COST Action Distant Reading for European Literary History

Corpus Design Principles and Challenges

Carolin Odebrecht (Humboldt-Universität zu Berlin) together with Christof Schöch, Lou Burnard, Borja Navarro-Colorado et al.

PARTHENOS Workshop for CEE countries, 2019-10-08

Outline

Schedule

9:00-9:15	Introduction COST and ELTeC
9:15-9:30	Introduction Romanian novels / literary contexts
9:30-09:55	Corpus design
09:55-10:30	Romanian language collection
10:30-11:00	Break
11:00-12:00	Introduction to TEI XML and ELTeC schema
12:00-13:00	Transcribus demonstration

Schedule

Goals of our sessions

- Present our research approach in Digital Humanities
- Concepts on corpus design and annotation model
- Language-specific contexts on text selection and balancing
- First steps encoding TEI XML
- First steps text digitization with Transcribus

→ Focus on data design and creation

→ Looking forward to discussing each part in the break out sessions!

Introduction

Corpus linguistics
Every linguistic analysis is an interpretation of the data.
(Lüdeling 2011)

Digital literary studies
Two scholars can read the same dataset - like the same literary work - and derive different meanings.
(Bode 2018)

Outline

- COST Actions are research networks[1]
  - for any scientific field,
  - for workshops, conferences, working group meetings, training, schools, short-term scientific missions, and dissemination and communication activities,
  - for fostering Inclusiveness Target Countries (ITC).
- Each country member has a national supporting institution

[1] www.cost.eu

Distant Reading

- Christof Schöch, University of Trier
- CA16204 will
  - "create a vibrant and diverse network of researchers jointly developing the resources and methods necessary to change the way European literary history is written"
  - "contribute to the development and distribution of methods, competencies, data, best practices, standards and tools relevant to Distant Reading research"[2]
- Working groups
  - WG 1: Scholarly Resources
  - WG 2: Methods and tools
  - WG 3: Literary Theory and History
  - WG 4: Dissemination

[2] www.distant-reading.net

COST Action Distant Reading

[3]

[3] https://www.distant-reading.net/about/network/

WG1 Scholarly Resources

- Creating an open source multi-lingual benchmark corpus for European literature: European Literary Text Collection (ELTeC)[4]
- (currently) 34 Members of 22 countries
- Main tasks are
  - defining corpus design,
  - developing basic encoding schemas,
  - developing workflows.

[4] https://www.distant-reading.net/wg-1/

ELTeC

- Digitized and annotated European novels of the 19th century
- Uniform sampling and balancing criteria
- Uniform and consistent encoding schemas in TEI XML
  - Basic encoding to facilitate distant reading
  - Applicable for different languages
  - Currently working on English, German, French, Spanish, Italian, Romanian, Slovenian, Polish, Hungarian, Portuguese, Serbian, Greek, Norwegian, Czech

Outline

Corpus

Historical corpora (cf. for example Claridge 2008; Kytö 2011)

- Digitized and annotated (encoded) historical texts
- Resources with a complex publication history and often conflicting texts definitions (cf. for example Caton 2013; De Rose et al. 2002; van Zundert and Andrews 2017)
- Divers methods and approaches towards corpus creation in relation to corpus architecture, annotations etc.
- Widely-used complex subtype of text corpora in (digital) humanities

Corpus design

Corpus design defines two things (cf. a.o. Hunston 2008; Lüdeling et al. 2016):

- Candidates: Which text(s) can be included in the corpus? Which don't?
- Proportion: How many texts with which characteristics should the corpus contain?

Corpus design – Action's purpose

- Benchmark corpus for distant reading
- Methods for data creation and analysis, e.g.
  - Part-of-speech tagging
  - Lemmatization
  - Morphological information
  - Authorship attribution
  - Network analysis
  - Topic modelling
  - Sentiment analysis

Corpus design – challenges

- Different publication histories in Europe
- Different literary scholars and traditions
- Accessibility of information and resources

Is it possible to define criteria for selecting novels from all over Europe?

Corpus design – Action's approach

- Sampling and balancing criteria[5] will
  - not define what a novel is,
  - follow a non-normative but metadata-based approach (not canon-based),
  - aim to represent the variety of a population[6],
  - allow for a comparability of texts and individual sub-collections according to different metadata set(s).

[5] `https://distantreading.github.io/sampling_proposal.html`

[6] Cf. for discussion of representativeness Biber (1993) and canonicity (Herrmann 2011) and corpus design Algee-Hewitt and McGurl (2018), Bode (2018), Hunston (2008), and Lüdeling et al. (2016).

ELTeC – sampling criteria

- language: European languages, no translations
- prose: narrative fictional prose
- period: 1840-1920
- length: min. 10.000 words
- publication: prefer books over novels published in serial publications
- access: only freely available digitizations

ELTeC – balancing criteria

- 100 texts per language (language collection)
- period: distribution over time
  - T1: 1840-1859
  - T2: 1860-1879
  - T3: 1880-1899
  - T4: 1900-1920
- gender: min. 10% and max. 50% have been written by female authors for the language subcollection
- authorship: 9 - 11 authors with exact three novels
- length: min. 20% are short novels (10-50k word tokens), min. 20% are long novels (>200k word tokens).
- reprint: min. 30% are highly canonized novels, min. 30% should be non-canonized novels, based reprint counts within the period 1970-2009

ELTeC – current state

Overview on ELTeC Language Collections:
`https://distantreading.github.io/ELTeC/index.html`

Research data management for ELTeC

- "Research data management is an explicit process covering the creation and stewardship of research materials to enable their use for as long as they retain value." Whyte, A. and Rans, J., Glossary of Digital Curation Center[7]
- (meta) data should findable, accessible, interoperable and reusable (Wilkinson et al. 2016)

[7] http://www.dcc.ac.uk/digital-curation/glossary#R

Research data management for ELTeC

- Data creation and update on GitHub[8]
- Encoding schema developed and documented with TEI ODD[9]
- Data and workflow documentation on GitHub[10]
- Persistent referencing and archiving on Zenodo[11]
- Free licence to foster re-usability: CC-BY 4.0[12]
- Further dissemination strategies are currently evaluated

[8] https://github.com/COST-ELTeC

[9] ODD https://github.com/distantreading/WG1/ and schema https://github.com/COST-ELTeC/Schemas

[10] https://github.com/distantreading/WG1/wiki

[11] https://zenodo.org/communities/eltec/

[12] https://creativecommons.org/licenses/by/4.0/

Outline

Corpus data

- Different starting points for data creation, e.g.:
  - Exemplar of the book
  - Digitized book
  - Plain text
  - Previously encoded data set
- Metadata describe the digital or/and analogue source(s) of the data set
  - Library catalogues
  - Online databases for texts, ebooks, corpora

Encoding | Annotation

- Annotation: explicit assignment of categories to one or more exponents in a corpus, always interpretation (c.f. a.o. Lüdeling 2011; McEnery and Hardie 2012; Zinsmeister et al. 2008)
- Tag set and guidelines: defining categories (and values) and formulate guidelines on how and when to assign them
- Motivation: Research question/context!

XML

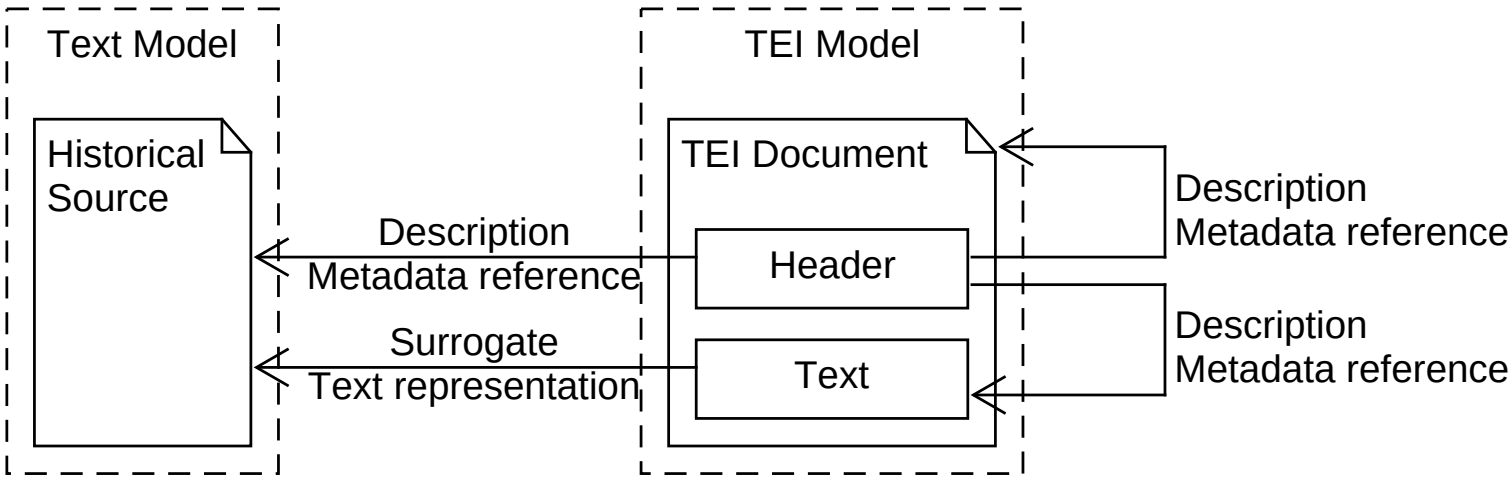
XML Extensible Markup Language

- W3C standard since 1998
- For structuring and organizing information.
- Metalanguage for defining domain-specific XML vocabularies
  $\rightarrow$ TEI XML

Text Encoding Initiative

- Encoding standard and guidelines for the representation of texts for humanities
- TEI consortium (since 1987) for developing and maintaining the standard (TEI-Consortium 2019)
- For various texts, e.g. manuscripts and prints, books, letters, poems, and dictionaries
  - Text-internal categories, text-external categories
  - Mark up, text structure(s) and divisions, content and references
- Guidelines provide ca. 500 elements and various specifying attributes

TEI document



- Consist of `teiHeader` and `text`
- Text contains e.g. `front`, `body`, `trailor`, `back`
- Customization TEI for domain-specific purpose (e.g. select elements and attributes, building subsets, defining new elements)
- Validation mechanisms
- Customization, documentation and validation via ODD (Burnard and Rahtz 2004)

TEI XML

- start tag:
  - `<title>`
- end tag:
  - `</title>`
- single composite tag:
  - `<lb/>`
- attributes
  - `type="..."`

→ Hierarchy of XML elements

Encoding XML

- We start with plain text (e.g. transcribed, OCR)
- We will encode manually[13]
- data for tutorial on `https://github.com/distantreading/WG1/tree/master/Training/2019-10-08-Sofia`

[13] Many approaches on transformation processes, see e.g. Distant Reading Training School Budapest 2019 `https://distantreading.github.io/Training/Budapest/#(2)`

XML – post card example

```
Warm greetings
from Sofia
Carolin
```

XML – example

- Open start.xml
- Encode
  - line
  - place
  - name

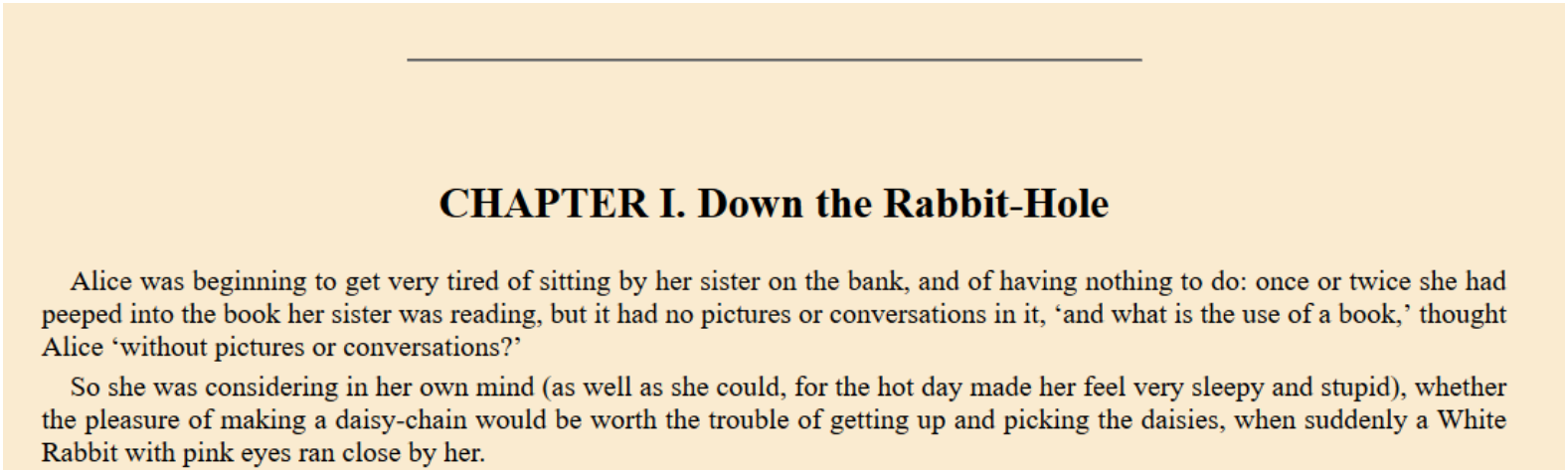
Hands on tutorial – data creation

- How do we encode texts for ELTeC?
- First steps: using TEI XML

→ Today's examples (taken from ELTeC)

  - *Why Paul Ferroll Killed His Wife* by Clive, Caroline, (1801-1873) Saunders, Otley, and Co. London 1860.
  - *Alice's Adventures in Wonderland*, by Lewis Carroll, (1832-1898).London: Macmillan 1865.

Alice



CHAPTER I. Down the Rabbit-Hole

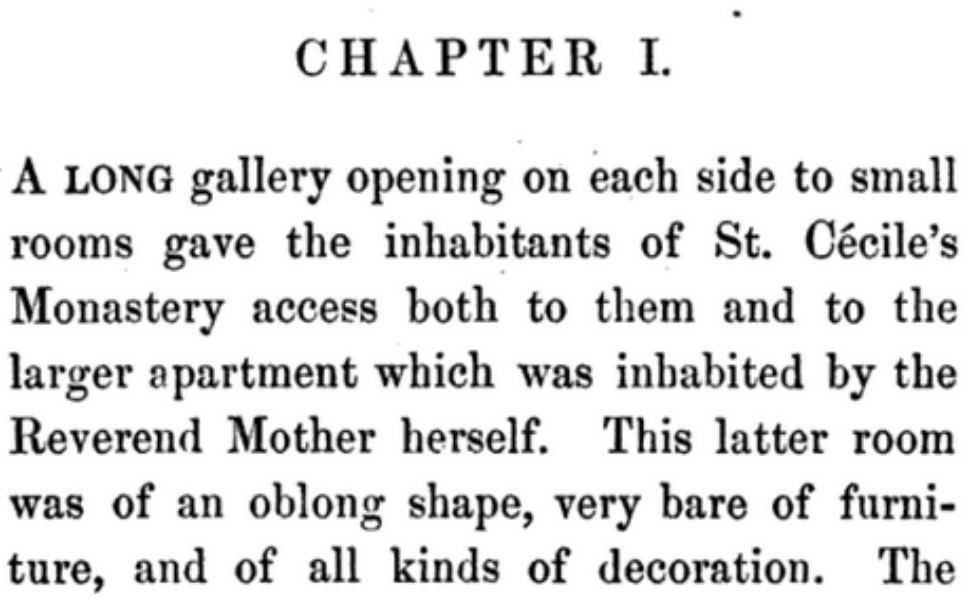
Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, ‘and what is the use of a book,’ thought Alice ‘without pictures or conversations?’

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

Figure: Gutenberg ebook Produced by Arthur DiBianca and David Widger. Gutenberg ebook

```
<div type="chapter">
<head>CHAPTER I. Down the Rabbit-Hole</head>
<p>Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing
to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or
conversations in it, ‘and what is the use of a book,’ thought Alice ‘without pictures or
conversations?’</p>
<p>So she was considering in her own mind (as well as she could, for the hot day made her feel very
sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of
getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by
her.</p>
```

Figure: ELTeC text in English language collection ELTeC version



CHAPTER I.

A LONG gallery opening on each side to small rooms gave the inhabitants of St. Cécile's Monastery access both to them and to the larger apartment which was inhabited by the Reverend Mother herself. This latter room was of an oblong shape, very bare of furniture, and of all kinds of decoration. The

Figure: Google's digitization of the novel. Google book

```
<head>CHAPTER I.</head>
<p><hi>A LONG</hi> gallery opening on each side to small rooms gave the inhabitants
    of St. Cécile's Monastery access both to them and to the larger apartment which
    was inhabited by the Reverend Mother herself. This latter room was of an oblong
    shape, very bare of furniture, and of all kinds of decoration. The windows were
```

Figure: ELTeC text in English language collection ELTeC version

Defining encoding schema

- Brain storming:
  Which text features can to be encoded for analysing European novels?

ELTeC metadata

- `teiHeader`
  - Bibliographic information
  - Balancing information
  - Data processing information

ELTeC encoding

- `text`
  - paragraphs
  - highlighted
  - head
  - division
  - chapter
  - page breaks
  - ...

ELTeC encoding schemas

- Not to represent texts in all their original complexity[14]
- Not aiming for duplicating the work of scholarly editors
- Aim to facilitate a richer and better-informed distant reading than a transcription of lexical content alone would permit
- Encoding levels (via ODD chaining)
  - level0: basic encoding
  - level1: richer encoding
  - level2: tokenization and linguistic annotation (work in progress)

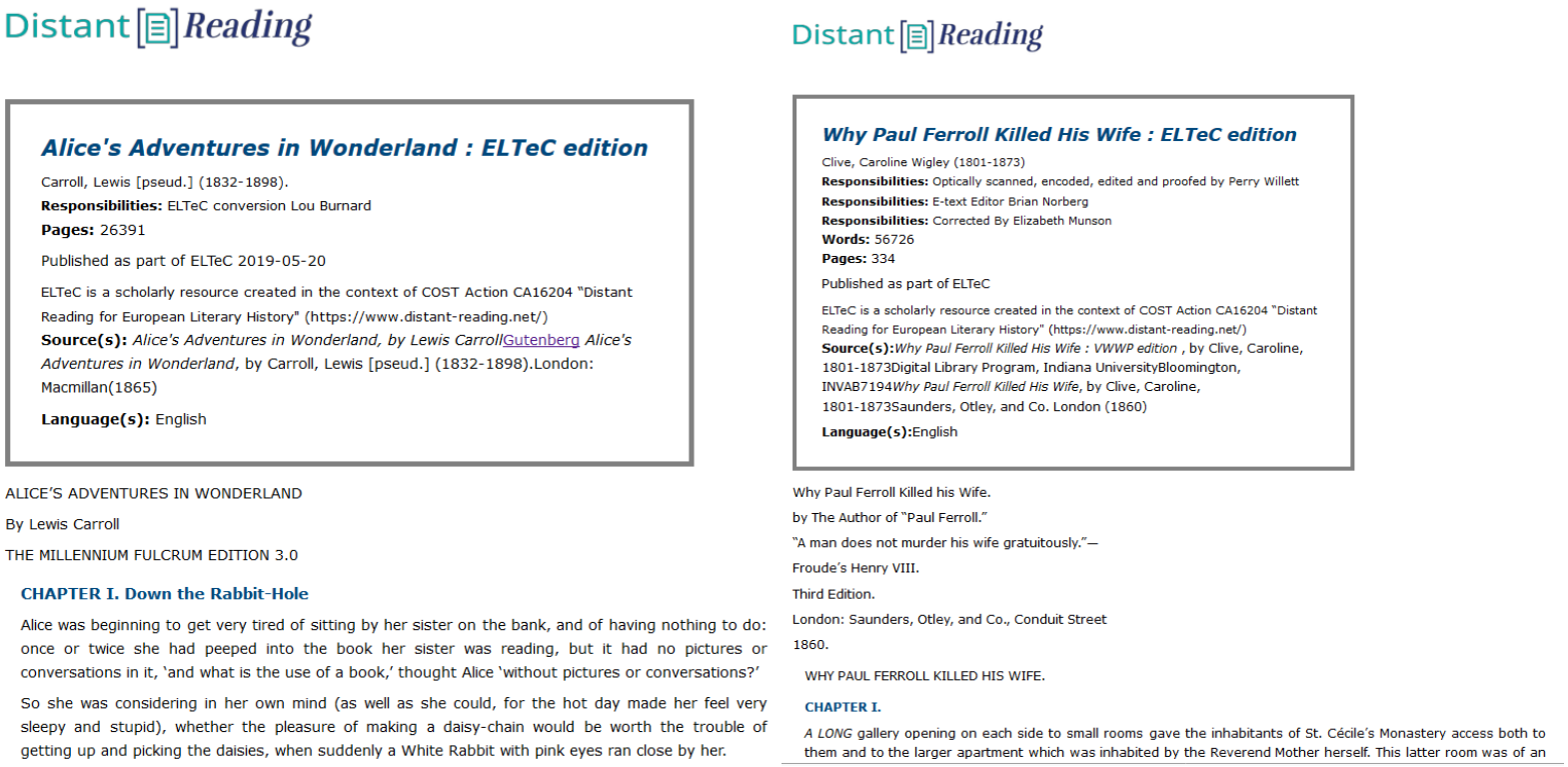
[14] cf. contribution to TEI Conference 2020 Burnard, Schöch and Odebrecht `http://gams.uni-graz.at/context:tei2019`

Outline

Current state

ELTeC Language Collections:
`https://distantreading.github.io/ELTeC/index.html`

Uniform text display



Distant Reading

Alice's Adventures in Wonderland : ELTeC edition

Carroll, Lewis [pseud.] (1832-1898).
Responsibilities: ELTeC conversion Lou Burnard
Pages: 26391

Published as part of ELTeC 2019-05-20

ELTeC is a scholarly resource created in the context of COST Action CA16204 "Distant Reading for European Literary History" (https://www.distant-reading.net/)
Source(s): *Alice's Adventures in Wonderland, by Lewis Carroll*Gutenberg *Alice's Adventures in Wonderland*, by Carroll, Lewis [pseud.] (1832-1898).London: Macmillan(1865)

Language(s): English

ALICE'S ADVENTURES IN WONDERLAND

By Lewis Carroll

THE MILLENNIUM FULCRUM EDITION 3.0

CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversations?'

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.

Distant Reading

Why Paul Ferroll Killed His Wife : ELTeC edition

Clive, Caroline Wigley (1801-1873)
Responsibilities: Optically scanned, encoded, edited and proofed by Perry Willett
Responsibilities: E-text Editor Brian Norberg
Responsibilities: Corrected By Elizabeth Munson
Words: 56726
Pages: 334

Published as part of ELTeC

ELTeC is a scholarly resource created in the context of COST Action CA16204 "Distant Reading for European Literary History" (https://www.distant-reading.net/)
Source(s):*Why Paul Ferroll Killed His Wife : VWWP edition* , by Clive, Caroline, 1801-1873Digital Library Program, Indiana UniversityBloomington, INVAB7194*Why Paul Ferroll Killed His Wife*, by Clive, Caroline, 1801-1873Saunders, Otley, and Co. London (1860)

Language(s): English

Why Paul Ferroll Killed his Wife.

by The Author of "Paul Ferroll."

"A man does not murder his wife gratuitously."—

Froude's Henry VIII.

Third Edition.

London: Saunders, Otley, and Co., Conduit Street

1860.

WHY PAUL FERROLL KILLED HIS WIFE.

CHAPTER I.

A LONG gallery opening on each side to small rooms gave the inhabitants of St. Cécile's Monastery access both to them and to the larger apartment which was inhabited by the Reverend Mother herself. This latter room was of an

Figure: Text display is based on TEI encoded files: HTML version for Alice, HTML version for Paul

Metadata composition plot

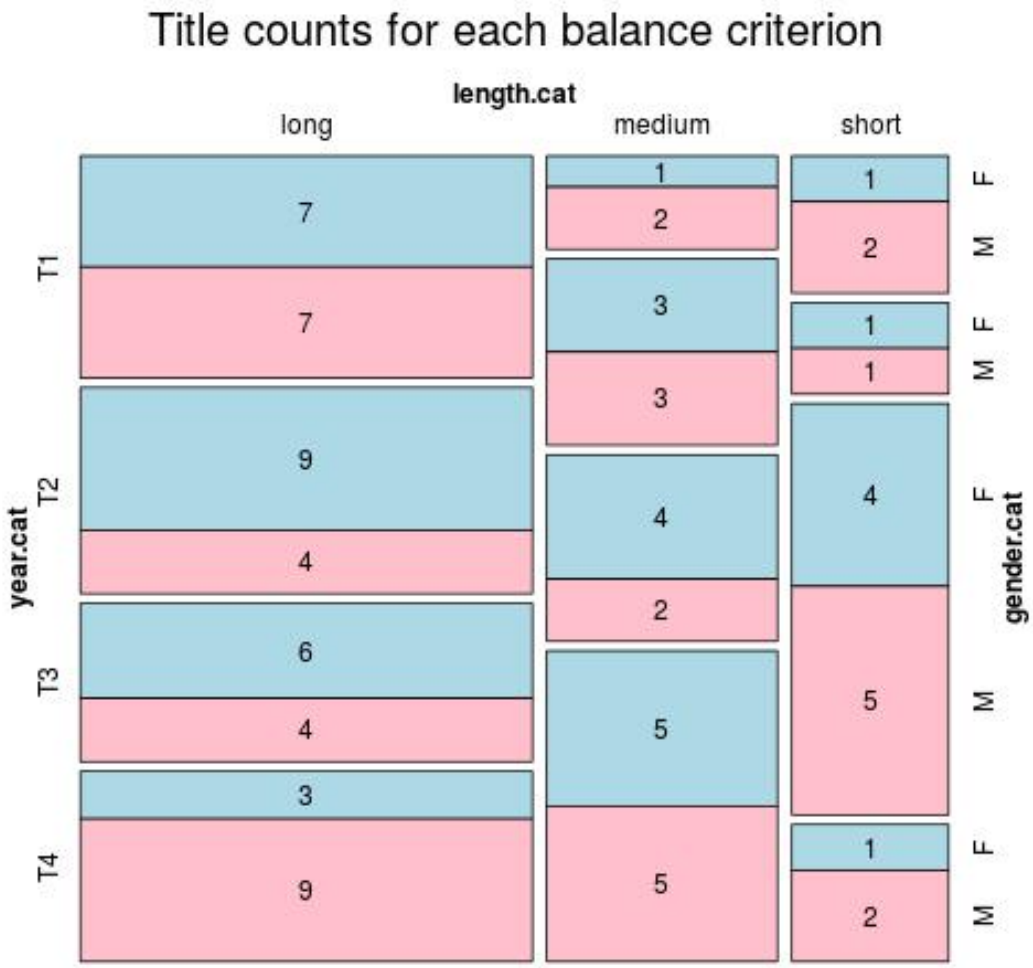


Figure: ELTeC-eng: Metadata in `teiheader` are parsed for each encoded file. Data is aggregated and visualized for corpus monitoring. Produced with ELTeC metadata and R package vcd by David Meyer [aut, cre], Achim Zeileis [aut], Kurt Hornik [aut], Florian Gerber [ctb], Michael Friendly [ctb][16].

[15] `https://cran.r-project.org/web/packages/vcd/index.html`

Outline

References COST Action Distant Reading

- COST Action Distant Reading homepage
  `https://www.distant-reading.net/`
- Documentation on `https://distantreading.github.io/`
  - Corpus design
  - Encoding guidelines
  - Working Group
  - Training schools
- ELTeC on `https://github.com/COST-ELTeC`
- ELTeC releases on Zenodo
  `https://zenodo.org/communities/eltec/`

References for introductions to XML and TEI

- Lou Burnard's Introduction to oxygen
- Martina Scholger's Introduction to TEI XML
- DARIAH's Training Digital Editions
- Lou Burnard's book on What is the Text Encoding Initiative
- Customizing TEI with ODD

References I


Algee-Hewitt, Mark and Mark McGurl (2018). *Between canon and corpus: six perspectives on 20th-century novels*. URL: https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.


Biber, Douglas (1993). "Representativeness in Corpus Design". In: *Literary and Linguistic Computing* (8), pp. 243–257.


Bode, Katherine (2018). *A World of Fiction - Digital Collections and the Future of Literary History*. eng. University of Michigan Press.


Burnard, Lou and Sebastian Rahtz (2004). *RelaxNG with Son of ODD*. Extreme Markup Languages. URL: http://www.tei-c.org/cms/Talks/extreme2004/paper.html (visited on 01/08/2017).


Caton, Paul (2013). "On the term text in digital humanities". In: *Literary and Linguistic Computing* 28.2, pp. 209–220.


Claridge, Claudia (2008). "Historical Corpora". In: *Corpus Linguistics. An International Handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 1. 2 vols. 1. Berlin: De Gruyter, pp. 242–259.


De Rose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear (2002). "What is Text, Really?" In: *Journal of Computing in Higher Education* I(2), pp. 3–26. (Visited on 04/05/2016).


Herrmann, Leonhard (2011). "System? Kanon? Epoche?" In: *Kanon, Wertung und Vermittlung. Literatur in der Wissensgesellschaft*. Ed. by Claudia Stockinger Matthias Beilein and Simone Winko. Berlin: De Gruyter, pp. 59–75.


Hunston, Susan (2008). "Collection strategies and design decisions". In: *Corpus Linguistics. An International Handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 1. 2 vols. 1. Berlin: De Gruyter, pp. 154–168.


Kytö, Merja (2011). "Corpora and historical linguistics". In: *Revista Brasileira de Linguistica Aplicada* 11, pp. 417–457.

References II

Lüdeling, Anke (2011). "Corpora in Linguistics. Sampling and Annotation". In: *Going Digital. Evolutionary and Revolutionary Aspects of Digitization*. Ed. by Karl Grandin. Vol. 147. Nobel Symposium 147. New York: Science History Publications, pp. 220–243.

Lüdeling, Anke, Julia Ritz, Manfred Stede, and Amir Zeldes (2016). "Corpus Linguistics". In: *OUP Handbook of Information Structure*. Ed. by Caroline Fery and Shinishiro Ishihara. Oxford: Oxford University Press, pp. 599–617.

McEnery, Tony and Andrew Hardie (2012). *Corpus Linguistics. Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambride [u.a.]: Cambridge University Press.

TEI-Consortium (2019). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. URL: http://www.tei-c.org/Guidelines/P5/ (visited on 06/21/2019).

van Zundert, Joris and Tara L. Andrews (2017). "Qu'est-ce qu'un texte numérique? A new rationale for the digital representation of text". In: *Digital Scholarship in the Humanities* 32, pp. 78–88.

Wilkinson, Mark D., Michel Dumontier, Aalbersberg, I Jsbrand Jan, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, da Silva Santos, Luiz Bonino, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific data* 3, p. 160018.

Zinsmeister, Heike, Erhard W. Hinrichs, Sandra Kübler, and Andreas Witt (2008). "Linguistically annotated corpora. Quality assurance, reusability and sustainability". In: *Corpus Linguistics. An International Handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 1. 2 vols. 1. Berlin: De Gruyter, pp. 759–776.