Research Methods

Transcriptions & Annotations (ELAN Tool)

Amandine Decker (*amandine.decker@gu.se*)

UNIVERSITY OF
GOTHENBURG

Table of contents

1. Transcriptions and Annotations
    - Examples of Transcription / Annotation Uses
    - Definition(s)
    - What is an Annotation Campaign?
2. One Transcription / Annotation Tool: ELAN
    - ELAN?
    - Examples
    - Core Concept
    - Application and Assignment

Transcriptions and Annotations

Application for daily life

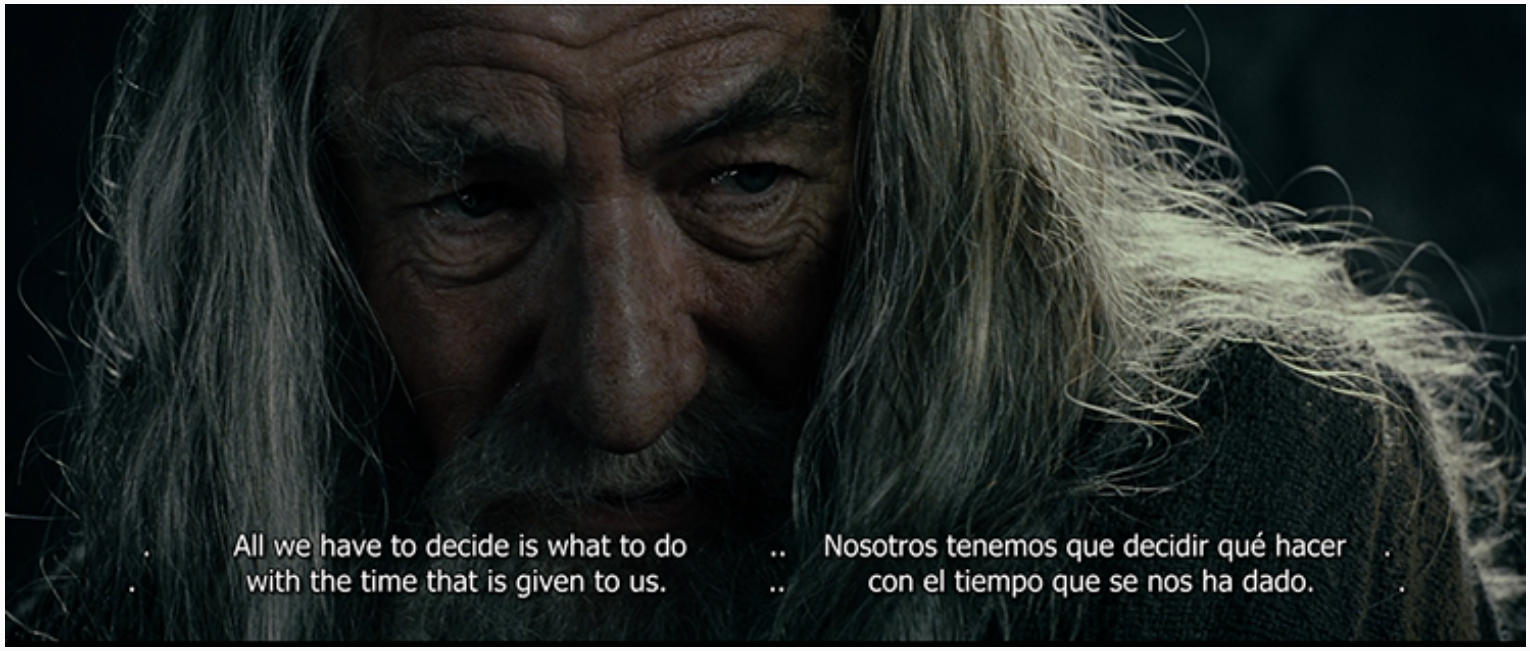
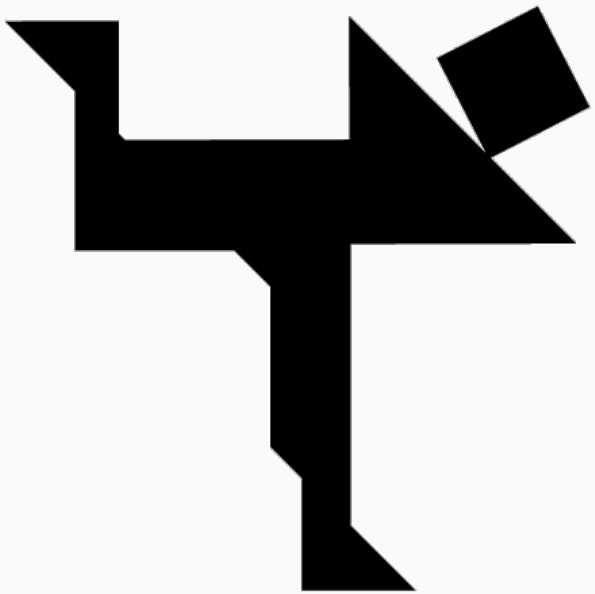
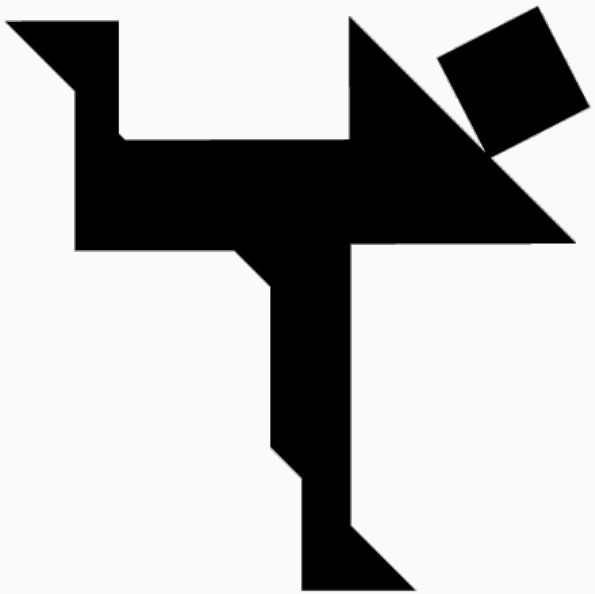


Image from *https://bonigarcia.dev/dualsub/*

Application for Linguistic Research

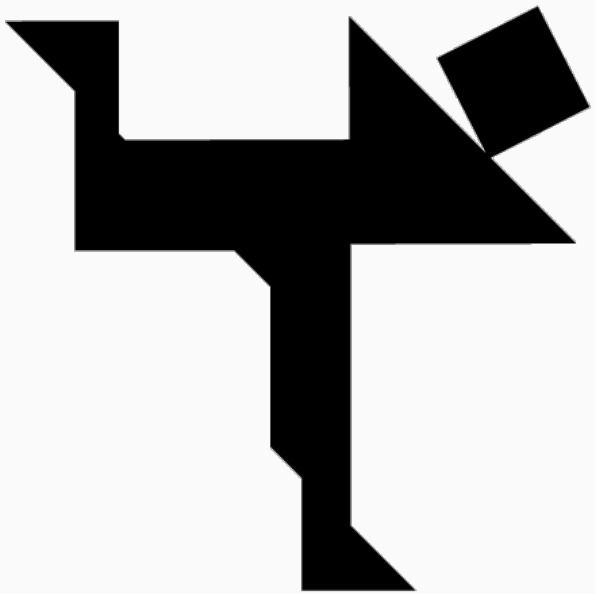


Application for Linguistic Research



“The guy with one foot on the ground and one leg up.”

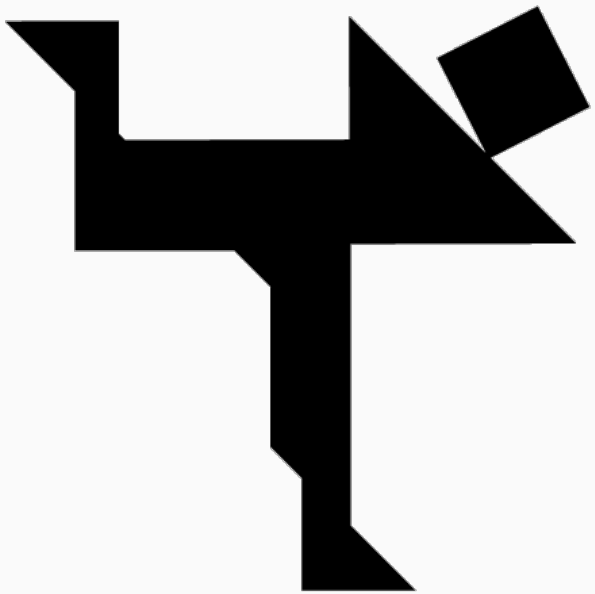
Application for Linguistic Research



“The guy with one foot on the ground and one leg up.”

“The ice skater guy.”

Application for Linguistic Research

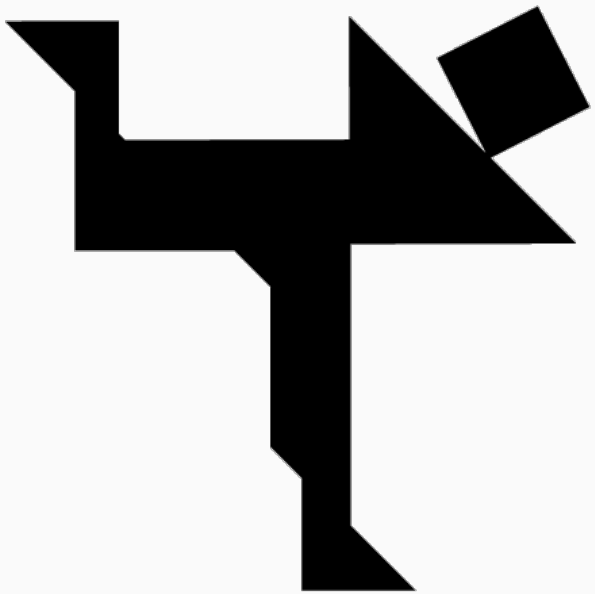


"The guy with one foot on the ground and one leg up."

"The ice skater guy."

"Someone trying to throw a ball."

Application for Linguistic Research



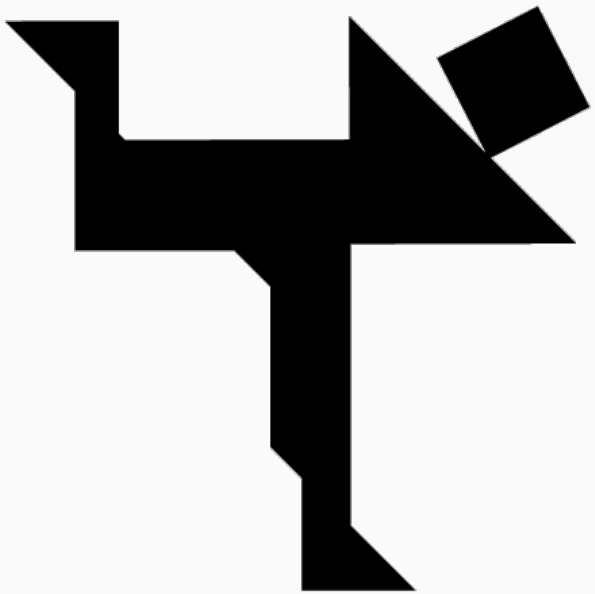
"The guy with one foot on the ground and one leg up."

"The ice skater guy."

"Someone trying to throw a ball."

"Someone dancing or in balance on one foot."

Application for Linguistic Research



"The guy with one foot on the ground and one leg up."

"The ice skater guy."

"Someone trying to throw a ball."

"Someone dancing or in balance on one foot."

"A yoga posture."

Application for Language Technologies

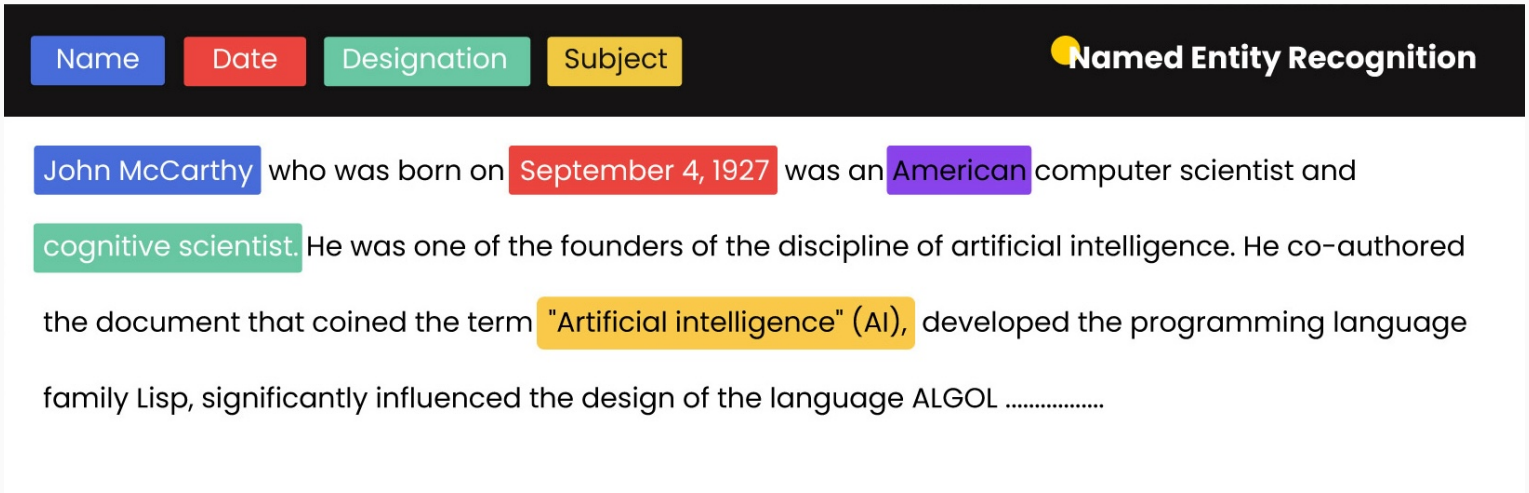


Image from *https://www.amygb.ai/blog/what-is-named-entity-recognition-in-nlp*

Annotation?

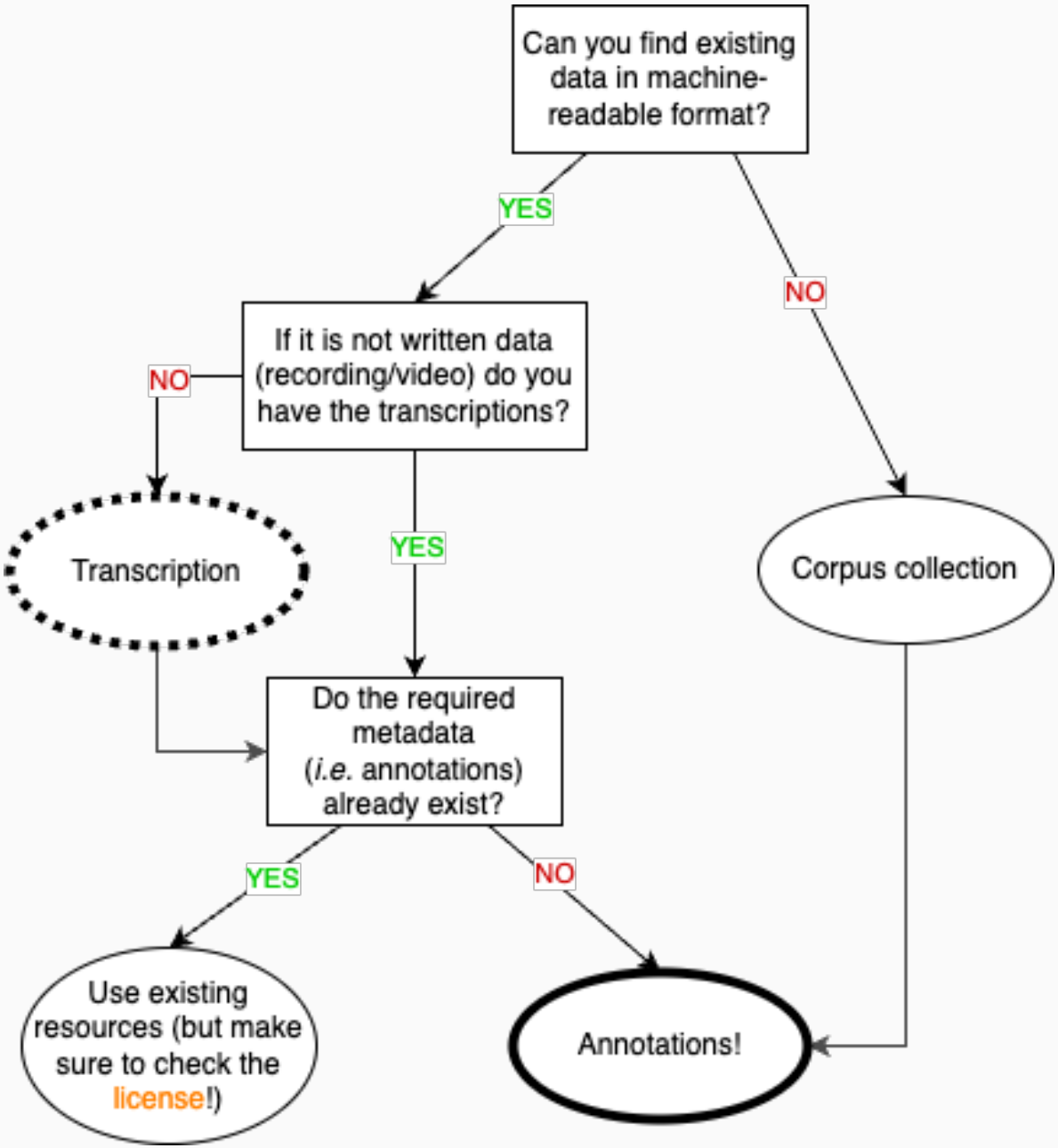
[Leech, 1997]

"[corpus annotation] can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process"

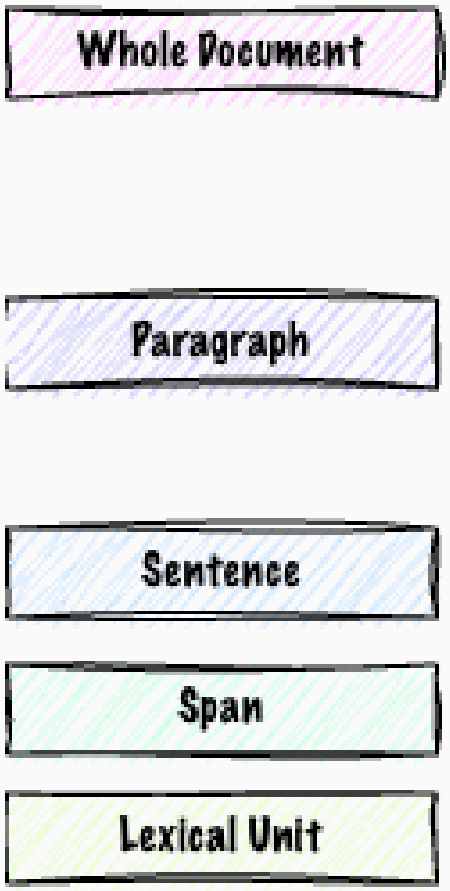
[Bird and Liberman, 2001]

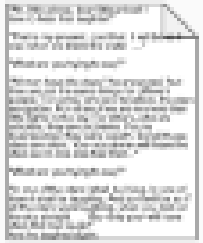
"'Linguistic annotation' covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions - audio, video and/or physiological recordings - or it may be textual."

Define the Goal



Choose the Annotation Unit(s)





"All men have the stars," he answered, "but they are not the same things for different people. For some, who are travellers, the stars are guides. For others they are no more than little lights in the sky. For others, who are scholars, they are problems. For my businessman they were wealth. But all these stars are silent. You--you alone--will have the stars as no one else has them--"

"'All men have the stars,' he answered, 'but they are not the same things for different people."

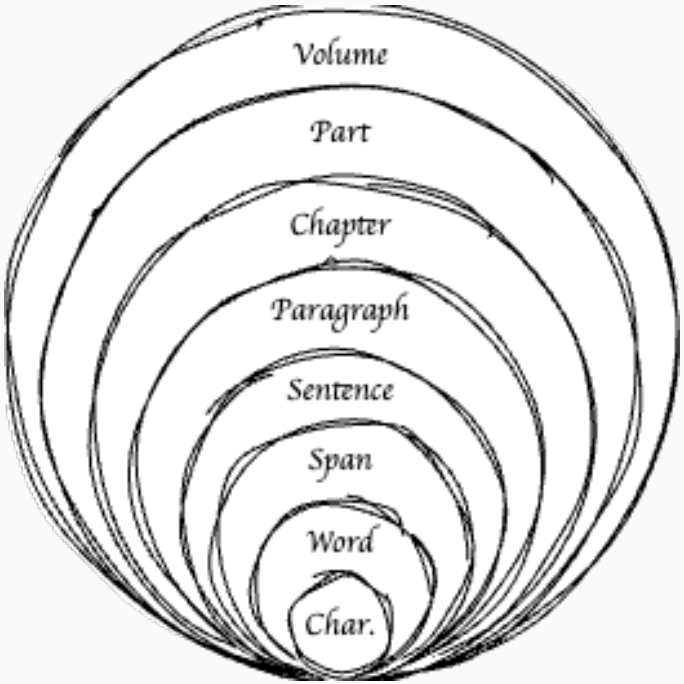
"'All men have the stars,' he answered, 'but they are not the same things for different people."

"'All men have the stars,' he answered, 'but they are not the same things for different people."

Inspired from the slides of Valerio Basile and Enrica Troiano on Data Perspectivism at ESSLLI 2023.

Choose the Annotation Unit(s)

E.g. Recognise if parts of narrative texts describe a fact or a character's beliefs:



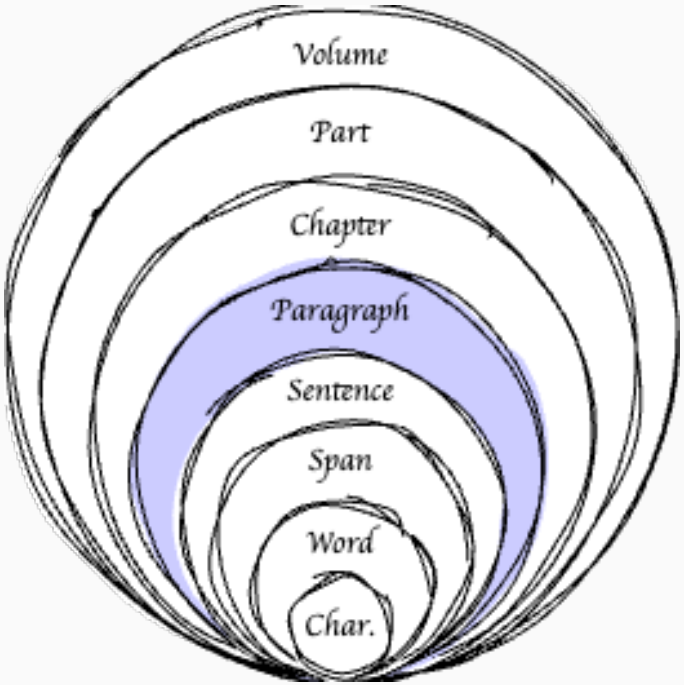
Alice's Adventures in Wonderland, Chapter II

'I'm sure I'm not Ada,' she said, 'for her hair goes in such long ringlets, and mine doesn't go in ringlets at all; and I'm sure I can't be Mabel, for I know all sorts of things, and she, oh! she knows such a very little! Besides, she's she, and I'm I, and–oh dear, how puzzling it all is!

Inspired from the slides of Valerio Basile and Enrica Troiano on Data Perspectivism at ESSLLI 2023.

Choose the Annotation Unit(s)

E.g. Recognise if parts of narrative texts describe a fact or a character's beliefs:



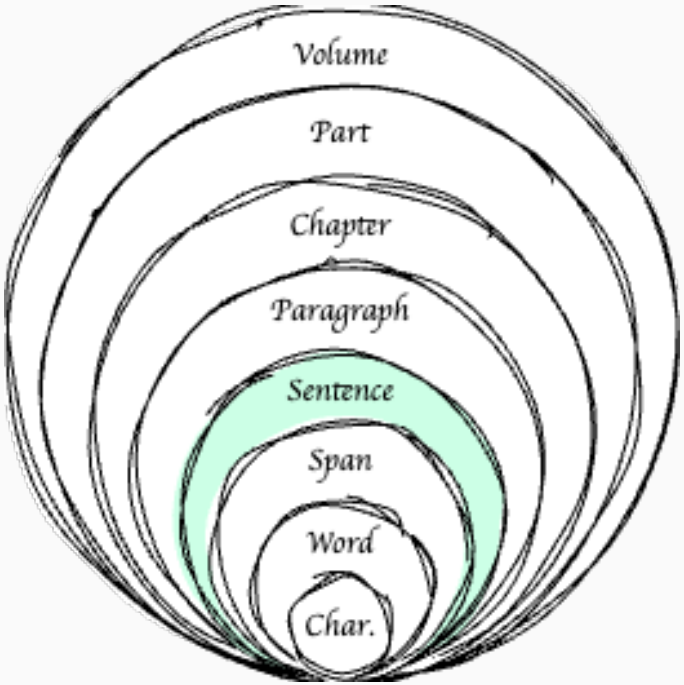
Alice's Adventures in Wonderland, Chapter II

'I'm sure I'm not Ada,' she said, 'for her hair goes in such long ringlets, and mine doesn't go in ringlets at all; and I'm sure I can't be Mabel, for I know all sorts of things, and she, oh! she knows such a very little! Besides, she's she, and I'm I, and–oh dear, how puzzling it all is!

Inspired from the slides of Valerio Basile and Enrica Troiano on Data Perspectivism at ESSLLI 2023.

Choose the Annotation Unit(s)

E.g. Recognise if parts of narrative texts describe a fact or a character's beliefs:



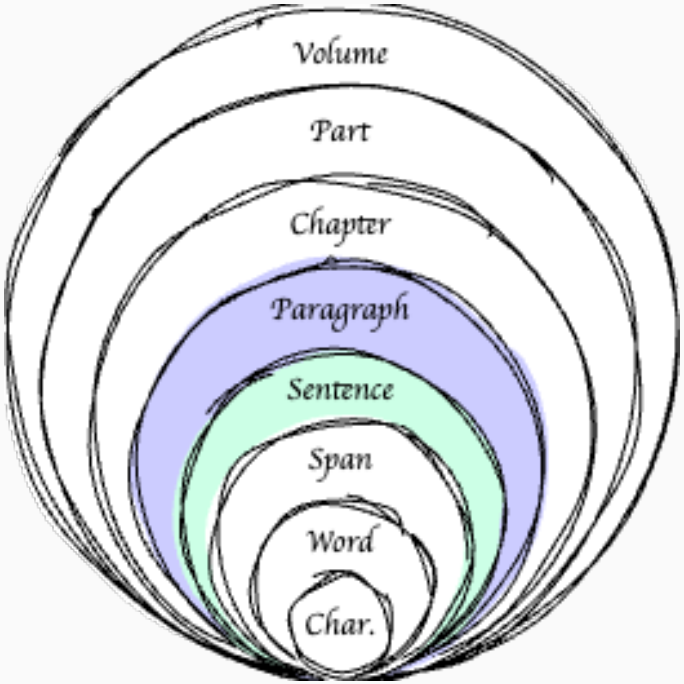
Alice's Adventures in Wonderland, Chapter II

'I'm sure I'm not Ada,' she said, 'for her hair goes in such long ringlets, and mine doesn't go in ringlets at all; and I'm sure I can't be Mabel, for I know all sorts of things, and she, oh! she knows such a very little! Besides, she's she, and I'm I, and–oh dear, how puzzling it all is!

Inspired from the slides of Valerio Basile and Enrica Troiano on Data Perspectivism at ESSLLI 2023.

Choose the Annotation Unit(s)

E.g. Recognise if parts of narrative texts describe a fact or a character's beliefs:



Alice's Adventures in Wonderland, Chapter II

'I'm sure I'm not Ada,' she said, 'for her hair goes in such long ringlets, and mine doesn't go in ringlets at all; and I'm sure I can't be Mabel, for I know all sorts of things, and she, oh! she knows such a very little! Besides, she's she, and I'm I, and–oh dear, how puzzling it all is!

Inspired from the slides of Valerio Basile and Enrica Troiano on Data Perspectivism at ESSLLI 2023.

Write the Guidelines

How to ask for the target information?

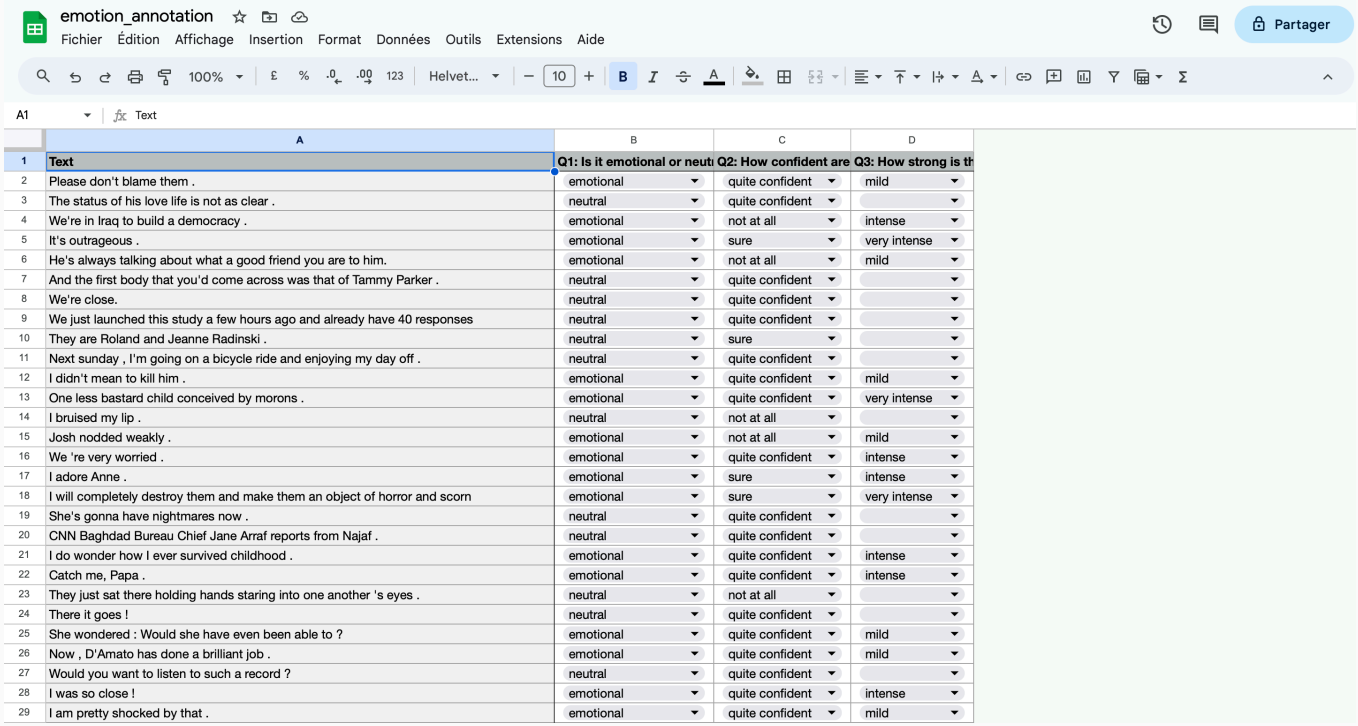
→ The guidelines are instructions to apply your coding scheme to the data. They include:

- Goal;
- How to use the tagset:
  - Where to appy the tags (*i.e.*, the annotation units);
  - Examples of good/bad uses;
  - Grey areas and how to deal with ambiguities.

Be informative and keep things clear and simple!

Inspired from the slides of Valerio Basile and Enrica Troiano on Data Perspectivism at ESSLLI 2023.

Choose the Right Tool


emotion_annotation

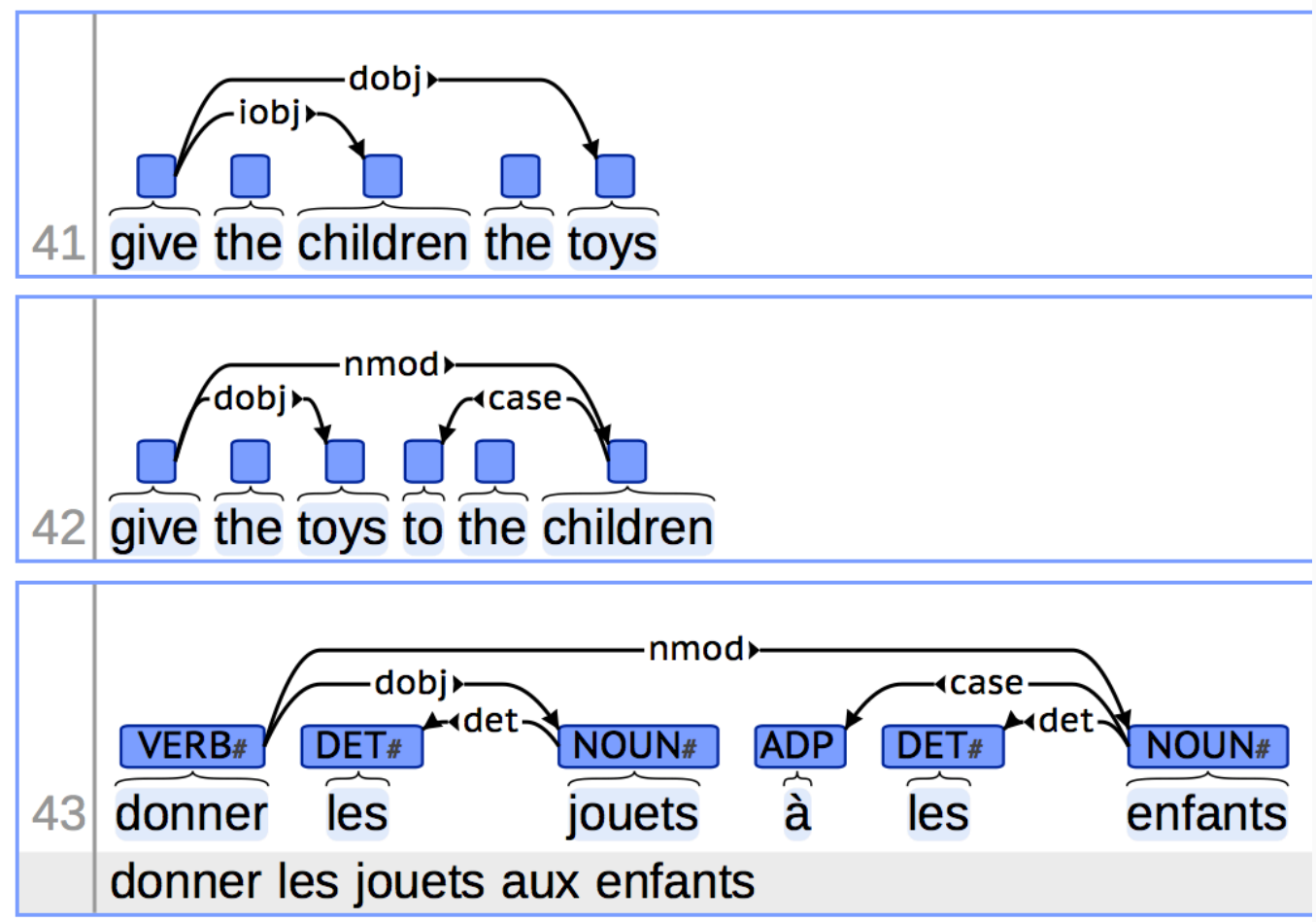
Fichier Édition Affichage Insertion Format Données Outils Extensions Aide

Partager

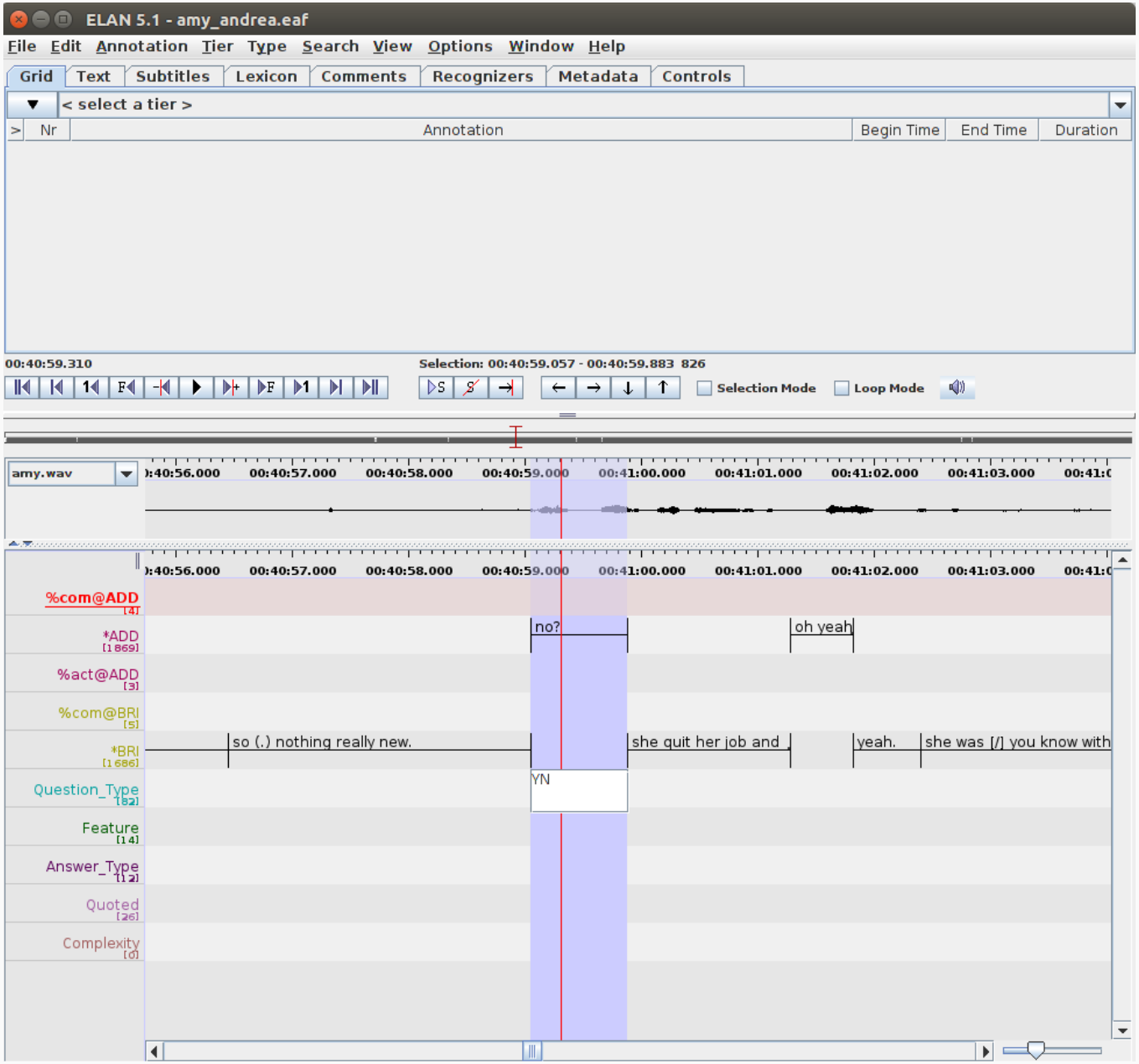
A1 | Text

	A	B	C	D
1	**Text**	**Q1: Is it emotional or neut**	**Q2: How confident are**	**Q3: How strong is th**
2	Please don't blame them .	emotional	quite confident	mild
3	The status of his love life is not as clear .	neutral	quite confident	
4	We're in Iraq to build a democracy .	emotional	not at all	intense
5	It's outrageous .	emotional	sure	very intense
6	He's always talking about what a good friend you are to him.	emotional	not at all	mild
7	And the first body that you'd come across was that of Tammy Parker .	neutral	quite confident	
8	We're close.	neutral	quite confident	
9	We just launched this study a few hours ago and already have 40 responses	neutral	quite confident	
10	They are Roland and Jeanne Radinski .	neutral	sure	
11	Next sunday , I'm going on a bicycle ride and enjoying my day off .	neutral	quite confident	
12	I didn't mean to kill him .	emotional	quite confident	mild
13	One less bastard child conceived by morons .	emotional	quite confident	very intense
14	I bruised my lip .	neutral	not at all	
15	Josh nodded weakly .	emotional	not at all	mild
16	We 're very worried .	emotional	quite confident	intense
17	I adore Anne .	emotional	sure	intense
18	I will completely destroy them and make them an object of horror and scorn	emotional	sure	very intense
19	She's gonna have nightmares now .	neutral	quite confident	
20	CNN Baghdad Bureau Chief Jane Arraf reports from Najaf .	neutral	quite confident	
21	I do wonder how I ever survived childhood .	emotional	quite confident	intense
22	Catch me, Papa .	emotional	quite confident	intense
23	They just sat there holding hands staring into one another 's eyes .	neutral	not at all	
24	There it goes !	neutral	quite confident	
25	She wondered : Would she have even been able to ?	emotional	quite confident	mild
26	Now , D'Amato has done a brilliant job .	emotional	quite confident	mild
27	Would you want to listen to such a record ?	neutral	quite confident	
28	I was so close !	emotional	quite confident	intense
29	I am pretty shocked by that .	emotional	quite confident	mild

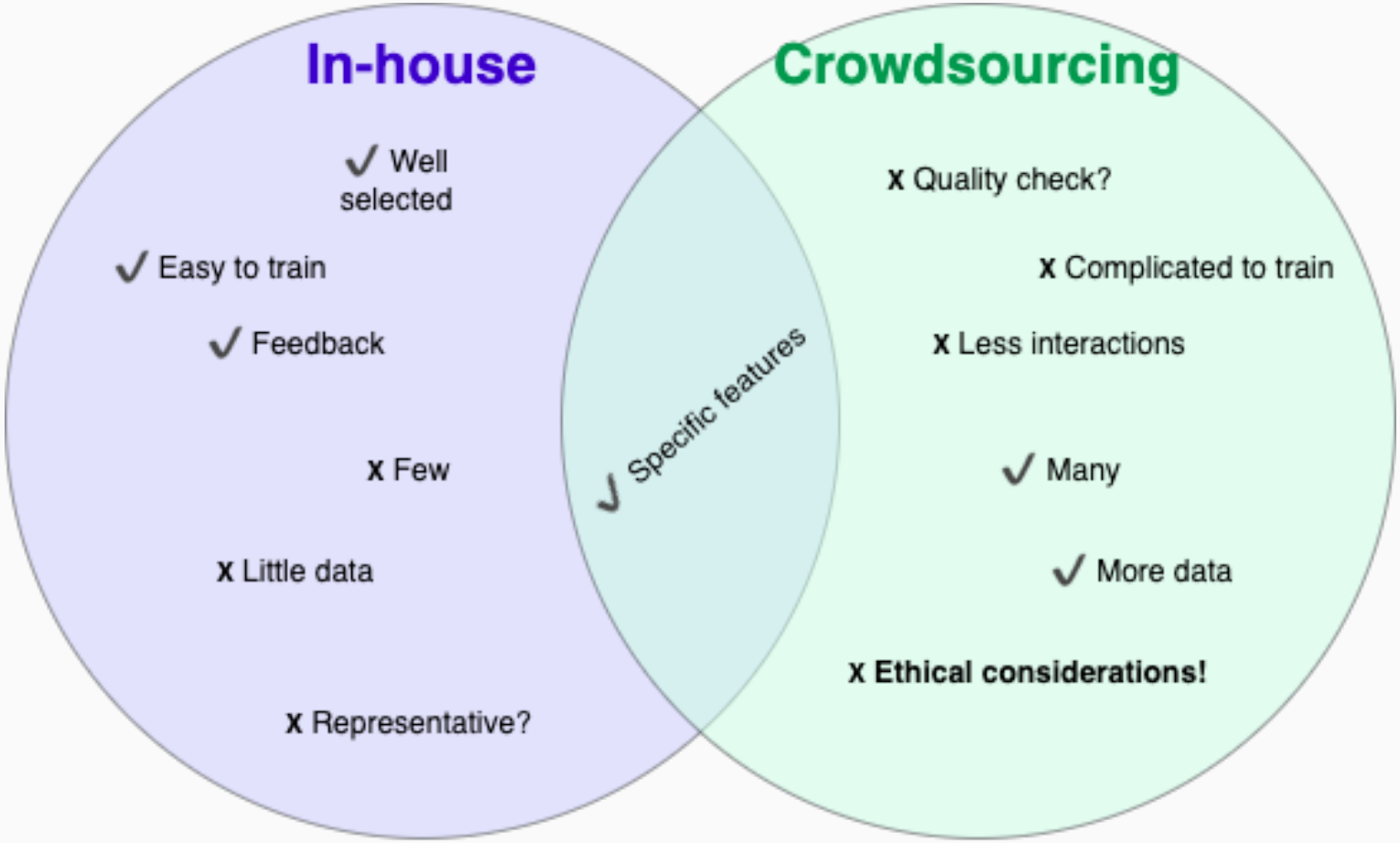
Choose the Right Tool



Choose the Right Tool



Recruit Participants (Annotators)



Inspired from the slides of Valerio Basile and Enrica Troiano on Data Perspectivism at ESSLLI 2023.

Evaluate the Annotations

How to assess the quality of each annotation?

$\rightarrow$ Assumption = Some annotators produce better annotations overall. We find them by:

- Having a common subset of data annotated by everyone;
- Comparing all the annotators' answers pairwise;
  $\rightarrow$ Annotators with very low agreement are sometimes ignored.
- larger the common subset $\leftrightarrow$ better assessment of the agreement
  BUT less annotated data in total

Summary so far

What you should remember:

- Define your goal clearly;
- Guidelines are crucial for the annotators;
- Transcribing and annotating is time-consuming and expensive!
  $\rightarrow$ Do not create a resource if it already exists.

One Transcription / Annotation Tool: ELAN

Table of contents

1. Transcriptions and Annotations
   - Examples of Transcription / Annotation Uses
   - Definition(s)
   - What is an Annotation Campaign?

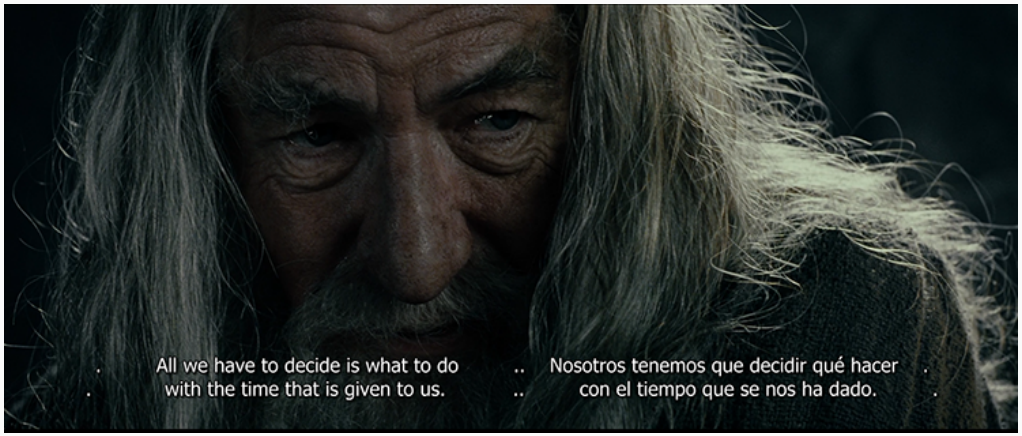
2. One Transcription / Annotation Tool: ELAN
   - ELAN?
   - Examples
   - Core Concept
   - Application and Assignment

What Does it Do?

ELAN makes time-aligned transcripts:

What Does it Do?

ELAN makes time-aligned transcripts:

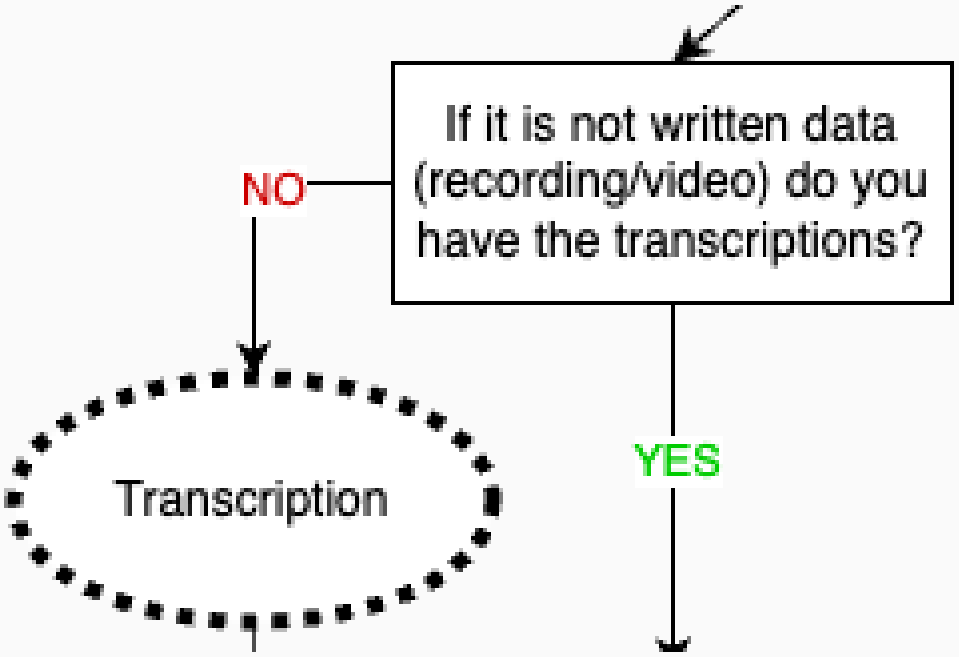


Match text (transcription and/or annotations)

with sections of an audio or video recording.

What Does it Do?

ELAN makes time-aligned transcripts:

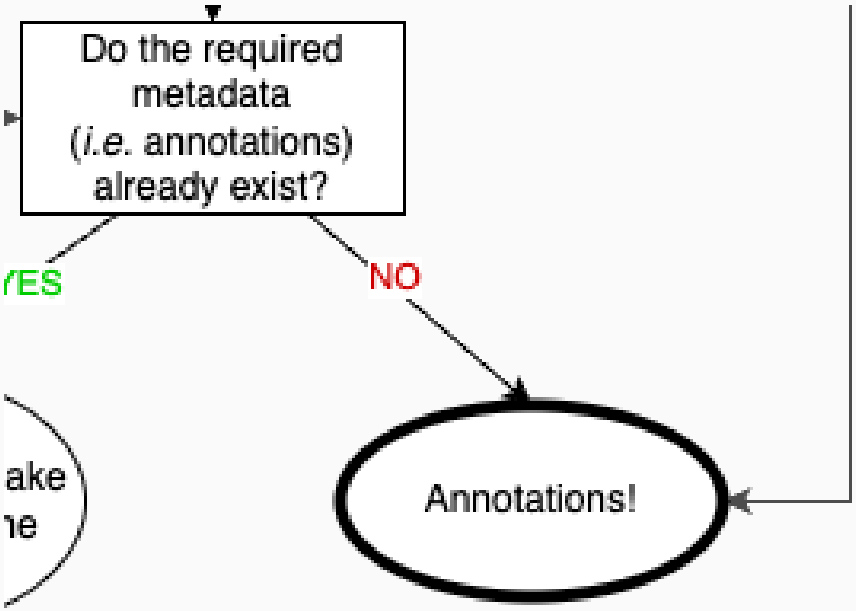


Match text (transcription and/or annotations)

with sections of an audio or video recording.

What Does it Do?

ELAN makes time-aligned transcripts:



Match text (transcription and/or annotations)

with sections of an audio or video recording.

What Does it NOT Do?

It is not an audio/video editor:

- Cannot edit media;
- If you change the media (with another software) you must change the ELAN file as well;

It is not a text editor:

- Output is ***plain text***;
- Cannot add embellishments.

Examples of Applications

Example 1: One language, one speaker

(Morality, provided by RadioLab *http://www.wnyc.org/shows/radiolab*)

Josh: "How do people make this judgment? Forget whether or not these judgments are right or wrong, just, what's going on in the brain that makes people distinguish so naturally and intuitively between these two cases? Which, from an actuarial point of view, are very very very similar if not identical..."

Examples of Applications

Example 2: One language, multiple speakers

(Excerpt: Television)

ALVIN: yeah, I haven't –
like admittedly,
I haven't,
(SNIFF) (0.3)
It's funny, I haven't watched those in years, you know,
PETER: Yeah.
ALVIN: so I've thought, (0.7) I've thought,
it might be fun to see them again.

Examples of Applications

Example 3: Two languages, richer linguistic info

(Pear Story, Kannada. Speaker: Keshava Subramanya)

KAN Sentence: Ii kathe obba, haṇṇu maaliya bagge, mattu, obba huḍugana bagge ide.

Intonation Unit	Ii	kathe	obba,	haṇṇu	maaliya
Morphemes	ii	kathe	obba	haṇṇu	maali -ya
Gloss	this	story	one	fruit	gardener- SRC

bagge,	mattu,	obba	huḍugana	bagge	ide.
bagge	mattu	obba	huḍuga -na	bagge	ide
about	and	someone.M	boy -GEN	about	be.3SM

Free translation: 'This story is about a fruit farmer and a boy.'

Concept of Tier

ELAN enables you to add as many layers of annotations as you want

$\rightarrow$ Everything that is the same "kind" of information is part of the same tier:

- All the transcription of a monologue (Example 1);
- All the utterances of one speaker in a dialogue (Example 2);
- All the glosses of the Kannada words in Example 3.

Concept of Tier

You are filming a 2-character puppet show designed to teach English speakers sentences in Irish:

- How many tiers do you need?
- What are they?

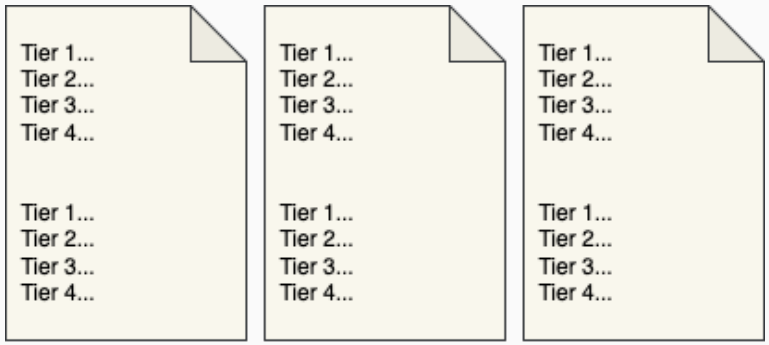
Concept of Tier

1. Fox's utterances in Irish
2. Translations of Fox's sentences into English.
3. Archer's utterances in Irish
4. Translations of Archer's sentences into English.

Concept of Tier

Printed transcripts have a fixed page width;

→ Forced to split content of single tier across many lines.



Concept of Tier

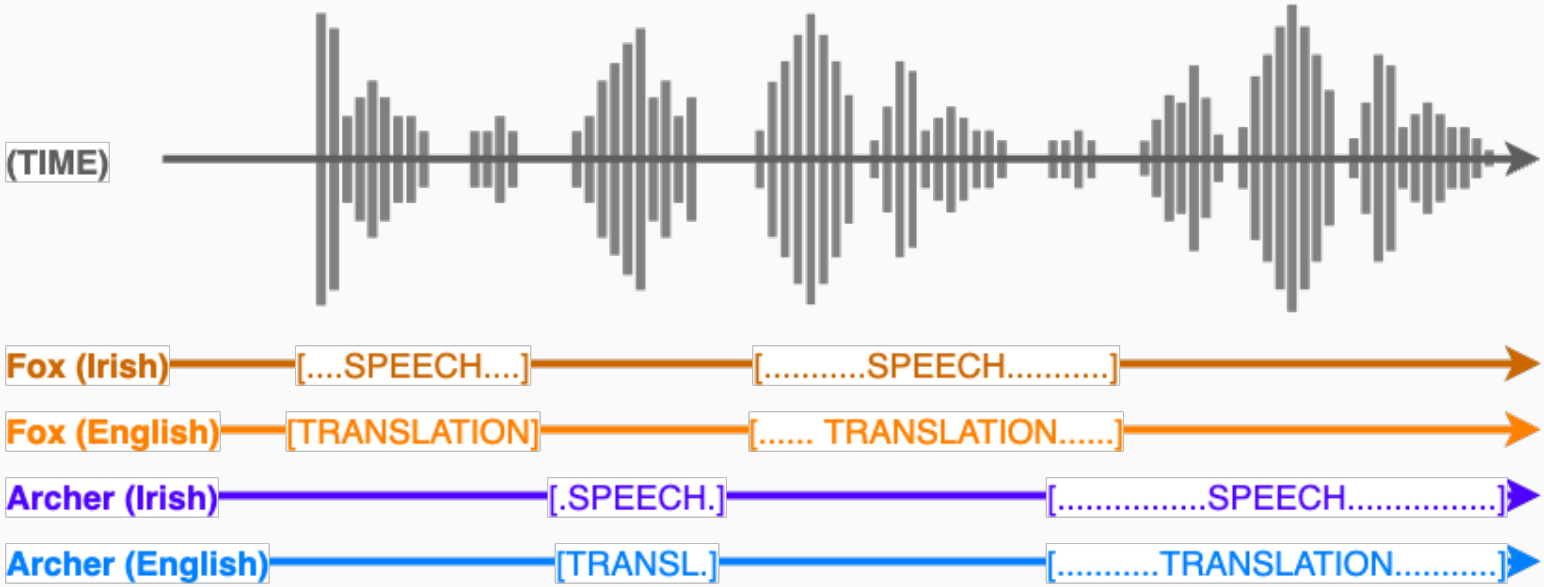
But ELAN is time-based!

Tiers are continuous and contiguous to the media timeline:

Concept of Tier

But ELAN is time-based!

Multiple continuous tiers can capture speaker overlap:



A Practice Example

Install ELAN:
https://archive.mpi.nl/tla/elan/download;

Download the recording we will work on (Canvas);

Open the recording in ELAN and:

- Transcribe it;
- Annotate the questions with question types (Wh-question, Yes-No-Question, Other);
- Consider that the goal of this experiment is to gather empirical data to determine the frequency at which people use the different question types;

You can find help on how to use ELAN on these slides.

Assignment

Write a paragraph on the transcription/annotation task you just did:

- How many tiers did you use and what were they?
- Why did you choose these tiers?
- Was the research question specific enough to make decisions regarding the tiers and the annotations? If not what additional information could have been helpful?
- What difficulties did you encounter?
- How much time did it take? Do you feel that it is a lot based on the length of the recording and the difficulty of the task?

It does not need to be very long, just show that you understand the concept of tier and how to take advantage of it for transcription and annotations. Also show that you are aware of the difficulties when it comes to creating transcriptions and defining an annotation task.

Send your paragraph by email to `amandine.decker@gu.se`

References i

- Berez-Kroeker, Andrea and Christopher Cox (2019). ***Linguistic Institute Course 353: Digital Methods in Language Documentation – Days 3&4: ELAN Lesson 1.*** Lecture Slides.
- Bird, Steven and Mark Liberman (2001). **"A formal framework for linguistic annotation".** In: *Speech Communication* 33.1. Speech Annotation and Corpus Tools, pp. 23–60.
- Dostal, Mateja (2021). ***Business English learner speech corpus SAPS.*** Slovenian language resource repository CLARIN.SI.
- Leech, Geoffrey (1997). **"Corpus annotation : Linguistic information from computer text corpora".** In: Longman, Londres, Angleterre. Chap. Introducing corpus annotation, pp. 1–18.