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

Nadezda Okinina, Lionel Nicolas, Verena Lyding
Eurac Research, viale Druso 1, 39100 Bolzano, Italy
nadezda.okinina@eurac.edu, lionel.nicolas@eurac.edu, verena.lyding@eurac.edu

Abstract

We present Transc&Anno, a web-based collaboration tool allowing the transcription of text images and their shallow on-the-fly annotation. Transc&Anno was originally developed in order to address the needs of learner corpora research so as to facilitate digitisation of handwritten learner essays. However, the tool can be used for the creation of any type of corpora requiring transcription and shallow on-the-fly annotation resulting in inline XML. Transc&Anno provides an intuitive environment that is explicitly designed to facilitate the transcription and annotation process for linguists. Transc&Anno ensures a high transcription output quality by validating the XML and only allowing predefined tags. It 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. Thereby we united the convenience of a collaborative transcription tool with its advanced image visualisation, centralised data storage, version control and inter-collaborator communication facilities with the precision of a linguistic annotation tool with its well-developed tag definition possibilities, easy tagging process and tagged-text visualisation. Transc&Anno is easily customisable, open source, and available on Github.

Keywords: transcription tools, annotation tools, learner corpora

1. Introduction

Linguistic resources are of fundamental importance for conducting linguistic research. Some of these resources are originally created in handwritten form but need to be digitised in order to be efficiently exploited. In the absence of automatic handwriting recognition systems capable of handling in a satisfactory fashion the inputted data, they have to be manually transcribed. In many cases, transcription can coincide with shallow annotation. Since the transcription and annotation process can be extremely laborious and time-consuming, Transc&Anno is designed to make transcription and annotation quick and intuitive. Transc&Anno was specifically developed in order to address the needs of learner corpora research (Glaznieks et al., 2014) with regard to the manual transcription of learner texts and their shallow on-the-fly annotation.

In the remainder of this paper, we present the Learner Corpus Infrastructure project that led to the creation of Transc&Anno (Section 2), review existing text transcription and annotation tools (Section 3), explain Transc&Anno's use and important features (Section 4), give some technical information (Section 5), point out future development possibilities and conclude the paper (Section 6).

2. The LCI Project

2.1 Overview

The Learner Corpus Infrastructure project (LCI) conducted at the Institute for Applied Linguistics at Eurac Research aims at creating a stable, systematic, and sustainable infrastructure for the collection, processing and maintenance of learner corpora (Nicolas et al., 2015). *"Computer learner corpora are electronic collections of authentic FL[1]/SL[2] textual data assembled according to explicit design criteria for a particular SLA[3]/FLT[4] purpose. They are encoded in a standardised and homogeneous way and documented as to their origin and provenance"* (Granger, 2002). They serve the needs of language acquisition studies and pedagogy development as well as help the creation of natural language processing tools such as automatic error detection and correction systems (Gamon et al., 2013) or automatic language proficiency level checking systems (Hasan et al., 2008). Due to their often handwritten text basis and non-standard language use, learner corpora are time consuming and difficult to create and therefore scarce. The LCI project implements the following learner corpora creation workflow which can be boldly divided into four steps: (1) transcription and basic on-the-fly error tagging, (2) highly elaborated manual and automatic annotation, (3) corpus statistics and exploration, (4) online corpus publication. One or more specific tools are used for each of these steps. Transc&Anno is developed for the purpose of the first step.

1 foreign language
2 second language
3 second language acquisition
4 foreign language teaching

2.2 Transc&Anno Objectives

Transc&Anno will be the first tool in the processing chain for the creation of digitised learner corpora at Eurac Research. Learner corpora created at Eurac Research are built from handwritten documents stored as digitised images. These text images have to be transformed into computer-readable text. In the absence of automatic handwriting recognition systems capable of accurately recognising collections of short samples of different handwritings containing multiple corrections, they are manually transcribed. While transcribing, transcribers also perform shallow on-the-fly error-tagging.

By shallow annotation, we designate string annotation of different sizes without discontinuities or references to other portions of text. On-the-fly annotation is done at the same time as the transcription. Experience has shown that on-the-fly annotation of elements that do not require deep linguistic knowledge saves time for linguists who afterwards perform detailed linguistic annotation. In that perspective, we do not aim at creating a fully fledged annotation tool that would allow to annotate any type of construction. It is not used for highly elaborated annotation which is a task handled in a subsequent step.

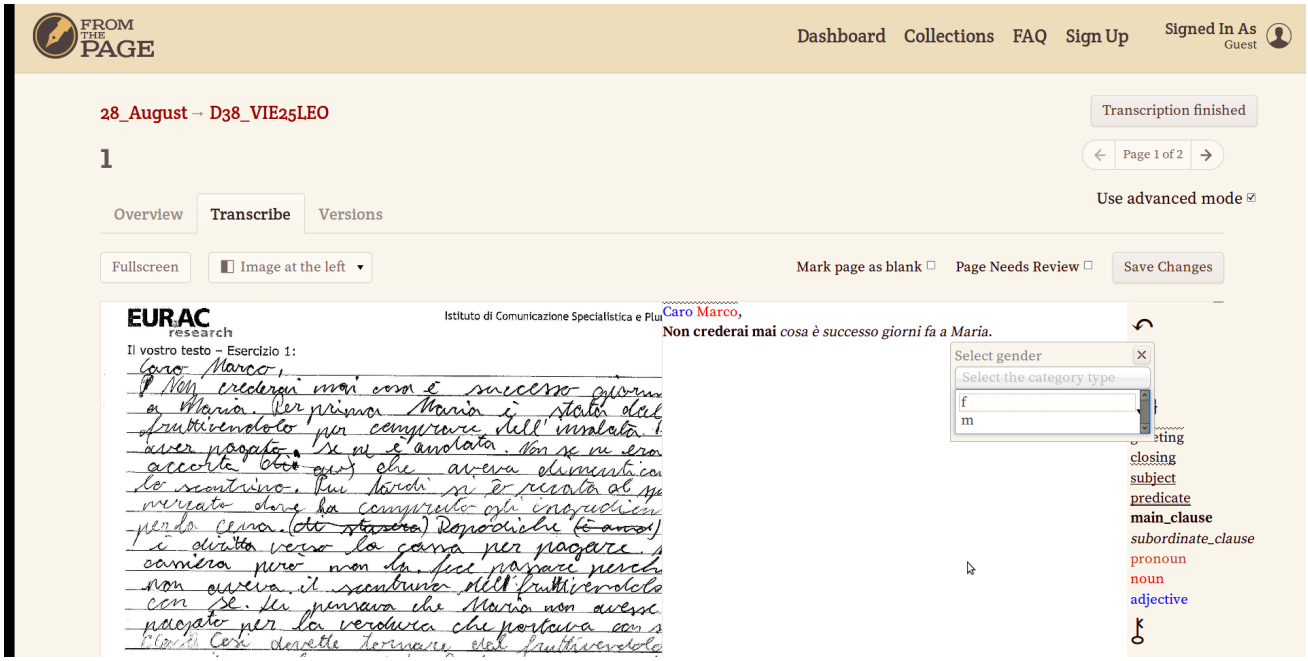

Figure 1: Transcription interface.

3. Related Work

According to our literature review there is no existent tool allowing both transcription of text images and their shallow on-the-fly annotation.

At present time, in learner corpus research, learner texts are typically transcribed using generic text editing tools such as XML Mind (Boyd et al., 2014), Oxygen (Tenford et al., 2006), Open Office Writer or Microsoft Word (Hana et al., 2010). These tools lack integrated collaboration facilities and are therefore a crucial drawback for big projects employing many transcribers and resulting in big quantities of data. Secondly, they do not have intuitive annotation facilities, because they were designed for general text-treatment purposes. That makes the annotation process even more laborious, time-consuming and error prone. Text editors such as Open Office Writer or Microsoft Word do not allow annotation tags' definition thus separating the transcription and annotation processes. They do not contain a text format verification functionality, so that the resulting transcription's format is entirely under the annotator's responsibility. XML editors allow tags' and text format definition, but they have to be hard coded into an XML schema with no front end facilitating this process.

In addition, a range of collaboration transcription tools exists within the field of digital humanities. Some of these tools have non-linguistic annotation functionalities. These tools were mostly conceived in order to create machine-readable versions of historical manuscripts and include, among others, the following projects: Europeana Transcribe[5], Transcribus[6], ProofreadPage[7], Transcribe Bentham[8], FromThePage[9], T-PEN[10] etc. The transcription projects are mainly organised by archives, libraries and museums that call for volunteers' help.

One of the main advantages of these crowdsourcing transcription tools is their web-based nature: they do not need local installation and can be easily used via a web browser. They also include advanced project managing functionalities and support text version control. Transcriptions are saved to a central database in a standardised format. Interpersonal communication is facilitated by the comment writing possibility. Many of the above mentioned tools support advanced image visualising thereby assuring the transcribers' comfort.

All these features can be useful for learner corpora transcription. However, as the final users of the digital texts produced via the above mentioned crowdsourcing tools are historians, the annotation systems of these tools are adapted to bibliographical, paleontological and historical needs. None of these tools offers annotation possibilities necessary for creating learner corpora.

Another tool worth mentioning is the multi-functional corpora editing platform TEITOK allowing document transcription and wiki-style annotation (Janssen 2016). The transcription module of this powerful platform was also inspired by FromThePage, but is not meant to evolve into a more user-friendly form.

Linguists also dispose of a very wide range of sophisticated annotation tools lacking the text transcription functionality such as MMAX (Müller et al., 2001) or WebAnno (Yimam et al., 2013) just to name a few. In creating Transc&Anno our goal is to unite both of these functionalities.

5 https://transcribathon.com
6 https://transkribus.eu/Transkribus
7 https://wikisource.org/wiki/Wikisource:ProofreadPage
8 http://blogs.ucl.ac.uk/transcribe-bentham
9 http://fromthepage.com
10 http://t-pen.org

4. Capabilities and Use

4.1 Basis Tool

Among all the existing tools we chose FromThePage as a basis for our application. First of all, FromThePage is open-source, continually updated and widely promoted by its creator. It is aimed at full-text transcription and not partial information collection or automatically recognised text proofreading. Moreover, it has a very well-developed user interface offering advanced visualisation customisation. It includes an annotation functionality serving historical research needs that can be transformed into a linguistic annotation functionality. FromThePage is built with standard web technologies and is easily customisable.

4.2 User Roles

Transc&Anno's internal structure, inherited from FromThePage, supposes the separation of different user functions. The database administration, the project preparation and the transcription tasks are distributed among different users having different user rights. Accordingly, there are three types of user rights: administrator, collection owner, and transcriber. The administrator is a technician who takes care of the database. The collection owner is the person who creates a collection: he uploads the scanned text to be transcribed and defines the annotation tags. One advantage of Transc&Anno is that the collection owner does not need to have technical knowledge in order to be able to set up the annotation scheme. Finally, the transcriber transcribes and annotates the texts, but is not allowed to modify the annotation system. The users' rights restriction protects sensitive information from accidental modification.

4.3 Setting Up Text Collections

Setting up a collection is the collection owner's responsibility. Documents intended for transcription are handwritten texts in the form of images. The image
loading part of the FromThePage tool has not been modified in Transc&Anno and is used as is. Text images for transcription can be loaded in PDF, PNG, GIF, JPG formats as well as in compressed archives. There are several views that show text images and transcriptions in format of different sizes.

Texts for transcription are organized in collections each of which can have its own annotation scheme and tagging categories.

4.4 Creating the Annotation Scheme

Transc&Anno contains an annotation scheme definition interface intended to be used by the collection owner. FromThePage possesses a menu system for categories creation that we extended in order to allow linguistic phenomena description. FromThePage offers the possibility to add, rename and delete a category. In addition to this, Transc&Anno lets the user define the category's style, description, attributes, its attributes' values and sequences.

The annotation system is defined for each collection of documents and consists of tagging categories and their attributes. The number of categories is unlimited. A category can have no attributes, one or many. An attribute may have a predefined set of values as well as allow the user to enter a new value. Such choice is made by the collection owner who designs the annotation system. Attributes of the same category may be independent or entertain causal relations between each other. Accordingly, the existence of one attribute can depend on the value of another attribute of the same category.

Let's take the example of a German *adjective* category. Depending on the value of its *number* attribute (*singular* or *plural* attribute value), defining the *gender* can be irrelevant. Indeed, plural adjectives have the same declension for all three genders (masculine, feminine and neutral) whereas singular adjectives do not. In linguistic annotations such interdependence between different characteristics is frequent, hence the need for the causal relations mechanism in the attributes' definition.

4.5 Customising Visualisation

FromThePage offers the transcribers some visualisation customisation features: while transcribing a page, the user can zoom in and out on the image text as well as choose its position in the browser window in respect to the transcription text.

FromThePage does not give any possibility to highlight tagged text, because it only allows wiki-style annotations that are performed by typing conventional characters. In order to give visual support to the new annotation functionality we added to Transc&Anno, we included a text highlighting feature. Tagged text highlighting is important for the annotator because it gives him the possibility to spot the existing tags in order to be able to modify or delete them. Each category can be assigned its own text style that will be visible during annotation. The text style includes a font colour, a font style and an underlining style offered by the CSS language. As different tags can overlap, the existence of these three types of highlighting lets two or three tags be visible at the same place. CSS doesn't allow coexistence of more than three styles in the same text fragment, but we intend to overcome this limitation in the future versions of Transc&Anno.

4.6 Transcription and Annotation Process

In this section, we describe the annotation process using Transc&Anno. Instead of typing wiki-style special characters as it is done in FromThePage, the user of Transc&Anno performs annotation using hot keys or buttons triggering pop-up menus.

One advantage of the interface being presented is the opacity of the possibly sophisticated annotation scheme for the transcriber. Once he chooses the tag to apply to a certain portion of text, he is automatically prompted to fill in the corresponding category's attributes' values (Figure 1).

The annotation can be performed by pushing buttons corresponding to categories defined for the current collection or by pressing hot keys. Using exclusively hot keys can considerably increase the annotation speed.
The annotation categories' menu is located on the right side of the screen: each category's name is shown using the style that has been assigned to it (Figure 1). Other buttons allow last operation cancellation, tag deletion, tag modification, hot keys modification. The user can also choose the frequency with which the system automatically saves the transcription.
If the user is unsatisfied with the work he has recently done, he has the possibility to revert to previous versions of the transcription.

4.7 Transcription and Annotation Result

The created resource format is inline TEI compatible XML, but it is shown as highlighted text, no XML markup is visible to the annotator (Figure 1). Despite of being implemented as inline XML, annotation tags can overlap. Overlapping is achieved by closing a tag and reopening it a second time preserving its id: the tagcode.
Figure 2 shows an example of a text tagged with Transc&Anno that corresponds to the text shown in Figure 1.

```
1    <greeting_id15 class="medium-greeting_id15" mode="1"
tagcode="1504275956554">
2      <adjective_id31 class="medium-adjective_id31" gender="m"
mode="1" number="sg"
tagcode="1504276017378">Caro</adjective_id31>
3      <noun_id30 class="medium-noun_id30" gender="m" mode="1"
number="sg" tagcode="1504276025907">Marco</noun_id30>
4    </greeting_id15>
5    ,
6    <div>
7      <main_clause_id19 class="medium-main_clause_id19"
mode="1" tagcode="1504275989612">Non crederai
mai</main_clause_id19>
8      <subordinate_clause_id20 class="medium-
subordinate_clause_id20" mode="1" tagcode="1504275994412">
       cosa e successo giorni fa a
9        <noun_id30 class="medium-noun_id30" gender="f"
mode="1" number="sg"
tagcode="1504276040872">Maria</noun_id30>
10     </subordinate_clause_id20>
       .
11   </div>
```

Figure 2: Transcription result.

This text has been tagged with some pragmatical (the *greeting* tag on lines 1-4), grammatical (the *adjective* and *noun* tags on lines 2, 3, and 9) and syntactical (the *main_clause* and *subordinate_clause* tags on lines 7-10) information. As multiple tags can be assigned to the same string sequence, all these different kinds of information could be added to the text. Most categories used in this annotation do not have attributes, but the noun and adjective categories do: the noun (lines 3 and 9) and adjective (line 2) tags contain the number and gender attributes. Their values have been chosen from a predefined set and therefore have the same form. The class attribute allows text styling via CSS stylesheets and the tagcode attribute uniquely identifies every single annotation. Each annotation category has its own id that distinguishes it from similar categories of other collections.

5. Technical Aspects

Transc&Anno was created on top of the FromThePage transcription tool developed entirely with standard web technologies – Ruby on Rails, Javascript, HTML, and CSS. In order to add the text tagging functionality to FromThePage, we used an open-source Javascript library medium.js[11]. Ruby on Rails interacts with a MySQL database. An automated transcription saving procedure has been set in order to avoid the loss of manual work. Transcriptions are registered in XML format with inline annotations. The validity of the XML is automatically checked before saving.
Transc&Anno is open source, easily customizable and available on Github[12].

6. Conclusion and Future Work

In this paper, we presented Transc&Anno, a web-based collaboration tool allowing the transcription of text images and their shallow on-the-fly annotation. The annotation results in inline TEI compatible XML and can possibly include overlapping annotation tags. Transc&Anno ensures a high transcription output quality by validating the XML and only allowing predefined tags. We intend to use Transc&Anno for transcribing learner corpora of German and Italian.
In creating Transc&Anno, our purpose is not to develop a fully fledged annotation tool, but to provide a user-friendly transcription tool allowing shallow on-the-fly annotation. Nonetheless, we are considering the possibility to extend annotation capacities of the tool by allowing discontinued constituents' tagging as well as pointing from a tag to another portion of text.
Since the visual representation of a transcription is crucial for users' comfort, our next step is to further improve it. We intend to allow an unlimited number of tags be visible in the same text fragment.

Bibliographical References

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B., Vettori, C.: The MERLIN corpus: Learner language and the CEFR, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 1281-1288.

Gamon, M., Chodorow, M., Leacock, C., Tetreault, J.: Using learner corpora for automatic error detection and correction. In Bariller, N., Thompson, P., Automatic Treatment and Analysis of Learner Corpus Data, Amsterdam & Philadelphia: Benjamins, 2013, pp. 127-149.

Glaznieks, A., Nicolas, L., Stemle, E., Abel, A., Lyding, V.: Establishing a Standardised Procedure for Building

11 http://jakiestfu.github.io/Medium.js/docs/

12 https://github.com/commul/fromthepage

Learner Corpora, *Journal of Applied Language Studies*, Vol. 8, 3, 2014, pp. 5-20.

Granger, S.: A Bird's Eye View of Learner Corpus Research. In Granger, S., Hung, J. & Petch-Tyson, S. (eds.), Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, Amsterdam & Philadelphia: Benjamins, 2002, pp. 3-33.

Hana, J., Rosen, A., Škodová, S., Štindlová, B.: Error-tagged Learner Corpus of Czech, *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala, Sweden, 15-16 July, 2010, pp. 11-19.

Hasan, M. M., Khaing, H. O.: Learner Corpus and its Application to Automatic Level Checking using Machine Learning Algorithms, *Proceedings of ECTI-CON*, 2008, pp. 25-28.

Janssen, M.: TEITOK: Text-Faithful Annotated Corpora, *Proceedings of LREC 2016*, Portorož, Slovenia, pp. 4037-4043.

Nicolas, L., Stemle, E., Glaznieks, A., Abel, A.: A Generic Data Workflow for Building Annotated Text Corpora, In Castello, E., Ackerley, K., Coccetta, F., Studies in Learner Corpus Linguistics, 2015, pp. 337-351.

Tenford, K., Meurer, P., Hofland, K.: The ASK Corpus – a Language Learner Corpus of Norwegian as a Second Language, *Proceedings of LREC 2006*, Genoa, Italy, pp. 1821-1824.

Yimam, S.M., Gurevych, I., Eckart de Castilho, R., Biemann C.: WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations, *Proceedings of ACL-2013*, Sofia, Bulgaria, pp. 1-6.

Müller, C., Strube, M.: MMAX: A tool for the annotation of multi-modal corpora, *2nd IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, 2001, pp. 45-50.