«One of the Very Non-Canonical Literatures:

Ukrainian Subcollection in the ELTeC»

1- how was the project organized ? (personnel, management…) (c. 100 words)

2- what is the current state of the collection ? (E5C, encoding, open issues, etc.) (c. 200 words)

3- how were texts selected? (source/s used; balancing methods etc,) (c. 200 words)

4- how did you learn how to use TEI ? (self-study, training, previous knowledge?) (c. 200 words)

5- how were texts processed? (software used etc.) (c. 200 words)

6- what problems did you encounter? (c. 200 words)

SLIDE 1: HELLO-TITLE

Ukraine joined the action as one of the last countries, namely at the end of 2019. I am the only member of national the team, which of course increases the workload but also gives certain advantages. Several colleagues from Ukraine occasionally helped me with the search for full texts.

Thus, to date, the formation of the collection took place over the course of a year. The pace of work could be faster, but of course, like all of you, I combine this work with several other workloads and ongoing research.

SLIDE 2: CURRENT STATE OF THE UKRAINIAN SUBCOLLECTION

Now the Ukrainian collection contains … texts. At the turn of this year, I hope to have a figure closer to 50 and go beyond this indicator next year. Almost all texts are encoded at level 1. The ELTeC Corpus Composition Criteria Compliance Calculator shows the level of 65 points for Ukrainian subcollection. What can improve this indicator is, of course, an increase in the number of novels, especially long ones. It is also would be nice to have more texts for the first time slot, 1840-1859.

SLIDE 3: TEXTS SELECTION: IMAGES OF SITES

As I said, I didn't have many live opportunities to learn from the experience of fellow compilers of other national collections. Thanks to the meeting in Malaga, the selection of texts became more clear to me. For example, for the Ukrainian situation, it is important to have the opportunity to take texts not only from separate editions, but also from periodicals and almanacs.
It goes without saying that I used every possible source to increase the subcollection. But in general, there are 7-8 sites with the most of the digital copies-scans of the first editions and / or plain texts.

For Ukrainian subcollection, private initiatives of making electronic libraries appeared to be surprisingly useful.[1] In some aspects, they not only do not lag behind but are even ahead in comparison with similar collections prepared by national libraries with their big teams involved. It should be also noted that a large number of publications of Ukrainian literature have been digitized in the archives and libraries of North America. Later, many of them appeared on free accessible sites, including those for Ukrainian users.

In any case, I did not bring the texts to the ELTeC without checking them against the first edition. And of course, I did mention them and make reference to the Internet source if there was any.

SLIDE 4: TEI: LINK TO TEI BY EXAMPLE

As for my first meeting with TEI, I learned about the existence of the initiative about 5 years ago. One of my scholarly interests is the publication of literary works and digital editing in particular. In other words, the field that remains rather new for both Ukrainian and Slavic studies.

I also had the opportunity to hear briefly about TEI encoding at the Digital Humanities School in Antwerp in 2018. Then Dirk van Huelle showed us some examples, and at exactly that time I first run Oxygen editor on my laptop.

I first started using TEI principles exactly for encoding texts for the Ukrainian sub-collection in ELTeC. It was not easy at first, but

[1] For instance, e-libraries *Chtyvo* https://chtyvo.org.ua/help/biblioteka/pro-nas or *Izbornyk* http://litopys.org.ua/. An interesting detail is that it is not easy to find information about some of the authors who initiated such large-scale (and noble by their intention) projects. Such anonymity may be related, among other things, to the fact that publications do not always adhere to copyright regulations.

Carolin Odebrecht helped me. In particular, she provided links to very effective slides from a training school in Budapest.

And a little later I learn about the TEI by Example course and worked with it as well.

https://teibyexample.org/

SLIDE 5: TEXTS PROCESSING, ISSUE, CONVERTOR TOOL, TOTAL TIME
A few words on how the texts were processed.

The way I worked with them depended on whether there was a plain text or not.

If not, then OCR via Abby Fine Reader. I'd mention here that the Ukrainian orthography was at the stage of formation and Ukrainian graphical characters within the Russian Empire were prohibited until 1905. Therefore, it was often simply impossible to make high-quality recognition, not because of the quality of scans, but also due to these linguistic features.

If I already had a ready plain text, I checked it against the version of the first edition. I also made notes in the text in order to simplify the subsequent annotation of the XML file.

After that, I converted doc text to tei using the online tool

http://nl.ijs.si/tei/convert/

I would like to take this opportunity to thank Tomaz Erjavec, who maintains this platform. It is really useful, including for Cyrillic texts.

I open the resulting XML file via oxygen, clean up the conversion flaws, check the annotation and upload the file to the github. The duration of work with texts is quite different, sometimes it takes quite a lot of time to find the necessary literary text that would meet all the criteria, and also have an available scan of the first publication. The text processing takes from 6-7 hours to twenty hours. This means that in a week at best I am able to add 2-3 pieces.

SLIDE 6: PROBLEMS: LACK OF SIMILAR PROJECTS, DIFFERENCE BETWEEN FIRST AND THE LAST LIFETIME PUBLICATION

What problems arise during the preparation of the Ukrainian collection? I would name two: I think they are close or similar to many other literatures.

We do have a huge number of novels published in periodicals or as separate editions and available on the shelves of libraries. We also have relatively many digitized versions of these publications available on various sites. However, when it comes to novels already recognized and conversed into plain texts and therefore ready for quick and convenient inclusion into the *ELTeC*-like corpora, their number is rather low.

As we all know, it is desirable to have the text from the first edition for including a novel into *ELTeC*. I must say that this circumstance added much to my workload as a compiler of the Ukrainian collection. (My guess colleagues also paid extra efforts to this). The Ukrainian editorial practice preferred texts from later editions. As a rule, their choice falls in favor of the last lifetime publication containing the author's final intentions.[2] Therefore, most of free available Ukrainian texts represent exactly such version. That means I have to check them in comparison with those from first editions, making minor and significant changes.

But it does not seem right to blame the project for this last limitation. as this is a more general issue. Katherine Bode aptly noted that both close and distant reading dismiss the fact that literary works are not a single and stable texts.[3]

SLIDE 7: THANK YOU

Discussion

- if you were starting again, what would you do differently in the light of your experience so far?

Perhaps the only change in the way of dealing with Ukrainian sub-collection might be if Ukraine has been included in the number of

[2] The principle of *editio ultima* (the last authorial edition) has dominated not only in the scholarly editing of Ukrainian literature, but most of the Slavic literatures (Yesypenko 2020, 50, 60).

[3] Bode 2018, 33.

participants earlier. Maybe I could have involved more colleagues in the process and obtain I mean internal Ukrainian funds for this.

- what do you plan to do for the remainder of the Action?

As was already mentioned, the Ukrainian sub-collection needs to be enriched with a number of additional texts. It's obvious that further, it will be more difficult to comply with the balance of criteria. I also plan to prepare several presentations and one or two articles (individually or in co-authorship) on the material of the ELTeC.

- how should the ELTeC be further developed?

It seems to me that one of the key issues of ELTEC is the incomplete presentation of literatures as a result of the applying principle of national literature - just in the national language.

Therefore, significant and at times extensive branches of literatures written in other languages are cutting off. For the Ukrainian subcollection, these are primarily texts in Russian, as well as in Polish and German, written not only by the authors of a ‘second rank’ but also by those considered the national classics. For instance, the central figure of Ukrainian literarary and broader cultural canon, Ukrainian language poet Taras Shevchenko has no chance to be represented in *ELTEC* with his Russian-language novels.

I mean that a revision of the principle of belonging of texts / authors to one or another literature, based not only on language, can result in a more complete picture.

Another obstacle that limited the range of included Ukrainian texts was the place of publication. This should be only in Europe. That means, novels published, for example, in North America, those published in the relevant time period cannot enter the subcollection. Although, on the other hand, I understand that North American publications were hardly a significant factor in the history of European literatures of the second half of the 19th century - beginning of the 20th c.

Thank you for your attention