

Sampling Criteria for ELTeC – Proposal

COST Action Distant Reading
WG 1 expert meeting 2018-02-13 Prague



① Task

② Method

③ Representativeness

④ Balance

⑤ Sampling Criteria

Outline

- 1 Task
- 2 Method
- 3 Representativeness
- 4 Balance
- 5 Sampling Criteria

Defining our tasks

- Defining selection criteria
- Developing guidelines for data and metadata
- Creating a workflow for corpus preparation
- resulting in three detailed proposals

proposals available at

<https://github.com/distantreading/distantreading.github.io>

Defining our tasks

- Developing criteria for corpus design means to decide which kind of sample of the world shall be included in the data base.
- Finding a compromise between what we would like to have in the corpus (all literature) and what we can put in the corpus (sample)
- *It is a truism that there is no such thing as a good or a bad corpus, because how a corpus is designed depends on what kind of corpus it is and how it is going to be used.* (Hunston 2008, p. 155)

Goal

- Creating a benchmark corpus
 - Allow for the creation of individual subcollections/subsets of the corpus by COST Action members or others
 - For specific tasks and research questions
 - Clear, operationalized, transparent and motivated selection criteria
 - Please note that we can use many different kinds of metadata to describe the novels in the corpus!
 - We will define just a few metadata as selection criteria.

- Creating a benchmark corpus
 - Allow for the creation of individual subcollections/subsets of the corpus by COST Action members or others
 - For specific tasks and research questions
 - Clear, operationalized, transparent and motivated selection criteria
 - Please note that we can use many different kinds of metadata to describe the novels in the corpus!
 - We will define just a few metadata as selection criteria.

Outline

- ① Task
- ② Method
- ③ Representativeness
- ④ Balance
- ⑤ Sampling Criteria

Basis for text selection

- Canon-based corpus design vs. metadata-based corpus design
- Differences in motivation of criteria, text definitions and perspectives on literature in Europe

Motivation for selection

- Canon-based
 - A portrait of someones prestigious social, cultural, economic status reflecting normative self-promotional legitimating and rating decisions (Algee-Hewitt and McGurl 2015; Beilein et al. 2011; Herrmann 2011)
- Metadata-based
 - Follows a research question or context and is therefore more research goal oriented (Hunston 2008; Lüdeling 2011)

Canon-based

- Varieties of canons
 - For each language
 - For each cultural context
 - Reflect different prestigious groups (e.g. publishers, authors and readers)
 - Multi-relational decisions and ratings
 - Intellectual rating, economical rating or readers rating
- Change over time, reflect different interpretations of famous, important or influential texts
- Not overall comparable, not categorial

Metadata-based

- Oriented towards research question or/and contexts (distant reading!)
- Defined by a distinct set of relevant metadata (bibliographic, technical, administrative)
- Sampling instances of a population
- Selection without reading the texts

Goal of the Action

This Action will develop the resources and methods necessary to change the way European literary history is written. (MoU, p.2)

- Canons provide traditional and normative access to the history of literature.
 - *The current canon sets limits to our understanding of literature, in several ways* (Fowler 2002, p. 214)
 - Choosing between canons can then mean choosing between tastes of (current and past) literature (in past and present)
- Metadata-based corpus design allows to create a new perspective on novels on Europe.

Different texts definitions

- Different ways of considering the actual texts
 - Consider the manifestation or the extension or the work of a text (cf. IFLA 2009)
 - A canon can contain an extension of a certain text which is available in different languages and prints.
 - Ontologically, these different levels of text are different from what a text in a corpus might be (van Zundert and Andrews 2017).
 - Data is ontologically different from the world. (Moisl 2009, p. 876)
 - Digitization is a kind of annotation, hence interpretation (Odebrecht et al. 2017).

Outline

- ① Task
- ② Method
- ③ Representativeness
- ④ Balance
- ⑤ Sampling Criteria

Representativeness

- *Representativeness refers to the extent to which a sample includes the full range of variability in a population.* (Biber 1993, p. 243)
- Is not an innate quality, like colour or weight
- Requires knowledge about the whole population of literature
- No knowledge about every book of every language published/read/discussed in Europe in the period in question
- *impossible to identify a complete list of categories that would exhaustively account for all texts produced in a given language* (Hunston 2008, p. 161)

Representativeness

- Basic Question: Would we like to use each criterion, with the intention to represent the variety of possible values, or should the sample represent the distribution of those values across the population?
- Distribution means to represent the population statistically
- Variety means to represent possible features of the population
- Proposal: Represent the variety of possible values for each criterion, and hence aim for approximately equal proportions for each of them

Outline

- ① Task
- ② Method
- ③ Representativeness
- ④ Balance
- ⑤ Sampling Criteria

Balance

- Control the proportion of texts in a collection concerning one or more features
- Decide which feature is balanced in which way
- One feature can interplay with other features (ranking?)
- Following the approach of representing variety of a population

Balance

- Defining the main feature for balancing
- Defining the relation between these features
 - For example balancing gender over the whole period or balancing gender in sub-periods
 - e.g., equal number of male and female authors in the whole period (potentially more female authors at the end of the period)
 - e.g., equal number of male and female author in sub-periods such as decades (potentially equal number of gender in every decade)

Outline

- ① Task
- ② Method
- ③ Representativeness
- ④ Balance
- ⑤ Sampling Criteria

Requirements according to the MoU

- Languages: Dutch, English, French, German, Greek, Italian, Polish, Portuguese, Russian, Spanish (ELTeC core)
- Homogeneous with respect to genre
- Balanced with respect to language

Requirements for criteria – interim summary

- Clear, operationalizable
- Represent the variety of the population
- Decidable without reading the novel
- Text-external and text-internal criteria
- Comparable

Criteria – text definition

- Including only first editions of a novel published as a book
 - No translations
 - Excluding publications in journals
 - Prefer electronically available texts (no funding for digitizations)
- First editions are free (no copy right issues)
- First editions are interesting from a philological point of view (might require additional normalizations, next steps)
- Books are comparable, printed full text (no potential text excerpts in different journal editions)

Criteria – text definition

- Including only first editions of a novel published as a book
- No translations
- Excluding publications in journals
- Prefer electronically available texts (no funding for digitizations)
- First editions are free (no copy right issues)
- First editions are interesting from a philological point of view (might require additional normalizations, next steps)
- Books are comparable, printed full text (no potential text excerpts in different journal editions)

First list of criteria

- Language
- Date
- Reprint
- Author Gender
- Length
- Kind of novel/topic

Language

- First iteration: 6 subcollections (100 novels per language) 1850 to 1920 starting with British, French, Spanish, German, Greek, Polish
- Second iteration: 4 subcollections (100 novels per language) 1850 to 1920
- Third iteration: 6 subcollections in additional languages and subcollections for all 16 languages 1780-1850

- Defining sub-periods
- group A (1850-1863)
- group B (1864-1877)
- group C (1878-1891)
- group D (1892-1905)
- group E (1906-1920)

Reprint count

- Low: reprinted less than 10 times
- Medium: reprinted 10 to 100 times
- High: reprinted more than 100 times

Gender

- male
- female
- mixed (undefined, more than one author)

Length

We should maybe try to include a variety of lengths

- short (less than 5000 words)
- medium (5000 to 20000 words)
- long (more than 20000 words)

Balancing example

Table 1: Minimum and maximum numbers of titles to be selected for each criterion

Language	Date	Reprint	Author
80-120	20	5-10	5-8

Topic/Genre

- We may try to include a variety of genre/topic?
- Or include this category for describing the novels?
- Difficult criterion, depending on scholars, normative theoretical definitions
- *texts that theory currently identifies as novels were before not circulated and understood as novels*
 - Classifications in bibliography, e.g. keywords
 - others?
- First draft from WG3
https://docs.google.com/document/d/1D_6lYieLPuXu7FV50KVBsjPA70h2QIw7HP0BVTSPQKM/edit?usp=sharing

Many thanks to Antonija and WG3!

Ready for discussion!

References I



Algee-Hewitt, Mark and Mark McGurl (2015). "Between Canon and Corpus: Six Perspectives on the 20th-Century Novels". In: *Stanford Literary Pamphlet*.



Beilein, Matthias, Claudia Stockinger, and Simone Winko, eds. (2011). *Kanon, Wertung und Vermittlung: Literatur in der Wissensgesellschaft*. Vol. Bd. 129. Studien und Texte zur Sozialgeschichte der Literatur. Berlin: De Gruyter.



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



Fowler, Alastair, ed. (2002). *Kinds of Literature: An Introduction to the Theory of Genres and Modes*. Oxford.



Herrmann, Leonhard (2011). "System? Kanon? Epoche?" In: *Kanon, Wertung und Vermittlung*. Ed. by Matthias Beilein, Claudia Stockinger, and Simone Winko. Vol. Bd. 129. Studien und Texte zur Sozialgeschichte der Literatur. Berlin: De Gruyter, pp. 59–75.



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



IFLA (2009). *Functional Requirements for Bibliographic Records*. URL: <http://www.ifla.org/publications/functional-requirements-for-bibliographic-records> (visited on 12/23/2016).



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



Moisl, Hermann (2009). "Exploratory Multivariate Analysis". In: *Corpus Linguistics*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 2. Berlin: De Gruyter, pp. 874–899.

References II



Odebrecht, Carolin, Malte Belz, Amir Zeldes, Anke Lüdeling, and Thomas Krause (2017). "RIDGES Herbology: Designing a Diachronic Multi-Layer Corpus". In: *Language Resources and Evaluation* 51.3, pp. 695–725.



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.