Literary Evidence for the Cultural Development of a Theory of Mind

Andrew S. Gordon (gordon@ict.usc.edu) and Anish Nair (anair@usc.edu)
Institute for Creative Technologies, University of Southern California
13274 Fiji Way, Marina del Rey CA 90292 USA

Abstract

The term Theory of Mind is used within the cognitive sciences to refer to the abilities that people have to reason about their own mental states and the mental states of others. An important question is whether these abilities are culturally acquired or innate to our species. This paper outlines the argument that the mental models that serve as the basis for Theory of Mind abilities are the product of cultural development. To support this thesis, we present evidence gathered from the large-scale automated analysis of text corpora. We show that the Freudian conception of a subconscious desire is a relatively modern addition to our culturally shared Theory of Mind, as evidenced by a shift in the way these ideas appeared in 19th and 20th century English language novels.

A Cultural Theory of Mind

One topic that is strikingly pervasive across the cognitive sciences is that of Theory of Mind, in reference to the abilities that people have in reasoning about their own mental states and those of others. It is the set of Theory of Mind abilities that enable people to reflect introspectively on their own reasoning, to empathize with other people by imagining what it would be like to be in their position, and to generate reasonable expectations and inferences about mental states and processes.

Although there are inherent difficulties involved in investigating behavior that is largely unobservable, a relatively sophisticated understanding of Theory of Mind abilities has emerged through the synthesis of widely disparate sources of evidence. This evidence suggests that Theory of Mind abilities progressively develop in children and adults (Wellman & Lagattuata, 2000; Happe et al., 1998), are degraded in people diagnosed with the illness of autism (Baron-Cohen, 2000), have a relationship to localized brain regions (Happe et al., 1999; Frith & Frith, 2000), and are a uniquely human cognitive faculty not available to other primates, e.g. chimpanzees and orangutans (Call & Tomasello, 1999). This last contribution to our understanding of Theory of Mind suggests that these abilities must have arisen in the human lineage only after a split from that of chimpanzees some 6-8 million years ago.

Although in may be reasonable to assume that Theory of Mind abilities emerged in humans through a combination of natural evolution and cultural evolution, the relative importance that one ascribes to either of these two forces can radically change one's conception of the mental lives of humans that are contemporary on a genetic time scale and primitive on a cultural time scale. If genetic evolution is the prime contributor to human Theory of Mind abilities, then we could imagine that human beings tens of thousands of years ago reflected introspectively on their own reasoning, empathized with other humans they had contact with, and were able to generate expectations and inferences about mental states and processes. If instead cultural is seen as the prime contributor, then human beings tens of thousands of years ago would not be capable of any of these behaviors, among many others.

In attempting to determine whether the emergence of Theory of Mind abilities is genetic or cultural, researchers are immediately faced with the problem of evidence. Drawing comparisons between Theory of Mind abilities is an extremely difficult task, even between people that participate in controlled psychological experiments, let alone across cultures separated by distance and/or time. Several researchers have thought that the strongest evidence for a cultural Theory of Mind would be the discovery of significantly different Theory of Mind abilities across contemporary cultures. Lillard's review of research in cultural variation in Theory of Mind (1998) suggests that meaningful variations may exist among the peoples of the world, but argues that there is little evidence available to draw firm conclusions, and that the methodologies employed in the past to study mental representations in other cultures have been problematic.

A second type of evidence for a cultural Theory of Mind has looked for significant variation within a culture across time. One of the more provocative of these historical analyses was that of Julian Jaynes in support of his ideas on the emergence of consciousness (1976). In this work, Jaynes examines references to psychological concepts as they appear in a variety of early narratives, including the Iliad and the Odyssey. By comparing how these texts and others use terms such as *thumos*, *phrenes*, *kradie*, *etor*, *noos*, and *psyche*, Jaynes advances his claim that there was a shift in the way the people of ancient Greece thought about the role that mental phenomena played in controlling behavior.

Although Jaynes successfully argues that there was a shift in the way that mental concepts were referenced in these early texts, one could argue that these changes can

be attributed solely to changes in linguistic convention, rather than to changes in the underlying semantics of the language or the cognitive abilities that are based on these representations. For this criticism, Jaynes makes the controversial retort, "Let no one think these are *just* word changes. Word changes are concept changes and concept changes are behavioral changes." (p. 292, original emphasis) While this comment raises a much larger philosophical debate concerning mental representation, we believe that arguments made on either side would be improved if the evidence were stronger. Among other points, it is tempting to argue that the changes exhibited in Jaynes' sampling of early texts are not representative of the cultural environment in which these texts were produced.

In this paper, we explore the role that language evidence can play a role in answering questions concerning the cultural development of a Theory of Mind. Our interest is in identifying evidence of a specific change in the way that a broad range of people within a culture refer to mental states or processes over a period of time. As we describe in this paper, we felt we could obtain compelling evidence by examining a large corpus of 19th and 20th century English-language novels, with particular attention to references to the Freudian notion of a subconscious desire.

The primary tool that we apply toward exploring these issues is that of automated corpus analysis, where computer programs are constructed to gather statistics about the linguistic elements found in large collections of text documents. Automated techniques for corpus analysis enable researchers to study text collections that are larger than could be reasonably processed by human analysts, but the tradeoff is that effort must be directed toward the development of the tool. In the next section, we describe our efforts to construct a corpus analysis tool capable of identifying references to Theory of Mind concepts in English text in a robust manner. Then, in the he sections that follow, we discuss the application of this tool to the investigation of the cultural development of a Theory of Mind.

The Theory of Mind in Language

The relationship between representation and language has been the subject of one of the most significant and long-standing philosophical debates of our time, and concerns some of the most fundamental hypotheses that we may draw concerning human cognition. In the field of computational linguistics, however, a pragmatic approach to this issue has been adopted to allow for the engineering of useful natural-language processing systems. It has become commonplace, particularly in the areas of automated question-answering and knowledge extraction, to parse natural language expressions into unambiguous, formal representations. Generally, the semantics that are employed in these approaches are strictly utilitarian, in that they are designed to address particular engineering problems. However, others have adopted many of the same methods and technologies to construct models of language understanding that employ representations that are more cognitively inspired (Baker et al., 1998).

In our work, we also sought to employ some of the methods and technologies that are used in computational linguistics to explore issues related to the Theory of Mind. Our aim was to develop a system that would be able to recognize all occurrences in English text of expressions related to the Theory of Mind. That is, our system takes as input a text document and produces as output an annotated version, where every reference to a Theory of Mind concept is tagged and categorized as a particular sort. For example, the following passage (from William Makepeace Thackeray's 1917 novel, *Vanity Fair*) illustrates the format of the output of this system, where references to Theory of Mind concepts are underlined, and their semantic type is inserted into the text following the reference delimited by square brackets:

> Perhaps [*partially-justified-proposition*] she had mentioned the fact [*proposition*] already to Rebecca, but that young lady did not appear to [*partially-justified-proposition*] have remembered it [*memory-retrieval*]; indeed, vowed and protested that she expected [*add-expectation*] to see a number of Amelia's nephews and nieces. She was quite disappointed [*disappointment-emotion*] that Mr. Sedley was not married; she was sure [*justified-proposition*] Amelia had said he was, and she doted so on [*liking-emotion*] little children.

The first challenge in developing this system was to answer the question: What are the representational elements of the Theory of Mind that we should be looking for in English text? We believe that the representational elements for Theory of Mind described in our previous work (Gordon, 2002) offer the greatest breadth and depth of concepts compared with previous representation research. This work, which was based on the large-scale analysis of analogous planning cases (Gordon, 2001), identified 30 representational areas and 635 individual concepts that participate in our commonsense mental models of human reasoning. These 30 representational areas are largely identified by cognitive function, and include memory retrieval, similarity judgments, belief management, explanation, prediction, planning, scheduling, monitoring, and execution. For our current purposes, we felt that these representational areas and specific individual concepts could serve as the basis for tagging references to Theory of Mind concepts in English text.

To develop the automated tool for identifying and tagging references to these concepts in English text, we

followed a methodology that consisted of three main steps, each requiring the efforts of a team of linguistics and computational linguistics graduate students of the University of Southern California. The first step, *expression elicitation*, involved collaborative brainstorming among members of our team to identify an initial set of English language expressions for each of the concepts in a particular representational area. This initial list was then enhanced in the second step, *lexical expansion*, where a variety of linguistic resources (thesauri, thematic dictionaries, phrase dictionaries) were consulted to identify a more complete set of synonymous referents to every given concept. In the third step, *rule authoring*, our team created (by hand) a set of generalized linguistic rules (finite state automata) that could be applied to a text document to recognize expressions like those in the expanded set of referents.

To aid in the development of linguistic rules of a high degree of accuracy, our group utilized the Intex Corpus Processor software (Silberztein, 1999), which allowed us to author these rules as finite-state automata using a graphical user interface. To simplify the specification of patterns, we employed a large-coverage English dictionary that allows generalization over linguistic variations, such as verb inflections for a given lemma. Hundreds of generalized linguistic patterns were authored, generalizing over each expression that was identified in the previous step in the process, and then combined into a single finite state automaton that could be applied to any English text document.

At the time of writing of this paper, our group had executed this methodology for 10 of the original 30 Theory of Mind representational areas that were identified. These areas are Memory (12 concepts), Similarity comparisons (13 concepts), Explanations (20 concepts), Managing knowledge (37 concepts), Goals (20 concepts), and Goal Management (17 concepts), Envisionment (35 concepts), Plans (21 concepts), Plan elements (27 concepts), and Planning modalities (17 concepts). An evaluation of the accuracy (precision and recall) of our rule sets in identifying Theory of Mind expressions is reported in Gordon et al. (2003). The results of this evaluation demonstrate that this approach allows us to recognize and tag 82% of the actual number of expressions that appear in a document for any given representational area (recall score), and that 95% of expressions that our system identifies should actually be counted as positive instances (precision score).

Expressions of Subconscious Desires

The value of developing a robust lexical-semantic resource for Theory of Mind concepts is that it allows us to automatically analyze textual data sources that are larger than could be examined given limited human effort (e.g. Dyer et al., 2000). With this resource in hand, we sought evidence for the cultural development of a Theory of Mind by examining references to these concepts in large text corpora.

Due to the nature of this tool, we were forced to adhere to a number of constraints about the sorts of questions that can be pursued. This tool could only be applied to English language documents, and only those that were authored recent enough to share common lexical and grammatical conventions with contemporary English. As developing the tool involved the hand authoring of grammatical rules by speakers of contemporary English, its performance may be stronger for contemporary texts than for older texts, particularly with regard to idioms and slang. Also, when comparing the frequency of occurrences of Theory of Mind expressions across different texts, it is important that the texts are of the same genre (e.g. newspaper articles, political speeches, poetry, spoken language transcripts), as the tool performs somewhat differently for each type.

These constraints restricted the application of this tool to the study of cultural changes that have occurred within the last few hundred years. Has a significant change in Theory of Mind occurred among English speaking peoples during this period of time?

Probably the strongest argument for recent, widespread cultural change in Theory of Mind involves the impact of the ideas of Sigmund Freud. In *The Interpretation of Dreams*, first published in German in 1900 and translated to English in 1913, Freud outlined an understanding of human cognition that involved the influence of subconscious desires in behavior. Although Freud's use of symbolism in explaining dreams is often joked about in everyday conversation, the more central idea of a subconscious desire has become ingrained in the way people talk about their behavior. It is uncontroversial when a person reports, "I must have wanted to leave my wife all along, but I only recently became aware of this subconscious desire of mine."

In the English language, there are, in fact, many ways of expressing the Freudian notion of a subconscious desire without using this term – or any special terminology at all. In our work in developing a lexical semantic resource for Theory of Mind, we explored references to this concept within the representational area of Goals (one of the 30 representational areas that we aimed to complete). The following list gives some specific examples of the general linguistic forms that our hand-authored rules are able to identify and tag as a subconscious goal:

- Andrew *realized he hoped* the plan would fail.
- He *guessed that what he really wished for* was something different entirely.

- He *didn't figure out what it was that he desired* until it was too late.
- I'm *not sure I want* to go down that route.
- That man *wasn't aware of what he yearned* for.
- It was what I *actually wanted all along*.
- His *subconscious wish* now came to the surface.
- We later learned of his *unconscious desire*.

The ability to automatically identify and collect references to subconscious desires in English text allows us to explore whether this Freudian idea is a modern addition to our understanding of mental states and processes – the product of a cultural innovation and dissemination. If so, we hypothesize that we would not be able to find in English text documents any of the linguistic patterns listed above before the period in history when this Freudian idea was disseminated. If instead we view this concept as one that was already well established in the pre-Freudian mind, then we should be able to find some variations of the above expressions before this time.

The Pre-Freudian Shift

To explore the way that people use language related to the Freudian notion of a subconscious goal, it was necessary to construct a very large text corpus of uniform genre that spanned the relevant years in history. For this, we quickly settled on the genre of the novel as the most appropriate data source. First, electronic texts of novels (necessary for automated analysis) are readily available in the public domain, particularly from Project Gutenberg online resource (http://www.gutenberg.net). Second, the mix of narration and dialogue that is found in the genre makes it likely that expression related to desires of all sorts will occur in abundance, largely because of an emphasis on character development. Third, the genre of the novel has remained remarkably constant over the course of the last several hundred years in format, aiding in the comparison of novels over time.

The main downside of the novel as a genre of analysis is that U.S. copyright law has made it difficult to obtain electronic texts of works published after 1922. Freud's *The Interpretation of Dreams* was certainly well known and discussed among English-language writers by this date, but we imagine that the cultural impact of this work would have continued to grow after this time. Still, we felt that this cut-off date for novel data was sufficiently late in history to answer the fundamental question.

We assembled a corpus of 176 English-language novels with publication dates spanning the years between 1813 and 1922. This corpus consisted primarily of American and British novels, all of which appeared on at least one "great books list" of the 19th and 20th century and which were available electronically from Project Gutenberg. With an average of 118 thousand words per novel, our corpus consisted of over 20 million words.

To analyze this corpus, we applied our hand-authored rule set for subconscious desires to each novel, collecting each sentence from the corpus that matched any of the linguistic patterns for this concept. In all, 69 occurrences were identified.

Figure 1 displays the distribution of the occurrences that were identified. To normalize this data across novels of varying length, each data point on the graph indicates the number of references to subconscious goals found in a single novel divided by the word count of the novel. While this normalization permits us to compare the frequency of references in a more valid manner, the relatively constant size of novels in our corpus yields a normalization that has little affect on the shape of the graph.

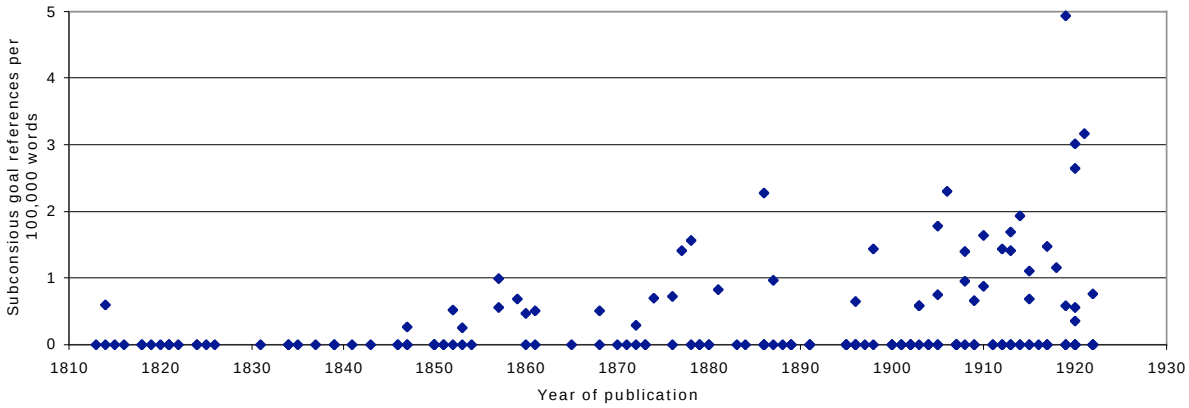

Figure 1: Subconscious goal references per 100,000 words in 176 English-language novels published between 1813 and 1922

The evidence from this corpus supports neither of our original two hypotheses. There is certainly a shift in the frequency of references to subconscious goals within our data set, but this shift is not an immediate one following the widespread cultural dissemination of Sigmund Freud's ideas. Instead, we find a gradual shift beginning in the middle of the 19th century that continues to grow into the 20th century through the last of our data points. We believe this evidence argues for a "Pre-Freudian Shift" in the mental models that people had about their capacity to be aware of their own desires. From this perspective, the work of Freud can be seen in a context of widespread cultural change, as an effect rather than an instigating force.

Figure 2 further supports the idea of a Pre-Freudian Shift by grouping the 69 references into 10 linguistically similar forms, and showing the first occurrence of each. The first form, "I am sure I hope", accounts for most instances in the latter half of the 19th century and could be viewed as a verbose way of expressing "I hope". Henry James' negation of this phrase in 1881 begins to connote a sense of doubt about one's desires, and gives rise to the 20th century usage first seen in 1896, "I don't know that I want". This trend becomes fully developed in 1913 when the certainty of someone else's desires is questioned, and in 1914 when one's own desires must be guessed. Most striking in Figure 2 is the direct references to *unconscious yearning* by George Eliot (pseudonym of Mary Anne Evans) in 1860 and similarly to *unconscious desire* by Thomas Hardy in 1874: "There was no necessity for any continuance of speech, and the fact that she did add more seemed to proceed from an unconscious desire to show unconcern by making a remark, which is noticeable in the ingenuous when they are acting by stealth."

1814 Jane Austen, *Mansfield Park*: "Her year!" cried Mrs. Price; "I am sure I hope I shall be rid of her before she has staid a year, for that will not be up till November."
Similar usage is found later in 1847, 1852, 1853, 1857(x2), 1861, 1868, 1872, 1876, 1877, 1886, 1887(x2), and 1905.

1857 Charles Dickens, *Little Dorrit*: Mr. Clennam got it him to do, and gives him odd jobs besides in at the Works next door-- makes 'em for him, in short, when he knows he wants 'em.
Similar usage is found later in 1857, 1877, 1878, 1886, 1898, 1906, 1908(x2), 1910(x2), 1913, 1914, 1915(x2), 1919, 1920(x4), and 1922.

1859 Charles Dickens, *Tale of Two Cities*: He'd never have no good of it; he'd want all along to be out of the line, if he, could see his way out, being once in-- even if it wos so.
Similar usage is found later in 1903 and 1914.

1860 George Eliot, *Mill on the Floss*: Maggie, in her brown frock, with her eyes reddened and her heavy hair pushed back, looking from the bed where her father lay to the dull walls of this sad chamber which was the centre of her world, was a creature full of eager, passionate longings for all that was beautiful and glad; thirsty for all knowledge; with an ear straining after dreamy music that died away and would not come near to her; with a blind, unconscious yearning for something that would link together the wonderful impressions of this mysterious life, and give her soul a sense of home in it.
Similar usage is found later in 1874 (Thomas Hardy's *Far from the Madding Crowd*).

1881 Henry James, *Portrait of a Lady*: "It's not only that," said Isabel; "but I'm not sure I wish to marry any one."

1896 Thomas Hardy, *Jude the Obscure*: "I don't care to go into them," she replied evasively. "I make a very good living, and I don't know that I want your company."
Similar usage is found later in 1903, 1905, 1909, 1910, 1912(x3), 1913(x2), 1914, 1915, 1919(x4), 1920(x3), and 1921(x2).

1913 D. H. Lawrence, *Sons and Lovers*: "You are sure you want me?" he asked, as if a cold shadow had come over him.

1914 Theodore Dreiser, *The Titan*: "Oh, I don't know," replied Cowperwood, easily; "I guess I want you as much as ever."
Similar usage is found later in 1918 (Willa Cather's *My Antonia*).

1917 Edith Wharton, *Summer*: "What's all this about wanting?" he said as she paused. "Do you know what you really want?"

1920 James Joyce, *Ulysses*: …ah yes I know I hope the old press doesn't creak ah I knew it would…

Figure 2: Examples of subconscious goal references in 176 English-language novels published between 1813 and 1922

Conclusions

The appearance of a Pre-Freudian Shift in the way that people refer to the idea of a subconscious goal does not, in itself, prove that Theory of Mind abilities are the result of cultural development. Nor does this shift provide direct evidence for a change in the abilities that 19th and 20th century English-speaking people had in reasoning about their own goals or the goals of other people. The evidence that is presented is simply that people have changed the way they refer to goals in English over this period in history. The argument for the cultural development of a Theory of Mind, however, is that this shift is indicative of a significant cultural change in the mental models of psychology that people hold – one that now includes the concept of a subconscious goal. If we believe that our Theory of Mind abilities are predicated on tacit representational models of mental states and processes, then we would expect that some change in our Theory of Mind abilities might have also occurred. If so, it could be viewed as just one recent change in a long history of cultural innovations that would have included the more fundamental commonsense psychology concepts of memories, beliefs, emotions, goals, and plans, among others that would not have participated in the mental lives of early man.

While the approach of automated corpus analysis taken in this research was appropriate for investigating the use of references to subconscious goals, this same approach may not be appropriate for these other fundamental concepts. The genre of the novel was particularly important to the success of this approach, in that many data points are electronically available that span a period in history where an interesting Theory of Mind question can be raised. However, it is clear that the automated recognition of Theory of Mind concepts in natural language text can be an important tool for answering cognitive science questions, and that the continued development of such a resource will be an important pursuit.

Acknowledgments

Abe Kazemzadeh and Milena Petrova participated in the language analysis portion of this research. This paper was developed in part with funds from the U.S. Army Research Institute for the Behavioral and Social Sciences under ARO contract number DAAD 19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Department of the Army.

References

Baker, C., Fillmore, C., & Lowe, J. (1998) The Berkeley FrameNet Project. *Proceedings of COLING-ACL 1998*, Montreal, Canada.

Baron-Cohen, S. (2000) Theory of mind and autism: a fifteen year review. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition*. Oxford, UK: Oxford University Press.

Call, J. & Tomasello, M. (1999) A Nonverbal False Belief Task: The Performance of Children and Great Apes. *Child Development 70*(2):381-395.

Dyer, J., Shatz, M., & Wellman, H. (2000) Young children's storybooks as a source of mental state information. *Cognitive Development* 15, 17-37.

Frith, C. & Frith, U. (2000) The physiological basis of theory of mind: functional neuroimaging studies. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition*. Oxford, UK: Oxford University Press.

Gordon, A. (2001) Strategies in Analogous Planning Cases. In J. Moore & K. Stenning (eds.) *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Gordon, A. (2002) The Theory of Mind in Strategy Representations. In W. Gray & C. Schunn (eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, Mahwah, NJ: Lawrence Erlbaum Associates.

Gordon, A., Kazemzadeh, A., Nair, A., & Petrova, M., (2003) Recognizing Expressions of Commonsense Psychology in English Text. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03) Sapporo, Japan.

Happé, F., Brownell, H., & Winner, E. (1998) The getting of wisdom: Theory of mind in old age. *Developmental Psychology*, 34 (2), 358-362.

Happé, F., Brownell, H., & Winner, E. (1999) Acquired 'theory of mind' impairments following stroke. *Cognition* 70, 211-240.

Jaynes, J. (1976) The Origin of Consciousness in the Breakdown of the Bicameral Mind. 2000 paperback edition from Mariner Books, Boston.

Lillard, A. (1998) Ethnopsychologies: Cultural Variations in Theories of Mind. *Psychological Bulletin 123*(1):3-32.

Silberztein, M. (1999) Text Indexing with INTEX. *Computers and the Humanities 33*(3).

Wellman, H.M., & Lagattuta, K. H. (2000). Developing understandings of mind. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from developmental cognitive neuroscience, second edition*. Oxford, UK: Oxford University Press.