

Understanding Spoken Language of Children Interacting with an Embodied Conversational Character

Manish Mehta¹ and Andrea Corradini^{2 1}

Abstract. Children spoken language understanding and recognition is a little researched area. It is a difficult task and more error-prone than adult speech processing because typical problems such as utterance variation, and a variety of disfluencies are accentuated in the case of children language. In this paper, we present a spoken language understanding architecture, which attempts to effectively deal with children speech. The approach is tailored for a real-time application where children interact with a conversational character via spoken language and 2D gesture. A user study is conducted to evaluate the effectiveness of the approach and the feasibility of constructing a robust spoken dialogue system for a children application. The results of the evaluation are encouraging and suggest that we are on the right track in terms of our approach for robust understanding of children speech.

1 Introduction

Advances in human language technologies and computer graphics techniques, have led to a growing trend towards designing agent based interfaces which exhibit human-like behavior and appearance. These interfaces, termed embodied conversational agents (ECAs) [5], have the ability to communicate through multiple modalities including spoken language, facial expression and gestures. They aim to use and realize cues inherently peculiar to human-human communication, such as sense of presence, mixed initiative and non-verbal behaviors to hold up their end of the dialogue with the user. A growing research community is working on the development of ECAs [4, 19]. So far though, there has been little reported work in spoken language understanding for ECAs. Much of research on embodied agents has concentrated on achieving lifelike qualities to meet the goals of believability. Effective understanding of verbal input is a key requirement for ensuring a smooth conversation with ECAs. To present an approach for understanding spoken language of children, communicating with an ECA is the prime motivation for this paper.

Advances in speech technologies in the 1990s led to the deployment of several prototype and commercial spoken dialogue systems. These systems were used, for example, to provide weather forecast information over the telephone [30], automatic call routing followed by processing for information retrieval and form filling purposes [7], access to information in a large directory database [3], and getting train travel information [17]. The primary users of these systems were adults and the main goal was to help the user complete one or several tasks. Although we have still not been able to develop perfect speech recognition for adults, these systems are able to establish the feasibility of developing successful spoken dialogue systems

even with imperfect recognition. They owe their success in part to robust natural language understanding and efficient dialogue recovery techniques. Speech technology has also been used in recent years to develop systems for children. These systems have ranged from tutoring on pronunciation [26], coach for oral reading [21] and commercial toys [28, 27]. Due to inherent issues related to children speech recognition, these systems have a limited vocabulary and understanding abilities. Apart from a few attempts [22, 8], there has been very little development on robust understanding techniques though. Robustness approaches have mainly been presented for a general spoken dialogue system [29, 9]. The issue of robustness becomes much more pronounced for children spoken dialogue applications as in this case speech recognition is subject to error rates that are two to five times higher than adult speech [1, 6, 18].

The development of our natural language understanding system, supported by a careful evaluation, represents a contribution in this direction. We envision a game scenario of a player interacting with the fairy-tale character Hans Christian Andersen (HCA) in an entertaining and educational way. There is no visible user avatar, as the user perceives the world around him in a first-person perspective. She can explore HCA's study and talk to him, in any order, about any topic within HCA's knowledge domains, using spontaneous speech and mixed-initiative dialogue. The user can change the camera view, refer to and talk about objects in the study, and also point at or gesture to them. Typical input gestures are markers like, e.g., lines, points, and circles. entered at will via a mouse-compatible input device or using a touch-sensitive screen. HCA's domains of discourse are: HCA's fairy tales, his life, his physical presence in his study, the user, HCA's role as gate-keeper for access to the fairy tale world, and the meta domain of solving problems of meta-communication during speech/gesture conversation. The targeted users of the system are kids between the age group of 10-18 years including both native and non-native English speakers.

The rest of the paper is organized as follows. Section 2 presents the spoken language understanding approach. We then report on the results from our evaluation study in Section 3 and conclude in Section 4.

2 Natural Language Understanding

The natural language understanding (NLU) method is based on a layered approach where each stage adds to the information of the previous stage and does a further deeper analysis of the spoken utterance. The goal is to send the results from any stage across to the dialogue module, in case any succeeding stage fails to analyze the utterance due to the presence of speech recognition errors. This results in a graceful degradation of the understanding approach and thereby supports the objective of robustness. The first stage inside

¹ 1. Georgia Institute of Technology, Atlanta, 2. University of Potsdam, Germany email: mehtama1@cc.gatech.edu, andrea@ling.uni-potsdam.de

the NLU performs a shallow level of analysis by searching for certain domain-related key phrases in the user utterance. This ensures that when faced with misrecognized utterances, the key phrases are extracted, and a wider acceptance of utterances is achieved. The key motivation for a shallow level of analysis is that children spoken language utterances is usually casual and full of ungrammatical phenomena. Thus focusing on the key phrases at first, ensures that a partial (and in sometime even full) understanding of user intention can still be achieved in case of a ungrammatical utterance. In the next stage, a combined rule-based and data-driven semantic/syntactic analysis is used. Here, the syntactic constraints are relaxed as needed through parameterized rule settings, semantically meaningful parts of the input utterance are extracted and the domain irrelevant parts are ignored. The use of a complete syntactic analysis of the utterance would have little success for casual spoken language utterances and a sole semantic driven approach would lose on important cues provided by syntax. The use of a combined approach helps us achieve the objective of accommodating spontaneous speech input and utilize syntactic knowledge as needed.

In terms of constituent components, the NLU module (see Fig. 1) can be broken down into four main components: a Keyphrase Spotter, a Semantic Analyzer, a Concept Finder, and a Domain Spotter. The NLU Manager is responsible for communication across different components of the entire module. It receives the user utterance from the Speech Recognizer (SR) and sends it to the keyphrase spotter.

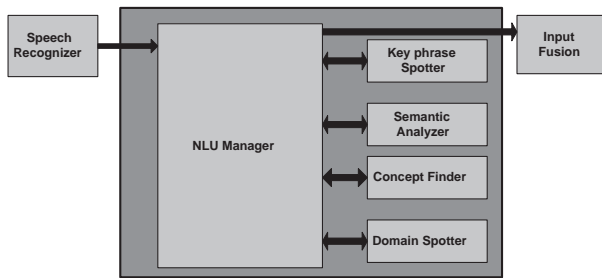


Figure 1. Schematic view of the NLU module.

2.1 Keyphrase Spotter

The Keyphrase Spotter forms the first stage of shallow level of processing inside the NLU and detects multi word expressions from the spoken utterance. It stores a pre-fixed set of words labelled with semantic and syntactic tags. These tags are also used in the sentence-level parsing. Table 1 shows keyphrases used for two semantic categories. Keyphrases are categorized per domain. The Keyphrase Spotter uses about 1200 domain dependent and 150 domain independent keyphrases that have been extracted after manual transcription of dialogue data collected in several Wizard of Oz (WOZ) studies and artificial domain related corpora. After detecting the keyphrases from the user utterance, the processed utterance is sent to the semantic analyzer for further processing.

2.2 Semantic Analyzer

Roughly speaking, the Semantic Analyzer consists of the following four components: a Number Finder, a Lexicon, a Rule Engine, and a Finite State Automaton (FSA) processor. As the first step, the

SEMANTIC : <Location:birthplace>	SEMANTIC : <Time:past>
"place of birth"	"in the past"
"birth place"	"long time back"
"where you were born"	"some time back"
"where you was born"	"few years ago"
"your place of origin"	"long time ago"
"place you were born"	"long long years back"
"place you grew"	"long years back"
"where you grew"	"back in time"

Table 1. A few keyphrases for two different semantic categories.

Number Finder detects all possible dates, age, and numerals in the user utterance. Next, the syntactic and semantic categories for single words are retrieved from a Lexicon. An entry in the Lexicon contains knowledge about each word in terms of its syntactic and semantic category. Words that occur in more than one semantic category have their actual meaning detected through rules as described later. Inflected forms of verb and morphological variants of a word such as plurals of a noun are provided with a separate entry. Relationships of synonymy are encoded by having the same semantic category for the involved words. These synonyms are near synonyms in the strictest sense. Words like 'a', 'the', 'between', 'anything' among others have not been provided with a lexical entry as they are domain irrelevant and do not help in formulating the rules based on semantic/syntactic categories. This also follows from the general principle of the approach to extract only the meaningful parts of the utterance.

Semantic relationships are encoded through a parent-child relationship in the Lexicon. This semantic categorization allows for inheriting the properties of the categories and subcategories of which a word is a member. Some 1400 lexical entries have been created manually after careful analysis of data collected from the WOZ studies and artificial corpora. Words without a lexical entry are not analyzed further and dropped. Nevertheless, these words are recorded so that they can be analyzed later on for making a lexical entry, if needed.

After retrieving the lexical entries the original utterance turns into a sequence of semantic and syntactic categories. This is next sent to the rule engine. The analysis performed here involves choosing an appropriate meaning representation according to a set of grammar rules that are defined on the basis of the existing semantic and syntactic categories. The hierarchical representation of lexical knowledge through the lexicon helps in inferencing through these rules. The rules also help in performing Word Sense Disambiguation (WSD); a crucial component of any natural language understanding module. As part of the human language understanding process, WSD represents the part which finds the appropriate meaning for an ambiguous word within a sentence from a range of possibilities [16]. For instance, the rules help in disambiguating between HCA's mother and the mother of a fairy tale character as shown in example 3 and 4 in Table 6.

The rule engine rewrites certain semantic/syntactic categories (or category sequences) in terms of other semantic/syntactic categories (or category sequences). The conversion is dependent on certain predefined conditions such as e.g. the presence or absence of given categories at specific positions in the processed user utterance. A typical rule is of the form

```
<aux:all><hca>      :- <question:yes/no><hca>
apply_at_position    :- [beginning]
Number_of_conditions :- 0
```

It makes sure that wherever the sequence of categories

Dialogue act	Dialogue act type	Meaning of dialogue act
User_Opinion	Positive	User agreement (e.g. "yeah sure I like Ugly Duckling")
User_Opinion	Negative	User disagreement (e.g. "No I don't want to talk about it")
User_Opinion	General	User opinion (e.g. "I like Ugly Duckling")
User_Opinion	Praise	User appreciation (e.g. "that is cool", "your fairytales are great")
User_Opinion	Thanks	User expression of gratitude (e.g. "Thanks for telling the story")
User_Opinion	Insulting	User response is socially inappropriate (e.g. "none of your business," "are you dumb")
Question	Location	User question concerning a location (e.g. "where did you write Ugly Duckling", "which place were you born")
Question	Reason	User question inquiring the reasons about something (e.g. "why did you leave Copenhagen", "how come you like Ugly Duckling")
Question	Yes/no	User question of yes/no type (e.g. "do you like Ugly Duckling", "is this your study")
Question	Person	User question about a person (e.g. "who was your father")
Question	General	General user question (e.g. "what are you doing")
Question	Time	User question concerning a temporal event (e.g. "when did you write that book", "in which year did move to Copenhagen")
Request	Listen	User request of hearing something (e.g. "could you tell me...", "I want to hear about...")
Request	Tell	User request of telling something (e.g. "I want to tell you...", "can I tell you...")
Meta	Repetition	User repetition/rephrase request (e.g. "could you please repeat", "say it again")
Meta	Clarification	User clarification request (e.g. "I don't get your point", "What does that mean")
Meta	Correction	User correction (e.g. "I think you misunderstand...")
Greeting	Beginning	Conversation start/resume (e.g. "Hey there")
Greeting	Ending	Conversation end (e.g. "see you later", "catch you later")

Table 2. Dialogue acts detected by the NLU from the user utterances.

<aux:all><hca> is detected at the 'beginning' of a processed user sentence it is converted into <question:yes/no><hca>. The rule also states that there is no condition which affects its application. The subfield 'all' in category <aux:all> specifies that the rule is applicable for all the auxiliary verbs like 'can', 'shall' etc. This provides a good generalization mechanism and the rule need not be created for all the auxiliaries individually. We have defined about 290 rules inside the rule engine out of which about 90 rules are domain independent.

This stage also involves detecting the dialogue acts listed in Table 2. The figure also shows the representation of social cues like thanks and praises in user utterance in terms of dialogue acts along with three types of user-initiated meta communication i.e. clarifications, repetitions and corrections.

The tussle for robust semantic interpretation generally results in an entire disregard for syntactic constructs and knowledge in the understanding component for spoken dialogue systems. We combine syntactic knowledge as needed for our purposes. Although the grammatical correctness of the sentence is ignored inside the rule engine, rules based on the syntactic categories are used to infer, for instance, the yes/no types of questions, declarative and imperative structures. The rule considered for detecting the dialogue act 'question' of type 'yes/no' uses the syntactic information regarding auxiliaries. Tense also holds importance as in e.g. distinguishing between "It was a pleasure meeting you" and "it is a pleasure meeting you" where in the former utterance the user bids goodbye and in the latter the user praises seeing HCA.

The next stage of processing inside the FSA processor involves a match procedure on the sequence of semantic/syntactic categories with a set of FSA developed through an offline process at design time. If the sequence of semantic and syntactic categories is able to traverse an FSA, the resultant corresponding to that FSA is a frame containing slot-value pairs filled with information extrapolated from the sequence at hand. The resultant, then, becomes the concise semantic representation of the original user utterance. In case of no match, the original sequence of semantic/syntactic categories

remains unchanged.

The FSAs are developed offline using training utterances from the Woz data and handcrafted artificial corpora related to the domain. To create the FSAs, the training utterances are sent through the key phrase spotter and the lexical entries for the words present in the utterances are retrieved from the lexicon to convert them into a sequence of semantic/syntactic categories. Next, the domain independent rules are applied on the sequence. The resultant is then stored in the form of an FSA. Each set of FSAs is tagged at design time with its own semantic representation. The semantic representation is a combination of fixed semantic categories with values already defined and semantic categories with placeholders. These placeholders are filled from the values obtained through the processed user utterance at run time.

Id.	Training Sentence
45.	What is the relationship between your fairy tales and your life
46.	Which one of your fairy tale is inspired from you own life
47.	Where did inspired you to write the Ugly Duckling
48.	Are any of your fairy tales which reflects your life
49.	Why did you write the Ugly Duckling
50.	What inspired you to write the Ugly Duckling
51.	Can you tell me what Ugly Duckling means to you
52.	What does your fairy tale about the Ugly Duckling mean to you

Table 3. Set of training utterances for constructing FSAs: fsaexample

In order to illustrate the process of creating the FSAs, let us consider the sentences given in Table ???. The resultant sequences corresponding to the training utterances are shown in Table 4. Certain words like "any" and "between" are not provided with a lexical entry and are dropped during this processing. The resultant sequences are further processed to remove duplicate sequences and to combine together often occurring category sub-sequences. This processing stage yields the results shown in Figure 2 which, for

sake of simplicity, does not show the node corresponding to the dialogue act at the beginning of each FSA. These FSAs sets are tagged with a semantic representation like <question:x><verb:x> <fairytale:x><fairytale:relationshiplifefairytale> where the slot x is filled in with the actual values from the user utterances at run time. Table 5 represents a simplified internal storage format.

This data-driven methodology helps in cases where sufficient data is available to develop the FSAs for that particular domain or a subset of it and thus helps avoid the need of handcrafting the rules. This also allows for a rule based and a data driven approach to co-exist so that the objectives of a) realistic development time upfront and b) the use of more data driven methodology as more data becomes available through WOZ experiments can be simultaneously supported. Currently, we use FSAs for a subset of HCAs 'life' and 'fairytale' domains.

No.	Resulting Representation
45.	<question:general><relation><hca><fairytale:general><hca><life>
46.	<question:general><hca><fairytale:x><inspire><hca><life>
47.	<question:location><inspire><hca><verb:x><hca><life>
48.	<question:yes/no><hca><fairytale:x><relation><hca><life>
49.	<question:reason><hca><verb:x><fairytale:x>
50.	<question:general><inspire><hca><verb:x><fairytale:x>
51.	<question:yes/no><hca><fairytale:x><relation>
52.	<question:general><hca><fairytale:x><fairytale:x><relation>

Table 4. The set of training utterances after they have been processed through the keyphrase spotter and the lexicon.

No.	Stored as	No.	Stored as	Id	Encoding
45.	[3][1][2]	46.	[1][3][2]	[1]	<hca><fairytale:x>
47.	[3][4][2]	48.	[1][3][2]	[2]	<hca><life>
49.	[4][5]	50.	[3][4][5]	[3]	<relation><inspire>
51.	[1][3]	52.	[1][5][3]	[4]	<hca><verb:x>
				[5]	<fairytale:x>

Table 5. Simplified format to store sequences. Numbers in brackets are shortcuts to sequence of categories as defined in the rightmost column.

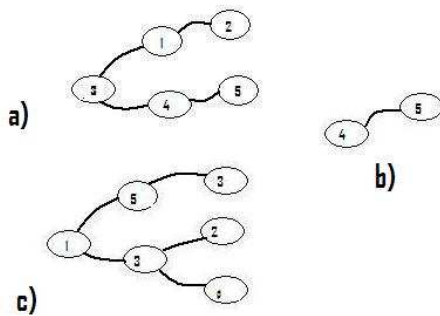


Figure 2. FSA for sentences through 45 to 52 as defined in Table 5.

The output representation of the user utterance in terms of the dialogue acts and semantical output representation is next sent to the concept finder for further processing.

2.3 Concept Finder

In a conversational system, the domain knowledge has to be connected to linguistic system levels of organization such as grammar and lexicon. Domain ontology captures knowledge of one particular domain and serves as a more direct representation of the world. A domain ontology would be limited when used in applications such as language processing. Ontological relationships 'is-a' and 'a-kind-of' have their lexical counterparts in hyponymy. The part-whole relationships meronymy and holonymy [15] also form hierarchies. This relationship parallelism would suggest that lexical relationships and ontology are the same but a lexical hierarchy might only serve as a basis for a useful ontology and can at most be called an ersatz ontology [14]. [2] provides an interesting discussion of the relationship between domain and linguistic ontologies. In our architecture we use two different sets of representations to support the two contrasting objectives of semantical and domain level representation. The concept finder helps in providing a link between the two representations.

The concept finder works by providing a mapping from the semantic categories present in the user utterance to the domain level concepts and properties used inside the dialogue module through rules defined on the semantic categories. To illustrate better, let us consider the fourth example of input processing in Table 6.

NLU Sub-module	Submodule Output
SR Output	what did your father do to earn a living
Keyphrase Sp.	what did your father do to <profession:gen>
Rule engine	<question:general><family:father><profession:gen>
FSA Proc.	<question:general><family:father><profession:gen>
SR Output	What is the big deal about your life
Keyphrase Sp.	What is the big deal about your life
Rule engine	<Question:general><size:big><deal><hca><life:general>
FSA Proc.	<Question:general><personal_attribute:self_identity>
SR Output	Who is the mother of the swan in the fairytale The Ugly Duckling
Keyphrase Sp.	Who is the mother of the swan in the fairytale the <fairytale:ugly_duckling >
Rule engine	<question:person><fairytalecharacter:mother_duck><fairytale:ugly_duckling >
FSA Proc.	<question:person><fairytalecharacter:mother_duck><fairytale:ugly_duckling>
SR Output	What do you remember about your mom
Keyphrase Sp.	What do you remember about your mom
Rule engine	<question:general><verb:remember><family:mother>
FSA Proc.	<question:general><verb:remember><family:mother>

Table 6. Four examples of processing through the NLU.

The concept finder maps the categories after the semantic analyzer into appropriate domain level concept/property used in the dialogue module and converts them into categories of the type e.g. dialogue_act:question, dialogue_act_type:general, property:cognitive, property_type:remember, concept:family and sub_concept:mother. This connects the two different representation used inside the NLU and the dialogue module. Some of the concepts and properties used by the dialogue module and NLU as part of its output representation are shown in Table 7.

The output from the concept finder hence, consists of the following three categories:

a) dialogue_act/dialogue_act_type: These categories are filled with

Concept	Sub_concept
Lifetime	birth,childhood,youth,adult,old,death
Family	parents,mother,father,children,grandfather,grandmother,wife,general,brother,sister, children,son,daughter
Personal_Attribute	self_identities,knowledge,interests capabilities,intelligence
Location	school,neighbourhood,house,study,apartment,birthplace,hometown
Property	Property_type
Number	few,lot,all,cardinals
Quality	good,bad
Emotion	funny,scary,happy,sad
Cognitive	know,remember,forgot
Activity	read,write,play,sing,act,dance,jump,walk,run,current_activity,hobby,travel

Table 7. *Life, physical-self concepts and shared properties.*

the actual values identified from the user utterance. The values provided by the NLU are listed in Figure 2. All in all, 5 `dialogue_act` and 19 `dialogue_act_type` are delivered from the NLU.

b) `concept/sub_concept`: These categories are filled with the actual values of `concept` and the corresponding `sub_concept` from the user utterance. A few examples of `concept/sub_concept` couples are listed in Table 7. The NLU delivers about 18 `concepts` and 150 `sub_concepts` as part of its output.

c) `property/property_type`: These categories are filled with the actual values of `property` and the corresponding `property_types` detected from the user utterance. A few examples of `property/property_type` couples are listed in Table 7. The NLU delivers about 25 `property` values and 55 `property_type` values as part of its output.

2.4 Domain Spotter

At the next stage, the domain spotter uses the output of the the previous processing step to find the domain(s) corresponding to the user utterance. A domain helps to provide a categorization of the character's knowledge set. HCA's domains of discourse are: HCA's fairy tales, his life, his physical presence in his study, the user, HCA's role as gate-keeper for access to the fairy tale world, and the meta domain of solving problems of meta-communication during speech/gesture conversation. Concepts and properties are grouped according to the domain of conversation. The domain spotter processes the output from the concept finder and finds the domain(s) by mapping the `concept/property` according to their respective domains. As an example, the `concept 'family'` shown in Table 7 is grouped under the domain 'life' at design time. Sticking to example 4 in Table 6, when the domain spotter spots the `concept:family` in the utterance, it detects that the utterance belongs to the domain 'life'.

The final output consisting of `concept/subconcept`, `property/property_type`, `dialogue_act/dialogue_act_type` and `domain` is used by the dialogue module to find the next conversational move of the character.

2.5 Conversational Mover

One of the challenges in developing a spoken dialogue system for conversational characters is to make system components communicate with each other. In our architecture, the ontological representation of the user input in terms of the `concepts`, `properties` and `dia-`

logue acts addressed in the utterance is delivered by the natural language understanding to the dialogue module. In the dialogue module, this information is used to find the next conversational move of the character. This stage of processing is performed inside a module called the conversational mover. For each conversational move of the character, rules are defined using the `concept/sub_concept`, `property/property_type` and `dialogue_act/dialogue_act_type` delivered by the NLU. This provides a systematic way to connect the user intention to the characters output move that is influenced by the personality of the character as well as facts about his life. Table 8 shows an example of rule while turn 8 in Figure 3 shows the corresponding user utterance and the NLU output before this move is triggered inside the conversational mover. Table 8 lists other examples of processing starting from the user utterance, the NLU and the conversational mover.

This stage is also responsible for detecting whether the user utterance is overdetermined, i.e., if it is a question whose answer the character currently does not have knowledge about. The following three types of overdeterminations are often encountered: a) `location_unknown`, for example when the user says "where did you write Ugly Duckling". b) `time_unknown`, for example when the user utters "when did you go to Italy" c) `reason_unknown`, for example when the user asks "why did you become a writer". To make clearer the role of overdeterminations, let us consider a sequence of conversation turns from the user study. At some point during conversation the user says "tell me about the Little Mermaid", the utterance is processed by the NLU and the conversational mover and HCA replies correctly with the story of the Little Mermaid. In the next turn, the user inquires further on the topic by asking "why you write Little Mermaid about". In that case, the conversational mover detects the conversational move `reason_unknown` and HCA correspondingly replies "i don't know why but we could talk about topics like...". This provides a generic way of replying to questions for which the system has not been designed. Any time HCA has to produce a response or initiate a new conversational turn, the dialogue module selects a contextually appropriate output in accordance with the determined conversational move, the conversational history and the emotional state. The processing after this stage is outside the scope of this paper.

```
define conv_move :- fairytale_little_mermaid_story
{
  dialog_act :- request or question and
  dialog_act_type :- listen or general and
  concept :- fairytale and sub_concept:little_mermaid and
  property :- no_value or summary
}
```

Table 8. An example of a rule for a conversational move.

3 Evaluation Experiment

The success of a multimodal conversational system depends on many factors. Among others, these include a good recognition performance, provided by an appropriate level of acoustic and language modelling, a robust understanding of the user utterance, and an efficient communication among the different system components that make up the complete architecture. Despite its importance, the evaluation of a multimodal dialogue systems is a poorly defined and understood task. Due to lack of both a consistent framework and sound

U(1):	how is your mother (<i>how was your mother</i>)
S(1):	The name of my mother was Anne Marie Andersdatter. Oh my poor mother!
U(2):	tell me about your family (<i>tell me about your family</i>)
S(2):	I had a nice and small family. Apart from my parents I only saw my grandparents....
U(3):	do you like to play games (<i>would you like to play games</i>)
S(3):	My favorite childhood game was to play theater. When I was eight I saw an opera at the theater in German about a mermaid which I...
U(4):	do you like the fairy tales that you have been written (<i>do you like the fairytale of them yes Andersen</i>)
S(4):	I like several of my fairytales very much. One of my favorites is the Ugly Duckling....
U(5):	Yeah Ok (<i>yeah i</i>)
S(5):	You want to hear about my father
U(6):	Yeah ok (<i>i ok</i>)
S(6):	Excuse me
U(7):	Yes i would like to hear about your father (<i>no yes i would like to hear about your father</i>)
S(7):	Journeyman was my fathers occupation. A journeyman you see is a kind of workman or craftsman employed by somebody....
U(8):	Can you tell me something about the fairy tale the Little Mermaid
SR(8):	<i>(did you tell me something about it to be the fairy tale the Little Mermaid)</i>
NLU(8):	<dialogue_act:request> <dialogue_act_type:listen> <property:summary> <dialogue_act_subject:hca> <concept:fairytale> <sub_concept:little_mermaid>
C. Mover(8):	conversation move: fairytale_little_mermaid_story
S(8):	The Little Mermaid is a story about a mermaid who saves a prince from drowning and falls in love with him. To become human....

Figure 3. Example of interaction from the user study. SR output is in italics. NLU and CM output is shown in turn 8.

theoretical foundations, evaluation is mainly performed through intuitive measures tailored to the application under investigation. In the past, various approaches have been put forward. Some of them deal with the performance of the system as a whole [12, 24] and others are based on the success of single components. The former approaches usually employ a variety of metrics such as task success rate, turn correction ratio, inappropriate utterance ratio, number of turns, concept accuracy, elapsed time and many others [13, 23] in an attempt to evaluate the degree to which the system is accepted by the user. Some other approaches use language input/answer pairs as an evaluation criterion [10], where the correct understanding is defined in terms of the number of right replies to the input sentences. There have also been evaluation methodologies based on the assumption that the performance of the global dialogue system depends on the quality of its single components and their interactions and cooperation with one another [11, 20]. These approaches have employed the subsystem evaluation techniques that consists of breaking down the entire dialogue system into its components and evaluating them independently.

In order to measure the effectiveness of our approach, we conducted an evaluation study with 7 users (4 boys, 3 girls) from the target user group of 10-18 year old kids and teenagers. The users were asked to interact with the fairytale author Hans Christian Andersen in a complete system setup. The recording setup consisted of a ME3 headset microphone, SK100 body pack transmitter and EW100 receiver. Each user test session had a duration of 45 minutes. A session included conversation for about 20 minutes with HCA followed by a post-test interview to evaluate the overall interaction experience of the user. A detailed evaluation based on a structured questionnaire approach was carried out. In this section, we concentrate on the evaluation of the understanding components and report on the parts of the questionnaire related to it. The objects of our attention is the natural language understanding agent that maps input utterances into a formal internal ontological representation as well as the conversational mover agent (see section 2 and subsection 2.5) that selects an adequate move in response to this representation of user intention.

In our investigation, we distinguish between a qualitative and a quantitative evaluation. For the qualitative analysis, we use the classical criterion of precision i.e. the extent to which the selected move,

given the input utterance at hand, is the correct one. We measure the precision as percentage of moves which are correctly selected by the conversational mover, given all user utterances as well as only the user utterances that are successfully processed by the speech recognizer. When humans interact with others, they tend to ignore the fragmented and disfluent qualities of the utterances and focus on extracting meaning in order to make sense of it. The robustness of language (recognition, parsing, etc.) is thus defined as a measure of the ability of human speakers to communicate despite incomplete information and ambiguity. Using that as a benchmark for the success of the NLU and the conversational mover components, two human judges were given the task to independently evaluate the degree of success of the speech recognizer and decide whether a third human being would be able to retrieve the meaning of the input sentence given the speech recognizer output. Interrater reliability for this task was 98%. A third human judge opinion was used to resolve the problem cases.

For the quantitative analysis we look at the single components and how they interact with each other. Starting from the speech recognition subsystems, we calculate well established measure like word error rate (see Table 9). We then analyze speech recognition output in relation with the linguistic aspects of the user utterance in terms of NLU output representation and the correct conversational move, which is the quantitative measurement on the semantic and pragmatic level of our natural language processor that expresses the quality of the understanding system. Our system achieve a 83.9% precision measure (see Table 9) when the speech recognizer provides the right results for a human wizard to make the correct move while this value drop to a 55.3% in case of wrong recognition of speech. From Table 9, it can be easily evinced that, on average, only 7.7% of the input utterances not properly processed by the speech recognizer give rise to a wrong move while the remaining 92.3% do not give rise to any move at all. Regarding the input that is correctly recognized by the speech processor, Table 9 shows how an average 16.1% of it could not be adequately processed by either the NLU (4% of the cases) or the conversational mover (12.1% of the cases). The relationship between the speech recognition performance and robustness can be stressed from the following cases that we encountered during the user interaction with the system.

Subject No.	Total utterances	WER in %	Speech recognition success (percentage)	Speech recognition failure		# of Correct Moves (percentage)	# of no Moves	
				Causing a wrong move	Causing no move		due to NLU	due to Con. mover
1	23	45.5	15 (65.2%)	0	8	13 (86.7%)	0	2
2	27	44.7	15 (55.6%)	1	11	14 (93.3%)	0	1
3	28	47.5	22 (78.6%)	1	5	17 (77.3%)	1	4
4	28	42.9	18 (64.3%)	1	9	15 (83.3%)	1	2
5	33	42.0	24 (72.7%)	1	8	21 (87.5%)	1	2
6	34	42.3	22 (64.7%)	1	11	16 (72.7%)	2	4
7	12	52.7	6 (50.0%)	0	6	6 (100%)	0	0
Average	26.4	45.4	17.4 (65.9%)	0.7	8.3	14.6 (83.9%)	0.7	2.1

Table 9. NLU and Conversational Mover evaluation results from user study.

Use case 1: In this case the input sentence contains a filled pause as well as a repeat. The NLU is able to correctly extract that the user wants to listen to the fairytale Ugly Duckling.

Input: *tell me about <filled pause>tell me about the Ugly Duckling*

SR output: *tell me about you tell me about the Ugly Duckling*

NLU Output : <dialogue_act:request>
<dialogue_act_type:listen><concept:fairytale>
<sub_concept:ugly_duckling>

Conv. Mover : fairytale_ugly_duckling_story

Use case 2: In this case the input sentence is correctly recognized by the speech processor and understood by the NLU. Based on that information the conversational mover is able to find the correct move

Input: *yeah what about games do you like to play games*

SR output: *yeah what about games do like to play games not*

NLU Output : <dialogue_act:question>
<dialogue_act_type:general>
<property:liketoplay><concept:games>
<sub_concept:general>

Conv. Mover : game_story

Use case 3: In this case the input sentence is not correctly recognized by the speech processor and thereby not understood by the NLU, which result in a wrong move being made by the system.

Input: *why*

SR output: *bye*

NLU Output : <dialogue_act:greeting>
<dialogue_act_type:ending>

Use case 4: In this case the recognizer errors change the phrase "I am" into "and" which does not trigger the correct rules in the rule engine to detect the right concept. If the input were not corrupted to change the phrase "I am" at the beginning the NLU representation would have been able to extract the right concepts and the correct move would have been made by the system.

Input: *i am twelve years old*

SR output: *and twelve years old*

NLU Output : <property:age><property_type:general>
<property:number><property_type:12>

Conv. Mover : low_confidence

SR output: *i am twelve years old*

NLU Output : <concept:user_info><sub_concept:age>
<property:age><property_type:general>
<property:number><property_type:12>

Conv. Mover : user_age_info

From the results of the evaluation experiments, two sets of problem patterns are identified. These problem patterns form the principal categories of errors from the SR. They are typically classified as SR failures in Table 9. These cause problems in the processing of the NLU and the conversational mover agents and thereby either results in a wrong move or no move from the system. These two problem categories occur either when SR errors change the important keyword which results in a complete change of user intent (as in use case 3 in the evaluation experiment) or when SR errors result in an ambiguous utterance (as the example in use case 4 from the evaluation experiment). The first category is a more serious issue as was reported in comments during the interview session from a few users. The users complained that some responses from the system did not correspond to their input. These were the cases when the system responded with a wrong move as shown in Table 9. The second category of errors result in the system expressing inability to understand the user as shown in turn 6 in Table 3. This typically occurs when the user repeats or rephrases his utterance. In the second turn, if still a 'no move' is produced, the system uses the strategy of suggesting a different conversation topic to discuss. From Table 9, it can be seen that the first category of errors forms about 2.7% of the total interaction turns whereas the second category forms about 31.5%. Other cases of a wrong move are due to missing rules and keyphrases in 2.6% and missing conversational rules in the conversational mover in 8% of all the conversational turns. Figure 3 shows example of user interaction with HCA from the evaluation study. Turn 8 shows the processing through the internal components.

From the results it can be seen that the speech recognizer needs improvement. The word error rate of 45.4% (see Figure 9) is on higher side even with issues of higher speech recognition error rate for children. This provides us with a future goal of improving its performance through further data collection. However, despite issues with speech recognition, the users were satisfied with the overall interaction experience with the system. In response to the question of how they would evaluate the overall system, 5 users found the system satisfactory. The other two users found the system very good. From the results of the second evaluation study, the positive response could be explained through two factors. Firstly, the methodology used in the NLU of a shallow analysis through the keyphrase spotter at the

first stage and rules which ignore the grammatical correctness inside the rule engine and attempt to extract meaningful information from the user utterance is able to recover on an average about 14.6 utterances from a total of 26.4 (i.e. about 55.3%). Secondly, in presence of recognition errors, the system does not say anything absurd because the NLU and conversational mover did not produce a wrong move. As explained above already, this happens where a 'no move' is selected (31.5% of the cases) and the system expresses the inability to understand by proposing to talk about a different topic in the next turn.

4 Conclusion

We described the spoken language understanding approach for a real time application where children and teenagers interact with an animated character. We presented a method that combines a shallow level of analysis at the first stage followed by a combined rule-based and data-driven semantic/syntactic analysis. In order to test the robustness of the approach, we conducted an evaluation study with the complete system to see the effectiveness of the robustness approach in order to achieve user satisfaction. The results seem to confirm the robustness of the approach. Although the speech recognition performance needs improvement, the users were satisfied with the overall interaction with the system. The successful development of a robust conversational system for children is a major challenge. Viewed in this light, our results seem to suggest that we have taken first steps towards building a robust understanding system for our children application.

Speech recognition performance is a problem in our system. Current evaluation tests have been conducted in a controlled lab setting. Further improvements in speech recognizer would let us carry out a test in a real museum setting. Another development issue in our system is the need of defining several rules to cover all the possible constructs for the spoken language in our domain. Although there have been reports of more robustness of data driven approaches[9], there is still no conclusive agreement upon the preferred use of a data driven approach over a rule based method [25]. We intend to test the two approaches to compare them for our system and use the most effective for our future efforts.

REFERENCES

- [1] G. Aist, P. Chan, X. D. Huang, L. Jiang, R. Kennedy, D. Latimer, J. Mostow, and C. Yeung, 'How effective is unsupervised data collection for children's speech recognition?', in *Proc. of International Conference on Spoken Language Processing*, pp. 3171–3174, (1998).
- [2] J. A. Bateman, 'The theoretical status of ontologies in natural language processing', in *Proc. of Workshop on Text Representation and Domain Modelling - Ideas from Linguistics and AI*, pp. 50–99, (1991).
- [3] B. Buntschuh, C. Kamm, G. Di Fabbrizio, A. Abella, M. Mohri, S. Narayanan, I. Zeljikovic, R. D. Sharp, J. Wright, S. Marcus, J. Shaffer, R. Duncan, and J. G. Wilpon, 'Vpq: A spoken language interface to large scale directory information', in *Proc. of International Conference of Spoken Language Processing*, pp. 2863–2867, (Nov 1998).
- [4] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan, 'Embodiment in conversational interfaces: Rea', in *Proc. of Computer Human Interaction conference*, pp. 520–527, (1999).
- [5] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, *Embodied Conversational Agents*, Cambridge, MA, MIT press, 2000.
- [6] S. Das, D. Nix, and M. Picheny, 'Improvements in children's speech recognition performance', in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, pp. 433–436, (1998).
- [7] A. Gorin, G. Riccardi, and J. Wright, 'How may i help you?', *Speech Communication*, **23**, 113–127, (1997).
- [8] A. Hagen, B. Pellam, and R. Cole, 'Children's speech recognition with application to interactive books and tutors', in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, (Dec 2003).
- [9] Y. He and S. Young, 'Robustness issues in a data-driven spoken language understanding system', in *Proc. of HLT-NAACL 2004 Workshop on Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing*, (2004).
- [10] L. Hirschman, 'Human language evaluation', in *Proc. of ARPA Human Language Technology Workshop*, pp. 99–101, (1994).
- [11] L. Hirschman, 'The evolution of evaluation: Lessons from the message understanding conferences', *Computer Speech and Language*, **12**(4), 249–262, (1998).
- [12] L. Hirschman, M. Bates, D. Dahl, W. Fisher, J. Garofolo, D. Pallet, K. H. Smith, P. Price, A. Rudnicky, and E. Tzoukermann, 'Multi-site data collection and evaluation in spoken language understanding', in *Proc. of Human Language Technology of a ARPA Workshop*, pp. 19–24, (1993).
- [13] L. Hirschman and C. Pao, 'The cost of errors in a spoken language system', in *Proc. of the Third European Conference on Speech Communication and Technology*, pp. 1419–1422, (1993).
- [14] G. Hirst, *Handbook on Ontologies*, chapter Ontology and the Lexicon, 209–230, Springer, 2004.
- [15] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, chapter Lexical Semantics, Prentice-Hall, 2000.
- [16] A. Kilgaraff, 'I don't believe in word senses', *Computers and the Humanities*, **31**(2), 91–113, (1997).
- [17] L. Lamel, S. Rosset, J. L. Gauvin, S. Bennacef, M. Garnier-Rizet, and B. Prouts, 'The limsi arise system', in *Proc. of IEEE 4th workshop on Interactive Voice Technology for Telecommunications Applications*, pp. 209–214, (Sept 1998).
- [18] S. Lee, A. Potamianos, and S. Narayanan, 'Acoustics of children's speech: Developmental changes of temporal and spectral parameters', *Journal of Acoustical Society of America*, **105**, 1455–1468, (1999).
- [19] D. W. Massaro, A. Bosseler, and J. Light, 'Development and evaluation of a computer-animated tutor for language and vocabulary learning', in *Proc. of 15th International Conference of Phonetic Sciences*, (2003).
- [20] C. Mellish and R. Dale, 'Evaluation in the context of natural language generation', *Computer Speech and Language*, **12**, 349–373, (1998).
- [21] J. Mostow, A. G. Hauptmann, and F. S. Roth, 'Demonstration of a reading coach that listens', in *Proc. of ACM symposium, User Interface Software Technology*, pp. 77–78, (1995).
- [22] S. Narayanan and A. Potamianos, 'Creating conversational interfaces for children', *IEEE transaction on Speech and Audio Processing*, **10**(2), 65–78, (2002).
- [23] J. Polifroni, L. Hirschman, S. Seneff, and V. Zue, 'Experiments in evaluating interactive spoken language systems', in *Proc. of the DARPA Speech and NL Workshop*, pp. 28–33, (1992).
- [24] P. Price, L. Hirschman, E. Shriberg, and E. Wade, 'Subject-based evaluation measures for interactive spoken language systems', in *Proc. of DARPA Speech and Natural Language Workshop*, pp. 34–39, (1992).
- [25] M. Rayner, P. Bouillon, B. A. Hockey, N. Chatzichrisafis, and M. Starlander, 'Comparing rule-based and statistical approaches to speech understanding in a limited domain speech translation system', in *Proc. of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*, (2004).
- [26] M. Russel, B. Brown, A. Skilling, R. Series, J. Wallace, B. Bonham, and P. Barker, 'Applications of automatic speech recognition to speech and language development in young children', in *Proc. of International Conference of Spoken Language Processing*, (October 1996).
- [27] Sensory. Sensory's speech technology helps holiday gift items "talk their talk", http://www.sensoryinc.com/html/company/pr01_11.html, November 2001.
- [28] A. Walker. How new technologies are changing the nature of play, <http://www.cyberwalker.net/features/kids-tech-toys.shtml>, December 2000.
- [29] Y. Wang, 'A robust parser for spoken language understanding', in *Proc. of Eurospeech Conference*, (Sept. 1999).
- [30] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and L. Hetherington, 'Jupiter: A telephone based conversational interface for weather information', *IEEE Transactions of Speech and Audio Processing*, **8**, 85–96, (January 2000).