



Natural Language Processing: international standards to bridge the old and the new

Lauriane Aufrant
(ISO/IEC JTC 1/SC 42/JWG 5
Convenor)



Natural Language Processing

Natural Language Processing

- The field of AI that is concerned with (human) language
 - Text, speech... sign languages... or multimodal!
- Very diverse!
 - Machine translation, speech recognition, OCR, text-to-speech, automatic summarization...
 - But also: chatbots, question answering, information retrieval, document clustering...
 - As well as: tokenization, syntactic parsing, language identification, named entity recognition, speaker diarization, information extraction, language error correction, terminology extraction...

NLP & LLMs

- NLP has worked on language modeling for decades
 - What changed in the 2020s: model performance, more diverse uses, the exact methods, model size...
 - But: still language models → hence already applicable good practices
- Our challenge: standardize what already exists, but aligning with the new interests
 - ... and suddenly increasing interests

NLP & Generative AI

- GenAI is mostly NLP
 - Ex: “Chatbot” = “dialogue management” task. “Question answering” is a long-standing task.
- GenAI is not only NLP
 - Ex: deep fake generation of faces. Multimodal LLMs: in part NLP, in part not NLP...
- NLP is not only GenAI
 - Ex: information extraction, language identification, document clustering...
 - Blurrier: automatic summarization, machine translation... depends on the definition of GenAI?



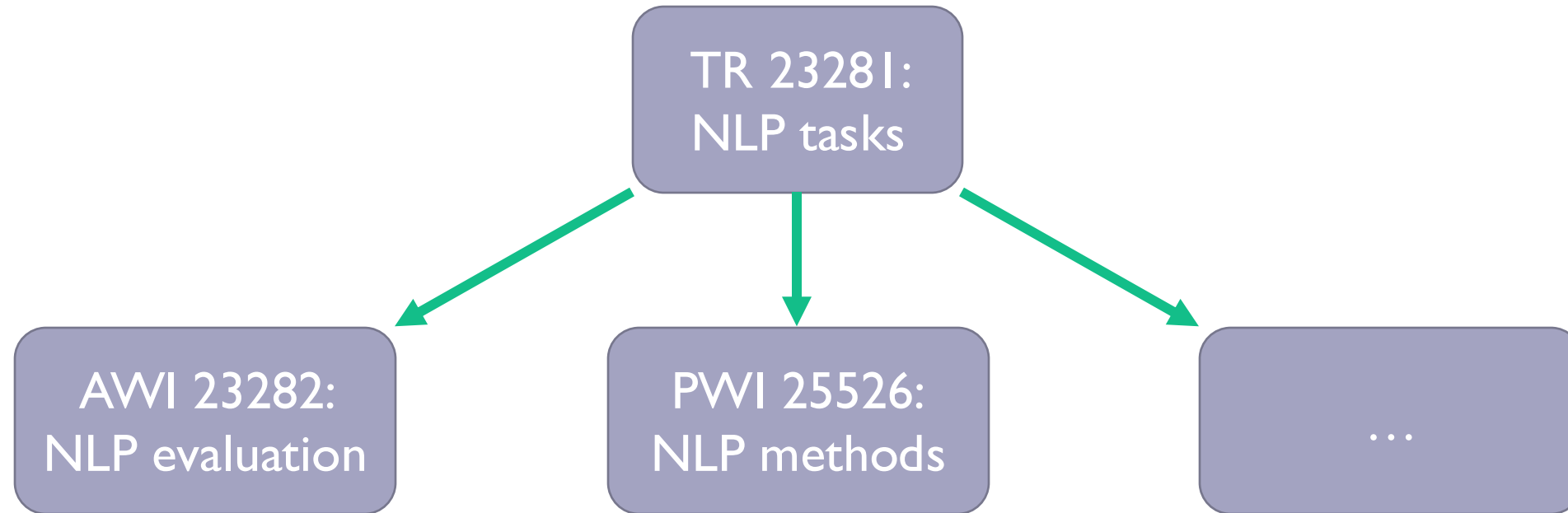
Ongoing work on NLP

ISO/IEC JTC 1/SC 42/JWG 5

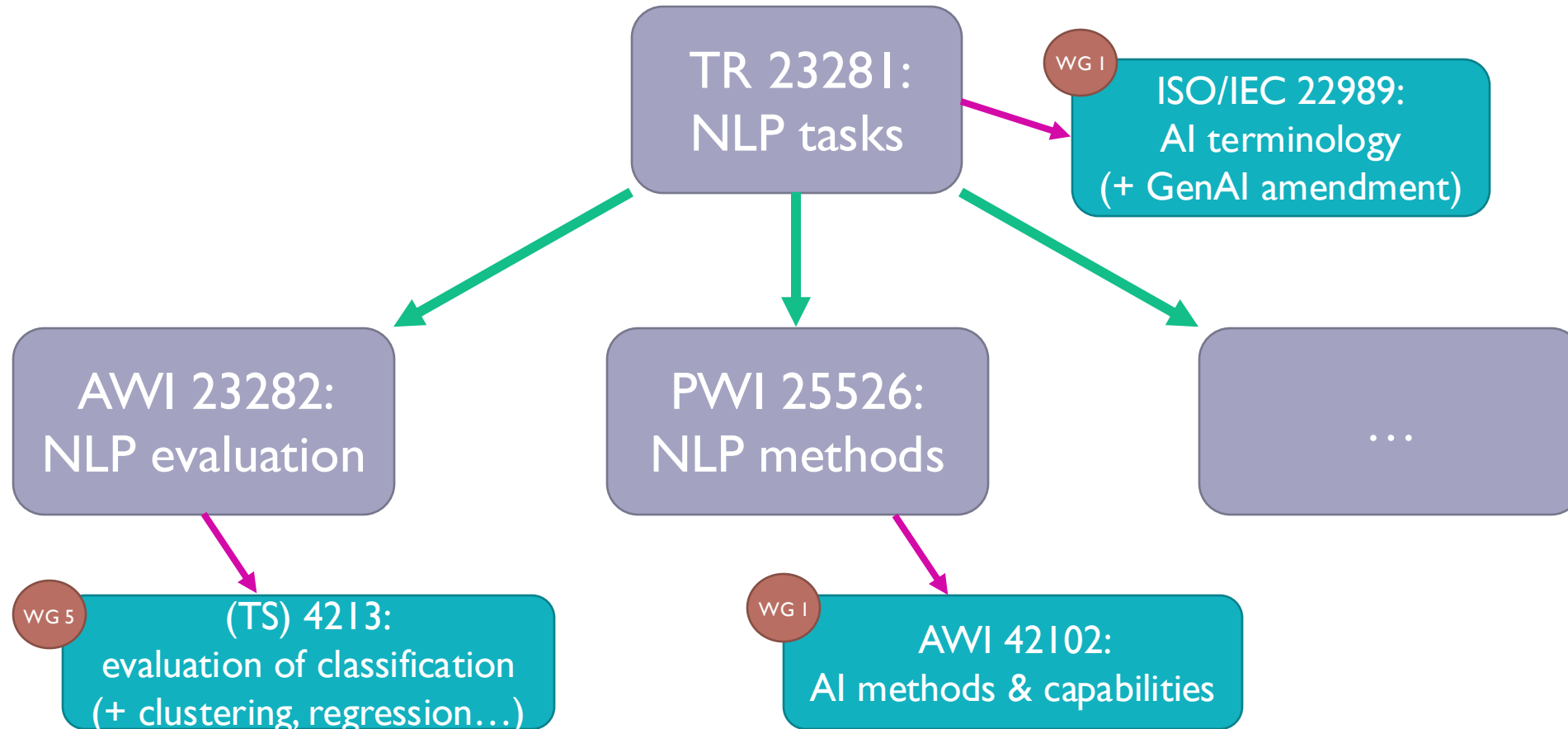
“Natural Language processing”

- Joint Working Group with ISO/TC 37 “Language and terminology”
- Incubated in 2023, kick-off in 2024
- Co-Convenors: Lauriane Aufrant (SC 42), Avashlin Moodley (TC 37)
- Editors and subgroup leaders: Rania Wazir, Fred Bane, Peter Chiu, Wanzhong Ma
- Collaboration with CEN-CENELEC JTC 21 and the EU AI Act

NLP tasks as the foundation



NLP tasks as the foundation



In incubation

- Study item on NLP corpora
 - ➔ Interplay with work items in SC42 (5259 on data), TC37 (corpus annotation), in JTC21 (EN Datasets)
- FDIS 12792 on Transparency ➔ Transparency for NLP?
- Robustness? Bias? Others?
 - ➔ Coordinating with new proposals throughout SC 42 on GenAI: risk management, red teaming, data quality... while building on TC37 work

JWG 5 philosophy

- Internally: NLP-specific work items
 - Not always needed
- Advisory role to all other groups: incubating items and content
 - AHG on Cross-group NLP content
- Extensive efforts on gap analysis
 - GenAI or NLP? Existing practices or new practices?

Thank you

Lauriane Aufrant

first.last@inria.fr

