10.12871/CLICIT2014160

Il corpus Speaky

Fabio Poroli, Massimiliano Todisco, Michele Cornacchia,
Cristina Delogu, Andrea Paoloni, Mauro Falcone
Fondazione Ugo Bordoni
Viale del Policlinico 147 – 00161 Roma
{ fporoli,mtodisco,mcornacchia,cdelogu,apaoloni,mfalcone @fub.it }

Abstract

Italiano. In questo lavoro presentiamo un corpus di dialogo uomo-macchina acquisito nell'ambito del progetto SpeakyAcutattile con la tecnica del Mago di Oz. Il corpus contiene più di 60 ore di registrazione di audio, trascritto ortograficamente, e di video. Si presta in particolar modo all'analisi della gestione del turno e della risoluzione degli errori. La simulazione del sistema con il Mago di Oz è stata orientata a una produzione di dialogo vocale senza vincoli da parte del soggetto, sia a livello di gestione del turno, sia a livello di composizione della frase.

English. *In this paper we describe a corpus of man-machine dialogue achieved in the context of SpeakyAcutattile project by the Wizard-of-Oz technique. The corpus consists of more than 60 hours of audio, orthographically transcribed, and video recording. It is particularly suited for the analysis of both turn managing and errors recovering. The system simulation by Wizard-of Oz has been oriented to support a restrictions-free vocal production by subjects, whether for turn managing or for input string composition. .*

1 Introduzione

In questo lavoro presentiamo un corpus di dialogo uomo-macchina acquisito nell'ambito del progetto Speaky Acutattile, una piattaforma digitale per la domotica pensata per il sostegno all'utenza debole (anziani, non vedenti, ecc.), in cui la Fondazione Ugo Bordoni ha introdotto un'interfaccia utente basata sul riconoscimento della voce (Poroli et al., 2013). La piattaforma è stata progettata per fornire agli utenti uno strumento semplificato per la gestione degli elettrodomestici e degli altri dispositivi multimediali presenti in casa (televisione, stereo, etc.), ma anche per l'accesso in rete ai molti servizi di pubblica utilità, come l'assistenza sanitaria, i pagamenti online, le prenotazioni, l'acquisto di titoli di viaggio, ecc. Per la raccolta dati è stata utilizzata la tecnica del Mago di Oz (Fraser and Gilbert, 1991; Dahlback et al., 1993). La tecnica, sebbene richieda maggiori attenzioni e risorse rispetto ad altre strategie di elicitazione del parlato, viene comunemente collocata fra i sistemi più affidabili per la prototipazione di interfacce vocali *user-oriented* e la raccolta dati sulle modalità di interazione con gli utenti.
Eccettuati alcuni vizi strutturali legati al contesto sperimentale (come ad esempio, il minor coinvolgimento del soggetto rispetto all'utente reale), la rilevanza di un *corpus* di dialogo uomo-macchina raccolto con tale metodo viene determinata dalla definizione di alcuni parametri che fissano *a priori* il comportamento del *Mago*, di fatto rendendolo da parte dell'utente il più possibile assimilabile ad una macchina (*machine-like)*.
In questo lavoro è stato inoltre applicato un modello di simulazione di sistema a iniziativa mista (Allen et al., 2001) con grammatiche "*frame-and-slot*" (Bobrow et al., 1977), comprensivo del protocollo di comportamento del dialogo.

La tecnica del Mago di Oz ha consentito pertanto di elaborare le grammatiche di comprensione del dialogo con alcune varianti, verificando nel contempo le reazioni dei soggetti di fronte a un sistema che appariva come reale e non costringeva a percorsi di interazione obbligati per la risoluzione dei compiti.

2 Allestimento dell'acquisizione

2.1 Ambiente sperimentale e soggetti

L'acquisizione dei dati sperimentali è stata condotta nel laboratorio di usabilità del Ministero dello Sviluppo Economico a Roma. Il laboratorio era formato da due stanze, separate da una finestra con specchio riflettente a una via. Analoghe sessioni di registrazione sono state realizzate anche nelle città di Palermo, Torino e Padova , con il Mago di Oz connesso in remoto per il controllo dell'interazione utente.
Ogni soggetto veniva accompagnato e fatto sedere a un tavolo su cui si trovava una lista riepilogativa dei compiti da svolgere. Uno sperimenta-

tore coordinava l'accoglienza, compilava la liberatoria di *privacy* per la sessione, forniva le istruzioni di base e assistenza su richiesta anche durante la fase attiva dell'interlocuzione tra utente e Mago.
Il soggetto, nel caso di appartenenza alla classe Anziani, poteva usufruire di feedback informativi su uno schermo 42" (a distanza di 3m circa) che visualizzava in un angolo un avatar umanoide parlante (Figura 1) denominato Lucia (Cosi et al., 2003). Un ambiente associabile al dominio coinvolto e al compito da svolgere completava il *setting* grafico delle videate, per esempio un menu di prodotti da acquistare fra quelli menzionati nel compito o i canali TV preselezionati un una lista di preferenze.

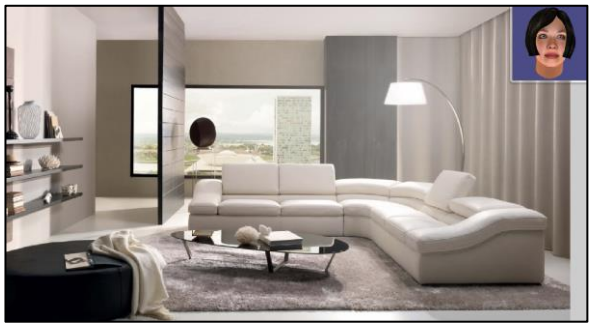

Figura 1: Schermata di lavoro di Speaky-WOz (lato utente)

Il Progetto Speaky Acutattile ha sviluppato dunque l'idea di una piattaforma digitale avanzata per la domotica, costituita da più moduli o dispositivi polifunzionali integrabili, conforme agli standard vigenti e con interfaccia semplice controllata per mezzo della voce.

Il programma di Progetto ha richiesto nello specifico che i servizi fossero rivolti a un'utenza diversamente abile non-vedente (o ipo-vedente) e agli anziani in *digital divide*, cioè persone con età nell'intervallo 65-80 anni, di media scolarizzazione e non dotati di competenze informatiche di base. Per ognuna delle quattro città partecipanti hanno partecipato 20 soggetti (bipartiti per genere M/F, con istruzione medio-bassa e senza conoscenze pregresse in materia di ICT), di cui tipicamente 10 anziani e 10 non-vedenti, per una totale complessivo sul territorio nazionale di 80 individui (oltre a 9 soggetti utilizzati nel *pretest*).

La Tabella 1 riassume le caratteristiche delle due classi utenza.

Città	Soggetti	M/F	Età Media	DS Età	SMB	NO ICT
Roma AN	10	1,0	66,7	6,4	X	X
Roma NV	10	0,7	64,1	16,6	X	X
Padova AN	10	0,3	72,0	5,3	X	X
Padova NV	10	0,9	56,1	16,2	X	X
Palermo AN	10	1,0	69,0	14,0	X	X
Palermo NV	10	0,4	50,8	20,8	X	X
Torino AN	10	1,0	70,1	4,0	X	X
Torino NV	10	1,5	53,3	11,2	X	X

Tabella 1: Utenza sperimentale (Legenda: AN=Anziani, NV=Non-Vedenti, SMB=Scolarizzazione Medio-Bassa, NO ICT=nessuna esperienza ICT pregressa)

2.2 Compiti

Sono stati redatti in totale 48 compiti: ogni soggetto ha svolto circa 20 compiti diversi, composti ognuno da 2-3 attività connesse tra loro. I compiti sono stati progettati in conformità delle caratteristiche del modulo di comprensione del futuro sistema Speaky, secondo il modello *frame-and-slot*: ogni sotto-compito prevedeva perciò un certo numero di variabili da fornire al sistema (di cui alcune obbligatorie e altre facoltative) per il completamento dell'attività. Le istruzioni ai soggetti sono state impartite in due momenti o fasi:
- all'accoglienza con una descrizione a voce del compito da svolgere, ai fini della contestualizzazione degli obiettivi da raggiungere;
- durante il compito, quando il soggetto poteva consultare un promemoria riepilogativo delle richieste all'esecuzione del compito (Tabella 2).

Descrizione	Descrizione estesa	Variabili
Impostare gli orari per l'assunzione di alcuni medicinali.	Il soggetto deve dare il nome del medicinale, la quantità, l'orario d'assunzione ed eventualmente il giorno.	(S1) nome, (S2) quantità, (S3) orario, (S4) giorni della settimana.

Tabella 2: Esempio di promemoria riepilogativo di un compito

2.3 Frasi del Mago di Oz verso i soggetti

Per ogni compito è stato predisposto un insieme di frasi predefinite (Tabella 3) e dipendenti dal dominio (*domain-dependent*), che il Mago inviava ai soggetti di volta in volta, in consonanza con gli obiettivi generali e l'occorrenza specifica dell'azione richiesta.

Compito	Sub	Fase	Out	Tipo	TESTO DA INVIARE
1	1	1	1	1	Ciao! Come posso aiutarti?
1	1	2	1	4	Non riesco a comprendere, puoi ripetere?
1	1	3	1	6	Sono aperte le finestre del salotto e della cucina, le altre sono chiuse.
1	1	3	2	6	Nel salotto la finestra è aperta.
1	1	3	3	6	La finestra della cucina è aperta.
1	1	3	4	6	In bagno la finestra è chiusa.
1	1	3	5	6	La finestra della camera da letto è chiusa.
1	2	1	1	1	Ti serve altro?
1	2	1	2	2	Se vuoi posso chiuderle o aprirle.
1	2	1	3	2	Vuoi chiuderne o aprirne qualcuna?
1	2	2	1	3	Vuoi aprire la finestra del bagno?
1	2	3	1	6	Ho chiuso le finestre del salotto e della cucina.
1	2	3	2	6	Ho aperto la finestra della camera da letto
1	2	3	3	6	Ho chiuso le finestre di salotto e cucina, e aperto la camera da letto
1	2	3	4	6	La finestra del bagno è già chiusa.
1	3	1	1	1	Posso esserti ancora utile?
1	3	2	1	3	L'antifurto non è impostato.
1	3	2	2	3	Vuoi che l'antifurto si attivi quando esci di casa o impostare un orario?
1	3	3	1	6	L'antifurto si attiverà quando esci di casa.
1	3	3	2	6	L'antifurto si attiverà all'ora impostata.

Tabella 3: Numerazione delle risposte

Ogni insieme di compiti è diviso rispettivamente in sotto-compiti, fase del dialogo e tipologia delle frasi, a partire dalla sintesi del dialogo pratico (Allen et al., 2000) proposta da Alexandersson et al. (1997). Come illustrato nella Tabella 3, la prima colonna definisce il compito, la seconda il sotto-compito, la terza la fase del dialogo (1 = apertura, 2 = negoziazione, 3 = chiusura) mentre la quarta il tipo di frasi (1 = apertura generica, 2 = apertura guidata, 3 = richiesta di completamento, 4 = richiesta di ripetizione, 5 = richiesta di conferma errata, 6 = di completamento). Il set generico, *domain-independent*, è invece uguale per ogni compito e comprende le frasi il cui uso è esteso a ogni interazione, come i feedback di accordo, i saluti e le risposte a richieste fuori dominio. Durante l'interazione il Mago usa un'interfaccia grafica, vedi Figura 2, per la selezione manuale dei testi audio da inviare agli altoparlanti del sistema collocati nella stanza utenti del laboratorio.

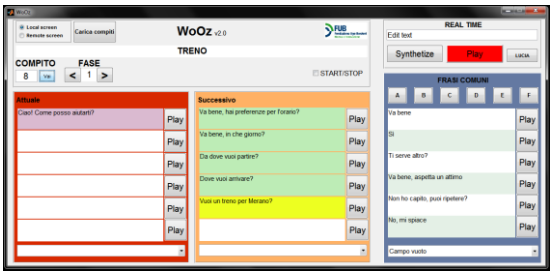

Figura 2: interfaccia grafica del Mago

2.3 Svolgimento dell'interazione

Ogni dialogo inizia con una frase di attivazione del parlante, a cui segue una risposta del tipo "*How may I help you*?" (Gorin et al., 1997), con cui viene lasciata l'iniziativa al parlante per indicare l'attività da svolgere e, potenzialmente, per organizzarne la risoluzione in un solo turno. La fase di negoziazione, collocata tra l'apertura del compito e il suo completamento, è caratterizzata da diversi tipi di frasi: richieste di completamento, richieste di riformulazione, richieste di conferma. Successivamente all'apertura del compito, l'iniziativa passa al parlante, la cui frase può o meno includere tutte le informazioni necessarie; nel caso non vi siano tutte le informazioni necessarie, l'iniziativa torna al mago, il cui compito è elicitare i dati mancanti con frasi di completamento predisposti per coprire ogni caso possibile di assenza di informazioni. La fase di negoziazione prevede anche errori simulati tramite richieste di ripetizione e/o di conferma. Anche in questo caso è stato rispettato per gran parte un protocollo definito a priori: ogni compito prevedeva, infatti, l'uso di una richiesta di ripetizione (ex: «Non ho capito, puoi ripetere?») e di una richiesta di conferma errata, scritta appositamente per ogni compito, da usare coerentemente con le informazioni presenti nella frase del parlante. A seguito del completamento della prima attività, mancando un'eventuale apertura di quella successiva da parte dell'utente (entro tre secondi), è compito del Mago indirizzare il dialogo verso il secondo sotto-compito con una richiesta di apertura generica («Ti serve altro?»).
Per la gestione del dialogo è stato usato un modello a iniziativa mista. Ad esempio, a fronte di un richiesta di conferma errata, il parlante può, infatti, correggere egli stesso l'informazione direttamente nel turno successivo a quello del Mago (es. W: «Vuoi avere informazioni sui treni da Roma a Torino?» – U: «No, da Roma a Mila-

no»); allo stesso modo può prendere il turno (e l'iniziativa) subito dopo la chiusura dell'attività per aprire l'attività successiva. In assenza di un'apertura, il Mago imposta l'avvio di una seconda attività dopo 2-3 secondi di silenzio.

2.4 Descrizione del corpus

Il *corpus* (disponibile in formato audio, video e testuale) è costituito dalle registrazioni delle 80 sessioni di interazione con il sistema simulato, condotte con altrettanti utenti. Ogni sessione comprende circa 20 [1] dialoghi pratici tra il soggetto e il sistema simulato, oltre alle istruzioni iniziali fornite dallo sperimentatore al soggetto e le brevi interazioni tra un dialogo e l'altro. La durata media di ogni sessione è stata di 43 minuti, per un totale di più di 60 ore di registrazione.
Il segnale vocale utile pronunciato dai soggetti è stimabile in circa 16 ore (circa il 25% del registrato disponibile). Tale segnale è stato acquisito da cinque diversi canali a Roma, per tutte le altre città si hanno solo due canali: microfonico e da ripresa video frontale. La tabella 4 mostra i formati utilizzati per tutti i dispositivi e le relative dimensioni dei file per soggetto.

DISPOSITIVO	FORMATO	DIMENSIONI x SOGGETTO
Radiomicrofono Sennheiser XSW 12	PCM wav 48kHz @24bit mono	~ 600 MB
Telefono cellulare Samsung Galaxy SII	PCM wav 32kHz @16bit mono	~ 200 MB
Array microfonico Microsoft Kinect	PCM wav 16kHz @16bit mono	~ 100 MB
Video front ZOOM Q3HD	MPEG-4 1280x720 @25fps PCM wav 48kHz @24bit stereo	~ 3 GB
Video back ZOOM Q3HD	MPEG-4 1280x720 @25fps PCM wav 48kHz @24bit stereo	~ 3 GB

Tabella 4: Dispositivi e formati di acquisizione

Il *corpus* è disponibile anche in formato testuale, trascritto a partire dalla registrazione effettuata tramite il radiomicrofono. Al momento non sono state presi in considerazione le analisi dei dati video (che riprendono i movimenti e le espressioni del soggetto da due diverse angolazioni). Il *corpus* testuale è stato sincronizzato alle tracce audio tramite il software Transcriber 1.5.1 (Barras et al., 2000). Considerato l'allineamento del testo con i file audio, che consente un rapido recupero dei segmenti di dialogo, la trascrizione è stata di tipo ortografico, organizzata per turni. Sono tuttavia stati annotati fenomeni dialogici tipici, come pause, pause piene, false partenze e sovrapposizioni.

3 Ulteriori considerazioni sul corpus

Il *corpus* si presta particolarmente a studi sulla gestione del turno e dell'iniziativa, e sulla gestione degli errori. Tali analisi, oltre a darci informazioni su alcune meccaniche dialogiche di una particolare situazione comunicativa (il dialogo uomo-macchina), possono costituire un utile supporto conoscitivo per integrare le grammatiche di comprensione e le architetture del gestore di dialogo. Ovviamente, il *corpus* raccolto presenta alcuni limiti su altri livelli di analisi linguistica. Infatti, l'utenza principale del sistema, composta da anziani e non vedenti, ha reso necessaria la presenza di uno sperimentatore nella stanza del soggetto e l'uso di un foglio riepilogativo delle attività, variabili che potevano condizionare le scelte lessicali e morfologiche da parte dei soggetti. Da un punto di vista applicativo, tale condizionamento non è un problema: l'ampliamento del dizionario e delle possibili situazioni nel singolo turno di dialogo andranno certamente implementati in una fase successiva del progetto, con dati ottenuti dall'uso reale del sistema reale. Al contrario, il comportamento degli utenti nelle situazioni d'errore e in relazione alla gestione del turno sembra essere meno sensibile al contesto sperimentale, e fornisce valide informazioni per la progettazione del sistema, sia nell'ambito del progetto Speaky, sia, più in generale, per lo studio dell'interazione uomo-macchina.

4 Conclusioni e future attività

La tecnica del Mago di Oz ci ha permesso di ottenere un *corpus* controllato su alcuni aspetti dell'interazione che forniscono indicazioni per l'architettura del sistema di dialogo. I dati attuali verranno integrati con l'acquisizione di un nuovo *corpus* in cui il Mago di Oz "umano" verrà sostituito dal prototipo del sistema, a fronte dello stesso tipo di utenza sperimentale e degli stessi scenari d'uso, allo scopo di ottenere dati confrontabili con gli attuali, sia per migliorare le prestazioni del sistema, sia per ottenere preziosi informazioni sulla tecnica del Mago di Oz.
Le politiche di distribuzione del database saranno definite al termine del progetto (giugno 2015), e auspicabilmente saranno di gratuità per attività di ricerca, ovviamente previo accordo NDA (*Non Disclosure Agreement*).

[1] Variazione dovuta alla presenza o meno dell'ultimo compito sul controllo delle funzioni domotiche interattive (p.e. regolazione altezza delle tapparelle).

Bibliografia

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1997. *Dialogue Acts in VERBMOBIL-2*. Verbmobil-Report, 204.

James F. Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu and Amanda Stent. 2000. An architecture for a generic dialogue shell. *Natural Language Engineering*, 6:1-16

James F. Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu and Amanda Stent. 2001. Towards Conversational Human-Computer Interaction. *AI Magazine*, 22 (4):27-37

Claude Barras, Edouard Geoffrois, Zhibiao Wu and Mark Liberman. 2000. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication (special issue on Speech Annotation and Corpus Tools)*, 33 (1–2).

Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry S. Thompson and Terry Winograd. 1977. GUS, A frame driven dialog system. *Artificial Intelligence*, 8:155–173

Piero Cosi, Andrea Fusaro, Graziano Tisato. 2003. LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro's Labial Coarticulation Model. *Proceedings of Eurospeech 2003*, Geneve, Switzerland.

Nils Dahlback, Arne Jonsson and Lars Ahrenberg. 1993. Wizard of Oz studies–why and how. *Knowledge-based systems*, 6(4):258–266

Allen L. Gorin, Giuseppe Riccardi and Jeremy H. Wright. 1997. How may I Help You?. *Speech Communication*, 23:113-127

Norman Fraser and Nigel Gilbert. 1991. Simulating speech systems. *Computer Speech & Language*, 5(1): 81–99

Fabio Poroli, Cristina Delogu, Mauro Falcone, Andrea Paoloni, Massimiliano Todisco. 2013. Prime indagini su un corpus di dialogo uomo-macchina raccolto nell'ambito del Progetto SpeakyAcutattile. *Atti del IX Convegno Nazionale AISV - Associazione Italiana di Scienze della Voce*, Venezia, Italy.

Fabio Poroli, Andrea Paoloni, Massimiliano Todisco. 2014 (in corso di stampa). Gestione degli errori in un corpus di dialogo uomo-macchina: strategie di riformulazione. *Atti del X Convegno Nazionale AISV – Associazione Italiana di Scienze della Voce*, Torino, Italy.