Data Statement for the Corpus of Basque Simplified Texts

Itziar Gonzalez-Dios

10/03/2024

1 Header

Dataset Title: The Corpus of Basque Simplified Texts (CBST) - Euskarazko Testu Sinplifikatuen Corpusa (ETSC)

Dataset Curator(s): Itziar Gonzalez-Dios, HiTZ Basque Center for Language Technologies - Ixa, University of the Basque Country UPV/EHU, data compiling, formatting and annotation

Dataset Version: V1, 2015

Dataset Citation: 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 Statement Authors: Itziar Gonzalez-Dios, HiTZ Basque Center for Language Technologies - Ixa, University of the Basque Country UPV/EHU, main author

Data Statement Version: 3, 10th March, 2024

Data Statement Citation and DOI: Gonzalez-Dios, Itziar (2024) Data Statement for the Corpus of Basque Simplified Texts. Version 3. University of the Basque Country (UPV/EHU). `http://www.ixa.eus/node/13302`

Links to versions of this data statement in other languages: None

2 Executive Summary

The corpus of Basque Simplified Texts (CBST) - Euskarazko Testu Sinplifikatuen Corpusa (ETSC) is a corpus collected in order to study the linguistic operations carried out when making text more simple, accessible for a general target audience. It contains 227 original sentences in standard Basque. For each original sentence, two simplified versions are available: one of them intuitively simplified and the other structurally simplified.

3 Curation Rationale

The corpus of Basque Simplified Texts is a corpus that compiles original texts from the Scientific popularisation magazine Elhuyar aldizkaria[1] (chosen because it is the only one of this topic for adults; texts obtained in 2015 with the permission and collaboration of the editors, which have the license CC-BY-SA-3.0 in 2020[2], and updated in 2024 to CC BY-SA 4.0 Deed[3] 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).

This corpus was created with the aim of studying simplification operations in different approaches 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 [2, 1].
2. The original texts were simplified by two experts: a court translator simplified the texts ac-

[1] `https://aldizkaria.elhuyar.eus/`
[2] `https://creativecommons.org/licenses/by-sa/3.0/`
[3] `https://creativecommons.org/licenses/by-sa/4.0/deed.en)`

cording to easy-to-read guidelines (structural approach) and a teacher simplified the texts based on her experience (intuitive approach). They perform the simplifications at the same time (Spring 2015), but independently. That is, each expert simplified the texts only following one approach.

3. An annotator tagged the simplification operations, namely the changes undergone from the original texts to the simplified texts.

In Table 1 we show an example of an original sentence and its two simplifications.

4 Documentation for Source Datasets

The original texts have been obtained from the T-comp corpus[4], which was crawled in 2014 to train the readability assessment system for Basque [2].

5 Language Varieties

The language of this corpus is Standard Basque (ISO 639-1 eu). Most of Basque speakers are natively bilingual. In addition to Basque they speak French or Spanish. Basque is written in Latin alphabet.

6 Language User Demographic

The original texts were written by three journalists specialised in science popularisation. The simplified texts were produced by a court translator and a Basque as foreign language teacher and university professor. See Table 2 for more details.

7 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 (Caucasian) woman natively bilingual Basque/Spanish. The guidelines were based on other similar works and are explained in the paper [3].

8 Linguistic Situation and Text Characteristics

These texts are scientific popularisation reports on science topics, exactly astronomy, medicine, physics. 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. All the images that were in the original texts were removed from the dataset.

9 Preprocessing and Data Formatting

The dataset is offered in txt (raw texts), csv (aligned) text and ann (annotation) formats. Texts were not anonymized.

The original texts were crawled from the web (html format) and the text was extracted (raw texts). We provide the links obtained in 2024 for the original texts:

- Bernoulli gabe hegan: `https://zientzia.eus/artikuluak/bernoulli-gabe-hegan/` (last accessed 10/03/2024)
- Etxeko historiak agintzen duenean: `https://zientzia.eus/artikuluak/etxeko-historiak-agintzen-duenean/` (last accessed 10/03/2024)
- Exoplanetak. Lurraren parekoen bila espazioan `https://zientzia.eus/artikuluak/exoplanetak-lurraren-parekoen-bila-espazioan/` (last accessed 10/03/2024) The simplified texts

[4]The whole T-comp corpus is available in GitHub together with the readability assessment tool MultiAztertest [1] under the GPL-3.0 license `https://github.com/kepaxabier/MultiAzterTest/tree/master/corpus/eu/simplecomplejo/complejo`

	Original sentence (complex)	**Structural simplification**	**Intuitive simplification**
Corpus sentence	Eguzki-sistemaz kanpoko planetei esaten zaie exoplaneta, eta Kanarietako Astrofisika Institutuko Enric Pallé astrofisikariak dioen bezala, "une honetan, egunero-egunero aurkitzen dira exoplaneta berriak; ez nintzateke harrituko hemendik bost urtera 20-25 mila exoplaneta ezagutuko bagenitu".	Exoplanetak dira eguzki-sistematik kanpo dauden planetak. Kanarietako Astrofisika Institutuan lan egiten du Enric Pallé astrofisikariak. Astrofisikari horrek hau esaten du: "une honetan, egunero-egunero aurkitzen dira exoplaneta berriak; ez nintzateke harrituko hemendik bost urtera 20-25 mila exoplaneta ezagutuko bagenitu".	Eguzki-sistemaz kanpoko planetei esaten zaie exoplaneta. Enric Pallé Kanarietako Astrofisika Institutuko astrofisikaria da; Pallék dio: egunero-egunero aurkitzen dira exoplaneta berriak; ez da harritzekoa hemendik bost urtera 20-25 mila exoplaneta ezagutzea.
English Translation	Planets out of the solar system are called exoplanets and, as astrophysicist Enric Pallé of the Canary Institute of Astrophysics says, "at this moment, every day new exoplanets are found, I would not be surprised if within five years we knew 20-25 thousand exoplanets".	Exoplanets are planets out of the solar system. The astrophysicist Enric Pallé works at the Canary Institute of Astrophysics. This astrophysicist says: "at this moment, every day new exoplanets are found, I would not be surprised if within five years we knew 20-25 thousand exoplanets".	Planets out of the solar system are called exoplanets. Enric Pallé is an astrophysicist at the Canary Institute of Astrophysics; Pallé says: "every day new exoplanets are found, I would not be surprised if within five years we knew 20-25 thousand exoplanets".

Table 1: Example of the corpus and its translation

	Original texts	**Simplified texts**
Age	30-50	45-50
Gender	1 man, 2 women	2 women
Race/ethnicity	Basque/ Caucasian	Basque/ Caucasian
First languages	Basque (Spanish?)	Basque
Socio-economic status	Studies: University degree	Studies: University degree
Number of speakers	3	2
Presence disordered speech	No	No

Table 2: Language User Demographic (summary)

were created in different word processing systems (Word and Writer). The alignments of the texts were done manually with LibreOffice Calc. The annotation was done with BrAT [4].

10 Capture Quality

N/A

11 Limitations

This dataset addresses general simplification. But, as text simplification is target audience dependent, this dataset may not suit the different needs of the particular target audiences.

12 Metadata

Annotation Guidelines: Chapter 7 from Itziar Gonzalez-Dios thesis Basque version `http://ixa.si.ehu.es/sites/default/files/dokumentuak/4102/TESIS_GONZALEZ_DIOS_ITZIAR%28eus%29.pdf`, English version (shorter) `http://ixa.si.ehu.es/sites/default/files/dokumentuak/4102/TESIS_GONZALEZ_DIOS_ITZIAR%28eng%29.pdf`

Annotation Process: The creators of the simplified texts wrote the texts on a voluntary basis and they were invited to lunch as compensation; the annotator of the simplification operations did the work in the contexts of her PhD thesis. No automatic process was carried out.

Dataset Quality Metrics: The dataset was annotated by only one annotator and, therefore, no inter-annotator agreement measures are reported

13 Disclosures and Ethical Review

This dataset was created during Itziar Gonzalez-Dios's PhD., who was funded by a grant from the Basque Government (BFI-2011-392). This research was also supported by the Basque Government (IT344-10), and the Spanish Ministry of Economy and Competitiveness, EXTRECM Project (TIN2013-46616-C2-1-R).

14 Distribution

 It can be downloaded in a zip file from `http://www.ixa.eus/node/13007`. In this zip file, the texts, the annotations and a readme file are included.

15 Maintenance

So far, this dataset has only one version and there is no intention to make another version of it or to augment it in the near future. The dataset will be maintained by Itziar Gonzalez-Dios and she will make the updates, correct the errors... if necessary.

16 Other

N/A

17 Glossary

- **Original texts/sentences:** Texts/sentences written for a general target audience.
- **Simplified texts/sentences:** Modified original texts/sentences with the aim of making them more simple.
- **Intuitive simplification:** Simplification approach, where the person who simplifies follows their intuition and experience to simplify the texts.
- **Structural simplification:** Simplification approach, where the person who simplifies follows some guidelines

About this document

A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software.

This data statement was written based on the template for the Data Statements Version 3 Schema. The template was prepared by Angelina McMillan-Major and Emily M. Bender and can be found at `http://techpolicylab.uw.edu/data-statements`.

References

[1] Kepa Bengoetxea and Itziar Gonzalez-Dios. Multiaztertest: A multilingual analyzer on multiple levels of language for readability assessment. *arXiv preprint arXiv:2109.04870*, 2021.

[2] Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. Simple or Complex? Assessing the Readability of Basque Texts. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 334–344, 2014.

[3] Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. The Corpus of Basque Simplified Texts (CBST). *Language Resources and Evaluation*, 52(1):217–247, 2018.

[4] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. BRAT: a Web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, 2012.