Notes on design considerations for the ELTeC.

I agree with Carolin's assessment that our selection criteria should not primarily be the "canonicity" or otherwise of a text. We need to construct a collection which has many uses, and thus may well need to include some material which has not been considered of any particular cultural value by some social group at some particular time. Nevertheless, the reception of a text (by which I mean all aspects of its distribution, but particularly its longevity as an object of interest in a culture) seems to me to be an important characteristic we should not neglect. I suggest that "frequency of reprint" might be a simple way of approaching this difficult topic: we might for example 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.

This of course raises immediately what seems to me (perhaps because I am a statistical ignoramus) a crucial question about our sampling principles. Many of the more difficult issues in corpus design are already settled for us :

- We are collecting whole texts from a single genre : so we don't have to worry about sample size, or text type.
- We are collecting a predefined number of texts from a predefined number of languages: so we don't have to worry about corpus size or language usage.
- The population from which we are sampling is well defined (novels written in a given language from a given time period) and thus we can easily assess the representativeness or balance of our sample, in the sense that Biber and Hunston use these terms.

Nevertheless, we should make explicit the principles underlying our sampling procedure. Our collection will be made by consciously selecting instances according to some set of criteria. Our intention is to produce a stratified representative sample of the population. For each stratum, each criterion, is the intention to represent the variety of possible values, or should the sample represent the distribution of those values across the population?

To clarify this distinction, consider a selection made according to one criterion: date of first publication. 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 prefering 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. The drawback is, of course, that interpreting the significance of statistical evidence derived from our collection will be more complex. As a trivial example, suppose that we are investigating a phenomenon (say the number of four letter words in a sentence) which we might expect to be normally distributed, and detect that this phenomenon occurs ten times more frequently in texts from decade x than from decade y. If we

wish to extrapolate from this observation (which relates to properties of our constructed collection) to make a judgment about the population as a whole, it makes a great deal of difference if the population of texts from decade x is twenty times as big as the population of texts from decade y. We may simply be witnessing an effect of an imbalance introduced by our "balanced" corpus.

In addition to date and language, which are already decided upon, Carolin proposes the following additional selection criteria:

- gender of author
- length
- kind of novel

And, as noted above, I would like to add a fourth:

- reprint frequency

We need to decide whether (if we apply them all) they are independent of each other, as well as whether we are sampling for variety or for frequency. Assuming the former, for example, our sample should reflect the variety of available values. For gender therefore, roughly one third of the texts chosen should be of female authorship, one third of male authorship, one third of unknown or unspecified gender. But we might achieve this in two ways: we might ensure that one third of each decade's texts have (say) female authors, or we might ensure that one third of the whole collection, irrespective of decade, should have female authors. We do not know (though it seems possible) that female authorship is more or less probable at different periods, or for different kinds of novel, or for different lengths; and in fact this question arises for all the sampling criteria.

As regards the proposed criteria, I agree that authorship is an important consideration, particularly for the novel. I wonder whether geographical or social origins of the author are not as significant as (claimed) gender though. Length seems to me to be of lesser importance: if something is identified as a novel, this will normally imply something about its length, and so we might not find much variation in the population here beyond outlying cases (very much longer or shorter than the average). I am not sure what is meant by "kind of novel" and the reference to *falkentheorie* did not help me much (even after I looked it up in Wikipedia), but I am guessing that the idea might be to provide a set of predefined categories pertaining to topic or style. There is much debate about the concept of "genre", and while it is clearly useful to distinguish what is called in English "genre" fiction (wild west stories, detective novels, historical fantasy, science fiction) from "literary" or "high brow" fiction, it's not a distinction that is always easy to make, or reliably objective as a sampling criterion.

Whichever set of sampling criteria we decide upon, there will of course be many more descriptive criteria available for those wishing to select particular profiles of text for analysis. Some of these are listed at the end of Carolin's document and others will appear in my proposal for metadata encoding. In general I think we should aim to identify a useful list of metadata categories which we can guarantee to document for each text, even if the value is "unknown". This is essentially a bibliographic task, and we are fortunate in that extensive bibliographies for the majority of the works we will be including already exist.