USING ELAN TO ANNOTATE ORAL DISCOURSE

Elena Pascual Aliaga
University of Valencia / Val.Es.Co. Research Group
elena.pascual@uv.es

i. ELAN interface
ii. Transcription sample
1. Transcription procedure
2. Annotation of DRD

---

i. ELAN interface

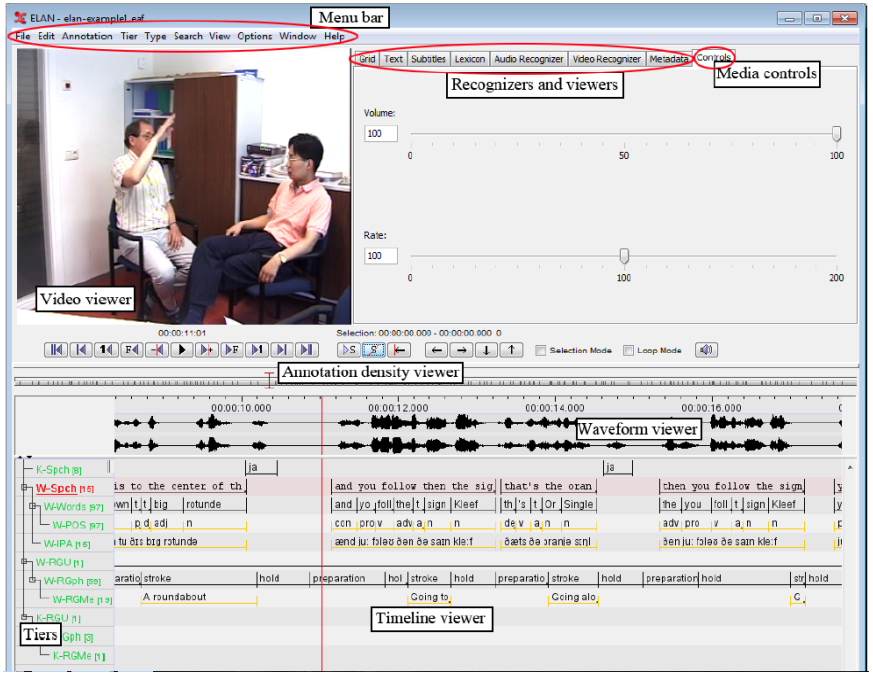

Figure 1. ELAN interface (Taken from Hellwing [2017])

ii. Transcription sample

P: hey Raj! /// (5'') hey listen // I don't know if you heard about what happened last night with Leonard and Sheldon but I am REALLY upset about it / I mean they just- they let themselves into my place↑ / and then they cleaned it! Can you even believe that?! How weird [is that?! (( )) =]

R: **[**oh! // she's standing very close to me /// (5'') oh my she does smell good // what is that? / vanilla?]

P: [= strangers but I mean→] you know? // where I come from / someone comes into your house at night↑ you shoot / OKAY? and you don't shoot to wound // I mean / right↓ my sister shot her husband but / it was an accident they were drunk // what was I saying? [(( )) =]

Extracted from *The Big Bang Theory*, season 1, episode 2

YouTube video clip (0:19 to 0:58):
https://www.youtube.com/watch?v=thpFZp5LS_E

1. Transcription procedure

1) Open, create or import files. Select an option clicking on "*File*" at the menu bar:

• "New" if you want to open a media file in ELAN (.mpg, .wav), but this is not for opening an annotation file (.eaf). Then select the media files (audio and/or video).

• "Open" if you want to open an ELAN file (.eaf).

• "Import" if you want to import an annotation file (Toolbox, FLEx, CHAT, Transcrier, .txt, .TextGrid, Recognizer, Shoebox…).

2) Select a *Working mode* in the "*Options*" button from the menu bar:

• "Synchronization mode": to synchronize media files (video, audio, time series).

• "Transcription mode": to type text in annotations (no time segmentation is allowed).

• "Segmentation mode": to create rapidly empty annotations (no typing is allowed).

• "Interlinearization mode": to tokenize, parse and gloss annotations.

• "Annotation mode": generic mode that contains several functions from the previous models.

→ We will use the *Annotation Mode*.

3) Add new tiers by clicking the option "*Add New Tire*" in the "*Tier*" button from the menu. Set the options and type.

(see different types of tiers in the next section)

→ We will use the type "*None*" for the transcription.

4) Select a segment on the timeline to enter the annotation:

1º Select the tier that you want to annotate (double click on the tier name).

2º Select a segment by dragging the mouse or using the "Selection mode" from the Media controls.

3º In order to type, double click on the selected segment or right click and select the "New Annotation here" option.

2. Annotation procedure

1) Add new tiers by clicking the option "*Add New Tire*" in the "*Tier*" button from the menu. Set the options and type:

<u>Independent tiers:</u>

• "None": an independent tier without a parent.

Dependent tiers:

• "Time Subdivision": can be subdivided into smaller segments – no gaps allowed between them – which can be linked to time intervals. E.g., tokens of words.

• "Symbolic Subdivision": can be subdivided into smaller segments – no gaps allowed between them – which cannot be linked to time intervals. E.g., morphemes.

• "Included-in": can be subdivided into smaller segments –gaps are allowed between them – which can be linked to time intervals. E.g., silences between words.

• "Symbolic Association": cannot be subdivided into smaller segments, it has an exact one-to-one correspondence with the parent annotation. E.g., a translation.

2) Create a Controlled Vocabulary (CV)

A *Controlled Vocabulary* is a list of annotation values (or codes) that can be used on one or more tiers.

1º Set up a *CV* by clicking on "*Edit > Edit Controlled Vocabulary*". Enter the values (labels) and their description.

2º Set up a linguistic *Type* that uses the *CV* by clicking on "*Type > Add New Tier Type*". Enter the name of the *Type* and select the *CV* in the "*Use Controlled Vocabulary*" option.

3º Add a new tier (or modify an existing one) and select the created type in the "*Tier Type"* option.

3) Tokenize a tier (segmentation into words)

1º Click the option "*Tokenize tier*" on the "*Tier*" button from the menu and select the source tier – the one which will be tokenized – and the destination tier – the one where the tokenization will be displayed –. Select the option "Create New Tier" if you want to add a new tier for the tokenization output.

2º Choose the tokenization options (e.g. select the elements delimiting tokens – spaces, punctuation, etc. –).

4) Create annotation templates

Once you have a set of tiers, linguistic types, and controlled vocabularies, you can save them in a template so that you can use them later for creating new annotation files.

• Select "File > Save as Template". A template file (.etf) will be created.

• In order to create a new annotation document (.eaf) based on a template,

1º Follow the steps to create a new document.

2º In the "File" browser click on "Add Template file" and select the .etf file.

Dialogue segmentation into acts and subacts[1] (Val.Es.Co. model)

P: # IAS {hey}IAS DSS{Raj!}DSS # /// (5'') # IAS {hey} IAS / IAS{listen}IAS // DSS{I don't know if you heard about what happened last night with Leonard and Sheldon}DSS SSS{but I am REALLY upset about it}SSS / # # TAS{I mean}TAS DSS{they just- they let themselves into my place↑}DSS SSS{and then they cleaned it!}SSS # # DSS{Can you SAM{even}SAM believe that?!}DSS # # DSS{How weird [is that?!}DSS (( )) =] #

R: # MAS{[oh!}MAS // DSS{she's standing very close to me}DSS # /// (5'') # MAS{oh my}MAS DSS{she does smell good}DSS # // # DSS{what is that? / vainilla?]}DSS #

P: # TAS{[= I mean→]}TAS IAS{you know?}IAS // DSS{where I come from / someone comes into your house at night↑ you shoot}DSS / IAS/M{OKEY?}IAS/M SSS{and you don't shoot to wound}SSS // # TAS{I mean}TAS / MAS/T{right↓}MAS/T DSS{my sister shot her husband}DSS SSS{but / it was an accident they were drunk}SSS # // # DSS{what was I saying? }DSS [(( )) =] #

References:

Briz, A. and Val.Es.Co. Group (2014): "Las unidades del discurso oral. La propuesta Val.Es.Co. de segmentación de la conversación (coloquial)". *Estudios de Lingüística del Español*, 35 (1), p. 11-7.

Briz, A. and Val.Es.Co. Group (2003): "Un sistema de unidades para el estudio del lenguaje coloquial". Oralia, 6, p. 7-61.

Hellwing, Birgit (2017): "ELAN - Linguistic Annotator version 5.0.0-alpha" (manual). https://tla.mpi.nl/tools/tla-tools/elan/.

Pons, S. (2016): "Cómo dividir una conversación en actos y subactos", in A. M. Bañón et al. (eds.) *Oralidd y Análisis del discurso: Homenaje a Luis Rodríguez Cortés*, Almería, Editorial Universidad de Almería, pp.545-566.

Pons y Estellés (2014): "Absolute Initial Poition", in S. Pons (ed.) *Discourse segmentation in Romance languages*, Amsterdam/Philadelphia, John Benjamins.

[1] The symbol "#" is used to mark act boudaries; subact boundaries are marked using "{}".