BaSCo: An Annotated Basque-Spanish Code-Switching Corpus for Natural Language Understanding

Maia Aguirre, Laura García-Sardiña, Manex Serras, Ariane Méndez, Jacobo López
Speech and Natural Language Technologies Department at Vicomtech, Donostia/San Sebastián, Spain
{magirre, lgarcias, mserras, amendez, jlopez}@vicomtech.org

Introduction

BaSCo –*Ba*sque-*S*panish *Co*de-Switching– is the first corpus with annotated linguistic resources encompassing Basque-Spanish code-switching. Publicly available at: **https://github.com/Vicomtech/BaSCo-Corpus**

Source Data and Compilation

- **Departure point**: the texts used for training the NLU modules of four task-oriented bilingual chatbots.
- **Domains**: administration, transport, fiscal, generic, and social.
- **Size**: 1936 reference sentences in Spanish and 2216 in Basque.
- **BaSCo gathering**: distribution of a web interface for submitting code-switched proposals given a reference utterance.

Data Curation and Annotation

Three Basque-Spanish bilingual annotators would consider an utterance valid if:

- It is, to whatsoever extent, in a mixture of Spanish and Basque.
- Its semantic content remains the same as its reference text's.
- It sounds natural.

The corpus is annotated at three levels:

- **NLU Annotation:** intents and entities.
- **Code-Switching Level Annotation:** annotators' subjective perspectives on the proportion of Basque and Spanish.
- **Domain of Origin:** administration, transport, fiscal, generic or social.

Corpus Structure

```
"referent": "dónde está la casa del
             deporte?",
"source_lang": "es",
"domain": "administration",
"intents": [
       "preguntar|ubicacion",
       "informar|tipo-oficina"
],
"entities": [
    {
      "entity": "tipo-oficina",
      "value": "casa del deporte",
      "normative_value": "deportes",
      "start": 14,
      "end": 29,
      "type": "bounded"
    }
],
"code_switching": [
    {
      "text": "Casa de deporte non
               dago?",
      "entities": [
        {
          "entity": "tipo-oficina",
          "value": "Casa de deporte",
          "normative_value": "deportes",
          "start": 0,
          "end": 14,
          "type": "bounded"
        }
      ],
      "lang_proportion": "balanced"
    }
]
```

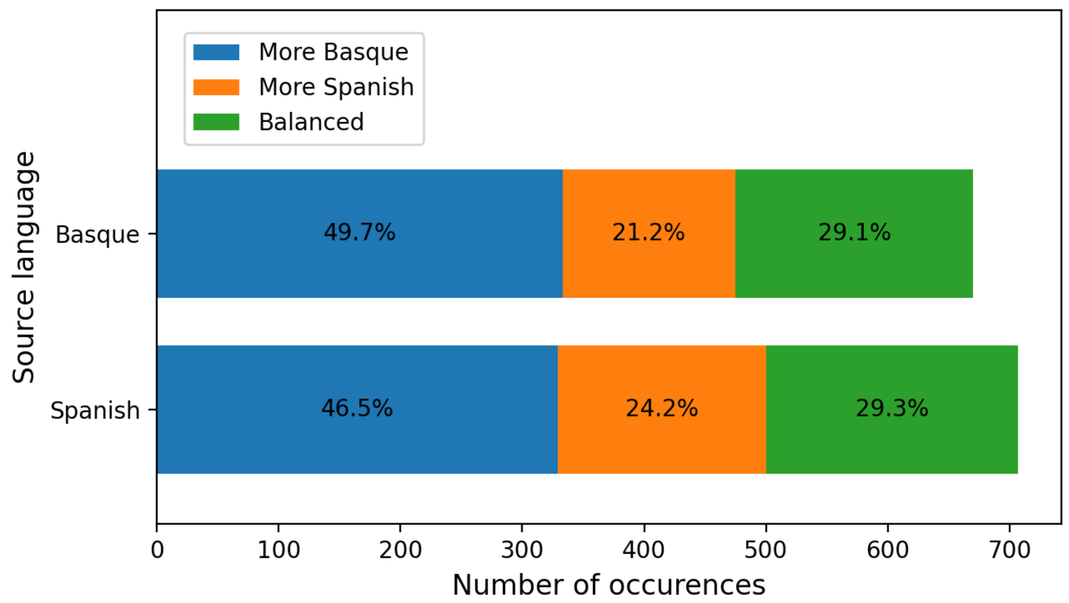
Statistics

Domain	Basque	Spanish	Euskañol
Generic	517	383	119
Social	250	271	205
Administration	488	490	538
Transport	141	136	55
Fiscal	820	656	460
Total	2216	1936	1377

Domain	Intents (Unique/Freq.)	Entities (Unique/Freq.)
Generic	9 / 191	0
Social	18 / 205	0
Administration	36 / 956	7 / 389
Transport	6 / 55	2 / 9
Fiscal	26 / 924	11 / 552
Total	90 / 2331	20 / 950

- 1377 utterances
- Inter-Annotator Agreement (IAA) of determining valid/invalid utterances is $\kappa = 0.4998$, moderate agreement.
- IIA on the language proportion label is $\kappa = 0.6083$, substantial agreement.
- Labelled intents per sample between 1 and 4.
- Labelled entities per sample between 0 and 4.
- Average number of words per utterance in the corpus is 5.43

- Final corpus:
  - `more-es`: 313
  - `more-eu`: 662
  - `balanced`: 402
- Reference samples:
  - Spanish: 707
  - Basque: 670



Potential Uses

- **Multilingual chatbots.** Compare the capacity of different language representation models when it comes to understanding intents and entities in the case of a corpus containing Basque-Spanish code-switching.
- **Speech recognition.** Develop speech to text systems that can perform adequately when the input audio source is given in Euskañol.
- **Linguistic analysis.** Explore the most common structures, characteristics, and patterns that emerge in Basque-Spanish code switching.
- **Performance evaluator.** Evaluate how much services –like dialogue systems or speech recognisers– are degraded by the phenomenon of code-switching.
- **Language identifier evaluator**. Evaluate the performance of language detectors by using the *proportion* labels and analyse their behaviour in cases where the label is tagged as *balanced*.

Acknowledgements

The authors of this work would like to thank the volunteers who participated in the compilation phase providing their Euskañol proposals.
This project has received funding from the the Department of Economic Development and Infrastructure of the Basque Government under grant number KK-2020/00055 (EKIN).

vicomtech
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE