Sampling criteria for the ELTeC

Cost Action CA16204 – WG1 *2018-01*

Sampling criteria for the ELTeC

1 Outline

2 Task

The task for WG1 is to develop guidelines for data and metadata for the creation of the ELTeC. This task can be split up into several distinct tasks: Guidelines for corpus design, basic annotation and metadata schemes and workflow. This discussion paper focuses on corpus design and metadata because both tasks interplay with each other.

The goal of CA16204 is to create a big benchmark corpus of literature from 1850-1920 (first period) for different computational distant reading methods for corpus annotation and analysis. The task of creating annotation guidelines of WG1 needs to be closely communicated and coordinated with WG2 in order to know which methods and tools needs which kind(s) of annotation model and format. The same holds for the development of the metadata scheme.

For creating such a benchmark corpus, we need a corpus design which allows for a comparability of texts and individual sub-collections according to different metadata set(s). It should be possible for every COST Action member to sample sub-collections from the ELTeC for specific tasks and research questions. In a first step, we focus on the development of clear, operationalized, transparent and motivated selection criteria for the corpus.

3 Method

First, we would like to discuss whether the task of corpus building requires normative selection criteria (canon) or methodical selection criteria (corpus sampling). The first and foremost important difference between the approaches is the motivation of criteria. A canon is the portrait of someone's prestigious social, cultural, economic status and it reflects normative self-promotional legitimating and rating decisions. In contrast, corpus design follows a research question or context and is therefore more research goal oriented. A second important aspect is the way of considering the actual texts. As Moisl (2009: 876) puts it 'Data is ontologically different from the world.' So there is a difference between texts in the world and data we create. By texts, we may 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 (cf. van Zundert and Andrews 2017). This means, that digitization is a kind of annotation, hence interpretation (Odebrecht et al 2017). A representation of a text

in a corpus (e.g. transcription, OCR) is the result of interpretation. A corpus design needs to consider this for sampling and digitization issues.

At the end of the Action, ELTeC should contain literary texts (novels) from a distinct period and in several languages. For each language (and thus for each cultural context) there exists a diversity of canons which reflect different prestigious groups such as publishers, authors and readers. These actors and their possible influence on canon building is multi-relational. Each canon is a result of rating texts from different perspectives. The assessment can reflect intellectual rating (a text is a representative of a certain literature period, is influential, is important), economical rating (a text is published in more than one print run), or readers rating (a text is most popular within a certain reader group) (cf. Hermann 2011 or Winko 1996). All these ratings can change over time and may also interplay which each other. A canon can therefore reflect different interpretation of 'famous', 'important' or 'influential' texts. These criteria are not overall comparable. For example, texts from a smaller language community such as Czech are less likely to be frequently reprinted than English texts of the same period and genre. Additionally, the awareness level of texts depends on the influence of the country and their publishing houses.

The criteria derived from a canon are not completely comparable and categorical. Which prestigious group's canon should be considered, which should be excluded, and why? Are there comparable canons for novels in all countries of the language in question? Algee-Hewitt and McGurl (2015) show an approach to corpus design based on several canons and which kind of problems occur. Each analysis of the corpus than only shows the different effects of the decisions made by the normative group. Considering national canons is also very difficult and somewhat problematic. An example is the German National Canon of literature which was developed in the 18th century and was promoted by the national educational system until the 1990. Since German reunification, the educational system has not promoted a strict canon and does not recommend a list of books to be read in school or at university (cf. Winko 1996). Thus, taking such types of canons as part of a sampling base of ELTeC would mean reflecting a political and social past of German education and politics. Choosing between canons can then mean choosing between tastes of (current) literature (in past and present) and tastes of past literature when the canon builder rates historical texts. Finally, these canons are not built to be the sampling guidelines for a corpus which we would like to build in the Action.

The MoU of CA16204 formulates the goal as follows: "The main aim and objective of the Action is to This Action will develop the resources and methods necessary to change the way European literary history is written." This goal requires a new approach to corpus design, metadata design and annotation models. As Fowler (2002, 214) puts it: 'The current canon sets limits to our understanding of literature, in several ways'. Relying on canons will obstruct the Action's goal in a fundamental way. Canons provide traditional and normative access to the history of literature. In contrast, the Action focuses on new approaches to tell another story. Instead, we might decide that our collection should contain a mixture of works that have never been reprinted since their first appearance, works that have been reprinted a small number of times within one or two decades of their first appearance, and works that have been reprinted in almost every decade since their first appearance.

Therefore, we argue for a non-normative but metadata-based approach of sampling criteria which will follow a corpus design approach. Corpus sampling criteria are mostly oriented/developed by the research question or/and contexts of the corpus creators group. In CA16204, we have neither a distinct research question nor a fixed and previously known corpus creator group. The research context of the Action is more interested in knowledge production in a methodological sense and does not prefer a single method, model or theory. Furthermore, the member group of the Action will fluctuate and consist of researches from different disciplines with different theoretical and cultural contexts. Thus, we need to build the corpus design on a

methodical basis. With this method, we will also be able to select canonical texts as well but not exclusively.

4 Representativeness and balance

Additional to the aspect of 'prestige' (canon), the aspect of representativeness is problematic for corpus design. Developing criteria for corpus design means to decide which kind of sample of the world shall be included in the data base. Obviously, including the whole population of 9th century literature in several languages is impossible. So we need to make 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, 155)

Following Hunston (2008), a corpus design needs to follow the research goal. For the Action, representativeness may be the relationship between the corpus and the body of literature in question. 'Representativeness refers to the extent to which a sample includes the full range of variability in a population' (Biber 1993, 243). To say something about the representativeness of a corpus requires knowledge about the whole population of literature (in the period in question). Actually, we don't know every book of every language published/read/discussed in Europe in the period in question. It is further 'impossible to identify a complete list of 'categories' that would exhaustively account for all texts produced in a given language', (Hunston 2008, 161) or context. Such categories can refer to various factors such as characteristics of authors, e.g., gender or place of birth, publishers, topics of the texts, readership etc. Against the background of canon building, there is also 'no true measure of the 'significance' of a type of discourse to a community'. The chance, that a corpus represents the whole population of something increases with the size of a corpus. In this way, size and representativeness are connected. [1] Representativeness is therefore a kind of ideal which we would like to pursue but which cannot be achieved as whole. In line with the MoU, the ELTeC can be designed as a monitor corpus where texts (from different languages and periods) can be added over time.

Balance refers to the internal proportion of the corpus. Note that a fully balanced corpus is an ideal which we only can try to achieve next to the ideal of representativeness (Hunston 2008, 163). According the MoU, the corpus shall contain 2,500 full-texts novels at least in 10 different languages:

- Languages: Dutch, English, French, German, Greek, Italian, Polish, Portuguese, Russian, Spanish (ELTeC core)
- 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,

In this way, ELTeC is balanced with respect to language. With respect to genre, the corpus is not balanced but homogenous; all texts in the corpus shall be full novels. With respect to time, the corpus design shall focus on the period 1850 to 1920.

Before discussing the criteria in more detail, we would like to ask another methodical question concerning corpus sampling: 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?

[1] See Biber (1993) and Hunston (2008) for a detailed discussion.

Let's say we wish to select 100 texts from a population of texts published over a period of (say) 20 decades. We might select five texts from the first decade, five from the second, and so on, making up our 100 titles, evenly spread across the possible decades. The probability that a text in our corpus will come from any given decade will always be the same: 1 in 5. This selection represents the *variety* of possible values for the criterion. Suppose now that we look more closely at the number of titles from each decade actually available in the population we are sampling. It's more than likely that this number will vary significantly: for example, we might notice that there are 2000 titles published in decade x, and only 100 in decade y. To represent this population *statistically* we should therefore make it 20 times more probable that a randomly chosen title will come from decade x than from decade y. Since the total number of titles we can choose is quite small relative to the total number available in the population, strict application of this principle may mean that we cannot choose any titles at all from some decades. This is one reason for preferring to make our sampling represent variety rather than frequency; another is that we cannot choose fractional numbers of titles. When we start considering more than one criterion, the task of ensuring that the numbers in our sample accurately reflect the distribution of all values across the population becomes prohibitively complex.

Following the approach of representing the variety of a population, we then need to decide which criterion is balanced in which way and interplays with other criteria. For example, we may want to choose novels from male and female authors in a balanced way. This may mean that in the total of all novels one half will be from female authors. Without any further regulation, we might have more female authors in one decade than in other decades. If we would like to have an equal number of male and female authors in every decade, we need to link the criterion of the author's gender with the criterion of time. Doing this, might complicate the selection process (cf. finding novels for this proportion in every decade of the period in question). So we have to decide which categories shall be present in a balanced way in the corpus.

5 Requirements of sampling criteria

According to the MoU, the corpus design should be balanced with respect to language and publication date of the texts. This means that the corpus should not be based solely on chronological criteria, meaning that we need a text from each year of the period in question. The main sampling criterion 'language' will require not to including translations at all. We will then take the first edition of a novel. By a novel, we mean the first edition of the book, hence excluding novels (or versions of them) printed only in journals. Considering only the first editions has two advantages: the first edition should be free in most cases, meaning they carry no copyright. Later editions and reprints may have such restrictions. The first edition is more interesting from a philological point of view. It represents the authentic texts of the authors. Dealing with historical texts requires some cleaning up and normalizations (we need to discuss these steps later on).

Electronically availability should not be a leading sampling criterion althoug availability is a limiting factor. A text should not be excluded from ELTeC because it is not digitized, but it should be excluded if the text cannot be made freely available in ELTeC. If we only use availability as a selecting criterion, we are at risk of copying projects such as 'Gutenberg' for example. We then need additional criteria which can be applied without having to know (read) the texts in question. The criteria should be checked without a deep knowledge about the texts. Otherwise, this will oppose the goal of the whole Action and the methodical approach of distant reading. The criteria should be operationalizable, meaning decidable from text metadata. Here, we define text metadata in a wider scope than only the classical bibliographical metadata. In this way corpus design interacts with metadata. Some of the text's metadata can be used a sampling criteria. These criteria are text-external and -internal criteria (cf. Hunston 2008) on which we then need to rely.

We suggest using an online table as a means of collecting nominations for inclusion in the ELTeC but other methods are feasible.

6 Sampling criteria

Principles:

- Represent the variation of production
- First publication of the novel as a book
- No translations

Criteria:

Date : 1850 to 1920 (first iteration)

Rather than dividing by decade we propose longer time slots, dividing the period into five subgroups

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

Language of the text

The MoA defines the languages to be sampled. It does not propose distinguishing regional variation (e.g. in German), nor geographical variation (e.g. the French of Belgium or France or Switzerland). It assumes only European varieties, so English excludes US English; French excludes Quebecois.

- Dutch, English, French, German, Greek, Italian, Polish, Portuguese, Russian, Spanish

Reprint count

We propose to use the number of times a work is reprinted as an objective measure of its reception, using categories like the following:

- low: reprinted less than 10 times
- medium: reprinted 10 to 100 times
- high: reprinted more than 100 times

Author gender

We use the following three categories for actual (not claimed) author gender

- male
- female
- mixed (undefined or 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)

Kind of novel

It is an open issue how to classify novels by topic. We are not sure how to use this as a sampling criterion.

- Main theme

The following table suggests minimum and maximum numbers of titles to be selected for each criterion. For example, for each language, we would aim to select between 80 and 120 different titles. Of these, no more than 20 will come from each date period defined above. Of these, we will aim to have between 5 and 10 examples from each of the reprint categories defined (high, medium, low). Of these, the number produced by male authors will be between 5 and 8, as will the number produced by female authors, and the number produced by authors of unknown or indeterminate sex.

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

7 Metadata for texts in ELTeC

We list here the metadata items to be collected for each text. These will be provided by specific components of the TEI Header structure, as definding in WG paper on Encoding.

- title
- subtitle
- publication date
- publication place
- publisher
- series
- editor
- author
  - name
  - sex
  - first language
  - place of birth
  - entry of a person database such as GND if available
- size

  - in words / in tokens
- pages
- topics (cf. Falkentheorie)
  - politics, crime, espionage, provincial, school, adventure, war, faith, domestic, nature, factory
- subgenre
  - work-in-progress novel
- Narrator
  - first person
  - third person
  - authorial narrator
- canonicity/reception
  - rated by a canon: y/n
  - canon
- keywords
  - from OPACs? others? Own?
- language
  - possible: language area, language type (e.g. German, Bavarian)
- Source reference (if available)
  - e.g. DTA, textgrid etc.

8 Literature

[1] Algee-Hewitt, Mark; McGurl, Mark (2015): *Between Canon and Corpus. Six Perspectives on the 20th-Century Novels.* Stanford Literary Lab Pamphlet no 8.

[2] Biber, Douglas (1993): 'Representativeness in Corpus Design.' In: *Literary and Linguistic Computing* (8), 243–257.

[3] Herrmann, Leonhard (2011): 'System? Kanon? Epoche?' In: Matthias Beilein, Claudia Stockinger und Simone Winko (Hg.): *Kanon, Wertung und Vermittlung. Literatur in der Wissensgesellschaft.* Berlin: De Gruyter (Studien und Texte zur Sozialgeschichte der Literatur, Bd. 129), S. 59–75.

[4] Hunston, Susan (2008): 'Collection strategies and design decisions.' In: Anke Lüdeling und Merja Kytö (Hg.): *Corpus Linguistics. An International Handbook.* 2 Bände. Berlin: De Gruyter (1), S. 154–168.

[5] IFLA (2009): *Functional Requirements for Bibliographic Records* (Technical Report). Online verfügbar unter http://www.ifla.org/publications/functional-requirements-for-bibliographic-records, zuletzt geprüft am 23.12.2016.

[6] Lüdeling, Anke (2011): 'Corpora in Linguistics. Sampling and Annotation'. In: Karl Grandin (Hg.): *Going Digital. Evolutionary and Revolutionary Aspects of Digitization.* New York: Science History Publications (Nobel Symposium, 147), 220–243.

[7] Moisl, Hermann (2009): 'Exploratory Multivariate Analysis'. In: Anke Lüdeling und Merja Kytö (Hg.): *Corpus Linguistics. An International Handbook.* 2 Bände. Berlin: De Gruyter (2), S. 874–899.

[8] Winko, Simone (1996): 'Literarische Wertung und Kanonbildung'. In: *Grundzüge der Literaturwissenschaft.* Hrsg. v. H. L. Arnold und H. Detering. München, 585–600.

[9] van Zundert, Joris; Andrews, Tara L. (2017): 'Qu'est-ce qu'un texte numérique? A new rationale for the digital representation of text.' In: *Digital Scholarship in the Humanities* (32), S. 78–88. DOI: 10.1093/llc/fqx039.