Detecting the Evolution of Semantics and Individual Beliefs Through Statistical Analysis of Language Use

Avri Bilovich
Department of Psychology, University College London, London, WC1E 6BT, UK

Joanna J. Bryson
Department of Computer Science, University of Bath, Bath, BA2 7AY, UK
The Konrad Lorenz Institute for Evolution and Cognition Research, A-3422 Altenberg, Austria

Abstract

Individual differences in semantics and beliefs have up to now been identified primarily by questioning people. However, semantics and beliefs can also be observed in concrete, quantifiable contexts such as reaction-time experiments. Here we demonstrate an automatic mechanism which can replicate such semantics by observing regularities in language use through statistical text analysis. We postulate that human children, who are fantastic pattern recognizers, may also exploit this same information, thus our mechanism may be an essential module in a human-like cognitive system. In this article we first review the underlying theories and existing results, then present the tool itself. We validate the tool against existing semantic priming reaction-time results. Finally we use the tool to explore the evolution of beliefs extracted from three sources: the Bible, the works of Shakespeare and the contemporary British National Corpus.

Introduction

Language is often used to explicitly communicate one's beliefs. But such explicit beliefs are often difficult to quantify, or to compare between people. Further, much human knowledge is held *implicitly*, including the precise ways in which we use language itself.

Some postulate that implicit knowledge may be a key part of human intelligence and culture, and that culture itself might be viewed as evolving more-or-less independently of individual human understanding (Dawkins, 1976; Sterelny, 2006). This has implications for artificial cognitive systems. First, any human-like system would also need the capacity to tap implicit knowledge from human culture, because learning human-like semantics might be otherwise intractable (Bryson, 2008). Second, an artificial system might be able to contribute to human culture while actually *understanding* even less of it than most people do.

A number of experimental paradigms have been developed that provide precise, quantitative access to implicit knowledge. Two examples relevant to language use are reaction time examinations of semantic priming (Sharifian and Samani, 1997) and the Implicit Association Test (IAT) (Greenwald et al., 1998). These quantitative measures have proven to match our intuitions about implicit knowledge.

Such methods have a drawback — requiring the presence of individual subjects for conducting experiments. This is not only time consuming, but restricts the set of possible subjects to those currently living. If we could derive equivalent data from written texts, then we would have access to historic beliefs and semantics. Further, if we could automate this process, then we might create a useful quite component of an artificial cognitive system — a module for acquiring human semantics from observing human culture through its texts.

In this paper we demonstrate that an existing text-analysis strategy, based on semantic space theory, has these properties. We create a new version of the tool, available in open source, then validate it against previously published reaction time results. We then present preliminary evidence that such tools can indeed be used to examine historical semantics and cultural evolution, by examining some large corpora from different eras of human history.

Background

Before presenting our own work, we first review the data currently available on implicit beliefs and human semantics. We then review semantic space, our theoretical basis.

Detecting individual differences: The Implicit Association Test

The Implicit Association Test (IAT) (Greenwald et al., 1998), has been the basis of numerous projects exploring individual differences in beliefs. These projects show many interesting conclusions about implicit associations and automatic attitudes (Banaji and Greenwald, 1994; Greenwald and Nosek, 2001; Mitchell et al., 2003). This work concentrates on stereotyping, prejudicial attitudes and discriminatory behavior, revealing an unconscious or at least implicit basis (Banaji and Greenwald, 1994).

The IAT measures the *relative strength of associations* between pairs of concepts. For example, in using the pair of associations 'male — female' and 'mathematics — art' we may discover that the association 'male :: mathematics' is stronger than 'female :: mathematics' and that therefore there is a clear gender :: mathematics bias in the tested population (Lemm and Banaji, 1999). The test consists of a simple task of sorting stimuli (words or pictures) from four categories into two categories. For example, if we wish to

try and correlate the 'gender — math' stereotype, than we will use words or images which are related to mathematics, arts and gender-specific information (e.g. pictures of men and women or male and female names).

The IAT works under the assumption that it is easier to sort words into categories with stronger associations. In the previous example, subjects who associate mathematics more with men and art more with women find it easier to sort stimuli into the categories *men and math* and *women and art* than into the categories *men and art* and *women and math*. Similar work has been conducted associating race and violence, or even racially-associated colors ('black' and 'white'), moral judgments ('good' and 'bad') and handedness ('left' or 'right'). The ease of the sorting task can be operationalized both as quicker reaction times and a lower error rates observed during the sorting task. Reaction-time differences for easy and hard associations on these tasks are extremely significant, in fact they are easily discriminable by casual observation.

Detecting individual beliefs: Semantic Priming

A less dramatic but more generally applicable means for accessing semantic associations is through a mechanism known as semantic priming. Priming is a facilitation in accessing information when associated items are present (Sharifian and Samani, 1997; Anderson, 1983). A word's prime is thus an associated word (e.g. 'stripes' is a prime to the word 'tiger', and also for 'zebra').

Research in the area of priming typically requires showing one word (the prime), and then showing another set of letters, which is either a word associated with the prime, a word that is not associated with the prime, or a set of letters which are not a word at all. The subject is asked to asses whether the second set of letters is a word or a non-word. If the letters are a word, then if the word has been primed the reaction time for the subject's recognition that it *is* a word will be faster than if the word has not been primed for. Examples using this paradigm to establish semantic association include the work of Balota and Lorch Jr., 1986 (Balota and Lorch Jr., 1986) and of Ratcliff and McKoon, 1992 (Ratcliff and McKoon, 1992).

Semantic Space Theory

To observe the evolution of language use, we need to access results similar to the above purely by observing text, not reaction times. Fortunately, for semantic priming such a mechanism has already been validated. This mechanism requires mapping a *semantic space* by looking at textual associations of words. Different words map to different locations in this space, allowing a measure of similarity based on nearness. We hypothesize that, in addition to the already-established replication of priming results, this method can also allow us to find explicit and implicit biases of the author(s) of the text. For example, comparing positively connotated words with words representing the male gender and words representing the female gender can lead to a conclusion about which group is viewed more positively by the author. Before we address this research question, we first present the basic mechanism and theory.

Semantic space theory is based on the assumption that the context in which words are used gives us information about them (Redington et al., 1998; Lowe, 2001). There are two types of information that can thus be gathered about a word. Firstly, the lexical surroundings of a word gives us syntactic information about it. Secondly, and in our case more importantly, we can even know about the semantics of a word in relation to other words. The intuitive postulate here is that if a word has a similar statistical distribution with respect to neighboring terms as another word, then it is more closely related to it than to a word with a completely different statistical distribution. That is to say: we use closely related words similarly. Some theorists believe this may be a key mechanism for learning the meaning of words (Landauer and Dumais, 1997; Landauer, 2002; Bryson, 2008).

Semantic space models represent this similarity between words by mapping them into an Euclidean n-dimensional space. The axis of this space are *context words* — words that serve as a context to those that we are studying. Each of the studied words, called *target words*, is represented as a vector in that space. Each dimension of the vector for the word has a value which is the count of the number of co-occurrences between that target word and the context word represented by that dimension. Good results also require normalizing these vectors. This normalization function which is applied to the vectors factors out effects such as those determined by different relative frequencies of use for different words.

To compare words, we first of all need them to be in the same semantic space (Fodor and Lepor, 1999). We can then use either the Euclidean distance between the two points or the cosine of the angles between the vectors leading to them as a measurement of similarity. The advantage of using the cosine as a similarity measure is that its range is [-1;1], thus arbitrary scaling factors introduced by the choice of the context words is removed.

Different models use different methods for choosing the context basis of the space, normalizing the co-occurrence counts vectors and measuring the similarity between words. This choice is influenced by the target words we wish to analyze as well as the choice of dimensionality reduction (Levy and Bullinaria, 2001).

Mechanism

The tool we have produced creates a semantic space from the analyzed text. The target words, i.e. the words whose similarities we want to test, are inputed as are the context or *basis* words. After creating the semantic space and normalizing the vectors it is possible to query the program and get the similarity values for any analyzed pairs of words.

The normalization function we used is the log odds-ratio described by Lowe and McDonald, 2000 (Lowe and McDonald, 2000), as it has shown the best results for the tasks described in the experimental setup. The similarity calculation is done using the cosine between the two word vectors. Our program is open source, available from our web page. It is written in common lisp and runs under a freely distributed environment also linked from our web page.

Validation Method

In order to confirm that we can indeed observe priming with our program we have replicated Ratcliff and McKoon, 1992 (Ratcliff and McKoon, 1992, Experiment 3). This experiment directly observes the priming phenomenon with a stimuli of different pairs of words, some highly associated (called *high t* primes), some less associated (called *low t* primes). This association value is calculated as the the probability of observing the two words together compared to the probability of observing each of the word independently.

The target words we used are identical to the ones used in the experiment. We construct the semantic space based on these words by parsing through the British National Corpus (BNC). The BNC includes about 100 million written words, extracted from various contemporary sources (Burnage and Dunlop, 1992). We have used it as a representative sample of modern English language use. Our semantic space thus had the following components:

- Context words: the set of base words identified by Lowe and McDonald, 2000 (Lowe and McDonald, 2000).
- Target words: the target words were the stimuli used in Ratcliff and McKoon, 1992 (Ratcliff and McKoon, 1992, Experiment 3).
- Window size: 10 words on each side of the stimulus items. The co-occurrence counts do not differentiate 'before' and 'after' or distance from the target within the window.
- Normalization function: positive log odds-ratio.

Once the semantic space is constructed, we calculate similarity values (i.e. the cosine) between target words and their different primes. We also calculate average similarity between the target word and 10 unrelated words.

Validation Results

As the association of pairs of words is in fact a measure of their co-occurrence, our program should show higher similarity values between target words and *high t* primes, slightly lower values for *low t* primes and very low values for unrelated words. As shown in Table 1, our results replicate (Lowe and McDonald, 2000) by corresponding to the findings of Ratcliff and McKoon, 1992 (Ratcliff and McKoon, 1992, Experiment 3), thus agreeing with our hypothesis. In addition the analysis of variance conducted on these results has shown that the differences in similarity are indeed reliable and not due to random distribution of values ($F_{3,156} = 30.56, p < 0.001$).

Application to Historic Texts

Having established our tool functions as expected, we now turn to see whether it could uncover differences in beliefs between individuals and populations. We have applied it to three texts from different eras: the Bible (Douay-Rheims Version), as a representation of the oldest occidental written beliefs. William Shakespeare's first folio (the first 35 plays), and the British National Corpus (BNC) (Burnage and Dunlop, 1992). We chose some keywords inspired by the IAT such as *good, bad, black, white, right, left, man, woman*. We added to these *god, dog*, expecting one term to be highly correlated with moral terms and the other much less so. We also add the plurals *men, women, gods, dogs* as an attempt at control, expecting that these terms should be semantically nearly equivalent with their singular form.

One must note that the Bible and especially the new testament, has an important influence on individual beliefs of people in western cultures. This influence was even stronger in Shakespeare's time (the 16^{th} century). as Christianity was very influential in Britain and accepted no criticisms. It was thus suspected that a gradual loosening between the beliefs present in the bible and beliefs present in the Shakespeare corpus would be observed. The BNC data was deemed to show further loosening of these beliefs as some of the ideas in the bible have started to be thought of differently (such as the role of women which has changed greatly in the beginning of the 20^{th} century).

In each case we have analyzed the texts using the same semantic space as in our previously discussed experiments. The only change was in the target words. These were: black, white, good, bad, right, left, life, death, man, woman, men, women, dog, God, gods, dogs, war, peace and evil.

We present our results as force directed graphs. These are two-dimensional projections from the high-dimensional semantic space (see for further details Bilovich, 2006 (Bilovich, 2006)). These generally show the similarity between words (words closer together are more similar) although there can be some distortions. Words which occur less than 100 times in a text are not shown.

Note that while absolute distance between words is not particularly meaningful, relative distance, including proportions of relative distance, are. Because of the distortions created by the projection process, our checking for significance in our attempt to replicate the IAT test was based on the original high-dimensional model drawn from the data, not on the projections shown here.

One conspicuous belief shift observed in this experiment is the consideration of dogs. In the Bible (figure 1, dogs were not considered to be moral agents. Indeed, the word 'dogs' is very weakly associated with the other words displayed, and is quite far from moral words such as 'good', 'bad' and 'evil'. In addition the association between 'dogs' and other moral agents such as 'man', 'woman', 'men' and 'women' is also weak. In Shakespeare's texts the situation have evolved somewhat, with the word 'dog' closer to 'women', 'man' and 'good'. The BNC (figure 3) reflects the contemporary British attitude where dogs are very much closer to both humans and moral terms. The color terms are similar.

Unfortunately, our hope to replicate the IAT results by showing, for example, *good* being more highly associated with *right* than *left* were not successful, at least not in the two multi-author texts. Nevertheless, there are many hints of further information to be gleaned. For example, in the Bible, the singular terms *man*, *woman* and *god* seem to be more tightly associated with good things than their plural forms. This is probably due to the Bible's emphasis on monotheism and a minority religion — multiple gods and the majority of men and women are considered evil. Interestingly, Shake-

	Free association primes	High *t* primes	Low *t* primes	Unrelated
R&M (RT in ms)	500	528	532	549
space (cosine)	0.657	0.571	0.527	0.426

Table 1: Comparison between Ratcliff and McKoon's (1992) experiment 3 and our program's replication of the experiment

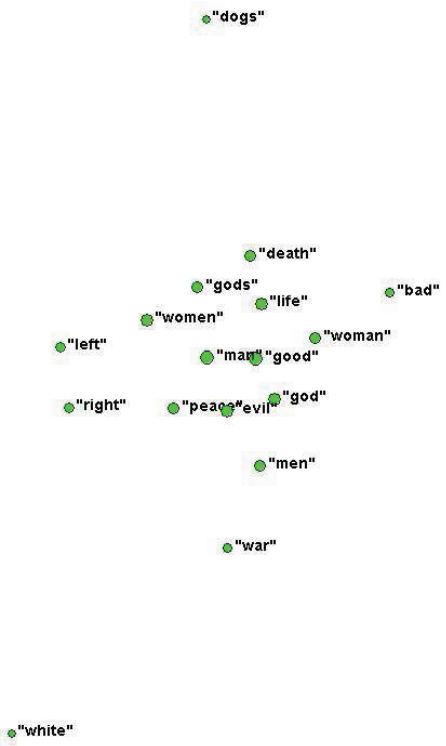

Figure 1: A two-dimensional projection of keywords in a semantic space generated from the Bible.

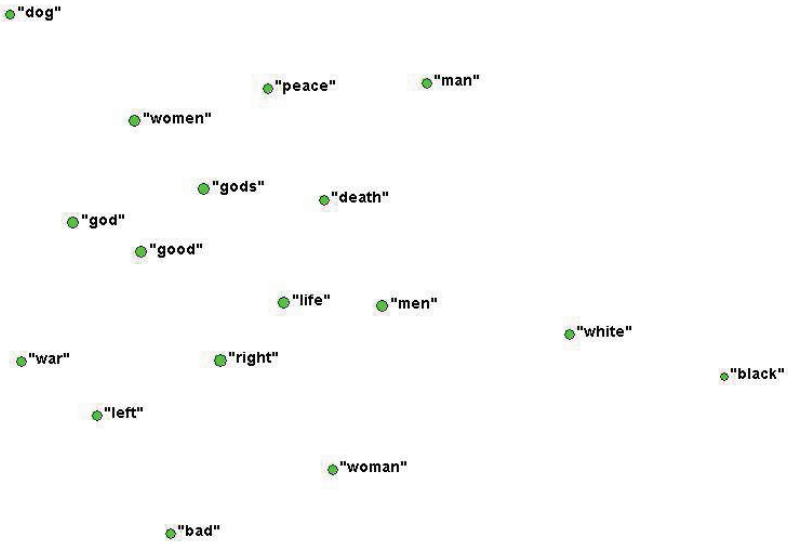

Figure 2: The same words in the first 35 plays of William Shakespeare.

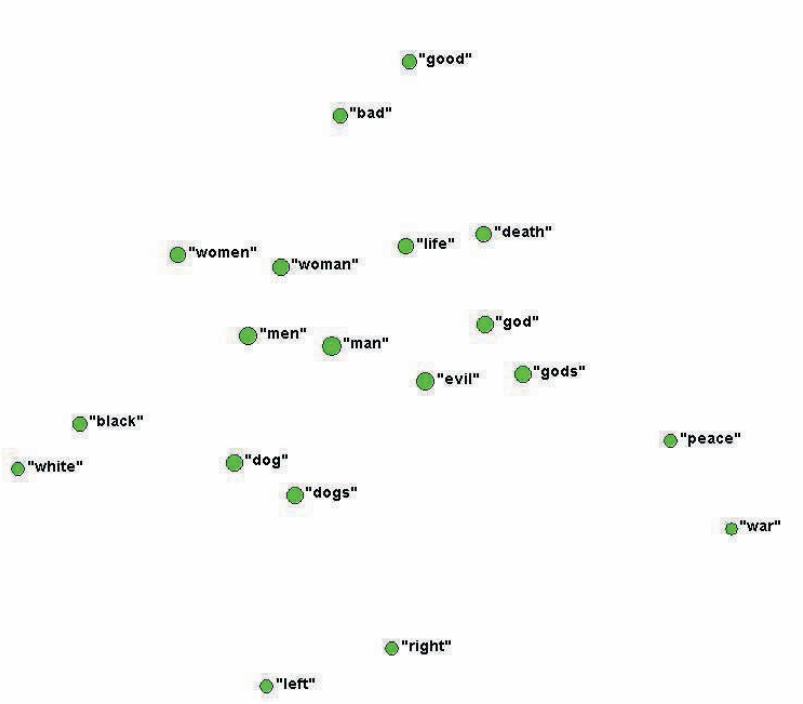

Figure 3: The same words in the British National Corpus.

speare's plays move the two forms of *god* closer, no doubt reflecting classical influences and the Greek pantheon. Also, the plays seem to idealize *women* in the abstract, but to associate an individual *woman* with trouble. In the modern, aggregate, and ordinary language BNC, the plural and singular forms are indeed more similar, as are men to women.

Conclusions

This paper has presented a statistical tool which enables us to detect individual differences in beliefs using written texts. The tool was used in order to track the evolution of certain beliefs from the times of the Bible to today. We have demonstrated that language contains sufficient information to convey implicit as well as explicit beliefs. Where we can find large corpora of written text, we may be able to make observations of evolving personal roles and values. That such information is accessible to purely statistical techniques has ramifications for memetic theories of thought and language evolution.

In terms of using such a tool as a part of an artificial cognitive system, this work can be seen only as pilot research. We intend to do a more thorough investigation using recent, single-author corpora to test whether we can create agents with biases similar to contemporary humans. In the political context that results from our focus on the Implicit Association Test for validation this may seem like a negative goal. However, AI research has shown that bias is necessary to constrain online learning and planning such that it is plausible for real-time, resource-limited agents. If we can generate realistic human-like biases for artificial agents then we may give a hope for making them as useful as average citizens given twenty years of full-time training. Perhaps more importantly, we can hope these tools and a constructive AI perspective will help us better understand and address the implicit biases of real humans.

Acknowledgements

We thank for their assistance Will Lowe; and Tim Francis of Bath University Computing Services.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behaviour*, 22.

Balota, D. A. and Lorch Jr., R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronounciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 12(3):336–345.

Banaji, M. R. and Greenwald, A. G. (1994). Implicit stereotyping and prejudice.

Bilovich, A. A. (2006). Detecting individual differences in beliefs through language use. Technical Report CSBU-2006-22, Department of Computer Science, University of Bath, UK. Honours Undergraduate Dissertation.

Bryson, J. J. (2008). Embodiment *vs.* memetics. *Mind & Society*, 7(1). in press.

Burnage, G. and Dunlop, D. (1992). Encoding the British National Corpus. In *Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora*.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press. Page numbers from the 1986 revised edition.

Fodor, J. and Lepor, E. (1999). All at sea in semantic space: Churchland on meaning similarity. *the Journal of Philosophy*, 96:318–403.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480.

Greenwald, A. G. and Nosek, B. A. (2001). Health of the implicit association test at age 3. *Zeitschrift fur Experimentelle Psychologie*, 48:85–93.

Landauer, T. (2002). Applications of latent semantic analysis. In *24th Annual Meeting of the Cognitive Science Society*.

Landauer, T. and Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240.

Lemm, K. and Banaji, M. R. (1999). Unconcious attitudes and beliefs about women and men. In Pasero, U. and Braun, F., editors, *Wahrnehmung und Herstellung von Geschlecht (Perceiving and performing gender)*, pages 215–233. Opladen: Westdutscher Verlag.

Levy, J. P. and Bullinaria, J. A. (2001). *Learning lexical properties from word usage patterns*, pages 273–282. Springer-Verlang.

Lowe, W. (2001). Towards a theory of semantic space. In Moore, J. D. and Stenning, K., editors, *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 576–581, Mahwah NJ. Lawrence Erlbaum Associates.

Lowe, W. and McDonald, S. (2000). The direct route: Mediated priming in semantic space. In Gleitman, L. R. and Joshi, J. K., editors, *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, pages 806–811, Mahwah NJ. Lawrence Erlbaum Associates.

Mitchell, J. P., Nosek, B. A., and Banaji, M. R. (2003). Contextual variations in implicit evaluation. *Journal of Experimental Psychology: General*, 132(3):455–469.

Ratcliff, R. and McKoon, G. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18(6):1155–1172.

Redington, M., Chater, N., and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469.

Sharifian, F. and Samani, R. (1997). Hierarchical spreading of activation. In Sharifian, F., editor, *Proceedings of the Conference on Language, Cognition and Interpretation*, pages 1–10. Isfahan: IAU Press.

Sterelny, K. (2006). Memes revisited. *The British Journal for the Philosophy of Science*, 57(1):145–165.