Deconstructing Human-assisted Video Transcription and Annotation for Legislative Proceedings

THORSTEN RUPRECHTER, Graz University of Technology, Austria

FOAAD KHOSMOOD, California Polytechnic State University, USA

CHRISTIAN GUETL, Graz University of Technology, Austria

Legislative proceedings present a rich source of multidimensional information that is crucial to citizens and journalists in a democratic system. At present, no fully automated solution exists that is capable of capturing all the necessary information during such proceedings. Even if professional-quality automated transcriptions existed, other tasks such as speaker or rhetorical position identifications are not fully automatable. This work focuses on improving and evaluating the transcription software used by the Digital Democracy initiative, named Transcription Tool. Human transcribers work to up-level state legislative proceedings using this tool. Five phases of tool improvements are introduced and for each phase, the resulting change in efficiency is measured. We investigate over 12,000 individual transcription sessions (2,300 hours of video), where each session is the record of one bill discussion. A set of about 3,200 sessions belonging to a single cohort of 20 transcribers is further evaluated. Through introduction of new tool features, human-assisted transcription efficiency can be improved by 19.4% over five phases. Furthermore, investigation into transcriber usage patterns reveals that transcription time is composed of passive time, speaker identification, text correction, tool startup, as well as splitting and merging utterances. We analyze and rank these as a contribution.

CCS Concepts: • **Applied computing** → **E-government**; • **Human-centered computing** → *Information visualization*; *User interface management systems*;

Additional Key Words and Phrases: Government transparency, transcription system, transcription software, speech annotation, speaker recognition, speaker detection, user behavior, log analysis

ACM Reference format:

Thorsten Ruprechter, Foaad Khosmood, and Christian Guetl. 2020. Deconstructing Human-assisted Video Transcription and Annotation for Legislative Proceedings. *Digit. Gov. Res. Pract.* 1, 3, Article 19 (November 2020), 24 pages.
<https://doi.org/10.1145/3395316>

1 INTRODUCTION

In addition to the federal legislature, over 50 state and territorial legislatures operate in the USA, holding considerable power. States are situated between federal and local government in the political system of the United

This study was funded in part by the Cal Poly Institute for Advanced Technology and Public Policy (IATPP). Also supported in part by the TU Graz Open Access Publishing Fund.

Authors' addresses: T. Ruprechter and C. Guetl, Graz University of Technology, Graz, Austria; emails: {ruprechter, c.guetl}@tugraz.at; F. Khosmood, California Polytechnic State University, San Luis Obispo; email: foad@calpoly.edu.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2020 Copyright held by the owner/author(s).

2639-0175/2020/11-ART19

<https://doi.org/10.1145/3395316>

States. State legislatures handle many important issues touching everyday life of citizens, such as welfare, education, and law enforcement.

Unfortunately, monitoring all state legislative activity and therefore holding representatives to account is currently difficult. It can be a time-consuming task for the general public to gain insight into the workings of state government. Official records about the content of state legislative hearings are only sparsely available and no official transcriptions are produced. Due to the massive number of laws passed and amended each year in these hearings, lobbyists are interested in influencing state legislators. Lobbyists can generally afford to travel to state capitols and be in the room when important discussions are had. Without a platform offering insight into the content of legislative hearings, the public is shut out of an important part of the democratic process. There are also implications for the free press, which has had a traditional role of recording important information and representing the public interest. Recent industry consolidations have reduced total coverage for state-level legislative news [18, 42, 52, 65]. Such media organizations have difficulty discovering contents of talks and political negotiations without proper textual records being created. The stance of lawmakers discussing an issue might be hard to guess without knowing details about the proceedings of state house or senate.

This work is based on a master's thesis by Ruprecht [62]. Some of the preliminary results were previously published as a conference poster [63]. This article further elaborates upon previous work by clearly defining metrics, introducing more sophisticated approaches for analysis, and using a larger dataset.

1.1 Digital Democracy Initiative

The Digital Democracy initiative, which was introduced in 2015, attempts to alleviate the above-mentioned access difficulties [4, 5, 14, 22, 23, 28]. The Digital Democracy initiative succeeds in creating a model online platform featuring professional-grade, searchable transcriptions synced with video content. This system brings transparency into state legislative hearings and government proceedings that would otherwise not be accessible to the broader public. In general, digital democracy describes the use of information and communication technology (ICT) for enhancing democracy and the participation of citizens in politics [19]. Other authors simply define digital democracy as “the practice of democracy using digital tools and technologies” [39]. In contrast to these general definitions, in this work, we use the term “Digital Democracy” to refer to the aforementioned transparency initiative.

With Digital Democracy, citizens, journalists, and researchers can more easily gain information about the content of legislative hearings and debates. Digital Democracy processes information for four states: California, Florida, New York, and Texas. In addition to giving insight into the proposal of new bills and laws, Digital Democracy also presents a chance for the general public to monitor lobbyists, lawmakers, and advocates. All information can be searched, queried, and made subject of email alerts. Contents of legislative hearings are presented through interactive transcripts, each of which contains valuable information about a hearing, for example the video of the hearing, spoken text, speaker information, and discussed bills. Figure 1 shows how transcripts are displayed on the Digital Democracy website.^{1,2}

1.2 Research Goals and Contributions

The Digital Democracy initiative uses a human-assisted approach for generating transcription texts and metadata [4]. While automatic transcription might be sufficient in other areas, the legislative domain requires high-quality, professional transcripts. This is achieved by human transcribers manually up-leveling automated transcriptions and performing annotations such as speaker and position identification. As a result, transcripts are high quality and enriched with further metadata, enabling the general public to access flawless transcripts and gain a complete understanding of the underlying political context.

¹<https://iatpp.calpoly.edu/digital-democracy>.

²As of 2019, the Digital Democracy website has suspended operations.

Fig. 1. Interactive transcript on the Digital Democracy website.

Transcribers use the Transcription Tool, a software developed by the Digital Democracy project team to enhance transcripts (see Section 2.4). However, manual correction of transcription texts and finding speaker assignments is a tedious process. It takes human transcribers a formidable amount of time to work through transcripts. This bottleneck created by the manual transcription process leads to a considerable obstacle for Digital Democracy, both in terms of monetary cost as well as time delay before final transcripts are available. Furthermore, at the time of initiating this research there is no way for the initiative's administrators to evaluate ongoing costs or estimate future workload. Therefore, tool improvements and metrics enabling detailed analysis of the transcription process must be introduced. To address these issues, three major contributions are made.

1.2.1 Tool Improvements. The Digital Democracy Transcription Tool, operated by human transcribers, poses a cost factor for the Digital Democracy initiative [61]. Data-processing overhead and necessary human labor is high. Therefore, introducing new tool features represents a main priority for the initiative.

1.2.2 Definition of Efficiency Metrics. Efficiency of the tool and newly introduced features must be analyzed and quantified. Prior to this work, there was no reliable way for the Digital Democracy initiative to measure performance changes that new tool releases cause. Nevertheless, workflow disturbances or disruptions in the transcription process introduced by new releases must be detectable. Finally, it should be possible to derive future transcription costs from efficiency results.

1.2.3 Transcriber Interaction Analysis. The lack of clarity regarding concrete composition of transcription time signifies another issue. It is up to debate which aspects of transcription work yield the largest efficiency gain. To further improve performance, the crucial question of which transcriber interactions contribute how much to overall transcription time must be answered.

2 BACKGROUND AND RELATED WORK

In this section, literature review for various areas are presented. First, we give insight into the terms government transparency and open government. Second, existing government transparency efforts and websites that expose legislative proceedings are listed. Third, the tools used to further enrich this data are introduced. This survey concentrates on major tools available for transcription and annotation purposes similar to those of the Digital Democracy initiative. Last, the Digital Democracy Transcription Tool is explained in detail.

2.1 Government Transparency and Open Government

In the era of virally spreading information, it is sometimes hard for the general public to distinguish between facts and misinformation, the so-called “fake news” problem. Confirming claims made by politicians and actually holding them accountable can be a difficult task [4]. Information and data necessary to achieve this are often unstructured, for example in the form of interviews or videos of political debates [24]. A possible solution to this problem is to expand access for citizens, allowing direct verification of crucial legislative information and therefore providing more transparency into the political system [28].

Government transparency has been described as “*publicizing of incumbent policy choices*” [16]. Dawes and Helbig [11] put it more precisely when they state that it means providing possibilities to “*hold elected officials and public agencies accountable for their decisions and actions.*” In the era of the Internet, this translates to making government information easier accessible online [24]. Utilizing ICT to achieve better government transparency is frequently associated with the term “open government.”

Open government is a phrase often used in the context of government transparency, sometimes even as a synonym [35]. Wijnhoven et al. [78] define it as having two dimensions. On the one hand, it should provide access to information. On the other hand, it should also enable citizens to participate in decision making. These authors also mention that the idea of open government is often simplified to collaboration of the public sector with the crowd. According to Janssen et al. [21], open government should interact with its environment and act as an open system. Furthermore, the authors agree with Wijnhoven et al. [78] that it must allow effective oversight by promoting transparency and participation. Khosmood et al. [24] as well as Dawes and Helbig [11] also mention the characteristic goals of the open government principles outlined by the US government under President Obama: collaboration, participation, and transparency.

2.2 State of Government Transparency around the World

Over the past few years, governments all around the world attempted improvements in open government policies and government transparency. Recent studies investigated such improvements for multiple countries, for example Brazil [12, 27, 31, 33], China [8], Denmark [36], Ecuador [37], Estonia [44], Germany [66], Malaysia [41], Mexico [55], Nigeria [40], South Africa [34], South Korea [79], Sweden [69], and Spain [2]. However, even though such prevalent open government literature exists, assessing the ability of citizens to access interactive transcripts of legislative hearings is a tedious task.

To the best of our knowledge, no elaborate survey has been conducted yet for comparing transcription systems and citizen accessibility to interactive transcripts for multiple parts of the world. We therefore investigate some of the transcription systems that allow monitoring of lawmakers in the congress of the USA, the Commonwealth of Nations, and selected countries in Europe. Government transparency in US state governments was already briefly addressed in Section 1. Table 1 shows a summary of our survey, giving an overview of how governments and parliamentary systems around the world provide information to the public.

The situation regarding government transparency in US state governments is problematic, as already mentioned in Section 1. A study by the CALPIRG Education fund exposes the lack of online access to information regarding spending information in many states over multiple years [10, 59, 70]. States ranked especially low were Alaska, California, and Idaho. Although some states, such as California, attempted to increase transparency

Table 1. Government Transparency for Selected Countries around the World, as of Section 2.2

Government	Records about Legislative Hearings				Search Filters and Scope			
	Transcript	Videos	Linked	Votes	Full-Text	Title	Date	Speaker
European Union	OL	✓	✓	✓		✓	✓	✓
Australia	✓	✓		✓	✓	✓	✓	✓
New Zealand	✓	✓	✓	✓		✓	✓	✓
Canada	✓	✓	✓	✓	✓	✓	✓	✓
United Kingdom	✓	✓	✓	✓	✓	✓	✓	✓
Switzerland	OL	✓		✓	✓	✓	✓	✓
Federal Government of the USA	✓	✓		✓	✓	✓	✓	✓
State of California		✓						
DigitalDemocracy (CA, FL, NY, TX)	✓	✓	✓	✓	✓	✓	✓	✓
Austria	OL	< 7 days		TP	✓	✓	✓	

OL = non-English records that are only available in the originally spoken language; TP = third-party sites.

through propositions, records are lacking in quality [29]. Some publicly available resources to gain insight into hearings exist, but most of them do not offer rich functionality like full-text search in transcripts [4, 24, 61]. On a separate note, to this day some US states violate the “Americans with Disabilities Act” (ADA), which became law in 1990, by not providing any captioning on live-streamed or archived proceedings [57].

In contrast to American state governments, searchable transcripts are available for hearings of the federal government. C-SPAN is a cable TV channel that broadcasts congressional hearings and creates interactive transcripts that are linked to the video source [7]. Hearing records include vote, bill, and speaker data. This archive provided by C-SPAN allows for search in federal proceedings of Congress, the Executive Branch, as well as the Supreme Court. In addition, citizens can access proceedings and debates online through the Congressional Record website [30].

The British parliament provides official transcripts of parliamentary debates in a searchable manner. These records, officially called “Hansard,” are accessible to the public via the parliament’s online presence [51]. During a sitting day, an online version of this day’s proceedings is published gradually, with the full Hansard being available the next morning. Similar to Digital Democracy, the website provides a search interface that allows to query transcripts for specific terms. It directly links the search results to the exact position in both textual transcripts and video recordings. Some of the Commonwealth countries also implement searchable Hansards. For example, the Parliament of Australia [49], the New Zealand Parliament [43], and the Parliament of Canada [50] all offer Hansards on their websites while creating links between plain text records and corresponding videos. They also integrate advanced interfaces for multi-purpose search.

The European Union provides online access to debate videos and verbatim texts [73]. However, segments of speech in textual transcripts of debates are not translated to all languages. This effectively means that debate transcripts only contain speaker utterances in the original spoken language, which is mostly not English. Although speeches contained in reports are linked to video sources and therefore translated audio, no full text search of plenary sittings is available. However, agendas, reports, adopted texts, votes, and audio recordings of debates are made available in all official languages of the EU.

Similar to the EU, the parliament of Switzerland only creates textual transcripts in the original language of the speaker. However, a searchable official bulletin is made available that allows users to issue full-text queries on statements of council members (in their original language) and provides links to transcripts and videos [74].

In Austria, stenographic protocols of plenary sittings of the parliament are made available online in German [72]. The full-text of these records can be searched using the advanced search functionality provided on the website of the Austrian Parliament, but there is no direct link from found text to video recordings [71]. Currently, there also exists no permanent video archive of past parliamentary sittings. However, sittings are streamed live on the parliamentary website and are stored for seven-day on-demand access in the archive of the Austrian national public service broadcaster [47]. Most of the functionality and information is only available in German. Since the official services of the parliament in Austria do not keep track of votes, the website Addendum.org provides this service to the public by manually counting the votes of the representatives attending a sitting [56].

2.3 Transcription Editing and Annotation Tools

Although automatic algorithms nowadays manage to generate reasonable transcripts as well as additional contextual data from digital sources, human-assistance is still mostly necessary to ensure correctness on a professional level [4]. Due to this, tools for both research and commercial use emerged in past years. However, transcription software differs strongly from field to field.

Entity tagging or annotation tools such as those introduced by Stenetorp et al. [67], Papazian et al. [48], or Widlöcher and Mathet [77] focus on linking entity metadata to plain text. They allow users to create links between existing or newly created entities as well as their text occurrences. In addition, further details such as description of relationships can be specified. Although such annotation tools fulfill a different purpose than transcription tools do overall, both generate metadata for plain text records. Entity information created by transcription tools can be used to derive more in-depth information, such as speaker identification and alignment.

Audio transcription tools must allow the user to pause, rewind, or in other ways manipulate the currently investigated files while editing transcripts. Even though audio tracks are a different media type than video, many audio transcription systems are similar to those handling video in that the initial text presented to the user is also created by automatic speech recognition [3, 6, 32, 60, 76]. Some tools provide the option to investigate waveform representation of played audio, such as the one implemented by Luz et al. [32]. In contrast to plain audio transcription, videos can be used to derive additional data from information such as physical on-screen appearance of speakers.

Tools used to create and edit video subtitles, captions, or transcripts are those most similar to the Digital Democracy Transcription Tool (see Section 2.4). One of the earliest approaches of providing a software package for creating and synchronizing video transcripts was introduced by Nivre et al. [45]. Seps [64] created *NanoTrans*, a tool that allows for creation of both textual and phonetic transcriptions, including tracks in waveform. Deshpande et al. [13] developed *ICS Caption Editor*, a crowd-sourced caption generator that enables students to collaborate on correction of preprocessed captions generated from lecture videos.

Besides these research-oriented transcription tools, commercial software exists that provides assistance in editing video transcripts, captions, or subtitles. The automatic transcription services provided by cielo24 [9] come with a sophisticated transcription editor. This tool is the one most similar to the Digital Democracy Transcription Tool functionality-wise. Some of the features of this editor are: allowing users to navigate between utterances using a button panel, jumping to specific parts of a video by entering a timestamp, adding speaker information to utterances, the option of auto-pausing the video player while a user is typing, hotkeys for navigating the transcript, and shortcuts for adding sound tags.

In some cases, it is difficult to clearly distinguish between tool categories. Software used to transcribe audio may also be useful to create and export subtitles or video captions. Dependent on exact tool features and requirements, the same could be said about subtitle, caption, and video transcription editors. Example use cases for such editors include: records of lectures or educational videos [13, 64], subtitles and captions for television, movies, or presentations [9], rich transcriptions of legislative hearings [61], or any other arbitrary video resource.

Fig. 2. Utterance editing screen of the Transcription Tool prior to improvements.

2.4 Digital Democracy Transcription Tool

The Digital Democracy Transcription Tool serves as the main source of generating semi-structured data for most of the information and content provided by the Digital Democracy initiative and its website [28, 61]. Transcription Tool is a web application, developed in-house. One of the tool's purposes is to handle administrative tasks such as importing hearing videos, updating hearing metadata, or supervising the transcription process. Despite the importance of its administrative features, Transcription Tool is mainly used by human transcribers to verify automatically generated transcripts in a distributed, asynchronous way, allowing for workforce flexibility. The necessity for human transcription stems from the fact that the textual transcripts produced by automatic systems are not high quality enough for professional and government purposes. In such settings, even slight inaccuracies are highly problematic. Legislative bill information and personal names are not always correct. To fulfill the professional requirements demanded for proceedings of legislative hearings, these mistakes have to be corrected. Additionally, transcribers must identify speakers and decide on their alignment regarding the current issue discussed in the hearing. Last, they work to standardize utterance³ length by merging or splitting utterances. This allows for proper presentation of transcripts on the Digital Democracy website. Transcribers have to make sure no utterances exist that are too short to stand alone.

Figure 2 shows parts of the interface that transcribers use to enhance transcripts. The "Utterance Editing" element seen in this figure is the most relevant part of the transcription screen for this work. It allows transcribers

³In this context, an utterance is defined as one to many sentences spoken by the same person containing a line of thought.

Table 2. Transcription Tool Versions with Feature Descriptions

#	Name	Description
0	Baseline	Overhaul of old Transcription Tool, no major functionality changes
1	Profile Preview	Hovering icon near people’s names shows preview of their profile picture
2	Video Features	Fullscreen for video player, UI to change playback rate of video
3	Utterance Navigation	Buttons for directly jumping to utterance in video and going to next or previous utterance, interactive error messages, UI for manually setting video time
4	VFT	Incorporation of voice, face, and text analysis for speaker suggestion

to modify the automatically generated transcript and its utterances. Transcribers also use this interface to assign speakers to separate utterances. Furthermore, start and end time as well as type of utterances can be changed via these UI elements.

3 TRANSCRIPTION TOOL IMPROVEMENTS

To improve efficiency of the Digital Democracy Transcription Tool, new features are developed. We present these new tool features as four separate tool versions. Feature development and tool releases were performed in a sequential manner. These releases are titled “versions” for the remainder of this article. Table 2 shows an overview of all released tool versions. The following sections describe each new version and their functionality in detail.

3.1 Overhaul of Original Transcription Tool

This overhaul replaces the deprecated Stripes framework as well as other deprecated components with Spring MVC, Spring Boot, Spring Data JPA, and Hibernate [17, 53, 54, 58, 68, 75]. No major new features were introduced over the course of this overhaul. Besides the expected increase in stability and speed, only minor changes such as a dialog box for person creation and editing were realized. Front-end code was not subject to further revisions.

3.2 Profile Preview

One of the biggest efforts for transcribers was found to be identifying speakers solely by voice or their appearance on video. Therefore, a speaker profile preview including pictures was implemented. Through JavaScript and jQuery, this profile picture preview was made available in multiple UI elements containing speaker names. Hovering over the icon added next to a name now displays the speaker profile picture. If no picture is available, only the person’s name and classification are shown (for example, legislator, lobbyist, etc.). Figure 3 demonstrates usage of the profile picture preview.

3.3 Video Features

In this version, the option to toggle “full window mode” was implemented. Through this, currently visible speakers can be compared to people on official websites and third-party platforms. In addition, possibility for transcribers to change video playback rate was enabled. On the one hand, this feature aims at resolving issues with understanding mumbled speech. Some people speak really fast, particularly during routine interactions in the legislature. Slowing down videos helps transcribers to properly understand problematic audio sequences. On the other hand, some people speak very slowly, particularly during testimonies. Experienced transcribers can increase video playback rate to speed up their work.

3.4 Utterance Navigation

Transcribers face difficulties when trying to jump to specific utterances or sections in a video. Therefore, each utterance element in the transcription screen was extended with three icon buttons. The “Play”-button next to

Fig. 3. Profile picture preview for the speaker selection box implemented in version 1.

Fig. 4. Updated utterance element with navigation buttons introduced in version 3.

an utterance directly sets the video time to this utterance's start time. Next to the currently played utterance, two extra buttons are displayed. These allow transcribers to seamlessly jump to the previous or next utterance in the video. Figure 4 shows modifications of the utterance interface element. Furthermore, timestamp input fields were appended below the video player, which can be used to directly jump to the entered video time. Additionally, error messages displayed in the "Task"-tab of the transcription screen were made interactive. Clicking an interactive error message causes the utterance list to automatically scroll to the erroneous utterance. Figure 5 visualizes these UI changes.

3.5 Incorporation of Voice, Face, and Text Analysis for Speaker Suggestion

This version links the Transcription Tool to a system developed by Kauffman et al. [23] that implements voice, face, and text (VFT) analysis. The VFT algorithm is able to guess with a certain probability which speaker made an utterance. Due to the lack of data for non-legislators, speaker recognition is mostly only available for legislators

Fig. 5. Interface for setting video time and display of interactive error messages introduced in version 3.

Fig. 6. VFT feature and speaker suggestion dialog as of version 4.

and ex-legislators. To include VFT results in Transcription Tool, an icon (magnifying glass) was placed in each utterance interface element on the transcription screen. Clicking this icon brings up a dialog box suggesting speakers to the transcriber, as seen in Figure 6. Transcribers can then choose one of the suggestions in the dialog.

4 EVALUATION METHODOLOGY

To fulfill the research goals, metrics enabling evaluation of the Transcription Tool and its separate versions must be defined. Furthermore, the dataset and user groups as well as utilized tools have to be specified.

4.1 Metrics

For our evaluation, we estimate transcription tool efficiency and try to identify transcriber interaction patterns. Therefore, metrics covering both of these areas are defined.

Table 3. Comparison of Video Duration, Video Speech Time (VS_{tid}), and Their Relation over All States for Tasks Completed from October 2017 to March 2018

State	Tasks	Duration (h)		% of VS_{tid} in Video	
		Video	Speech	per Task	Overall
CA	697	250.26	227.79	86.86	91.02
FL	4,091	683.19	645.44	94.03	94.47
NY	924	288.92	262.16	81.6	90.74
TX	6,483	1,071.05	887.81	82.02	82.89
Average	3,045	573.35	505.80	86.13	89.78
All	12,195	2,293.41	2,023.21	86.3	88.22

4.1.1 Transcription Tool Efficiency. Performance of the Transcription Tool is evaluated by how efficiently transcribers handle transcription tasks. To precisely measure this efficiency, well-defined metrics are necessary. Their specification is a requirement for enabling meaningful statistical analysis.

Two basic premises of the Digital Democracy transcription process must be considered. Primarily, plain video duration is not a reliable base for calculating efficiency. Some videos are uploaded untrimmed and contain silent periods, which do not require any transcription. Consequently, only the duration of a video containing speech is chosen as the transcription task's actual length. It is recorded for all tasks (distinguished by their identifier— tid) and labeled as video speech time (VS_{tid}). The actual value of VS_{tid} is derived from utterance timestamps produced by the automatic transcription process. As Table 3 shows, VS_{tid} accounts for only 86.13% of video time in a task, on average. However, it varies strongly across different states. Florida's videos seem to be more appropriately cut by default, since ratio of VS_{tid} to video duration is high (94.03%). All other states hold values below 87%, with New York being especially low (81.6%). On a different note, we base our calculations on values computed through task averages (macro-average) instead of overall mean (micro-average). When using micro-average, longer tasks would have a bigger impact on results than shorter ones. We therefore utilize macro-average, because it weighs each task equally regardless of its duration. The use of VS_{tid} and valuing all tasks the same are prerequisites for defining further efficiency metrics.

For our experiments, the base metric is *Transcription time* (T_{tid}). It measures the time needed to complete a specific transcription task in minutes. Next, *Transcription ratio per task* (TR_{tid}) describes the time in minutes it takes a transcriber to work on a minute of video speech of a specific task (Equation (1)). Subsequently, *Editing ratio* (ER_{eid}) is the average TR_{tid} over all tasks for each specific transcriber (distinguished by their edit identifier— eid), while $EditTasks_{eid}$ captures all tasks edited by a single transcriber (Equation (2)). ER_{eid} can be seen as a value depicting how well a specific transcriber performs the transcription process. The smaller this value, the more efficient a transcriber is. Finally, the main metric used to measure efficiency of the tool and its separate versions is called *Transcription cost in time* (TC). As seen in Equation (3), TC represents the average transcriber editing ratio over the number of all transcribers ($|Transcribers|$). Basically, it is the average amount of time needed to transcribe one minute of video speech. Lower TC naturally corresponds to higher tool efficiency.

$$TR_{tid} = \frac{T_{tid}}{VS_{tid}} \quad (1) \quad ER_{eid} = \frac{\sum_{tid}^{EditTasks_{eid}} TR_{tid}}{|EditTasks_{eid}|} \quad (2) \quad TC = \frac{\sum_{eid}^{Transcribers} ER_{eid}}{|Transcribers|} \quad (3)$$

All metrics can also be calculated over an arbitrary subset of transcription tasks and transcribers (e.g., only specific tool versions). For example, when diminishing task selection according to a specific criteria (c), the resulting transcription cost is referred to as TC_c .

4.1.2 Transcriber Interactions. Interacting with the Transcription Tool can be split up into several interaction categories. The transcription process, and therefore T_{tid} , is composed of: startup time, text correction, speaker identification, splitting and merging utterances, as well as passive interactions. Equations (4) and (5) show a summarization of how T_{tid} is dividable into components for this analysis.

$$T_{tid} = Startup_{tid} + TextCorrection_{tid} + SpeakerId_{tid} + Split_{tid} + Merge_{tid} + Passive_{tid} \quad (4) \quad \begin{aligned} Passive_{tid} = Idle_{tid} \\ + Proofread_{tid} \end{aligned} \quad (5)$$

Startup time can be classified as the time span between loading the transcription screen and first user interaction. *Text correction* time is probably the most elementary interaction between transcriber and tool. It records time spent in text areas editing utterance texts. Time needed for *speaker identification* includes all activities belonging to speaker selection, speaker search, and correction of wrongly assigned speakers. More complex interactions have to also be accounted for, such as leaving the transcription screen to look up people on other websites. *Splitting* and *merging* utterances are relatively simple operations. However, counting time necessary to complete them requires some consideration. Clicking the split or merge button barely takes any time, but deciding whether or not to do so does. Usually, it is obvious to transcribers which very short utterances should be merged right away. However, splitting up utterances requires more effort. Transcribers must not separate a speaker's line of thought while making sure utterance texts are of appropriate length. Due to these reasons, both operations are given fixed costs. Based on real-world observations, it was decided that merging accounts for half a second of transcription time, while splitting accounts for one second.

In contrast to the active interactions explained up until now, others can not be measured reliably using quantitative data. Such activities include the transcriber *proofreading* text and simply watching the video, or not doing any work at all (being *idle*). These *passive* activities were combined into a single parameter (Equation (5)).

Values making up transcriber interaction time in a task can also be collected in an interaction vector I_{tid} (Equation (6)). Consequently, $I_{tid,i}$ refers to the time spent for a specific interaction type i . Using this representation, Equation (4) can be shortened to Equation (7), which shows another definition for T_{tid} . Considering this, *Interaction ratios* ($IR_{tid,i}$) represent the relation between $I_{tid,i}$ and VS_{tid} per task and interaction type (Equation (8)). Last, *Average interaction ratios* ($AIR_{v,i}$) are computable for all tasks per tool version v ($VersionTasks_v$), as seen in Equation (9). $AIR_{v,i}$ makes changes in interaction patterns detectable and is a general representation of how time-consuming they were.

$$I_{tid} = \begin{pmatrix} Startup_{tid} \\ TextCorrection_{tid} \\ SpeakerId_{tid} \\ Split_{tid} \\ Merge_{tid} \\ Passive_{tid} \end{pmatrix} \quad (6) \quad \begin{aligned} T_{tid} = |I_{tid}| = I_{tid,1} + I_{tid,2} + I_{tid,3} \\ + I_{tid,4} + I_{tid,5} + I_{tid,6} \end{aligned} \quad (7)$$

$$IR_{tid,i} = \frac{I_{tid,i}}{VS_{tid}} \quad (8) \quad AIR_{v,i} = \frac{\sum_{tid}^{VersionTasks_v} IR_{tid,i}}{|VersionTasks_v|} \quad (9)$$

To further quantify how different interactions attribute to concrete changes in timing, a parameter $\alpha_{v,i}$ is introduced for each interaction feature. $\alpha_{v,i}$ acts as a multiplier denoting the increase or decrease in time used for an interaction produced in this version compared to the base version 0. This task-independent metric is calculated using $AIR_{v,i}$ (Equation (10)) and measures interaction changes over different versions. Values less than 1 represent an increase in efficiency (i.e., reduction of cost) compared to version 0. Consecutive performance changes over multiple versions illustrated by $\alpha_{v,i}$ can be best presented in a matrix IT . Equation (11) shows the layout of such a matrix for m versions and n interaction types. As already defined above, calculations for $\alpha_{v,i}$ are

carried out for five different versions and six interaction types. In practice, we therefore compute 5×6 matrices.

$$\alpha_{v,i} = \frac{AIR_{v,i}}{AIR_{0,i}} \quad (10) \quad IT = \begin{matrix} & \xrightarrow{\text{interaction types}} \\ \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m,1} & \alpha_{m,2} & \cdots & \alpha_{m,n} \end{pmatrix} & \downarrow \text{versions} \end{matrix} \quad (11)$$

For the sake of completeness, we introduce an alternative transcription time model $M_{tid,v}$, using $\alpha_{v,i}$ and the novel variable $\beta_{tid,i}$. The metric $\beta_{tid,i}$ models the base effort of specific interactions regardless of version. It is dependent on specific task characteristics such as transcriber, video time, number of speakers, and others. Due to the difficult estimation of the influence these characteristics have on transcription time, $\beta_{tid,i}$ is not fully computable. Nevertheless, we utilize this metric in combination with $\alpha_{v,i}$ to model $M_{tid,v}$ for arbitrary tool versions and tasks. As a result, Equation (12) infers a statistically informed model of transcription time, using $\alpha_{v,i}$ and $\beta_{tid,i}$.

$$M_{tid,v} = \alpha_{v,1} \cdot \beta_{tid,1} + \alpha_{v,2} \cdot \beta_{tid,2} + \alpha_{v,3} \cdot \beta_{tid,3} + \alpha_{v,4} \cdot \beta_{tid,4} + \alpha_{v,5} \cdot \beta_{tid,5} + \alpha_{v,6} \cdot \beta_{tid,6} \quad (12)$$

4.2 Data Sources and User Cohort

To properly evaluate tool metrics, interactions between transcribers and the Transcription Tool must be recorded. The implementation of client-side logging achieves this. Meaningful analysis is performed by investigating workflow of a well-selected transcriber cohort. Logs provide insight into the transcription process of these transcribers.

4.2.1 Logging System. A client-side logging system enables tracking of the transcription progress of separate videos, measuring performance and identifying interaction patterns. It also provides additional information for debugging. Therefore, a logging system was implemented that records the following JavaScript events: *focusin*, *focusout*, *click*, *change*, and *keydown* [38]. In addition to such standardized browser interactions, custom events interesting to our use case were also added to the recordings. These interactions include minimizing or leaving the browser window and using keyboard shortcuts. For each event, up to eight values are recorded:

- **event:** String value describing the triggered event. This can either be a standardized JavaScript event or a custom event defined for our purpose. Events are explained in more detail below.
- **element:** Either the tag name of the HTML element for which the event was recorded or additional information for custom events.
- **tagId:** Unique identifier given to this element. This ID might contain information such as the database ID of specific entities to enable further analysis.
- **value:** Text content of the targeted HTML element when the event was triggered. In case of a custom event, this field is used for providing additional information about the specific interaction.
- **keyPressed:** Integer value describing the key pressed to trigger this event. For keyboard interactions, ASCII decimals are used. Left, middle, and right mouse buttons are represented by values 0, 1, and 2.
- **timestamp:** UNIX timestamp in milliseconds (UTC).
- **editorId:** Unique identifier for the current editor.
- **ip:** IP address, used to detect if a person was working remotely or from the project lab.

Logs are stored in JSON format. A separate log file is created for each transcription task and contains all events triggered by the transcriber. In addition, Logs are saved to a MySQL database to enable faster processing and easier data access.

4.2.2 Dataset. As seen in Table 3, 12,195 transcription tasks were completed between October 2017 and March 2018. This accounts for logs about over 2,293 hours of hearings. Since the old Transcription Tool technology was still in use through the beginning of October, only 10,464 tasks were completed with tool versions 0 to 4. A cohort of 20 transcribers was selected to compute tool efficiency metrics and transcriber interactions. These transcribers worked with all tool iterations, including the old tool prior to the overhaul. This selection aims at reducing effects of external factors such as transcriber experience. Working with tool versions 0 to 4, this transcriber cohort completed 3,374 tasks. Furthermore, tasks with $1 < TR_{tid} < 12$ were excluded from the analysis to reduce effects of outliers. After this, 3,272 tasks remained, equaling 652 hours of hearing video and 1,605 hours of transcription work. T_{tid} was calculated by analyzing timestamps of collected logs. Whenever possible, transcriber inactivity was removed from T_{tid} (e.g., closing or minimizing the browser window). Any interaction timeout longer than seven minutes was also excluded, since we assumed that transcribers stopped working on a task after such a pause. This concrete inactivity threshold was defined after discussions with project administrators who know the transcription process well. Analyzing these logs for 1,605 hours of work provides meaningful information, from which conclusions about the transcription process can be drawn.

4.3 Tools and Instruments

The five new tool versions specified in Section 3 were released sequentially. Therefore, we perform a quantitative performance analysis of the separate versions using sequential A/B testing [25, 26, 46].

All computations were carried out using Python3, while accessing data stored in the Digital Democracy MySQL database.

5 FINDINGS AND DISCUSSION

In this section, log analysis results are presented. Similar to the metric definition in Section 4.1.1, we first show-case results for tool efficiency and then transcriber interaction patterns.

5.1 Transcription Tool Efficiency

We achieved a *Transcription cost in time (TC)* reduction of 16.74% from version 0 to version 2 for the selected transcriber cohort, as Table 4(a) shows. Although still producing better performance values than version 0, version 3 caused a cost increase of 7.35% in comparison to version 2. Version 4 again led to decreasing costs and therefore an overall increase of efficiency. Figure 7(a) visualizes these results by comparing *TC* for each version and all transcribers in the cohort. Performance changes over different versions in relation to version 0 can be seen in Figure 7(b).

It has to be mentioned that not every state had hearings eligible for transcription when new versions were released. Therefore, a separate analysis was performed to compare tool efficiency per state. Table 5 presents an overview for *TC* per state and version. New York and California hearings took considerably longer to transcribe in version 3 than they did otherwise. While values for Florida barely changed over different versions, Texas results showed a steady decrease in cost. As a result of these state-specific findings, an additional efficiency analysis excluding New York and tasks containing more than 15 minutes of video was performed. For this analysis of 2,237 tasks (158 hours of video), Table 4(b) shows a 14.80% cost decrease from first to last version. Figure 8(a) visualizes *TC* over all versions for this filtered data. Relation of each version's performance to that of version 0 can be observed in Figure 8(b).

5.2 Transcriber Interactions

Interaction results confirm the feasibility of the proposed division of transcription time. Performing text corrections represented an inexpensive task, taking up only 15.2% of transcription time. This had to be suspected,

Table 4. Transcription Cost for Specific Transcriber Cohort Selections

(a) Original transcriber cohort				(b) Filtered transcriber cohort			
Version	TC	TC Change		Version	TC	TC Change	
		To Version 0	To Previous			To Version 0	To Previous
0	3.935	-	-	0	4.168	-	-
1	3.626	7.85%	7.85%	1	3.934	5.59%	5.59%
2	3.276	16.74%	8.88%	2	3.705	11.11%	5.84%
3	3.566	9.38%	-7.35%	3	3.604	13.52%	2.71%
4	3.172	19.40%	10.01%	4	3.551	14.80%	1.48%

(a) Transcription cost per version.

(b) Performance change in comparison to version 0.

Fig. 7. Transcription cost and performance change over all versions for the original cohort.

since automatic text transcription mostly provides good text quality. By contrast, speaker identification consumed 34.6% of a transcriber's work time and therefore poses a problem. The amount of time passing until a transcriber first interacts with the Transcription Tool was surprisingly high. Ratio of this startup time was observed to be 9.5%. Since timestamps have to be adjusted manually, splitting up utterances is considered more time-consuming than merging (4.3% versus 0.7%). However, measuring these two operations proved difficult, and the approximation described in Section 4.1.2 was used. Passive interactions such as proofreading, watching the video, and inactivity accounted for 35.6%. The inactivity threshold was set to seven minutes, the same as for the efficiency analysis described in Section 5.1 above. These initial findings about transcriber interactions confirm assumptions of administrators. Figure 9 summarizes these findings, which confirm initial assumptions about transcriber interactions.

To provide further perspective, Figure 10 displays interaction analysis results over all versions. It summarizes how much each interaction type contributed to overall transcription time of all tasks. For this, results for each interaction type are clustered together. One of the striking changes observable in the graph is the shift in split and merge operations after version 2. Over the first three versions, merging utterances took up 0.83% of time on average, while splitting accounted for about 3.8%. For versions 3 and 4, merging heavily increased to 3% on average, whereas splitting decreased to 1.7%. Number of text corrections also declined, reaching a 11.5% low in version 4. Being especially high in version 0 with 9.5%, startup time steadily decreased to 6.2% in version

Table 5. Transcription Cost Per State and Version for the Selected Transcriber Cohort

Version	State Results			
	CA	FL	NY	TX
0	2.410	-	2.438	3.925
1	2.478	3.147	-	3.514
2	-	3.336	2.042	3.447
3	3.176	3.134	3.756	3.292
4	2.645	3.045	3.060	2.933

(a) Transcription cost per version.

(b) Performance change in comparison to version 0.

Fig. 8. Transcription cost and performance change over all versions for the filtered cohort.

4. Speaker identification fluctuated between 35% and 28% over all versions. Versions 0 and 1 produced values around 34%, before version 2 reduced the speaker identification ratio to 28.8%. Version 3 increased it to 33.4% again, before it fell to 29.8% in version 4.

In Equation (13), IT_{Cohort} creates a representation of the described cohort results across all versions. However, Equation (14) displays a matrix for results exclusively produced by tasks handling Texas hearings ($IT_{CohortTexas}$).

As explained in Section 4.1.2, rows in IT matrices stand for different versions, while columns represent interaction types. From left to right, these interactions are: startup, text correction, speaker identification, splitting, merging, and passive. Values in the last row of $IT_{CohortTexas}$ are inconclusive, because only seven tasks concerning Texas legislature were completed with version 4. In contrast to IT_{Cohort} , $IT_{CohortTexas}$ shows a constant decrease for speaker identification time in the third column (0.954 to 0.327). Furthermore, Texas tasks completed with version 3 had significantly higher splitting and lower merging ratios than tasks of other states. Startup

Fig. 9. Average interaction ratios (AIR) for transcription time in version 0.

Fig. 10. Average interaction ratios (AIR) by interaction type.

time was also lower for version 3 in $IT_{CohortTexas}$. These results again suggest that state selection influences interaction ratios.

$$IT_{Cohort} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ .78 & .96 & .98 & 1.12 & .95 & 1.10 \\ .77 & .97 & .83 & 1.30 & .68 & 1.27 \\ .73 & .88 & .97 & 3.98 & .48 & 1.16 \\ .65 & .76 & .86 & 4.34 & .33 & 1.35 \end{pmatrix} \quad (13) \quad IT_{CohortTexas} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ .89 & .93 & .93 & 1.04 & .86 & 1.18 \\ .98 & .86 & .92 & 1.30 & .72 & 1.20 \\ .48 & 1.12 & .75 & 1.42 & 1.07 & 1.39 \\ .15 & .84 & .48 & 3.61 & .46 & 2.00 \end{pmatrix} \quad (14)$$

5.3 Discussion

Overall, it was shown that new features decreased *TC* for the given transcriber cohort. However, some discrepancies in efficiency and interaction pattern results need further investigation. It is especially relevant to research why costs for version 3 increased. For this, a closer look at changes of interaction type ratios over multiple versions is taken. Discussion about how past tool evaluation methodologies compare to the log analysis attempted in this thesis is also carried out. Furthermore, an estimate of monetary savings, resulting from the achieved cost reduction, is presented. In the end, threats to validity of the analysis and countermeasures taken to prevent them are elaborated.

5.3.1 Speaker Identification and Passive Interactions. Ratio changes of passive interactions seem to be connected to speaker identification ratios. When the amount of time used to identify speakers increased, passive time also decreased in return. This can be explained due to several reasons. First, determining if transcriber inactivity was actually devoted to identifying a speaker using another browser window is difficult. Time spent outside the transcription screen was only attributed to speaker identification if elements belonging to speaker selection, speaker search, or profile modification were clicked after coming back. Second, interaction ratio results are influenced by existence of relevant speaker profile pictures. If officials participating in a task's hearing have pictures assigned to them, they can be viewed using the profile picture preview introduced in version 1. Opening and closing the preview is observed by logs and can be attributed to the speaker identification parameter. However, transcribers must search for speaker names on third-party sites if no pictures exist. This presents another potential reason for a gain in passive interaction time. Furthermore, passive interactions were used as a "catch all" parameter to collect time that could not be attributed to any other interaction type. Results could also be skewed because of this. In addition, time necessary for speaker identification is highly dependent on state. As seen in Table 5, New York tasks are more expensive than others in version 3. Transcribers might have had problems identifying legislators and persons unknown to them. New York hearing videos also provide the worst video and sound quality, further complicating speaker identification.

5.3.2 Merging, Splitting, and Version 3 Efficiency. The drastic changes in merge and split operations as well as the cost increase for version 3 were closely investigated. It was discovered that a configuration change in the preprocessing pipeline caused these abnormalities. A modified parameter in a preprocessing script led to generation of shorter utterances in tasks. This produced an increase in merge and decrease in split operations. The vast number of utterances also led to browsers being slowed down by longer tasks, decelerating the transcription process. These findings can be seen as proof that logs can also be utilized to detect software defects and unexpected behavior in the tool.

On a side note, a decline in splitting operations also led to a decrease in text corrections. Basically, sentence boundaries must be modified nearly every time an utterance is split. Therefore, less text correction time can be explained by fewer splitting operations occurring. Considering these assumptions and efficiency results for filtered cohort data (Table 4(b)), it seems probable that side effects arose from non-ideal preprocessing parameters. Excluding longer tasks from this filtered data also led to an increase in *TC*. This could be explained by a number of reasons. First, transcribers can fast-forward through longer speeches using video features introduced in version 2. Second, long testimonies and debates between legislators do not influence factors such as speaker identification. Naturally, speakers only have to be identified once, no matter how often they appear in a video. From all of this it can be concluded that preprocessing strongly affects the transcription process, especially for longer tasks.

5.3.3 Monetary Savings. Tool improvements yield a considerable decrease in monetary cost. As presented in Section 5.1, version 0 produced a *TC* of 3.935 (3 minutes 56 seconds). Using version 4, experienced transcribers managed to achieve a *TC* of only 3.172 minutes (3 minutes 10 seconds), reducing costs by 19.4%. Assuming an hourly labor rate of \$10, such a reduction equates to cost savings of \$7.63 per hour of video speech. In March

Table 6. Comparison between Rovin’s Tool Evaluation and the Study Conducted in This Work

Characteristics	Rovin	Rupprechter et al.
Transcription Time per Minute of Video Speech	5.441 Minutes	3.935 Minutes
Accomplished Efficiency Improvement	16.89%	19.40%
Number of Transcribers for Evaluation	15	20
Number of Tasks for Evaluation	5	>3,200
Text Quality Evaluation	Yes	No
Evaluation Method	Lab Study	Log Analysis
Date of Data Collection	Spring 2016	October 2017 – March 2018

Rovin’s results were normalized using average VS_{tid} .

2018, the Digital Democracy initiative created transcriptions for 545.17 hours of hearing video speech. Applying the presented estimations, a claim can be made that \$4159.65 was saved for this one month alone. This would result into a yearly benefit of nearly \$50,000 for the initiative, assuming hearing video input would stay similar for all months. Therefore, the implemented improvements clearly decreased process cost.

5.3.4 Comparison to Previous Evaluations. Rovin [61] previously attempted an evaluation of transcription cost for Digital Democracy. He introduced improvements developed during multiple projects into the transcription process, aiming towards increasing transcription speed and quality. For evaluation of transcription time reduction as well as quality enhancement, Rovin conducted a laboratory user study. In contrast to his analysis, no hypothesis regarding transcription quality can be constructed using our system. Log analysis, as attempted in this work, is unfit to detect mistakes regarding speaker assignments or transcription texts for arbitrary tasks. Such a quality evaluation has to be performed in a controlled environment, where an immaculate test set of tasks is available. However, log analysis provides better insight into actual transcription cost and tool efficiency. While helpful for an initial evaluation, lab tests using a small number of transcribers and tasks do not capture real-world working conditions [1, 15].

Due to new developments being implemented after publication of Rovin’s work, results differ vastly. Back in 2016, his improvements led to an average transcription time of 6.317 minutes per minute of video (6 minutes 19 seconds), increasing efficiency by 16.89% from when he first started his work. Since TC considers video speech time instead of video length, Rovin’s results are normalized to 5.441 minutes per minute of video for a comparison (5 minutes 27 seconds). This value is calculated by using the average proportion of VS_{tid} to video length as presented in Section 4.1.1 (86.13%). Table 6 shows a final comparison between the characteristics of Rovin’s tool evaluations and the evaluation study conducted here.

5.3.5 Socio-technical Aspects of Production. There has been a great amount of ongoing discussion between supervisors and managers of the Digital Democracy initiative about budgeting and staffing.

One of the recurring issues is the trade-off between quality of third-party transcription services, especially regarding automatically transcribed texts and cost of human transcription. On the one hand, high-quality services such as cielo24 [9] are relatively expensive, but produce close to perfect transcription texts along with accurate utterance timestamps. On the other hand, alternatives such as Watson [20] might slightly decrease quality of generated texts, but could conceivably reduce costs. In addition, human labor for text correction purposes is comparatively cheap. Through our interaction analysis, we have shown that only a minor fraction of transcribers’ time is actually attributed to correcting texts. Contrarily, speaker detection is the biggest challenge in human-assisted transcription. Automatic services still fail to provide correct data for this purpose. This issue is currently not solvable automatically, no matter the investment in transcription services. Therefore, human transcribers’ primary task is to perform correct speaker assignments. During this process, they read transcription texts only superficially and merely correct minor mistakes anyway. A marginal reduction in quality of

automatically transcribed texts might therefore only slightly increase costs for human transcribers, while greatly diminishing the cost of automatic transcription services.

Moreover, Digital Democracy utilized college student employees for transcription up-leveling and to conduct user studies. This proved to be advantageous, as students tended to be tech-savvy and absorb their short training material quite efficiently. Still, according to operational managers, it was difficult employing the right person for such tasks and turn-over was fairly high. While the transcription process is largely automated, the additional text correction, speaker identification, and utterance management can be considered repetitive. In addition, the subject matter is often stodgy and humdrum for college students. Some guidelines communicated by the initiative's managers were to first identify if the student had the interest and stamina for this kind of labor. Additionally, managers scheduled sessions in one- to three-hour periods at the longest. Furthermore, average transcription tasks assigned to students were only about 20 minutes long, so students could set achievable goals. Last, it was beneficial to create a common working space for students where they could perform tasks in flexible hours. Nonetheless, it is debatable if utilization of professional workforce would have been preferable to student employees, considering both costs and throughput.

5.3.6 Threats to Validity. The presented analysis can only be considered valid if the majority of interaction and efficiency changes are indeed caused by tool improvements. There are three definite threats we identify. First, side effects and non-measurable factors can occur in a real-world testing scenario, as featured in this research. If such factors had major influence on results, validity of the presented analysis would be threatened. Minimizing effects of unexpected threats was addressed by choosing a large task subset to evaluate tool versions. Although tasks are vastly different, analyzing over 3,200 of them aims at compensating for outliers in the data. Another major threat to validity of the analysis could be the different skill levels of transcribers. Choosing a cohort of 20 experienced transcribers who worked with all versions of the tool mitigates this. Through this selection, it can be assumed that all test participants possess a similar skill level. Last, states strongly influence nature of transcription tasks. Although not directly addressed in the initial analysis in Section 5.1, results for separate states were also computed to confirm this. Another possibility of solving this issue was presented in Section 5.3.2, where cohort data was filtered to exclude long tasks and New York hearings. We believe that apparent threats to the validity of this analysis were addressed accordingly.

6 CONCLUSIONS AND FUTURE WORK

In the absence of fully automated, highly accurate transcription technology, human-assisted transcription is at present a viable and cost-effective option for capturing speech from legislative proceedings. This work focuses on building a deeper understanding of human-assisted transcription and annotation systems, how to make them more efficient, and, finally, how the human contribution is distributed across various interaction types expected from transcribers. Gaining understanding in the transcription process was achieved by examining results of the implemented logging system.

First, we describe the Digital Democracy initiative and the Transcription Tool necessary for the bulk of input data preparation for this project. A technology overhaul and four sets of improvements were implemented for this tool, which aimed at increasing overall efficiency. Improvements were sequentially released as four separate tool versions. While baseline version 0 represented the developed technology overhaul, versions 1 to 4 introduced new features. Version 1 incorporated speaker profile preview into the tool. In version 2, enhanced video player functionality was implemented. Version 3 enabled better utterance navigation and interactivity with the transcription screen. Finally, version 4 connected the Transcription Tool to an automated speaker identification service implemented in another student project. Furthermore, a logging system was developed to allow sophisticated analysis of the transcription process. This system collected data for over 12,000 transcription tasks and about 2,300 hours of video, resulting in roughly 8,000 hours of transcription work. Using the collected logs,

a study of how tool improvements affected efficiency and cost of the entire operation was performed. A cohort of 20 transcribers was selected for this analysis.

Transcription cost (*TC*) was defined as the main metric measuring tool efficiency. *TC* stands for the average time it takes a human transcriber to transcribe one minute of video speech using the Transcription Tool. It was found that on average a 19.4% decrease in *TC*, and therefore increase in efficiency, could be realized over four versions. During this analysis, it was also shown that side effects and factors such as state, preprocessing parameters, and task duration influence the transcription process and consequently *TC*. To analyze how human transcribers interact with the Transcription Tool, transcription time was split up into six interaction types. The following interaction types were investigated: speaker identification, text correction, tool startup, splitting and merging utterances, as well as the transcriber being idle (passive). By utilizing log data, these interactions were quantified to assess their impact on the overall transcription process. It was confirmed that speaker identification and passive time accounted for over 65% of combined transcription time in all versions. Text correction only took up around 12%–15% of time on average, meaning that automatic transcription services produce good-quality texts. Splitting and merging operations are heavily dependent on preprocessing parameters such as utterance length. Startup time was higher than expected, reaching close to 10% for nearly all versions. Although the Digital Democracy initiative temporarily suspended operations as of January 2019, we believe this work presents valuable findings for similar projects and Digital Government as a whole.

As for future work, providing correct, professional, and fully contextual primary-sourced legislative information to citizens and media is essential for government transparency. Thus, we envision future systems will be even more integrated and easier to navigate. Further research into better user interfaces, multi-faceted search systems, and alerting capability will undoubtedly continue. Basic transcription technology, natural language annotation, and voice processing will become better and cheaper, leading to easier access to more information by citizens. Face recognition and other visual sense-making algorithms have already improved tremendously even since the start of the present study. Future versions could take advantage of such technology to better and more cheaply annotate the transcriptions. Though we see the need for human-based workforce diminishing, we do not see it disappearing. When it comes to important issues such as democracy and government transparency, some level of quality assurance using a knowledgeable workforce is going to be necessary for the foreseeable future.

However, as information becomes more voluminous and the access easier, another problem is exacerbated: Average working adults are not afforded the time and resources to study the legislature and fully consider all implications of what is occurring. Indeed, as any frequent viewer of digital product license agreements can attest, the volume of information and the time it takes to process it is a significant barrier, even when the crucial information is fully disclosed. We therefore suggest there is great need for unbiased algorithmic analysis and summarization of legislative content that must scale along with the volume of such content. Such automated analysis can take the form of intelligent chat bots or voice-interfaced assistants that would allow users to focus on the data necessary to comprehend political issues that are of interest to them.

ACKNOWLEDGMENTS

The authors thank the IATPP founder Hon. Sam Blakeslee, IATPP associate director Christine Robertson and staff members Toshihiro Kuboi and Lisa Nunes. The authors further thank the Austrian Marshall Plan Foundation and the Institute of Interactive Systems and Data Science (ISDS) at Graz University of Technology.

REFERENCES

- [1] Maristella Agosti, Franco Crivellari, and Giorgio Maria Di Nunzio. 2012. Web log analysis: A review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining Knowl. Disc.* 24, 3 (2012), 663–696.
- [2] Edgar Alejandro Ruvalcaba-Gomez, J. Ignacio Criado, Rodrigo Sandoval-Almazan, and David Valle-Cruz. 2019. Understanding the perspectives of open government: Exploratory factor analysis in Spain and Mexico. In *Proceedings of the 20th Annual International Conference on Digital Government Research*. ACM, 507–508.

- [3] J. Basu, M. S. Bepari, S. Nandi, S. Khan, and R. Roy. 2013. SATT: Semi-automatic transcription tool. In *Proceedings of the International Conference Oriental COCOSA Held Jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSA/CASLRE'13)*. IEEE, 1–6. DOI : <https://doi.org/10.1109/ICSDA.2013.6709903>
- [4] Sam Blakeslee, Alex Dekhtyar, Foaad Khosmood, Franz Kurfess, Toshihiro Kuboi, Hans Poschman, Giovanni Prinzivalli, Christine Robertson, and Skylar Durst. 2015. Digital Democracy Project: Making Government More Transparent one Video at a Time. In *Digital Humanities 2015*. Sydney, Australia. Retrieved from <https://dh-abstracts.library.cmu.edu/works/2221>.
- [5] Aditya Budhwar, Toshihiro Kuboi, Alex Dekhtyar, and Foaad Khosmood. 2018. Predicting the vote using legislative speech. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 35.
- [6] Moira Burke, Brian Amento, and Philip Isenhour. 2006. Error correction of voicemail transcripts in SCANMail. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, 339–348. DOI : <https://doi.org/10.1145/1124772.1124823>
- [7] C-SPAN. 2018. Congressional Chronicle—Members of Congress, Hearings and More. Retrieved from c-span.org/congress.
- [8] Yumei Chen, Sharon S. Dawes, and Shanshan Chen. 2017. E-government support for administrative reform in China. In *Proceedings of the 18th Annual International Conference on Digital Government Research*. ACM, 329–335.
- [9] cielo24. 2018. cielo24. Retrieved from cielo24.com.
- [10] Benjamin Davis and Phineas Baxandall. 2014. Following the Money 2014. *CALPIRG Education Fund*. Retrieved from calpirgedfund.org/reports/caf/following-money-2014.
- [11] Sharon S. Dawes and Natalie Helbig. 2010. Information strategies for open government: Challenges and prospects for deriving public value from government transparency. In *Proceedings of the International Conference on Electronic Government*. Springer, 50–60.
- [12] Emilio Feliciano de Oliveira and Milene Selbach Silveira. 2018. Open government data in Brazil—A systematic review of its uses and issues. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 1–9.
- [13] R. Deshpande, T. Tuna, J. Subhlok, and L. Barker. 2014. A crowdsourcing caption editor for educational videos. In *Proceedings of the IEEE Frontiers in Education Conference (FIE'14)*. IEEE, 1–8. DOI : <https://doi.org/10.1109/FIE.2014.7044040>
- [14] Digital Democracy initiative. 2017. Digital Democracy homepage. Retrieved from <https://iatpp.calpoly.edu/digital-democracy>.
- [15] Susan Dumais, Robin Jeffries, Daniel M. Russell, Diane Tang, and Jaime Teevan. 2014. *Understanding User Behavior through Log Data and Analysis*. Springer, New York, NY. 349–372.
- [16] Justin Fox. 2007. Government transparency and policymaking. *Public Choice* 131, 1–2 (2007), 23–44.
- [17] Oliver Gierke, Thomas Darimont, Christoph Strobl, Mark Paluch, and Jay Bryant. 2018. Spring Data JPA - Reference Documentation. Retrieved from docs.spring.io/spring-data/jpa/docs/current/reference/html.
- [18] Governing.com. 2014. The New Reality of Statehouse Reporting. Retrieved from governing.com/topics/politics/gov-the-new-realities-of-statehouse-press-corps.html.
- [19] Kenneth L. Hacker and Jan van Dijk. 2000. *Digital Democracy: Issues of Theory and Practice*. ACM.
- [20] IBM. 2018. Watson Speech to Text. Retrieved from ibm.com/watson/services/speech-to-text.
- [21] Marijn Janssen, Yannis Charalabidis, and Anneke Zuidewijk. 2012. Benefits, adoption barriers, and myths of open data and open government. *Inf. Syst. Manag.* 29, 4 (2012), 258–268.
- [22] Daniel Kauffman, Foaad Khosmood, Toshihiro Kuboi, and Alex Dekhtyar. 2018. Learning alignments from legislative discourse. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 119.
- [23] Daniel Kauffman, Michael Williams, Calin Washington, Gudrun Socher, and Foaad Khosmood. 2018. Multimodal speaker identification in legislative discourse. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM. Retrieved from dl.acm.org/citation.cfm?id=3209375.
- [24] Foaad Khosmood, Alex Dekhtyar, Hisham Assai, Franz Kurfess, and Joanna Snyder. 2014. *Making California Legislative Process Transparent*. Technical Report. California Polytechnic State University, San Luis Obispo, CA.
- [25] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Chicago, IL, 1168–1176.
- [26] Ron Kohavi and Roger Longbotham. 2017. *Online Controlled Experiments and A/B Testing*. Springer, New York, NY. 922–929 pages.
- [27] Lahis Pasquali Kurtz, Paloma Maria Santos, and Aires José Rover. 2018. Access to information on websites of Brazilian superior courts of justice: Comparative analysis between data from 2013 and 2017. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 1–5.
- [28] Michael Latner, Alexander M. Dekhtyar, Foaad Khosmood, Nicole Angelini, and Andrew Voorhees. 2017. Measuring legislative behavior: An exploration of Digitaldemocracy.org. *California Journal of Politics and Policy* 9, 3 (2017). Retrieved from <https://escholarship.org/uc/item/3g73v00t>.
- [29] Legislative Analyst's Office. 2016. Proposition 54. Retrieved from lao.ca.gov/BallotAnalysis/Proposition?number=54&year=2016.
- [30] Library of Congress. 2018. Congressional Record. Retrieved from congress.gov/congressional-record.
- [31] Edimara Mezzomo Luciano, Guilherme Costa Wiedenhöft, and Fabio Pinheiro dos Santos. 2018. Promoting social participation through digital governance: Identifying barriers in the Brazilian public administration. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 1–9.

- [32] Saturnino Luz, Masood Masoodian, Bill Rogers, and Chris Deering. 2008. Interface design strategies for computer-assisted speech transcription. In *Proceedings of the 20th Australasian Conference on Computer-human Interaction: Designing for Habitus and Habitat (OZCHI'08)*. ACM, New York, NY, 203–210. DOI: <https://doi.org/10.1145/1517744.1517812>
- [33] Vivianne Machado, Gabriel Mantini, José Viterbo, Flavia Bernardini, and Raissa Barcellos. 2018. An instrument for evaluating open data portals: A case study in Brazilian cities. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 1–10.
- [34] Vukosi Marivate and Nyalleng Moorosi. 2018. Exploring data science for public good in South Africa: Evaluating factors that lead to success. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 1–6.
- [35] Albert Meijer. 2015. Government transparency in historical perspective: From the ancient regime to open data in the Netherlands. *Int. J. Pub. Admin.* 38, 3 (2015), 189–199. DOI: <https://doi.org/10.1080/01900692.2014.934837>
- [36] Morten Meyerhoff Nielsen. 2019. Governance lessons from Denmark’s digital transformation. In *Proceedings of the 20th Annual International Conference on Digital Government Research*. 456–461.
- [37] Diómedes Eduardo Palacios Morales. 2018. Analysis of digital government in Ecuador: Review of Ecuadorian agenda with regard to the digital government stage framework. In *Proceedings of the International Conference on eDemocracy & eGovernment (ICEDEG'18)*. IEEE, 345–350.
- [38] Mozilla and individual contributors. 2018. —Event reference. Retrieved from developer.mozilla.org/en-US/docs/Web/Events.
- [39] Geoff Mulgan. 2017. *Digital Democracy: The Tools Transforming Political Engagement*. Nesta.
- [40] Abdulsalam Salihu Mustafa and Mahammad Sharifov. 2018. Towards enhancing citizens engagement: A review of parliamentary web-sites in the 36 state houses of assembly in Nigeria. *JeDEM* 10, 2 (2018), 97–111.
- [41] Rini Yudesia Naswir, Nurazeen Maarop, Mahmudul Hasan, Salwani Daud, Ganthan Narayana Samy, and Prithheega Magalingam. 2019. Towards a conceptual model to evaluate usability of digital government services in Malaysia. *High. Educ.* 10, 4 (2019).
- [42] National Public Radio. 2014. Study: Statehouse Press Corps in Decline. Retrieved from [npr.org/sections/itsallpolitics/2014/07/10/330456166/study-statehouse-press-corps-in-decline?t=155678344382](https://www.npr.org/sections/itsallpolitics/2014/07/10/330456166/study-statehouse-press-corps-in-decline?t=155678344382).
- [43] New Zealand Parliament. 2018. Read Hansard Reports. Retrieved from parliament.nz/en/pb/hansard-debates/rhr.
- [44] Morten Meyerhoff Nielsen. 2017. eGovernance and online service delivery in Estonia. In *Proceedings of the 18th Annual International Conference on Digital Government Research*. ACM, 300–309.
- [45] Joakim Nivre, Kristina Tullgren, Jens Allwood, Elisabeth Ahlsén, Jenny Holm, Leif Gronqvist, Dario Lopez-Kästen, and Sylvana Sofkova. 1998. Towards multimodal spoken language corpora: TransTool and synctool. In *Proceedings of the Workshop on Partially Automated Techniques for Transcribing Naturally Occurring Continuous Speech (Transcribe'98)*. Association for Computational Linguistics, 11–23.
- [46] Helena Holmström Olsson and Jan Bosch. 2014. From opinions to data-driven software R&D: A multi-case study on how to close the “open loop” problem. In *Proceedings of the Conference on Software Engineering and Advanced Applications (SEAA'14) and 40th EUROMICRO*. IEEE, 9–16.
- [47] ORF. 2018. ORF TVthek. Retrieved from tvthek.orf.at.
- [48] Frédéric Papazian, Robert Bossy, and Claire Nédellec. 2012. AlvisAE: A collaborative web text annotation editor for knowledge acquisition. In *Proceedings of the 6th Linguistic Annotation Workshop (LAW'12)*. Association for Computational Linguistics, 149–152.
- [49] Parliament of Australia. 2018. Hansard. Retrieved from aph.gov.au/Parliamentary_Business/Hansard.
- [50] Parliament of Canada. 2018. Publication Search. Retrieved from ourcommons.ca/parliamentarians/en/publicationsearch.
- [51] Parliament of the United Kingdom. 2018. Hansard Online. Retrieved from hansard.parliament.uk.
- [52] Pew Research Center. 2014. America’s Shifting Statehouse Press. Retrieved from [journalism.org/2014/07/10/americas-shifting-statehouse-press](https://www.journalism.org/2014/07/10/americas-shifting-statehouse-press).
- [53] Pivotal Software. 2018. Spring AOP APIs. Retrieved from docs.spring.io/spring/docs/current/spring-framework-reference/core.html#aop-api.
- [54] Pivotal Software. 2018. Web on Servlet Stack. Retrieved from docs.spring.io/spring/docs/current/spring-framework-reference/web.html.
- [55] Gabriel Puron-Cid. 2017. From technology to social development: Applying a public value perspective to digital government in local governments in Mexico. In *Proceedings of the 18th Annual International Conference on Digital Government Research*. ACM, 336–345.
- [56] Quo Vadis Veritas Redaktions GmbH. 2018. Addendum – Politometer. Retrieved from <http://tvthek.orf.at>.
- [57] Raychel Lean. 2020. Court Addresses “Fundamental Right” to Closed Captioning in Politics. Retrieved from [law.com/dailybusinessreview/2020/01/06/court-addresses-fundamental-right-to-closed-captioning-in-politics/](https://www.law.com/dailybusinessreview/2020/01/06/court-addresses-fundamental-right-to-closed-captioning-in-politics/).
- [58] Red Hat. 2018. Hibernate. Retrieved from hibernate.org.
- [59] Christopher G. Reddick, Akemi T. Chatfield, and Gabriel Puron-Cid. 2017. Online budget transparency innovation in government: A case study of the US state governments. In *Proceedings of the 18th Annual International Conference on Digital Government Research*. ACM, 232–241.

- [60] Alejandro Revuelta-Martinez, Luis Rodriguez, and Ismael Garcia-Varea. 2012. A computer assisted speech transcription system. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*. Association for Computational Linguistics, 41–45.
- [61] Justin Rovin. 2016. *Reducing Costs in Human Assisted Speech Transcription*. Master's thesis. California Polytechnic State University, San Luis Obispo, CA. Retrieved from digitalcommons.calpoly.edu/theses/1538.
- [62] Thorsten Ruprecht. 2019. *Increasing Efficiency in Video Transcription*. Master's thesis. Graz University of Technology, Graz.
- [63] Thorsten Ruprecht, Foaad Khosmood, Toshihiro Kuboi, Alex Dekhtyar, and Christian Gütl. 2018. Gaining efficiency in human assisted transcription and speech annotation in legislative proceedings. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age (dg.o'18)*. ACM, New York, NY. DOI: <https://doi.org/10.1145/3209281.3209410>
- [64] Ladislav Šeps. 2013. NanoTrans—Editor for orthographic and phonetic transcriptions. In *Proceedings of the 36th International Conference on Telecommunications and Signal Processing (TSP'13)*. IEEE, 479–483.
- [65] Lee Shaker. 2014. Dead newspapers and citizens' civic engagement. *Polit. Commun.* 31, 1 (2014), 131–148. DOI: <https://doi.org/10.1080/10584609.2012.762817>
- [66] Tobias Siebenlist and Agnes Mainka. 2018. Culture of openness: The beginning of open government in municipalities in Germany. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 1–6.
- [67] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*. Association for Computational Linguistics, 102–107.
- [68] Stripes. 2018. Stripes Wiki - Stripes Framework Wiki. Retrieved from stripesframework.org.
- [69] Leif Sundberg. 2019. Value positions and relationships in the Swedish digital government. *Admin. Sci.* 9, 1 (2019), 24.
- [70] Michelle Surka and Elizabeth Ridlington. 2016. Following the Money 2016. *CALPIRG Education Fund*. Retrieved from calpirg.org/reports/cap/following-money-2016.
- [71] The Austrian Parliament. 2018. Erweiterte Suche. Retrieved from parlament.gv.at/SUCH.
- [72] The Austrian Parliament. 2018. Stenographische Protokolle. Retrieved from parlament.gv.at/PAKT/STPROT.
- [73] The European Parliament. 2018. Debates and Videos | Planary. Retrieved from europarl.europa.eu/plenary/en/debates-video.html.
- [74] The Swiss Parliament. 2018. Search Official Bulletin. Retrieved from parlament.ch/en/ratsbetrieb/suche-amtliches-bulletin.
- [75] Phillip Webb, Dave Syer, Josh Longand Stéphane Nicoll, Rob Winch, Andy Wilkinson, Marcel Overdijk, Christian Dupuis, Sébastien Deleuze, Michael Simons, Vedran Pavić, Jay Bryant, and Madhura Bhawe. 2018. Spring Boot Reference Guide. Retrieved from docs.spring.io/spring-boot/docs/current-SNAPSHOT/reference/htmlsingle.
- [76] Steve Whittaker and Brian Amento. 2004. Semantic speech editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'04)*. ACM, New York, NY, 527–534. DOI: <https://doi.org/10.1145/985692.985759>
- [77] Antoine Widlöcher and Yann Mathet. 2012. The glozz platform: A corpus annotation and mining tool. In *Proceedings of the ACM Symposium on Document Engineering (DocEng'12)*. ACM, New York, NY, 171–180. DOI: <https://doi.org/10.1145/2361354.2361394>
- [78] Fons Wijnhoven, Michel Ehrenhard, and Johannes Kuhn. 2015. Open government objectives and participation motivations. *Gov. Inf. Quart.* 32, 1 (2015), 30–42.
- [79] Sang-Pil Yoon, Moon-Ho Joo, and Hun-Yeong Kwon. 2018. Role of law as a guardian of the right to use public sector information: Case study of Korean government. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 1–10.

Received July 2019; revised March 2020; accepted April 2020