1 The QALL-ME benchmark

As described in D9.2 and D9.3, the QALL-ME benchmark is a collection of several thousand spoken utterances related to the domain of tourism, both audio files and their corresponding transcriptions, in the four languages involved in the project: English, German, Italian and Spanish. These utterances ask for information about cultural events, accommodation, movies, gastro, etc. and have been transcribed according to guidelines set out by the QALL-ME consortium. The German, Italian and Spanish questions are also translated into English. Annotation of speech acts and expected answer type (EAT) have been completed. The benchmark is a complete, reusable multilingual resource which can be used for training and testing in the field of QA.

1.1 Data

1.1.1 Spoken data acquisition

The QALL-ME benchmark consists of 16722 utterances in the four languages English, German, Italian and Spanish. Tables 1-3 below give details of precise numbers of utterances for each language, and a breakdown of the utterances into scenarios (created) and pre-prepared (read). Other information, such as number and type of speakers, average utterance duration and total speech recording time, is also given.

The data was collected using scenarios and pre-prepared questions presented to participants in the acquisition process in a supervised call-centre environment, in which the spoken requests were recorded. The scenarios functioned to elicit requests which included certain types of information, allowing participants to phrase their requests in a wide variety of ways, rather than being restricted to a particular form. This is necessary to model linguistic variability. These scenarios were alternated with pre-prepared questions which simply had to be read out, to ease the burden of utterance production on the participants. For more information see D9.2, Section 2.1.1.

	Speakers	Female	Male	Native	Non-native
Italian	161	93	68	149	12
English	113	63	46	88	21
Spanish	150	41	109	142	8
German	66	58	8	63	3

Table 1: Speakers in the QALL-ME benchmark

	Total duration	Average duration/utterance
Italian	9hr 20m	7s
English	7hr 35m	6.1s
Spanish	6hr 26m	5.14s
German	4hr 39m	5.5s
All languages	28 hours	-

Table 2: Utterance duration in the QALL-ME benchmark

		Number of questions	Total length (words)	Average length (words)
Italian	Read	2290	25715	11.2
	Created	2374	33492	14.1
	Total	4664	59207	12.7
English	Read	2215	26626	12
	Created	2286	36000	15.8
	Total	4501	62626	13.9
Spanish	Read	2250	25919	11.5
	Created	2250	26327	11.7
	Total	4500	52246	11.6
German	Read	1966	24368	12.4
	Created	1091	11298	10.4
	Total	3057	35666	11.7
All languages	Read	8721	102628	-
	Created	8001	107117	-
	Total	16722	209745	-

Table 3: Read and created questions in the QALL-ME benchmark

Due to the nature of our data collection and the program used to generate the scenarios and present the pre-prepared utterances to the participants, there are some repetitions in utterances across the collection. Of the 16722 utterances, 11748 distinct, i.e., not repeated, and 4974 are repeated at least once more in the benchmark. However, almost all the repetitions appeared in the read utterances, which means that the scenarios were effective; they did not suggest certain realisations of the information to the speakers, and, as predicted, allowed participants to structure their requests in a wide variety of ways.

All utterances were manually transcribed using the Transcriber tool[1], and the German, Italian and Spanish data translated into English, according to guidelines specified in the project (see D9.2, Section 2.2).

1.1.2 SMS data acquisition

An SMS acquisition activity was carried out for Italian by Comdata, as an exercise to identify issues involved in this novel kind of question acquisition. 515 SMS messages were collected from 103 participants by employing the same web sessions as for the spoken data acquisition (see above). We found that the use of standard Italian prevailed (80%), although there were instances of the use of 'telegraphic' forms of language and of abbreviations. We also found that there was a limited use of typical SMS syntax and symbols. Because the addressee is an automatic system, the participants tried to be as clear as possible to be more likely to obtain the information required. Had they been dealing with a person and not a machine, it is possible that they would have used more typical SMS language. D9.2, Section 2.1.3 provides more details of the SMS data acquisition.

[1] http://trans.sourceforge.net

1.2 Annotation of the benchmark

Two main levels of annotation were carried out after the recording, transcription and translation into English (where necessary) of the data in the QALL-ME benchmark: speech acts annotation and expected answer type (EAT) annotation. Each of these levels is dealt with in turn below.

1.2.1 Speech acts annotation

The first level of annotation completed is the speech acts annotation, where different types of speech acts within the benchmark utterances are marked. This stage of annotation has been completed for Italian, Spanish and German according to a set of annotation guidelines specified in the first cycle (see D9.2, Section 2.3), and for English using an adapted set of guidelines developed in the second cycle (see D9.3, Section 2.2). The tools CLaRK[2] and PALinkA (Orasan 2003) were used for speech acts annotation of Italian, Spanish and German, and English, respectively.

The speech acts annotation of the QALL-ME benchmark incorporates marking information at the utterance level, to identify utterances suitable for further annotation, and at the speech acts level. Speech acts are described as the actions performed by particular utterances, and our annotation of these relates to requesting and questioning as well as other types of speech acts such as stating/asserting, greeting and thanking (introducing and ending). This annotation allows us to identify the precise part of the utterance which requests information and therefore expects an answer, and to separate this from other parts of the same utterance which are used to contextualise the request and do not need an answer as such. Based on these labels, and on the notion of primary and secondary speech acts, we also annotated direct speech acts and indirect speech acts. In the latter, the form and the function of the utterance do not conform to their typical matching as they do in the case of direct speech acts. This element of the annotation is important as it captures the different ways in which the same (or similar) request(s) can be realised by speakers. D9.2 and D9.3 contain more information about the speech acts annotation of the benchmark.

1.2.2 Expected answer type (EAT) annotation

The second level of annotation completed is that of expected answer type (EAT). The EAT is the semantic category associated with the desired answer to a particular question/request. Question types were annotated first, which allowed us to separate out FACTOID questions from other types. Only FACTOID questions were further annotated with EAT information. Two EAT annotations were performed: one using a general taxonomy, Sekine's Extended Named Entity Hierarchy[3], and one using a domain-specific ontology, the QALL-ME ontology. D9.2, Section 2.4 presents details of EAT annotation according to Sekine's Extended Named Entity Hierarchy and D9.3, Section 2.2.3 discusses the annotation based on the QALL-ME ontology, which uses ontology concepts as expected answer types and establishes rules to guide the annotation. The conventions used in the QALL-ME project could, by extension, be used in any other ontology-based QA system. The tools FuuTag[4] and CLaRK[5] were used for the general and domain-specific EAT annotations, respectively.

---

[2] http://www.bultreebank.org/clark/index.html
[3] http://nlp.cs.nyu.edu/ene/
[4] Also available at http://nlp.cs.nyu.edu/ene/
[5] http://www.bultreebank.org/clark/index.html