Guidelines for the Annotation of Spoken Language Corpora with Multimodal Traits v.1

Author: Daniela Trotta, Università di Salerno, dtrotta@unisa.it

Introduction

In this document we present the guidelines for the creation and annotation of corpora of spoken language (composed only of the audio source) or multimodal (composed of audio-visual material) in monological or dialogical form. These guidelines have been followed for the creation and the annotation of the PoliModal corpus (Trotta et al., 2019).

The annotation scheme presented here aims to integrate phonological, morpho-syntactic and proxemic information in order to give a description of the language as complete as possible. It can therefore be further extended with additional levels of language description.

The annotation scheme that we present is divided into three levels:

1. annotation of phonological or speech constants (e.g. dialogical organization in turns, use of repetition, use of speech signals, etc.) performed manually following the TEI standard for spontaneous speech*;*

2. morpho-syntactic annotation automatically performed using the TINT syntactic parser (Aprosio & Moretti, 2016);

3. proxemics annotation automatically carried out with ANVIL tool (Kipp, 2001) following the MUMIN annotation scheme (Allwood et al., 2007).

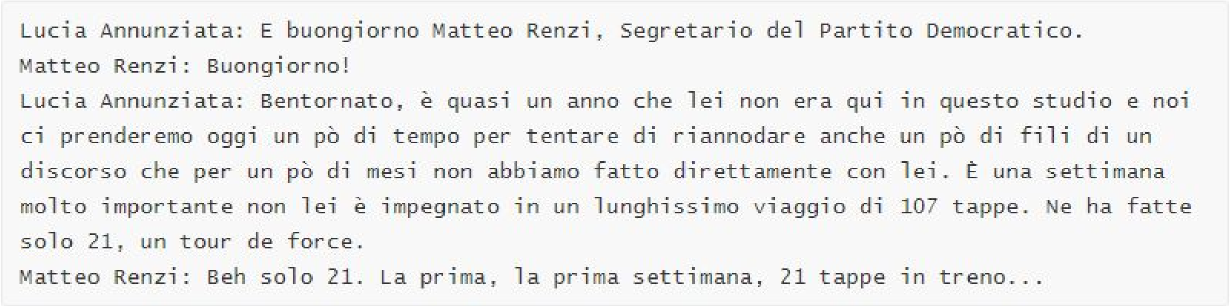
Preliminary step: transcription

Before moving on to annotation, the audio (.pcm, .wav, .aiff, .mp3, .aac, .ogg, .wma, .flac, .alac etc.) or audio-video (.webm, .mpg, .mp2, .mpeg, .mpe, .mpv, .mp4, .mp4, .m4v, .avi, .wmv, .mov, .qt etc.) source must be transcribed in such a way as to obtain a raw text (.txt, .doc, .docx etc.) preferably encoded in UTF-8 (Unicode Transformation Format, 8 bit).

Although there are many tools that can be used for transcription (e.g. PRAAT, OH Portal, OTranscribe, Transcribe etc.) we propose a semi-automatic speech-to-text methodology using Web Using API and manual correction. For using Web Speech API we recommend installing Virtual Audio Cable in this way the audio source can be transcribed.

At the end of the transcription process the raw text will appear without punctuation and there will probably be some errors, so the annotator will make spelling corrections, insert punctuation and in the case of dialogical forms, divide into turns.

At the end of the human review process described above, the text in .txt format and coded in UTF-8 will be ready to be annotated. It will then appear as follows, with a turn per line:



```
Lucia Annunziata: E buongiorno Matteo Renzi, Segretario del Partito Democratico.
Matteo Renzi: Buongiorno!
Lucia Annunziata: Bentornato, è quasi un anno che lei non era qui in questo studio e noi
ci prenderemo oggi un pò di tempo per tentare di riannodare anche un pò di fili di un
discorso che per un pò di mesi non abbiamo fatto direttamente con lei. È una settimana
molto importante non lei è impegnato in un lunghissimo viaggio di 107 tappe. Ne ha fatte
solo 21, un tour de force.
Matteo Renzi: Beh solo 21. La prima, la prima settimana, 21 tappe in treno...
```

The example presented above as well as all the other examples reported in these guidelines are taken from the PoliModal corpus of political interviews in Italian (Trotta et al., 2019).

First level of annotation: speech constants

The first level of annotation wants to keep track of so-called “speech constants” (e.g. dialogical organization in turns, use of repetition, use of speech signals, etc.). Indeed, spoken language deviates systematically and regularly from the written form, and researchers (Biber 1995; Miller e Weinert 1998) have observed that there are constants of speech that make two spoken texts in a natural and spontaneous context similar, even if belonging to different diastratic or diaphasic levels. Thanks to these constants, spoken texts tend to resemble each other more than a spoken and a written text belonging to the same diaphasic and diastratic level.

This level of annotation is largely inspired by the TEI standard for spontaneous speech. However, the tags that we will present are not exactly the same as those present in the standard. In some cases, functional changes have been made in compliance with the research objectives of the PoliModal corpus.

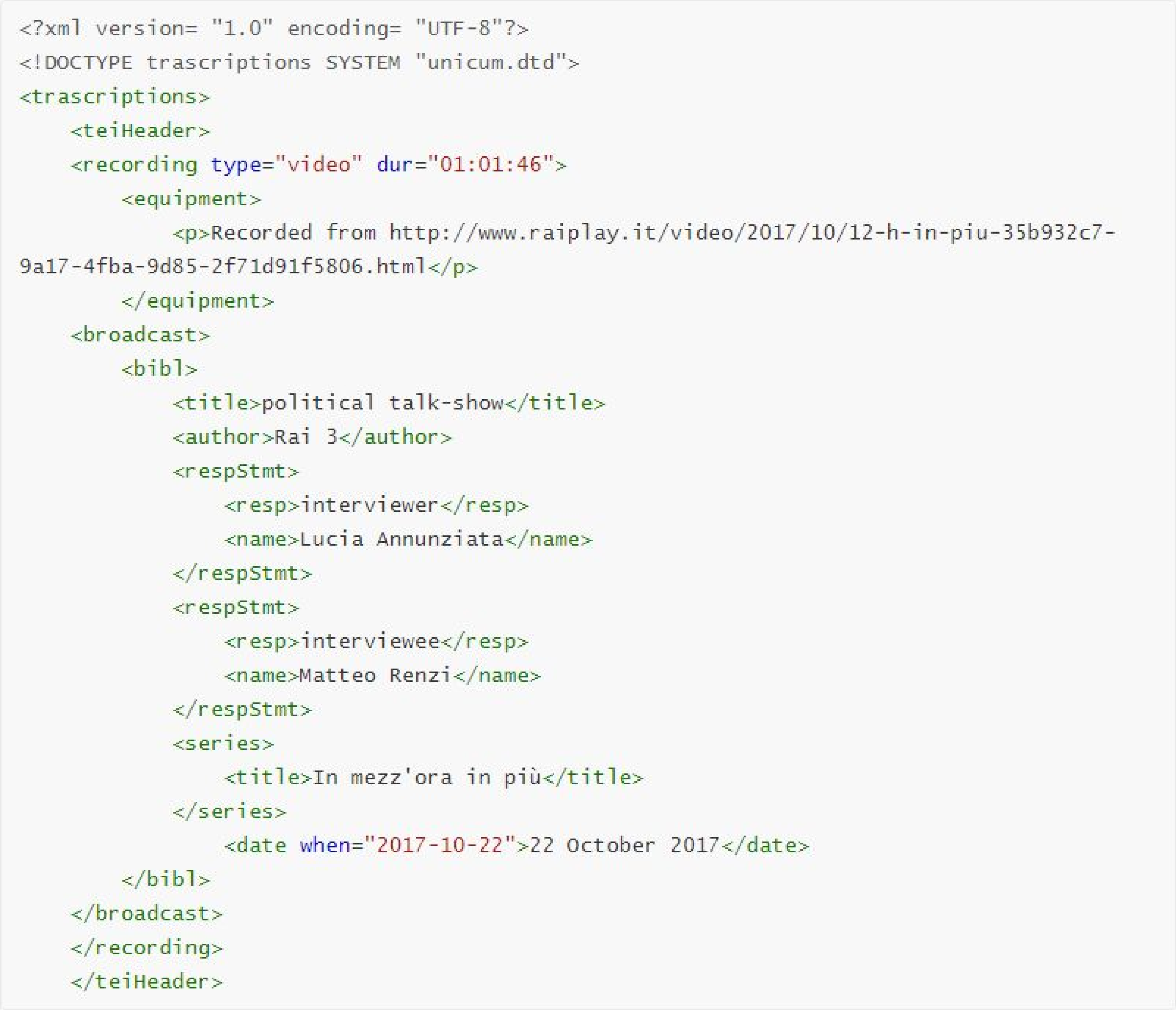
This annotation step will be conducted manually by the annotator using the xml language. Any XML editor can be used for this purpose.

Step 1. Metadata annotation

As recommended by TEI: "*Where a computer file is derived from a spoken text rather than a written one, it will usually be desirable to record additional information about the recording or broadcast which constitutes its source".*

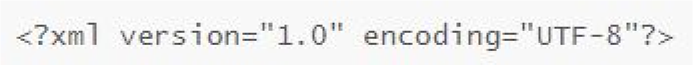
So first of all, for each raw file some metadata must be defined. These include useful information for a quick identification of transcriptions, for example the tools used for the transcription, a link to the file, the owner account, the title, the date etc.

To explain how to write metadata, we use an example taken from the corpus we created, in particular the transcript of the interview with Matteo Renzi, previously mentioned, aired on 22 October 2017.



```
<?xml version= "1.0" encoding= "UTF-8"?>
<!DOCTYPE trascriptions SYSTEM "unicum.dtd">
<trascriptions>
    <teiHeader>
    <recording type="video" dur="01:01:46">
        <equipment>
            <p>Recorded from http://www.raiplay.it/video/2017/10/12-h-in-piu-35b932c7-
9a17-4fba-9d85-2f71d91f5806.html</p>
        </equipment>
    <broadcast>
        <bibl>
            <title>political talk-show</title>
            <author>Rai 3</author>
            <respStmt>
                <resp>interviewer</resp>
                <name>Lucia Annunziata</name>
            </respStmt>
            <respStmt>
                <resp>interviewee</resp>
                <name>Matteo Renzi</name>
            </respStmt>
            <series>
                <title>In mezz'ora in più</title>
            </series>
                <date when="2017-10-22">22 October 2017</date>
        </bibl>
    </broadcast>
    </recording>
    </teiHeader>
```

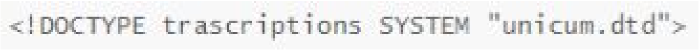
More specifically, the first line of any transcription must have the following format:



```
<?xml version="1.0" encoding="UTF-8"?>
```

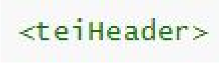
The attribute **version** declares the version of the XML language used, and it is required.

The attribute **encoding** specifies the code used in the document XML to represent the characters, and it refers in general to Unicode. In particular, it can be UTF-8 or UTF-16 (Unicode Transformation Format) 8 or 16 bit, or ISO-8859-1 for ISO-8859 Latin 1 characters. If the encoding is UTF-8 or UTF-16, it can be omitted. However, it is often included for more precise documentation of the XML file.



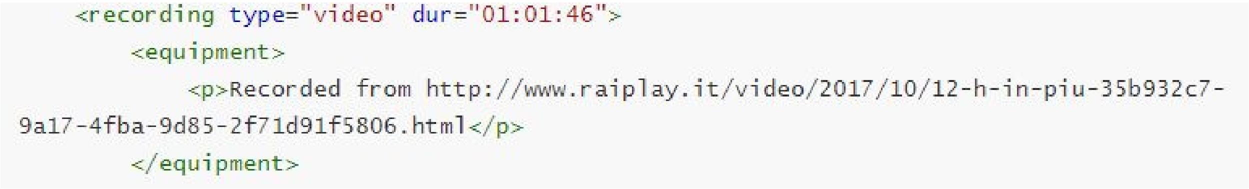
```
<!DOCTYPE trascriptions SYSTEM "unicum.dtd">
```

The **DOCTYPE** is used to tell the browser what type of document it is (in this case "trascrizioni", transcriptions) followed by the declaration of the **DTD (Document Type Definition)** or the description of the elements that can be used in the XML document, their relationship to each other in relation to the structure of the document and other information on the attributes of each element which in this case is external and is called **unicum.dtd**.



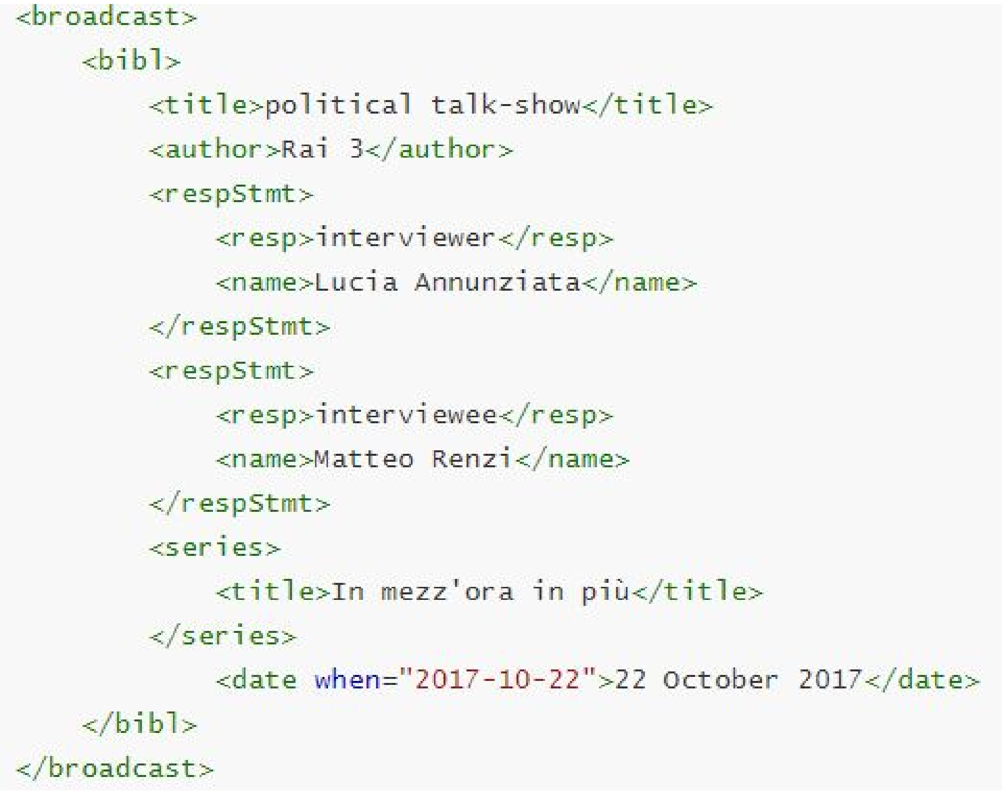
```
<teiHeader>
```

teiHeader provided detailed contextual information such as the source of the transcript, the identity of the participants, whether the speech is scripted or spontaneous, the physical and social setting in which the discourse takes place and a range of other aspects.



```
    <recording type="video" dur="01:01:46">
        <equipment>
            <p>Recorded from http://www.raiplay.it/video/2017/10/12-h-in-piu-35b932c7-
9a17-4fba-9d85-2f71d91f5806.html</p>
        </equipment>
```

Recording (recording event) provides details of an audio or video (as in the example presented) recording event used as the source of a spoken text, either directly or from a public broadcast. It therefore includes first of all: the **duration** (shortened to **dur**) of the registration taken into account and includes within it the tag **equipment** that provides technical details of the equipment and media used for an audio or video recording used as the source for a spoken text. The example shows the link through which it is possible to access the resource, so that any third party researchers can also easily access the resource.



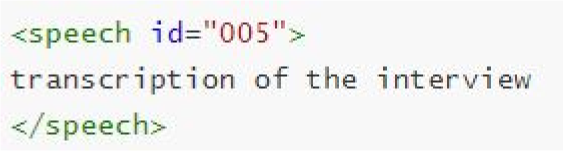
```
<broadcast>
    <bibl>
        <title>political talk-show</title>
        <author>Rai 3</author>
        <respStmt>
            <resp>interviewer</resp>
            <name>Lucia Annunziata</name>
        </respStmt>
        <respStmt>
            <resp>interviewee</resp>
            <name>Matteo Renzi</name>
        </respStmt>
        <series>
            <title>In mezz'ora in più</title>
        </series>
            <date when="2017-10-22">22 October 2017</date>
    </bibl>
</broadcast>
```

According to TEI when a recording has been made from a public broadcast, details of the broadcast itself should be supplied within the **recording** element, as a nested **broadcast** element. A broadcast is closely analogous to a publication and the **broadcast** element should therefore contain one or the other of the bibliographic citation elements **bibl**, **biblStruct**, or **biblFull**. The broadcasting agency responsible for a broadcast is regarded as its **author**, while other participants (for example **interviewers**, **interviewees**, etc.) should be specified using the **respStmt** or **editor** element with an appropriate **resp**.

In the example shown we indicate with **title** the format of television program from which the interviews are taken; with **author** the broadcasting agency responsible for a broadcast; with **resp** the main roles of the interview are indicated (**interviewer** and **interviewee),** followed by the name and surname of corresponding role; finally with **series** we indicate the television program from which the interviews are taken and with **date** is indicated the date of airing.

Step 2. Utterances and speaking turns

At the beginning of each transcription, whatever the form (monological or dialogical), a tag is inserted with the file ID. For example, the following tag introduces the **id = 005**. The numbers can follow any order and convention chosen by the annotator:

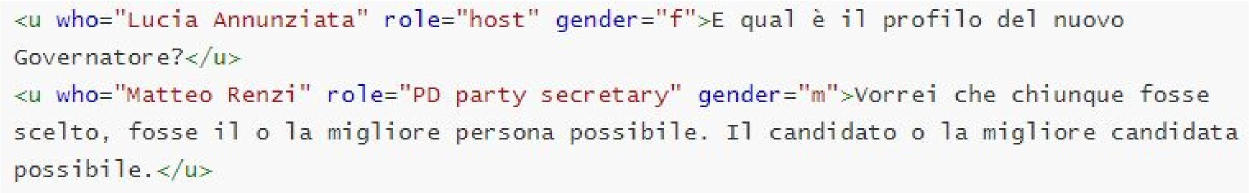


```
<speech id="005">
transcription of the interview
</speech>
```

The next step is the annotation of the utterances or turns of individual speakers. As explained by TEI:

Most researchers agree that the utterances or turns of individual speakers form an important structural component in most kinds of speech, but these are rarely as well-behaved (in the structural sense) as paragraphs or other analogous units in written texts: speakers frequently interrupt each other, use gestures as well as words, leave remarks unfinished and so on. Speech itself, though it may be represented as words, frequently contains items such as vocalized pauses which, although only semi-lexical, have immense importance in the analysis of spoken text. Even non-vocal elements such as gestures may be regarded as forming a component of spoken text for some analytic purposes. Below the level of the individual utterance, speech may be segmented into units defined by phonological, prosodic, or syntactic phenomena; no clear agreement exists, however, even as to appropriate names for such segments.

Each transcription consists of alternating turns between the sender and receiver in the case of dialogic form, and utterance by utterance in the case of monological form. Therefore, annotators are asked to segment the document marking this kind of components. See for example the following excerpt:



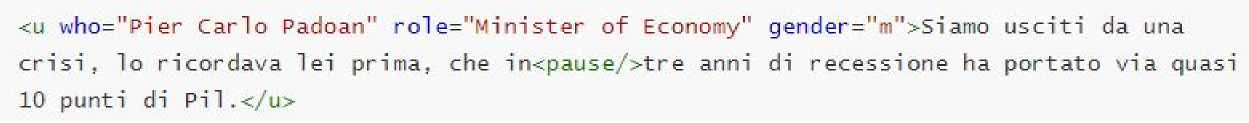
```
<u who="Lucia Annunziata" role="host" gender="f">E qual è il profilo del nuovo
Governatore?</u>
<u who="Matteo Renzi" role="PD party secretary" gender="m">Vorrei che chiunque fosse
scelto, fosse il o la migliore persona possibile. Il candidato o la migliore candidata
possibile.</u>
```

u (utterance) contains a stretch of speech usually preceded and followed by silence or by a change of speaker. **who** indicates the person (name and surname) who holds the turn. In order to analyse additional socio-linguistic dimensions, we also introduce in the annotation scheme of PoliModal the following attributes: **role** indicates the work of the person who holds the turn; **gender** indicates male or female gender.

Step 3. Pausing

Speakers differ very much in their rhythm and in particular in the amount of time left between words or utterances. The **pause** tag is therefore used to mark when the speech has been paused, irrespective of the actual amount of silence. The tag can be put either between or within utterances.

For example:



```
<u who="Pier Carlo Padoan" role="Minister of Economy" gender="m">Siamo usciti da una
crisi, lo ricordava lei prima, che in<pause/>tre anni di recessione ha portato via quasi
10 punti di Pil.</u>
```

A pause marked within an utterance applies to the speaker of that utterance. The attribute type may be used to categorize the pause, for example as short, medium, or long; alternatively the attribute **dur** (i.e. duration) may be used to indicate its length more exactly.

Step 4. Vocal

A typical aspect of spoken language is the use of semi-lexical and non-lexical expressions. Lexical expressions consist mainly of interjections (lexical category that conveys the meaning of an entire sentence, so it alone constitutes a complete linguistic act demonstrated by the fact that it is paraphrasable). Instead non-lexical expressions costint of phenomena such as coughing, exhaling, sniffing, snorting, huffing etc.

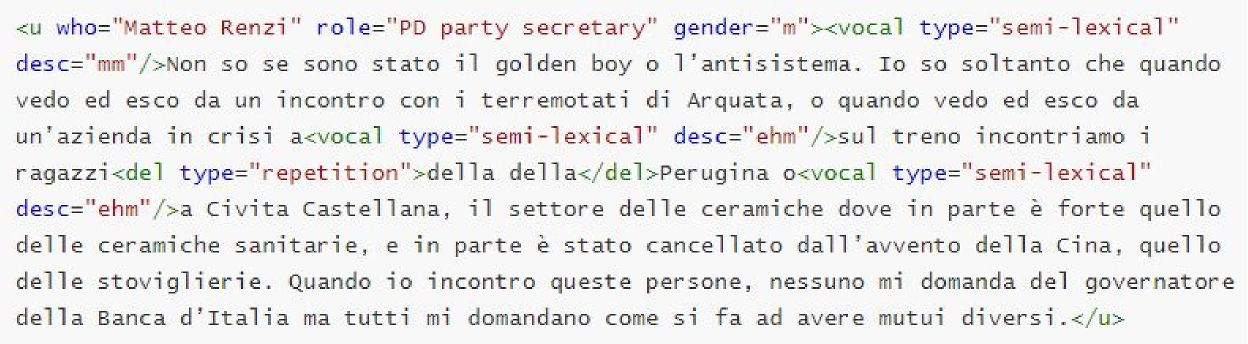
The presence of non-transcribed semi-lexical or non-lexical phenomena either between or within utterances is foreseen also by theTEI standard. It can be marked using the following tags:

- **vocal** marks any vocalized but not necessarily lexical phenomenon, for example voiced pauses, non-lexical backchannels, etc.

- **kinesic** marks any communicative phenomenon, not necessarily vocalized, for example a gesture, frown, etc.
- **incident** marks any phenomenon or occurrence, not necessarily vocalized or communicative, for example incidental noises or other events affecting communication.

Note that the phenomena foreseen by the **kinesic** tag have been the object of a separate annotation that will be discussed separately in Section 7. The phenomena described by the **incident** tag have never been detected in the PoliModal corpus because the alternation of shifts foreseen by the format under examination makes its presence very unlikely.

Instead, the phenomena marked with the **vocal** tag have been detected in the corpus and correspond - in most cases - to the phenomena usually defined as "interjections". For example:



```
<u who="Matteo Renzi" role="PD party secretary" gender="m"><vocal type="semi-lexical"
desc="mm"/>Non so se sono stato il golden boy o l'antisistema. Io so soltanto che quando
vedo ed esco da un incontro con i terremotati di Arquata, o quando vedo ed esco da
un'azienda in crisi a<vocal type="semi-lexical" desc="ehm"/>sul treno incontriamo i
ragazzi<del type="repetition">della della</del>Perugina o<vocal type="semi-lexical"
desc="ehm"/>a Civita Castellana, il settore delle ceramiche dove in parte è forte quello
delle ceramiche sanitarie, e in parte è stato cancellato dall'avvento della Cina, quello
delle stoviglierie. Quando io incontro queste persone, nessuno mi domanda del governatore
della Banca d'Italia ma tutti mi domandano come si fa ad avere mutui diversi.</u>
```

As can be observed from this excerpt, the **vocal** tag enables two types of attributes: **type**, which admits **semi-lexical** or **non-lexical** values, and **desc** (i.e. description), which may be used to supply a conventional representation for the phenomenon, for example non-lexical e.g. burp, click, cough, exhale, giggle, gulp, inhale, laugh, sneeze, sniff, snort, sob, swallow, throat, yawn | semi-lexical e.g. ah, aha, aw, eh, ehm, er, erm, hmm, huh, mm, mmhm, oh, ooh, oops, phew, tsk, uh, uh-huh, uh-uh, um, urgh, yup.

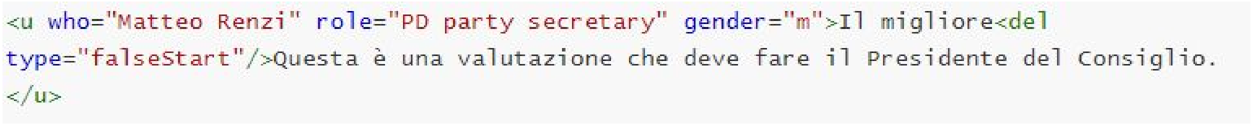
Step 5. False starts, repetition and truncated words

Phenomena of **speech management** include disfluencies such as filled and unfilled pauses, interrupted or repeated words, corrections, and reformulations as well as interactional devices asking for or providing feedback. These phenomena are marked as editorially deleted i.e. **del** in the annotation.

5.1. False start

Although spoken texts are the product of a physically continuous process, their structure shows a strong discontinuity: false starts, interruptions, project changes are common to all spontaneous speech texts.

The false starts are therefore noted as follows:



```
<u who="Matteo Renzi" role="PD party secretary" gender="m">Il migliore<del
type="falseStart"/>Questa è una valutazione che deve fare il Presidente del Consiglio.
</u>
```

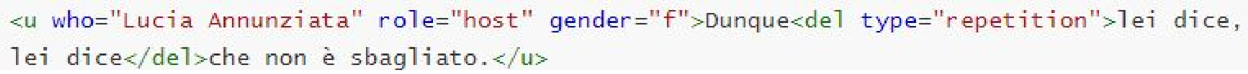
5.2. Repetition

As noted by Voghera (2001):

Si è da più parti notato (Simone 1990, Bazzanella 1992; Tannen 1989; Voghera 1992b) che nel parlato spontaneo vi è un'alta percentuale di ripetizioni. [...] Esistono infatti vari tipi di ripetizione con funzioni diverse, che possiamo ricondurre a due macrocategorie (Voghera 1992b): ripetizione di enunciati altrui per dare coerenza e coesione al discorso; autoripetizione di tipo automatico come meccanismo di controllo della programmazione del discorso. Tanto il primo quanto il secondo tipo di ripetizione sono funzionali al controllo della progettazione testuale in fieri del parlato.

(en. It has been noted from many partes (Simone 1990, Bazzanella 1992; Tannen 1989; Voghera 1992b) that in spontaneous speech there is a high percentage of repetition. [...] In fact, there are various types of repetition with different functions, which can be traced back to two macro-categories (Voghera 1992b): repetition of other people's statements in order to give coherence and cohesion to speech; repetition of him/herself as a mechanism for controlling the programming of speech. Both the first and the second type of repetition are functional to the control of the verbal design in progress of speech.)

In the annotation scheme presented in these guidelines, the type of pause does not affect the tag used, which will be unique, as follows:

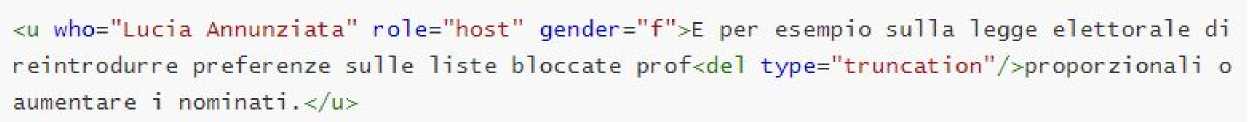


```
<u who="Lucia Annunziata" role="host" gender="f">Dunque<del type="repetition">lei dice,
lei dice</del>che non è sbagliato.</u>
```

5.3. Truncation

Spontaneous dialogical texts present a frequent and somewhat 'compulsory' use of deictic elements (Givon 1995). Some forms of ellipses can also be traced back to deictic or indessical phenomena (Berretta 1994). The same need for indessicality can also be traced back to cases of reduction and truncation.

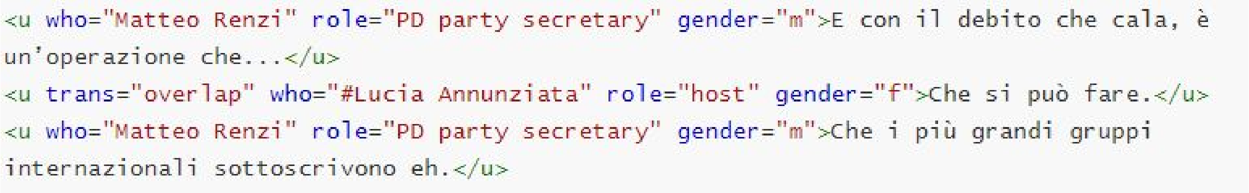
Truncations - which also fall into the category of editorially deleted - are annotated as follows:



```
<u who="Lucia Annunziata" role="host" gender="f">E per esempio sulla legge elettorale di
reintrodurre preferenze sulle liste bloccate prof<del type="truncation"/>proporzionali o
aumentare i nominati.</u>
```

Step 6. Overlap

This element is obviously only present in the case of transcriptions in dialogical form. As Voghera (2001) points out again, the conditions of construction and reception of texts mean that speech needs a lot of redundancy because it is more exposed to noise than writing. A source of noise is the alternation of turns, which can lead to overlapping of the participants in the communication and, therefore, partial or total loss of information. So this phenomenon is present when the speaker conveys (in a verbal or non-verbal manner) that he/she is about to finish his/her turn and the co-locutor starts speaking so that there is a slight overlap of utterances. Overlays will be annotated as follows:



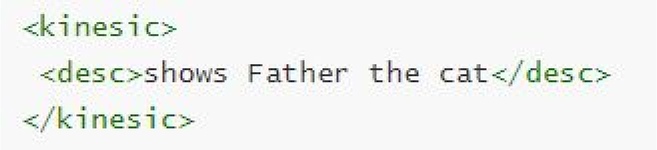
```
<u who="Matteo Renzi" role="PD party secretary" gender="m">E con il debito che cala, è
un'operazione che...</u>
<u trans="overlap" who="#Lucia Annunziata" role="host" gender="f">Che si può fare.</u>
<u who="Matteo Renzi" role="PD party secretary" gender="m">Che i più grandi gruppi
internazionali sottoscrivono eh.</u>
```

In the example above, the journalist Lucia Annunziata breaks Matteo Renzi's turn to speak, completing the sentence through an overlap. The simplest way of representing this **overlap** is to use the **trans** attribute that is provided as a means of characterizing the transition from one utterance to the next. The tag is completed by entering the name and surname of the person **who** overlaps preceded by #, then adding **role** and **gender** below.

7. Kinesic tag details

The TEI standard provides the possibility to mark any communicative phenomenon, not necessarily vocalized, for example a gesture, frown, etc. using the tag **kinesic**, which should be followed by a **desc** tag asking the annotator to describe the observed proxemic phenomenon.

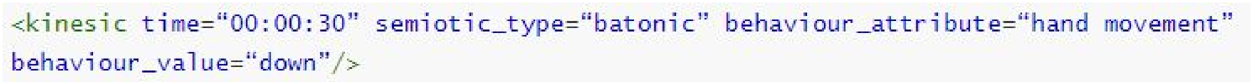
The tag would look like this:



```
<kinesic>
 <desc>shows Father the cat</desc>
</kinesic>
```

In our opinion, this would leave out a set of useful information for further linguistic analysis. This is why - in case of manual annotation - we propose to overcome this limitation by introducing a specific tagset inspired by the well-known MUMIN multimodal annotation standard (Allwood et al. 2007). We suggest that the tagset for **kinesic** should consist of four elements: a) *time*: with which the specific moment when the observed phenomenon is needed is marked; b) *semiotic_type*: semiotic categories relevant to all types of gestures defined on the basis of Pierce's semiotic types (Pierce, 1931) therefore *deictic gesture* (indicate an object or a person with the index finger or open hand), *iconic* (depicting the shape in the air or imitating the typical movements of an object, an animal, a person), *symbolic* (gesture that in a given culture has a meaning easily translated into words or sentences, for example index and V-shaped middle finger in front of the mouth which means "smoking" or "cigarette"), *batonic* (the hands go from top to bottom to scan and emphasize speech); c) *behaviour_attribute*: facial expressions, eyebrow and eye movements, lip movements, head and hand movements; d) *behaviour_value*: values that can be assigned to each of the previous movements observed according to MUMIN.

So the tag would look like this:



```
<kinesic time=“00:00:30” semiotic_type=“batonic” behaviour_attribute=“hand movement”
behaviour_value=“down”/>
```

REFERENCES

Allwood, Jens, et al. "The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena." *Language Resources and Evaluation* 41.3-4 (2007): 273-287.

Aprosio, Alessio Palmero, and Giovanni Moretti. "Italy goes to Stanford: a collection of CoreNLP modules for Italian." *arXiv preprint arXiv:1609.06204* (2016).

Berretta, Monica. "Il parlato italiano contemporaneo." *Storia della lingua italiana* 2 (1994): 239-270.

Bazzanella, Carla. "Aspetti pragmatici della ripetizione dialogica." (1992): 433-454.

Biber, Douglas. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press, 1995.

Givón, Talmy, and Morton Ann Gernsbacher, eds. *Coherence in spontaneous text*. J. Benjamins, 1995.

Kipp, Michael. "Anvil-a generic annotation tool for multimodal dialogue." *Seventh European Conference on Speech Communication and Technology*. 2001.

Miller, James Edward, Jim Miller, and Regina Weinert. *Spontaneous spoken language: Syntax and discourse*. Oxford University Press on Demand, 1998..

Hartshorne, Charles. *Collected Papers of Charles Sanders Peirce*. Eds. Paul Weiss, and Arthur W. Burks. Vol. 1. Cambridge: Harvard University Press, 1931.

Simone, Raffaele. "Effetto copia e effetto quasi-copia." *AIΩN. Annali del Dipartimento di Studi del Mondo Classico e del Mediterraneo Antico* 12 (1990): 69-83.

Tannen, Deborah. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Vol. 26. Cambridge University Press, 2007.

Trotta, Daniela, Palmero Aprosio, Alessio, Tonelli, Sara, and Elia, Annibale. Adding Gesture, Posture and Facial Displays to the PoliModal Corpus of Political Interviews. In Proceedings of LREC 2020, Marseille, France.

Trotta, Daniela, Tonelli Sara, Palmero Aprosio, Alessio,and Elia, Annibale. Annotation and Analysis of the PoliModal Corpus of Political Interviews. In Proceedings of CliC-it 2019, Bari, Italy.

Voghera, Miriam. "Repetita iuvant." *Italiano e oltre* 3 (1992): 121-125.

Voghera, Miriam. *Teorie linguistiche e dati di parlato*. na, 2001.