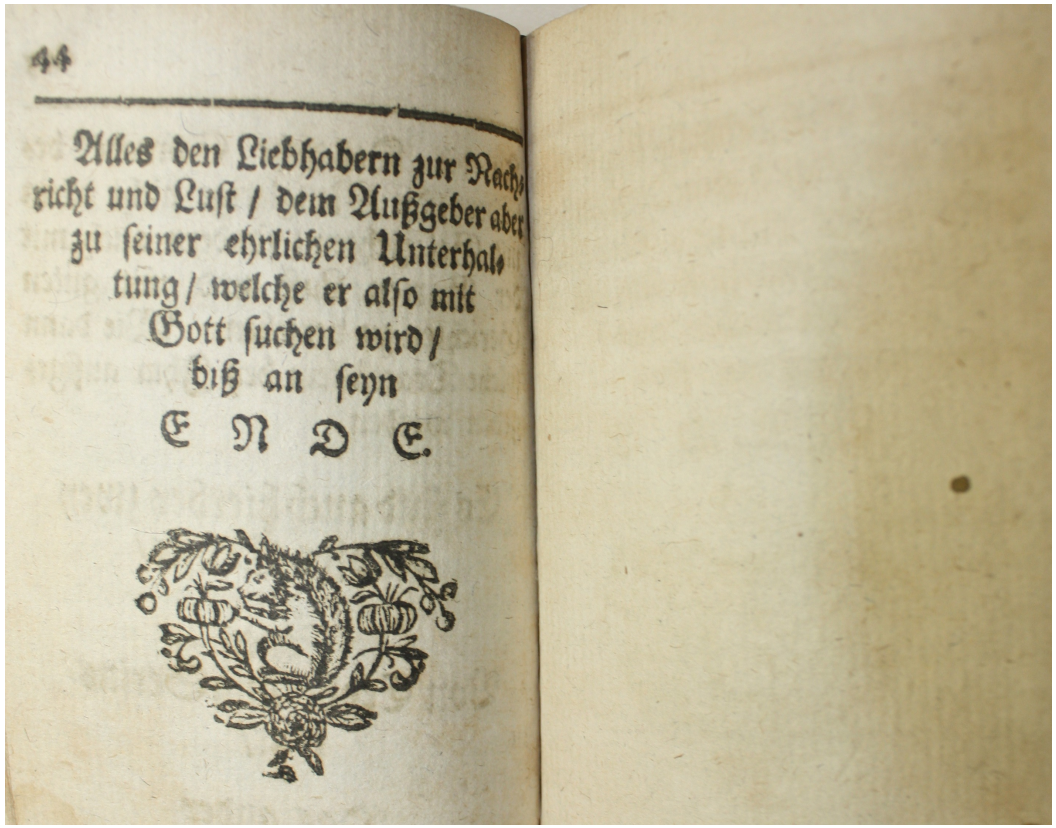
- typographical means alone are not an adequate means of news item segmentation (maybe in 60% of the correspondences they are helpful)
- at least 7% of phrasal entities at the beginning of news items refer to entities in the preceding news item(s)
  $=>$ implies no thematic change
- but: considering typographical AND functional means could lead to successive segmentation methods

With help of the Historical Newspaper Corpus we can get a better understanding of the textual and publicistic structure of these interesting old newspapers.

Conclusions

- typographical means alone are not an adequate means of news item segmentation (maybe in 60% of the correspondences they are helpful)
- at least 7% of phrasal entities at the beginning of news items refer to entities in the preceding news item(s)
  $=>$ implies no thematic change
- but: considering typographical AND functional means could lead to successive segmentation methods

With help of the Historical Newspaper Corpus we can get a better understanding of the textual and publicistic structure of these interesting old newspapers.


44

Alles den Liebhabern zur Nach-
richt und Luſt / dem Außgeber aber
zu ſeiner ehrlichen Unterhal-
tung / welche er alſo mit
Gott ſuchen wird /
biß an ſeyn
E N D E.

Nordischer Mercurius 1664

„Everything [serves] to the [news] lovers [as] a message and a pleasure / but to the editor [it serves as] his honest entertainment, which he therefore will seek with God / until his END.“

References

- Fritz, G. & E. Straßner (eds.) (1996): Die Sprache der ersten deutschen Wochenzeitungen im 17. Jahrhundert. Tübingen.
- Demske, U. (2007): Das Mercurius-Projekt: eine Baumbank für das Frühneuhochdeutsche. In: G. Zifonun & W. Kallmeyer (eds.): Sprachkorpora - Datenmengen und Erkenntnisfortschritt (= Jahrbuch des Instituts für Deutsche Sprache 2006). Berlin, 91-104.
- Krause, T. & A. Zeldes (2014): ANNIS3: A New Architecture for Generic Corpus Query and Visualization. Literary and Linguistic Computing.
- Lefèvre, M. (2013): Textgestaltung, Äußerungsstruktur und Syntax in deutschen Zeitungen des 17. Jahrhunderts. Zwischen barocker Polyphonie und solistischem Journalismus. Berlin.
- Schröder, T. (1995): Die ersten Zeitungen. Textgestaltung und Nachrichtenauswahl. Tübingen.

Special thanks to all who supported me with the newspaper records (Österreiche Nationalbibliothek, Staatsbibliothek zu Berlin, Staats- und Universitätsbibliothek Bremen, Institut für Deutsche Sprache Mannheim) and Prof. Ulrike Demske for providing the Mercurius Corpus.

Additional corpus :: NM_1667-01_synt

- all January issues of Nordischer Mercurius 1667
- integrated syntactical annotation from the Mercurius corpus