Simplification Strategies in French Spontaneous Speech

DeTermIt! Workshop: Evaluating Text Difficulty in a Multilingual Context

Lucía Ormaechea[1,2], Nikos Tsourakis[1], Didier Schwab[2], Pierrette Bouillon[1] and Benjamin Lecouteux[2]

[1] Department of Translation Technology – University of Geneva – Switzerland
[2] GETALP Team – University of Grenoble-Alpes – France

Overview

1. Introduction

1.1. ATS: a written-text-centered task?

- **A**utomatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** and **grammatical coherence** [Horn et al, 2014; Stajner, 2021].

1.1. ATS: a written-text-centered task?

- **A**utomatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** and **grammatical coherence** [Horn et al, 2014; Stajner, 2021].

Complex

FORM

1.1. ATS: a written-text-centered task?

- Automatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** and **grammatical coherence** [Horn et al, 2014; Stajner, 2021].

Complex

Simple

!=

FORM

1.1. ATS: a written-text-centered task?

- **A**utomatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** and **grammatical coherence** [Horn et al, 2014; Stajner, 2021].

1.1. ATS: a written-text-centered task?

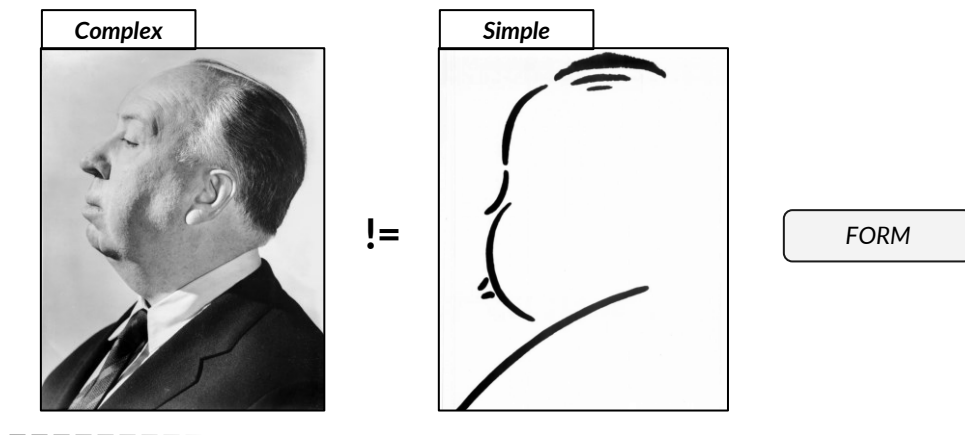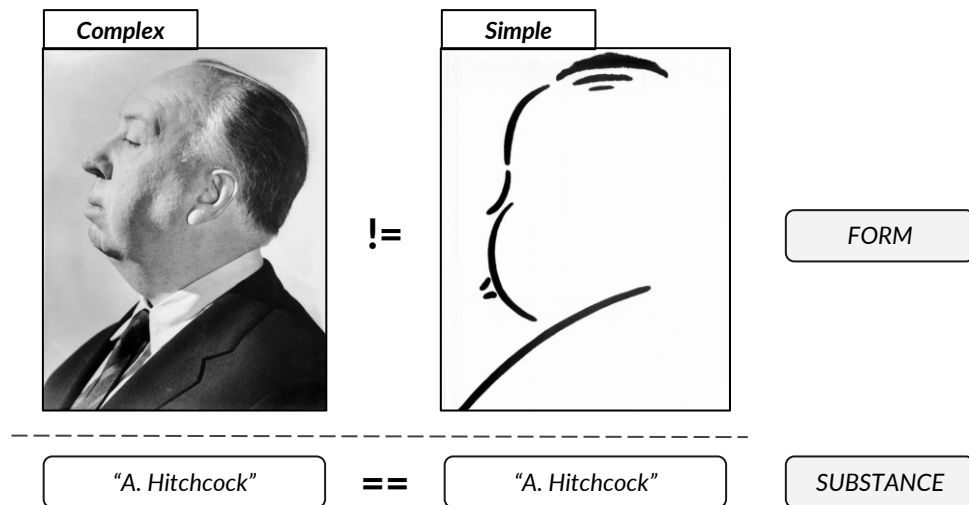- **A**utomatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** and **grammatical coherence** [Horn et al, 2014; Stajner, 2021].
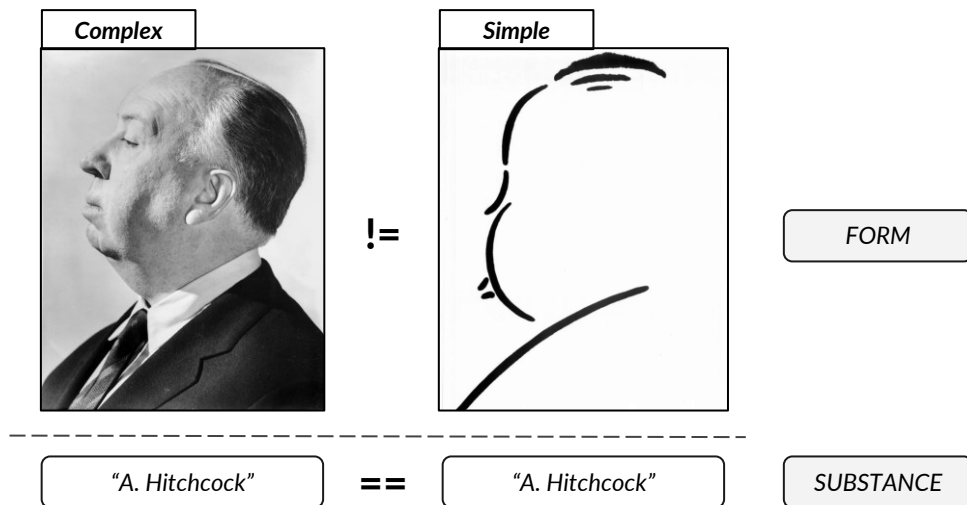
- Providing **simplified versions** of texts has been typically applied for a **written modality**:

  - **Newswire articles**, *i.e.,* Newsela [Xu et al, 2015].

Complex		Simple	

!=   FORM

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

1.1. ATS: a written-text-centered task?

- **A**utomatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** and **grammatical coherence** [Horn et al, 2014; Stajner, 2021].
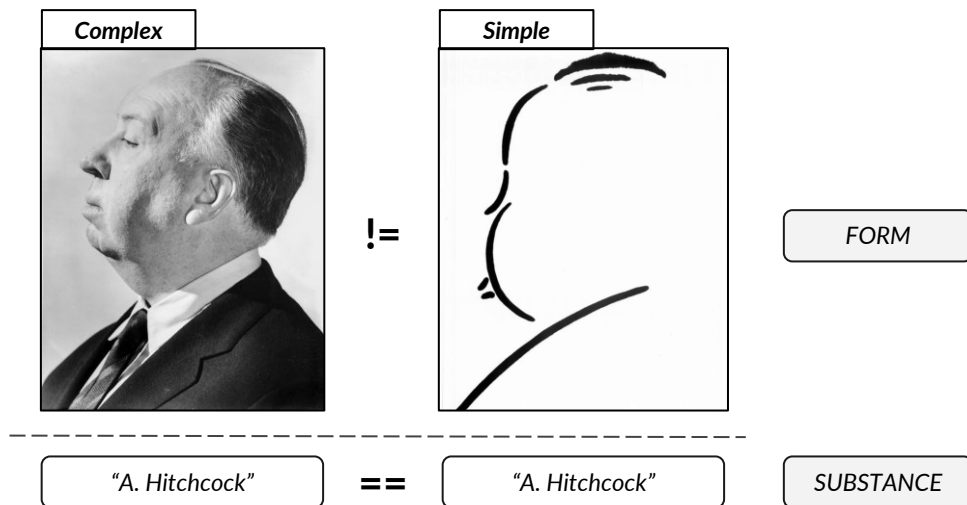
- Providing **simplified versions** of texts has been typically applied for a **written modality:**

 - **Newswire articles**, *i.e.,* Newsela [Xu et al, 2015].

  - **Wiki-based content**, *i.e.,* WikiLarge [Zhang & Lapata, 2017], WiViCo [Ormaechea & Tsourakis, 2023].

Complex		Simple

!= FORM

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

1.1. ATS: a written-text-centered task?

- **A**utomatic **T**ext **S**implification (**ATS**) is an area of NLP that aims at automatically converting texts into **simpler variants**, by **reducing their linguistic complexity**, albeit **preserving their original meaning** and **grammatical coherence** [Horn et al, 2014; Stajner, 2021].

- Providing **simplified versions** of texts has been typically applied for a **written modality:**

  - **Newswire articles**, *i.e.,* Newsela [Xu et al, 2015].

  - **Wiki-based content**, *i.e.,* WikiLarge [Zhang & Lapata, 2017], WiViCo [Ormaechea & Tsourakis, 2023].

  - **Healthcare documents**, *i.e.,* PharmMT [Li et al, 2020], BioLaySumm [Goldsack et al, 2022].
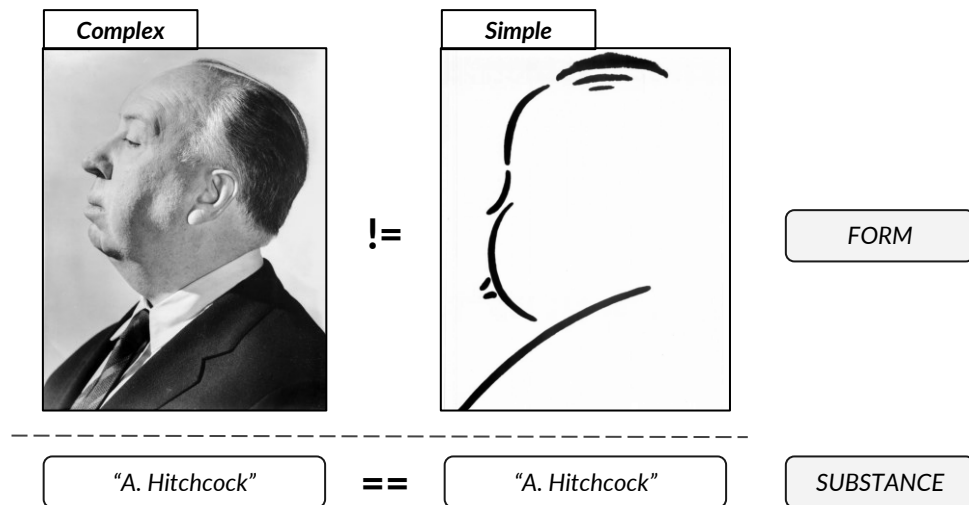
Complex		*Simple*	
	!=		FORM
"A. Hitchcock"	==	"A. Hitchcock"	SUBSTANCE

1.2. Challenges in Speech Simplification (SpeechSimp)

What about applying simplification to a speech input instead?

- We often associate **complexity** in language with the **features** commonly found in **formal written texts**.

- Yet, **speech** is **not exempt from complexity**, but might be **reflected differently** in reason of:

1.2. Challenges in Speech Simplification (SpeechSimp)

What about applying simplification to a speech input instead?

- We often associate **complexity** in language with the **features** commonly found in **formal written texts**.

- Yet, **speech** is **not exempt from complexity**, but might be **reflected differently** in reason of:

  1. **Information structure**

     ❑ **Written –unspontaneous– texts** = Result, finalized version of a planned language production.

     ❑ **Spoken –spontaneous– texts** = On-the-fly process [Carter & McCarthy, 2017].

1.2. Challenges in Speech Simplification (SpeechSimp)

What about applying simplification to a speech input instead?

- We often associate **complexity** in language with the **features** commonly found in **formal written texts**.

- Yet, **speech** is **not exempt from complexity**, but might be **reflected differently** in reason of:

1. Information structure

❑ **Written –unspontaneous– texts** = Result, finalized version of a planned language production.

❑ **Spoken –spontaneous– texts** = On-the-fly process [Carter & McCarthy, 2017].

2. Spontaneity and grammaticality

❑ **Traces of real-time construction of speech:** revisions, false starts, repetitions and self-corrections.

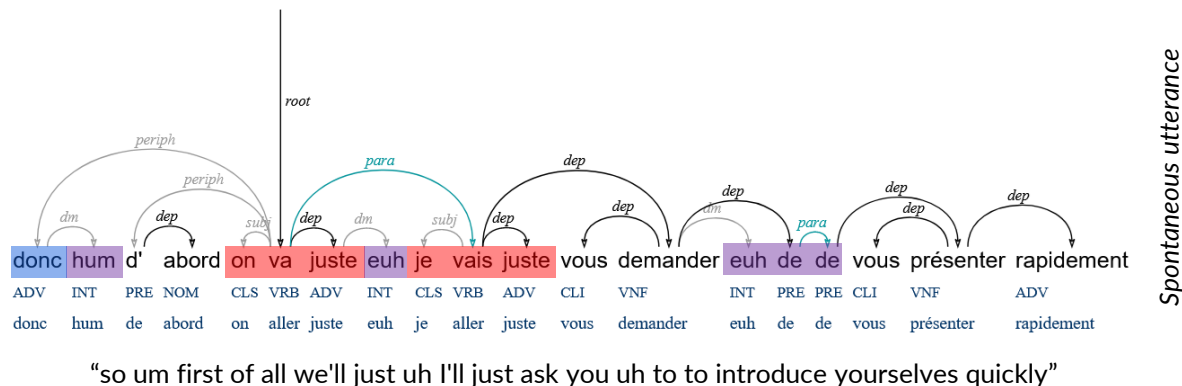❑ **Concatenations** of elements having a **paradigmatic relation** along the syntagmatic axis.

1.2. Challenges in Speech Simplification (SpeechSimp)

> **What about applying simplification to a speech input instead?**

- We often associate **complexity** in language with the **features** commonly found in **formal written texts**.

- Yet, **speech** is **not exempt from complexity**, but might be **reflected differently** in reason of:

1. Information structure   **2. Spontaneity and grammaticality**

"so um first of all we'll just uh I'll just ask you uh to to introduce yourselves quickly"
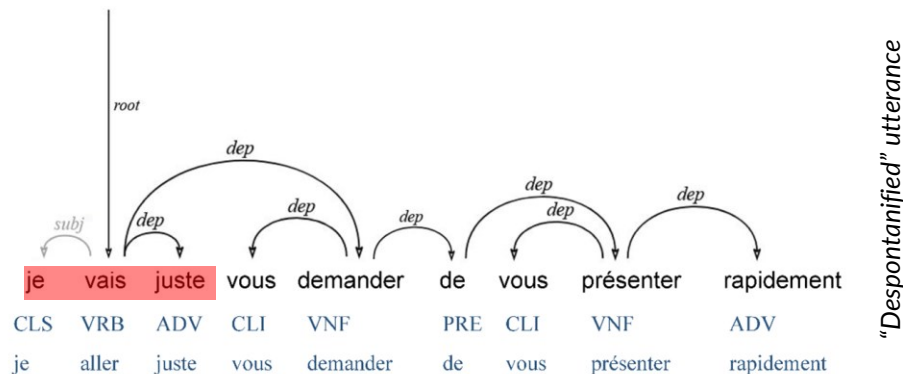
1.2. Challenges in Speech Simplification (SpeechSimp)

What about applying simplification to a speech input instead?

- We often associate **complexity** in language with the **features** commonly found in **formal written texts**.

- Yet, **speech** is **not exempt from complexity**, but might be **reflected differently** in reason of:

1. **Information structure**   2. **Spontaneity and grammaticality**

"I'll just ask you to introduce yourselves quickly"

1.3. Bridging the gap: characterizing SpeechSimp strategies

The idea of decomplexifying speech can be in turn be interesting from different perspectives:

1.3. Bridging the gap: characterizing SpeechSimp strategies

The idea of decomplexifying speech can be in turn be interesting from different perspectives:

Accessibility purposes	**Simplified transcripts** → Help clarify the input message for different target audiences.
Ancillary purposes	**Raw transcripts** → **Difficult to process by NLP pipelines.** **Simpler transcripts** → **May be helpful for other tasks**, *i.e.*, subtitle or speech-to-pictograph translation.

1.3. Bridging the gap: characterizing SpeechSimp strategies

The idea of decomplexifying speech can be in turn be interesting from different perspectives:

Accessibility purposes	**Simplified transcripts** → Help clarify the input message for different target audiences.
Ancillary purposes	**Raw transcripts** → **Difficult to process by NLP pipelines.** **Simpler transcripts** → **May be helpful for other tasks**, *i.e.*, subtitle or speech-to-pictograph translation.

- **Main caveat about SpeechSimp→ No guideline to steer the simplification process.**

1.3. Bridging the gap: characterizing SpeechSimp strategies

The idea of decomplexifying speech can be in turn be interesting from different perspectives:

Accessibility purposes	**Simplified transcripts** → Help clarify the input message for different target audiences.
Ancillary purposes	**Raw transcripts** → **Difficult to process by NLP pipelines.** **Simpler transcripts** → **May be helpful for other tasks**, *i.e.*, subtitle or speech-to-pictograph translation.

- **Main caveat about SpeechSimp→ No guideline 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**.

RESEARCH QUESTIONS

RQ1: What are the **edit operations** performed to obtain a **simplified version** of a **French spontaneous speech transcript**?

RQ2: How do **human simplification strategies align** with those adopted by **ChatGPT** and how **suitable** are they for a **non-native audience**?

2. Methodology

2.1. Source corpus and sampling

To analyze SpeechSimp strategies, we resort to...

2.1. Source corpus and sampling

To analyze SpeechSimp strategies, we resort to...

- **Orféo-CÉFC** [Benzitoun et al, 2016] → French spontaneous speech dataset:
  - Covering a **wide range** of **communicative situations**.
  - Various **degrees** of **spontaneity**.
  - Amounts to **~200k segments**.

2.1. Source corpus and sampling

To analyze SpeechSimp strategies, we resort to...

- **Orféo-CÉFC** [Benzitoun et al, 2016] → French spontaneous speech dataset:
  - Covering a **wide range** of **communicative situations**.
  - Various **degrees** of **spontaneity**.
  - Amounts to **~200k segments**.
- **Infeasible to manually simplify** → Resort to **sampling technique**:
  - **Proportionate random stratified sampling**.
  - **Population** → Orféo test set in [Pupier et al, 2022].
  - **Stratum** → Each Orféo subcorpus (12).

2.1. Source corpus and sampling

To analyze SpeechSimp strategies, we resort to...

- **Orféo-CÉFC** [Benzitoun et al, 2016] → French spontaneous speech dataset:
  - Covering a **wide range** of **communicative situations**.
  - Various **degrees** of **spontaneity**.
  - Amounts to **~200k segments**.

- **Infeasible to manually simplify** → Resort to **sampling technique**:
  - **Proportionate random stratified sampling**.
  - **Population** → Orféo test set in [Pupier et al, 2022].
  - **Stratum** → Each Orféo subcorpus (12).

- Resulting sample size → **100 utterances**.
  - **Reasonable workload** for respondents.
  - Not compromise the task's **stability** and **consistency**.

	Orféo-Test		
subcorpus	# utterances	%	# samples
Cfpb	362	1.69	2
Cfpp	3,232	15.06	15
Clapi	967	4.51	5
Coralrom	1,376	6.41	6
Crfp	2,259	10.53	10
Fleuron	217	1.01	1
Oral-Narr.	1,050	4.89	5
Ofrom	1,476	6.88	7
Reunions	1,245	5.80	6
Tcof	1,997	9.31	9
Tufs	4,525	21.09	21
Valibel	2,753	12.83	13
Total	**21,459**	**100**	**100**

2.2. Expert-based SpeechSimp: survey design

Due to the **absence** of **speech simplification guidelines**:

- Intuitive approach leveraging **human experts'** criteria.

2.2. Expert-based SpeechSimp: survey design

Due to the **absence** of **speech simplification guidelines**:

- Intuitive approach leveraging **human experts'** criteria.

- Linguists' **profile**:
  - European **French-native** speakers.
  - **Solid background** on linguistics.
  - **Current dedication** to the latter.

- –

2.2. Expert-based SpeechSimp: survey design

Due to the **absence** of **speech simplification guidelines**:

- Intuitive approach leveraging **human experts'** criteria.

- Linguists' **profile**:
  - European **French-native** speakers.
  - **Solid background** on linguistics.
  - **Current dedication** to the latter.

- - - - - - - - - - - - - - - - - - - -

- Setting a **manual simplification task**: **LimeSurvey platform**.

- **Survey** including **two questions** to respondents:

  1. Provide a **candidate simplified version** of the utterance.
  2. Explain the **chain-of-thought** followed to make such transformations.

Simplification du français parlé spontané

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. Notez que les mots avec le symbole « ~ » sont des amorces (termes inachevés).

2. **Expliquer votre raisonnement.** Après chaque simplification, énumérez et expliquez les transformations que vous avez effectuées. Par exemple, le remplacement de mots complexes par des synonymes plus simples ou encore la restructuration de la phrase.

*Voici la phrase à simplifier :
"ouais c'est ça sauf que moi on m'a jamais expliqué le rythme du coup"

Simplification :

Raisonnement :

2.3. Machine-based SpeechSimp: ChatGPT prompting

Human-crafted simplifications agree or differ against machine-generated ones?

- **Experts annotations** are **expensive** to produce.
- We sought to compare **human- vs. machine-based simplified outputs**.

2.3. Machine-based SpeechSimp: ChatGPT prompting

Human-crafted simplifications agree or differ against machine-generated ones?

- **Experts annotations** are **expensive** to produce.
- We sought to compare **human- vs. machine-based simplified outputs**.

To this respect:

- **ChatGPT** has been utilized for **text-annotation tasks** [Gilardi et al, 2023].
- We **prompted** one of the latest models (`gpt-4-0125-preview`) to **collect artificially-generated simplified transcripts**.

2.3. Machine-based SpeechSimp: ChatGPT prompting

Human-crafted simplifications agree or differ against machine-generated ones?

- **Experts annotations** are **expensive** to produce.
- We sought to compare **human- vs. machine-based simplified outputs**.

To this respect:

- **ChatGPT** has been utilized for **text-annotation tasks** [Gilardi et al, 2023].
- We **prompted** one of the latest models (`gpt-4-0125-preview`) to collect **artificially-generated simplified transcripts**.
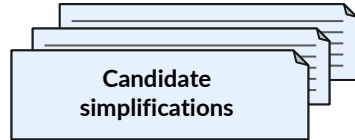
  - ○ `temperature=0` → Ensuring **consistency**.
  - ○ **One-by-one** basis → Avoiding history **influence**.
  - ○ **Identical prompt** than one used with **humans**.

```
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
)
```
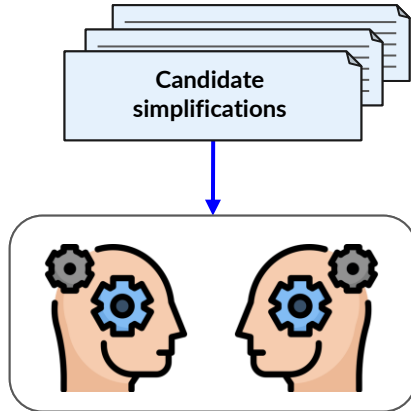
3. Results

3.1. Quantitative evaluation

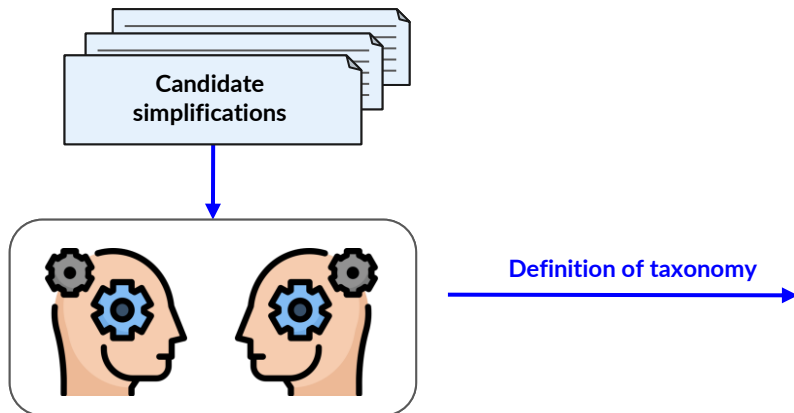**Once the collection of simplifications was completed:**

Candidate simplifications

3.1. Quantitative evaluation

Once the collection of simplifications was completed:

3.1. Quantitative evaluation

Once the collection of simplifications was completed:

Candidate simplifications
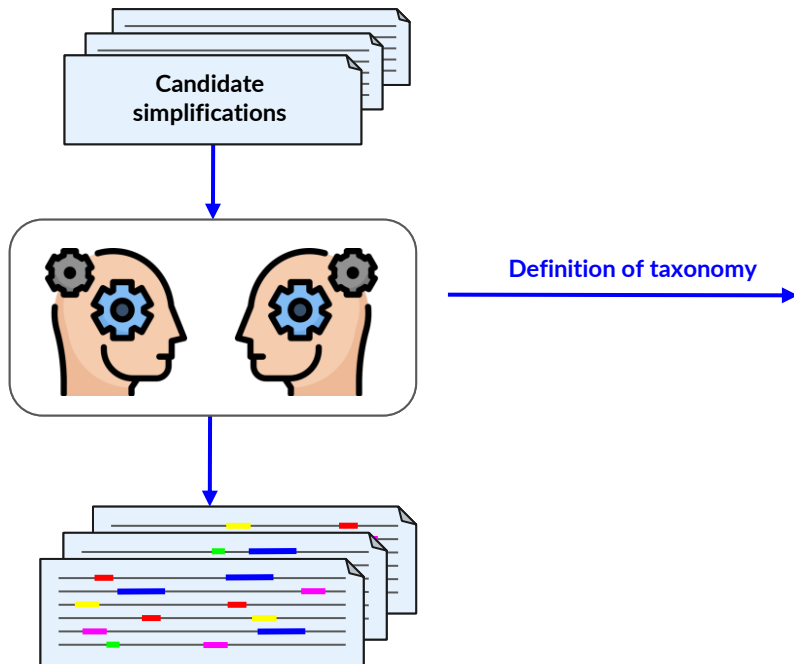
Definition of taxonomy

Edit	Level	ID	Linguistic unit(s) affected, operation
Deletion		1	Repetitions
		2	Affirmation and negation words
		3	Interjections
		4	Conjunctions
		5	Discourse markers
		6	Restarts and reformulations
		7	Adverbs and adjectives
		8	Incomplete words
		9	Statement verbs
		10	Pronouns
		11	Verbs with little semantic value
Replacement	Lexical	12	Simpler synonyms for content words
		13	Compression of nominal phrases
		14	More standard equivalents for content words
		15	Smoothing of swear words
	Verbal morphology	16	Intransitive to transitive verbs
		17	Pronominal to non-pronominal verbs
		18	Change of verbal tense
		19	Compression of verbal locutions
	Syntactic	20	Passive to active voice
		21	Cleft to canonical constructions
		22	Neutralization of dislocated subjects
		23	Pronoun transformations
Restructuration		24	Reorder
		25	Sentence splitting
		26	Sentence merging
Addition		27	Explicitation or disambiguation of a word
		28	Completion of truncated sentences
		29	Clarification of uncommon terms
Copy		30	Input sentence is left unchanged

3.1. Quantitative evaluation
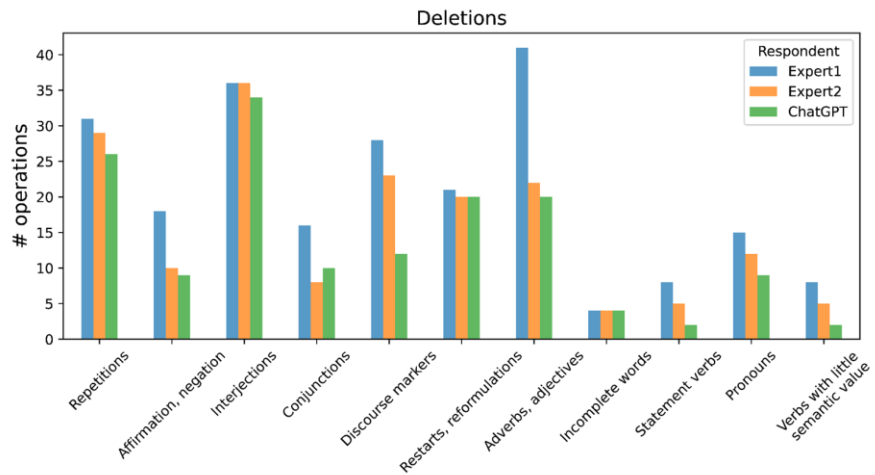
Once the collection of simplifications was completed:

Edit	Level	ID	Linguistic unit(s) affected, operation
Deletion		1	Repetitions
		2	Affirmation and negation words
		3	Interjections
		4	Conjunctions
		5	Discourse markers
		6	Restarts and reformulations
		7	Adverbs and adjectives
		8	Incomplete words
		9	Statement verbs
		10	Pronouns
		11	Verbs with little semantic value
Replacement	Lexical	12	Simpler synonyms for content words
		13	Compression of nominal phrases
		14	More standard equivalents for content words
		15	Smoothing of swear words
	Verbal morphology	16	Intransitive to transitive verbs
		17	Pronominal to non-pronominal verbs
		18	Change of verbal tense
		19	Compression of verbal locutions
	Syntactic	20	Passive to active voice
		21	Cleft to canonical constructions
		22	Neutralization of dislocated subjects
		23	Pronoun transformations
Restructuration		24	Reorder
		25	Sentence splitting
		26	Sentence merging
Addition		27	Explicitation or disambiguation of a word
		28	Completion of truncated sentences
		29	Clarification of uncommon terms
Copy		30	Input sentence is left unchanged

3.1. Quantitative evaluation

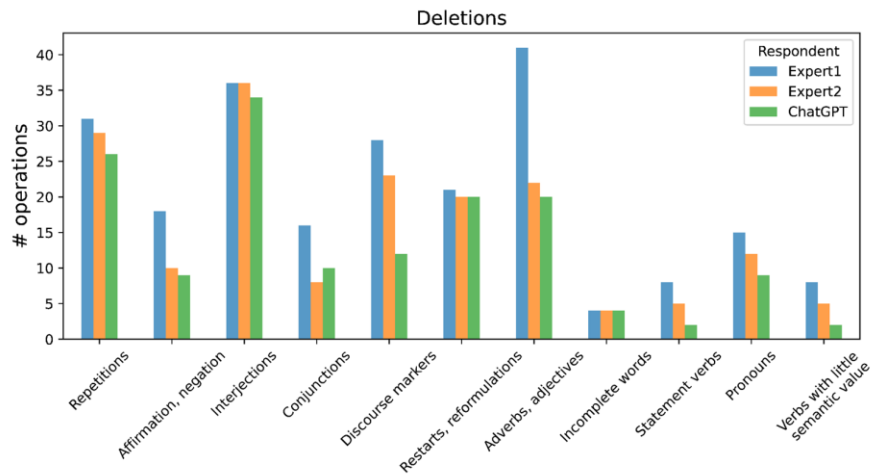**Once the collection of simplifications was completed:**

3.1. Quantitative evaluation: special focus on deletions

Deletions

- By far, the **most common operation** happening in SpeechSimp.
- Especially affecting:

3.1. Quantitative evaluation: special focus on deletions

Deletions

- By far, the **most common operation** happening in SpeechSimp.
- Especially affecting:
  - **Redundant elements (from a semantic perspective)**: restarts, repetitions.

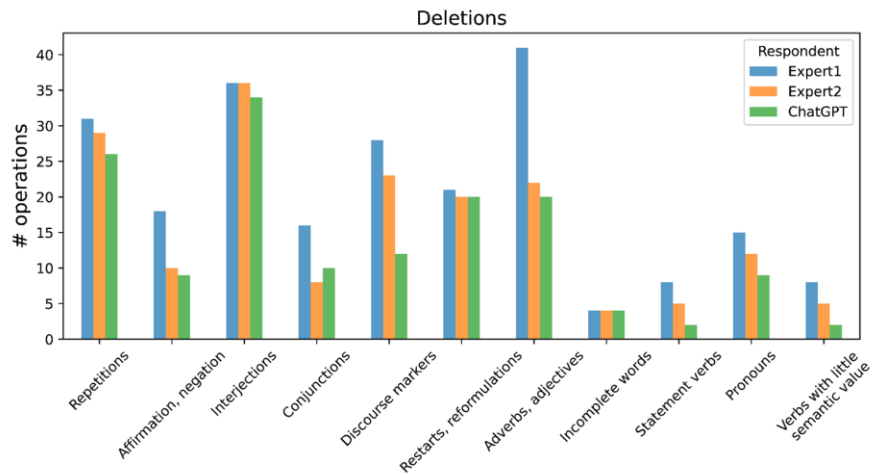| Input	ouais c'est ça sauf que **moi** on m'a jamais expliqué le rythme du coup
Expert 1	*On ne m'a pas expliqué le rythme*
Expert 2	*Oui, c'est ça, sauf qu'on ne m'a jamais expliqué le rythme*
ChatGPT	*Oui, c'est vrai, mais personne ne m'a jamais expliqué le rythme*

3.1. Quantitative evaluation: special focus on deletions

Deletions
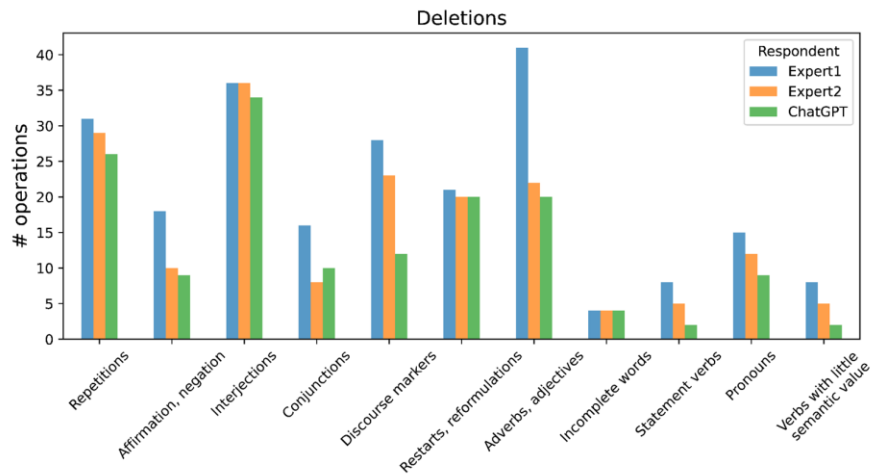
- By far, the **most common operation** happening in SpeechSimp.
- Especially affecting:
  - **Redundant elements (from a semantic perspective)**: restarts, repetitions.
  - Elements **related to enunciation**: yes/no adverbs (*non, voilà, ouais*), statement verbs (*tu sais, je tiens à dire*).

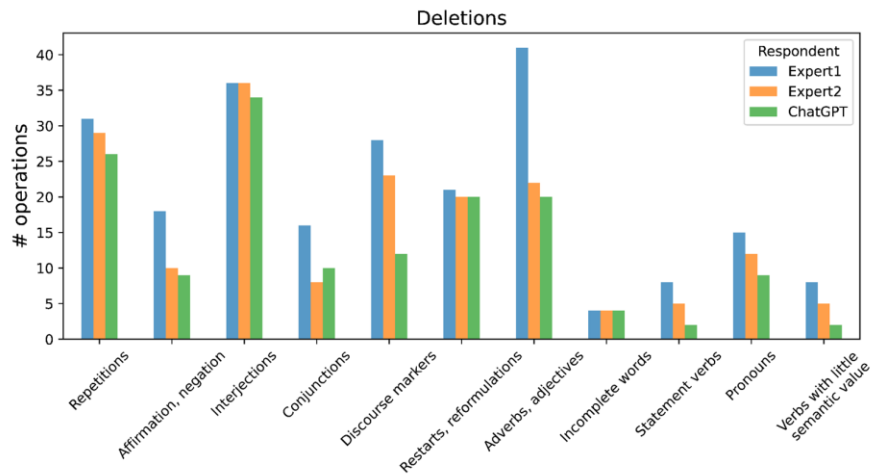| Input	**ouais** c'est ça sauf que moi on m'a jamais expliqué le rythme **du coup**
Expert 1	*On ne m'a pas expliqué le rythme*
Expert 2	*Oui, c'est ça, sauf qu'on ne m'a jamais expliqué le rythme*
ChatGPT	*Oui, c'est vrai, mais personne ne m'a jamais expliqué le rythme*

3.1. Quantitative evaluation: special focus on deletions

Deletions

- By far, the **most common operation** happening in SpeechSimp.
- Especially affecting:
  - **Redundant elements (from a semantic perspective)**: restarts, repetitions.
  - Elements **related to enunciation**: yes/no adverbs (*non, voilà, ouais*), statement verbs (*tu sais, je tiens à dire*).
  - Elements providing **little information**: adjectives, adverbs (*toutes* nos traditions → *nos traditions*).

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

3.1. Quantitative evaluation: special focus on deletions

Deletions

- By far, the **most common operation** happening in SpeechSimp.
- Especially affecting:
  - **Redundant elements (from a semantic perspective)**: restarts, repetitions.
  - Elements **related to enunciation**: yes/no adverbs (*non, voilà, ouais*), statement verbs (*tu sais, je tiens à dire*).
  - Elements providing **little information**: adjectives, adverbs (*toutes nos traditions → nos traditions*).

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

RESULT

Simplified utterances seem to be "**writified**" or **register-standardized versions** of the **inputs** that just include their **propositional content**.

3.1. Quantitative evaluation: special focus on replacements

Category subdivided according to the **linguistic transformations** observed:

3.1. Quantitative evaluation: special focus on replacements

Category subdivided according to the **linguistic transformations** observed:

- **Lexical** (more frequent):
 - ○ Propensity to find **simpler** equivalents for **content words**.

Replacements
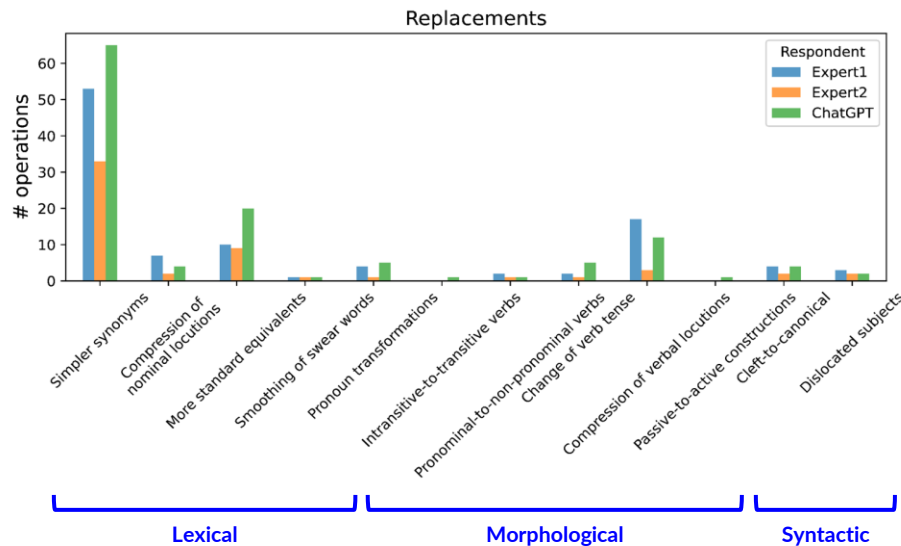
Input	mais enfin cette **confrérie** elle avait quand même un un avantage c' est que elle venait en aide euh aux malheureux
Expert 1	*Cette association avait un avantage, elle a aidé les démunis*
Expert 2	*Cette association avait l'avantage de venir en aide aux malheureux*
ChatGPT	*Mais cette confrérie aidait quand même les malheureux*

3.1. Quantitative evaluation: special focus on replacements

Category subdivided according to the **linguistic transformations** observed:

- **Lexical** (more frequent):
  - Propensity to find **simpler** equivalents for **content words**.
  - But also: **smoothing** of **slang**, **colloquialisms** and **profanity** (*monde* → *personnes*, *bouquins* → *livres*).
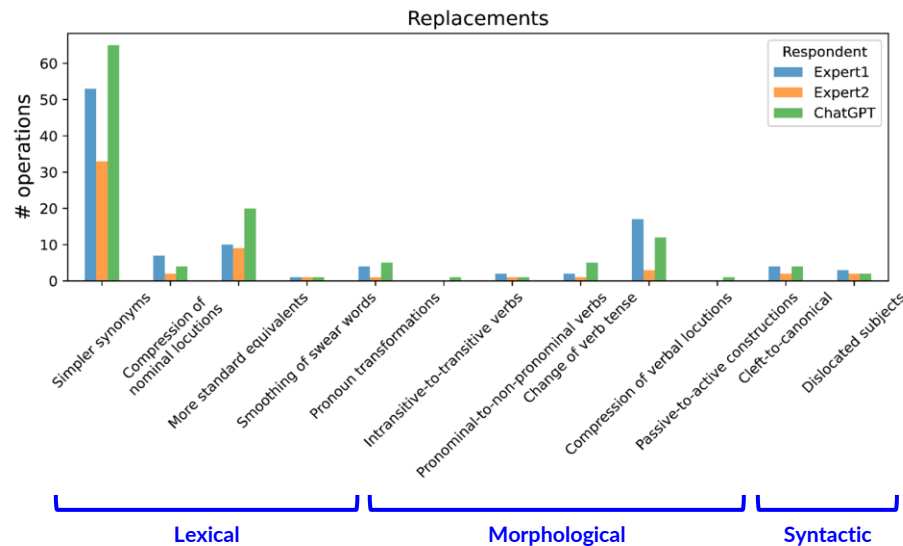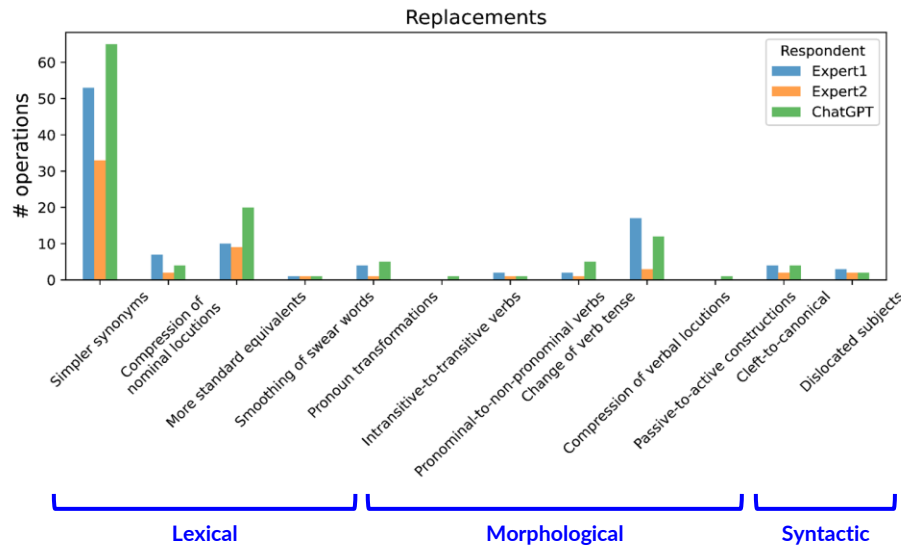

Input	et c' était pas son style de **bouffer** de la **merde**
Expert 1	c'était pas à son goût de manger mal
Expert 2	Et il/elle n'est pas habitué à manger des aliments de mauvaise qualité
ChatGPT	Il n'aimait pas manger de mauvaises choses

3.1. Quantitative evaluation: special focus on replacements

Category subdivided according to the **linguistic transformations** observed:

- **Lexical** (more frequent):
  - Propensity to find **simpler** equivalents for **content words**.
  - But also: **smoothing** of **slang**, **colloquialisms** and **profanity** (*monde → personnes, bouquins → livres*).

- **Morphological** (less frequent):
  - **Compression** of **constituents**: nominal (*monde du travail →* *travail*) and **verbal** (*faire la demande → demander*) groups.

3.1. Quantitative evaluation: special focus on replacements

Category subdivided according to the **linguistic transformations** observed:

- **Lexical** (more frequent):
  - Propensity to find **simpler** equivalents for **content words**.
  - But also: **smoothing** of **slang**, **colloquialisms** and **profanity** (*monde → personnes*, *bouquins → livres*).

- **Morphological** (less frequent):
  - **Compression** of **constituents**: **nominal** (*monde du travail → travail*) and **verbal** (*faire la demande → demander*) groups.

- **Syntactic** (less frequent):
  - **Standardization** of **marked** information structures (cleft clauses and dislocated subjects).
  - **Passive voice** changes: **anecdotal**.

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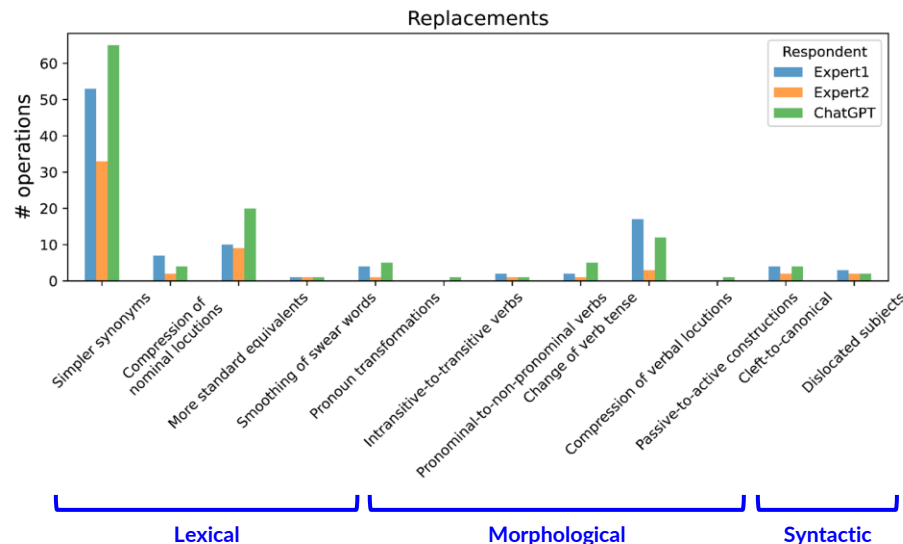
Input	**on sent que** la prise de conscience de ce genre de choses **elle s' est faite** tard
Expert 1	Nous pensons que la compréhension de ce problème est arrivée tard
Expert 2	La prise de conscience de ces choses-là est arrivée tard
ChatGPT	Les gens ont commencé à comprendre ces choses tard

3.2. Qualitative intrinsic evaluation

To assess the **suitability** of the **produced** human- and machine-based **outputs** for a **foreign-speaking audience**:

- We conducted a **qualitative intrinsic evaluation**.
- **Respondents**: a group of **3 non-native master students** (at University of Geneva).

3.2. Qualitative intrinsic evaluation

To assess the **suitability** of the **produced** human- and machine-based **outputs** for a **foreign-speaking audience**:

- We conducted a **qualitative intrinsic evaluation**.
- **Respondents**: a group of **3 non-native master students** (at University of Geneva).

Survey design:

- Also set up on the **LimeSurvey** platform.
- We asked to **score** the produced **simplifications** on a **5-point Likert scale**.
- Under two typical **ATS dimensions** [Yamaguchi et al, 2023]:

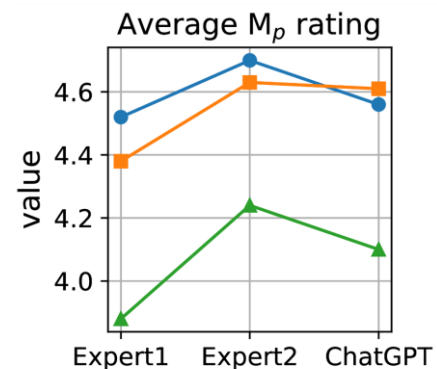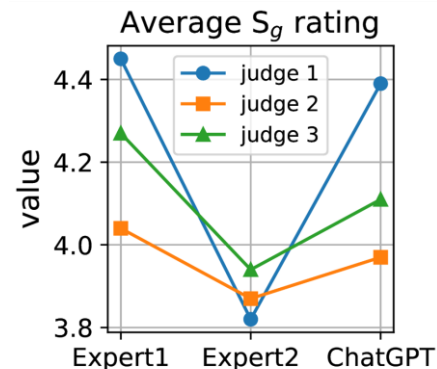| Simplicity gain (S_g)	Meaning preservation (M_p)
5 – Much simpler	5 – Fully preserved
4 – Somewhat simpler	4 – Mostly preserved
3 – Same difficulty	3 – Partially preserved
2 – More difficult	2 – Completely different
1 – Unintelligible	1 – Unintelligible

3.2. Qualitative intrinsic evaluation

To assess the **suitability** of the **produced** human- and machine-based **outputs** for a **foreign-speaking audience**:

- We conducted a **qualitative intrinsic evaluation**.
- **Respondents**: a group of **3 non-native master students** (at University of Geneva).

Survey design:

- Also set up on the **LimeSurvey** platform.
- We asked to **score** the produced **simplifications** on a **5-point Likert scale**.
- Under two typical **ATS dimensions** [Yamaguchi et al, 2023]:

Simplicity gain (S_g)	Meaning preservation (M_p)
5 – Much simpler	5 – Fully preserved
4 – Somewhat simpler	4 – Mostly preserved
3 – Same difficulty	3 – Partially preserved
2 – More difficult	2 – Completely different
1 – Unintelligible	1 – Unintelligible

Average S_g rating

Average M_p rating

4. Conclusions and further work

- **Taxonomy** and **quantification** of simplification operations applied to French spontaneous transcripts.
- **Intuitive-based approach**, due to the lack of guidelines to steer the simplification process.

- **Taxonomy** and **quantification** of simplification operations applied to French spontaneous transcripts.

- **Intuitive-based approach**, due to the lack of guidelines to steer the simplification process.

- **Speech decomplexification** ≅ Speech "**despontanefication**".

- **Taxonomy** and **quantification** of simplification operations applied to French spontaneous transcripts.

- **Intuitive-based approach**, due to the lack of guidelines to steer the simplification process.

- **Speech decomplexification** ≅ Speech "**despontanefication**".

- Provision of a dataset: **Propicto-Orféo-Simple*** → mapping Orféo-based audio-transcripts with simplified versions.

---

* https://www.ortolang.fr/market/corpora/propicto

- **Taxonomy** and **quantification** of simplification operations applied to French spontaneous transcripts.
- **Intuitive-based approach**, due to the lack of guidelines to steer the simplification process.
- **Speech decomplexification** ≅ Speech "**despontanefication**".
- Provision of a dataset: **Propicto-Orféo-Simple*** → mapping Orféo-based audio-transcripts with simplified versions.

LIMITATIONS

- **Small scale study** → Due to the **costly process** required to produce such data.
- **Lack of context** as an obstacle for the completion of the task **→ BUT motivated by representativeness purposes.**

---

* https://www.ortolang.fr/market/corpora/propicto

CONTRIBUTIONS

- **Taxonomy** and **quantification** of simplification operations applied to French spontaneous transcripts.

- **Intuitive-based approach**, due to the lack of guidelines to steer the simplification process.

- **Speech decomplexification** ≅ Speech "**despontanefication**".

- Provision of a dataset: **Propicto-Orféo-Simple*** → mapping Orféo-based audio-transcripts with simplified versions.

LIMITATIONS

- **Small scale study** → Due to the **costly process** required to produce such data.

- **Lack of context** as an obstacle for the completion of the task **→ BUT motivated by representativeness purposes.**

FURTHER WORK

- **Propicto-Orféo-Simple** → Further use as an **evaluation set** to assess the performance of **speech simplification models**.

---

* https://www.ortolang.fr/market/corpora/propicto

Thank you for your attention!

Lucía Ormaechea

Ph.D. Candidate

Lucia.OrmaecheaGrijalba@unige.ch

https://luciaormaechea.com/

References (1/3)

C. Horn, C. Manduca and D. Kauchak (2014)

Learning a Lexical Simplifier using Wikipedia

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 458-463.

S. Stajner (2021)

Automatic Text Simplification for Social Good: Progress and Challenges

Findings of the Association for Computational Linguistics, 2637-2652.

X. Zhang and M. Lapata (2017)

Sentence Simplification with Deep Reinforcement Learning

Proceedings of the Conference on Empirical Methods in Natural Language Processing, 584-594.

L. Ormaechea and 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.

References (2/3)

J. Li, C. Lester, X. Zhao, Y. Ding, Y. Jiang, V.G.V. Vydiswaran (2020)

PharmMT: A Neural Machine Translation Approach to Simplify Prescription Directions

Findings of the Association for Computational Linguistics: EMNLP, 2785-2796.

W. Xu, C. Callison-Burch, C. Napoles (2015)

Problems in Current Text Simplification Research: New Data Can Help

Transactions of the Association for Computational Linguistics, 283-297.

T. Goldsack, Z. Zhang, C. Lin, C. Scarton (2022)

Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature

Proceedings of the Conference on Empirical Methods in Natural Language Processing, 10589-10604.

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.

R. Carter and M. McCarthy (2017)

Spoken Grammar: Where Are We and Where Are We Going?

Applied Linguistics, Volume 38, Issue 1, 1-20.

References (3/3)

A. Pupier, M. Coavoux, B. Lecouteux and J. Goulian (2022)

End-to-End Dependency Parsing of Spoken French

Proceedings of Interspeech 2022, 1816-1820.

C. Benzitoun, J.M. Debaisieux and H.J. Deulofeu (2016)

Le projet ORFÉO : un corpus d'étude pour le français contemporain

Corpus, 15, 91-114.

F. Gilardi, M. Alizadeh and M. Kubli (2023)

ChatGPT outperforms crowd workers for text-annotation tasks

Proceedings of the National Academy of Sciences, 120 (30).

D. Yamaguchi, R. Miyata, S. Shimada and S. Sato (2023)

Gauging the Gap Between Human and Machine Text Simplification Through Analytical Evaluation of Simplification Strategies and Errors

Findings of the Association for Computational Linguistics: EACL 2023, 359-375.