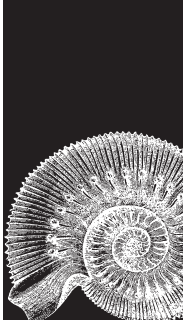

JMR

Journal of Maritime Research, Vol. VIII. No. 2, pp. 13-28, 2011

ISSN: 1697-4840

ECOMARS: SPEECH DATABASE FOR EXTERNAL COMMUNICATIONS IN MARITIME SETTINGS

R.M. de la Campa[1,2] and **B.A. Rodríguez**[1,3]

Received 13 October 2010; in revised form 20 October 2010; accepted 7 May 2011

Abstract

Results obtained from different international research projects clearly show the feasibility of implementing simultaneous automatic translation systems to key areas when there is limited content. One of these controlled areas, in regards to vocabulary and syntactic structure, is that of routine oral conversations that take place in maritime settings between ships or between ships and shore services through the use of Standard Marine Communication Phrases.

Hence, the University of A Coruña has conducted a study titled "Language industries applied to oral communications in maritime settings" which had as its main goal to study the legal and technical possibilities and the commercial suitability of the development of an automatic language translator for oral communications in that area.

Among the many activities to be developed in this project, is the compilation of actual oral samples of such communications; samples that have given rise to the emergence of the speech database ECOMARS.

ECOMARS is the first and unique spontaneous speech database for external communications in maritime settings in Spanish and English that has been compiled and dealt with in accordance with the needs stated in the aforementioned project. The specifications corresponding to its design, realization and treatment will be broadly discussed in this article.

Key words: speech database, maritime communications, controlled languages, speech translation.

[1] ETSNM University of A Coruña, Paseo de Ronda 51, 15011 A Coruña, Spain. [2] Professor, Email: rosamary@udc.es, Tel. 981167000- 4252, Fax. 981167001. [3]Professor, Email: benigno@udc.es, Tel. 981167000- 4208, Fax. 981167001.

INRODUCTION

As it is widely known, the main goal of speaking automatic translation is to allow conversations among and between diverse peoples who only speak their mother tongue. Automatic translations thus position the speaking technologies as mediators between two people, exceeding the traditional role of interface between users and machinery (Mariño, 2002).

In general, we can talk about two approaches for the realization of automatic speech translation: the classical approach and the statistical approach. (Casacuberta, 2001, Casacuberta and Vidal, 2006; Llisterri, 2004).The latter, which is newer and nowadays more often used, is also known as an integrated and consistent approach in the simultaneous realization of voice recognition and automatic translation. Regarding this approach, Casacubierta and Vidal (2006) state, "the acoustic models of words that are part of a speaking recognition system are integrated in the translation model." From a statistical point of view, the goal of the translation system is to generate the most probable replies to a given phrase. The most significant advantages of this approach are the capacity of recognition and the simultaneous and homogeneous translation, and the possibility to automatically generate and – based on examples – sources of acoustic, lexical and translation knowledge (Casacubierta, 2001; Casacubierta and Vidal, 2006.).

The main problems associated with this system are:

- They can only be used in specific and restricted areas.
- Vocabulary must be restricted to a range of 5000 to 10000 words.
- Great volumes of learning data or training corpus are needed and its compilation is very expensive.
- The size of the models obtained also constitutes a problem, which, once more, make it necessary to restrict the application environment.
- Even the systems and approaches that are working most efficiently have had extremely high error rates.
- That they are based on language pairs, that is, different translators must be compiled for each language pair. This implies the extra difficulty of composing an aligned training corpus for both languages (Mariño, 2002.).

On the other hand it seems logical to assume that systems of voice recognition will deteriorate under severe environment conditions, such as noise, the use of variable channels such as telephones and radiotelephones, and the usual vocal problems in human speakers such as hoarseness or impaired pronunciation due to a cold.

In this sense, and in order to improve the stability and accuracy in voice recognition systems, several techniques are being researched for the purpose of reducing disturbing effects due to noise. Villarubia (2004) indicates that the techniques currently being employed in this area are the spectral subtraction of noise, the standardization of durations, and the extraction of robust features of AURORA.

For his part, Mariño (2002) proposes the training of Markov's hidden models based on a database that includes the highest number of possible situations where the recognizing party will be able to find an actual application as a way of resolving the problem of attaining robustness in the system. This implies the availability of a large database, which includes speaking samples from different interlocutors and communication channels.

AUTOMATIC TRANSLATION APPLIED TO MARITIME SETTINGS

Results obtained from different projects such as Verbmovil, EuTrans or ALIADO clearly show the feasibility in the implementation of simultaneous automatic translation systems for specific tasks whose discourse commands are limited. One of these controlled areas, specifically with regards to vocabulary and syntactic structure, is that of routine oral conversations that take place in maritime settings between ships or between ships and shore services through the use of Standard Marine Communication Phrases.

Therefore all these processes, previously studied, may be applied, to a certain extent, to maritime communications in order to reduce the problems arising from the multiple languages spoken in the environment, by and large in regards to those communications that are made through radio devices, where the introduction of an automatic translator would permit, for example, that two people of different nationalities may communicate with each other using their mother tongue. This system could be very useful when the messages emitted are standardized, thus, assuring a quality translation and a perfect understanding.

On the other hand and given the increasing prevalence of automated systems on board ships, such equipment could help to reduce bridge officers' workload (Hanz-Pazara et al 2008; Meck, Strohschneider and Brüggermann, 2009).

In order to perform this task, three types of corpus – that would generate the respective models – would be necessary:

- An acoustic corpus that could be obtained by the recording of actual conversations on board ships.
- A grammatical corpus or set of grammatical rules that allows the recognition of the source phrases and the construction of phrases translated into the target language. These grammatical rules could be extracted from the same structure used by the Standard Marine Communication Phrases.
- And thirdly, vocabulary, whose base would be the contents of these standardized phrases.

For this reason, the University of A Coruña is conducting a study called "Language industries applied to oral communications in maritime settings" whose main goal is to study the legal and technical possibilities and the commercial suitability of the development of an automatic translator for oral communications in that area

and, as the case may be, to set the basis for its development and further study the implications that such a device would have in maritime safety.

Among the activities to be developed in this project is the compilation of actual oral samples of such communications, samples that has given rise to the construction of the speech data base ECOMARS that will be described below.

ECOMARS: DATABASE PROJECT FOR EXTERNAL COMMUNICATIONS IN MARITIME SETTINGS

ECOMARS is the first and unique oral database of spontaneous speaking correspondence for external communications in maritime settings in Spanish and English that has been compiled and dealt with in accordance with the needs stated in the aforementioned project. The specifications corresponding to its design, realization and treatment are detailed below.

Basic design specifications

Taking into account that one of the goals of this project "Language industries applied to oral communications in maritime settings", which also serves as a support for the realization of other goals that are no less significant, is the obtaining and analysis of actual oral samples in communications ship to ship and ship to shore, the need for the design and creation of the database or oral corpus arose and its main uses would be the following:

- Study the actual use of standardized vocabularies in maritime communications.
- Coordination of the recognition of and association between standardized phrases and those that are not standardized.
- The possibility of non-standardized phrase conversion into standardized phrases.
- Identification of noise present in maritime communications made through the VHF in order to set an adequate filtering method.

Aside from the objectives of the project, and that the database was designed at the onset to comply with these goals, it shall be possible to use this corpus – with the proper treatment of its data – in other applications corresponding to the automatic speech translation, and the training and test of systems of speech recognition designed for its use in these settings.

Therefore, and always with the objective of serving the abovementioned goals, the database to be designed should comply with the following specifications:

- It must be an oral database, that is to say, it should contain sound, so transcription is not necessary at first. Currently, the database contains a total of 120 acoustic records of which only 20% have been transcribed in a basic way in order to comply with the other previously stated goals of the project.

- It should be a specialized corpus due to the fact that the recordings should be performed on external communications made through VHF between ships and between ships and shore services. The database, then, contains examples of unilateral and bilateral communications performed in maritime settings. Among the former, we find conversations between the vessel and harbour services; reports to traffic separation schemes and communications between ships, while the latter case basically consist of bulletins with weather forecasts and notice to mariners.
- It should be a spontaneous speech corpus. Compiled conversations had to be, and they have actually been, compiled in the natural settings where they take place and at the time and way in which they were performed. Communications were recorded in the working environment, including those that are usually read, such as meteorological bulletins, but they were not specially performed for their recording even when they were recorded at the time of being retransmitted.
- It should be a multilingual corpus as the goal of the project is the study of the Spanish and the English languages. It could be said that we are dealing with a comparable corpus, as it is compiled in different languages that share similar origin, thematic and extension. In this regard, the number of texts in Spanish is 18, there are 83 texts in English, and 19 texts are expressed in both languages.
- Finally, the channel to be used should be radiotelephony, as it is through this means that most of the oral communications between ships and between ships and shore services are performed.

Selection of speakers' features

In a database as specific as that proposed here, the selection of speakers' characteristics is almost mandatory. In the first place and in regards to the number of speakers, we would want it to be as high as possible, as the greater the number of speakers, the more variety of communication, that is, due to the fact that communications in these settings are quite limited in regards to topic and that the way they are performed should be standardized, the highest number of speakers performing the same type of communication will give a more accurate idea of the deviation experienced by actual communications in comparison with the standard way in which they should be performed. This database has around 231 speakers, out of which, 190 are men and only 41 are women. This recent data is particularly significant, as we must take into account that the proportion of men and women who work in maritime settings is far from 50/50 and thus it is reflected in the communications recorded in the corpus. It is also important to remark that the same nature of maritime trade and the special characteristics of the communications performed in these settings makes it difficult for us to know the identity of the speakers, their nationality or mother tongue.

Another important aspect about the data of the speakers has to do with their age. Because the recorded conversations have been performed through the VHF between ships or between ships and shore services, it is logical to assume that duty officers and captains of ships, or duly qualified and professional inshore personnel perform communications. It is for the this reason that we estimate that the minimum age of the speakers should not be less than 20 years old in any cases, nor, except for occasional cases, over 60 years. This would be a perfectly appropriate age range in order to avoid voice degradation.

Also related to these speakers' characteristics, we found little social diversity, as we should assume that speakers are all people who have attained at least a medium or higher academic education and who are working in a common environment, therefore their social status should be similar.

Finally, in regards to the diversity of dialects, we will distinguish among the recordings performed: those that are in Spanish and those that are in English. In relation to the recordings taken in Spanish, we can guarantee that there are only 40 speakers and that the conversations held – all of which are related to the settings in question – do not leave much room for differences in dialects.

On the other side, and in relation to the conversations held in English, we can state that, taking into account the statistics, which show that the number of non-native English speakers in maritime settings is near to 60%, we can assure that most of our speakers in ships (71 over 118) are non-native English speakers and that the entire sample of land speakers (113) are also non-native English speakers due to the fact that they belong to services located in Spain and Italy.

Data compilation

As this is a corpus of spontaneous speech of the communications between ships and between ships and shore services, the conversations should be recorded from either vessels or from shore services, that is, maritime traffic control schemes and other currently used ship-shore communication devices.

In order to take the samples from the vessels, the presence of several members of the research team on board was necessary. These team members would take the samples and make the initial database. To achieve it, three members of the team have been recording communication exchanges that would take place during navigation duties and prepare, while on board, a first database related to the conditions under which each of the conversations was held. 95 recordings have been compiled and this constitutes 78.5% of the corpus.

In order to acquire samples of shore devices, a formal request was made for collaboration with the institutions of maritime rescue (SASEMAR) in this project through the contribution of recordings conducted from the maritime traffic control schemes and Spanish rescue coordination centres; a request that was favourably received and which provided us with 26 recordings, constituting 21.5% of the total database.

Recording

In the database, the time and date of each of the recordings from ships are registered while there is no data for those given by the shore services. The method of data collection was through a digital recorder for conversations held in ships and in the case of recordings performed on land, direct connection to a CPU. For this reason, the quality of the former is lower than that of the latter, while the level of noise is – in general – higher in the first case. Also, in the case of conversations taken from the ship, these were later transferred to a CPU in order to be coded and treated. The samples acquired constitute a total of 3 hours 50 minutes and 49 seconds of acoustic material in total.

Treatment

As was mentioned previously the treatment to be performed on the acoustic data depends upon the use we want to give to the corpus. Out of the four processes that have been studied: transcription, notes, labels and alignment, we have only applied two of them to our corpus: a partial transcription and labelling.

Transcription was made because it is the study of standardized phrases in actual conversations, one of the main goals for the compilation of this corpus. This transcription has been made partially and orthographically.

The justification for a partial transcription is mostly due to lack of human and technological resources that are needed to perform this task completely. At the same time, the phonetic and acoustic transcription needs many hours of work on the part of specialized research personnel, and neither of them is attainable within the parameters of the current project, for which, nonetheless, this specific type of transcription was unnecessary. In sum, a total of 24 conversations were transcribed, constituting approximately 20% of the available material. We must remark that those transcriptions were not made on conversations randomly chosen but that they were selected from among those with the best acoustic features.

With regard to labelling, a series of data detailed below were registered in each conversation. Some of these data, such as date, time or ship's status were registered *in situ* at the same time that the conversations were recorded. Other information, such as the type of communication, number of speakers or number of *turns*, was obtained later during the first treatment of data. The labelling/coding exercise explained below was further conducted on the entries in the database.

Organization

All the compiled acoustic data in addition to those features added during the labelling stage, have been organized in an easy to use Microsoft Access database. Data corresponding to labelling features are the following: record ID, date, time, duration, type of communication, communication conditions, ships' conditions, status

between ships, reason for communication, environmental features, distance, location, quality, number of speakers, gender of speakers, language, recording place, speaking features, recording method, number of *turns*, transcription, number of words. Of all these characteristics, the following can be remarked upon:

- Type of communication: refers to whether the communication was bilateral, ship to ship or ship to land services, or unilateral, that is, made from the ship or from land. In our corpus, 78.3% of the communications are bilateral, while 21.7% are unilateral.
- Communication conditions: in this case we explain whether the bilateral communication was ship to ship or ship to shore and if the unilateral communication was from the ship or from shore. In this case 3.2% of the bilateral communications were ship to ship and 96.8% were ship to shore, while in the case of unilateral communications, 84.6% were made from shore and only 15.4% were made from ships, as it is shown in Figure 1.

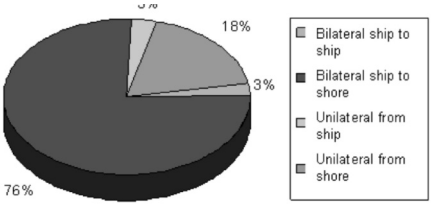

Figure 1: Communication conditions.

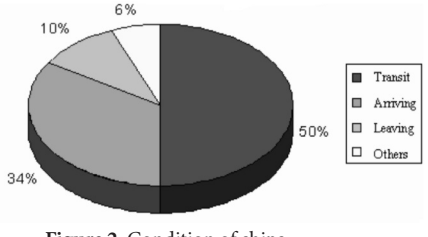

Figure 2: Condition of ships.

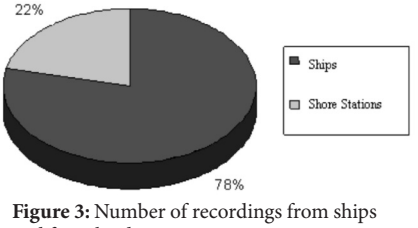

Figure 3: Number of recordings from ships and from land services.

- Ship/s Status: in the case that one ship took part of the conversation, data should be compiled in relation to its navigation circumstances such as transit, anchorage, arriving port, leaving port, or any others. Figure 2 shows the distribution of results where most conversations were observed taking place by ships in transit 50%, arriving port 34%, leaving port 10% and other circumstances 6%. Figure 3 shows the number of ships participating, 78.58%, in contrast to the number of recordings in shore stations, 21.5% .
- Reason for communication: this sections gives us an idea of the type of message transmitted through communication, that is, it indicates the reason why the communication was made, such as calls to pilot stations 28.4%; call to port control stations 10.8%; report to maritime traffic control schemes 25.8%; weather forecasts 11.7%; and others (EVAMED, SECURITÉ, collision report, request for anchor permission, report to navy ship, notices to mariners, etc.), 23.3%. This information is represented in Figure 4.

- Environmental features: refers to the current meteorological conditions at the time of recording. The percentages corresponding to these conditions were the following: clear 70%; cloudy 15.5%; precipitation 3.3%; and others 11.2%.
- Recording quality: refers to the final acoustic quality of the recording obtained under the criteria of noise presence in the sample. The range of quality is indicated as good when maximum noise level is up to 700 Hz, fair when maximum noise level in the sample is between 700 and 1000 Hz , and bad when noise peak levels are over 1000 Hz. As Figure 5 reveals, the average quality of recordings have been deemed of good (58.3%) quality while only 10% are considered to be bad quality.
- Number of speakers: we refer to the number of speakers in each conversation which will vary between 1, as in unilateral communications and up to 4 registered speakers in a particular bilateral conversation. The total number of participating speakers is 231.
- Gender of speakers: we have already remarked that maritime settings constitute a professional environment mainly composed of men, circumstance that is reflected in the data obtained for this corpus where the proportion of male speakers is much higher than that of women. In concrete, 34% women have participated in the conversations.

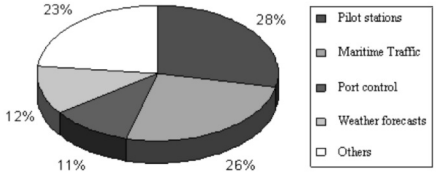

Figure 4. Reason for communication.

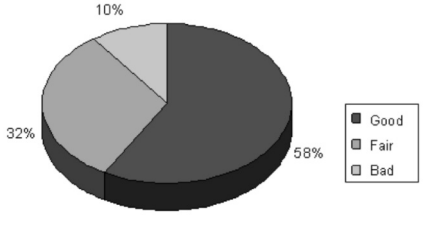

Figure 5. Recording quality.

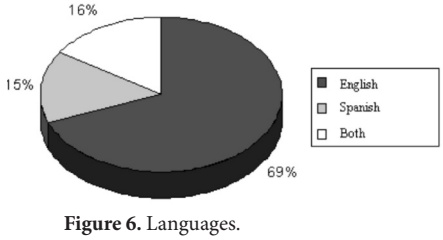

Figure 6. Languages.

- Language: languages compiled in this corpus are Spanish and English. In Figure 6 we can compare the number of recordings made in Spanish, 15% with those made in English, 69.2%. Furthermore, we have added a section called "mixed" which comprise conversations made in both languages with a total of 15.8%.
- Location and recording method: the location of recording could be a ship, 78.5% of the conversations, or the Rescue Coordination Centre of Finisterre, with 21.5%. The recording location also conditions the recording method, as conversations recorded on the ship could only be performed with a portable digital recorder, and the recordings corresponding to the Rescue Coordination Centre of Finisterre were directly made with a PC.

- Speaking features: as we have remarked upon earlier, is basically a corpus of spontaneous speech and 82.5% of communications of this type are to be found. However, there is a small percentage of communications made on the basis of a written script (17.5%), corresponding to weather forecasts and notice to mariners, which, although they not intended to be recorded, were conducted nonetheless at the time the information was transmitted to the rest of the stations.
- Number of *turns*: at this point, the number of turns in each conversation is compiled so that we can obtain the number of interventions per recording and furthermore, the total number of interventions.
- Transcription and number of words: the orthographical transcription of the conversation is attached, if necessary. Furthermore, for those transcribed conversations, we know the number of words for each record as well as the total number of words and the number of distinct words contained in the database.

Table 1. shows some of the general data finally obtained in this corpus.

Table 1: General data of ECOMARS

Duration	3h 50m 49s
Number of speakers	231
Total number of turns	985
Number of transcribed words	2648
Number of diferent transcribed words	562

FUTURE PERSPECTIVES

The ECOMARS database was created in order to comply with the goals of the project "Language industries applied to oral communications in maritime settings". However, for the purpose of its design it was decided to register the most likely variety of data about recorded conversations in order to obtain a usable database for the future compiling and treatment of useful data for the training and test of voice recognition systems specially designed to the improvement of oral external communications in maritime settings.

Therefore, in spite of the fact that the amount of compiled data is little more than symbolic, being far from the registered quantities from other national and international projects, the variety of information on this data is valid and its format allows, furthermore, the realization of a later and much broader treatment.

At the same time, there is enormous potential to use this corpus as a database for the compiling and treatment of conversations of this type on the basis of more specific future projects, with human and economic resources to that effect, where the cooperation of official institutions, such as those in charge of maritime traffic control in separation schemes, port control and pilotage services along national and international lines, would be the most interesting due to the amount of data that these institutions obtain and manage on a daily basis.

Finally, we would like to add that a database with these features could be useful in and for the creation of didactic materials, applicable to the teaching of the English language in maritime settings, as these samples reflect the actual "language in use" in communications made in these particular settings.

REFERENCES

Casacubera, F. (2001). Confluencias entre el procesamiento del lenguaje natural y las tecnologías del habla: traducción automática del habla. In: *Seminario sobre Industrias de la lengua*. Fundación Duques de Soria.

Casacuberta, F and Vidal, E. (2006). La traducción automática del habla. EUROMAP Tecnologías del lenguaje España.

Hanzu-Pazara, R.; Barsan, E.; Arsenie, P.; Chiotoroiu, L. and Raicu, G. (2008). Reducing of maritime accidents caused by human factors using simulators in training process. *Journal of Maritime Research*, Vol. 5, N.1, pp. 3-18.

Llisterri, J. (2004). Las tecnologías del habla para el español. In: Sequera, R. (Ed.) *Ciencia, tecnología y lengua española: la terminología científica en español*. Madrid: Fundación Española para la ciencia y la Tecnología, pp. 123-141

Mariño, J.B. (2002). Memoria científico técnica del proyecto ALIADO. Unpublished.

Meck, U.; Strohschneider, S. and Brüggermann, U. (2009). Interaction design in shipbuilding: an investigation into the integration of the user perspective into ship bridge design. *Journal of Maritime Research*, Vol. 6 N. 1, pp 15-32.

Villarubia, L. (2004). Eliminación de barreras mediante la tecnología del habla. *Jornada Tecnología del habla y discapacidad visual*. Facultad de informática, Universidad Complutense de Madrid, marzo 2004.

ECOMARS: BASE DE DATOS ORAL PARA LAS COMUNICACIONES EXTERNAS EN EL ÁMBITO MARÍTIMO

Resumen

Los resultados obtenidos en diversos proyectos internacionales muestran claramente la viabilidad de implementar sistemas automáticos de traducción simultánea en tareas concretas, cuyo dominio del discurso es limitado. Una de estas áreas restringidas, tanto en lo que a vocabulario se refiere como a estructura sintáctica, son las conversaciones orales rutinarias realizadas en el ámbito marítimo entre buques o entre buques y dispositivos de tierra mediante el uso de las Frases Normalizadas para las Comunicaciones Marítimas.

Así pues, desde la Universidad de A Coruña se ha afrontado un estudio denominado "Industrias de la lengua aplicadas a las comunicaciones orales en el ámbito marítimo" cuyo objetivo es estudiar las posibilidades legales y técnicas, así como la conveniencia comercial del desarrollo de un traductor automático para las comunicaciones orales en este.

Entre las acciones a desarrollar en este proyecto se encuentra la recopilación de muestras orales reales de tales comunicaciones, muestras que han dado lugar a la base de datos acústicos ECOMARS.

ECOMARS es una la primera y única base de datos oral del habla espontánea correspondiente a las comunicaciones externas marítimas en lenguas española e inglesa, que se ha recogido y tratado conforme las necesidades establecidas en el mencionado proyecto. Las especificaciones correspondientes a su diseño, realización y tratamiento serán ampliamente comentadas a en el presente artículo.

Palabras clave: Base de datos orales, comunicaciones marítimas, lenguajes controlados, traducción automática del habla.

ECOMARS: PROYECTO DE BASE DE DATOS PARA LAS COMUNICACIONES EXTERNAS MARÍTIMAS (External communications in maritime settings)

ECOMARS es la primera y única base de datos oral del habla espontánea correspondiente a las comunicaciones externas marítimas en lenguas española e inglesa, que se ha recogido y tratado conforme las necesidades establecidas en el anteriormente mencionado proyecto: "Industrias de la lengua aplicadas al ámbito marítimo". Las especificaciones correspondientes a su diseño, realización y tratamiento serán ampliamente comentadas a continuación.

Especificaciones básicas de diseño

Teniendo en cuenta que uno de los objetivos de trabajo del proyecto "Industrias de la lengua aplicada a las comunicaciones orales en el ámbito marítimo", que sirve además de apoyo para la realización de otros objetivos no menos importantes, es la obtención y análisis de muestras orales reales sobre las comunicaciones marítimas buque-buque y buque – tierra, se nos planteó la necesidad del diseño y creación de una base de datos o corpus oral, cuyas principales utilidades serían:

- Estudio sobre el uso real de vocabularios normalizados en las comunicaciones marítimas.
- Ayuda al reconocimiento y asociación entre frases normalizadas y no normalizadas.
- Posibilidad de conversión de frases no normalizadas en frases normalizadas, y
- Caracterización del ruido presente en las comunicaciones marítimas realizadas a través del VHF, con el fin de establecer un método de filtrado adecuado.

Si bien los puntos marcados son objeto del mencionado proyecto, y la base de datos se diseñó en principio para cumplir estos objetivos, será posible utilizar este corpus, con el debido tratamiento previo de los datos, en otras aplicaciones propias de la traducción automática del habla, como el entrenamiento y prueba de sistemas de reconocimiento del habla diseñados para su uso en este ámbito.

Así pues, y siempre con objeto de servir a los fines anteriormente marcados, la base de datos a diseñar debería cumplir las siguientes especificaciones:

- Ser una base de datos oral, es decir, debería contener sonidos, no siendo necesaria, a priori la transliteración. De hecho la base de datos contiene un total de 120 registros acústicos, de los cuales sólo un 20 % han sido transliterados de forma básica con el fin de cumplir otros objetivos del anteriormente mencionado proyecto.
- Ser un corpus especializado, ya que los registros se realizarían sobre las comunicaciones externas realizadas a través del VHF entre buques y entre éstos y los servicios de tierra. Así pues, la base de datos contiene comunicaciones bilaterales y unilaterales realizadas en el ámbito marítimo. Entre las primeras encontramos, entre otras, conversaciones entre el buque y los servicios de practicaje, notificaciones al paso por dispositivos de separación de tráfico, y comunicaciones entre buque, mientras que las segundas consisten básicamente en boletines con información meteorológica y avisos a los navegantes.
- Ser un corpus del habla espontánea. Las conversaciones recogidas debían, y de hecho así ha sido, recogerse en el ambiente natural en que éstas se producían, en el momento y de la forma que se realizaban. Las comunicaciones fueron registradas en el ambiente de trabajo, incluso aquellas que son leídas de forma habitual, como los boletines meteorológicos, no fueron realizadas de especialmente para su registro, si que se grabaron en el momento en que eran retransmitidas.

- Ser un corpus multilingüe, ya que el objetivo del proyecto es el estudio de las lenguas española e inglesa. Podría decirse que se trata éste de un corpus comparable, ya que recoge textos en distintas lenguas que comparten similar origen, temática y extensión. Por otro lado, el número de textos en español es de 18, 83 se hayan únicamente en lengua inglesa, y 19 textos están expresados en ambas lenguas.
- Finalmente el canal utilizado debía ser la radiotelefonía, ya que es a través de este medio como se realizan la mayoría de las comunicaciones orales entre buques y entre éstos y los servicios de tierra.

Organización

Tanto los datos acústicos recogidos como todas aquellas características otorgadas a los mismos durante el etiquetado se han organizado en una base de datos de Access fácilmente consultable. Los datos de características correspondientes al etiquetado son los siguientes: número de registro, fecha, hora, duración, tipo de comunicación, condición de comunicación, condición de los buques, situación entre los buques, motivo de la comunicación, características ambientales, distancia, situación, calidad, número de locutores, sexo de los locutores, lengua, lugar de grabación, características del habla, método de grabación, número de turnos, trascripción, número de palabras. De éstas, las siguientes características merecen una mención especial:

- Tipo de comunicación: se refiere este dato a si la comunicación era bilateral, bien buque-buque o buque- tierra, o unilateral, realizadas desde el buque o desde tierra. En nuestro corpus un 78.3% de las comunicaciones son bilaterales, mientras que el 21.7% son unilaterales.
- Condición de comunicación: en este caso se explicita si la comunicación bilateral era buque –buque o buque– tierra, y si la comunicación unilateral era desde un buque o desde tierra. En este caso el 3.2% de las comunicaciones bilaterales eran buque –buque y el 96.8% buque– tierra, mientras que para las unilaterales el 84.6% fueron realizadas desde tierra y sólo el 15.4% se realizaron desde un buque. El Gráfico 1 muestra los datos de estas dos características.
- Condición del buque o buques: en caso de que uno o más buques participasen en la conversación, debían recogerse datos sobre su situación de navegación: tránsito, fondeo, entrada, salida u otros. El Gráfico 2 muestra la distribución de resultados, donde puede apreciarse que la mayoría de las conversaciones fueron realizadas por buques en tránsito, 50%, Entrada, 34%, salida 10 %, y otras circunstancias 6%. Por otro lado el Gráfico 3 muestra el número de buques participantes, 78.5%, frente al número de grabaciones de estaciones de tierra , 21.5%.
- Motivo de la comunicación: este apartado nos da idea del tipo de mensaje transmitido en la comunicación, es decir, nos indica le motivo que originó la

comunicación: llamada a prácticos . 28.4%, llamada a control de tráfico portuario, 10.8%, notificación a control de tráfico marítimo, 25.8%, boletines meteorológicos, 11.7%, y otros (EVAMED, SECURITÉ, notificación de colisión, permiso para fondear, notificación a buque militar, avisos a los navegantes, etc.), 23.3%. Estos datos se muestran en el Gráfico 4.

- Características ambientales: este punto hace referencia a las condiciones meteorológicas reinantes en el momento de la grabación. Estas condiciones podían ser: despejado, 70%, nublado , 15.5%, lluvia, 3.3%, y otros, 11.2%.
- Calidad de la grabación: hace referencia este dato a la calidad acústica de la grabación final obtenida, según el nivel de ruido encontrado en la misma. El rango de calidades debería encontrarse entre bueno, con un nivel de ruido inferior a 700Hz, regular, con un nivel de ruido entre 700 Hz y 100 Hz, y mala con picos máximos de ruido superiores a 1000 Hz. Como se muestra en el Gráfico 5, la calidad media de las grabaciones ha sido considerada como buena, 58.3 %, mientras que sólo el 10% de las grabaciones poseían mala calidad.
- Número de locutores: se hace mención aquí al número de locutores de cada conversación, que variará entre 1, para las comunicaciones unilaterales, y hasta 4 locutores registrados en alguna conversación bilateral puntual. El número total de locutores participantes es de 231. Aproximadamente el 80% de estas voces aparecen únicamente en un texto.
- Sexo de los locutores: ya hemos señalado con anterioridad que el ámbito marítimo es un ámbito profesional preferentemente masculino, circunstancia que se refleja en los datos obtenidos para este corpus, donde el número de locutores masculino es mucho mayor que el número de locutores femenino. Concretamente, han participado mujeres en 34% de las conversaciones.
- Lengua: Las lenguas recogidas en este corpus son española e inglesa. En el Gráfico 6 podemos comparar la cantidad de grabaciones realizadas en español, 15%, con las realizadas en inglés, 69.2%. Además hemos añadido un apartado al que hemos dado en llamar mixto, y que recoge conversaciones realizadas en ambas lenguas con un total del 15.8%.
- Lugar y método de grabación: el lugar de grabación sólo podía ser, o bien un buque , 78.5% de las conversaciones, o bien desde el Centro de Coordinación de Salvamento de Finisterre, 21.5 %. El lugar de grabación condiciona, así mismo, el método de grabación, ya que las conversaciones grabadas desde el buque sólo podían hacerse con una grabadora digital portátil, y las grabaciones desde el Centro de Coordinación de Salvamento de Finisterre se realizaron directamente a un PC.
- Características del habla: como hemos comentado con anterioridad, este es un corpus básicamente del habla espontánea, y el 82.5% de las comunicaciones de este tipo así lo corroboran. Sin embargo existe un pequeño porcentaje de

comunicaciones realizadas desde la base de un guión escrito, 17.5%, correspondientes a los boletines meteorológicos y avisos a los navegantes, cuya realización, por otra parte, no fue ex profeso para su grabación, sino que ésta se realizó en el momento en que la información era transmitida al resto de las estaciones.

- Número de turnos: se recogen en este punto el número de turnos de cada conversación, de forma que podamos obtener el número de intervenciones por grabación, el número total de intervenciones.
- Trascripción y número de palabras. Se acompaña aquí la trascripción ortográfica de la conversación, si procede. Además para aquellas conversaciones transcritas podemos conocer el número de palabras, así como el número total de palabras y número de palabras diferentes contenidas en la base.

PERSPECTIVAS PARA EL FUTURO

La base de datos ECOMARS fue creada para cumplir los objetivos del proyecto "Estudio sobre la aplicación de las industrias de la lengua a las comunicaciones orales marítimas". Sin embargo para su diseño se pensó en registrar la mayor variedad posible de datos sobre las conversaciones grabadas, con el fin e obtener una base utilizable, en un futuro, para la recolección y tratamiento de datos útiles para el entrenamiento y prueba de sistemas de reconocimiento de voz especialmente diseñados para su uso en la mejora de las comunicaciones orales externas en el ámbito marítimo.

Así pues, a pesar de que la cantidad de datos recogida es poco más que simbólica, distando mucho de las cantidades registradas en otros proyectos nacionales e internacionales, la variedad de informaciones sobre dichos datos no es despreciable, y su formato permite, además, la realización de un tratamiento posterior más amplio.

Así mismo, sería posible utilizar este corpus como base para la recogida y tratamiento de conversaciones de este tipo en el seno de un posible proyecto futuro más específico, dotado económica y humanamente para tal fin, donde la colaboración de las instituciones oficiales, tales como las encargadas del control de tráfico marítimo en dispositivos de separación e tráfico, tráfico portuario y corporaciones de prácticos, tanto a nivel nacional como internacional, sería de lo más interesante, debido a al cantidad de datos que estas instituciones obtienen y manejan diariamente.

Finalmente señalar que una base de datos de estas características podría ser utilizada para la creación de materiales didácticos aplicables a la enseñanza de la lengua inglesa en el ámbito marítimo, ya que estas muestras reflejan el uso real de la lengua en las comunicaciones orales realizadas en dicho ámbito.