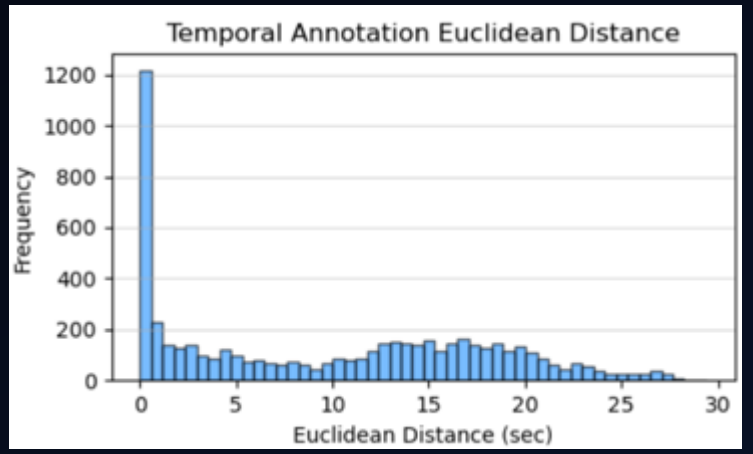
Annotation Quality

TEAM IMPORTED

Lóránd Heidrich
Gergő Márk Sere
Gergely Terényi
Diego Caparros Vaquer

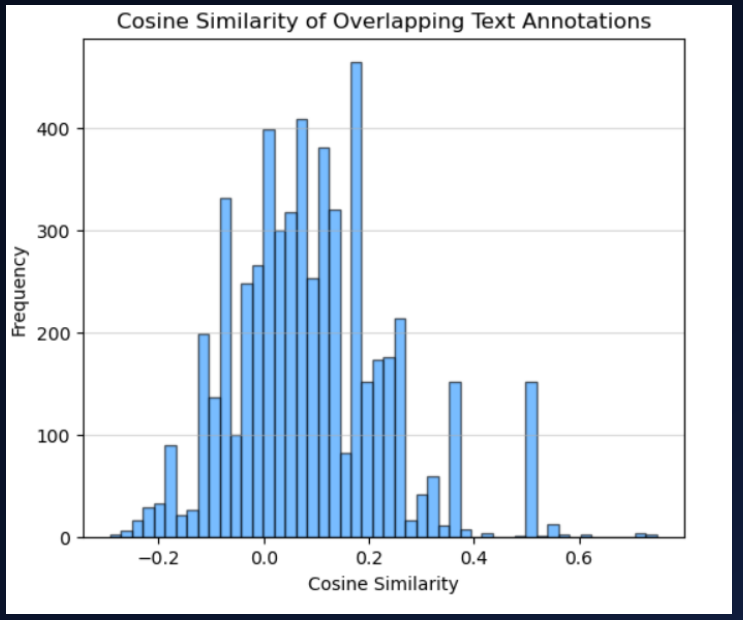
How precise are the temporal annotations?

- Majority agree on start/end time
- Long positive skew with gradual decay
- Euclidean distance: strong disagreement pattern (15-20 sec range)



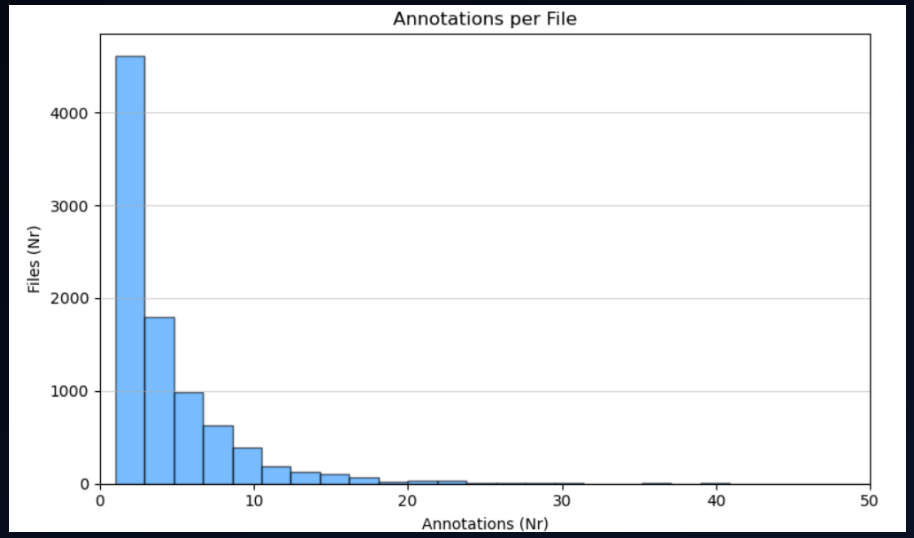
How similar are the text annotations that correspond to the same region?

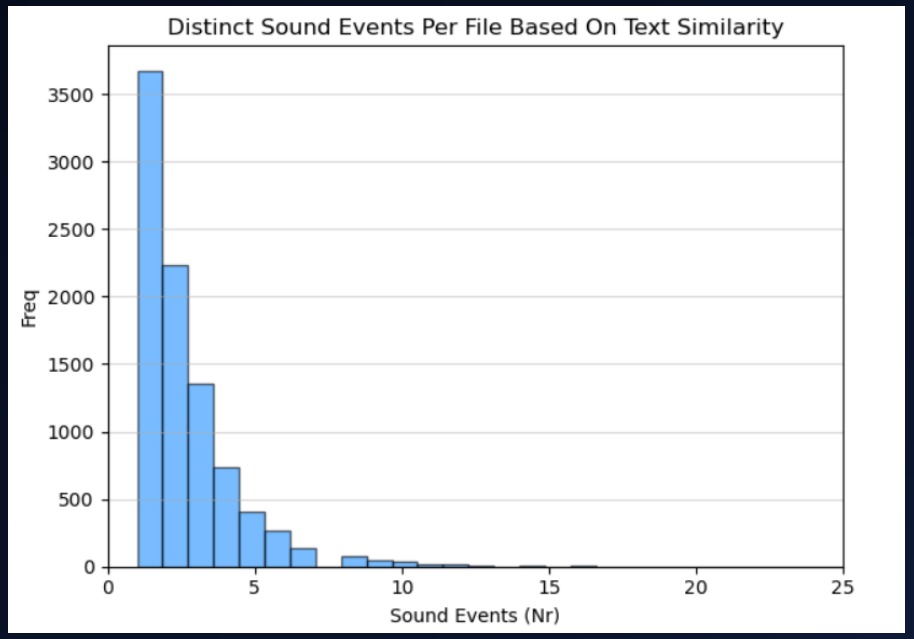
- Cosine similarity mostly 0-0.2 (mu=0.088, median=0.072)
- Annotators utilize different wording for overlapping timespans



How many annotations did we collect per file? How many distinct sound events per file?

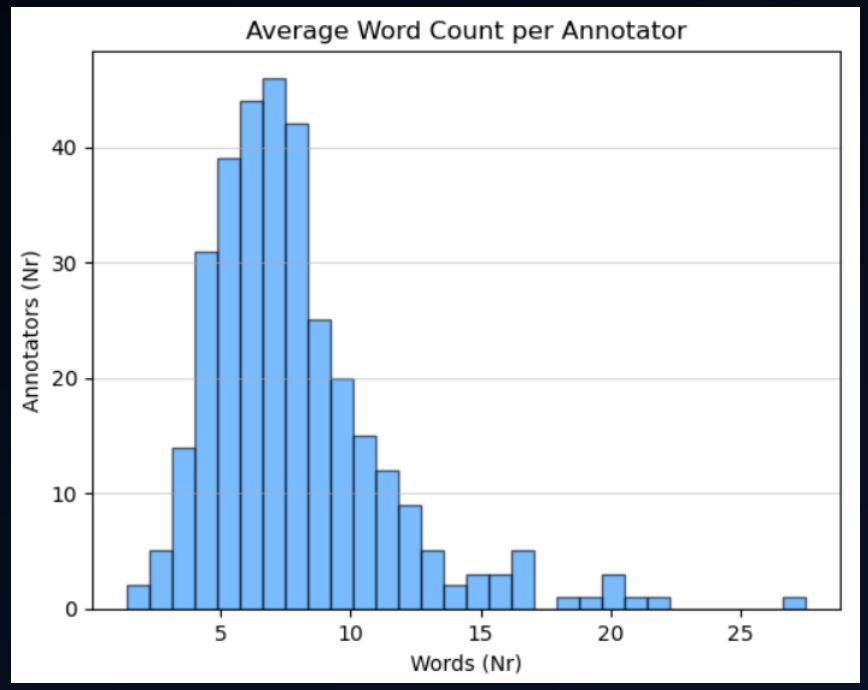
	Annotations /File	Distinct Sound Events/File
Total Unique Files	9026	9026
Total Annotations	35826	35826
Mean	3.9692	2.4253
Median	2	2
STD	4.4254	1.9063
Min	1	1
Q1	1	1
Q2	2	2
Q3	5	3
Max	96	27

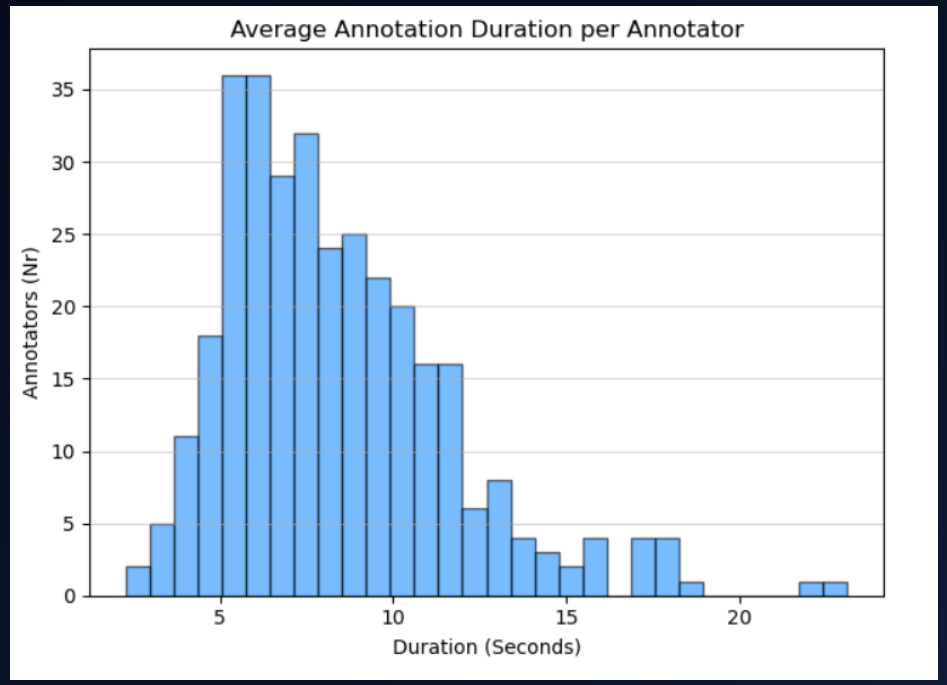




How detailed are the text annotations? How much does the quality of annotations vary between different annotators?

- Average annotation = 7.85 words, average duration = 8.38 sec
- Big differences across annotators: some use >20 words or long segments
- 5-10 sec and 5-10 word range is most common





Are there any obvious inconsistencies, outliers, or poor-quality annotations in the data? Propose a simple method to filter or fix annotations

- 15.25%–22.14% flagged poor using duration, word count, and spelling
- 26.6% of annotations under 5 words
- 3% had over 3 spelling mistakes, 49.56% had none
- Proposed filter:
  - Flag extreme durations
  - Flag short annotation texts
  - Check spelling with pyspellchecker

Duration Outcomes	IQR	99%	95%
Text too short (annotations)	0	355	1792
Text too long (annotations)	0	359	1792
Text too short (%)	0%	0.9909%	5.0020%
Text too long (%)	0%	1.0021%	5.0020%

Word Count Outcomes	1%	5%	IQR Lower	Explicit 5
Word threshold	2	2	-3.5	5
Annotation Count	335	335	0	9537
Annotation Proportion	0.9351 %	0.9351 %	0%	26.6203 %