Using Language Tests and Emotional Expressions to Determine the Learnability of Artificial Languages

Omar Mubin
Department of Industrial Design.
Eindhoven University of
Technology (TU/e).
Eindhoven, The Netherlands.
o.mubin@tue.nl

Suleman Shahid
Tilburg Center for Creative
Computing, Dept. of Comm. &
Information Sciences. Tilburg
University. The Netherlands.
s.shahid@uvt.nl

Christoph Bartneck
Department of Industrial Design.
Eindhoven University of
Technology (TU/e).
Eindhoven, The Netherlands.
c.bartneck@tue.nl

Emiel Krahmer
Tilburg Center for Creative
Computing, Dept. of Comm. &
Information Sciences. Tilburg
University. The Netherlands.
e.j.krahmer@uvt.nl

Marc Swerts
Tilburg Center for Creative
Computing, Dept. of Comm. &
Information Sciences. Tilburg
University. The Netherlands.
m.j.g.swerts@uvt.nl

Loe Feijs
Department of Industrial Design.
Eindhoven University of
Technology (TU/e).
Eindhoven, The Netherlands.
l.m.g.feijs@tue.nl

CHI 2009, April 4 – 9, 2009, Boston, Massachusetts, USA
ACM 978-1-60558-247-4/09/04.

Abstract

The study described hereunder lies within the context of a larger project focusing on the design and implementation of a "Robotic Interaction Language". The research goal of this project is to find the right balance between the effort necessary from the user to learn a new or artificial language and the resulting benefit of robust communication between a robot and the user as a direct consequence of optimized speech recognition. To measure the first criteria we have explored two methods to evaluate language learnability, namely Language Tests and analyzing expressed emotions during interaction in an artificial language. Our results indicate that both have potential in being used as measurement tools for evaluating the learnability of artificial languages.

Keywords

Artificial Languages, Speech Interaction, Emotions, Human Robot Interaction, Language Learnability.

ACM Classification Keywords

H5.2. User Interfaces.

Introduction

Speech is one of the primary modalities utilized in Human Robot Interaction (HRI) and is a vital and

Mubin, O., Shahid, S., Bartneck, C., Krahmer, E., Swerts, M., & Feijs, L. (2009). Using Language Tests and Emotional Expressions
to Determine the Learnability of Artificial Languages. Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI2009),
Boston pp. 4075-4080. | DOI: 10.1145/1520340.1520620

natural means of information exchange [1]. Therefore, improving the status of speech interaction in HRI could lead to more pleasant user-robot-interaction. Speech in HRI is confronted and troubled by various issues pertaining mainly to the use of natural language. These include: Ambiguity in natural dialogue, un-robust speech recognition and un-synchronization between software and hardware [2]. In principle, natural language interaction is intuitive and requires little or no learning. But the scales are tipped over in a hurry when there is a break down, leading to frustration from the user. Generally in speech interfaces the focus is on using natural language and given their unsuitability to HRI, it is perhaps time to find a different balance in the form of a new language. Therefore, using the methodology of research through design we aim for a "Robotic Interaction Language” that addresses various problems of natural language interaction in not only HRI but all systems alike.

Our research model is constructed on the basis of two main goals given that we design a new or artificial language. Firstly it should be learnable by the user and secondly, the language should be optimized for efficient recognition by the robot. In this paper we articulate our efforts in measuring the learnability of a new language for humans in reference to the first criterion. For this we have explored two objective measurement tools: self designed language tests and the intensity/level of emotional expressions while interacting in a language; measured via a Perception Test (where independent judges analyze emotional response). Our study was carried out in the context of a game for children, where they interacted in two artificial languages of varying difficulty: Toki Pona [3] and Klingon [4]. Our hypothesis was that an easier to learn language would result in children scoring higher on our version of the language test and expressing richer emotions. The second hypothesis emerged from the following argument that, when people are interacting in their native language, they tend to be so spontaneous that they can very naturally express additional emotions. However, it could be that a difficult language reduces this possibility, because the use of the language is too cognitively demanding which leads to a very constrained interaction style with no room for the display of emotions. Various methods are used to measure natural language learnability, for e.g. computer simulations [5]. Language Aptitude as defined in [6] is the ability of an individual to learn a foreign language. The use of various standardized tests for measuring aptitude in natural languages is fairly common (for e.g. the TOEFL). However in our case we required an engineering effort to adapt a similar tool for an artificial language. Numerous paradigms exist, that assist in the design of language tests, for e.g. MLAT, VORD and CANAL-F see [7] and [8]. In [7] a comparison concluded that the MLAT testing framework was the most appropriate and efficient instrument to predict language learnability. Consequently, adapting the MLAT testing methodology and utilizing it as a starting point for creating a test for artificial languages became one of the goals of our reported study.

Experimental Design

As a scenario and case study option upon which the "Robotic Interaction language" could be evaluated, we adopted the interaction mechanism in game play for children as suggested in [9], where it was argued that games are an effective tool to elicit emotional response. To evaluate the user perspective with respect to the learnability of foreign languages an experimental study

Mubin, O., Shahid, S., Bartneck, C., Krahmer, E., Swerts, M., & Feijs, L. (2009). Using Language Tests and Emotional Expressions
to Determine the Learnability of Artificial Languages. Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI2009),
Boston pp. 4075-4080. | DOI: 10.1145/1520340.1520620

using a game was carried out in Lahore, Pakistan, with 36 children aged 8-12 years (Male=22, Female =14, average age =10.36, std dev=1.42). In total 18 game sessions were executed, each involving a pair of children. All children had sufficient knowledge of both English and Urdu: their native language.

Game Design

We designed a simple wizard of oz card game where children would guess whether the subsequent number in a sequence would be larger or smaller than the previous number (see Figure 1). When the game would start, players would see a row of six cards on the screen where the number of the first card was visible and the other five cards were placed face down. All the cards ranged from 1 to 10 and a card displayed once was not repeated in a particular sequence. Once players would make a guess, the relevant card was revealed. Players were informed about the correctness or incorrectness of their answer via a characteristic non-speech sound (booing or clapping). A correct guess would earn positive points (+1) and an incorrect guess would result in loosing the particular game. The children were encouraged to discuss with each other in order to attain a consensus about their final guess. All interaction in the game was speech based.

Procedure

The game was run as a power point presentation on a laptop. Entire experiment sessions were video recorded. Written consent was given by the teachers and parents to use the recordings for research purposes. A pair of children was asked to sit in front of a desk on which the laptop was placed. Above the laptop, a camcorder was placed to record the children's faces and their upper body. A monitor connected to the laptop facilitated the wizard in controlling the game. The wizard was located out of the visual field of the game-playing children in another room. After an introductory round, the children were given game instructions and were informed about the points that they could win or lose. After this, the experimenter left the children's field of vision and started the game. At the end of the game session, the experimenter rewarded the children with gifts based on their points.

Interaction in the game via Artificial Languages

Pairs of children would play the number guessing game where the permissible set of game commands and utterances that could be said to the co-player (~ 10 words such as larger, smaller, equal, go to the next number, etc) was predefined in an artificial language (see figure 1 and 2). There were two artificial languages chosen: Klingon (*apparently difficult to learn*) [4] and Toki Pona (*apparently easy to learn*) [3]. Toki Pona is designed on a principle of simplicity and aims to reduce complexity. Its lexical and phonetic vocabulary is considerably small in size (118 words and 14 phonemes). In contrast, Klingon is a unique apriori language and therefore has no influence from any existing natural language. In fact most of its phonemes within the International Phonetic Alphabet (IPA) framework [10] are not found in any of the major natural languages, which is precisely the reason why Klingon has words that should be difficult to pronounce, besides being longer in length. This led us to set the ground truth for our experiment that Toki Pona should be relatively easy to learn and Klingon in comparison more difficult. Each pair of children would play two rounds of the game each in either of the two artificial languages and in a natural language, in this case their native language Urdu. The four orders of language

Mubin, O., Shahid, S., Bartneck, C., Krahmer, E., Swerts, M., & Feijs, L. (2009). Using Language Tests and Emotional Expressions to Determine the Learnability of Artificial Languages. Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI2009), Boston pp. 4075-4080. | DOI: 10.1145/1520340.1520620

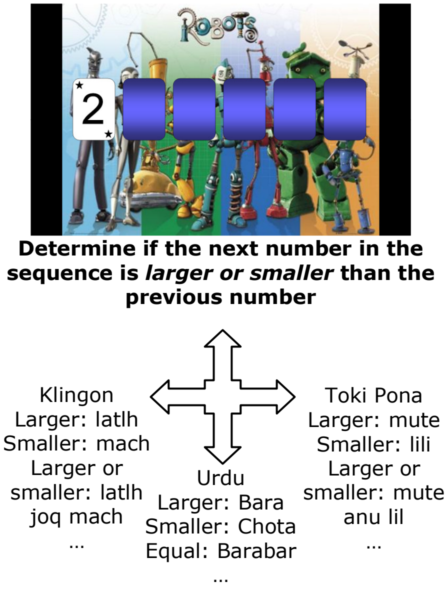
ROBOTS
2
Determine if the next number in the sequence is *larger or smaller* than the previous number
Klingon
Larger: latlh
Smaller: mach
Larger or smaller: latlh joq mach
…
Urdu
Larger: Bara
Smaller: Chota
Equal: Barabar
…
Toki Pona
Larger: mute
Smaller: lili
Larger or smaller: mute anu lil
…

Figure 1. Game Design.

presentation were counter balanced. In total, 9 pairs of children played the game in Klingon and Urdu and 9 pairs in Toki Pona and Urdu. The wizard had knowledge of all the relevant game commands in each of the three languages and would direct the flow of the game accordingly. Prior to playing the game and during the explanation of the instructions phase, each child was given exactly 10 minutes to learn the game commands with the aid of audio clips. During the game, if the children would forget the commands of the artificial language, they could call out for help, but at the cost of exponentially increasing negative points.

A Learnability Test for Artificial Languages

One of the primary objectives of the study was to design and evaluate a self developed language learnability measurement test. Similar language tests were constructed for both Klingon (KN) and Toki Pona (TP), by adapting the framework as in [6]. The tests included ten questions each having only one correct answer. The questions tested the learnability of the artificial language in terms of vocabulary and pronunciation via semantics and rhyming respectively; two of the four language learning abilities advocated in [6]. The tests were handed out at the end of the game playing session. An independent samples t-test revealed that the children who learnt TP performed significantly better on their version of the test than those who played the game in KN, (t (34) = 2.04, p < .05). There was no order effect upon the test scores, (F (3, 32) = 2.2, p = 0.11).

Evaluating Emotional Expressions via a Perception Test

A secondary method to evaluate the learnability of languages was to run a perception test on the game videos. We hypothesized that a language that is easier to learn would be much more enjoyable to interact with and would hence elicit richer emotional expressions.

Procedure and Participants

From the children that played the game, we selected video snippets of a random winning guess and a random loosing guess from each individual game. This was done twice for each of the languages. In addition, from the clips we randomly selected one child from each of the 18 pairs by zooming in on his/her face. In this selection, half of the children sitting on the right chair and half of the children sitting on the left chair were selected. The stimuli were recorded from the moment the card in question was overturned till the primary response of the child was completed. This resulted in 72 stimuli: 2[win/lost] X 2[(KN or TP) and Urdu] X 18 children. Stimuli were presented to participants in a random order, in vision-only format to avoid participants from relying on auditory cues. 30 Dutch adults, with a roughly equal number of men and women, participated in the perception experiments. For the perception test, participants were seated in a classroom where the stimuli were projected on a wall. The participants were informed that they would see stimuli of children who had just won or lost a game. As viewers, they were instructed to guess from the children's facial expression whether the children had won or lost. Each stimulus was preceded by an ID and followed by a 6 second pause during which participants could fill on a form firstly whether they thought it was a wining or losing situation and secondly how expressive did they think the children were on a scale of 1 to 7, with 7 being the most expressive.

Figure 2. Children involved in the game.

Statistical Analysis and Results

Tests for significance were performed using a repeated measures analysis of variance (REMANOVAs) and the Bonferroni method was used for pairwise comparisons. The experiment had a within subject design with language type (levels: natural language Urdu-NL, TP and KN), being the independent variable. The percentage of correct classifications and level of expressiveness as ranked by the participants were recorded as the dependent variables. The REMANOVAs showed a significant main effect of language type (F (2, 58) = 143.220, p < .001, ηp^2 = .832) on correct classification. Pair-wise comparisons revealed a significant difference between the languages. The average of correct classifications was the highest for NL (M = .793), followed by TP (M = .602) and lastly KN (M = .546). Similarly, for the variable 'level of expressiveness', a significant main effect of language type (F (2, 58) = 67.629, p < .001, ηp^2 = .700) was found. Pair-wise comparisons illustrated a significant difference between the languages. The level of expressiveness was ranked to be the highest for NL (M = 5.04), lowest for KN (M = 4.05), with TP (M = 4.40) lying in the middle.

Discussion

The goal of our study was to explore two methods to evaluate the learnability of artificial languages. The results from both objective methods agree with each other. Both were able to highlight the difference between the artificial languages in terms of learnability and consequently engagement and expressiveness.

Firstly, our initial results illustrate that there is potential in utilizing a language test as an instrument to detect the learnability of an artificial language, because it was able to highlight the difference in ease of learnability between TP and KN. However our developed tests were not extensive and care must be taken while interpreting the results as the lexical vocabulary of each of the two languages was not of sizeable proportions. Moreover, mainly due to the aforementioned reason, our language tests did not test all the four cognitive abilities advocated in [6], which can be termed as another limitation of our test.

Secondly, from the results of the perception test it is evident that the children were significantly more expressive in TP as compared to KN, which can also be adjudged by observing the higher number of correct classifications for the case of TP. In the case of TP, it was relatively easier to guess whether children have won or lost the game based on their facial expressions. Therefore our hypothesis that an easier to learn artificial language would elicit more emotions is confirmed. The differences in emotional expression and the accuracy of independent observers to perceive these differences is in line with existing work where emotional expressions of Pakistani children were used as a mean to judge fun or engagement in a game [9]. The level of expressiveness and the average of correct classification were the highest for NL. We would expect this to be the case because, the children were quite comfortable while playing the game in their own language and there was no cognitive load in terms of recalling a new word from a new language which could hinder their expressiveness. The use of natural language in our design was mainly a control condition to check the expressiveness of children across two game sessions and the high expressiveness in natural language and relatively low expressiveness in artificial languages confirms our choice of not only this condition

Mubin, O., Shahid, S., Bartneck, C., Krahmer, E., Swerts, M., & Feijs, L. (2009). Using Language Tests and Emotional Expressions
to Determine the Learnability of Artificial Languages. Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI2009),
Boston pp. 4075-4080. | DOI: 10.1145/1520340.1520620

but also the perception test method. An interesting and potential limitation in the perception test was a cross-cultural element. The stimuli were of Pakistani children and the observers were Dutch adults. This cultural incongruence could have resulted in a winning guess being perceived as a losing one or a losing guess as a winning one, while perceiving the emotional reactions of children. It is known that complex emotional states e.g. shame and guilt in particular, are packed in cultural wrappers and it is sometimes difficult to judge such emotional expressions across cultures [11].

Future Work

As an extension of our reported study we plan to carry out certain steps in the future. Firstly, we would replicate the same study in another culture to validate if language tests and emotional expressiveness can be used cross culturally to determine language learnability. Secondly, we aim to validate and extend our language test for larger vocabularies and larger grammar constructs of artificial languages using validation techniques such as in [12]. Thirdly we aim to explore subjective measurement tools for assessing language learnability. By doing so, we can quantify *the precise criteria that makes a language easy to learn for humans*. Fourthly, we shall investigate speech recognition engines, so that we can move a step towards determining *the exact criteria that makes speech easy for machines to recognize*. As a combination of our endeavors we will then move towards designing the "Robotic Interaction Language".

References

[1] Goodrich, M. A. and Schultz, A. Human Robot Interaction: A Survey. *Foundations and Trends in Human–Computer Interaction*, 1, 3 2007, 203-275.

[2] Kulyukin, V. A. On natural language dialogue with assistive robots. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction* Salt Lake City, Utah, USA, 2006. ACM, 164-171.

[3] Toki Pona. http://www.tokipona.org/.

[4] The Klingon Language Institute. http://www.kli.org/.

[5] Lupyan, G. and Christiansen, M. *Case, word order, and language learnability: insights from connectionist modeling*. 2002.

[6] Carroll, J., Sapon, S. and Corporation, P. *Modern language aptitude test*. Psychological Co New York, 1959.

[7] Parry, T. and Child, J. Preliminary Investigation of the Relationship between VORD, MLAT and Language Proficiency. In *Proceedings of the Symposium on Language Aptitude Testing*, 1990. Prentice Hall, 30-67.

[8] Grigornko, E., Sternberg, R. and Ehrman, M. A Theory-Based Approach to the Measurement of Foreign Language Learning Ability: The Canal-F Theory and Test. *The Modern Language Journal*, 84, 3 2000, 390-405.

[9] Shahid, S., Krahmer, E. and Swerts, M. Alone or Together: Exploring the Effect of Physical Co-presence on the Emotional Expressions of Game Playing Children Across Cultures. In *Proceedings of the Fun and Games* Eindhoven, the Netherlands, 2008. Springer, 94-105.

[10] Ladefoged, P. and Maddieson, I. *The Sounds of the World's Languages*. Blackwell Publishers, 1996.

[11] Breugelmans, S. and Poortinga, Y. Emotion Without a Word: Shame and Guilt Among Raramuri Indians and Rural Javanese. *Journal of Personality and Social Psychology*, 91, 6 2006, 1111-1122.

[12] Alderson, J., Clapham, C. and Wall, D. *Language test construction and evaluation*. Cambridge University Press, 1995.

Mubin, O., Shahid, S., Bartneck, C., Krahmer, E., Swerts, M., & Feijs, L. (2009). Using Language Tests and Emotional Expressions
to Determine the Learnability of Artificial Languages. Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI2009),
Boston pp. 4075-4080. | DOI: 10.1145/1520340.1520620