**Dialogs Re-enacted Across Languages: Multiple-References Supplement (DRAL-MRS)**

University of Texas at El Paso

Vanessa Bolado Garcia, Nigel Ward

May 2025

=== Overview

The DRAL corpus is a bilingual speech dataset consisting of parallel Spanish-English utterances, and this supplement, DRAL-MRS, includes alternative re-enactments of the original utterances.

=== Background

DRAL is a bilingual speech corpus of parallel utterances, using recorded conversations and fragments re-enacted in a different language. It is intended as a resource for research, especially for training and evaluating speech-to-speech translation models and systems.

The original releases include 2893 short matched Spanish-English pairs (> 2 hours) taken from 104 conversations with 70 unique participants. All are packaged together with the full original conversations and full re-enactment recording sessions.

The original DRAL data collection has only one re-enacted version of each utterance. However there can be more than one way to validly translate an utterance, so we decided to augment it with additional versions.

We accordingly set out to collect an alternative translation for each of the short utterances in the DRAL corpus. The new release includes the original 2893 short Spanish-English pairs, each re-created by another person who wasn’t in the original conversation. The release includes the re-enactment audio, paired with the annotated Elan file for the utterances. We gathered this data by guiding 12 different participants to re-create the original utterances as closely as possible in the other language.

=== Motivation

Our specific motivation is to support meaningful evaluations of system performance, in two ways. First, to evaluate the output of a system, rather than measure how closely it matches the single reference, we could measure the match to all the references, and use the minimum distance as the true estimate of system performance. Second, to estimate the upper bound on system performance, we could measure inter-annotator agreement.

=== Participants

For the re-enactments we relied on four lab members and eight external participants, all Spanish-English bilinguals.

Our re-enactors were varied male and female participants all around the ages of 19-25. Occasionally the original speech seemed gendered, such that a different-gender speaker felt awkward producing the utterance or felt that it was not highly equivalent. We made an effort to have more gender-matched recordings: if we had both a male and a female reenactor, and both a male and a female in the original conversation, they each re-enacted the same-gender speaker. Otherwise we didn't worry about it. To support modeling or evaluation conditioned on gender, the metadata for the reenactments includes a gender field, with four possible values, MM, FF, MF, FM, indicating the gender of the original speaker and the gender of the re-enactor. We also included a field for the participants' initials, to help us keep better track of their gender and information during the post processing phase, and because we decided to use their initials for our naming conventions. If two participants had the same set of initials, we would add a third identifying letter(e.g. the last letter of their last name or the first letter of their second last name if they had one). In addition, we considered adding a character A for the code in cases where the re-enactor felt awkward in naturally translating an other-gender utterance, or felt that in doing so they had to leave out a lot of the nuances, but in fact this was never really felt necessary.

=== Collection Strategy

We considered two possible ways to collect these alternative translations. One possibility would be to have the original speakers produce alternatives, prompting them to ``now do it again, but with different words or in a slightly different style.'' The other possibility would be to have a third-party speaker produce the translations after the fact. We chose the latter, for three reasons: it could provide higher throughput, is more natural, and doesn't require us to track down the original speakers.

== Recording Steps

1. Set up the audio recording equipment in the lab in a quiet area, such that the microphones are away from any background noise like computer fans. The recommended location is the room on the right of the lab. You’ll need the digital recorder (TASCAM DR-40) and two single sided headsets and microphones (Shure BRH441M). Both should be found on the shelves next to the front desk at the lab. Be sure to keep the microphone away.

2. If you need further information about how to connect these devices, refer to the ***Dialogs Re-enacted Across Languages, Version 2*** technical report.

3. Recruit and schedule participants to come in. While in theory this could be done by a single participant speaking to themself, in fact it’s much better to have two speakers involved in re-enacting the recordings. When this is hard to schedule, it’s possible for you, to serve as a re-enactor while simultaneously performing the producer’s functions.

4. Have the volunteer(s) come into the lab and sit down where you have chosen to set up. Have them read and sign the consent forms. Introduce everyone to the equipment that you will be using and provide them with the headsets.

5. The producer will open the original conversation in Elan. Per utterance, go back 10 seconds into the conversation to gather context for re-enactment.

6. Start recording and start replicating the utterances, from before the utterance happens, to acquire the most natural result. In the original annotations there are “long” and “short” utterances, where short utterances are the split up version of the long ones. We chose to replicate long utterances as they were easier to re-create naturally, with separation of the long utterances into short ones done in the post-processing phase.

7. The utterance should be replicated in the opposing language (if the original conversation is in English, it should be re-created in Spanish and vice versa). Remember to specify that they should try and re-create not only the words as closely as possible, but also the feeling, prosody, meaning, etc.

8. Have everyone evaluate, namely the producer and re-enactor(s), evaluate each re-creation. Everyone should discuss whether the re-creation was good or not, and if the decision is negative, try again until everyone is happy with the result. If after 5 or 6 tries the re-creation isn’t satisfactory, you may skip that re-creation and try it again later on.

9. Repeat the procedure for as many utterances as desired.

10. When you are finished, upload the recordings to your computer for storage using the cable included in the recorder’s box and connecting it to your computer. Store away the equipment.

=== Post-processing

1. Create a datasheet to store participant information, including gender of both speakers, audio re-created, and gender of original pair of speakers.

2. Open each file on ELAN and annotate each utterance. Remember to use the same numbers and tiers as the original conversation.

2. Follow steps from section 4.6 of the technical report ***Dialogs Re-enacted Across Languages, Version 2*** if needed in order to make a release with each utterance as an individual file.

File naming conventions:

*[original language]\_[original Conversation Number]\_[Reenactor 1 Initials]\_[Reenactor 2 Initials]*

**Ex. EN\_011\_VB\_NW**

=== Data Release

* Temporarily at nigelward.com/dral-mrs.tgz
* About 5GB.
* The data release should be phrase pairs, re-enactment pairs, short fragment audios, track, a copy of the complete original and re-enacted audio recordings, and updated metadata.

=== Observations

During the data collection we noticed the following:

* Each set of utterances, originally derived from one 10-minute conversation and consisting of approximately 20–40 utterances, took about 20 minutes to re-enact, with an additional 20 minutes required for post-processing. With a total of 104 recordings, the entire data recollection procedure required roughly 70 hours: approximately 35 hours for recording and 35 hours for post-processing.
* Sometimes utterances were simply too long and the re-enactors struggled to re-create them in one go. In these cases, we would split up the long utterance, which was typically already split up into several short utterances, and have them re-create each short utterance. During the post-processing annotations, we would add a decimal to the long utterance number to represent that the utterances belonged together to one original long one.
* The hardest utterances to re-create included laughter. We believe that this expression can never fully be re-created naturally, and the re-creations always sounded a little off.
* Context is needed for a natural re-creation result. Many prosodic expressions included in the utterance related to prior context in the conversation. Hearing the context before replicating helps create the setting and replicate feelings the best.
* Trying to replicate a conversation by oneself may feel odd and unnatural to the person, thus having someone else present when re-creating utterances to replicate the original conversation improves the flow and overall quality.
* A producer present during recordings is vital in order to operate equipment and let the re-enactors focus on producing the best audio possible.
* During the post-processing phase, we annotated the recordings using ELAN. This involved identifying the exact span where the re-enacted utterance occurred. We found the process to be significantly more efficient when leveraging the spectrogram, which allowed us to visually distinguish the speaker audio. Since the microphones captured both the playback of the original utterance and the participant’s re-enactments, the visual display of the former conveniently always louder and thus darker, served as a reliable cue for where to jump to, since participants typically began their re-enactment immediately afterward.
* When we had a new participant come in, they would often seem to feel unsure at first whether they were doing it right and look at the producer for approval and add a rising final intonation. Because of this, we found it better to have repeat participants rather than getting a huge variety of people, in order to save time.

=== References

https://www.cs.utep.edu/nigel/dral/

LDC2024S08, Linguistic Data Consortium

Technical Report UTEP-CS-23-27 Dialogs Re-enacted Across Languages, Version 2. Nigel G. Ward, Jonathan E. Avila, Emilia Rivas, Divette Marco. 2023