Data statement of the Corpus of Basque Simplified Texts

Data set name: The Corpus of Basque Simplified Texts (CBST) - *Euskarazko Testu Sinplifikatuen Corpusa* (ETSC)
Citation (if available): Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza. (2018). The Corpus of Basque Simplified Texts (CBST). Language Resources and Evaluation 52. pp. 217-247.
Data set developer(s): Itziar Gonzalez-Dios, María Jesús Aranzabe, Edurne Aldasoro, Arantza Díaz de Ilarraza
Data statement author(s): Itziar Gonzalez-Dios
Others who contributed to this document: Totok Suhardijanto, Renny Pradina Kusumawardani, Jerry Spanakis, Catalina Goanta, Surangika Ranathunga, María Jesús Aranzabe, Arantza Díaz de Ilarraza
Link to the dataset: http://ixa.si.ehu.es/node/13007
Dataset license: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

A. CURATION RATIONALE

The corpus of Basque Simplified Texts is a corpus that compiles original texts from the Scientific popularisation magazine *Elhuyar aldizkaria* (chosen because it is the only one of this topic for adults; texts obtained in 2025 with the permission and collaboration of the editors, license in 2020 CC-BY-SA-3.0) and manually simplified texts. The original texts were randomly selected (similar size) from three different domains. It contains 227 original sentences. The simplified texts are two versions of each original text according to simplification strategies (structural and intuitive). The aim of this corpus is to study simplification operations in different strategies and domains. In order to create this corpus, we followed these steps: 1) the original texts (considered as complex texts) were selected from the *T-comp* corpus (Gonzalez-Dios et al., 2014); 2) the original texts were simplified by two experts (see Section C) and 3) an annotator tagged the simplication operations (see Section D), namely the changes undergone from the original texts to the simplified texts.

B. LANGUAGE VARIETY/VARIETIES

The language of this corpus is Standard Basque (ISO 639-1 eu).

C. SPEAKER DEMOGRAPHIC

The original texts were written by three journalists specialised in science popularisation. The simplified texts were produced by a court translator and as Basque as FL teacher and university professor. See Table 1 for more details.

	Original texts	**Simplified texts**
Age	30-50	45-50

Gender	1 man, 2 woman	2 women
Race/ethnicity	Basque/ Caucasian	Basque/ Caucasian
First languages	Basque (Spanish?)	Basque
Socioeconomic status	Studies: University degree	Studies: University degree
Number of speakers	3	2
Presence disorder speech	No	No

TABLE 1: speaker demographics (summary)

D. ANNOTATOR DEMOGRAPHIC

The simplification operations were annotated in 2015. The annotator was a last-year PhD student on computational linguistics (Linguistics BA and Computational Linguistics MS), 26 year old, Basque woman natively bilingual Basque/Spanish. The guidelines were based on other similar works and are explained in the paper (Gonzalez-Dios et al. 2018).

E. SPEECH SITUATION

In the case of the original texts, the texts were created in 2011 and 2012 in the Basque Country. The simplified texts were created in 2015 in the Basque Country. The modality is written and the texts were edited. The interaction is asynchronous. The intended audiences are i) for the original texts, the readers of the magazine, and people generally interested in science (the wider public); and ii) for the simplified text, this was an experiment for the PhD thesis, but the simplified text the audience was people learning Basque.

F. TEXT CHARACTERISTICS

These texts are scientific popularisation reports on science topics, exactly astronomy, medicine, physics.

G. RECORDING QUALITY

N/A

H. OTHER

We provide the links obtained in 2020 for the original texts:

- Bernoulli gabe hegan: https://zientzia.eus/artikuluak/bernoulli-gabe-hegan/
- Etxeko historiak agintzen duenean (last accessed 2020/05/13)
- https://zientzia.eus/artikuluak/etxeko-historiak-agintzen-duenean/ (last accessed 2020/05/13)
- Exoplanetak. Lurraren parekoen bila espazioan https://zientzia.eus/artikuluak/exoplanetak-lurraren-parekoen-bila-espazioan/ last accessed 2020/05/13)

I. PROVENANCE APPENDIX

The original texts have been obtained from the *T-comp corpus*, which was crawled in 2014 to train the readability assessment system for Basque (Gonzalez-Dios et al., 2014).

REFERENCES

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza. (2018). The corpus of Basque simplified texts (CBST). *Language Resources and Evaluation.* 52. pp. 217-247.

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, Haritz Salaberri. (2014). Simple or Complex? Assessing the readability of Basque Texts. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.* 334--344, Dublin City University and Association for Computational Linguistics.