Invited CLARIN Workshop: Transc&Anno: A Graphical Tool for the Transcription and On-the-Fly Annotation of Handwritten Documents

Nadezda Okinina and Lionel Nicolas

Institute for Applied Linguistics, Eurac Research

Wednesday, 10 September 2019, 14.30-18.00

This workshop centres on the use of Transc&Anno, a tool for transcription and annotation of handwritten documents. Transcription and annotation of handwritten documents is crucial in learner corpora research as it is an inevitable step in learner corpus creation. Nowadays the majority of learner productions are collected in handwritten form. After that, they are manually transcribed and annotated using generic text editing tools such as XML Mind, Oxygen, Open Office Writer or Microsoft Word. Those tools were not specifically conceived for transcription and annotation of handwritten documents, whereas Transc&Anno was developed exactly for that purpose.

Transc&Anno provides an intuitive environment that is explicitly designed to facilitate the transcription and annotation process for linguists. It facilitates the work of learner corpus creators as well as of its transcribers. Learner corpus creators can easily upload scanned learner essays into Transc&Anno, define an annotation scheme for this collection, give access to transcribers, and monitor their progress. As for transcribers, they enjoy an intuitive working environment with an area displaying the scanned document and a transcription area next to it. They can add annotations to the text they are transcribing on-the-fly, by using the mouse or the hot keys. They have an opportunity to go back to previous versions of their transcriptions as well as to write comments about the texts they are transcribing.

Transc&Anno ensures a high transcription output quality by validating the XML and only allowing predefined tags.

Transc&Anno was created on top of the FromThePage transcription tool developed entirely with standard web technologies – Ruby on Rails, Javascript, HTML, and CSS. We adapted this open-source web-based tool to linguistic research purposes by adding linguistic annotation functionalities to it. Transc&Anno is easily customisable, open source, and available on Github.

For this workshop, we will begin with an overview of the functionalities of Transc&Anno. Next, each participant will upload a small learner corpus (a couple of learner essays) into Transc&Anno and transcribe it. In order to do that, he or she will first create a document collection, then upload a couple of scanned documents into it, then define annotation categories (for example, types of learner errors) and finally transcribe and annotate the uploaded documents. At the end, the participant will be able to download the transcription.

This workshop is supported by CLARIN ERIC.

Nadezda Okinina, Eurac Research

Nadezda is a researcher in the Institute for Applied Linguistics of Eurac Research, where numerous efforts are dedicated to learner corpora research. Nadezda helps her colleagues in corpus creation and exploration by providing technical expertise.

Lionel Nicolas, Eurac Research

Lionel is a senior researcher in Natural Language Processing (NLP) at the Institute for Applied Linguistics of Eurac Research. His research objectives focus, among other things, on practical guidelines for saving efforts when creating or improving linguistic resources, on the automatized transfer of linguistic knowledge between two closely related languages, on the automatized creation and extension of linguistic resources and on the combination of language learning and crowdsourcing for the purpose of creating linguistic resources.