

The Twins Corpus of Museum Visitor Questions

Priti Aggarwal,¹ Ron Artstein,¹ Jillian Gerten,¹ Athanasios Katsamanis,²
Shrikanth Narayanan,² Angela Nazarian,¹ David Traum¹

¹USC Institute for Creative Technologies, 12015 Waterfront Drive, Playa Vista CA 90094-2536, USA
last_name@ict.usc.edu

²USC Signal Analysis and Interpretation Laboratory, 3740 S. McClintock Ave, RTH 320, Los Angeles CA 90089, USA
{shri,nkatsam}@sipi.usc.edu

Abstract

The Twins corpus is a collection of utterances spoken in interactions with two virtual characters who serve as guides at the Museum of Science in Boston. The corpus contains about 200,000 spoken utterances from museum visitors (primarily children) as well as from trained handlers who work at the museum. In addition to speech recordings, the corpus contains the outputs of speech recognition performed at the time of utterance as well as the system interpretation of the utterances. Parts of the corpus have been manually transcribed and annotated for question interpretation. The corpus has been used for improving performance of the museum characters and for a variety of research projects, such as phonetic-based Natural Language Understanding, creation of conversational characters from text resources, dialogue policy learning, and research on patterns of user interaction. It has the potential to be used for research on children's speech and on language used when talking to a virtual human.

Keywords: speech corpora, children's speech, virtual characters, museum

1. Background

The Twins corpus is a collection of utterances from museum visitors and handlers, spoken in interactions with the Twins, Ada and Grace, a pair of life-size virtual characters who serve as guides at the Museum of Science in Boston (Swartout et al., 2010, see Figure 1). The characters promote interest in Science, Technology, Engineering and Mathematics (STEM) in children between the ages of 7 and 14. The corpus consists of audio recordings of user utterances and logs of user and system actions. A portion of the recordings has been transcribed and annotated.

The Twins process utterances using a pipeline architecture:

1. A user presses a button and speaks into a microphone.
2. An acquisition client sends the audio to a bank of automatic speech recognizers (ASR), which create hypotheses of the words said.

3. The resulting text is sent to a Natural Language Understanding (NLU) module, which identifies a set of appropriate responses from a library of pre-recorded animated character utterances.
4. A dialogue manager (DM) module selects a response based on the NLU output and dialogue history, and sends the response to the animation components.

Speech recognition (2) was performed by the SONIC toolkit (Pellom and Hacıoğlu, 2001/2005) until December 2010, and thereafter by OtoSense, a recognition engine that is currently being developed by USC; using OtoSense allowed appropriate customization of the speech processing front-end for the needs of the setup at the museum. Natural language understanding (3) and dialogue management (4) are integrated in a single component, NPCEditor (Leuski and Traum, 2010), a text classification system that drives virtual characters and is freely available for research



Figure 1: Visitors talking to the Twins at the Museum of Science. Picture from Swartout et al. (2010), Figure 6, page 297.

purposes as part of the ICT Virtual Human Toolkit (<http://vh toolkit.ict.usc.edu>). These components work on individual user utterances, defined by a press and release of the push-to-talk button (1).

2. User Utterances

2.1. Recording

The audio portion of the corpus is composed of utterances recorded in the course of normal operation of the Twins system at the museum. Utterances are recorded directly on the computer hosting the characters, as WAV files sampled at 16 kHz. Initial recordings were made with a wired USB microphone, which was replaced in June 2010 by a wireless Sennheiser microphone; in February 2011 a Shure 522 microphone was added as a second, fixed microphone.

2.2. Composition

Two kinds of users speak to the characters: visitors to the museum and handlers. The latter consist of museum staffers and volunteers that serve as interpreters who help visitors explore the exhibits in Cahners ComputerPlace, the space that also hosts the Twins. In the first year of the Twins' deployment (November 2009–November 2010), most of the interaction between visitors and the characters was mediated by handlers: visitors would present a question to a handler, who would then relay it to the characters by talking into a microphone (possibly modifying the visitor's original question). This allowed for reliable speech recognition while we collected data for training visitor models. Since December 2010 (concurrent with the deployment of the OtoSense speech recognition engine) visitors have been talking to the characters directly, and since February 2011 handlers and visitors use separate, dedicated microphones.

No demographic data are collected about the speakers, and the only information about their identity is what can be inferred from listening to their voice. A separate survey of visitors to the exhibit space, conducted by the Institute for Learning Innovation, found that approximately 50% of the visitors live in Massachusetts, 75% live in the Northeast, and 90% live in the United States.

Parts of the corpus were manually transcribed and annotated; most of the corpus, however, remains unexplored. Figure 2 shows the total number of utterances collected, broken down by month. A total of about 40,000 utterances were collected using the one-microphone setup, mostly from handlers. Close to 250,000 utterances have been collected to date with the two-microphone setup (almost 98% from the visitor microphone). A sampling of 17,244 utterances from April and May of 2011 revealed the following composition (identified by listening to the voice): 47% children, 13% adult male, 8% adult female, and 31% blank. Assuming similar proportions hold across the corpus, we expect the full corpus collected to date to contain about 114,000 utterances by children, 32,000 utterances by adult males, and 20,000 utterances by adult females.

2.3. Transcription

More than 55,000 utterances have been transcribed. Of these, almost 20,000 have been identified as uttered by

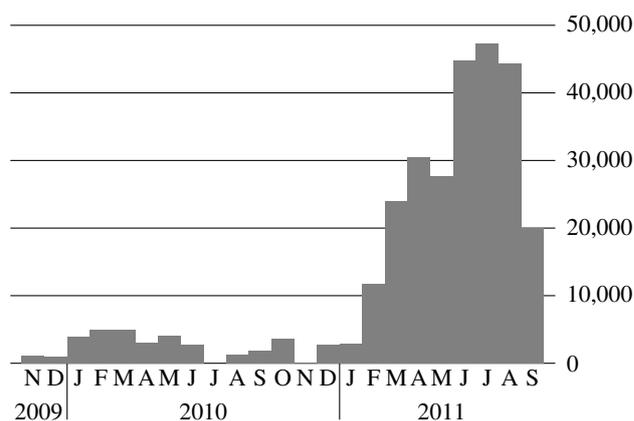


Figure 2: Utterances collected, by month

one of two designated handlers, a little over 13,000 were marked with a speaker identification as child, adult male, or adult female, and for the remaining 22,000 the speaker was not identified. About 32,000 of the transcribed utterances come from the earlier, one-microphone setup, and consist mostly of handler speech; the remaining transcriptions are of recordings made using the later, two-microphone setup, with a higher share of visitor speech.

Many of the utterances spoken to the Twins are known to them, that is they are identical to utterances in the NLU training data; this happens because of the prompting techniques discussed below. Table 1 shows the division of the transcribed data into known and novel utterances. The newer recordings are more varied – they contain a higher proportion of novel utterances, and a similar number of utterances represents more utterance types. Even known utterances are sometimes difficult for the Twins to understand, because speech recognizer output may diverge substantially from the words uttered by the user (see section 4.1). The distribution of utterances is highly skewed; the most common ones are listed in appendix A.

Typical utterances are questions addressed to the Twins, for example *What are you called?* or *What are fossils made of?* The Twins are designed to answer questions about themselves and their technology, about exhibits in the museum, and about science in general. Visitors are prompted to ask such questions through several means.

1. Suggested questions are posted next to the microphone that visitors use to talk to the Twins (Figure 3). Of the 22637 transcribed utterances collected with the double microphone setup (Table 1), 6891 (30%) are identical to one of the posted questions; see also appendix A.
2. Museum staffers make suggestions to visitors about what questions to ask.

Microphone Setup	Utterance types		Utterance tokens	
	known	novel	known	novel
Single	317	4007	21452	11546
Double	176	7393	10660	11977

Table 1: Transcribed utterances

Hi there!

You can ask Ada and Grace whatever you like, but they mostly know about computers, communications, robots, this exhibit hall, and of course, themselves.

Here are some examples to get you started:

1. Hello.
2. Who are you named after?
3. What can you tell me about?
4. Tell me about computers.
5. Are you a computer?
6. What is a virtual human?
7. What do you like to do for fun?
8. Where do you go to school?
9. What is artificial intelligence?
10. What can I do at Robot Park?

Figure 3: Questions posted next to the Twins

3. The Twins themselves prompt visitors for follow-up questions (Figure 4).

Some of the visitor utterances are not phrased as questions, for example *I like to play video games*. There are also visitor utterances that appear to be intentionally inappropriate such as *I'd like a double cheeseburger hold the lettuce*, utterances containing obscene language, and utterances that appear to be directed at someone other than the Twins, for example *Pushing it I'm pushing it pushing it* (presumably referring to the push-to-talk button). The vast majority of utterances are in English, but occasionally visitors try to talk to the characters in a different language – for example, among the transcribed utterances are 8 instances of *¿Cómo estás?* (Spanish for “How are you?”).

2.4. Annotation

In addition to transcription, sample utterances were annotated in order to serve as performance benchmarks for the characters. In the first round of annotation, utterances were run through the Natural Language Understanding module to generate responses; each system output was then manually annotated as to whether it constituted an appropriate response to the user utterance (a binary decision). A reliability sample of 99 utterances was given to 3 annotators; agreement was calculated using α (Krippendorff, 1980) with the result $\alpha = 0.76$, that is fairly good agreement. The remaining utterances were annotated by one coder each; overall, 1321 utterances of 2718 received an inappropriate system response (49%).

In a second round of annotation, those utterances from the first round which received inappropriate responses were annotated as to whether they were appropriate as questions for the Twins (that is, whether they constitute the kind of questions the Twins are expected to answer). A small reliability sample of 100 utterances was given to 3 annotators who achieved a reliability of $\alpha = 0.51$, that is well above chance but not particularly high. This shows that the decision on whether the Twins should be able to answer a particular question is a difficult and somewhat subjective judgment.

Ada	Let's see. What are you interested in? Do you want to learn more about computers, robots, cell phones, hands-on activities. . .
Ada	I sometimes wish more visitors would ask about how we work, instead of asking about everything else.
Grace	Here at ComputerPlace, you can look inside a computer, create your own electronic inventions, find out why there's no such thing as white light, learn the magic behind magnetism, play with robots, and share your own computer animations online. So, what do you want to do?

Figure 4: Some prompts used by the Twins

The full set of utterances with inappropriate responses was then either marked as an inappropriate question, or mapped to the desired responses; overall, 268 utterances of the 1321 were marked as inappropriate questions (20%).

3. Character utterances and interaction

In addition to the user utterances, the corpus contains system logs generated during the interaction between the users and the characters. These logs record a variety of events taking place during the interaction.

1. User actions such as pressing and releasing the push-to-talk button.
2. Speech recognition results delivered by the system in real time.
3. Character actions such as utterances and animations.

All of the events are timestamped, allowing the reconstruction of conversations by combining the logs and the manual transcriptions (Figure 5). All the logging is done within the system that operates the characters, so we do not have records of external aspects of the conversation such as interactions between visitors and handlers, side conversations among visitors, or speech uttered when the microphone is in the off position. For additional examples of interaction between users and the characters, see appendix B.

13:40:35	User	Hello
13:40:36	Ada	Hey there!
13:40:38	Grace	How can we help you?
13:40:43	User	Who are you
13:40:43	Grace	Could you say that again?
13:40:54	User	Who are you named after
13:40:58	Ada	Our namesakes are Ada Lovelace and Grace Hopper.
13:41:02	Grace	Yeah, they may be old fashioned names. But both are from the most modern computing women of their time.

Figure 5: Conversation reconstructed from the logs and transcriptions

4. Usage

The corpus has been used for improving the system deployed at the museum, as well as for general research.

4.1. Speech recognition adaptation

The transcribed audio recordings from the museum visitors have been used for the adaptation of three separate acoustic models, namely for children's, male and female speech (Potamianos and Narayanan, 2003). Adaptation was performed using Maximum Likelihood Linear Regression while the original children's models were trained on the Colorado University children's speech database (Hagen et al., 2003) and the two adult speech models were trained on the Wall Street Journal corpus (Paul and Baker, 1992). Automatic speech recognition using these three models has been systematically evaluated only for a small but representative portion of the Twins corpus comprising 1003 utterances recorded on a single day. The average word error rate was found to be 43% when the best acoustic model is chosen for each utterance. The performance drops by approximately 14% when the acoustic model is automatically selected for each recording based on the confidence of the speech recognition result of three recognition engines running in parallel, each using one of the three acoustic models (see appendix C for additional details). The relatively high word error rate is illustrative of the challenges posed by the specific characteristics of the corpus:

1. Speech is spontaneous, i.e., with frequent hesitations, mispronunciations, and repetitions.
2. Speech is coming mainly from children.
3. There are no vocabulary constraints.

The above characteristics make the corpus an ideal testbed for speech recognition research.

Automatic recognition of the handlers' speech using speaker-adapted models is significantly better on the other hand, with an average word error rate close to 5%. The handlers however mostly select their questions from a large but predefined set and they have been trained to interact with the Twins, which makes their utterances much easier to recognize automatically.

4.2. Character interaction

The transcriptions have been further used for identifying common questions, in order to increase the characters' repertoire and improve their ability to appropriately respond to visitor utterances. Common questions were added to the characters' training data on the input side, and were either linked to existing responses or to newly created responses when there was no appropriate response available. The annotations were used to create a benchmark test set for the NLU component while developing the training data.

4.3. Standard test set for Natural Language Understanding

A portion of the handler utterances (about 6,000) have been used as a test set in other research projects. These are utterances which are identical to ones in the NLU training data,

therefore the exact desired responses are known, but actual NLU performance varies due to errors in speech recognition. This portion of the corpus thus serves as a good test set for the NLU robustness against noise. Wang et al. (2011) used this test set to compare methods for incorporating phonetic knowledge into Natural Language Understanding, finding that adding a phonetic dictionary to the NLU component allows it to recover from certain speech recognition errors. Nouri et al. (2011) tested the effects of expanding the character knowledge base with automatically generated questions and answers on additional topics, effectively adding noise to the NLU training data; the performance of the NLU did deteriorate somewhat, but in return it was able to also answer questions on the new topics.

4.4. Integration of speech recognition and NLU

A portion of the user utterances (about 14,000) have been transcribed and annotated with the appropriate character responses (section 2.4). This allows for research on how to use speech recognition and Natural Language Understanding together in order to achieve the best results. One method involves generating multiple speech recognition hypotheses, interpreting them individually, and then ranking the results using joint information from the speech recognizer and the NLU. Initial experiments show that this method results in better NLU accuracy than interpreting the best ASR result.

4.5. Dialogue policy learning

The corpus is being used for research on policies for dialogue management: individual sessions are identified from timing information in the logs, and user models are trained to characterize actions such as engagement in conversation, disengagement, and topic progression. The user models are then used to train dialogue policies for the character using reinforcement learning. This research is ongoing.

4.6. User interaction patterns

We are conducting research on how character utterances affect subsequent user reaction, using information from the logs and transcriptions of the recorded utterances. This research is in its preliminary stages.

4.7. Future uses

The corpus has potential for use in research on children's speech, from the Boston area in particular. It can be particularly useful for the study and further development of automatic recognition of children's speech. Other potential uses are for studying how museum visitors phrase questions, how they talk to virtual humans, and what topics interest visitors when they face a virtual museum guide.

5. Conclusion

This paper presented a large corpus of utterances spoken by human users, primarily children, to a pair of virtual humans. The corpus consists of speech data and is partly transcribed and annotated. The corpus has already been used in several completed and ongoing research projects, and we anticipate that it will be useful to the research community

at large. We are presently working on organizing and improving the corpus, and we hope to be able to release it. We welcome inquiries about the corpus, which should be sent to the second author.

The Twins corpus serves as a very good example of how Virtual Agents can facilitate the collection of significant amounts of rich and targeted linguistic data from a wide range of speakers.

6. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant 0813541. We thank the staff and volunteers of Cahners ComputerPlace for their support.

7. References

- Andreas Hagen, Bryan Pellom, and Ronald Cole. 2003. Children’s speech recognition with application to interactive books and tutors. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU ’03)*, pages 186–191.
- Klaus Krippendorff, 1980. *Content Analysis: An Introduction to Its Methodology*, chapter 12, pages 129–154. Sage, Beverly Hills, California.
- Anton Leuski and David Traum. 2010. Practical language processing for virtual humans. In *Proceedings of the 22nd Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-10)*.
- Elnaz Nouri, Ron Artstein, Anton Leuski, and David Traum. 2011. Augmenting conversational characters with generated question-answer pairs. In *Question Generation: Papers from the AAAI Fall Symposium*, pages 49–52, Arlington, Virginia, November. AAAI Press.
- Douglas B. Paul and Janet M. Baker. 1992. The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 357–362, Harriman, New York, February.
- Bryan Pellom and Kadri Hacıoğlu. 2001/2005. SONIC: The University of Colorado continuous speech recognizer. Technical Report TR-CSLR-2001-01, University of Colorado, Boulder.
- Alexandros Potamianos and Shrikanth Narayanan. 2003. Robust recognition of children’s speech. *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616, November.
- William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, Chad Lane, Jacquelyn Morie, Priti Aggarwal, Matt Liewer, Jen-Yuan Chiang, Jillian Gerten, Selina Chu, and Kyle White. 2010. Ada and Grace: Toward realistic and engaging virtual museum guides. In Jan Allbeck, Norman Badler, Timothy Bickmore, and Alla Pelachaud, Catherine Safonova, editors, *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20–22, 2010 Proceedings*, volume 6356 of *Lecture Notes in Artificial Intelligence*, pages 286–300. Springer, Heidelberg.

- William Yang Wang, Ron Artstein, Anton Leuski, and David Traum. 2011. Improving spoken dialogue understanding using phonetic mixture models. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference*, pages 329–334, Palm Beach, Florida, May. AAAI Press.

Appendix A. Most common user utterances

The following are the most common user utterances collected with the double microphone setup (Table 1). Percentages are calculated from a total of 22637; boldfaced utterances are identical to one of the posted suggestions (Figure 3).

N	%	Utterance
3076	13.6	hello
1118	4.9	who are you named after
1074	4.7	are you a computer
798	3.5	what do you like to do for fun
417	1.8	hi
399	1.8	what is artificial intelligence
272	1.2	how old are you
256	1.1	what’s your name
175	0.8	what is your name
169	0.7	what are your names
166	0.7	no
155	0.7	where do you go to school
148	0.7	what do you do for fun
147	0.6	what’s your favorite color
134	0.6	where did you go to school
117	0.5	how are you
108	0.5	who are you
102	0.5	what is a virtual human
89	0.4	bye
78	0.3	good morning
76	0.3	goodbye
75	0.3	what
74	0.3	yes
73	0.3	are you twins
73	0.3	what is your favorite color
72	0.3	what’s your favorite food
68	0.3	computers
67	0.3	hello hello
63	0.3	tell me about computers
63	0.3	what is your favorite food
62	0.3	shut up
61	0.3	where are you from
60	0.3	what’s up
57	0.3	robots
56	0.2	okay
56	0.2	what can i do at robot park
55	0.2	cell phones
55	0.2	where do you live
53	0.2	oh
52	0.2	hello who are you named after
51	0.2	thank you
50	0.2	what can you tell me about

Appendix B. Sample interactions

The following examples are reconstructed from the logs and transcriptions, as explained in section 3. Each transcribed user utterance is followed by the speech recognizer output (as logged by the system in real time), which is what the Twins are reacting to.

B.1. Coherent conversation segment

Coherent conversations between a visitor and the Twins happen when speech recognition is good, allowing the characters to select an appropriate response, and when the user follows up on previous utterances by the Twins.

User: Who are you named after (ASR: what are your names after)
Ada: My name's Ada.
Grace: And I'm Grace. We're your Virtual Museum Guides. With your help, we can suggest exhibits that will get you thinking! Or answer questions about things you may have seen here.
Ada: What do you want to learn about?
User: Artificial intelligence (ASR: is artificial intelligence)
Grace: One example of AI, or Artificial Intelligence, is 20Q, an online computer activity here at Computer-Place that asks you questions to guess what you're thinking.
Ada: I wish we'd been programmed to do that. Nah... on second thought, I prefer just answering your questions.
Grace: That takes AI too. Check out 20Q and tell a volunteer what you think we have in common and what makes us different.
User: Cell phones (ASR: can cell phones)
Ada: We have activities and resources based on existing cell phone research that will help you make your own informed decisions about cell phone usage. Go check them out at Cell Nation, and share your findings with a volunteer.

B.2. Inappropriate best available response

The Twins have a fixed set of available responses. When the user utterance is close enough to something they think they understand, they will respond with the best match, which may not quite answer the user's question.

User: What's your cell phone number (ASR: what do cell phone number)
Ada: The quality of your cell phone's signal depends on how many other people are using their cell phones, how close you are to a cell tower, and a few other factors. Experiment with our Cell Phone exhibit to see if you can figure out what the other factors are.

B.3. Non-understanding and misunderstanding

Non-understanding happens when the Twins are aware that they cannot interpret the user's utterances, that is when the Natural Language Understanding module has low confidence in its best match; in such cases the Dialogue Manager responds with an off-topic response such as a clarification

request. Misunderstanding happens when the NLU confidence is high, and results in an inappropriate response.

Grace: Is there something we can help you with?
User: Towers of Hanoi (ASR: how can a raise)
Ada: Could you repeat that?
User: Hanoi (ASR: what are your names)
Ada: My name's Ada.
Grace: And I'm Grace. We're your Virtual Museum Guides. With your help, we can suggest exhibits that will get you thinking! Or answer questions about things you may have seen here.

B.4. Users testing the Twins in foreign languages

Sometimes users try to see if the Twins can understand foreign languages. The transcriptions in the corpus only include transliterations (the non-Latin script transcriptions below have been added specifically for this paper).

User: Do you speak other languages (ASR: do you speak any other languages)
Ada: We love talking! When you talk to us, we figure out what to say, kinda like you would.
Grace: We listen to the words you use and the way you use them, and respond with what we think makes the most sense.
User: 对你好 (*dui ni shi hao*) (ASR: what are your names from)
Ada: My name's Ada.
Grace: And I'm Grace. We're your Virtual Museum Guides. With your help, we can suggest exhibits that will get you thinking! Or answer questions about things you may have seen here.
Ada: What do you want to learn about?
User: Ελληνικά (*ellinika*) (ASR: tell me about)
Grace: What did you say?
User: Ελληνικά (*ellinika*) (ASR: what can you talk)
Grace: A virtual human looks and acts like a human, but it isn't really alive.

Appendix C. Speech recognition

The Twins presently employ three speech recognizers with distinct acoustic models, for children, adult males, and adult females. The three recognizers are used in parallel, and at the end of each utterance, the text output of the recognizer with the highest confidence is sent to the Natural Language Understanding module. The following table shows the result of evaluating each of the recognizers on the test set mentioned in section 4.1; the "Oracle" line corresponds to selecting the appropriate recognizer for each utterance based on the speaker as identified in the transcription, and the "Switching" line shows the results of selecting the recognizer with the highest confidence, as done in the deployed system.

Model	WER (%)	WER per utterance (%)	
		Mean	Median
Children	53	64	63
Male	71	78	100
Female	63	72	83
Oracle	43	48	
Switching	57	67	

The fact that the children’s model is the best performer (even outperforming the automatic switching) is likely due to the fact that children’s utterances are an overwhelming majority in the test sample (764 children, 167 male and 72 female). The oracle condition improves on the children’s model by 10 percentage points, suggesting that better speaker identification should lead to better recognition overall.

The following table shows some individual results from the above test set. Boldface utterances indicate the actual speaker type, and italic utterances are the ones chosen by automatic model selection.

Model	Text	WER (%)	Conf (%)
Ref	are you a computer		
Child	what you stop retarded	75	0.98
Male	<i>who are you a not not you</i>	100	1.00
Female	who are you a top are you	100	1.00
Ref	i don’t like virtual museum guides		
Child	i don’t like virtual reason that	33	0.24
Male	i like virtual museum guides	17	0.79
Female	<i>i don’t like virtual museum guides</i>	0	0.99
Ref	what are you named after		
Child	<i>what are your names from</i>	60	0.99
Male	what are you named after	0	0.99
Female	what are you named after	0	0.78