A high speed transcription interface for annotating primary linguistic data

Mark Dingemanse, Jeremy Hammond, Herman Stehouwer,
Aarthy Somasundaram, Sebastian Drude
Max Planck Institute for Psycholinguistics
Nijmegen
{mark.dingemanse, jeremy.hammond, herman.stehouwer,
aarthy.somasundaram, sebastian.drude}@mpi.nl

Abstract

We present a new transcription mode for the annotation tool ELAN. This mode is designed to speed up the process of creating transcriptions of primary linguistic data (video and/or audio recordings of linguistic behaviour). We survey the basic transcription workflow of some commonly used tools (Transcriber, BlitzScribe, and ELAN) and describe how the new transcription interface improves on these existing implementations. We describe the design of the transcription interface and explore some further possibilities for improvement in the areas of segmentation and computational enrichment of annotations.

1 Introduction

Recent years have seen an increasing interest in language documentation: the creation and preservation of multipurpose records of linguistic primary data (Gippert et al., 2006; Himmelmann, 2008). The increasing availability of portable recording devices enables the collection of primary data even in the remotest field sites, and the exponential growth in storage makes it possible to store more of this data than ever before. However, without content annotation for searching and analysis, such corpora are of limited use. Advances in machine learning can bring some measure of automation to the process (Tschöpel et al., 2011), but the need for human annotation remains, especially in the case of primary data from undocumented languages. This paper describes the development and use of a new rapid transcription interface, its integration in an open source software framework for multimodality research, and the possibilities it opens up for computational uses of the annotated data.

Transcription, the production of a written representation of audio and video recordings of communicative behaviour, is one of the most time-intensive tasks faced by researchers working with language data. The resulting data is useful in many different scientific fields. Estimates for the ratio of transcription time to data time length range from 10:1 or 20:1 for English data (Tomasello and Stahl, 2004, p. 104), but may go up to 35:1 for data from lesser known and endangered languages (Auer et al., 2010). As in all fields of research, time is a most important limiting factor, so any significant improvement in this area will make available more data and resources for analysis and model building. The new transcription interface described here is designed for carrying out high-speed transcription of linguistic audiovisual material, with built-in support for multiple annotation tiers and for both audio and video streams.

Basic transcription is only the first step; further analysis often necessitates more fine-grained annotations, for instance part of speech tagging or morpheme glossing. Such operations are even more time intensive. Time spent on further annotations generally goes well over a 100:1 annotation time to media duration ratio[1] (Auer et al., 2010).The post-transcription work is also an area with numerous possibilities for further reduction of annotation time by applying semi-automated annotation suggestions, and some ongoing work

[1] Cf. a blog post by P.K.Austin http://blogs.usyd.edu.au /elac/2010/04/how_long_is_a_piece_of_string.html.

Proceedings of the 6th EACL Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pages 7–12,
Avignon, France, 24 April 2012. ©2012 Association for Computational Linguistics

to integrate such techniques in our annotation system is discussed below.

2 Semi-automatic transcription: terminology and existing tools

Transcription of linguistic primary data has long been a concern of researchers in linguistics and neighbouring fields, and accordingly several tools are available today for time-aligned annotation and transcription. To describe the different user interfaces these tools provide, we adopt a model of the transcription process by (Roy and Roy, 2009), adjusting its terminology to also cover the use case of transcribing sign language. According to this model, the transcription of primary linguistic data can be divided into four basic subtasks: 1) *find* linguistic utterances in the audio or video stream, 2) *segment* the stream into short chunks of utterances, 3) *play* the segment, and 4) *type* the transcription for the segment.

Existing transcription tools implement these four steps in different ways. To exemplify this we discuss three such tools below. All three can be used to create time-aligned annotations of audio and/or video recordings, but since they have different origins and were created for different goals, they present the user with interfaces that differ quite radically.

Transcriber (Barras et al., 2001) was "designed for the manual segmentation and transcription of long duration broadcast news recordings, including annotation of speech turns, topics and acoustic condition" (Barras et al., 2001, p. 5). It provides a graphical interface with a text editor at the top and a waveform viewer at the bottom. All four subtasks from the model above, FSPT, are done in this same interface. The text editor, where Segmenting and Typing are done, is a vertically oriented list of annotations. Strengths of the Transcriber implementation are the top-to-bottom orientation of the text editor, which is in line with the default layout of transcripts in the discipline, and the fact that it is possible to rely on only one input device (the keyboard) for all four subtasks. Weaknesses are the fact that it does not mark annotation ends, only beginnings,and that it treats the data as a single stream and insists on a strict partitioning, making it difficult to handle overlapping speech, common in conversational data (Barras et al., 2001, p. 18).

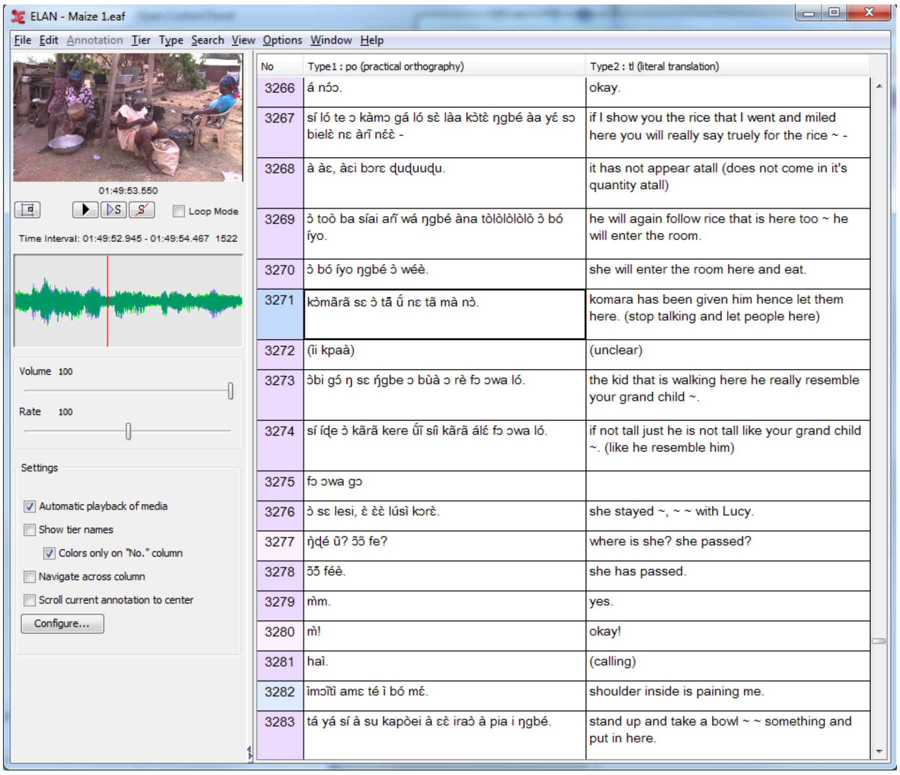
BlitzScribe (Roy and Roy, 2009) was developed in the context of the Human Speechome project at the MIT Media Lab as a custom solution for the transcription of massive amounts of unstructured English speech data collected over a period of three years (Roy et al., 2006). It is not available to the academic community, but we describe it here because its user interface presents significant improvements over previous models. BlitzScribe uses automatic speech detection for segmentation, and thus eliminates the first two steps of the FSPT model, Find and Segment, from the user interface. The result is a minimalist design which focuses only on Playing and Typing. The main strength of BlitzScribe is this streamlined interface, which measurably improves transcription speed — it is about four times as fast as Transcriber (Roy and Roy, 2009, p. 1649). Weaknesses include its monolingual, speech-centric focus, its lack of a mechanism for speaker identification, and its single-purpose design which ties it to the Human Speechome project and makes it unavailable to the wider academic community.

ELAN (Wittenburg et al., 2006) was developed as a multimedia linguistic annotation framework. Unlike most other tools it was built with multimodal linguistic data in mind, supporting the simultaneous display and annotation of multiple audio and video streams. Its data model is tier-based, with multiple tiers available for annotations of different speakers or different modalities (e.g. speech and gesture). Its strengths are its support for multimodal data, its handling of overlapping speech, its flexible tier structure, and its open source nature. Its noted weaknesses include a steep learning curve and a user interface that was, as of 2007, "not the best place to work on a 'first pass' of a transcript" (Berez, 2007, p. 288).

The new user interface we describe in this paper is integrated in ELAN as a separate "Transcription Mode", and was developed to combine the strengths of existing implementations while at the same time addressing their weaknesses. Figure 1 shows a screenshot of the new transcription mode.

3 Description of the interface

From the default Annotation Mode in ELAN, the user can switch to several other modes, one of which is Transcription Mode. Transcription Mode displays annotations in one or more columns. A column collects annotations of a single type. For


ELAN - Maize 1.eaf
File Edit Annotation Tier Type Search View Options Window Help
01:49:53.550
Loop Mode
Configure...

Time Interval: 01:49:52.945 - 01:49:54.467 1522

Volume 100

Rate 100

Settings

☑ Automatic playback of media
□ Show tier names
☑ Colors only on "No." column
□ Navigate across column
□ Scroll current annotation to center

No	Type1 : po (practical orthography)	Type2 : tl (literal translation)
3266	á nɔ́ɔ.	okay.
3267	sí ló te ɔ kàmɔ gá ló sɛ̀ làa kɔ̀tɛ̀ ŋgbé àa yɛ́ sɔ bielɛ̀ nɛ àrĩ nɛ́ɛ̀ -	if I show you the rice that I went and miled here you will really say truely for the rice ~ -
3268	à àɛ, àɛi bɔrɛ ɖuɖuuɖu.	it has not appear atall (does not come in it's quantity atall)
3269	ɔ̀ toò ba síai arĩ wá ŋgbé àna tòlòlòlòlò ɔ̀ bó íyo.	he will again follow rice that is here too ~ he will enter the room.
3270	ɔ̀ bó íyo ŋgbé ɔ̀ wéè.	she will enter the room here and eat.
3271	kɔ̀mãrã sɛ ɔ̀ tã̄ ű nɛ tã mà nɔ̀.	komara has been given him hence let them here. (stop talking and let people here)
3272	(ĩi kpaà)	(unclear)
3273	ɔ̀bi gɔ́ ŋ sɛ ŋ́gbe ɔ bùà ɔ rè fɔ ɔwa ló.	the kid that is walking here he really resemble your grand child ~.
3274	sí íɖe ɔ̀ kãrã kere ű̃ĩ sĩi kãrã álɛ́ fɔ ɔwa ló.	if not tall just he is not tall like your grand child ~. (like he resemble him)
3275	fɔ ɔwa gɔ	
3276	ɔ̀ sɛ lesi, ɛ̀ ɛ̀ɛ̀ lúsì kɔrɛ̀.	she stayed ~, ~ ~ with Lucy.
3277	ŋ̀ɖé ũ? ɔ̃ɔ̃ fe?	where is she? she passed?
3278	ɔ̃ɔ̃ féè.	she has passed.
3279	m̀m.	yes.
3280	m̀!	okay!
3281	haì.	(calling)
3282	ìmɔĩ̀tì amɛ té ì bó mɛ́.	shoulder inside is paining me.
3283	tá yá sí à su kapɔ̀ei à ɛɛ̀ iraɔ̀ à pia i ŋgbé.	stand up and take a bowl ~ ~ something and put in here.

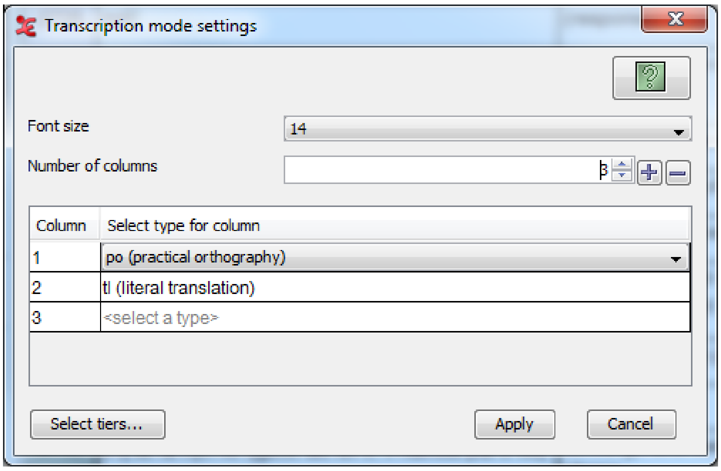
Figure 1: The interface of the transcription mode, showing two columns: transcriptions and the corresponding translations.

instance, the first column in Figure 1 displays all annotations of the type "practical orthography" in chronological order, colour-coding for different speakers. The second column displays corresponding, i.e., time aligned, annotations of the type "literal translation". Beside the annotation columns there is a pane showing the data (video and/or audio stream) for the selected utterance. Below it are basic playback parameters like volume and rate, some essential interface settings, and a button "Configure" which brings up the column selection dialog window. We provide an example of this preference pane in Figure 2.

The basic organisation of the Transcription Mode interface reflects its task-oriented design: the annotation columns occupy pride of place and only the most frequently accessed settings are directly visible. Throughout, the user interface is keyboard-driven and designed to minimise the number of actions the user needs to carry out. For instance, selecting a segment (by mouse or keyboard) will automatically trigger playback of that segment (the user can play and pause using the Tab key). Selecting a grey (non-existent) field in a dependent column will automatically create an annotation. Selection always opens up the field for immediate editing. Arrow keys as well as user-configurable shortcuts move to adjacent fields.

ELAN Transcription Mode improves the transcription workflow by taking apart the FSPT model and focusing only on the last two steps: Play and Type. In this respect it is like BlitzScribe; but it is more advanced than that and other tools in at least two important ways. First, it is agnostic to the type of data transcribed. Second, it does not presuppose monolingualism and is ready for multilingual work. It allows the display of multiple annotation layers and makes for easy navigation between them.Further, when transcription is done with the help of a native speaker it allows for them to provide other relevant information at the same time (such as cultural background explanations) keeping primary data and meta-data time aligned and linked.

Some less prominently visible features of the user interface design include: the ability to reorder annotation columns by drag and drop; a toggle for the position of the data streams (to the left or to the right of the annotation columns); the ability to detach the video stream (for instance for display on a secondary monitor); the option to show names (i.e. participant ID's) in the flow of anno-



Font size: 14

Number of columns: 3

Column	Select type for column
1	po (practical orthography)
2	tl (literal translation)
3	<select a type>

Figure 2: The interface of the transcription mode; the configuration dialog.

tations or to indicate them by colour-coding only; the option to keep the active annotation centered; and settings for font size and number of columns (in the ”Configure” pane). These features enable the user to customise the transcription experience to their own needs.

The overall design of Transcription Mode makes the process of transcription as smooth as possible by removing unnecessary clutter, foregrounding the interface elements that matter, and enabling a limited degree of customisation. Overall, the new interface has realised significant speedups for many people[2]. User feedback in response to the new transcription mode has been overwhelmingly positive, e.g., the members of mailing lists such as the Resource Network for Linguistic Diversity[3].

4 A prerequisite: semi-automatic segmentation

As we noted in the introduction, the most important step before transcription is that of segmentation (steps Find and Segment in the FSPT model). Segmentation is a large task that involves subdividing the audio or video stream in, possibly overlapping, segments. The segments each denote a distinct period of speech or any other communicative act and each segment is commonly assigned to a specific speaker. This step can potentially be sped up significantly by doing it semi-automatically using pattern recognition techniques, as pursued in the AVATecH project (Auer et al., 2010).

In the AVATecH project, audio and video streams can be sent to detection components called ‘recognisers’. Some detection components accept the output of other recognisers as additional input, next to the audio and/or video streams, thus facilitating cascaded processing of these streams. Amongst the tasks that can be performed by these recognisers is the segmentation of audio and video, including speaker assignment.

A special challenge for the recognisers in this project is the requirement of language independence (in contrast to the English-only situation in the Human Speechome project that produced Blitzscribe(Roy et al., 2006)). The recognisers should ideally accommodate the work of field linguists and other alike researchers and therefore cannot simply apply existing language and acoustic models. Furthermore, the conditions that are encountered in the field are often not ideal, e.g., loud and irregular background noises such as those from animals are common. Nevertheless, automatic segmentation has the potential to speed up the segmentation step greatly.

5 Future possibilities: computational approaches to data enrichment

While a basic transcription and translation is essential as a first way into the data, it is not sufficient for many research questions, linguistic or

[2] Including ourselves, Jeremy Hammond claims that: “*Based on my last two field work trips, I am getting my transcription time down below that of transcriber (but perhaps not by much) but still keeping the higher level of data that ELANs tiers provide - probably around 18-20 hours for an hour of somewhat detailed trilingual annotation.*”

[3] www.rnld.org

otherwise. Typically a morphological segmentation of the words and a labelling of each individual morph is required. This level of annotation is also known as basic glossing (Bow et al., 2003b; Bow et al., 2003a).

Automatically segmenting the words into their morphological parts, without resorting to the use of pre-existing knowledge has seen a wide variety of research (Hammarström and Borin, 2011). Based on the knowledge-free induction of morphological boundaries the linguist will usually perform corrections. Above all, a system must learn from the input of the linguist, and must incorporate it in the results, improving the segmentation of words going forward. However, it is well known from typological research that languages differ tremendously in their morphosyntactic organisation and the specific morphological means that are employed to construct complex meanings (Evans and Levinson, 2009; Hocket, 1954).

As far as we know, there is no current morphological segmentation or glossing system that deals well with all language types, in particular inflectional and polysynthetic languages or languages that heavily employ tonal patterns to mark different forms of the same word. Therefore, there is a need for an interactive, modular glossing system. For each step of the glossing task, one would use one, or a set of complementary modules. We call such modules “annotyzers”. They generate content on the basis of the source tiers and additional data, e.g. lexical data (or learnt states from earlier passes). Using such modules will result in a speedup for the researcher. We remark that there are existing modular NLP systems, such as GATE(Cunningham et al., 2011), however these are tied to different workflows, i.e., they are not as suitable for the multimodal multi-participant annotation process.

Currently a limited set of such functionality is available in Toolbox and FLEX. In the case of both Toolbox and FLEX the functionality is limited to a set of rules written by the linguist (i.e. in a database-lookup approach). Even though the ELAN modules will offer support for such rules, our focus is on the automation of machine-learning systems in order to scale the annotation process.

Our main aim for the future is to incorporate learning systems that support the linguists by improving the suggested new annotations on the bases of choices the linguist made earlier. The goal there is, again, to reduce annotation time, so that the linguist can work more on linguistic analysis and less on annotating. At the same time, a working set of annotyzers will promote more standardised glossing, which can then be used for further automated research, cf. automatic treebank production or similar (Bender et al., 2011).

6 Conclusions

The diversity of the world’s languages is in danger. Perhaps user interface design is not the first thing that comes to mind in response to this sobering fact. Yet in a field that increasingly works with digital annotations of primary linguistic data, it is imperative that the basic tools for annotation and transcription are optimally designed to get the job done.

We have described Transcription Mode, a new user interface in ELAN that accelerates the transcription process. This interface offers several advantages compared to similar tools in the software landscape. It automates actions wherever possible, displays multiple parallel information and annotation streams, is controllable with just the keyboard, and can handle sign language as well as spoken language data. Transcription Mode reduces the required transcription time by providing an optimised workflow.

The next step is to optimise the preceding and following stages in the annotation process. Preceding the transcription stage is segmentation and speaker labelling, which we address using automatic audio/video recogniser techniques that are independent of the language that is transcribed. Following transcription, we aim to support basic glossing (and similar additional annotations based on transcriptions) with a modular software architecture. These semi-automated steps lead to further time savings, allowing researchers to focus on the analysis of language data rather than on the production of annotations.

The overall goal of the developments described here is to help researchers working with primary language data to use their time more optimally. Ultimately, these improvements will lead to an increase in both quality and quantity of primary data available for analysis. Better data and better analyses for a stronger digital humanities.

References

Eric Auer, Peter Wittenburg, Han Sloetjes, Oliver Schreer, Stefano Masneri, Daniel Schneider, and Sebastian Tschöpel. 2010. Automatic annotation of media field recordings. In *Proceedings of the ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2010)*, pages 31–34.

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2001. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22, January.

Emily M. Bender, Dan Flickinger, Stephan Oepen, and Yi Zhang. 2011. Parser evaluation over local and non-local deep dependencies in a large corpus. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 397–408, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Andrea L. Berez. 2007. Review of EUDICO linguistic annotator (ELAN). *Language Documentation & Conservation*, 1(2):283–289, December.

Catherine Bow, Baden Hughes, and Steven Bird. 2003a. A four-level model for interlinear text.

Cathy Bow, Baden Hughes, and Steven Bird. 2003b. Towards a general model of interlinear text. In *Proceedings of EMELD Workshop 2003: Digitizing and Annotating Texts and Field Recordings*. Lansing MI, USA.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. 2011. *Text Processing with GATE (Version 6)*.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(05):429–448.

Jost Gippert, Nikolaus P. Himmelmann, and Ulrike Mosel, editors. 2006. *Essentials of language documentation*. Mouton de Gruyter, Berlin / New York.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. To Appear in Computational Linguistics.

Nikolaus P. Himmelmann. 2008. Reproduction and preservation of linguistic knowledge: Linguistics' response to language endangerment. In *Annual Review of Anthropology*, volume 37 (1), pages 337–350.

Charles F. Hocket. 1954. Two models of grammatical description. *Word 10*, pages 210–234.

Chris Rogers. 2010. Review of fieldworks language explorer (flex) 3.0. In *Language Documentation & Conservation 4*, pages 1934–5275. University of Hawai'i Press.

Brandon C. Roy and Deb Roy. 2009. Fast transcription of unstructured audio recordings. In *Proceedings of Interspeech 2009*, Brighton, England.

Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon C. Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, Micheal Levit, and Peter Gorniak. 2006. The human speechome project. In Paul Vogt, Yuuga Sugita, Elio Tuci, and Chrystopher Nehaniv, editors, *Symbol Grounding and Beyond*, volume 4211 of *Lecture Notes in Computer Science*, pages 192–196. Springer, Berlin / Heidelberg.

Michael Tomasello and Daniel Stahl. 2004. Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, 31(01):101–121.

Sebastian Tschöpel, Daniel Schneider, Rolf Bardeli, Peter Wittenburg, Han Sloetjes, Oliver Schreer, Stefano Masneri, Przemek Lenkiewicz, and Eric Auer. 2011. AVATecH: Audio/Video technology for humanities research. *Language Technologies for Digital Humanities and Cultural Heritage*, page 86.

Peter Wittenburg, Hennie Brugman, Albert Russel, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of LREC 2006*.