CHRISTOPH DRAXLER[1]

Automatic Transcription of Spoken Language Using Publicly Available Web Services

Abstract

This paper is an informal introduction to transcribing speech in general and the current state of the art in automatic speech recognition in particular, and it elaborates on the differences between commercial and academic speech recognisers. Based on a short extract of an oral history interview, it presents four different types of transcription, and compares the output of a commercial speech recognition system to a human-generated transcript. It proposes a simple graphical tool which allows potential users to estimate the quality of the recognition output. Finally, it introduces the speech processing web services offered by the Bavarian Archive for Speech Signals (BAS) and shows how they may be used to automate parts of the transcription workflow.

1. *Introduction*

In recent years, the performance of automatic speech recognition (ASR) has improved substantially. In many application areas, e. g. travel information systems, or medical or juridical dictation systems, it has reached human performance levels. Currently, voice-driven personal digital assistants such as Amazon Alexa, Apple Siri, or Google Assistant are becoming increasingly popular and commercially successful – they are available on mobile personal devices, do not require technical expertise to use, and perform well on tasks as diverse as retrieving information from the internet, online shopping, selecting song tracks from streaming services, or dictating personal messages.

The success of ASR has of course attracted the attention of scientific fields working on spoken language – wouldn't it be great if a machine could provide high-quality transcripts of interviews, field recordings, medical or foreign language learning tests, child speech recordings, etc.? Expectations are high!

In this informal introduction, I will first give an overview of the workflow when working with spoken language, and briefly describe the foundations of current ASR technology in non-technical terms. Then, I will look at the various types of transcriptions needed by different scientific fields, and compare these to the outcomes of currently available free or low-cost services provided by commercial ASR providers. Finally, I will present web services offered by the Bavarian Archive for Speech

[1] LMU München.

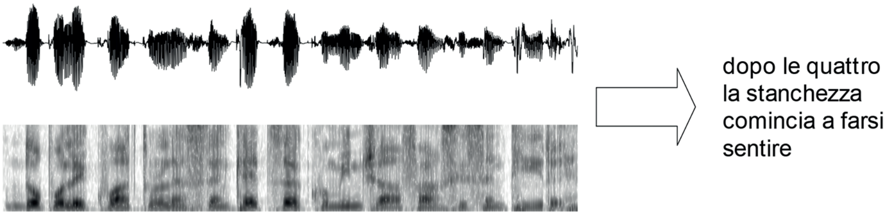
Signals (BAS) which support and even partly automate the transcription workflow, including access to ASR.

2. *Spoken Language Transcription Workflow*

Spoken language is a physical signal, i.e. a change in air pressure over time, produced by an airflow from the lungs and modulated by the vocal tract, and perceived via the ears. By nature, it is *volatile* – as soon as it is produced, it is gone.

For processing and analysis, the speech signal needs to be captured and stored. In a *categorisation process*, symbolic labels, e.g. phonemes or words, are mapped to time-delimited fragments of the signal (see e.g. (Spreafico 2020) for an in-depth discussion). The result of this process is a *transcript*, which contains the verbal content of a given utterance (see Figure 1 for an example[2]). This transcript is the basis of all further analysis and processing steps.

Figure 1 - *In a categorisation process, a continuous spoken language signal (shown in its waveform and the derived spectrogram) is labelled with discrete symbolic labels, e. g. words*



The workflow consists of five main consecutive steps:

1. Recording: The speech signal is recorded and stored as digital data
2. Transcription: The recorded speech is transcribed by human transcribers or ASR to produce a verbal transcript
3. Segmentation: The transcript is time-aligned with the signal, and annotation levels, e. g. fine-grained phonetic labels, are added
4. Data Management: Transcripts and the corresponding signals are compiled in speech databases for further analysis
5. Analysis, Processing, Presentation: Theory-guided semi-automatic processing, manual analysis, or visual presentation of audio signals, derived signal data, and annotations

Each of these steps requires specific expertise and software tools. Figure 2 shows this workflow; icons represent software tools typically used for the individual steps. Some of these tools, e.g. web browsers, database management systems, or statistics systems such as R (R Core Team 2015) are general purpose software tools.

[2] This recording is available under https://www.phonetik.uni-muenchen.de/forschung/Bas/Experimente/aitla/Test0001IT_S0.wav.

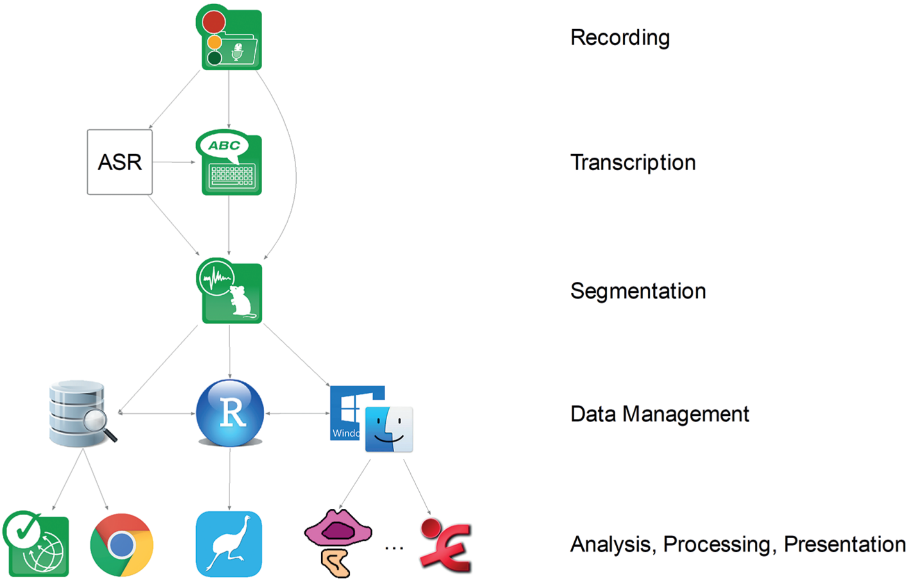
Others, such as Praat (Boersma 2001), ELAN (Sloetjes 2007), or Octra (Pömp & Draxler 2017), the segmentation tool MAUS (Schiel 2015, Kisler et al. 2012), or the recording software SpeechRecorder (Draxler & Jänsch 2004), are tailored to spoken language processing. Draxler et al. (2020) describes the T-chain web service, which provides a simple graphical user interface to the services used for transcribing recordings.

The costs associated with the workflow depend on the time and expertise needed to perform the different steps. A common measure is the *real time factor.* It states how much longer a given processing step takes than the duration of the spoken language signal. Table 1 contains estimates for recording, transcription, and manual or automatic segmentation.

Table 1 - *Real time factors and unit costs for recording, transcription, and segmentation (i.e. time-alignment) of spoken language*

task	*real time*	*unit cost*
recording	1 - 2	€
transcription	5 - 10	€€
manual segmentation	300 - 1000	€€€
automatic segmentation + manual correction	10 - 50	€€€

Figure 2 - *Transcription workflow and tools*



The real time factor for segmentation depends on the granularity of the transcription – orthographic vs. broad phonemic vs. detailed phonetic transcription – and the amount of manual validation and correction necessary.

Recording and transcription in general require only basic skills and thus they are relatively cheap.

The segmentation of speech requires phonetic or phonological expertise, which makes it expensive. This is especially true when working with dialects or less common languages, for which there are only a few experts in the world, or when working on recordings with low audio quality. Automatic segmentation dramatically reduces the time needed for segmentation, but still needs manual validation and correction, and thus requires expensive human experts.

3. *Transcription Types*

A transcript is a faithful written representation of the content of a spoken language signal. What do *content* and *faithful* mean? In spoken language, content is much more than the sequence of words of an utterance. Besides the verbal content, the signal contains information about the speakers: age, gender, socio-demographic background, emotional state, communication situation, etc.[3]

When creating a speech database, one has to weigh the options: on the one hand, a fine-grained transcription of the verbal, paraverbal and nonverbal content is expensive and may be of use to only a few researchers. On the other hand, a concise summary of the verbal content of the recording will serve to quickly select material of interest, but it is not sufficient for any in-depth analysis.

In the following, I will present the different types of transcriptions for the same short fragment of an oral history interview. The interview is part of the Anna Maria Bruzzone archive on the Ravensbrück concentration camp (Beccari Rolfi & Bruzzone 2020, Vangelisti et al. 2019). The interview was conducted by Anna Maria Bruzzone (AMB), the interviewee is Lidia Beccari Rolfi (L), a survivor of the concentration camp. The topic of this fragment is how using a spoon to eat one's soup counteracts the systematic dehumanisation in the camp.[4]

3.1 Broad transcript

This type of transcript is close to the original recording, focuses on the key issues, and features punctuation and text smoothing etc. to make it easy to read (see Figure 3 for an example).

[3] When listening to the utterance shown in Figure 1, you will quickly notice that it is a sentence read by a male, non-native – possibly German – speaker producing an Italian sentence. You might guess his age, and perhaps even his weight and height.

[4] The fragment is 1:04 minutes long. It was kindly provided by Silvia Calamai of Siena University and is available at https://www.phonetik.uni-muenchen.de/forschung/Bas/Experimente/aitla/cucchiaio.wav.

Figure 3 - *Broad transcript*[5]

Si tratta della storia del cucchiaio. Lidia spiega che l'obiettivo è quello di disumanizzare, di ridurre al livello degli animali, in modo psicologico di far sentire istintivamente come degli animali. Ed è per questo che la prima cosa che un deportato riceve nel campo è un cucchiaio, per sentirsi meno animale, perché c'è sempre e solo zuppa, e senza cucchiaio devi berla o leccarla via. La frase ripetuta più e più volte era: "non siamo le pulci di un cane".

This type of transcript is used to summarise the content of an interview, e. g. for a presentation, and to facilitate quick browsing through a collection of recordings.

3.2 Verbatim raw transcript

In general, a verbatim orthographic raw transcript uses standard orthography and optionally a (very small) set of mark-up symbols to denote specific paraverbal and nonverbal phenomena. The orthographic transcript is as close as possible to what was said, and includes word repetitions, hesitations, and repairs etc. (see Figure 4).

Figure 4 - *Raw verbatim transcript*

Il discorso del cucchiaio. Del cucchiaio ecco, del perché non danno il cucchiaio. Il discorso del cucchiaio rientra, rientra nello stesso tipo di discorso: volendoti disumanizzare, ridurti a livello di bestie: la bestia lecca. Allora istintivamente, tu già psicologicamente sei pronto a sentirti animale. E tant'è vero che una delle prime cose che il deportato acquista, in campo, è il cucchiaio. Per non sentirsi bestia. Cioè chi: tenta di reagire, o lo ruba, o o lo compra, o lo acquista con il pane, ma acquista 'sto cucchiaio per potersi sentire meno animale. E perché se fosse stato qualcosa che si mh potesse mangiare con le mani, le mani le mani tutto sommato sono uno strumento. Sì sì certo Ma, visto che è sempre la minestra, sempre minestra, la devi leccare. Bere o leccare. E lì allora ti senti e effettivamente a livello di di di, le frasi che ricorrevano che, che ricorrevano erano: "Non siamo mica dei cani". Eh già: "Non siamo mica dei cani".

This type of transcript is often used to compute the ASR word error rate (see section 4 for details), for language modelling with n-grams, or as input to time-alignment tools (see section 5.3 on WebMAUS for an example of such a tool). For interviews, this transcript generally focuses on the main speaker, i.e. backchannel feedback often is not transcribed.

3.3 Transcript with diarisation

Diarisation adds explicit information on who is speaking and how speaker roles change to the transcript. It divides the transcript into turns which are labelled with a code for the speaker (see Figure 5).

[5] English translation: «This is the story of the spoon. Lidia explains that the goal is to dehumanize, to reduce to the level of beasts, in a psychologic way, to make one feel instinctively, like animals. And this is why the first thing a deported person tries to obtain is a spoon, to feel less like an animal. It's always soup, and only soup, and without a spoon one has to drink it or lick it. The ever-repeated saying was: "we are not the fleas of a dog.»'. This summary was kindly provided by Lorenzo Spreafico.

Figure 5 - *Verbatim transcript with speaker diarisation*

Spk	*Manual verbatim transcript*
L	Il discorso del cucchiaio
AMB	Del cucchiaio ecco, del perché non danno il cucchiaio.
L	Del cucchiaio rientra, rientra nello stesso tipo di discorso:
AMB	Certo certo
L	volendoti disumanizzare, ridurti a livello di bestie: la bestia lecca.
AMB	Sì sì certo certo
L	Allora istintivamente, tu già psicologicamente sei pronto a sentirti animale.
AMB	Sì sì sì
L	E tant'è vero che una delle prime cose che il deportato acquista, in campo, è il cucchiaio. Per non sentirsi bestia. Cioè chi: tenta di reagire, o lo ruba, o o lo compra, o lo acquista con il pane, ma acquista 'sto cucchiaio per potersi sentire meno animale. E perché se fosse stato qualcosa che si mh potesse mangiare con le mani, le mani le mani tutto sommato sono uno strumento.
AMB	Sì sì certo certo
L	Ma, visto che è sempre la minestra, sempre minestra, la devi leccare. Bere o leccare. E lì allora ti senti e effettivamente a livello di di di, le frasi che ricorrevano che, che ricorrevano erano: "Non siamo mica dei cani".
AMB	Eh già: "Non siamo mica dei cani".

This type of transcription is often used to perform automatic processing of the individual speakers' contributions, for content analysis, and for statistical analyses.

3.4 Transcript with technical and interpretive mark-up

Mark-up adds information to the text in the form of reserved codes. In Figure 6, each speaker turn begins with a speaker code, followed by a timestamp and the transcript text. Within the text, tags written as <…> contain mark-up code, e. g. <OVL> for *overlapping speech*, <BCH> for *backchannel feedback*, or <REP> for *repetition*.

Figure 6 - *Transcript with speaker diarisation and technical and interpretive mark-up*

Spk	*Time (s)*	*Manual verbatim transcript with mark-up*
L	0.00	Il discorso del cucchiaio
AMB	1.14	<OVL> Del cucchiaio ecco, del perché non danno il cucchiaio.
L	1.14	<OVL> Del cucchiaio rientra, rientra nello stesso tipo di discorso:
AMB	7.36	<BCH> Certo certo
L	7.99	volendoti disumanizzare, ridurti a livello di bestie: la bestia lecca.
AMB	10.79	<BCH> Sì sì certo certo
L	12.55	Allora istintivamente, tu già psicologicamente sei pronto a sentirti animale.
AMB	17.90	<BCH> Sì sì sì
L	19.12	E tant'è vero che una delle prime cose che il deportato acquista, in campo, è il cucchiaio. Per non sentirsi bestia. Cioè chi: tenta di reagire, o lo ruba, o o lo compra, o lo acquista con il pane, ma acquista 'sto cucchiaio per potersi sentire meno animale. E perché se fosse stato qualcosa che si <FIL> potesse mangiare con <REP>le mani, le mani le mani</REP> tutto sommato sono uno strumento.
AMB	46.42	<BCH> Sì sì certo certo

L	48.59	Ma, visto che è sempre la minestra, sempre minestra, la devi leccare. Bere o leccare. E lì allora ti senti e effettivamente a livello <REP>di di di</REP>, le frasi che ricorrevano che, che ricorrevano erano: "Non siamo mica dei cani".
AMB	61.53	Eh già: "Non siamo mica dei cani".

This type of transcription is used for in-depth linguistic, discourse, content, and other analyses. The exact format and the extent of the mark-up varies with the research discipline – for linguistic analysis it will be different from that for sociological research. Transcription guidelines define the set of allowed mark-up tags, and their syntax.[6]

3.5 Discussion

The main reason for transcribing recordings is to obtain transcripts which can be used for further research, i.e. which can be browsed, searched, analysed, processed, or visualised.

In general, a raw verbatim transcript with diarisation is considered the basis for all subsequent processing steps. It is theory-neutral, supports the creation of a lexicon and word frequency lists, and creating such a transcript does not require special skills, apart from acute hearing and good knowledge of the language and its orthography. *Acute hearing* here means that the transcriber must be able to separate the different speakers by their voice, and should reliably distinguish and recognise the sounds of the language, even under adverse acoustic conditions or in accented speech.

Depending on the technical quality of the recording and the familiarity of the transcriber with the content, the transcription factor typically lies in the range of 5-10.

A broad transcript can be generated from a raw verbatim transcript, either manually or automatically. Topic detection identifies the main topics of the transcript and provides a structured, possibly statistical or graphical representation, whereas summarisation tools generate new text from the transcript. Both methods can be fine-tuned via parameters, e. g. to deliver only the *n* most relevant topic items, or to generate a summary of *m* words.

Mark-up adds para- and nonverbal information to a transcript. Technical information, such as timestamps, is easy to add. Basic syntactic information, such as part-of-speech tags can be provided automatically with a high degree of precision. Other information, e.g. named entity recognition, also achieves good results, but requires manual verification. Higher-level features such as e.g. discourse strategies or establishing a common semantic ground, require multi-facetted analyses and thus cannot be annotated automatically. Other analyses, e. g. emotion recognition, process the spoken language signal, and again may show their results in mark-up tags in the transcript.

[6] Note that any mark-up should be formatted in such a way that it can be removed mechanically without damaging the transcript, and that it can be searched using regular expressions.

Because each research discipline views the transcript under different aspects, there is no common agreement on what to mark-up. Transcribing a text with mark-up has a transcription factor of 50 or more (in the case of manual phonetic transcription, up to 1000!). It requires specific skills and knowledge, which make it expensive: phonetic analytical hearing, recognition of syntactic structures, discourse analysis, psychology, and others.

To summarize: because of the relatively low effort necessary, both in terms of required skills and cost, the minimum level of transcription should be a raw verbatim transcript with diarisation. This transcript is the prerequisite for identifying material of interest for further processing – this selection will dramatically reduce the amount of data that needs in-depth annotation. The question now is: can automatic speech recognition contribute to the creation of such transcripts?

4. *Automatic Speech Recognition*

Today, automatic speech recognition is everywhere: one can dictate messages on a handheld device, request services or buy products via the internet, or communicate with household devices or cars. ASR has become a business. On the one hand this means that the technology has become affordable and sufficiently reliable for everyday applications, on the other hand this means that by focusing on commercially interesting languages and applications, the requirements of non-commercial applications, e.g. humanities research, receive little or no attention.

The most frequently used measure for ASR performance is the *word error rate* (WER). It computes the number of insertions, deletions and substitutions necessary to transform a hypothesis, i.e. the ASR output, to a reference or *gold standard*, i.e. the transcript generated by a human expert.

In some domains, e.g. medical or juridical dictation, ASR has reached human performance levels, with WERs of 2-5%; the same is true for travel or other information systems. In other domains, e.g. dictating personal messages, performance is very often good, but occasionally – and for no apparent reason – ASR fails miserably. And in yet other domains, e.g. linguistic field recordings, dialogues, colloquial conversation at home, and others, ASR consistently performs badly, with WERs of 40% and more.

For us humans, this is difficult to understand: highly complex application areas such as medicine or law, with their difficult terminology and non-natural way of saying things – how can a machine do it? And seemingly simple things, like having a conversation with grandfather in the living room about some family event, are impossible? How can this be?

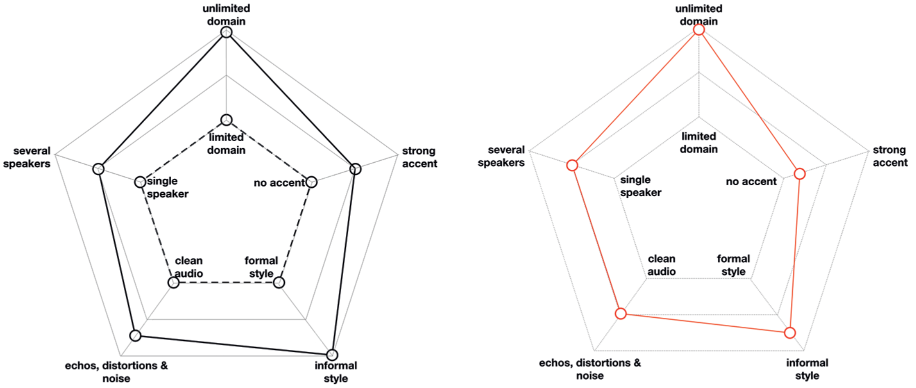
ASR performance depends on a number of factors. A simple graphical model, based on five factors, may help to estimate the quality of ASR for a given spoken language recording. Figure 7 shows a pentagon with the five factors (anti-clockwise):

1. Number of speakers
2. Audio quality

3. Speaking style
4. Dialect or accent
5. Domain of discourse

The innermost pentagon contains the values of the factors favourable to ASR: a single speaker, high quality recordings with very little background or technical noise, a formal style of speaking, e.g. reading or speech with no overlap, standard language, and a limited domain, i.e. a small vocabulary. The outermost pentagon displays values that are detrimental to ASR performance: several speakers, echoes and distortions in the signal, informal way of speaking, i.e. a lot of overlapping speech and frequent interruptions, language with a strong accent or dialect, and unlimited domains, i.e. virtually unrestricted vocabularies.

Figure 7 - *The left pentagon displays the the values of the factors for the medical dictation (dashed line) and the family conversation (solid line). The right pentagon displays the values for the interview with the cucchiaio story*



With the help of the pentagon it becomes clear why medical dictation performs well: only one person speaks at a time, often headset microphones are used to keep the hands free and to reduce background noise, the communication is highly structured, doctors employ a clear language in their everyday professional communication, and, finally, medical fields have a limited terminology.

For the conversation with grandfather at home, the picture is different: there are many speakers, there is only one microphone on the table, people enter and leave the room, and maybe even the TV is on, the participants are familiar to each other and thus interrupt each other frequently, the language may be accented, dialectal or a family language, and, of course, one talks about almost anything imaginable, i.e. the domain is unlimited.

Both the medical dictation and the family conversation are shown on the left in Figure 7.

The interview fragment is shown on the right in the same figure: two speakers, audio signal with the high frequencies missing and a soft noise, a dialog with crosstalk and frequent interruptions, Italian with a weak accent, and a potentially unlimited vocabulary with strong emotional content. The prediction is that the word error rate will be rather high for this particular recording.

In fact, when this interview fragment was processed with the Google Speech Cloud ASR using the BAS web services in January 2021, it generated the transcript in the right column of Figure 8. With text normalisation, i.e. converting all text to lower case, removing speaker labels, punctuation and other markers, the word error rate for this fragment is 48.5%[7].

4.1 Analysis of the ASR-generated transcript

The output of the commercial ASR provider is a normalised and moderately smoothened verbatim transcript of the interview recording, with punctuation and capitalisation. Quite a lot of the spoken language signal was recognised faithfully.

However, there are severe problems: the ASR transcript

- is approx. 33% shorter (114 vs. 169 words) than the manual transcript because entire phrases are missing, e.g. the first two sentences, 'la bestia lecca', and 'le mani sono uno strumento';
- contains some totally unexpected words, e.g. 'distrarre', 'Giusti', 'desti';
- contains a few word duplications, e.g. 'che che', but far fewer than the manual transcript;
- is not diarised.

For a human reader, especially the two first problems are surprising: there is no apparent reason for the omission of sentences or the selection of these words.

Other errors are explicable, e.g. 'tecnologicamente' vs. 'psicologicamente', or 'ma aquista 'sto' vs. 'ma qui sta sto' and 'Eli. Allora' vs. 'E li allora' because they sound similar, or 'Campo' as a capitalized named entity instead of a simple noun.

The transcript created by human transcribers is diarised, and it contains a number of interesting word or phrase repetitions that are missing in the automatically generated transcript. They are interesting because they illustrate the communication situation or give insight into the speaker's emotional state:

- 'sì sì certo certo' etc. as backchannel feedback to the speaker to continue,
- 'di di di' and 'mh' as a consequence of searching for the right words, and
- 'sempre la minestra, sempre minestra' places extra emphasis on this topic.

Higher-level phenomena, such as the citation status of "Non siamo mica dei cani.", require linguistic knowledge and insight into the communicative function of phrases within the utterance. They are thus outside the scope of ASR.

A transcript for research purposes must contain these phenomena to allow researchers to decide whether this recording will be analysed in more detail.

[7] For the entire interview, which has longer passages where only one person is speaking, the WER is 37.9%. WER was computed using the wer() function in the wersim R-package by Jens Wäckerle (Proksch et al. 2018).

Today's commercial ASR does not deliver this type of transcript for two main reasons:

1. For maximum performance and efficiency for a given task, ASR has to be customised to this task, and because this is expensive, there has to be commercial interest.
2. Many of the phenomena of interest to research, e.g. dialects, accents, or vernacular language, or specific linguistic structures, or cultural or sociological analyses, do not have direct commercial potential.

As a consequence, research in ASR needs to address the requirements of humanities disciplines, e.g. oral history, sociolinguistics, and others.

Figure 8 - *Comparison of the ASR output and a manual verbatim transcript. The transcripts were diarised and formatted manually to improve legibility*

#	*Spk*	*Manual verbatim transcript*	*Google Speech Cloud (Jan. 2021)*
1	AMB	Del cucchiaio ecco, del perché non danno il cucchiaio.	
2	L	Del cucchiaio rientra, rientra nello stesso tipo di discorso:	Rientra nello stesso tipo di distrarre
3	AMB	Certo certo	
4	L	volendoti disumanizzare, ridurti a livello di bestie: la bestia lecca.	Volendoti disumanizzare le Giusti a livello di bestia.
5	AMB	Sì sì certo certo	
6	L	Allora istintivamente, tu già psicologicamente sei pronto a sentirti animale.	Allora Istintivamente tu già tecnologicamente sei sotto a sentirti animale.
7	AMB	Sì sì sì	
8	L	E tant'è vero che una delle prime cose che il deportato acquista, in campo, è il cucchiaio.	Tanto è vero che una delle prime cose che il deportato acquista in Campo ecco che hai.
9		Per non sentirsi bestia.	Per non sentirsi desti.
10		Cioè chi: tenta di reagire, o lo ruba, o o lo compra, o lo acquista con il pane, ma acquista 'sto cucchiaio per potersi sentire meno animale.	Cioè chi è tenta di reagire o lo ruba o lo compra all acquista con il pane, ma qui sta sto cucchiaio per potersi sentire meno animale,
11		E perché se fosse stato qualcosa che si mh potesse mangiare con le mani, le mani le mani tutto sommato sono uno strumento.	Perché se fosse stato qualcosa che potresti mangiare con le mani in mano,
12	AMB	Sì sì certo certo	
13	L	Ma, visto che è sempre la minestra, sempre minestra, la devi leccare. Bere o leccare.	Visto che è sempre la minestra la devi leccare leccare
14		E lì allora ti senti e effettivamente a livello di di di, le frasi che ricorrevano che, che ricorrevano erano: "Non siamo mica dei cani".	Eli. Allora ti senti effettivamente a livello di Le frasi che che ricorre vero
15	AMB	Eh già: "Non siamo mica dei cani".	non siamo mica del cane.

4.2 Is transcribing based on ASR faster than purely manual transcription?

Currently, research is under way to estimate what level of ASR performance must be achieved to effectively speed up the generation of verbatim transcripts. In a pilot study performed in 2019 at our institute, 10 student presentations on the topic of 'communication' with a duration between 1:14 min and 5:34 min (total 38:14 min), recorded via a standard video camera in a lecture hall, were transcribed by two transcribers. Each transcriber generated 5 transcriptions from scratch, and 5 by correcting ASR output. Overall, the real time factor for the transcription from scratch was 9.43, and 8.52 for the transcription based on ASR – a speed-up of approximately 10%.

In this pilot study, the European Media Lab recogniser was used, and it achieved a WER of 68.9%. This might also explain why the real time factors are so similar: because of the high WER, the transcribers simply deleted some of the ASR-generated transcript and transcribed from scratch.

In 2021, the WER was computed for four of the student presentations and for three different ASR providers, and between the human-created transcripts. Table 2 shows that human-created transcripts are very close to each other, and that for this type of data – several speakers, low signal quality with reverberation, and unknown domain – the performance of ASR varies greatly.

Table 2 - *Averaged WERs for three different ASR systems and four selected student presentations; the last row displays the WER between the two human-generated transcripts*

ASR	*WER*
EML	68.9%
Google Speech Cloud	58.7%
Fraunhofer	23.2%
Human/human	3.7%

In a pilot study performed by Silvia Calamai's group for this paper[8], the interview of Anna Maria Bruzzone and Lidia Beccari Rolfi was re-transcribed (approximately 41 minutes). To compare transcription speed, the interview was split into five fragments of 5:28-11:25 minutes. Two fragments were transcribed from scratch, three by first running the Google Speech Cloud ASR and then manually correcting this transcript. Table 3 presents the results. Overall, re-transcription was very quick, with an average real time factor of 3.52. Transcribing from scratch was slightly faster than correcting the ASR generated transcript. As expected, the ASR WER is high. The relatively high WER in the comparison of the two transcripts created by human transcribers – especially with regard to the student presentations in Table 2 – may be attributed to the interview situation with overlapping speaker turns and the audio quality.

[8] The data presented here was provided by Fabio Ardolino from Siena University.

Table 3 - *Comparison of transcribing from scratch vs. correcting an ASR generated transcript*

Fragment	*Playtime*	*Type*	*Real time factor*	*WER*	*Human/human*
BRZTO061a_16_0_0002	7:04	scratch	3.28	35.5%	11.6%
BRZTO061a_16_0_0003	8:38	ASR	3.79	42.1%	14.5%
BRZTO061a_16_0_0004	8.51	scratch	3.22	41.3%	14.0%
BRZTO061a_16_0_0005	11:25	ASR	3.41	40.5%	14.5%
BRZTO061a_16_0_0006	5:28	ASR	3.90	37.6%	14.5%
Average			3.52	39.5%	13.8%

Gorisch et al. (2020) is one of the few published studies on the use of ASR for the transcription of spoken language corpora. For different German corpora, it reports WERs between 13.3% and 30.0% for the Fraunhofer ASR, and correlates regional accents and ASR performance. In a private communication, the head of the spoken language archive at the Leibniz-Institut für Deutsche Sprache Mannheim, Thomas Schmidt, states, that for TV and radio broadcast with a WER of approximately 10%, the use of ASR is generally worthwhile. For interview data, it is recommended to run ASR tests on a subset of the corpus and then decide, and for natural conversations in field recordings it is more efficient to transcribe from scratch.

The following section describes how state-of-the-art ASR systems can be accessed via easy-to-use web services by academic users.

5. *Web services at BAS*

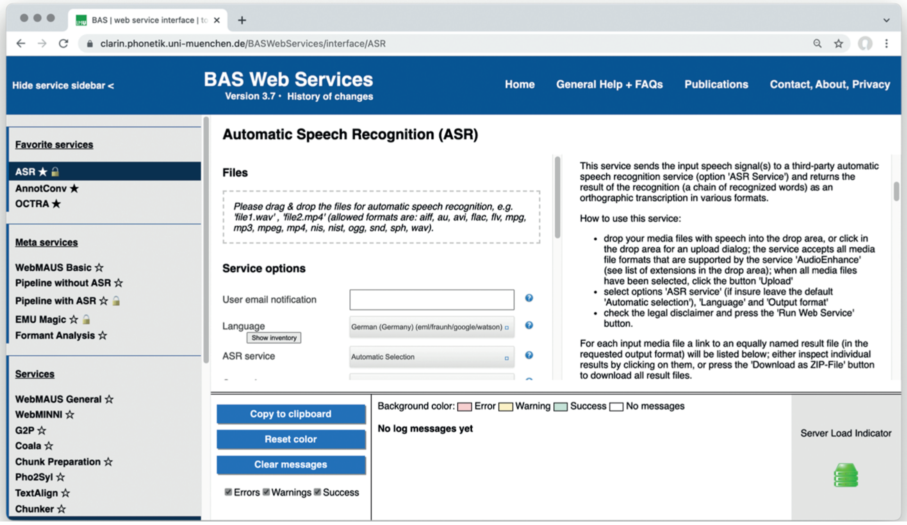
The Bavarian Archive for Speech Signals (BAS) is a German CLARIN centre. It operates a repository for spoken language resources, and it offers a range of web services which make available spoken language processing tools to non-technical users. The use of the BAS web services is restricted to members of academic institutions[9].

The web services user interface is basically the same for all services (see Figure 9 for an example): on the left side, there is a sidebar with all available services; this sidebar can be hidden or shown. In the top middle is a panel with the file upload area and a form to set the service parameters. A click on the question marks next to the parameter fields opens a description of the selected parameter. On the right is a documentation of the service; again, this documentation can be hidden or shown. The bottom row is the feedback area: here, colour-coded status, warning or error messages are displayed, and the load indicator displays the current workload of the server.

Using a web service is easy: upload files from the local computer to the server, select a service and set some parameters, accept the conditions of use and run the service. Once a result has been computed, it can be viewed in the browser, e.g. using the Emu WebApp (Winkelmann 2015, Winkelmann et al. 2017), or downloaded to the local computer. Note that the most important parameters or 'service options' are always shown; many services also have optional parameters which can be accessed via a click on 'expert options'.

[9] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface.

Figure 9 - *Detail page for the ASR web service. On the left, the sidebar with the list of all services is shown, and on the right the manual page for the selected service*



Automatic Speech Recognition (ASR)

Files

Please drag & drop the files for automatic speech recognition, e.g. 'file1.wav' , 'file2.mp4' (allowed formats are: aiff, au, avi, flac, flv, mpg, mp3, mpeg, mp4, nis, nist, ogg, snd, sph, wav).

Service options

User email notification

Language: German (Germany) (eml/fraunh/google/watson)

ASR service: Automatic Selection

This service sends the input speech signal(s) to a third-party automatic speech recognition service (option 'ASR Service') and returns the result of the recognition (a chain of recognized words) as an orthographic transcription in various formats.

How to use this service:

- drop your media files with speech into the drop area, or click in the drop area for an upload dialog; the service accepts all media file formats that are supported by the service 'AudioEnhance' (see list of extensions in the drop area); when all media files have been selected, click the button 'Upload'
- select options 'ASR service' (if insure leave the default 'Automatic selection'), 'Language' and 'Output format'
- check the legal disclaimer and press the 'Run Web Service' button.

For each input media file a link to an equally named result file (in the requested output format) will be listed below; either inspect individual results by clicking on them, or press the 'Download as ZIP-File' button to download all result files.

Background color: 🟥 Error 🟨 Warning 🟩 Success □ No messages

No log messages yet

☑ Errors ☑ Warnings ☑ Success

The following sections present a selection of web services of particular interest to linguists and phoneticians. The presentation follows the transcription workflow shown in Figure 2.

These services are (with the service names in parentheses)

1. Automatic Speech Recognition (ASR),
2. Grapheme-to-phoneme conversion (G2P),
3. Automatic phoneme and word alignment (WebMAUS), and
4. Anonymisation (Anonymizer)

Again, the interview with the cucchiaio story from the Bruzzone archive, or an excerpt from it, is used throughout this section to show how the services work and what they return.

5.1 Automatic Speech Recognition

The ASR web service calls on external ASR providers. Currently, these providers are

- European Media Lab (Germany),
- Fraunhofer Intelligente Analyse- und Informationssysteme (Germany),
- Google Speech Cloud (US),
- IBM Watson (US), and
- LST (Language and Speech Tools), Radboud University, (Netherlands).

Most providers support more than one language. The ASR service by LST is an academic service, and allows the selection of different content domains, e.g. conversational speech, parliamentary discussions, or oral history[10].

[10] Automatic Transcription of Dutch Speech Recordings, Language and Speech Tools, Radboud University, Nijmegen: https://webservices.cls.ru.nl/oralhistory.

The BAS web services use the free services of these providers. As a consequence, a number of limitations apply, depending on the provider. Such limitations concern the maximum audio duration, or a given quota per month, etc.[11] Note that some ASR providers keep the uploaded audio files, which is often not acceptable for privacy reasons.

The ASR web service requires an authentication – this makes sure that only academic users access the service. Members of a European academic institutions should be able to log in using credentials of the account at their home institution[12].

Once authenticated, the user can upload the audio files either via drag & drop to the marked area in the browser, or via a click on the marked area to select the files using the standard file selection dialog of the local computer. Note that several files can be loaded at once.

For ASR, the language to recognise, the provider of the service, and the output format must be set.

For the cucchiaio story, the language was set to Italian, Google was chosen as the ASR provider, and .txt as the output format. The result of ASR is shown in the middle column of Figure 8.

5.2 Grapheme-to-phoneme conversion

The tool G2P converts text in standard orthography into a phonemic representation. Technically, the service is based on statistic decision trees, part-of-speech tags and morphological segmentation. The service is trained on pronunciation dictionaries, or on a letter-sound table for languages with a unique correspondence between letters and sounds (e.g. Italian, Finnish). See (Reichel 2012) for details. G2P is available for 50+ languages and dialects, and users may upload their own letter-sound mapping table.

In Figure 10 the second sentence of the interview is processed by the G2P web service using Italian as the language, the output symbol inventory SAM-PA, and the output format .txt.

Figure 10 - *Grapheme-to-phoneme conversion for a sentence from the interview*

Orthography	*Pronunciation (SAM-PA)*
Del cucchiaio ecco, del perché non danno il cucchiaio.	d e l k u k: j a j o E k: o d e l p e r k e n O n d a n: o i l k u k: j a j o

[11] To lift some of these restrictions, users may purchase a key. With this key, they will be charged for the use of the service.

[12] As an alternative, users may apply for a CLARIN account with CLARIN-EU: https://idm.clarin.eu/user/home.

5.3 Automatic phoneme and word segmentation

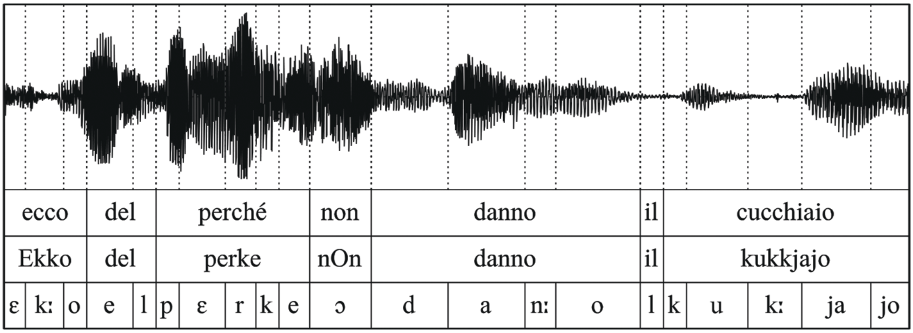
The Munich Automatic Segmentation (MAUS) time-aligns an audio file and its orthographic transcript to return a multi-tier phonetic transcript. MAUS internally generates pronunciation hypotheses from the orthographic transcript using G2P, and then computes the phoneme sequence that matches the audio file best. MAUS is available for 50+ languages. Furthermore, a language-independent mode takes as input a phoneme string in SAM-PA and thus allows the segmentation of languages for which there is no dedicated model. See (Kisler et al. 2012) for further information.

MAUS exports segmentations to different output formats, such as .TextGrid for Praat, .eaf for ELAN, or .csv for spreadsheets or statistics packages. MAUS comes in two flavours: WebMAUS Basic with a limited set of options, and WebMAUS General with a large number of configuration parameters.

Figure 11 displays a time-aligned transcript, with orthographic words in the top level, the canonic word pronunciation in the middle and phoneme segments on the bottom level. The input format was BAS partitur file format, the language Italian, the output format Praat TextGrid, and the output in the IPA alphabet in UTF-8 encoding.

Here, the actual pronunciation deviates from the canonical pronunciation: the /d/ in 'del', both /n/ in 'non' and the /i/ in 'il' are not in the phonemic segmentation. This may be due to coarticulation in fluent speech, but also due to low signal quality of the recording – here, it is probably both[13].

Figure 11 - *MAUS segmentation of the second sentence of the transcript of the cucchiaio story*



WebMAUS works best for high quality audio files with a single speaker. Because processing time is quadratic, WebMAUS works fast only for short files. In practice, 10 min or 3000 words are recommended as maximum file duration or transcript length. To process longer files, the use of a pipeline service with the chunker service is recommended (Pörner & Schiel 2018).

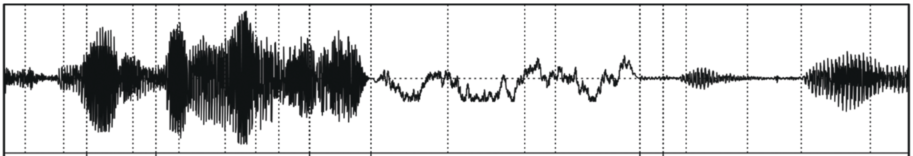
[13] Of course, we also cannot rule out an error in the segmentation algorithm.

The performance of automatic segmentation systems has been shown to depend strongly on the type of language and the method applied for the comparison (for details, see e.g. (McAuliffe et al. 2017, or Meer 2020)).

5.4 Anonymizer

For interview data, it is often necessary that both transcripts and audio files be anonymised. The Anonymizer web service takes as input an audio file and its time-aligned segmentation plus a list of terms to anonymise. The service searches for these terms in the segmentation, extracts the boundaries from the segmentation, replaces the transcript text with a marker symbol, and replaces the matching signal fragment with brown noise.

Figure 12 - *Output of the Anonymizer service with the word 'danno' anonymised. Note that in the signal, the fragment corresponding to the word is replaced by noise, and the word is removed from the transcript*



ecco	del	perché	non	ANON	il	cucchiaio
Ekko	del	perke	nOn	ANON	il	kukkjajo

ɛ | kː | o | e | l | p | ɛ | r | k | e | ɔ | _ | _ | _ | _ | l | k | u | kː | ja | jo

5.5 Pipeline services

The web services presented in the previous sections were run individually. This means that for every service, files have to be uploaded to the server, and parameters have to be set. In the transcription workflow, the output of one service often is the input to the next, and thus it makes sense to organise individual services into a pipeline to reduce the number of file transfers and manual interactions.

BAS provides such pipeline services as pre-configured sequences of individual services. Files are uploaded only once and then are passed on from one process to the next until done. Pipelines not only streamline the use of the services, they also make them safe by allowing only meaningful combinations of services and parameters (see Kisler et al. 2017 for details).

For example, to run the four services from the previous sections, one has to upload the audio file three times (for ASR, WebMAUS, and Anonymizer), and download a .txt, a .par and two .TextGrid files.

The same result can be achieved by using the BAS pipeline service *ASR→G2P→MAUS→Anonymizer*. The audio file needs to be uploaded only once, the services are executed sequentially on the server, and the final output file is returned. Of course, the pipeline can be parameterised. For example, by selecting the output format .csv (for comma-separated values), the resulting table can be imported directly into a spreadsheet or statistics package, e.g. to generate a frequency list, perform phoneme counts, or calculate mean durations and variance (see Table 4 for an example).

5.6 Final remarks on the BAS web services

The graphical user interface of the BAS services allows the user to upload any number of files; on the local computer, they may reside in different folders. During upload, and if required by a given service, the server will try to find matching files i.e. files with the same base name and different extensions. Non-matching files will be reported.

If processing is expected to take long, users may choose the email notification option: the BAS services notify the user via email when processing is done, and the email contains a link to the result in the format of a compressed archive file.

Finally, all BAS web services may be accessed programmatically via an API. This is particularly useful for very large numbers of files or repeated tasks, or when working in environments such as jupyter notebook, R Studio, or the command line. See the FAQ section and the documentation of the web services for further details. A number of transcription and annotation editors, e.g. Octra or ELAN, may call the BAS web services in the background. This allows running the services on the data currently open in the editor, without leaving the application.

Table 4 - *Segmentation data imported into a statistics package (formatted to improve legibility)*

Token	*Begin (s)*	*Label*	*Duration (ms)*
	...		
41	20.070	una	130
42	20.200	delle	500
43	20.700	prime	360
44	21.060	<P>	60
45	21.120	cose	480
46	21.600	<P>	80
47	21.680	che	140
48	21.820	il	30
49	21.850	<P>	50
50	21.900	deportato	700
51	22.600	<P>	100
52	22.700	acquista	430
53	23.130	in	90
54	23.220	<P>	90
55	23.310	Campo	550
	...		

6. *Summary and outlook*

The transcription of spoken language recordings is a time-consuming task, and ASR promises to speed up the generation of transcripts. Currently, ASR works well under specific conditions – the ASR pentagon can give an estimate of the expected quality of ASR output. For contemporary or future recordings, the specific requirements of ASR can be accounted for already during the recording, and thus ASR will speed up the generation of broad or verbatim orthographic transcripts. For legacy recordings, often with low audio quality, strongly accented speech, unlimited domains etc. ASR does not yet perform well enough to effectively increase transcription speed – correcting an ASR-generated transcript with a high word error rate is often slower than transcribing from scratch.

Two developments are needed:

- the workflow in the humanities needs to be adapted to make the best use of tools and resources, and
- research and development in ASR should focus on the specific requirements of the humanities.

Currently, orthographic transcriptions and in-depth annotations of recordings are often performed in one step by the same highly-trained transcriber. A modularisation of the workflow, which separates transcription from in-depth annotation, divides hour-long recordings into meaningful units of shorter length, and uses the proper tools at each stage of the workflow, promises more flexibility and a more efficient use of human resources. In such a modularised workflow, new technology, e.g. ASR, can be introduced for specific tasks without adverse effects on others. The key to this modularisation is a smooth flow of data from one processing step to the other, which in turn means that the tools used in the workflow must support each other's formats.

The specific requirements of the humanities with regard to ASR, such as truly verbatim orthographic transcripts with e.g. word repetitions, backchannel feedback, turn taking and pauses, are not well supported by current ASR systems. These topics are open research questions, and humanities scholars and speech technology researchers need to collaborate to answer them.

7. *Acknowledgements*

I thank the team at BAS (Markus Jochim, Thomas Kisler, Julian Pömp, Nina Pörner, Florian Schiel, Fritz Seebauer, Raphael Winkelmann) for developing and maintaining the web services and tools. This work was funded by the German Federal Ministry of Education and Research under contract #01UG1620B. I also thank the Oral History group (Jeannine Beeken, Silvia Calamai, Arjan van Hessen, Henk van den Heuvel, Stefania Scagliola) for the continued discussions and collaboration, and CLARIN-EU for funding parts of this work.

References

Beccaria Rolfi, Lidia & Bruzzone, Anna Maria. 2020. *Le donne di Ravensbrück. Testimonianze di deportate politiche italiane*. Torino, Einaudi.

Boersma, Paul. 2001. Praat, a System for doing Phonetics by Computer. *Glot International* 5(9/10). 341-245.

Draxler, Christoph & van den Heuvel, Henk & van Hessen, Arjan & Calamai, Silvia & Corti, Louise & Scagliola, Stefania. 2020. A CLARIN Transcription Portal for Interview Data. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, 3353-3359.

Draxler, Christoph & Jänsch, Klaus. 2004. SpeechRecorder – A Universal Platform Independent Multi-Channel Audio Recording Software. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, 559-562.

Gorisch, Jan & Gref, Michael & Schmidt, Thomas. 2020. Using Automatic Speech Recognition in Spoken Corpus Generation. In: *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, 6424-6428.

Kisler, Thomas & Reichel, Uwe D. & Schiel, Florian. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326-347.

Kisler, Thomas & Schiel, Florian & Sloetjes, Han. 2012. Signal Processing via Web Services: the Use Case WebMAUS. In *Proceedings of the Digital Humanities Conference 2012*, 30-34.

McAuliffe, Michael & Socolof, Michaela & Mihuc, Sarah & Wagner, Michael & Sonderegger, Morgan. 2017. Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 498-502.

Meer, Philipp. 2020. Automatic Alignment for New Englishes: Applying state-of-the-art Aligners to Trinidadian English. *Journal of the Acoustical Society of America*, 147(4). 2283-2294. DOI: 10.1121/10.0001069.

Pömp, Julian & Draxler, Christoph. 2017. OCTRA – A Configurable Browser-based Editor for Orthographic Transcription. In: *Tagungsband der 13. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, 145-148.

Pörner, Nina & Schiel, Florian. 2018. A Web Service for Presegmenting Very Long Transcribed Speech Recordings. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*.

Proksch, Sven-Oliver & Wratil, Christopher & Wäckerle, Jens. 2018. *Testing the Validity of Automatic Speech Recognition for Political Text Analysis*. Political Analysis.

Reichel, Uwe D. 2012. PermA and Balloon: Tools for String Alignment and Text Processing. In *Proceedings of the Annual Conference of the International Speech Communication Association*, paper no. 346.

R Core Team (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schiel, Florian. 2015. A Statistical Model for Predicting Pronunciation. In *Proceedings if the International Congress of Phonetic Sciences*, paper 195.

Sloetjes, Han. 2007. ELAN: a Free and Open-source Multimedia Annotation Tool. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 4015-4016.

Spreafico, Lorenzo. 2020. Corpora di parlato o corpora di ascoltato? *Rivista italiana di dialettologia* 44. 37-51. https://hdl.handle.net/10863/18479.

Vangelisti, Petra & Pesce, Caterina & Setaro, Marica & Bianchini, Greta & Gigli, Lucilla & Calamai, Silvia. 2019. *Ritrovare Voci: il lavoro intorno all'archivio di Anna Maria Bruzzone*. DOI: 10.17469/O2106AISV000009.

Winkelmann, Raphael & Harrington, Jonathan & Jänsch, Klaus. 2017. Emu-SDMS: Advanced Speech Database Management and Analysis in R. *Computer Speech and Language* 45. 392-410.

Winkelmann, Raphael. 2015. Managing Speech Databases with emuR and the Emu-webApp. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, 2611-2612.