**Cost Action CA16204 – WG1**

Sampling criteria for ELTeC

Outline

1. Task
2. Method
3. Representativeness
4. Requirements of sampling criteria
5. Criteria
6. Metadata for texts in ELTeC
7. Literature
8. **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, metadata scheme and annotations. This discussion paper focuses on corpus design and metadata because both tasks interplay with each other.

The CA16204 has the goal 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, operationalizable, transparent and motivated selection criteria for the corpus.

1. **Method**

First, I 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 a normative self-promotional legitimation and rating decisions. In contrast, corpus design follows a research question or context and is therefore more research goal oriented. A second important difference is the way of considering the actual texts. Canons and corpus design rely on different text definitions and interpretations. As Moisl (2009: 876) puts it ‘Data is ontologically different from the world.’ Canons select texts to be included. By texts, the canons may consider the manifestation, 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 to 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. transcriptions, 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 from several languages. For each language (and thus for each cultural contexts) there exists a diversity of canons which reflect different prestigious groups such as publishers, authors and readers. These actors and there possible influence on canon building is multi-relational. Each canon is a result of rating texts from different perspectives. The rating can reflect intellectual rating (a text is a representative of a certain literature period, is influencing, is important), by economical rating (a text is published in more than one print run), or by readers rating (a text is most popular by 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 such as Czech are less likely to be highly reprinted in a way English texts from the same period and genre are. 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 shall be considered which shall be excluded and why? Are their comparable canons for novels in all countries of the language in question? Algee-Hewitt and McGurl (2015) show an approach of a corpus design based on several canons and what kind of problems occur. Each analysis of the corpus than only shows the different effects of the decisions made by the normative canon 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. After the German reunification, the educational system does not promote a strict canon and does not recommend a list of book 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 to reflect 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 of 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 oppose to 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.

Therefore, I 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 no 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 these method, we will also be able to select canonical texts as well but not exclusively.

1. **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 literature of the 19th century of 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 qeustion). Actually, we don’t know every book of every language published/read/discussed 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]](#footnote-1)

Representativeness is therefore a kind of ideal which we would like to pursuit but which cannot not 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. 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)
* 1st iteration: 6 subcollections (100 novels per language) 1850 to 1920 starting with British, French, Spanish, German, Greek, Polish
* 2nd iteration: 4 subcollections (100 novels per language) 1850 to 1920
* 3rd 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, all texts in the corpus shall be full novels. With respect to time, the corpus design shall focus on the period 1850 to 1920. A fully balanced corpus is an ideal which we only can try to achieve next to the ideal of representativeness (Hunston 2008, 163). Balancing a corpus with respect to gender can mean either that we chose the same amount of novels from each gender or/and that we chose the same amount of words for each gender (if we can determine the gender of every author and those using pseudonyms). To achieve both balances it nearly impossible. So we have to decide which categories shall be present in a balanced way in the corpus.

1. **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 not that the corpus should be solely based 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 include translations at all. The electronical availability should not be a leading sampling criterion. But, 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 German historical texts. 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 on.

1. **Criteria**

In this section, we provide the sampling criteria for ELTeC. Overall, ELTeC should contain fictions – printed novels – no translations.

Authors should write in their first language?

Authors should be located in Europe?

Feeling, that it is somewhat not the most important issue which distinct novel is in the corpus as we have no curation approach, no thematic predominance due to research question etc.

Implementing a member vote when having a preselection?

* Date
  + 1850 to 1920 (first iterations)
  + text sample covering each decade in the period
* Language of the text
  + Dutch, English, French, German, Greek, Italian, Polish, Portuguese, Russian, Spanish
* Author
  + male, female, (undefined)
* Length
  + Average (cf. balancing)
* Kind of novel
  + Main theme (cf. Falkentheorie)

1. **Metadata for texts in ELTeC**

For every text in the ELTeC, the following metadata should be recorded.

* 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.

1. **Literature**

**Algee-Hewitt, Mark; McGurl, Mark** (2015): Between Canon and Corpus. Seix Perspectives on the 20th-Centurary Novels. Literary Lab.

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

**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.

**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.

**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.

**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.

**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.

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

**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.

1. See Biber (1993) and Hunston (2008) for a detailed discussion. [↑](#footnote-ref-1)