Information on Individual Language Pairs

For all languages, the domains of the training (“train”) data differ from that of the development (“dev”) and test sets. However, dev and test sets are taken from the same domain and use the same orthography. None of the dev and test sets are tokenized; in contrast, some of the training sets are tokenized.

Spanish–Wixarika

Development, training, and test sets belong to the same dialectic variation: Wixarika of Zoquipan, and use the same orthography. However, word boundaries are not always used with the same criteria among dev/test and train. Also, the training data [1] is a translation of the FairyTales of Hans Cristian Andersen; dev/test come from other domains. Please also note the word acquisitions from Spanish, and in some cases, also strong use of code-switching.

Translator:

- Silvino González de la Crúz

[1] Mager, M., Dionico, C., & Meza, I. The Wixarika-Spanish Parallel Corpus.

Spanish–Nahuatl

In general, Nahuatl is a language with a wide dialectal variation (around 30 variants). There is a lack of consensus regarding the orthographic standard.

The train corpus [2] has dialectal, domain, orthographic and diachronic variation (Nahuatl side). However, the majority of entries are closer to a Classical Nahuatl orthographic "standard".

The dev/test datasets were translated to modern Nahuatl. In particular, the translations belong to: Náhuatl Central/Náhuatl de la Huasteca (Hidalgo y San Luis Potosí). In order to be closer to the training corpus, an orthographic normalization was applied. Therefore, these texts are also closer to Classical Nahuatl orthography.

A simple rule based approach was used for normalization. This is based on the most predictable orthographic changes between modern varieties and Classical Nahuatl

Translators:

- Giovany Martinez Sebastián
- Pedro Kapoltitan
- José Antonio

[2] Nahuatl: Gutierrez-Vasques, X., Sierra, G., & Pompa, I. H. (2016). Axolotl: a Web Accessible Parallel Corpus for Spanish-Nahuatl. In *LREC*.

Spanish–Guaraní

The training corpus for Guaraní [3] was collected from web sources (blogs and news articles) that have a mix of dialects, from pure Guarani to more mixed Jopara which combines Guarani with Spanish neologisms. The dev and test corpora, on the other hand, are in pure Guarani.

Translator:

- Perla Alvarez Britez

[3] Guaraní: Chiruzzo, L., Amarilla, P., Ríos, A., & Lugo, G. G. (2020, May). Development of a Guarani-Spanish Parallel Corpus. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 2629-2633).

Spanish–Bribri

The training set for Bribri (spoken in southern Costa Rica) was extracted from six sources (see dataset_readme.md). These sources include a dictionary, a grammar, two language learning textbooks, one storybook and the transcribed sentences from one spoken corpus. The sentences come from three major dialects: Amubri, Coroma and Salitre.

There are numerous sources of variation in the Bribri sentences [4]: (1) There are several different orthographies, which use different diacritics for the same words (e.g. nasalization can be indicated with a line underneath the vowel, a tilde over the vowel or a Polish hook attached to the vowel). (2) The Unicode encoding of visually similar diacritics differs amongst authors. (3) There is phonetic and dialectal variation, and (4) there is considerable idiosyncratic variation between writers, including variation in word boundaries (e.g. *ikíe* vrs *i kie* "it is called").

In order to build a standardized training set, an intermediate orthography was used to make these different forms comparable and learning easier (see dataset conversion file). All of the training sentences are comparable in domain; they come from either traditional stories or language learning examples. Because of the nature of the texts, there is very little code-switching into Spanish. This is different from regular Bribri conversation, which would contain more borrowings from Spanish and more code-switching.

The dev/test texts belong to a very different domain. The dev/test sentences were translated by Francisco Morales, a speaker of the Amubri dialect. The dev/test sentences were also transformed into the intermediate orthography.

Translator:

- Francisco Morales

[4] Feldman, I., & Coto-Solano, R. (2020, December). Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3965-3976).

Spanish–Rarámuri

The training set of the Rarámuri dataset is a set of extracted phrases from Brambila's [5] dictionary in the Raramuri language. However, we could not find any description of the dialectic variation to which these examples belong. On the other hand, the dev and test set are translations from Spanish into the highlands Rarámuri. Another important aspect to consider is that the training set orthography is not the same as the dev/test sets. Please consider this to improve the performance of the system. Finally, as in many polysynthetic languages, the boundaries of a morpheme and a word are not clear and have no consensus. Native speakers, even with a standard orthography and from the same dialectical variation, have different standards to define word boundaries.

Translator:

- María del Cármen Sotelo Holguín

[5] Brambila, David. (1976) Diccionario Raramuri--Castellano (Tarahumara). Obra Nacional de la Buena Prensa, México.

Spanish–Quechua

The development and test sets are translated into the standard version of Southern Quechua, specifically the Quechua Chanka (Ayacucho, code: "quy") variety. This variety is spoken in different regions of Peru, and it can be understood in different areas of other countries, such as Bolivia or Argentina. This is the variant used on Wikipedia Quechua pages, and by Microsoft in its translations of software into Quechua. Southern Quechua includes different Quechua variants, where Quechua Cuzco (quz) and Quechua Ayacucho (quy) are very well-known ones. We provide training datasets for both variants:

1. JW300 [6] (quy, quz): Jehova's Witnesses texts, available in OPUS. We are also providing the parallel data aligned with English as an extra.
2. MINEDU (quy): Sentences extracted from the official dictionary of the Minister of Education (MINEDU) in Peru for Quechua Ayacucho.
3. Dict_misc (quy): Dictionary entries and samples collected and reviewed by Diego Huarcaya.

We encourage the participants to extract texts from different variants.

In the process of translating the development and test sets, the handling of pentavocalism from other regions, such as Collao Quechua and the apostrophe, is not used in Quechua Chanka. But normally both varieties are understood, since they share interculturality. In the

suffixes there will be some limitations. Example in Quechua of variety chanka, kachkanchik will end with K while in Quechua of variety Collao it ends with S as for example Kachkanchis.

There are different varieties of Quechua, they are Quechua Chanka, Quechua Collao, Quechua Wanka and Central Quechua. This language of variety Chanka and Collao share the lexicon, although Quechua Collao handles pantavocalism and Chanka trivocavolismo are understood correctly, while Central Quechua moves away. It is worth mentioning that these two varieties are the most widely spoken.

Translators:

- Facebook AI

[6] Agić, Ž., & Vulić, I. (2019, July). JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3204-3210).

Spanish–Aymara

The development and test sets are translated into the Central Aymara variant, specifically Aymara La Paz jilata, the largest variant. This is the similar variant of the available training set, which is obtained from Global Voices [7] (and published in OPUS). Global Voices is a news portal translated by volunteers. This means that the texts have potentially different writing styles that are not necessarily edited.

During the translation process for the development and test sets, some problems were presented in relation to terms that are not used in the Aymara language such as: missiles, bomb, artist, penny, coat of arms, enrollment rate, percentages, etc. Some have been worked with refonymisation and interpretation. Another problem encountered was the relationship of the translation with the interpretation of the context. Context is usually very important to have for accurate translations. With the lack of context, translations for some sentences may not be consistent or precise due to ambiguity in interpretation of the source texts.

On the other hand, many many neologisms are not spread to the speakers of the different regions of the Aymara population. Aymara is diverse according to different regions. Another difficulty was finding exact equivalences in the Aymara language and culture for some terms found in the Spanish source.

There are many variants in Aymara. The Aymara of Pacajes is different from Los Andes, Omasuyos, Camacho, Franz Tamayo, etc. Though there may be variations in some terminologies or even grammar, all or most Aymara dialects are mutually intelligible.

- *To leave* for example: Sarxma, sarawayxä, sarxä, etc.
- *Kid* for example: wawa, yaya, yuqallwawa, lala, etc.

Translators:

- Facebook AI

[7] Tiedemann, J. (2012, May). Parallel Data, Tools and Interfaces in OPUS. In LREC (Vol. 2012, pp. 2214-2218).

Spanish–Shipibo-Konibo

The training sets for Shipibo-Konibo [8] have been obtained from different sources and translators:

1. Flashcards: A sample from the Tatoeba dataset translated into Shipibo-Konibo
2. Educational: Translated sentences from books for bilingual education
3. Dictionary: SIL dictionary entries and examples

The first two sources are translated by a bilingual teacher, and follow the most recent guidelines of the Minister of Education in Peru (similar to the development and test sets). However, the third source is an extraction of parallel sentences from an old dictionary.

There are many neologisms that are not spread to the speakers of different communities. The translator of the development and test sets only translated the words and concepts that are well known in the communities. Other terms are preserved as in Spanish or English.

Translator:

- Liz Chávez

[8] Galarreta, A. P., Melgar, A., & Oncevay-Marcos, A. (2017, September). Corpus Creation and Initial SMT Experiments between Spanish and Shipibo-konibo. In RANLP (pp. 238-244).

Spanish–Asháninka

Three things are important to know about the texts that are included in the training data [9, 10, 11]:

1. The texts belong to domains such as: traditional stories, educational texts, environmental laws for the Amazonian region.
2. Not all the texts are translated into Spanish, there is a small fraction of these that are translated into Portuguese because a dialect of pan-Ashaninka is also spoken in the state of Acre in Brazil.
3. The texts come from different pan-Ashaninka dialects and have been normalized using the **AshMorph** tool mentioned in the article below.

There are many neologisms that are not spread to the speakers of different communities. The translator of the development and test sets only translated the words and concepts that are well known in the communities. Other terms are preserved in Spanish.

Translator:

- Feliciano Torres Ríos

[9] Ortega, J., Castro-Mamani, R. A., & Samame, J. R. M. (2020, December). Overcoming Resistance: The Normalization of an Amazonian Tribal Language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages* (pp. 1-13).
[10] Romano, R & Richer S. (2008) Ñaantsipeta asháninkaki birakochaki. Diccionario Asháninka-Castellano. Versión preliminar, www.lengamer.org/publicaciones/diccionarios/
[11] Mihas, E. (2011) Añaani katonkosatzi parenini, El idioma del alto Perené. Milwaukee, WI: Clarks Graphics.

Spanish–Hñähñu

The Hñähñu training [12] set was collected from a set of different sources, so this implies that the text contains more than one dialectal variation and orthographic standard. However, most texts belong to the Valle del Mezquital dialect. This can lead to challenges, since the development and test sets are from the Ñûhmû de Ixtenco, Tlaxcala, variant. This also leads to different orthographic conventions. This variant is especially endangered as less than 100 elders still speak the variant.

Translator:

- José Mateo Lino Cajero Velázquez

[12] Hñähñu online corpus: https://tsunkua.elotl.mx/about/