Evaluating the Impact of LLM-Assisted Annotation in a Perspectivized Setting

the Case of FrameNet Annotation

Authors:
Frederico Belcavello[1], Ely Matos[1], Arthur Lorenzi[1], Lisandra Bonoto[1], Lívia Pádua Ruiz[1], Luiz Fernando Pereira[1], Victor Herbst[1], Yulla Navarro[1], Helen de Andrade Abreu[1], Lívia Vicente Dutra[1 3] and Tiago Timponi Torrent[1 2]

Affiliations:
[1] FrameNet Brasil, Federal University of Juiz de Fora
[2] Brazilian National Council for Scientific and Technological Development – CNPq
[3] Gothenburg University

THE CONTEXT

Self_motion

Mark [SELF_MOVER] was **running** from home [SOURCE] to school [GOAL] along the road [PATH].

Buildings

Mark was running from **home** [BUILDING] to school along the road.

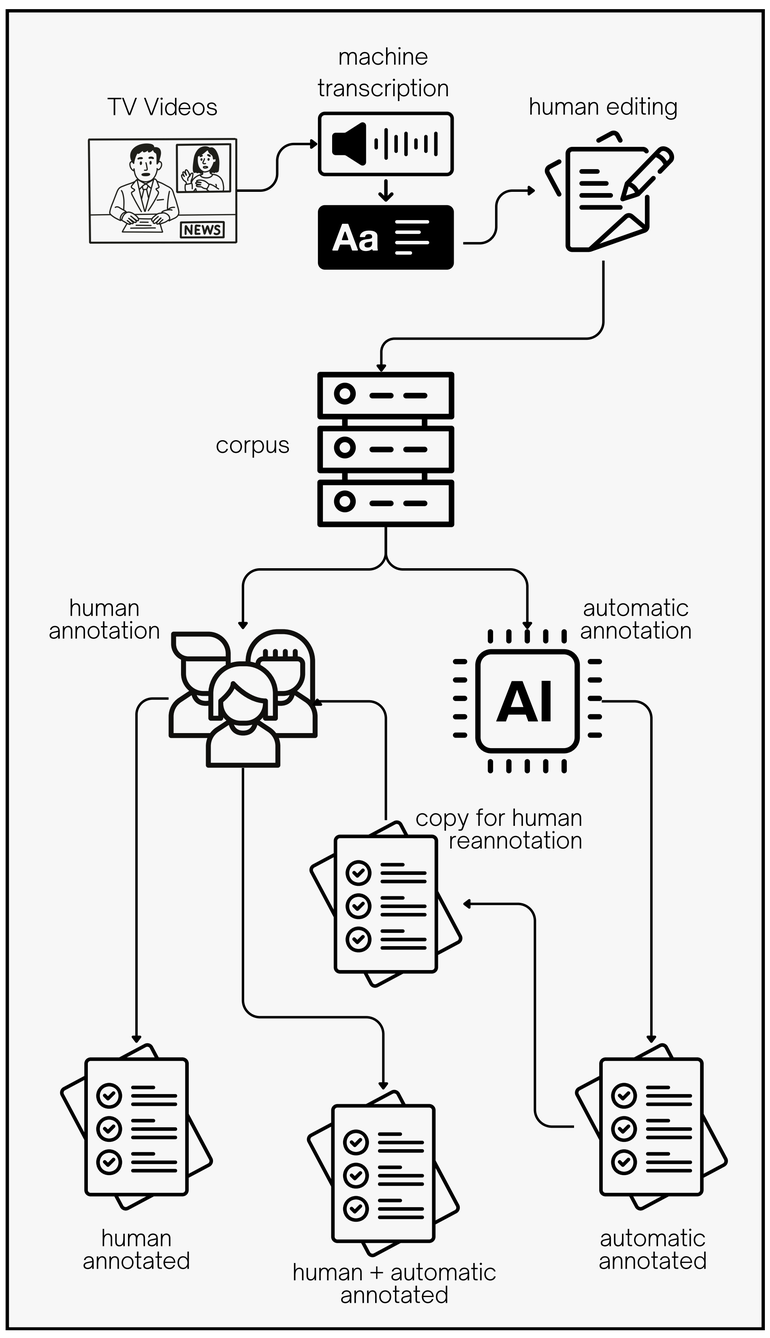
Meticulous manual curation by trained linguists ensures resource quality but limits scalability due to the significant human effort and time required.

THE QUESTION

Can LLMs be used as assistants to facilitate, accelerate, and improve the quality of FrameNet annotation?

THE EXPERIMENT DESIGN

- Data source:
  - 311 sentences from 12 documents from a Brazilian TV News corpus: **Framed News** (3,442 sentences)
- 3 annotation settings:
  - **Manual** (human-only) - phase 1
  - **Automatic** (machine-only) - phase 2/step 1
  - **Semi-automatic** (machine + human) - phase 2 / step 2
- Tool for machine annotation:
  - **LOME** (Xia et al., 2021) parser: chosen for its multilinguality and ability to provide both frame and Frame Element (FE) labels without requiring manual pre-processing of targets.



THE RESULTS

1. Increased diversity

The semi-automatic setting result suggests that LLM-generated suggestions may prompt human annotators to consider a wider range of frame interpretations and perspectives.

Annotation Setting	Avg. Sentences per Document	Avg. Unique Frames per Document	Avg. Frames per Sentence
Human	19.75	67.91	3.80
Machine (LOME)	21.83	52.66	2.50
Machine + Human	22.91	**80.91**	3.74
Human	19.75	67.91	3.80

2. Preservation of human judgment

The highest similarity score suggests that while humans refined the output, they tended to keep and build upon the frames originally suggested by the machine.

Comparison Pair	Average Cosine Similarity
Human vs. Machine	0.6320
Human vs. Machine + Human	0.7465
Machine vs. Machine + Human	**0.7818**

3. Annotation of core elements

Human annotators excel at identifying core FEs, including those that are inferred from context (null instantiations).

Annotation Setting	Avg. Core FEs Identified	Avg. Minimal Required FEs	Avg. Coverage (%)
Human	290.50	284.83	**95.79%**
Machine (LOME)	95.75	279.17	34.20%
Machine + Human	**338.00**	**353.67**	90.65%

4. Modest time saving

The average time saved was 1.99 min per sentence, but 7 annotators performed faster manualy, while 5 performed faster using pre-annotated sentences.

Annotation Setting	Avg. Time per Sentence (minutes)
Human (Manual)	14.96
Machine + Human (Semi-automatic)	12.97
Average Time Saved	**1.99**

5. Useful machine suggestions

The percentage of updated annotations indicates that machine suggestions provide a useful, albeit imperfect, foundation that humans can refine and improve.

Edit Type	Avg. Frequency (Count)	Percentage (%)
UPDATED (Modified machine output)	130.83	**65.45%**
DELETED (Rejected machine output)	42.75	19.68%
CREATED (New labels added by human)	17.50	10.08%
ACCEPTED (Kept as-is)	11.67	6.61%

THE CONCLUSION

It is a viable path for large-scale growth

Semi-automatic annotation increases coverage and diversity while preserving human judgment, even if speed gain remains modest.

THE OUTLOOK

Pipeline should be extended

Future work should look at the inclusion of other types of semantic role labelers—such as DAISY (Torrent et al., 2024) to serve as a post-processing step for LOME.

Ensure Core FEs annotation

Future work should also implement stricter automated policies to ensure the minimum required core FEs are always identified, which also implies recognizing and recording null instantiations.

DOWNLOAD THE PAPER

Acknowledgments

Research reported in this paper was developed under the ReINVenTA—Research and Innovation Network for Vision and Text Analysis of Multimodal Objects—initiative, funded by the Minas Gerais State Agency for Research and Development (FAPEMIG– grant RED-00106-21) and the Brazilian National Council for Scientific and Technological Development (CNPq– grant 420945/2022-9). The resulting dataset will be part of the data collection of the National Science and Technology Institute for Responsible Artificial Intelligence, Computational Linguistics and Information Treatment and Dissemination (INCT-TILDIAR, CNPq grant 408490/20241). Belcavello was supported by CNPq (grant 200270/2023-0). Lorenzi, Abreu and Pereira were supported by FAPEMIG. Bonoto, Ruiz, Herbst and Navarro were supported by CNPq. Torrent is a CNPq research productivity grantee (grant 311241/2025-5). The presentation of this paper at LREC2026 is supported by the Institute of Artificial Intelligence at the National Laboratory for Scientific Computing (IIA-LNCC), an initiative led by the Ministry of Science, Technology, and Innovation (MCTI) as part of the Brazilian Artificial Intelligence Plan (PBIA).