UNIVERSITY OF LEEDS

This is a repository copy of *Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/100801/

Version: Accepted Version

Article:

AbuShawar, B and Atwell, E orcid.org/0000-0001-9395-3764 (2016) Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems. International Journal of Speech Technology, 19 (2). pp. 373-383. ISSN 1381-2416

https://doi.org/10.1007/s10772-015-9330-4

The final publication is available at Springer via
http://dx.doi.org/10.1007/s10772-015-9330-4

Reuse
Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Usefulness, Localizability, Humanness, and Language-benefit: Additional evaluation criteria for Natural Language Dialogue systems.

Bayan Abu Shawar* — Eric Atwell**

**IT department*
Arab Open University
P.O. Box 1339 Amman 11953 Jordan
b_shawar@aou.edu.jo

*** School of Computing*
University of Leeds
LS2 9JT
Leeds-UK
eric@comp.leeds.ac.uk

ABSTRACT. *Human-computer dialogue systems interact with human users using natural language. We used the ALICE/AIML chatbot architecture as a platform to develop a range of chatbots covering different languages, genres, text-types, and user-groups, to illustrate qualitative aspects of natural language dialogue system evaluation. We present some of the different evaluation techniques used in natural language dialogue systems, including black box and glass box, comparative, quantitative, and qualitative evaluation. Four aspects of NLP dialogue system evaluation are often overlooked: "usefulness" in terms of a user's qualitative needs, "localizability" to new genres and languages, "humanness" compared to human-human dialogues, and "language benefit" compared to alternative interfaces. We illustrated these aspects with respect to our work on machine-learnt chatbot dialogue systems; we believe these aspects are worthwhile in impressing potential new users and customers.*

KEY WORDS: *chatbot, corpus, dialogue, evaluation, usefulness, localizability, humanness, language benefit*

1. Introduction

Practical applications and evaluation are key issues in language engineering. Cunningham (1999) characterises language engineering in terms of "…its focus on large-scale practical tasks and on quantitative evaluation of progress, and its willingness to embrace a diverse range of techniques".

Some NLP systems can be evaluated by comparing output against a "gold standard" perfect target output (e.g., Elliott et al., 2004; Hughes and Atwell, 1994; van Zaane et al., 2004). This is practicable in competitions when comparing rival systems attempting an agreed set task; but when developing systems for practical applications, there are other aspects worth evaluating, even if a single "score" is harder to find. We review some alternative approaches to evaluation of Natural Language Dialogue systems, and then suggest some additional criteria to take account of.

If any computer system is to be taken up by users and customers, it must be demonstrably useful, so "usefulness" is the first of the more qualitative evaluation criteria we look at. "Usefulness" is not a new concept, but it is hard to pin down; we illustrate our approach to measuring "usefulness" by asking users whether responses from our Qur'an-trained dialogue system were relevant to groups of Muslim and non-Muslim users.

One aspect of the evaluation of NLP systems is often overlooked: "Localizability" or portability to deal with dialogue in new languages. To go beyond a "toy" demonstrator, a system needs to be readily adaptable to new language genres or even new languages; it is important to see how easy this process is, otherwise the system is not demonstrably "re-usable" or portable in this sense. We localized our chatbot-generator to a range of new genres of English, and new languages (French, Spanish, Arabic); we have at least informal measures of time and effort this process required, indicative of future requirements for further adaptations or localizations.

Another approach to evaluation of human-computer dialogue systems is comparative linguistic analysis of a corpus of computer-user conversation against an existing corpus of "real" human-human dialogues, to evaluate the "humanness" of the generated natural language. We illustrate this approach by analysing the differences between a chatbot-human dialogue corpus and a human-human dialogue corpus, at lexical, syntactic and semantic levels. The linguistic differences can be detected by NLP analysis, and could be "explained" as a deliberate attempt by chatbot engineers to fake empathy by exaggerating interpersonal conversation cues.

A fourth aspect of evaluation is assessing the "Language-benefit" gained by adopting a natural language dialogue rather than a simpler keyword-based interface; again this can be done by direct comparison to an alternative. We illustrate this aspect by comparing a natural language chatbot interface to our web-FAQ to an alternative interface using Google-style keyword-search. We asked a number of FAQ users to try both interfaces on a range of information-seeking tasks; we were able to conclude that many users found the natural language interface more user-friendly, evidencing the "Language benefit".

These aspects of NLP dialogue systems evaluation, "usefulness" or relevance to user's needs, "localizability" to new genres and languages, "humanness" compared to human-human dialogues, and "language benefit" compared to alternative interfaces, are often overlooked in competitions, but are worthwhile in impressing potential new users and customers and helping them bring their searches to a successful close.

The rest of this paper is organised as follows. Section 2 presents the ALICE/AIML chatbot architecture underlying our natural language dialogue systems, which we use to illustrate evaluation in terms of usefulness, localizability, humanness, and language-benefit. Section 3 reviews some existing evaluation techniques for natural language dialogue systems. The evaluation of usefulness, localizability, humanness, and language benefit are discussed in sections 4, 5, 6, and 7; leading to our conclusion in Section 8.

2. The ALICE/AIML chatbot architecture

Human machine conversation as a technology integrates different areas where the core is language, and the computational methodologies facilitate communication between users and computers using natural language.

A related term to machine conversation is the chatbot, a conversational agent that interacts with user's turn by turn using natural language. Different chatbots or human-computer dialogue systems have been developed using text communication starting from ELIZA (Weizenbaum, 1966), which simulates a psychotherapist, then PARRY (Colby, 1973) which simulates a paranoid patient. "Colby regarded PARRY as a tool to study the nature of paranoia, and considered ELIZA as a potential clinical agent who could, within a time-sharing framework, autonomously handle several hundred patients an hour." (Güzeldere and Franchi, 1995: p. 295).

ALICE (Wallace, 2003) is the Artificial Linguistic Internet Computer Entity, first implemented by Wallace in 1995. ALICE's knowledge about English conversation patterns is stored in AIML files. AIML, or Artificial Intelligence Mark-up Language, is a derivative of Extensible Mark-up Language (XML). It was developed by Wallace and the Alicebot free software community during 1995-2000 to enable people to input dialogue pattern knowledge into chatbots based on the A.L.I.C.E. open-source software technology.

AIML consists of data objects called AIML objects, which are made up of units called topics and categories. The topic is an optional top-level element; it has a name attribute and a set of categories related to that topic. Categories are the basic units of knowledge in AIML. Each category is a rule for matching an input and converting to an output, and consists of a pattern, which matches against the user input, and a template, which is used in generating the Alice chatbot answer.

The AIML pattern is simple, consisting only of words, spaces, and the wildcard symbols _ and *. The words may consist of letters and numerals, but no other characters. Words are separated by a single space, and the wildcard characters function like words. The pattern language is case invariant. The idea of the pattern matching technique is based on finding the best, longest, pattern match. Three types of AIML categories are used: *atomic category* are those with patterns that do not have wildcard symbols, _ and *; *default categories* are those with patterns having wildcard symbols * or _. The wildcard symbols match any input but can differ in their alphabetical order. For example, given input 'hello robot', if ALICE does not find a category with exact matching atomic pattern, then it will try to find a category with a default pattern; The third type, *recursive categories,* are those with templates having <srai> and <sr> tags, which refer to simply recursive artificial intelligence and symbolic reduction. Recursive categories have many applications: symbolic reduction that reduces complex grammatical forms to simpler ones; divide and conquer that splits an input into two or more subparts, and combines the responses to each; and dealing with synonyms by mapping different ways of saying the same thing to the same reply.

The knowledge bases of almost all chatbots are edited manually which restricts users to specific languages and domains. We developed a Java program to read a text from a machine readable text corpus and convert it to ALICE chatbot format language (AIML). The chatbot-training-program was built to be general, the generality in this respect implies no restrictions on specific language, domain, or structure. Different languages were tested: English, Arabic, Afrikaans, French, and Spanish. We also trained with a range of different corpus genres and structures, including: dialogue, monologue, and structured text found in the Qur'an, and FAQ websites.

The chatbot-training-program is formed from several modules that are separated in more than one class, which interact with each other to achieve the specifications. The basic architecture of the system is composed of four phases as follows:

– Reading module which reads the dialogue text from the basic corpus and inserts it into a list.

– Text reprocessing module, where all corpus and linguistic annotations such as overlapping, fillers and others are filtered.

– Converter module, where the pre-processed text is passed to the converter to consider the first turn as a pattern and the second as a template. All punctuation is removed from the patterns, and the patterns are transformed to upper case.

– Producing the AIML files by copying the generated categories from the list to the AIML file.

3. Evaluation techniques used within spoken language dialogue systems (SLDs)

Dialogue systems could be classified in terms of modalities into spoken dialogue systems (SLDs) and textual dialogue ones. Spoken dialogue systems can be *system directed dialogue*, *user directed dialogue*, and *mixed initiative dialogues* dependent on who controls the conversation. In system directed dialogue, the system controls the conversation by requiring a specific piece of information from users to complete the task such as telephone train reservation systems. In user directed dialogue, users control the conversation, and in the mixed ones both parties the system and the user take the initiative at some points.

Hirschman and Thompson (1997) realised that "evaluation plays an important role for system developers (to tell if their system is improving), for system integrators (to determine which approaches should be used where) and for consumers (to identify which system will best meet a specific set of needs). Beyond this, evaluation plays a critical role in guiding and focusing research."

Walker, Litman, Kamm, and Abella, (1997) addressed four types of evaluations for spoken dialogue systems: *subjective* or *objective*, *quantitative* or *qualitative*. *Subjective evaluation* is based on user's judgment and includes metrics such as: contextual appropriateness which classifies the utterances in terms of appropriate, inappropriate and ambiguous; implicit recovery in which user could use context to recover from errors. *Objective evaluation* could be done without referring to human judgments and this includes: percentage of correct answers with respect to a set of reference answers, transaction success, task completion, and the number of utterances. *Quantitative evaluation* computes some statistics and allows comparison across systems. *Qualitative evaluation* uses some rules or experts to judge some parameters.

The purpose of evaluation for SLDs could be classified into: *adequacy evaluation*, which is the determination of the fitness of a system for a purpose; the *diagnostic evaluation* which is the production of a system performance profile, mostly done with test suits of exemplary input; the *performance evaluation* which is the measurement of system performance in one or more specific areas (Hirschman and Thompson, 1997).

Given these competing objectives for spoken dialogue system evaluation, (Hirschman and Thompson, 1997: p. 410) identify two ways to evaluate spoken language dialogue systems: "*glass box* and *black box* evaluation, which sometimes appears to differentiate between component-wise versus whole-system evaluation, and sometimes to a less clear-cut difference between a qualitative/descriptive approach (How does it do what it does) and a quantitative/analytic approach (How well does it do what it does)."

Hasida and Den (1999) agreed that human-computer dialogue systems must be evaluated in terms of the degree of fulfilment of the task achieved by the system, which reflects the efficiency of communication. They developed a framework named DiaLeague (Hasida and Den, 1999) to evaluate natural language dialogue systems on a black box, synthetic, objective, and quantitative basis. The evaluation mechanism is based on comparing between machine-machine dialogues and human-machine dialogues.

McTear (2002) illustrated that the glass box analysis of SLDs is based on evaluating individual components, with measures such as word accuracy, which determines the desired output of the component to be compared with its actual output; sentence accuracy, which measures the percentage of utterances in a corpus that have been correctly recognised; sentence understanding, on the other hand, measures the rate of understood sentences in comparison with a reference meaning representation. For example, glass box evaluation was applied on the ARPA Spoken Language system (Hirschman, 1995), and it shows that the error rate for sentence understanding was much lower than that for sentence recognition.

On the other hand, black box evaluation evaluates the system as a whole based on user satisfaction and acceptance. The black box approach evaluates the performance of the system in terms of achieving its task, the cost of achieving the task in terms of time taken and number of turns, and measures the quality of the interaction, normally summarised by the term 'user satisfaction', which indicates whether the user "gets the information s/he wants, is comfortable with the system, and gets the information within an acceptable elapsed time, etc." (Maier et al., 1996). Black box evaluation is used to evaluate many spoken systems that provide transactional services to customers and have specific tasks to achieve. For example: the Nuance automatic banking system that enables callers to conduct transactions over the phone (McTear, 2002); and the Philips Automatic Train Timetable Information System that provides information over the telephone about train connections between 1200 German cities (Aust, Oerder, Seide, and Steinbiss, 1995).

In 2000 Glass, Polifroni, Seneff, and Zue introduced two new understanding metrics called: *query density* and *concept efficiency metrics*. "The query density measures the mean number of new concepts introduced per user query, while the concept efficiency tabulates the average number of turns it took for a concept to be successfully understood”. They applied these metrics on the GALAXY client-server architecture, which is an air-travel information SLD.

4. Usefulness evaluation

If any computer system is to be taken up by users and customers, it must be demonstrably useful, so “usefulness” is the first of the more qualitative evaluation criteria we look at.

The YPA "is a natural language dialogue system that allows users to retrieve information from British Telecom's Yellow pages" (Kruschwitz et al., 1999, 2000). The yellow pages contain advertisements, with the advertiser name, and contact information. The YPA system returns addresses and if no address found, a conversation is started and the system asks the user for more details in order to give a user the required address. The YPA is composed of the Dialog Manager, the Natural Language Frontend, the Query Construction Component, and the Backend database. The Backend includes a relational database that contains tables extracted form the Yellow pages. The conversation starts by accepting user input through a graphical user interface, then the Dialogue Manager sends textual input through the Natural Language Frontend for parsing. If no addresses are found then the Dialog Manager sends the textual input to the Natural Language Frontend for parsing. After that, the parse tree is sent to the Query Construction Component, which translates the input into a database query, to query the Backend Database and return the retrieved address. If no addresses are found, then the Dialogue Manager starts putting more questions to the user to obtain further clarification.

To evaluate the YPA, 75 queries were extracted from a query corpus, and a response sheet was prepared to see if the returned addresses were appropriate or not, how many dialog steps were necessary, the total number of addresses recalled and he number of those relevant to the original query. Results show that 62 out of 75 queries managed to return addresses, and 74% of those addresses were relevant to the original query.

In a similar manner, we evaluated the “usefulness” of the responses generated by our Qur'an chatbot. The Qur'an chatbot was developed using our chatbot-training-program, where the English/Arabic corpus of the holy book of Islam the Qur'an is used. The Qur’an text is available via the Internet; and in principle the Qur’an provides guidance and answers to religious and other questions. The resulting system accepts user input in English, and answers with appropriate ayyas from the Qur’an in the English and Arabic languages as shown in Figure 1.

Human: How shall I pray?
Qur'an61: {The Resurrection (Al-Qiyama)} 31 So he gave nothing in charity, nor did he pray!-
31فَلَا صَدَّقَ وَلَا صَلَّى (سورة القيامة)

Human: Will Allah judge me?
Qur'an61:
{The Pen (Al-Qalam)} 36 What is the matter with you? How judge ye?
36مَا لَكُمْ كَيْفَ تَحْكُمُونَ (سورة القلم)

{Mankind (An-Nas)} 3 The God (for judge) of Mankind,-
3إِلَهِ النَّاسِ (سورة الناس)

Human: Are men judged in the same way as women?
Qur'an61: {The Resurrection (Al-Qiyama)} 40 Has not He, (the same), the power to give life to the dead?
40أَلَيْسَ ذَلِكَ بِقَادِرٍ عَلَى أَن يُحْيِيَ الْمَوْتَى (سورة القيامة)

Figure 1. Sample of chat with the Qur'an chatbot

To measure the quality of the answers given by the English/Arabic version of our Qur'an Chatbot version, random sentences were selected from Islamic sites and used as inputs to the English/Arabic version of the Qur'an. The resulting transcripts which have 67 turns were given to 5 Muslims and 6 non-Muslims, who were asked to label each turn in terms of: related (R) in case the answer was correct and in the same topic as the input; partially related (PR) in case the answer was not correct, but on the same topic; and not related (NR), in case the answer was not correct and on a different topic.

Proportions of each label and each class of users (Muslims and non-Muslims) were calculated as the total number over the number of users times the number of turns. Four out of the 67 turns returned no answers, therefore actually 63 turns were used as presented in Figure 2. In the transcripts used, more than half of the results were not related to their inputs. Approximately one half of answers in the sample were not related from non-Muslims' point of view, whereas this figure is 58% from the Muslims' perspective. This is because the Qur'an uses traditional Arabic language, which is sometimes difficult to understand without the meaning of some words, and the historical story behind it. Another reason is that the non-Muslims rely on the English translation of the Qur'an, which is not always enough to judge whether the verse is related or not, especially since non-Muslims do not have the background knowledge of the Qur'an that Muslims do.

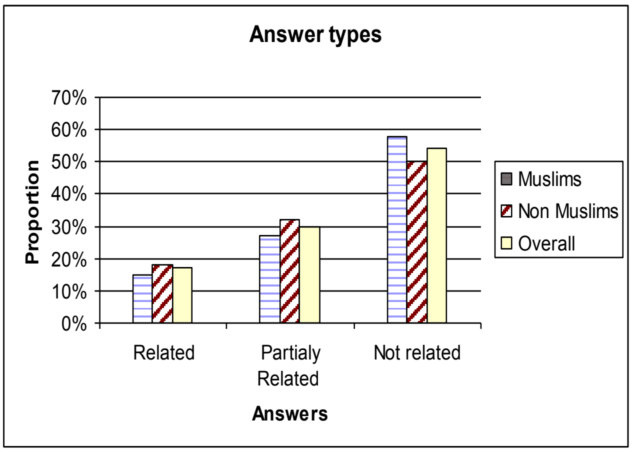

Figure 2. The proportion of each answer type identified by users of the Qur'an chatbot

Gandhe and Traum (2007) used SASO-ST (Traum et al., 2005) to create four types of chatbot prototypes to simulate a doctor in a Captain-Doctor scenario. The scenario used is based on a trainee acts as an army captain negotiating with a simulated doctor. The main goal of the system is to "retrieve one of the doctor's utterances from the corpus and present it to the user as the system response." (Gandhe and Traum, 2007). The aim of the four types is to capture different aspects of local and global coherence of dialogue. These four types have different approaches to retrieving answers ranging from selecting random answers, to using n-grams and accessing the dialogue history. The evaluation of these four types is based on subjective metrics by asking users to evaluate the doctor's (i.e., the Chatbox's) responses. For example, 1 is used for nonsensical replies and 5 is used for appropriate responses.

5. Localizability

The localizability aspect of evaluation tries to identify how easy it is to adapt a natural language dialogue system to new domain or language without affecting the way it works. With this goal in mind, some dialogue systems have been designed to be retrainable to a new domain via a domain corpus

Inui et al., (2003) introduced a natural language dialogue system based entirely on the use of corpora. The aim of this system is to be so general that it can be trained with any corpus in any domain and language. The system is mainly composed of three modules, the NL Parser, the Matcher, and the NL generator as displayed in Figure 3. The inputted sentence is sent to the natural language (NL) parser to analyze the input using the N-gram-based shallow parser (Inui et al., 2002). The matcher uses keyword matching and structural matching to find the dialogue most similar to the current flow in the Dialogue Corpus. The matcher uses the Context Data Base, in which each dialogue act is assigned an intention from a list containing greet, question, explain, etc. In the keyword-based matcher, the nouns and verbs identified by the NL parser are matched with the most similar nouns and verbs from the Dialogue Corpus. Before confirming this match, the matcher checks the intentions associated with those nouns and verbs in the Context Database. In the structural matcher (Koiso et al., 2002), the similarity dialogue is figured out by calculated the structural distance between two sentences. In this fully corpus-base approach, a user has the choice to select which matcher to use. The NL generator generates the system's responses and applies the necessary exchange on the response pronouns. However, no real evaluation found for this system.

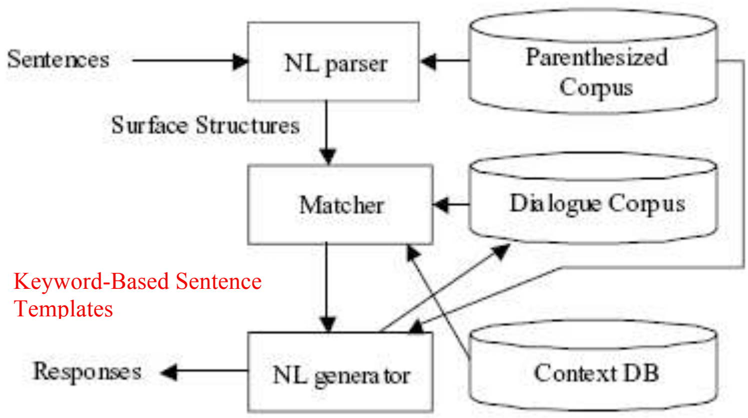

Figure 3. Corpus-Based Approach to Building a Natural Language Dialogue System

We built a generic Java program that reads a dialogue from a corpus and maps it to the AIML format used by the ALICE chatbot to produce different versions of the chatbot, which were evaluated using different techniques. Table 1 displays the corpora used to train our program.

After creating AIML files for the corpora types displayed in Table 1, the Pandorabot web-hosting service[1] was used to publish different versions of corpus-trained chatbots to make them available for use over the World Wide Web. Users were asked to chat with these versions and provide their feedback.

Based on user feedback and the retraining corpus, eight system prototypes were generated to satisfy users' expectations. The key issue in building these prototypes was how to expand the knowledge learned from the corpus to increase the chances of finding a match. The idea of matching is based on finding the best match, which is the longest one. Since the input will not necessary match exactly a whole sentence extracted from the corpus, other learning techniques were adopted.

Table 1. The Training Corpora

Corpus	Language	Content
Dialog Diversity Corpus (DDC)	English	A collection of spoken English dialogue corpora
Corpus of Spoken Afrikaans (KGA)	Afrikaans	Transcripts of General Afrikaans conversations
British National Corpus (BNC) spoken texts	English	Spoken English transcripts in different domains
Qur'an in Arabic	Arabic	Arabic monologue text
Parallel Qur'an in English and Arabic	English input, and Arabic-English output	Aligned English and Arabic monologue text
FAQ of the School of Computing at Leeds University	English	Frequent Asked Questions relating to the School of Computing
FAQ of different Websites	English	Multiple online FAQs: Perl, Linux and Python
QA obtained from health websites	Arabic	Questions and answers related to health issues such as teeth care, fasting, blood disease, and mothers and pregnancy.

In each prototype, machine-learning techniques were used and a new chatbot was tested. The machine learning techniques ranged from a primitive simple technique to more complicated ones. Building atomic categories and comparing the input with all atomic patterns to find a match is an instance based learning technique. However,

[1]http://www.pandorabots.com/

the learning approach does not stop at this level, but it improved the matching process by using the first word, and the most significant words. This increases the ability of finding a nearest match by extending the knowledge base which is used during the matching process. Four dialog transcripts generated by our Afrikaans prototype were used to measure efficiency of adopting learning techniques. The frequency of each type of matching (atomic, first word, significant word, and no match) in each generated dialogue was estimated and the absolute frequencies were normalised to relative probabilities as shown in Figure 4. The results proved that the first word and the most significant approach increase the ability to provide answers to users and to let conversation continue.

These prototypes demonstrated that the chatbot-learning system could be localized to a range of different languages and text-types, given appropriate corpora. The range of systems also demonstrated that the chatbot-learning approach was very versatile and flexible.

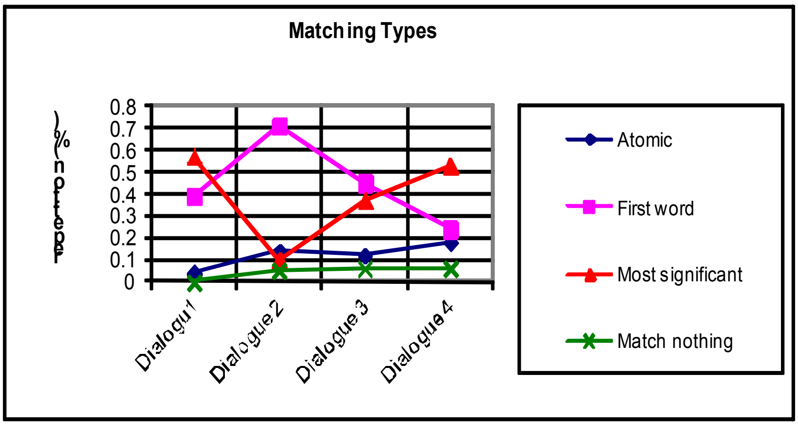

Figure 4. Matching types in the Afrikaans Prototype

Mikic et al., (2008) introduced a tutor bot (T-Bot) which answers students' questions about different courses and subjects that are available via e-learning platform using natural language. Freeling, an open-source linguistic analyser, to extract candidate keywords and concepts related to course contents. This analyser is supported by several languages: English, Spanish, Italian, Catalan, and Galician, which facilitate adapting T-Bot to different languages, and contents.

6. Humanness Evaluation

The humanness aspect of a chatbot is traditionally measured by the ability of the dialogue system to fool users into believing that they are interacting with a real human, not a virtual one. Colby (1975) used this strategy to evaluate his chatbot PARRY that simulates a paranoid patient. A blind test was applied by three psychiatrists questioning both PARRY and three other human patients diagnosed as paranoid. Psychiatrists were not able to distinguish PARRY chatbot from human patients.

The same policy was adopted in the Loebner prize competition, which allows users to chat with a conversational agent for 10 minutes: if this chatting gives the impression to users that they are dealing with a human and not a machine, that conversational agent succeeds in the competition.

However, this is a somewhat superficial and subjective measure: 10 minutes is not really enough to judge the humanness of a system, and the judgement depends on subjective opinions of a few users. We adopt a novel way to measure the humanness of a natural language dialogue system by comparing dialogues generated by the system, against “real” human dialogues. To do this, the Wmatrix tool (Rayson 2003) was used to compare a dialogue transcript generated via chatting with ALICE, and real conversations extracted from different dialogue corpuses. The comparison illustrates the strengths and weaknesses of ALICE as a human simulation, according to lthe inguistic features: lexical, part-of-speech, and semantic differences.

The semantic comparison illustrates that explicit speech act expressions are highly used within ALICE, an attempt to reinforce the impression that there is a real dialogue; pronouns (e.g. *he*, *she*, *it*, *they*) are used more in ALICE, to pretend personal knowledge and contact; discourse verbs (e.g. *I think*, *you know*, *I agree*) are overused in ALICE, to simulate human trust and opinions during the chat; liking expressions (e.g. *love*, *like*, *enjoy*) are overused in ALICE, to give an impression of human feelings.

In terms of Part-of-Speech analysis shows that singular first-person pronoun (e.g. *I*), second-person pronoun (e.g. *you*) and proper names (e.g. *Alice*) are used more in ALICE, to mark participant roles more explicitly and hence reinforce the illusion that the conversation really has two participants.

At lexical level, analysis results shows that ALICE transcripts made more use of specific proper names “Alice” (not surprisingly!) and “Emily”; and of “you_know”, where the underscore artificially creates a new single word from two real words. Table 2 illustrates the lexical comparison between ALICE transcripts file represented in column “O1”, and the real conversation file represented in column “O2”

Table 2. The lexical comparisons

Sorted by log-likelihood value					
Item	**01**	**%1**	**02**	**%2**	**LL**
Do	44	3.90	35	0.65 +	58.69
I	54	4.79	67	1.25 +	48.04
We	1	0.09	129	2.41 -	41.15
So	1	0.09	117	2.19 -	36.75
And	8	0.71	195	3.65 -	35.19
Emily	9	0.80	0	0.00 +	31.46
You	72	6.38	151	2.82 +	28.91
This	0	0.00	70	1.31 -	26.80
You_know	8	0.71	1	0.02 +	22.06

Here	0	0.00	55	1.03 -	21.06
Am	6	0.53	0	0.00 +	20.97
‘ll	1	0.09	71	1.33 -	20.14

Another way of measuring naturalness of a dialogue system was introduced by Dybkjaer et al., in 2004. They focussed on SLDs and claimed that users should talk to the system in an easy smoothly way. In order to do that the authors listed some aspects which must be considered in building dialogue systems: system’s output language should control user’s language so input becomes manageable for the system; output voice should be clear, intelligible and does not need extra effort to listen; contents of the system’s output should be correct and relevant to the topic; adequate feedback is essential for users to feel in control during interaction; and the structure of the dialogue should must be natural and reflects users’ intuitive expectations (Dybkjaer et al., 2004).

Recently et al., (2014, p. 1) discovered that "a chat bot that provides responses based on the participant's input dramatically increased the perceived humanness and engagement of the conversational agent." In their experiment researchers created a chat bot that asked participants to describe a series of images. The interaction was either static, in which the participants answer the base questions, or dynamic, where there is a follow-up question based on the participant's response. A survey was completed by each participant after answering questions about all images, In order to measure humanness, a question about chat partner was provided to see if it was a human or a computer, six option scale were used: definitely human; probably human; not sure but guess human; not sure but guess computer; probably computer; and definitely computer. Results reveal that 79.2% of static interview participants thought their partner was definitely a computer, while only 41.9% of those using the dynamic chatbot thought the same.

7. Language Benefit Evaluation

The Language-benefit evaluation is gained by adopting a natural language dialogue rather than a simpler keyword-based interface; again this can be done by direct comparison to an alternative.

Happy Assistant is "a natural language dialog-based navigation system that helps users access e-commerce sites to find relevant information about products and services" (Chai et al., 2001a). The system is composed of three main modules: the presentation manager (PM), the dialog Manager (DM), and the Action Manager (AC). The presentation manager applies a shallow parsing technique to identify semantic and syntactic information of interest from the user textual input. Then it translates the users input into a well formed XML message called the logical form. The dialogue manager is responsible for matching concepts from a user's query to business rules found in the knowledge domain. The business rules consist of a list of concepts together with some meta data about the target product or service. If a match is found, the webpage associate with that rule is presented to the user. Otherwise, the most important missing concept is figured out by introducing questions to the user. Control is now turned over to the action manager, which accesses the product that matched the query, and if the user provides special preferences, a sorting algorithm is applied to yield a ranked list of products. To make users trust the system, it must offer some explanation before producing a result, so the system summarizes the user's request by paraphrasing it using the context history. Figure 5 presents a sample of conversation with the Happy Assistant System taken from (Chai et al., 2001a).

U: I am looking for a notebook for my consulting business
S: Please describe your financial constraints.
In this stage a list of most popular ones is displayed
U: not important thing performance is essential
S: are you looking for something that is top of the line?
U: Yes, absolutely
S: I have found something for you.
The target notebook is displayed for the user. And beneath it a summary of the users request displayed to explain why this product is displayed.

Figure 5. Example of interaction with the Happy Assistant System

Usability in this system is evaluated based on a study that designed to explore how much the system meet users' expectations in terms of ease of use, system flow, validity of the system response, and user vocabulary. (Chai et al., 2001b).

The study compared the navigation process in the dialog system against a menu driven system to find target products. Results show that users preferred the dialog based search over the menu driven search (79% to 21% of the users) for the following reasons: ease of use, meeting the users' needs, users like the idea that they could express their needs in their language without being restricted to menu choices, users feel that the computer did all the work for them, and more over users found that the system reduce the interaction time. However, novice users preferred the menu driven system because there is no need for typing.

In a similar manner, we used the comparative evaluation to compare the results generated by Google with the results generated by the FAQchat system. FAQchat is another version of the chatbot-training-program described in Section 2, where the FAQ corpus of the School of Computing (SoC) at University of Leeds is used to train the program. The results returned from FAQchat are similar to ones generated by search engines such as Google, where the outcomes are links to exact or nearest match web pages. An evaluation sheet was prepared which contains 15 information-seeking tasks or questions on a range of different topics related to the FAQ database. The evaluation sheet was distributed among 21 members: nine of the staff and the rest postgraduate students. An interface was built, which has a box to accept the user input, and a button to send this to the system. The outcomes appear in two columns: one holds the FAQchat answers, and the other is holds the Google answers after filtering it to the FAQ database. Users were asked to try using the system, and state whether they were able to find answers using the FAQchat responses, or using the Google responses; and which of the two they preferred and why.

Results in Table 3 show that 68% overall of our sample of users managed to find answers with the FAQchat while 46% found them with Google. Since there are several ways to ask the same question, the success in finding answers is based on the way the questions were presented to FAQchat. Of the overall sample, the staff outcome shows that 61% were able to find answers with the FAQchat where 73% of students managed to do so.

Table 3. Proportion of users finding answers

Users /Tool	Mean of users finding answers		Proportion of users finding answers	
	FAQchat	Google	FAQchat	Google
Staff	5.53	3.87	61%	43%
Student	8.8	5.87	73%	49%
Overall	14.3	9.73	68%	46%

In terms of preferences, for each question, users were asked to state which tool they prefer to use to find the answer. The proportion of users who preferred each tool was calculated. Results in Figure 6, show that 51% of the staff, 41% of the students, and 47% overall preferred using FAQchat against 11% who preferred using Google. As a result, we can claim that our users found the chatbot a novel and interesting way to access the FAQ using natural language questions.

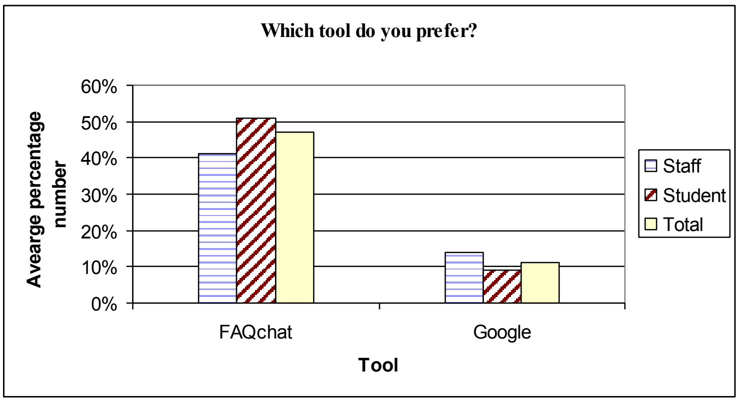

Figure 6. Which tool do you prefer?

In 2006, Quarteroni and Manandhar applied the same approach in using the chatbot as interface for an open-domain QA system. They developed a non-interactive QA system called YourQA (Quarteroni and Mandandhar, 2006), which accepts a textual natural language query from a user, takes the top 20 Google results for this query, retrieves the corresponding Web pages, analyses them to extract answers, ranks them by relevance to the question, and obtains a list of answers in the form of an HTML result page In order to make their system an interactive one, Quarteroni and Mandandhar (2007) used ALICE as an interface for their system. First of all they used the Wizard-of-Oz (WOz) experiment (Dahlbaeck et al., 1993) in which a human operator hides behind a computer interface to simulate the conversation with the user who thought that he is chatting with a computer system. A set of AIML categories were prepared and wizard role was to choose the appropriate category and utterance within the available set as a response to the user. In case there was o response, the wizard would create an utterance to keep the conversation alive. The aim of this experiment is to obtain initial data (conversations logs) to enrich knowledge base of ALICE chatbot. After this the authors build an interactive interface for YourQA system that consists of an applet window with a left panel where chat takes place and a right panel where results are visualised when ready. A sample of chatting with this version is shown in Figure 7.

System0: Welcome!

User1: Hi

S1: Hi! How can I help you?

U2: what is a thermometer?

S2: Possible answers are in the panel on the right. Are you happy with these answers?

U3: Yes!

S3: Good!

U4: How does it measure the temperature?

S4: Do you mean how does a thermometer measure the temperature?

U5: Yes

S5: I found the answer in the panel on the right. Can I help you further?

Figure 7. A sample of dialogue with YourQA chat version (taken from Quarteroni and Manandhar 2007)

A comparison has been made between the systems, the non-interactive YourQA system and the chat version. The evaluations were based on user satisfaction and questionnaires and interaction logs that are effective in evaluation QA systems (Kelly et al., 2006). The questionnaire involved questions to assess performance of the systems, others related to the ease of use and difficulties in reformulating questions, some questions related to the over all satisfaction of the user, and questions to show which version the users prefer. Some users found it slow in giving responses and this is because system performs document retrieval in real time, and hence it depends heavily on the network download speed. In general 83.3% users preferred the chat interface of the system because of its liveliness and ability to understand pronominal anaphora (Quarteroni and Manandhar, 2007). Quarteroni (2008) presented that making a QA system interactive is important to provide ability to chat with user in a natural manner.

The same strategy was adopted by Crockett et al., (2009) to evaluate Adam chatbot. Adam is a conversational agent that simulates a UK university student department Advisor who provides help to students who had department problems, and guides them through options for paying outstanding university fees. Crockett and other researchers compared using Adam with either phoning the university finance office, or visiting it. Results revealed that 80% of students were happy with speed of providing answers, and 70% were satisfied by the advice. As a whole 47% of students preferred using Adam than visiting the University Finance Office, 20% stated they would use Adam instead of telephoning the office.

Our success with using a chatbot as a tool to answer SoC FAQs encouraged us to try other FAQs, or Question Answering (QA) corpora to investigate the possibility of using a chatbot as a tool to access an information portal without the need for sophisticated natural language processing or logical inference. In 2008, an open ended FAQChat was built where the knowledge base was extracted from multiple FAQs: Perl, Linux and Python. In 2010, the TREC09 QA track was used to retrain ALICE and in 2011 Arabic QA corpora were used, a sample of chatting with our Arabic QA chatbot is presented in Figure 8. Even though Arabic language is different than English and other European languages because its: a highly derivative language; has a complex morphology, capitalization is not used, and many other characteristics as presented in (Shaalan, 2014), we found out that a simple (but large) set of pattern-template matching rules is sufficient to provide correct answers.

Overall, user trials with AskJeeves, Google and this generated chatbot demonstrate that a chatbot is a viable alternative, and, in fact, many users prefer it to Google as a tool to access FAQ databases (Abu Shawar, 2008, 2011; Abu Shawar and Atwell, 2010).

Human: ماهي أسباب السكري (What are the reasons of diabetes?)

Arabic_Q/A_blood_disease:

يحدث الداء السكري نتيجة فشل البنكرياس في إفراز ما يلزم من الأنسولين لخفض مستوى سكر الدم الذي يرتفع عادة بعد تناول الطعام وإجراء التمارين والتعرض للشدة

Human: ما سبب رائحة الفم؟ (What is the reason of mouth smell?)

Arabic_Q/A_Teeth_issues:

سبب هضمي 10% سبب فموي 80% عوز بعض الفيتامينات 10%.

Human: ماهي الثياب التي تلبسها الحامل؟ (What cloths should a pregnant woman wear?)

Arabic_Q/A_mother&kids_issues:

يجب أن تكون الثياب في أثناء الحمل عملية وبسيطة وفضفاضة ، لذلك تستبعد المشدات لعوقها الدوران الوريدي ومساعدتها في حدوث الدوالي ، وتجنب استعمال الأحذية ذات الكعب العالي التي تسيء إلى التوازن وتسبب آلام الظهر.

Figure 8. A sample of chatting with Arabic QAchatbot

8. Conclusion

Different evaluation techniques used in natural language dialogue systems has been presented in this paper, including black box and glass box, comparative, quantitative, and qualitative evaluation. Four aspects of NLP dialogue systems evaluation are often overlooked: “usefulness” in terms of user’s qualitative needs, “localizability" to new genres and languages, "humanness" compared to human-human dialogues, and "language benefit" compared to alternative interfaces. We illustrated these aspects with respect to our work on machine-learnt chatbot dialogue systems; we believe these aspects are worthwhile in impressing potential new users and customers.

9. References

Abu Shawar, B. (2008). Chatbots are natural web interface to information portals. Proceedings of INFOS2008, pp. NLP101-NLP107.

Abu Shawar, B. (2011). A Chatbot as a natural web Interface to Arabic web QA. International Journal of Emerging Technologies in Education (iJET), 6(1): 37-43.

Abu Shawar, B., and Atwell, E. (2010). Chatbots: Can they serve as natural language interfaces to QA corpus? Proceedings of the Sixth IASTED International Conference Advances in Computer Science and Engineering (ACSE 2010), pp. 183-188.

Aust, H., Oerder, M., Seide, F., and Steinbiss, V. (1995). The Philips Automatic Train Timetable Information System. Speech Communication 17: 249-262.

Chai, J., Horvath, V., Nicolov, N., Stys-Budzikowska, M., Kambhatla, N., and Zadrozny, W. (2001b). Natural language sales assistant – A web-based dialog system for online sales**.** In Proceedings of the Thirteenth Innovative Applications of Artificial Intelligence Conference, pp. 19-26. The AAAI Press.

Chai, J., Lin, J., Zadrozny, W., Ye, Y., Stys-Budzikowska, M., Horvath, V., Kambhatla, N., and Wolf, C. (2001a). The role of a natural language conversational interface in online sales: a case study". International Journal of Speech Technology, 4(3/4): 285-295.

Colby, K. (1073). Simulation of belief systems. In R. Schank and K. Colby (Eds.), Computer Models of Thought and Language, San Francisco: Freeman, pp. 251-286.

Colby, K., (1975). Artificial Paranoia: A Computer Simulation of Paranoid Processes. New York, NY, USA: Pergamon Press.

Crockett, K., Bandar, Z., O'Shea , J. and Mclean, D. (2009). Bullying and debt: Developing novel applications of dialogue systems. In A. Jönsson, J. Alexandersson, D. Traum and I. Zukerman (Eds.), Proceedings of the 6th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Palo Alto, CA, USA: AAAI (www.aaai.org).

Cunningham, H. (1999). A definition and short history of language engineering. Journal of Natural Language Engineering, 5(1): 1-16.

Dahlbaeck, N., Jonsson, A., and Ahrenberg, A. (1993). Wizard of Oz studies: Why and how? Proceedings of Intelligent User Interfaces (IUI 93). New York, NY, USA: ACM Press. pp. 193-200.

Dybkjaer, L., Bernsen, N.O. and Minker, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. Speech Communication 43(1-2): 33-54.

Elliott D., Atwell E., and Hartley A. (2004). Compiling and using a shareable parallel corpus for MT evaluation. Proceedings of the Workshop on The Amazing Utility of Parallel and Comparable Corpora, Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, pp. 18-21.

Ghande, S., Gordon, S. and Traum, D. R. (2006). Improving question-answering with linking dialogues. Proceedings of the 11th International Conference on Intelligent User Interfaces (IUI '06). New York, NY, USA: ACM Press, pp. 369-71.

Gandhe, S. and Traum, D. (2007). First steps towards dialogue modeling from an un-annotated human-human corpus. Proceedings of the 5th Workshop on Knowledge and Reasoning in Practical Dialogue Systems, Hyderabad, India, pp. 22-27.

Glass, J., Polifroni, J., Seneff, S., and Zue, V. (2000). Data collection and performance evaluation of spoken dialogue systems: The MIT experience. In Proceedings. Intl. Conf. on Spoken Language Processing, Beijing, China, October 2000.

Güzeldere, G., and Franchi, S. (1995). Dialogue with colourful "personalities" of early AI. Stanford Electronic Humanities Review, 4(2): 1-9.

Hasida, K. and Den, Y. (1999). A synthetic evaluation of dialogue systems. In Y. Wilks (Ed.), Machine Conversations. Boston/Dordrecht/London: Kluwer, pp. 113-126.

Hirschman, L. (1995). The roles of language processing in a spoken language interface. In D. Roe and J. Wilpon (Eds.), Voice Communication between Humans and Machines. Washington, DC: National Academy Press, pp. 217-237.

Hirschman, L. and Thompson, H. (1997). Overview of evaluation in speech and natural language processing. In R.A. Cole, J. Mariani, H. Uzkoreit, A. Zaanen, and V. Zue (Eds.), State of the Art in Natural Language Processing . Cambridge, UK: Cambridge University Press, pp. 475-518.

Hughes, J., and Atwell, E. (1994). The automated evaluation of inferred word classifications, In A.G. Cohn (Ed.) Proceedings of ECAI'94: 11th European Conference on Artificial Intelligence, Chichester, UK: John Wiley, pp. 535-540.

Inui, N., Koiso, T., and Kotani Y. (2002). Using patterns for syntactic parsing, Proceedings of IASTED International conference Artificial Intelligence and Applications, pp.522-527.

Inui, N., Koiso, T., Nakamura, J., and Kotani, Y. (2003). Fully corpus-based natural language dialogue system, AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue, AAAI Technical Report S-03-06, Palo Alto, CA, USA: AAAI, pp. 58-64.

Kelly, D., Kantor, P., Morse, E., Scholtz, J. and Sun, Y. (2006). User-centered evaluation of interactive question-answering Systems, Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006, June, Stroudsburg, PA: Association for Computational Linguistics, pp. 49-56.

Koiso, T., Ikeda, T., Inui, N., and Kotani, Y. (2002). A dialog system which chooses a response using similarity between a surface case rule patterns. Proceedings of the IPSJ Conference, IM-03, 2002.

Kruschwitz, U., De Roeck, A., Scott, P., Steel, S., Turner, R., and Webb, N. (1999). Natural language access to yellow pages. Proceedings of Third International conference on knowledge-based intelligent information engineering systems, pp.34-37 .

Kruschwitz, U., De Roeck, A., Scott, P., Steel, S., Turner, R., and Webb N. (2000). Extracting semistructured data-lessons learnt. Proceedings of the 2nd international conference on natural language processing (NLP2000), pp. 406-417.

Maier, E., Mast, M. , and Luperfoy, S. (1996). Overview. In E. Maier, M. Mast, and S. Luperfoy (Eds.), Dialogue Processing in Spoken Language Systems, Berlin: Springer, pp. 1-13.

McTear, M. (2002). Spoken dialogue technology: Enabling the conversational user interface. ACM Computing Surveys, 34(1): 90-169.

Mikic, F. A., Burguillo, J. C., Rodríguez, D. A., Rodríguez, E., and Llamas, M. (2008). T-BOT and Q-BOT: A couple of AIML-based bots for tutoring courses and evaluating students. Proceedings of 38th ASEE/IEEE Frontiers in Education Conference, pp.S3A-7-S3A-12.

Quarteroni, S. (2008). Personalized , interactive question answering on the Web. Proceedings of the Workshop on Knowledge and Reasoning for Answering Questions (KRAQ '08), COLING 2008, Stroudsburg, PA: ACL, pp. 33-40.

Quarteroni, S., and Manandhar, S. (2006). User modeling for adaptive question answering and Information retrieval. G.C.J. Sutcliiffe and R. G. Goebel (Eds.), Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS- 19), Melbourne Beach, FL, USA, May 2006, pp. np.

Quarteroni, S. and Manandhar, S. (2007). A chatbot-based interactive question answering system. Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 11), Rovereto, Italy (DECALOG 2007), pp. 83-90.

Rayson, P. (2003). Matrix: a statistical method and software tool for linguistic analysis through corpus comparison. Ph.D. thesis, Bailrigg, Lancaster, UK: Lancaster University.

Schuetzler, R., Grimes, G. M., Giboney, J., and Buckman, J. (2014). Facilitating natural conversational agent interactions: Lessons from a deception experiment. International Conference on Information Systems. Auckland, New Zealand. December 14-17, 2014, pp. 1-16.

Shaalan, K. (2014). A Survey of Arabic Named Entity Recognition and Classification. Computational Linguistics, 40(2): 469-510.

Traum, D.R., Swartout, W., Marsella, S. and Gratch, J. (2005). Virtual humans for non-team interaction training. Proceedings of the AAMAS Workshop on Creating Bonds with Embodied Conversational Agents.

Van Zaanen, M., Roberts, A., and Atwell, E.S. (2004). A multilingual parallel parsed corpus as gold standard for grammatical inference evaluation. Proceedings of the Workshop on The Amazing Utility of Parallel and Comparable Corpora. Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, pp. 58-61.

Walker, M., Litman, A., Kamm, D., and Abella, A. (1997). Evaluating interactive Dialogue systems: Extending component evaluation to integrated system evaluation. Proceedings of the ACL/EACL Workshop on Spoken dialogue Systems. pp. 1-8.

Wallace, R. (2003). The Elements of AIML Style. ALICE A.I. Foundation.

Weizenbaum, J. (1966). ELIZA-A computer program for the study of natural language communication between man and machine. Communications of the ACM, 10(8): 36-45.