ES-Port: a Spontaneous Spoken Human-Human Technical Support Corpus for Dialogue Research in Spanish

Laura García-Sardiña, Manex Serras, Arantza del Pozo
Speech and Natural Language Technologies, Vicomtech
Mikeletegi Pasealekua 57, 20009 Donostia-San Sebastián
{lgarcias, mserras, adelpozo}@vicomtech.org

Abstract

In this paper the ES-Port corpus is presented. ES-Port is a spontaneous spoken human-human dialogue corpus in Spanish that consists of 1170 dialogues from calls to the technical support department of a telecommunications provider. This paper describes its compilation process, from the transcription of the raw audio to the anonymisation of the sensitive data contained in the transcriptions. Because the anonymisation process was carried out through substitution by entities of the same type, coherence and readability are kept within the anonymised dialogues. In the resulting corpus, the replacements of the anonymised entities are labelled with their corresponding categories. In addition, the corpus is annotated with acoustic-related extralinguistic events such as background noise or laughter and linguistic phenomena such as false starts, use of filler words or code switching. The ES-Port corpus is now publicly available through the META-SHARE repository, with the main objective of promoting further research into more open domain data-driven dialogue systems in Spanish.

Keywords: spontaneous dialogue corpus, human-human dialogue, technical support, transcription, anonymisation, named entities

1. Introduction

Dialogue systems, often referred to as conversational agents or chatbots, are becoming increasingly popular as they allow users to directly interact with a wide range of information systems in a natural way. Customer support is an application scenario with a strong interest in dialogue system development, driven by the promise of intelligent digital assistants available 24x7 to resolve customer requests in a fast, cost-effective and consistent manner (Guzmán and Pathania, 2016).

Data-driven approaches to dialogue system development have shown to be more robust than rule-based techniques to variability in user behaviour, the performance of speech and language processing subcomponents and the dynamics of the task domain (Meena, 2015). Despite their promising results in recent years (Young et al., 2013; Wen et al., 2016; Li et al., 2017), most practical dialogue systems are still built by human experts through significant manual engineering. In most currently deployed systems, rule-based dialogue managers (DM) are combined with statistical natural language understanding (NLU) models capable of classifying intents and their related entities (Williams et al., 2015). In the customer support domain, this limits their application to frequent use cases in specific areas where solutions are well known, predictable and where scripted answers can be developed (Guzmán and Pathania, 2016).

Lack of annotated corpora is the main problem for the development of data-driven systems (Serban et al., 2015). To overcome this issue, it is usual to develop rule-based baselines or to employ the Wizard of Oz (WOZ) technique (Benedı et al., 2006; Rieser and Lemon, 2008) in which a human mimics the intended dialogue system, in order to gather interactions with real users. Although this kind of human-machine data is constrained by the employed baseline systems or the scenarios defined in the followed WOZ approaches, it is useful to bootstrap goal-driven dialogue systems whose policies can be optimised through user simulation and adaptive learning (Schatzmann et al., 2007; Gašić et al., 2013; Serras et al., 2017). On the other hand, human-human dialogue corpora contain unconstrained and unscripted natural dialogue interactions exhibiting traits different from human-machine dialogue (i.e. richer turn-taking and more common grounding phenomena) (Doran et al., 2003), which are more suitable to train more open domain dialogue systems. Human-human customer support corpora would allow progress towards the development of large-scale data-driven dialogue systems capable of handling a wider amount of customer queries.

A considerable amount of corpora are available for building data-driven dialogue systems (Serban et al., 2015). Unfortunately, because customer support interactions occur in commercial settings, most customer support datasets are proprietary and not released to the public due to privacy and data protection reasons. In practice, there are only a couple of publicly available technical support datasets (Lowe et al., 2015; Uthus and Aha, 2013) derived from the Ubuntu IRC channels[1], used to receive technical support for issues related to the Linux-based operating system. Although some data is available from the support channels in other languages, most of the compiled resources are in English.

In this paper, the Spanish Technical Support (ES-Port) Corpus is presented, a compilation of spontaneous spoken human-human dialogues from the technical customer support service of a Spanish telecom operator for companies. The corpus has been directly transcribed from call recordings, annotated at various linguistic and acoustic-related extralinguistic levels, and anonymised in order to comply with data protection legislation. Its release is intended to foster further research into more open domain data-driven di-

[1] These logs are available from 2004 to 2018 at http: //irclogs.ubuntu.com/

alogue systems in Spanish, capable of achieving more natural interactions in the technical support domain.

2. Compilation Process

The raw corpus was provided by an independent telecom operator, dedicated to providing tailor made cloud data centre, fixed voice, IP or mobile telephony and Internet connectivity solutions to companies. In order to serve their clients, they offer 24/7 customer support: 24 hours a day, 7 days a week, 365 days a year. Despite having a multichannel customer service and also providing support through web forms and email, the majority of the clients still prefer calling. Thus, the corpus provided consisted of raw audio recordings of such calls.

2.1. Transcribing the Audio

The first step of the corpus compilation process involved transcribing the provided raw audio data. Details regarding the characteristics of the audio and the followed transcription process are given next.

2.1.1. Audio characteristics

The recorded calls contain both speech and other background sounds, such as channel-associated noises or background music. The type of speech used is spontaneous, and so it includes phenomena such as false starts, mispronunciations, non standard forms, overlapping segments between speakers, unfinished sentences and, in general, speech more focused on conveying the message than on taking care of its form. The audio data consisted of a total of 40 hours , with an average length of 2 minutes per dialogue.

2.1.2. Transcription process

The provided recordings were transcribed using the Transcriber 1.5 annotation tool (Barras et al., 2001). In addition to the orthographic transcriptions, the following phenomena were also annotated:

- speaker turns
- non-speech events (e.g. coughing and laughter) and background acoustic conditions (e.g. noise and music)
- overlapping speech
- false starts, repetitions, unfinished words and non-words
- mispronunciations, lengthening in pronunciation, and typical spoken Spanish shortening of words (e.g. *pa* instead of *para*) or dropping of intervocalic *d* in final syllables (e.g. *demasiao** for *demasiado*, *entrao** for *entrado*)
- continuers and filler words (e.g. *o sea*, *eh*, *hala*, *mhm*, etc.)
- words in a language other than Spanish (when pronounced correctly)

Given the more challenging spontaneous and telephone nature of the data, attempts to follow incremental automation methodologies such as those described in (Pozo et al., 2014) to make the transcription process more productive were not feasible. The word error rates (WER) of generic large vocabulary continuous speech recognition (LVCSR) systems turned out too high to provide any time savings (77.21% in test set).

In the end, the transcription process was carried out fully manually and took a linguist six months working full-time to complete.

2.2. Anonymising the Dialogues

In order to comply with the European data protection legislation (Art29WP, 2014) and not to compromise the right to confidentiality of the individuals involved, the personal information contained in any dataset must be neutralised before releasing the data open to the public in order to be exploited for other purposes.

Data anonymisation is the process of treating personal data in such a way that it can no longer be used to identify the individuals involved, while preserving the value and usefulness of the original format.

2.2.1. Anonymisation practices and standards

Despite European legislation does not prescribe any particular anonymisation technique, randomisation and generalisation approaches are usually employed to anonymise structured datasets in the form of tables or graphs:

- Randomization: involves alteration of the data without losing its value and includes techniques such as noise addition and permutation.
- Generalisation: implies diluting or reducing the granularity of the data and comprises techniques such as aggregation and K-anonymity.

For unstructured text, such as the transcriptions of the technical support recordings in the ES-Port corpus, the following methods have also been proposed in (Dias, 2016):

- Suppression: the element to be anonymised is replaced by some neutral indicator, e.g. ’XXXXX’.
- Tagging: the element to be anonymised is replaced by a label which can refer to its class or identifier, e.g. ’ORGANISATION123’.
- Substitution: the entity to be anonymised is substituted by another entity, e.g. ’Juan’ for ’Pedro’. The choice of the new entity can be random from a dictionary, swapped with another entity within the document, ’intelligently’ substituted by an entity sharing the same features, or applying a generalisation technique to the item (e.g. replacing ’University of the Basque Country’ by ’University’).

The technique chosen to anonymise the ES-Port corpus was substitution because readability and coherence are kept and the result is a natural anonymised text. The process is described in Section 2.2.3..

Table 1: NER and NERC Precision (Pr), Recall (Rc), and F1 scores for the three taggers on our test set.

	NER			NERC		
	Pr	Rc	F1	Pr	Rc	F1
IXA Pipes	0.32	0.62	0.42	0.26	0.50	0.34
FreeLing	0.36	0.65	0.47	0.24	0.54	0.33
CoreNLP	0.47	0.96	0.63	0.36	0.99	0.53

Table 2: Entity tags used in the anonymisation process.

Utterance	Replacements	Tags
”Soy Bárbara de Cincode”	Bárbara	female_name
	Cincode	organisation
”Arturo Noriega arroba Hotmail punto es, tengo que poner?”	Arturo	male_name
	Noriega	surname
	Hotmail	mail
”te la digo, es M de Madrid,”	M	letter
	Madrid	place
”el último registro es del veintisiete de septiembre.”	veintisiete	number
	septiembre	month
”Inexistent punto com.”	Inexistent	domain
”tiene que entrar= a CompDNS”	CompDNS	product/ service

2.2.2. Identifying the features to be anonymised

The first step in the anonymisation process is to identify the type of elements in the dataset that refer to personal information or that could possibly be used in any way to identify the people involved, endangering their right to confidentiality. Considering the nature of the information given in the ES-Port corpus, we decided to anonymise the following types of elements:

- Elements referring to individuals' basic personal information: names, surnames, name diminutives or nicknames, personal identification numbers.
- Contact information and digital trace elements: phone numbers[2], IP addresses, user names and numbers, email addresses, postal addresses, web domains.
- Workplace and organisation-related elements: names of organisations, NIFs (tax identification number) and CIFs (tax code), easily linkable names of products and services, prices.
- Other elements: card numbers and bank accounts, dates, locations, trouble ticket numbers, dispatch notes, passwords, spellings of any of the previous elements.

2.2.3. Anonymisation process

Once the types of elements that needed to be anonymised were identified, the anonymisation process was carried out in a semi-automatic way.

First, the items to be anonymised were selected and categorised. We tried to automate the selection process by using different Named Entity Recognition and Classification (NERC) tools available for the Spanish language, namely IXA Pipes (Agerri et al., 2014), FreeLing (Carreras et al., 2004) and Stanford CoreNLP (Manning et al., 2014). Although these taggers have reported good results on planned written language such as news texts, trial tests on a small dataset of our spontaneous spoken technical support corpus were too poor to automate the process of selecting and categorising the items to be anonymised, as their use would still require considerable manual revision and correction. Results for the three taggers on the test set both considering entity recognition alone (NER) and entity classification as well (NERC) are presented in Table 1. In addition, none of the taggers covered all types of items that had to be anonymised, as is the case of numbers, months and individual letters in spellings.

The next step was automatic and consisted in randomly substituting the selected items for an element of the same characteristics according to its type (organisation, number, male/female name, etc.). Once an item was anonymised, its substitution was kept throughout the whole dialogue in order to maintain coherence, but not across dialogues so as to prevent possible linkability issues. New names provided for organisations and domains are made up and did not correspond to any existing entity at the time the anonymisation was carried out. The final step involved manual revision of the results and correction of coherence errors (e.g. non matching spellings).

The named entity categories of the elements anonymised following the process described above have been kept in the compiled corpus. As a result, ES-Port also includes named entity annotations. However, these were anonymisation-oriented and therefore the number of classes and specificity of the tags are more granular than in the typical NERC approaches. Nevertheless, the tags used are easily generalisable to the usual four (PERSON, LOCATION, ORGANISATION, MISCELANEA) or six (plus NUMBER and DATE) NERC classes. The tags used and real examples of their usage are shown in Table 2.

Overall, the described anonymisation process took a linguist eight months working part-time to complete.

3. The Compiled Corpus

This section describes the gathered corpus quantitatively and qualitatively.

3.1. The corpus in numbers

Table 3 summarizes the basic statistics of the ES-Port corpus regarding its number of dialogues, turns and overlaps, its vocabulary and its amount of filler and foreign words.

Table 3: Corpus Characteristics

Num. Dialogues	1170	Vocabulary size	11221
Num. Turns	65239	Labelled Filler Words	37
Avg. Turns per Dialogue	55.76	Filler Words Freq.	26574
Avg. Turn Length	8.20	Foreign Words Freq.	3294
Num. Overlap Turns	11329	English Words Freq.	3017

As can be seen, the corpus presents attributes typical of spontaneous spoken human-human interaction such as overlapping turns (around 17% of the turns) and rich use of

[2] Some prefixes were kept if relevant to the conversation.

Table 4: Excerpt from a dialogue, including turn index (T), speakers (S) and the actual annotated and anonymised utterance transcription (U).

T	S	U
29	spk1 spk2	De todas formas esto
30	spk1	si has +enviado el correo estate tranquilo porque <se=> se para.
31	spk2 spk2	(*EVENT*: noise-rire) <%mm> Es <lom-> <i-> incluso si lo <envi-> <%aver> <su-> supuestamente hasta las cinco y media, no?
32	spk1	Sí.
33	spk2 spk2	(*EVENT*: noise-rire) Y si lo envío a las cinco y diez se cancela?
34	spk1	Sí, sí.
35	spk2	Ay, dios (*EVENT*: pronounce-ch)
36	spk1	<%mhm>
37	spk2 spk2	Ay, mi madre (*EVENT*: pronounce-ch) no puedo largarme de aquí digamos.
38	spk1 spk2	<%eh> si quieres <%eh>
39	spk1	llamar un poquillo más tarde y te intento pasar con él de nuevo.
40	spk2	Es que no hay ninguna forma de que, ningún número que yo pueda=
41	spk1 spk2	No, no. llamarles
42	spk2	a ellos o
43	spk1 spk2	No, <%osea> algo?
44	spk1	que le estoy llamando yo y no me responde.
45	spk2 spk2 spk2 spk2	(*EVENT*: noise-nontrans) <ueh-> okay (*LANGUAGE*: en) <%pues> muchas gracias.
46	spk1	<%venga> a ti.

continuers and filler words (approximately 5% of total word occurrences). It is interesting to note that since the corpus is gathered from an IT domain, English foreign words are quite common, reaching up to 91.59% of all foreign words occurrences and constituting around 3.24% of the vocabulary.

3.2. Sample data description

In order to give an idea of the type of information and phenomena present in the ES-Port corpus, Table 4 shows an excerpt of one of its dialogues.

Different types of speech and non-speech events and background acoustic conditions occur often along the corpus. We find instances in turns 35 and 37, where the event tags indicate that the utterance was whispered, and in turn 45, indicating that the speech was unintelligible and could not be transcribed. Other instances can be found in turns 31 and 33, where the speaker laughs before continuing talking.

Overlapping speech is also common. Examples can be seen in turns 29, 41, and 43, where two different speakers intervene within the same turn. Speech, extralinguistic phenomena, or a combination of both can be overlapped.

False starts, repetitions, incomplete words, and nonwords are very common. False starts and repetitions are tagged simply using `<word>`, as in turn 30, while incomplete words and other nonwords are tagged as `<nonword->` and can be found in turns 31 and 45 of the excerpt.

Deviations from the standard pronunciation take place regularly in the corpus. An instance dropping the intervocalic phoneme /d/ in a final syllable can be found in turn 30, marked with a + symbol. Lengthening examples appear in turns 30 and 40, marked with an = symbol.

The words tagged following the pattern `<%word>` correspond to a set of 37 filler words (e.g. *o sea*, 'I mean') and continuers (e.g. *mhm*) frequently used in the corpus. Some instances of these can be found in turns 31, 36, 38, 43, 45, and 46.

Finally, words from a language other than Spanish appear quite often, especially in English. The language event in turn 45 indicates that a foreign word was used, in this case from English.

3.3. Potential Applications

The ES-Port corpus is a source of annotated spontaneous spoken human-human dialogues which may be valuable for several research tasks and applications. In this section, a few of them are mentioned.

Our main objective is to promote more open and natural dialogue interactions in Spanish customer support. Although it does not yet include dialogue act annotations, the corpus as is could be used to explore unsupervised approaches to dialogue system development in Spanish. These approaches include modelling the language of the system to generate more human-like prompts, modelling the language of the user to better detect the nuances of human-human communication or analysing the turn-taking dynamics for incremental dialogue processing.

The corpus annotated at the current level can also be exploited to develop supervised approaches for spontaneous LVCSR in Spanish (including automatic capitalization and punctuation) or to develop NERC tools that work better on dialogue text.

Finally, linguistic research in Spanish may use this data to study a wide range of issues, such as the use of discourse markers and filler words in conversation, their meaning in context, and how they influence the dialogue, or the strategies used for turn-taking and self-correction, among others. Other interesting phenomena for study are code switching and the abundant use of words from the English language in the IT domain.

3.4. Data sharing

The current version of the ES-Port dialogue corpus is available via META-SHARE[3] in the repository of the University

[3] The raw audio corpus cannot be released for public access, since it contains sensitive data which falls under the European General Data Protection Regulation (GDPR)

of the Basque Country UPV/EHU[4] under the name of ES-PORT.

4. Conclusions and Future Work

A spontaneous spoken human-human technical support dialogue corpus in Spanish has been transcribed and anonymised. At this point, the corpus contains annotations referring to linguistic and acoustic-related extralinguistic phenomena such as music, laughter, use of filler words and code switching in conversation. Named entities anonymised using the substitution technique are also annotated. The ES-Port corpus is now publicly released so it can be used for dialogue research or the adaptation of LVCSR and NERC systems to spontaneous dialogue. Future work includes dialogue act annotation of the corpus.

5. Acknowledgements

The authors would like to thank the telecom operator who has kindly provided the raw technical support call recordings. We would also like to thank Aholab, the Signal Processing Laboratory of the University of the Basque Country, for granting us access to their META-SHARE node to share our corpus.

6. Bibliographical References

Agerri, R., Bermudez, J., and Rigau, G. (2014). Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *LREC*, volume 2014, pages 3823–3828.

Art29WP. (2014). Article 29 data protection working party: Opinion 05/2014 on anonymisation techniques.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1):5–22.

Benedı, J.-M., Lleida, E., Varona, A., Castro, M.-J., Galiano, I., Justo, R., López, I., and Miguel, A. (2006). Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1636–1639.

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). Freeling: An open-source suite of language analyzers. In *LREC*, pages 239–242.

Dias, F. M. C. (2016). Multilingual automated text anonymization.

Doran, C., Aberdeen, J., Damianos, L., and Hirschman, L., (2003). *Comparing Several Aspects of Human-Computer and Human-Human Dialogues*, pages 133–159. Springer Netherlands, Dordrecht.

Gašić, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., Tsiakoulis, P., and Young, S. (2013). On-line policy optimisation of bayesian spoken dialogue systems via human interaction. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8367–8371. IEEE.

Guzmán, I. and Pathania, A. (2016). Chatbots in customer service. *Accenture Interactive*.

Li, X., Chen, Y.-N., Li, L., and Gao, J. (2017). End-to-End Task-Completion Neural Dialogue Systems. *ArXiv e-prints*, March.

Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *CoRR*, abs/1506.08909.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Meena, R. (2015). *Data-driven Methods for Spoken Dialogue Systems*. Ph.D. thesis, KTH, Royal Institute of Thechnology.

Pozo, A. D., Aliprandi, C., Álvarez, A., Mendes, C., Neto, J. P., Paulo, S., Piccinini, N., and Raffaelli, M. (2014). Savas: Collecting, annotating and sharing audiovisual language resources for automatic subtitling. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May.

Rieser, V. and Lemon, O. (2008). Learning effective multimodal dialogue strategies from wizard-of-oz data: Bootstrapping and evaluation. In *ACL*, pages 638–646.

Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H., and Young, S. (2007). Agenda-based user simulation for bootstrapping a pomdp dialogue system. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 149–152. Association for Computational Linguistics.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2015). A survey of available corpora for building data-driven dialogue systems. *CoRR*, abs/1512.05742.

Serras, M., Torres, M. I., and Del Pozo, A., (2017). *Online Learning of Attributed Bi-Automata for Dialogue Management in Spoken Dialogue Systems*, pages 22–31. Springer International Publishing, Cham.

Uthus, D. and Aha, D. (2013). The ubuntu chat corpus for multiparticipant chat analysis.

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-Barahona, L. M., Su, P.-H., Ultes, S., and Young, S. (2016). A Network-based End-to-End Trainable Task-oriented Dialogue System. *ArXiv e-prints*, April.

Williams, J., Kamal, E., Ashour, M., Amr, H., Miller, J., and Zweig, G. (2015). Fast and easy language understanding for dialog systems with microsoft language understanding intelligent service (luis). In *Proceedings of 2015 SIGDIAL Conference, Prague*. ACL – Association for Computational Linguistics, September.

Young, S., Gasic, M., Thomson, B., and Williams, J. D. (2013). Pomdp-based statistical spoken dialog systems: A review. In *Proceedings of the IEEE*, volume 101(5), pages 1160–1179.

[4] http://aholab.ehu.es/metashare/repository/search/