Replace, Paraphrase or Fine-tune? Evaluating Automatic Simplification for Medical Texts in Spanish

Leonardo Campillos Llanos,[1] Ana R. Terroba,[2]
Rocío Bartolomé,[3] Ana Valverde,[4] Cristina González,[4]
Adrián Capllonch[5] and Jónathan Heras[6]

[1] Instituto de Lengua, Literatura y Antropología, CSIC
[2] Fundación Rioja Salud
[3] Facultad de Filosofía y Letras, Universidad Autónoma de Madrid
[4] UTM – Spanish Royal Academy of Medicine (RANME)
[5] IPS Marañón & HGU Gregorio Marañón
[6] Facultad de Ciencia y Tecnología, Universidad de La Rioja

LREC-COLING 2024

Team

Leonardo Campillos

ILLA CSIC

Cristina V. González

UTM RANME

Ana Valverde

UTM RANME

Adrián Capllonch

IPS Marañón & HGU Gregorio Marañón

Ana Rosa Terroba

Fundación Rioja Salud

Rocío Bartolomé

Universidad Autónoma de Madrid

Jónathan Heras

Universidad de La Rioja

Funded by AEI/MICIN/10.13039/501100011033 in project call RETOS 2020 (PID2020-116001RA-C33) and partly by project PID2020-115225RB- I00 (JH)

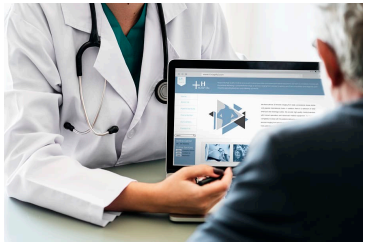
Introduction and Background

Introduction

- Medical records, medication leaflets or clinical trial announcements are written with **jargon** or **abbreviations/acronyms** that represent a communication barrier.

> Clinical trial to establish the effects of low-dose rtPA and the effects of intensive blood pressure lowering in patients with acute cerebrovascular accident

- Medical professionals need to **explain** details about conditions or procedures, but they often **lack the time** to give clear explanations.



→ **Automatic text simplification (ATS)** methods can alleviate the language gap and be a complement, provided that the content is transmitted rigurously.

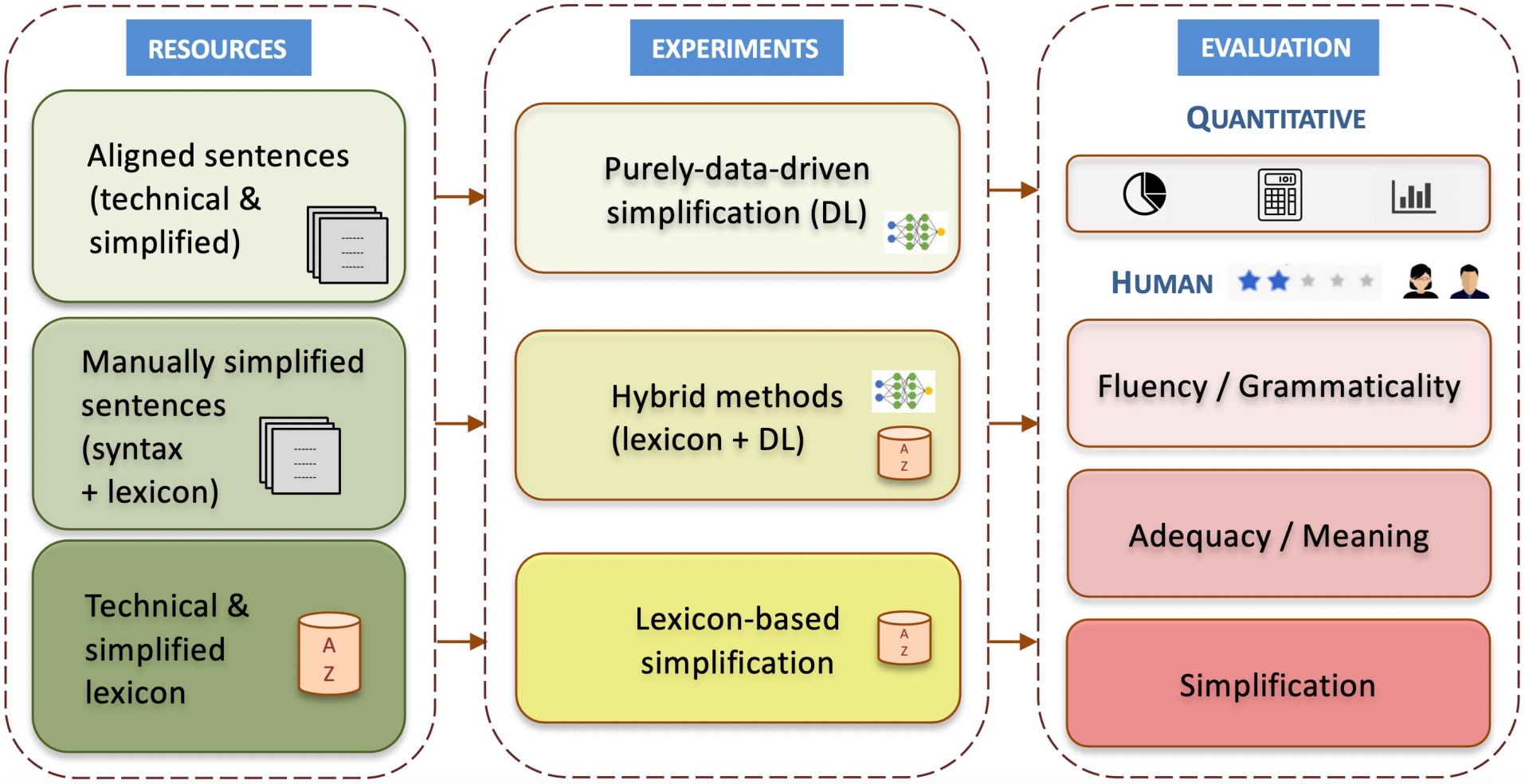
Background

- ATS methods rely on **lexicons, rules, machine learning/deep learning** or **prompt-based approaches** (Saggion, 2017; Shardlow & Nawaz, 2019; Wang et al., 2022)
- **Hybrid methods** (Bot et al. 2012; Cardon & Grabar 2020; Todirascu et al., 2022).
- Simplification applied at **all linguistic levels**: lexis, grammar, discourse...
- Resources and corpora are needed but scarce:
  - English: Newsela (Xu et al. 2015); medical corpora (Van den Bercken et al., 2019; Sakakini et al., 2020; Devaraj et al., 2021; Guo et al. 2022; Basu et al. 2023).
  - Spanish: EASIER (Alarcón et al. 2023), Saggion et al. (2011).
  - French (Grabar & Cardon, 2018; Gala et al. 2020).
  - German (Seiffe et al. 2020; Trienes et al. 2022).

Contributions: 1) First medical lexicon for simplification in Spanish
2) Experiments and human evaluation.
3) Trained models released on the Hugging Face hub

Methods

Methods


RESOURCES
Aligned sentences (technical & simplified)
Manually simplified sentences (syntax + lexicon)
Technical & simplified lexicon
EXPERIMENTS
Purely-data-driven simplification (DL)
Hybrid methods (lexicon + DL)
Lexicon-based simplification
EVALUATION
QUANTITATIVE
HUMAN
Fluency / Grammaticality
Adequacy / Meaning
Simplification

Experimental data

5000 parallel (technical/simplified) sentences:

- **3800** (149 862 tokens):
  - Extracted from the CLARA-MeD corpus (Campillos-Llanos et al. 2022).
  - Aligned with Sentence Transformers (Reimers & Gurevych 2019) and revised.
  - Data sources: Medication leaflets/summaries of product characteristics, cancer-related information (NCI), and clinical trial announcements (EudraCT).

- **1200** (144 019 tokens):
  - Extracted from EudraCT and manually simplified (Campillos-Llanos et al. 2024)
  - Created simplification guidelines.
  - Used version with both syntactic and lexical simplification.

Experimental data

- Samples

Original	*Se considera mujer en edad fértil como aquellas mujeres que no hayan sido sometidas a procedimientos de infertilidad permanente o que sean amenorreicas desde hace menos de 12 meses.* (2019-004871-38) 'Women of childbearing age are considered to be those women who have not undergone permanent infertility procedures or who have been amenorrheic for less than 12 months'
Syntactic simplification	*Se consideran mujeres en edad fértil aquellas mujeres que no han sido sometidas a procedimientos de infertilidad permanente.* ***También*** *se consideran en edad fértil a quienes sean* ***amenorreicas*** *desde hace menos de 12 meses.* 'Women of childbearing age are considered to be those women who have not undergone permanent infertility procedures. Women are **also** considered of childbearing age if they have been **amenorrheic** for less than 12 months'
Lexical and syntactic simplification	*Se consideran mujeres en edad fértil aquellas que no han sido sometidas a procedimientos de infertilidad permanente.* ***También*** *se consideran en edad fértil a quienes* ***no tienen menstruación*** *desde hace menos de 12 meses*. 'Women of childbearing age are considered those who have not undergone permanent infertility procedures. Women are **also** considered of childbearing age if they have **not been menstruating** for less than 12 months.'

Experimental data

- Samples

Original	*Se considera mujer en edad fértil como aquellas mujeres que no hayan sido sometidas a procedimientos de infertilidad permanente o que sean amenorreicas desde hace menos de 12 meses.* (2019-004871-38) 'Women of childbearing age are considered to be those women who have not undergone permanent infertility procedures or who have been amenorrheic for less than 12 months'
Syntactic simplification	*Se consideran mujeres en edad fértil aquellas mujeres que no han sido sometidas a procedimientos de infertilidad permanente.* ***También*** *se consideran en edad fértil a quienes sean* ***amenorreicas*** *desde hace menos de 12 meses.* 'Women of childbearing age are considered to be those women who have not undergone permanent infertility procedures. Women are **also** considered of childbearing age if they have been **amenorrheic** for less than 12 months'
Lexical and syntactic simplification	*Se consideran mujeres en edad fértil aquellas que no han sido sometidas a procedimientos de infertilidad permanente.* ***También*** *se consideran en edad fértil a quienes* ***no tienen menstruación*** *desde hace menos de 12 meses*. 'Women of childbearing age are considered those who have not undergone permanent infertility procedures. Women are **also** considered of childbearing age if they have **not been menstruating** for less than 12 months.'

SimpMedLexSp

- First patient-oriented Spanish lexicon of medical terms and simplified forms

```
CUI|TECHNICAL_TERM|PATIENT_TERM
C0004057|ácido acetilsalicílico|aspirina
C0006736|cálculo urinario|piedra en la orina
C0007137|ca epidermoide|carcinoma epidermoide
C0007137|cec|carcinoma escamocelular
C0040423|amigdalectomía|operación de anginas
C0231221|asintomático|sin síntomas
C0277797|apirético|sin fiebre
C0522523|percutáneo|a través de la piel
```

```
p.a.|presión arterial
peg|polietilenglicol
per os|por la boca
percutánea|a través de la piel
percutáneas|a través de la piel
percutáneo|a través de la piel
percutáneos|a través de la piel
peribucales|alrededor de la boca
peribucal|alrededor de la boca
```

- Subset mapped to Unified Medical Language System Concept Unique Identifiers
- 12 605 term pairs in the experiments (but >14000 variant forms to date)
- Available at: https://digital.csic.es/handle/10261/349662

SimpMedLexSp

- Data sources:
  - Terms with colloquial equivalents from the Dictionary of Medical Terms (RANME)
  - Eugloss glossary (https://users.ugent.be/~rvdstich/eugloss/welcome.html)
  - Abbreviations and acronyms from the MedLexSp lexicon
  - Paraphrase patterns and definitions extracted from the CLARA-MeD corpus

    Ej. *amenorrea (es decir, problemas con la regla)*

    'amenorrhea (that is to say, problems with menstruation)'

- Frequency-based filter (frequency > 100) to discard widely-used medical terms (Leroy and Endicott, 2012; Chen et al., 2018)

Lexicon-based simplification

- The lexicon was split into two files to conduct:
  - **Lexical substitution** → simpler synonyms

    `cephalea → headache`

    Technical: cephalea might occur

  - **Paraphrase, explanation or definition** → preserve term and append paraphrase

    `clomipramine → medical drug to treat depression`

    Technical: patients that were prescribed clomipramine

Lexicon-based simplification

- The lexicon was split in two files to conduct:
  - **Lexical substitution** → simpler synonyms

    `cephalea → headache`

    Simplified: headache might occur

  - **Paraphrase, explanation or definition** → preserve term and append paraphrase

    `clomipramine → medical drug to treat depression`

    Simplified: patients taking clomipramine (medical drug to treat depression)

Deep learning-based methods

- **Fine-tuned models:**
  - multilingual BART (mBART; Liu et al. 2020)
  - multilingual T5 (mT5; Xue et al. 2021)
  - NASES (BART model trained on news summarization dataset; Ahuir et al. 2021)
  - MariMari (RoBERTa model trained on the MLSUM dataset; Fandiño et al. 2022)
  - Pegasus XSUM (Zhang et al., 2020)
  - OpenNMT (Klein et al. 2017; results below the expected, finally not applied)
- **Prompt-based learning:**
  - BERTIN Alpaca (zero-shot prompting) (BERTIN-project 2023)*
  - BERTIN Alpaca (few-shot prompting)
  - BERTIN Alpaca fine-tuned with training set + lexicon (zero-shot prompting)

*https://huggingface.co/bertin-project/bertin-gpt-j-6B-alpaca

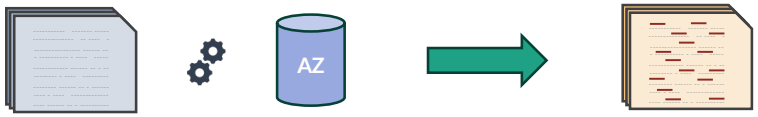
Deep learning-based methods

Zero-shot prompting
Below is an instruction that describes a task. Write a response that appropriately completes the request. ### Instruction: Write as easy-to-read text the following text: {asymptomatic} ### Response:
Few-shot prompting
Below there are some examples to simplify medical text. Write a response following the examples: ### Complex text: {asymptomatic} ### Simple text: {no symptoms} ### Complex text: {apyretic} ### Simple text:

Table 10: Samples of prompts used to train the BERTIN model

Experiments

- **Method A**: only use the lexicon (lexical substitution or append paraphrases)

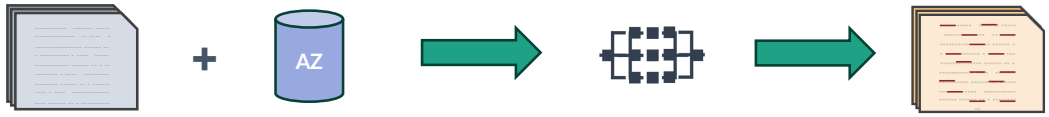


- **Method B**: only zero-shot or few-shot prompting (BERTIN Alpaca)

- **Method C**: fine-tuning without the lexicon

Experiments

- **Method D**: use the lexicon and sentences for training



- **Method E**: use the lexicon and sentences for training and append paraphrases

- **Method F**: use the lexicon for lexical substitution of terms in training data

- **Method G**: lexicon for lexical substitution in training data and for paraphrasing in predictions

Distribution of trained models

https://huggingface.co/CLARA-MeD

CLARA-MeD Community

https://clara-nlp.uned.es/home/med/ lcampillos · Upgrade to Enterprise

+ New | Activity Feed | Organization settings | Watch repos

AI & ML interests

Natural language processing for automatic text simplification of medical texts

Team members 2

Organization Card Edit org card

The CLARA-MeD project applies automatic natural language processing methods to enhance the accessibility of health information.

One of the approaches is term simplification; i.e. substituting a difficult-to-read word (e.g. *amigdalectomía*) with an easier or more explicative paraphrase (e.g. *operación de anginas*).

The CLARA-MeD repository includes the following work:

- A parallel corpus of technical-patient aligned sentences, extracted from the CLARA-MeD comparable corpus of technical and laymen documents.
- Models used in experiments to compare simplification approaches based on state-of-the-art neural-network algorithms.

Models 8 ↑↓ Sort: Recently updated

CLARA-MeD/flan-t5-large
Text2Text Generation • Updated 27 days ago • ↓ 10 • ♡ 1

CLARA-MeD/bertin-gpt
Updated Oct 11, 2023

CLARA-MeD/mbart-large-50
Text2Text Generation • Updated Mar 28, 2023 • ↓ 3

CLARA-MeD/pegasus-xsum
Text2Text Generation • Updated Feb 5, 2023 • ↓ 3

CLARA-MeD/flan-t5-base
Text2Text Generation • Updated Jan 31, 2023 • ↓ 5

CLARA-MeD/mt5-simplification-spanish
Text2Text Generation • Updated Jan 29, 2023 • ↓ 73 • ♡ 1

Evaluation

- 5-fold cross-validation:
  - From the 3800 subset: 80% (3040 sentences) train / 20% (760) test
  - From the 1200 subset: 80% (960 sentences) train / 20% (240) test
- Quantitative evaluation:
  - BLEU (Papineni et al., 2002)
  - ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004)
  - SARI (Xu et al. 2016)

Human evaluation

- **11 evaluators** (documentalists and linguists)
- **500 sentences** evaluated in two rounds:
  - **250** sentences from the 3800 subset (50 random sentences x 5 simplifications)
  - **250** from the 1200 subset (idem)
- Only evaluated models without high rate of severe hallucinations:
  - Lexicon approach (replace + paraphrase, method A)
  - Prompt—based (BERTIN Alpaca, method B)
  - mBART (method C)
  - mBART + fine-tuning with lexicon (method D)
  - mBART + fine-tuning with lexicon + post-processing with lexicon (method E)

Human evaluation

Score	Grammaticality / Fluency
	Is the simplified sentence grammatically correct or sufficiently readable? Does it have errors in syntax, agreement or too many words that make it difficult to read?
5	The sentence is fluent (native speaker level) and grammatically correct.
4	The sentence is nearly fluent (non-native speaker level) and grammatically correct.
3	The sentence is less fluent, with some ungrammatical but understandable parts.
2	The sentence is less fluent and with fewer grammatical parts, but is partially understandable.
1	The sentence is completely unintelligible.
Score	Semantic adequacy / Meaning preservation
	Does the simplified sentence adequately preserve the original meaning, and is it meaningless or inadequate in relation to the original?
5	The sentence adequately conveys the core meaning of the original sentence.
4	The sentence mostly conveys the essential meaning of the original sentence.
3	The central meaning of the original text is not conveyed, but some information from the original text is partially preserved.
2	The meaning of the sentence is markedly different from the original sentence, even contrary.
1	It is impossible to evaluate the meaning of the target sentence due to its unintelligibility.
Score	Simplification
	Is the resulting text simpler than the original? Is the simplification of good quality?
5	The target sentence is much simpler than the original and no information is lost.
4	The target sentence is simpler than the original and not much (or any) information is lost.
3	The target sentence is as simple/difficult as the original.
2	The target sentence is slightly more difficult than the original and/or the resulting sentence loses some relevant information.
1	The target sentence is more difficult and/or loses important information compared to the original; or it is impossible to evaluate the simplicity of the target sentence due to its unintelligibility.

Instructions for human evaluation, translated to English and adapted from (Yamaguchi et al., 2023)

Results

Results – Quantitative evaluation

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(A) Only lexicon	Replacements	**46.69 (±1.06)**	**28.64 (±0.82)**	**41.79 (±1.00)**	45.85 (±0.5)	**19.67 (±0.74)**
	Replacements + *par./def.*	46.24 (±0.8)	27.49 (±0.65)	40.29 (±0.79)	**47.41 (±0.53)**	14.49 (±0.53)

Table 1: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method A)

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(B) Only prompts	BERTIN (zero-shot prompting)	**41.89 (±0.62)**	**23.19 (±0.66)**	**35.60 (±0.72)**	43.07 (±0.27)	**15.75 (±0.78)**
	BERTIN (few-shot prompting)	39.27 (±0.59)	21.43 (±0.55)	34.08 (±0.60)	**43.29 (±0.52)**	12.01 (±0.41)

Table 2: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method B)

Results – Quantitative evaluation

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(C) Fine-tuning without lexicon	BERTIN	46.90	29.17	41.73	47.00	22.03
		(±1.95)	(±1.54)	(±1.79)	(±2.74)	(±0.74)
	MariMari	42.47	24.28	36.78	47.08	17.89
		(±0.66)	(±0.78)	(±0.67)	(±0.72)	(±0.94)
	mBART	**48.82**	**31.04**	**43.96**	**50.93**	**22.90**
		(±1.02)	**(±1.01)**	**(±1.02)**	**(±0.62)**	**(±0.98)**
	mT5	34.20	19.85	31.56	40.46	6.58
		(±0.79)	(±0.61)	(±0.66)	(±0.45)	(±0.39)
	NASES	44.33	26.51	38.54	48.68	20.47
		(±1.18)	(±1.25)	(±1.23)	(±0.64)	(±1.16)
	Pegasus	43.78	25.86	39.74	41.65	14.62
	XSUM	(±0.7)	(±0.85)	(±0.76)	(±0.23)	(±0.43)

Table 3: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method C)

Results – Quantitative evaluation

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(D) Fine-tuning with lexicon	*BERTIN (zero-shot)*	48.34 (±0.83)	29.27 (±0.85)	42.95 (±0.98)	49.09 (±0.52)	21.08 (±0.90)
	MariMari	43.35 (±0.92)	25.95 (±2.4)	37.75 (±1.06)	47.35 (±0.75)	18.30 (±1.07)
	mBART	**50.50 (±0.98)**	**32.65 (±1.00)**	**45.33 (±1.03)**	**51.20 (±0.63)**	**24.71 (±1.34)**
	NASES	30.10 (±16.07)	14.74 (±12.63)	25.52 (±13.97)	42.67 (±7.57)	11.16 (±9.31)

Table 4: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method D)

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(E) Fine-tuning with lexicon + post-processing (par./def.)	MariMari	41.88 (±0.99)	23.8 (±1.19)	35.83 (±1.04)	47.7 (±0.62)	14.27 (±1.16)
	mBART	**48.07 (±0.82)**	**30.07 (±0.83)**	**42.42 (±0.82)**	**51.11 (±0.43)**	**18.26 (±0.83)**
	NASES	39.74 (±1.66)	22.2 (±1.84)	33.37 (±1.74)	48.17 (±0.44)	13.14 (±1.08)

Table 5: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method E)

Results – Quantitative evaluation

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(F) Lexical substitution with lexicon + fine-tuning	MariMari	42.73 (±0.94)	24.41 (±1.00)	36,99 (±1.11)	47.44 (±0.72)	17.63 (±0.97)
	mBART	**48.83 (±1.16)**	**30.69 (±1.27)**	**43.75 (±1.14)**	**50.28 (±0.79)**	**22.90 (±0.98)**
	NASES	44.19 (±0.70)	22.26 (±0.85)	38.31 (±0.74)	48.16 (±0.50)	19.15 (±1.28)

Table 6: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method F)

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(G) Lexical substitution with lexicon + fine-tuning + post-processing	MariMari	44.0 (±0.82)	25.57 (±0.80)	38.28 (±1.03)	47.29 (±0.48)	17.9 (±1.00)
	mBART	**49.5 (±2.34)**	**31.42 (±2.34)**	**44.48 (±2.42)**	**50.12 (±1.01)**	**23.25 (±1.96)**
	NASES	45.32 (±1.02)	27.12 (±1.22)	39.2 (±1.12)	47.63 (±0.55)	19.73 (±0.89)

Table 7: Results of the quantitative evaluation with the 3800 aligned sentences pairs (Method G)

Results – Quantitative evaluation

Method		Rouge1	Rouge2	RougeL	SARI	BLEU
(A) Only lexicon	Replacements	73.95 (±0.86)	59.72 (±1.32)	68.77 (±0.98)	50.81 (±0.75)	41.11 (±1.13)
	Replacements + par./def.	69.29 (±1.0)	53.66 (±1.28)	62.71 (±0.94)	49.55 (±0.90)	36.41 (±1.98)
(C) Fine-tuning without lexicon	*BERTIN*	70.43 (±1.63)	54.21 (±1.71)	63.31 (±1.59)	51.87 (±1.73)	43.40 (±2.26)
	mBART	**76.96 (±0.82)**	**63.57 (±0.97)**	**71.37 (±0.77)**	**61.04 (±0.86)**	**52.49 (±1.06)**
(D) Fine-tuning with lexicon	*BERTIN (zero-shot)*	72.92 (±0.42)	59.77 (±0.70)	68.12 (±0.22)	55.66 (±0.41)	44.5 (±1.08)
	mBART	74.78 (±0.84)	61.21 (±1.21)	69.36 (±0.97)	57.81 (±0.87)	48.95 (±1.48)
(E) Fine-tuning with lexicon + post-processing (par./def.)	*mBART*	70.55 (±0.64)	55.75 (±1.1)	64.00 (±0.75)	55.71 (±0.47)	40.68 (±1.86)

Table 8: Results of the quantitative evaluation with the 1200 manually simplified sentences pairs (average score ± standard deviation of 5 experimental rounds); *par.*: 'paraphrase'; *def.*: 'definition'; the name of the methods that the human evaluators assessed qualitatively appears in italics; best results in bold

Results – Human evaluation

Method	Manually-aligned sentences (n=3800)				Manually-simplified sentences (n=1200)			
	G	M	S	Avg	G	M	S	Avg
Lexicon (replace + paraphrase/definition)	4.1	**4.3**	3.3	3.9	4.2	**4.5**	**3.3**	4.0
mBART	4.2	3.3	2.7	3.4	4.2	3.9	3.0	3.7
mBART + fine-tune with lexicon	4.3	3.7	3.1	3.7	4.2	3.9	3.0	3.7
mBART + fine-tune lexicon + post-proc.	4.0	3.7	3.1	3.6	3.9	3.9	3.0	3.6
Prompt learning fine-tuned with lexicon	**4.7**	**4.3**	**3.4**	**4.1**	**4.7**	**4.5**	3.1	**4.1**

Table 9: Results (average of 5-point Likert scale) of the human evaluation (n=11) of simplifying the 3800 sentence pairs (left) and 1200 sentences pairs (right) for Grammaticality / Fluency (G), Semantic adequacy / Meaning preservation (M), and Simplification (S); and average (Avg) of scores for the three aspects

Simplification samples

Original	*Disección radical de ganglios linfáticos.* 'Radical dissection of lymph nodes'
Only lexicon	*Disección radical de ganglios linfáticos (nódulos que forman parte del sistema inmunitario).* 'Radical dissection of lymph nodes (nodes that are part of the immune system)'
mBART	*Dificultad para abrir los ganglios linfáticos.* 'Difficulty opening the lymph nodes'
mBART + lexicon	*A veces se administran ganglios linfáticos a través de la piel.* 'Sometimes, lymph nodes are administered through the skin.'
mBART + lexicon + postproces.	*A veces se administran ganglios linfáticos (nódulos que forman parte del sistema inmunitario) a través de la piel.* 'Sometimes, lymph nodes (nodes that are part of the immune system) are administered through the skin.'
BERTIN + lexicon	*Disección radical de ganglios linfáticos (extirpación de todos los ganglios linfáticos de una zona del cuerpo)* 'Radical dissection of lymph nodes (removal of all lymph nodes from a body part)'

Discussion and Conclusions

Discussion

- **Findings:**
  - Using the **lexicon** tended to obtain **higher results**.
  - **BLEU**, **ROUGE** or **SARI** metrics did **not match human evaluation** scores
  - The **only-lexicon approach** and the **prompt-method** rated in **first position** wrt **semantic adequacy** and **simplification** → **a lexicon is preferred for correctedness**
  - The prompt-based method showed balanced scores across all rated aspects

- **Limitations:**
  - We did **not** apply any method for **candidate selection** before lexicon substitution
  - The **coverage of the lexicon** still needs to be **improved**

Conclusions

- **Experiments combining a patient-oriented medical lexicon and deep learning**
- **Quantitative evaluations and human assessments** (n = 11 subjects)
- Findings show that **using a lexicon increased results** wrt methods without it
- **New resource distributed and trained models released publicly**
- Results might be **similar** in **other languages** → **still to be confirmed**

Thanks for your attention!

Questions?

CSIC
Consejo Superior de Investigaciones Científicas

leonardo.campillos@csic.es

CLARA NLP MED

https://clara-nlp.uned.es/home/med/

10. Bibliographical References

Vicent Ahuir, Lluís-F Hurtado, José Ángel González, and Encarna Segarra. 2021. NASca and NASes: Two monolingual pre-trained models for abstractive summarization in Catalan and Spanish. *Applied Sciences*, 11(21):9872.

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Rodrigo Alarcón, Paloma Martínez, and Lourdes Moreno. 2023. Tuning BART models to simplify Spanish health-related content. *Procesamiento del Lenguaje Natural*, 70:111–122.

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2023. EASIER corpus: A lexical simplification resource for people with cognitive impairments. *Plos one*, 18(4):e0283622.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proc. of the 58th Annual Meeting of the ACL*, pages 4668–4679, Online. Association for Computational Linguistics.

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-EASi: Finely annotated dataset and models for controllable simplification of medical texts. *Proc. of AAAI 2023*, pages 14093–14101.

BERTIN-project. 2023. BERTIN-GPT-J-6B Alpaca. https://huggingface.co/bertin-project/bertin-gpt-j-6B-alpaca Accessed 26 June 2023.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Stefan Bott, Horacio Saggion, and David Figueroa. 2012. A hybrid system for Spanish text simplification. In *Proc. of the Third Workshop on Speech and Language Processing for Assistive Technologies*, pages 75–84.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2014. Syntactic sentence simplification for French. In *Proc. of the 3rd PITR Workshop@ EACL 2014*, pages 47–56.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Leonardo Campillos-Llanos. 2023. MedLexSp–A medical lexicon for Spanish medical natural language processing. *Journal of Biomedical Semantics*, 14(1):1–23.

Leonardo Campillos-Llanos, Rocío Bartolomé Rodríguez, and Ana R Terroba Reinares. 2024. Enhancing the understanding of clinical trials with a sentence-level simplification dataset. *Procesamiento del lenguaje natural*, 72.

Leonardo Campillos-Llanos, Ana R Terroba Reinares, Sofía Zakhir Puig, Ana Valverde, and Adrián Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for Spanish medical text simplification. *Procesamiento del lenguaje natural*, 69:189–196.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716.

Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, Hong Yu, et al. 2018. A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews. *Journal of medical Internet research*, 20(1):e8669.

José Camacho Collados. 2013. Splitting complex sentences for natural language processing applications: Building a simplified Spanish corpus. *Procedia-Social and Behavioral Sciences*, 95:464–472.

William Coster and David Kauchak. 2011. Simple English Wikipedia: a new text simplification task. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 665–669.

Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carriño, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68:39–60.

Yilu Fang, Jae Hyun Kim, Betina Ross S Idnay, Rebeca Aragon Garcia, Carmen E Castillo, Yingcheng Sun, Hao Liu, Cong Liu, Chi Yuan, and Chunhua Weng. 2021. Participatory design of a clinical trial eligibility criteria simplification method. In *Proc. of MIE*, pages 984–988.

Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: a dataset for lexical simplification in Spanish. In *Proceedings of LREC 2022*, pages 3582–94, Marseille, France.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Susannah Fox and Maeve Duggan. 2013. Health online 2013. *Health*, 2013:1–55.

Natalia Grabar and Rémi Cardon. 2018. CLEAR - Simple corpus for medical French. In *Proc. of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9.

Natalia Grabar and Thierry Hamon. 2016. A large rated lexicon with French medical words. In *Proc. of LREC 2016*, pages 2643–2648, Portorož, Slovenia.

Natalia Grabar and Horacio Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 453–463.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2022. Cells: A parallel corpus for biomedical lay language generation. *arXiv preprint arXiv:2211.03818*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 328–339.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh J Ramanathan, Wei Xu, Byron C Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. *arXiv preprint arXiv:2305.12532*.

Alla Keselman, Robert Logan, Catherine Arnott Smith, Gondy Leroy, and Qing Zeng-Treitler. 2008. Developing informatics tools and strategies for consumer-centered health communication. *Journal of the American Medical Informatics Association*, 15(4):473–483.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Anaïs Koptient and Natalia Grabar. 2020a. Fine-grained text simplification in French: steps towards a better grammaticality. In *Proc. of Int. Symp. on Health Information Management Research*.

Anaïs Koptient and Natalia Grabar. 2020b. Rated lexicon for the simplification of medical texts. In *Proc. of HEALTHINFO 2020*, Porto, Portugal.

Gondy Leroy and James E Endicott. 2012. Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 749–754.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proc. of Workshop on Text Summarization of ACL*, pages 74–81, Barcelona, Spain.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proc.of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3536–3553, Online. Association for Computational Linguistics.

Philip D Marshall. 2000. Bridging the terminology gap between health care professionals and patients with the Consumer Health Terminology (CHT). In *Proceedings of the AMIA Symposium*, page 1082. American Medical Informatics Association.

Francesco Moramarco, Damir Juric, Aleksandar Savkov, Jack Flann, Maria Lehl, Kristian Boda, Tessa Grafen, Vitalii Zhelezniak, Sunir Gohil, Alex Papadopoulos Korfiatis, et al. 2021. Towards more patient friendly clinical notes through

language models and ontologies. In *Proc. of the AMIA Annual Symposium*, pages 881–890.

Partha Mukherjee, Gondy Leroy, David Kauchak, Srinidhi Rajanarayanan, Damian Y Romero Diaz, Nicole P Yuan, T Gail Pritchard, and Sonia Colina. 2017. NegAIT: A new parser for medical text simplification using morphological, sentential and double negation. *Journal of biomedical informatics*, 69:55–62.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Yifan Peng, Catalina O Tudor, Manabu Torii, Cathy H Wu, and K Vijay-Shanker. 2012. iSimp: A sentence simplification system for biomedical text. In *2012 IEEE Int. Conference on Bioinformatics and Biomedicine*, pages 1–6. IEEE.

Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael Hogarth. 2017. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. *Journal of medical Internet research*, 19(12):e417.

RANME. 2011. *Diccionario de términos médicos*. Madrid: Panamericana.

Jessica Ross, Samson Tu, Simona Carini, and Ida Sim. 2010. Analysis of eligibility criteria complexity in clinical trials. *Proc. of the Summit on Translational Bioinformatics*, 2010:46.

Horacio Saggion. 2017. *Automatic text simplification*, volume 32. Synthesis Lectures on Human Language Technologies, Springer.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Tarek Sakakini, Jong Yoon Lee, Aditya Duri, Renato FL Azevedo, Victor Sadauskas, Kuangxiao Gu, Suma Bhat, Dan Morrow, James Graumlich, Saqib Walayat, et al. 2020. Context-aware automatic text simplification of health materials in low-resource domains. In *Proc. of the 11th LOUHI Workshop*, pages 115–126.

Carolina Scarton, Alessio Palmero Aprosio, Sara Tonelli, Tamara Martín Wanton, and Lucia Specia. 2017. MUSST: A multilingual syntactic simplification tool. In *Proc. of the IJCNLP 2017, System Demonstrations*, pages 25–28.

Laura Seiffe, Oliver Marten, Michael Mikhailov, Sven Schmeier, Sebastian Möller, and Roland Roller. 2020. From witch's shot to music making bones - resources for medical laymen to technical language and vice versa. In *Proc. of LREC 2020*, pages 6185–6192, Marseille, France.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Matthew Shardlow and Fernando Alva-Manchego. 2022. Simple TICO-19: A dataset for joint translation and simplification of COVID-19 texts. In *Proceedings of LREC 2022*, pages 3093–3102.

Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proc. of the 57th ACL*, pages 380–389, Florence, Italy. Association for Computational Linguistics (ACL).

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Julia Suter, Sarah Ebling, and Martin Volk. 2016. Rule-based automatic text simplification for German. In *Proc. of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 280–287, Bochum, Germany.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca` Accessed 26 June 2023.

Amalia Todirascu, Rodrigo Wilkens, Eva Rolin, Thomas François, Delphine Bernhard, and Nuria Gala. 2022. Hector: A hybrid text simplification tool for raw texts in French. In *Proc. of the 12th Language Resources and Evaluation (LREC)*, pages 4620—4630.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2022. Patient-friendly clinical notes: towards a new text simplification dataset. In *Proc. of the TSAR-2022 Workshop*, pages 19–27, Marseille, France.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *Proc. of the World Wide Web Conference*, pages 3286–3292.

Vinod Vydiswaran, Qiaozhu Mei, David A Hanauer, and Kai Zheng. 2014. Mining consumer health vocabulary from community-generated text. In *AMIA Annual Symposium Proceedings*, volume 2014, pages 1150–1159. American Medical Informatics Association.

Haochun Wang, Chi Liu, Nuwa Xi, Sendong Zhao, Meizhi Ju, Shiwei Zhang, Ziheng Zhang, Yefeng Zheng, Bing Qin, and Ting Liu. 2022. Prompt combines paraphrase: Teaching pre-trained models to understand rare biomedical words. In *Proc. of the 29th International Conference on Computational Linguistics*, pages 1422–1431, Gyeongju, Republic of Korea.

Rodrigo Wilkens, Bruno Oberle, and Amalia Todirascu. 2020. Coreference-based text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 93–100.

Rodrigo Wilkens and Amalia Todirascu. 2020. Un corpus d'évaluation pour un système de simplification discursive. In *6e conférence conjointe JEP 33e éd., TALN, 27e éd., RÉCITAL, 22e éd.*, pages 361–369. ATALA; AFCP.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. *Proc. EMNLP 2020: system demonstrations*, pages 38–45. Association for Computational Linguistics.

Danny TY Wu, David A Hanauer, Qiaozhu Mei, Patricia M Clark, Lawrence C An, Joshua Proulx, Qing T Zeng, VG Vinod Vydiswaran, Kevyn Collins-Thompson, and Kai Zheng. 2016. Assessing the readability of ClinicalTrials.gov. *Journal of the American Medical Informatics Association*, 23(2):269–275.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of NAACL 2021*, pages 483–498, Online. Association for Computational Linguistics.

Daichi Yamaguchi, Rei Miyata, Sayuka Shimada, and Satoshi Sato. 2023. Gauging the gap between human and machine text simplification through analytical evaluation of simplification strategies and errors. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 359–375.

Qing Zeng-Treitler, Hyeoneui Kim, Sergey Goryachev, Alla Keselman, Laura Slaughter, and Catherine-Arnott Smith. 2007. Text characteristics of clinical reports and their implications for the readability of personal health records. *Studies in health technology and informatics*, 129(2):1117.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *International Conference on Machine Learning*, pages 11328–11339.

11. Language Resource References

Leonardo Campillos-Llanos, Rocío Bartolomé and Ana R. Terroba. 2024. *CLARA-MeD simplified sentences*. CSIC. Distributed via Digital CSIC. PID https://doi.org/10.20350/digitalCSIC/16110.

Leonardo Campillos-Llanos, 2022. *CLARA-MeD corpus*. CSIC. Distributed via Digital CSIC. PID https://doi.org/10.20350/digitalCSIC/14644.

Daniel Ferrés and Horacio Saggion. 2022. *ALEXIS*. Universidad Pompeu Fabra. PID https://github.com/lastus-taln-upf/alexsis.

HIP. 1995. *EUGLOSS: Multilingual Glossary of technical and popular medical terms in nine European Languages*. Heymans Institute of Pharmacology.

Leonardo Campillos-Llanos. 2023. *Medical Lexicon for Spanish (MedLexSp)*. CSIC. Distributed via Digital CSIC, 1.0. PID https://digital.csic.es/handle/10261/270429.