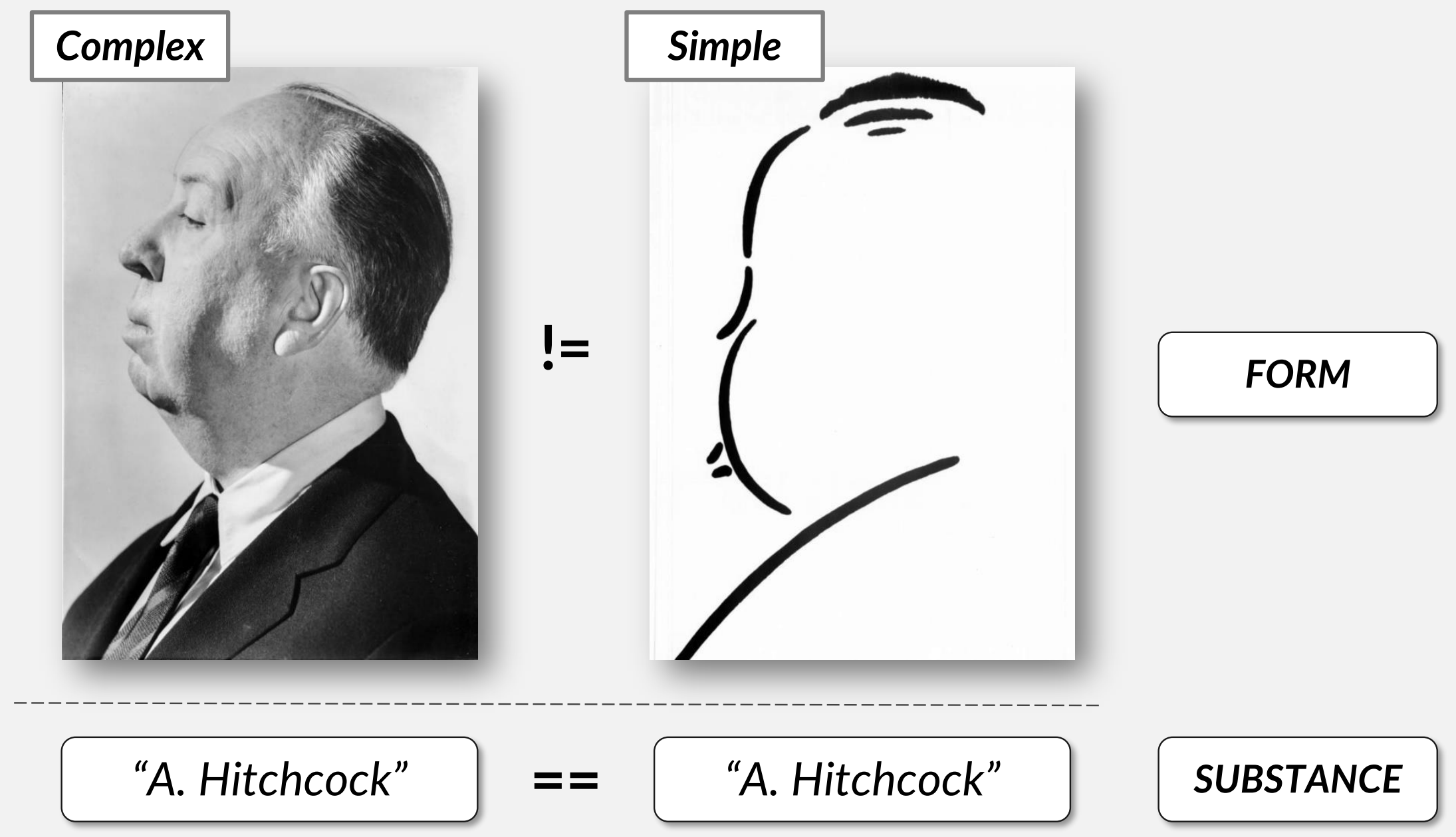
From Complex to Simpler Transcriptions: Simplifying Spontaneous French Speech

Lucía Ormaechea and **Nikos Tsourakis**

Department of Translation Technology – University of Geneva – Switzerland

Introduction

Automatic **T**ext **S**implification (**ATS**) → Conversion of texts into **simpler variants**, by reducing their **linguistic complexity**, while preserving their **original meaning** [Stajner, 2021].


Complex

Simple

!=

FORM

"A. Hitchcock" == "A. Hitchcock"

SUBSTANCE

Providing simplified versions of texts has seldom been applied to a speech input

Accessibility purposes	**Raw transcripts** → Challenging to understand **Simplified transcripts** → Helpful for diff. target audiences
Ancillary purposes	**Raw transcripts** → Hard to process by NLP pipelines **Simplified transcripts** → Helpful for subtitling, speech-to-pictograph translation

Challenges in speech simplification

1 Simplification: a written-text-centered task

Newswire articles [Xu et al, 2015]	Wiki-based content [Ormaechea & Tsourakis, 2023]	Healthcare documents [Goldsack et al, 2022]

✕ **Lack of speech-specific simplification data resources**

2 Complexity is represented differently

- **Information structure**: On-the-fly process
- **Spontaneity and grammaticality**: Traces of real-time construction

3 Lack of guidelines to steer the simplification process

Need to characterize the process empirically:

- **Intuitive approach** [Allen, 2009].
- Based on the **criteria** of **expert linguists**.
- **Output comparison** with **ChatGPT**.

Methodology

I) To analyze speech simplification strategies, we resort to:

SOURCE DATASET

	# utterances
Orféo-Test [Pupier et al, 2022]	21,459

↓ *Proportionate random stratified sampling*

	# utterances
Orféo-Simple	100

↓ *SIMPLIFICATION TASK*

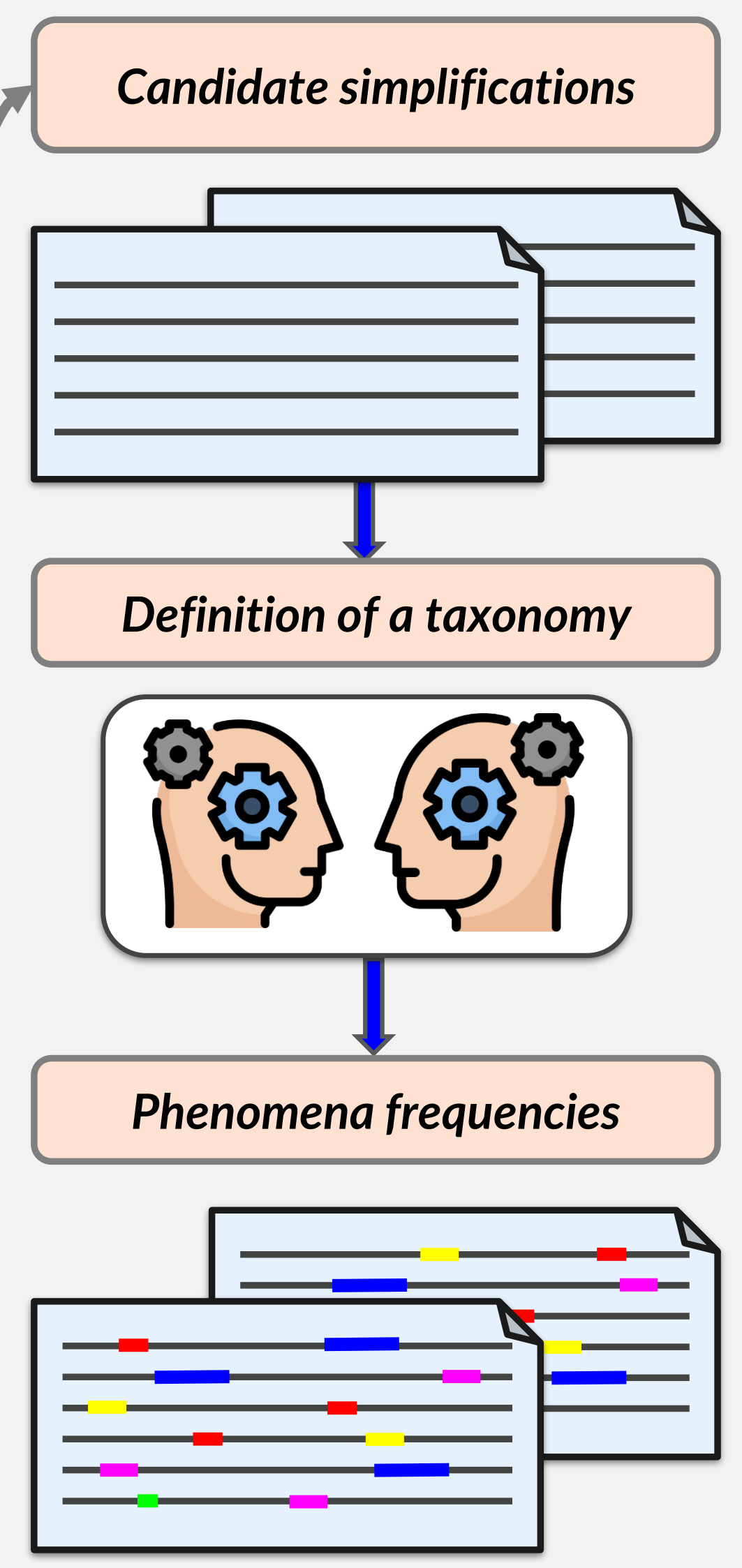
Machine-based (ChatGPT)
- **Identical prompt** than one used with humans.

Human-based
- **LimeSurvey** platform.
- **Linguists' profile**: French native speakers, background on linguistics.

PROMPT

```
completion = client.chat.completions.create(
    model="gpt-4-0125-preview",
    messages=[
  {"role": "user", "content": f"Notre corpus est
  constitué de phrases en français qui proviennent
  de transcriptions de discours spontané. \ Nous
  souhaiterions obtenir leur équivalent simplifié,
  c’est-à-dire, une phrase qui soit linguistiquement
  plus simple, sans pour autant perdre le sens et
  les informations originales. L’objectif est
  d’obtenir des phrases plus compréhensibles pour
  des locuteurs non natifs du français. \
  Pour chaque phrase, il vous est demandé de : \
  1. Transformer la phrase donnée en une version
     plus simple. Utilisez un langage clair, en
     évitant le jargon et les constructions
     grammaticales complexes. Vous pouvez également
     ajouter des signes de ponctuation si
     nécessaire. \
  2. Expliquer votre raisonnement. Après chaque
     simplification, énumérez et expliquez les
     transformations que vous avez effectuées. \
  Voici les phrases à simplifier : \
  {sentence} \
  Voici le modèle pour ta sortie : \
  SIMPLIFICATION : \
  RAISONNEMENT :"}],
  temperature=0)
```

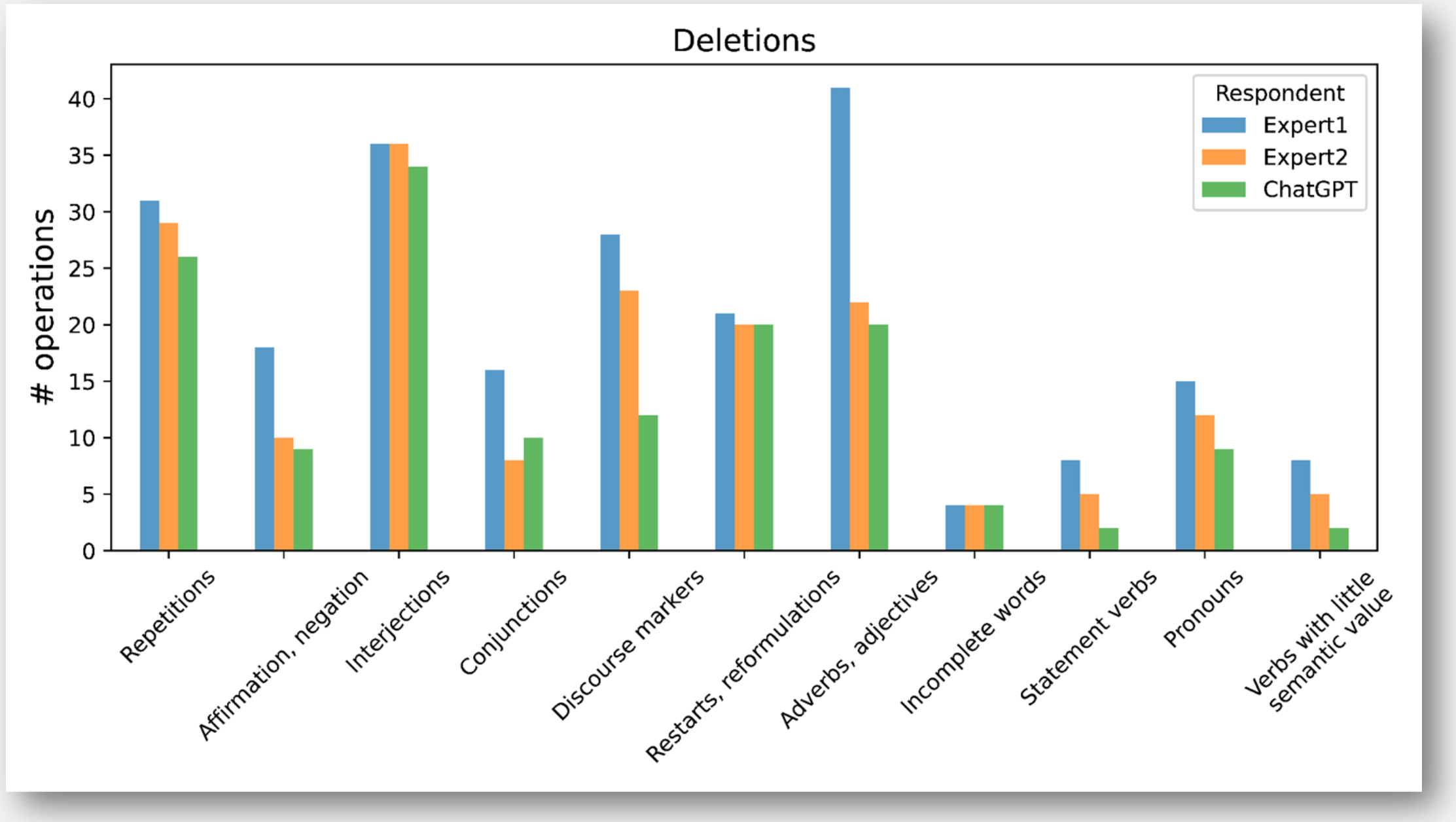
II) Once the collection is completed:

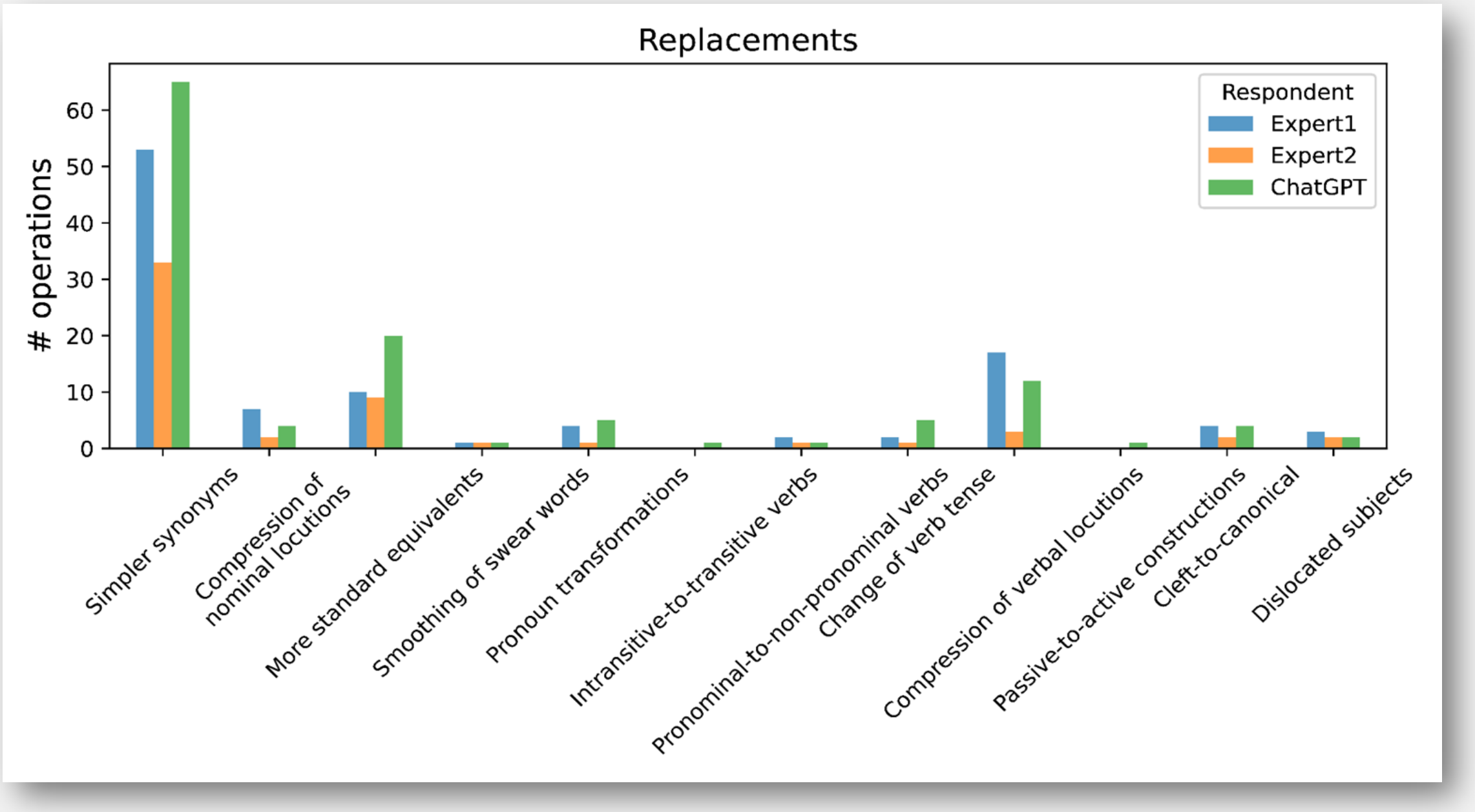

Candidate simplifications

Definition of a taxonomy

Phenomena frequencies

Quantitative evaluation





Input	**ouais** c'est ça sauf que **moi** on m'a jamais expliqué le rythme **du coup**
Exp. 1	*On ne m'a pas expliqué le rythme*
Exp. 2	*Oui, c'est ça, sauf qu'on ne m'a jamais expliqué le rythme*
GPT	*Oui, c'est vrai, mais personne ne m'a jamais expliqué le rythme*

Input	**on sent que** la prise de conscience de ce genre de choses **elle s' est faite** tard
Exp. 1	Nous pensons que la compréhension de ce problème est arrivée tard
Exp. 2	La prise de conscience de ces choses-là est arrivée tard
GPT	Les gens ont commencé à comprendre ces choses tard

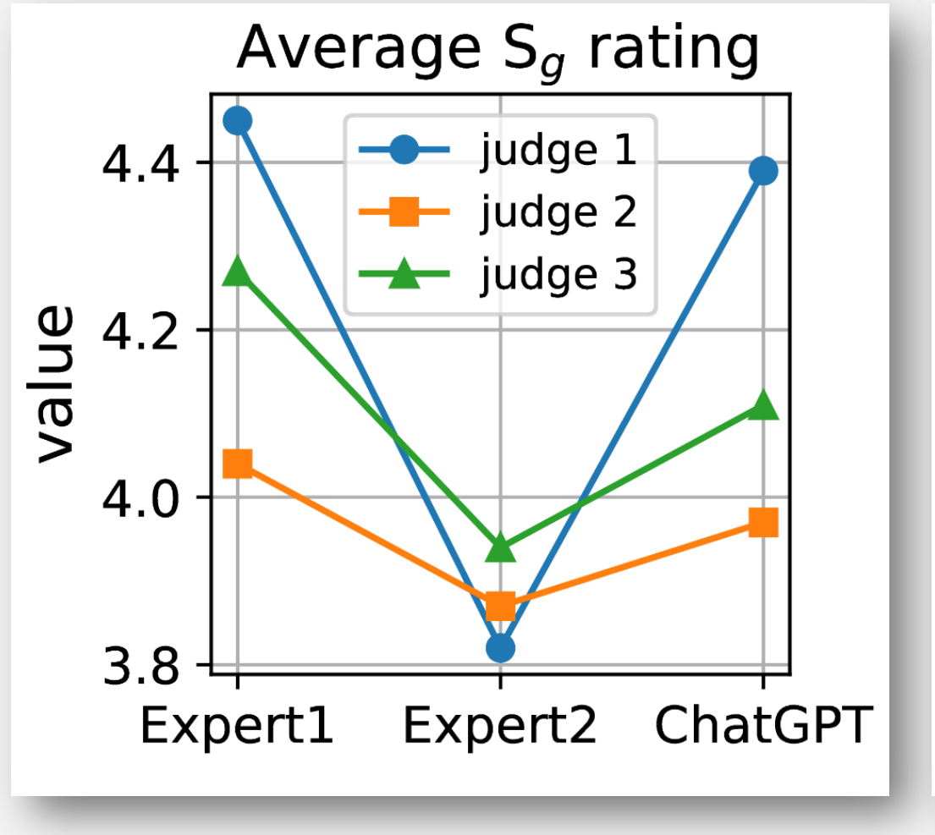
Simplified utterances seem to be "**writified**" or **register-standardized versions** of the **inputs** that just include their **propositional content** and erase all traces related to the **enonciation**.

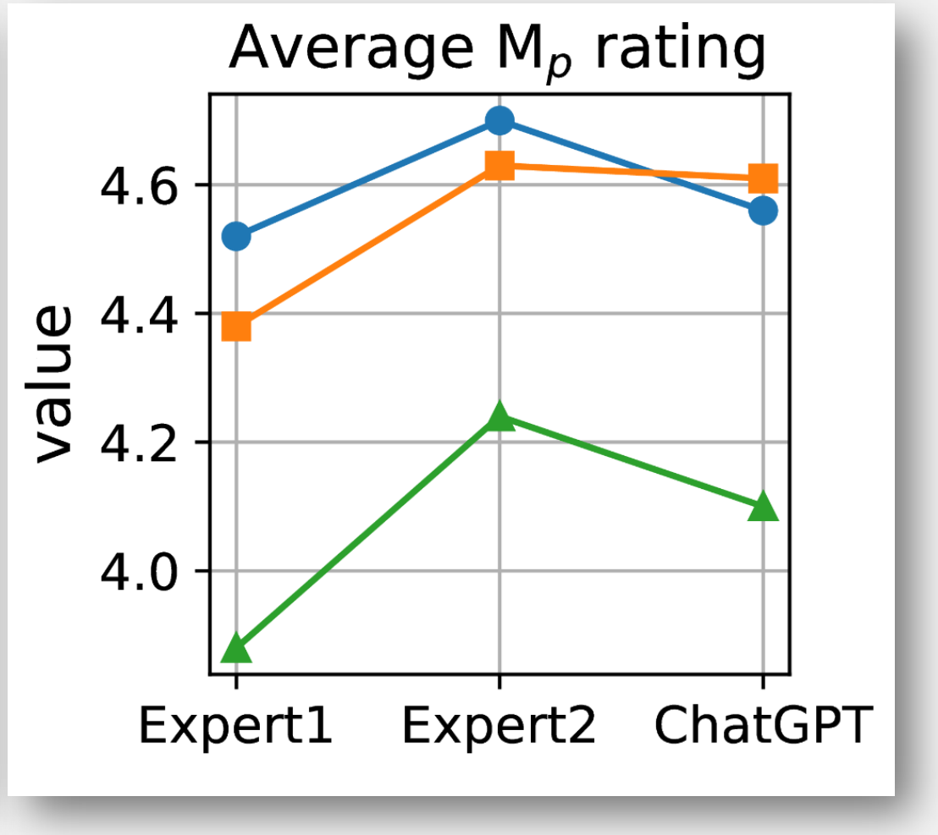
Lexical changes (more frequent):
- **Simpler** equivalents for **content words**.
- **Smoothing** of **slang** and **profanity**.

Syntactic changes (less frequent):
- **Standardization marked** information **structures**.

Qualitative evaluation

- **Respondents**: a group of **3 non-native master students.**
- **Scoring** of the simplifications on a **5-point Likert scale,** for two dimensions: **Simplicity gain (Sg)** and **Meaning preservation (Mp).**





Conclusions and further work

- **Taxonomy** and **quantification** of simplification operations applied to French spontaneous transcripts.
- **Small-scale study** → Due to the **costly process**.

Results show that: **speech decomplexification ≅ speech despontanefication**

New dataset: **Propicto-Orféo-Simple***
Further use as an **evaluation set** for **speech simplification models**

Acknowledgements

We would like to thank the linguists for the completion of the manual simplification task, and to the participants who completed the qualitative evaluation.

References

- **S. Stajner (2021)**. Automatic Text Simplification for Social Good: Progress and Challenges. *Findings of the Association for Computational Linguistics*, 2637-2652.
- **L. Ormaechea & N. Tsourakis (2023)**. Extracting Sentence Simplification Pairs from French Comparable Corpora Using a Two-Step Filtering Method. *Proceedings of the 8th Swiss Text Analytics Conference (SwissText)*, 30-40.
- **D. Allen (2009)**. A study of the role of relative clauses in the simplification of news texts for learners of English. *System, 37*, 585-599.

*This QR code links to the Ortolang repository where we made available the **Propicto-Orfeo-Simple** dataset.

Contact me: Lucia.OrmaecheaGrijalba@unige.ch