Collection of empirical data for standardization of generic vocabularies in speech driven ICT devices and services

Rosemary Orr1,3, Bernat González i Llinares2, Françoise Petersen3, Helge Hüttenrauch1, Martin Böcker3, Michael Tate3

1 Dept. of Lang. & Speech, Radboud University, Nijmegen, NL; 2University College Utrecht, NL; 3European Telecommunications Standards Institute (ETSI)

r.orr@uu.nl

Abstract

This paper describes a method of collecting multilingual speech data for use in the compilation of spoken command vocabularies for ICT devices and services in the EU, the EFTA countries and Turkey and Russia. The resulting vocabularies will be published as a European standard, for use by industry in the production of such applications. The context of this work is the EU i2010 framework for addressing the main challenges and developments in ICT up to 2010.

Index Terms: standardisation, elicitation, validation, phone discrimination

1. Introduction

Most modern ICT applications offer the option of speech as a form of command and control. Although the speech interface is not the most common form of user interface, there are a number of user groups for whom it is crucial. These are users with special needs, such as children who cannot yet read or write, visually impaired users, users with reduced ability to perceive tactile stimuli, and users with limited dexterity. As technology continues to spread to all sections of society, the proportion of users who can benefit from speech driven applications is increasing.

Users have to learn a specialized vocabulary before they can effectively make use of a speech interface. Uniformity in the basic interactive elements across technologies and brands increases the transference of learning between applications. Standardisation of vocabulary for spoken access to technology is greatly facilitated by this re-application of knowledge.

Such a standard has been published for 5 of the largest European languages, in terms of numbers of speakers, in 2002, i.e. English, Spanish, German, Italian and French [1]. In the interest of cultural diversity, this standardisation is to be extended to a further 23 languages, including the official EU and EFTA languages. Russian and Turkish are also included owing to the large number of speakers of these languages in the EU (see Appendix I).

The work is being carried out by the European Telecommunication Standards Institute (ETSI), with experts from industry, research and development, in the context of the i2010 framework for addressing the main challenges and developments in ICT up to 2010 [2].

2. Methods and Materials

The method of establishing generic vocabularies for this standard differs in some important respects from the general principles described in Annex A of the 2002 standard [1]. For specific details regarding that method, the reader is referred to that document.

2.1. Outline of the procedure for selecting commands for the standard.

The procedure for selecting commands for the making of this standard can broadly be divided into three phases, listed as follows.

- phase 1: elicitation of a command word or phrase from a potential end user using carefully worded descriptions (CWD) [3]
- phase 2: validation of the command as actually representing the CWD
- phase 3: phonetic discriminability testing, to ensure that the system can recognise the correct command

The selection of commands for the standard for each language requires a set of candidate commands. To collect these, a series of interviews with potential end-users, or subjects, is to be carried out in which subjects produce commands which they would use for a certain functionality. Descriptions of the functionalities are presented to them by an interviewer. The procedure for this elicitation process is outlined in section 2.2 below.

Once the candidate command sets have been compiled, they must be validated on two levels. Firstly, they should be rated for confidence that they actually reflect the functionality from whose description they were initially generated. Also, they should not be phonetically similar to other commands, such as to cause confusion in the speech recognizer. The validation process of the elicited command candidates is described in section 2.3 and testing for phonetic confusability is described in section 2.4.

The final choice of command for the standard will be based on results from both the confidence test rating and the acoustical discriminability.

2.2. Collecting candidate commands.

The extension of the language set from 5 to 30 entails a greater need for control over the elicitation and collection of the data. For this purpose, a series of direct interviews are to be conducted with subjects in each language.

2.2.1. Interviewers

In order to cover all 30 of the proposed languages, a team of international interviewers has to be recruited. Clearly, it is logistically challenging and expensive to locate, train, deploy and monitor such a large team over a geographical area covering all the languages. A solution to this challenge is proposed by using, for a large subset of the languages, a team
of students from the University College of Liberal Arts, which is part of the University of Utrecht. The college has a strong international and socially diverse focus and at the time of writing, 26 of the 30 languages are represented among its population of 650 students.

The students follow an interdisciplinary education covering exact sciences, social sciences and humanities. All have followed courses in research methodology. Ideally, at least two or three speakers for each language will be recruited, which should provide in the region of 60 to 70 interviewees. Ideally, where a language is spoken in more than one European country, speakers from all relevant countries would be sought. This should ensure a broad spread of cultural, academic and social backgrounds.

Interviewers are trained in a special induction session, given by the team of experts. In this induction session, they learn about the principles of establishing standards, the general method being used for the current work, their place in it, and the importance of their contribution. They will also be trained in understanding the functionalities for which the command sets are sought, and the use of speech for command and control of ICT applications. They receive official recognition for their work on this standard, and are compensated on an hourly basis.

2.2.2. Subjects

The subjects are drawn from the social network of the interviewers. A general subject-set profile is applied in which subjects are balanced for gender and are representative of three separate age groups, broadly categorised as the young, the working, and the mature generations. There are overlaps in identity between adjacent generations as many members of the younger and mature generations, for example, will also be part of the workforce. The divisions have been chosen to explicitly include the effects of language change, most often initiated by teenagers and young adults. Further, the ageing population, who receive specific mention in the i2010 framework, are also explicitly included. For each language, a target of at least 90 subjects has been set. The intended profile of subjects would therefore be as in Table 1 below.

Table 1. Intended number and profile of subjects per language

<table>
<thead>
<tr>
<th></th>
<th>young 15 - 25 yrs</th>
<th>working 26 - 49 yrs</th>
<th>mature 50 - 60 yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>female</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

2.2.3. Acquiring the subjects

The subjects are addressed in a general mailing by ordinary post, in which they are told about the proposed standard and encouraged to participate so that their language can be represented in the European standard. Further, they are informed that they will be compensated for participation by being entered in a lottery for a number of small prizes. They are further told that the elicitation task that is asked of them will be carried out with the interviewer, whom they know. It is hoped that the personal contact will encourage participation.

In this letter, they are also asked for their consent to have their language data used anonymously for the purpose of creating the standard. They are informed that their contact will call them within ten days of posting to ask if they will participate, and if so, to schedule a time for the interview. Where subjects are also available by e-mail, the same message is also sent out, to facilitate a quick response where possible.

2.2.4. Command set

The command set as put forward in 2002 will be used for this standard also. Although technology has advanced such that many functionalities are now built in to the most advanced ICT applications, such as navigation or search functions, the aim of the standard is to address generic functionalities, and this implies taking the standard applications as the basis for the command set. In total, 70 command words or phrases are required for each language.

2.2.5. Carefully Worded Descriptions CWDs

In order to elicit a potential command from a subject, carefully worded descriptions of each functionality will be read to the subject by the interviewer. The descriptions are such that they do not contain possible command words.

A master set, in English, is produced by the team for this standard. The choice of English simply reflects the fact that this is the working language of this project. A check is made for accuracy regarding the intended functionality, and fluency of the descriptions.

The master set is passed to each interviewer. For each language, the interviewers form a group, translate the CWDs into the target languages, and monitor each other's work. An independent check on these translations is not built in to the design, but this method is considered preferable to using professional translators. Since the interviewers all follow English language education at university level, their competence in English is assured. Furthermore, they will have had a thorough induction into the purpose of the work, and will understand the functions they have to describe, as well as the purpose for which the translations will be used. They are therefore considered more qualified and motivated for this particular translation task.

2.2.6. Interviews

The interviews will take place via Skype from a computer locale reserved for this purpose. The interviewers are provided with headsets to reduce distraction and noise interference with neighbouring interviewers. Previous work with headsets in this area indicates that up to 30 interviews can take place simultaneously.

Each interview begins with three minutes of unscripted informal conversation between the interviewer and their subject, in which the form of the questionnaire is explained. This is intended to put the subjects at their ease, and to give advance warning to the subject of the artificial nature of the questions, and the constraints on the interviewer, regarding the help he or she can give. Such constraints are necessary to reduce interviewer bias as much as possible. Explanation of the artificial nature of the questionnaire is necessary to avoid awkwardness due to the personal link between interviewer and subject.

The remaining text for the interviews is primarily generated as a questionnaire, using commercial software. The questionnaire begins with questions related to the background
and personal details of the interviewee, addressing in particular, age, familiarity with ICT applications, and familiarity with speech-driven user interfaces.

Following this, each CWD is produced as a question in the frame of “What command would you want to give in the following situation?” The time required to carry out a questionnaire with 70 CWDs would be in the region of 40 minutes.

At the end of each interview, the interviewer has three more minutes to close the session with informal conversation, with time to observe any difficulties, or remarkable data or subject behaviour.

The responses given to all the questions in the questionnaire are typed into a response box by the interviewer. The responses are then automatically stored in a comma separated value (csv) file format. When the interviews are completed for a language, the files resulting from each interview are merged, and the data can be further analysed.

In total, it is estimated that the interviewer will need 45 minutes for each interview, from start to finish, including making notes and beginning and ending with informal conversation.

2.3. Validation

Once the sets of candidate commands have been collected, and the data is checked for typographical errors, grammatically equal candidates and homonyms, a histogram of candidate frequencies can be made. With maximally five candidates per functionality, the top 85% in terms of frequency will be chosen for validation testing.

2.3.1. Test for validation

Although a high frequency candidate response suggests a good candidate, testing is necessary to ensure that the candidates which are produced will actually describe the functionality intended, as well as to rule out possible priming in the elicitation phase. For example, a confirmation could be given by “yes”, “sure” and “why not?!”. It is feasible that “why not!” would also suggest a request for a reason. A confidence rating should be carried out to test the validity of a candidate command.

2.3.2. Setting up the confidence rating

The interviewers now contact a different set of subjects than in the elicitation round, and present the same descriptions, this time with candidate answers. The subjects are asked to rate the candidate commands on a 5-point scale, indicating their level of confidence that the word presented implies the functionality described.

The initial interviewer pool remains unchanged. The validation interviews are a part of their tasks. The subject group is recruited in much the same way as for the elicitation interviews. The questionnaires are also produced in the same way as for the elicitation.

2.3.3. Producing the confidence-rated list

After the interviews, the confidence ratings are sorted according to confidence levels, and the top three will be eligible as candidates for phonetic confusability testing.

2.4. Phonetic confusability testing

For a speech-enabled application, it is essential that command words are recognized correctly. The context for discrimination in human perception is much richer than that present in any speech-enabled application. Therefore, once the confidence-rated commands are established, it is necessary to ensure that they are maximally phonetically discriminable.

The use of a recognizer field test, or a pronunciation dictionary test would both be good ways of examining this, but since such tools are not available for all the languages in the set, an alternative method has been chosen.

In a semi-anechoic chamber, recordings are made of each candidate command by the interviewers. The candidate command is spoken 5 times in isolation. A group of analysts, headed by a phonetician, uses these recordings to describe command words in terms of IPA transcriptions. The transcriptions are then used in testing for phonetic confusability.

The acoustic discriminability of two command words can be predicted on the basis of the phone sequences of the words. The number of different phones in a pair of words is used as a predictor of discriminability. A word confusion estimate is calculated, based on the Levenshtein phone edit distance, and generalised towards phone pair distances that are representative of their phonetic confusability. For instance, the edit distance from [m] to [n] would be much smaller than the edit distance from [h] to [i] [6, 7]. confusability testing is performed only on command words that can be simultaneously available to the user.

2.5. Selecting the command set

The selection of the final command set is based on the confidence ratings, confusability testing and common sense. For example, should the word for the digit 7 be confusable with one of two candidates for shut down in any language, then clearly, the digit takes priority, and the shut down functionality must use another candidate. In the final selection, top candidates will be presented with alternative candidates.

3. Discussion

The method described above contains a number of weaknesses. The student team covers, at the time of writing, 26 of the 30 languages. An alternative, but similar method has to be applied to the other four languages, probably in the countries in which they are spoken.

The issue of regional dialect and vocabulary may not be adequately addressed. For many of the languages in the set, they are spoken in more than one country, and different word use can be expected. The choice of which words will enter the standard is then dependent on the social network of the student team. While every effort will be made to ensure language groups which represent a fair spread of the language area, no guarantee of this can be made in the current design.

The length of the interview also merits attention. Even in reduced form, the subject may be expected to be occupied for 35 minutes answering questions. This is still a considerable length of time, and after the first round of interviews, it may be necessary to generate smaller CWD sets. Furthermore, interviewers will need to expand their list of subjects to deal with smaller questionnaires and still gather sufficient data.
While this will cause some delay to the work, it is not an insurmountable problem.

NOTE: At the time of writing, about 65% of the interviews have taken place. Interviewers are in the process of annotating the data, and some of the above issues can now be addressed.

4. Appendix

List of languages covered for the intended standard:
Spanish, Italian, German, English, French, Slovenian, Gaelic, Bulgarian, Slovakian, Czech, Turkish, Portuguese, Croatian, Estonian, Latvian, Romansh, Russian, Greek, Lithuanian, Norwegian, Swedish, Macedonian, Finnish, Danish, Icelandic, Romanian, Maltese, Hungarian, Dutch and Polish

5. Acknowledgements

We acknowledge the co-operation of the experts from the ES 202 076 standard [1] in the preparation on this work, namely Bruno van Niman, Catriona Chaplin, David van Leeuwen, Lutz Groh and Scott McGlashan.

6. References