Theoretical Assessment of the SOILIE Model of the Human Imagination

Michael O. Vertolli (michaelvertolli@gmail.com)
Vincent Breault (breault_vincent@gmail.com)
Sébastien Ouellet (sebouel@gmail.com)
Sterling Somers (sterling@sterlingsomers.com)
Jonathan Gagné (gagne.jonathan@gmail.com)
Jim Davies (jim@jimdavies.org)

Institute of Cognitive Science, Carleton University
1125 Colonel By Drive, Ottawa, Ontario K1S 5B6 Canada

Abstract

We describe the overall theory of the SOILIE model of the human imagination. In this description, we outline cognitive capacities for learning and storage, image component selection and placement, as well as analogical reasoning. The guiding theory behind SOILIE is that visual imagination is constrained by regularities in visual memories.

Keywords: imagination; spatial cognition; creativity; analogy; visualization; cognitive model.

Introduction

The cognitive literature on imagination involves two related capacities: general creativity and the ability to generate mental simulations of possible worlds, often using sensory data from memory or the environment. The current focus is on the latter, particularly in the visual modality.

This type of imagination is implicated in a number of cognitive activities, including reading a novel, planning future actions, recalling previous experiences, fantasizing about the future, and dreaming (Davies, Atance, & Martin Ordas, 2011). Although imagination of visual phenomena is often thought to be identical with pictographic, mental imagery, the view described here sees the rendering of a mental image as a final, optional stage. The process of rendering an imagined scene into neural "pixels" (colors at particular locations) is usually preceded by processes that determine what is to be placed in the image and where. For example, if one is asked to picture "a computer and a mouse," one is likely to also picture a keyboard, desk, and related objects in an office or similar environment. The question is how does a mind know to combine these particular objects in their appropriate spatial configurations?

To address this question, we chose to model a task in which a given agent takes a single word (e.g., "computer") as the trigger to engage in the act of imagination. The task of the agent is to imagine a "computer" in a realistic scene. Using visual and spatial long-term memories, the agent populates the scene with elements that are likely to appear in an image with the triggering word (such as a keyboard). Once the underlying cognitive processes of the agent have selected what should appear in the image and where those things should be located, the mental scene is passed on for further processing—perhaps mental imagery.

Figure 1: SOILIE's imagined output given the query 'mouse' and the returned labels: 'computer', 'keyboard,' 'monitor,' and 'screen.'

The Model

The Science of Imagination Laboratory Imagination Engine (SOILIE) is a computational model composed of multiple subsystems that together create the informational precursors of a 2D visual scene from an environmental trigger or query. In its current implementation, the engine takes a single word as input and returns a collection of object labels and their relative positions. The over-arching goal is for SOILIE to create visually imagined scenes in the same way that humans do.

Many of SOILIE's underlying subsystems have been discussed in previous work (Breault, Ouellet, Somers, & Davies, 2013; Davies & Gagné, 2010; Somers, Gagné, Astudillo & Davies, 2011; Vertolli & Davies, 2013). In what follows, we will take a step back and look at the entire model as a whole, including parts that are not explicitly used to determine SOILIE's output. These elements contribute to the overall theory and include what is currently being worked on or extended in the model. Each of the parts will be addressed in chronological order as they might occur in an act of imagination.

This chronological account will outline the following processes and structures. The first area is the agent-world interface, or the point at which information in the

environment enters the agent and is compressed for storage. This step is a model of high-level perception and learning.

The second step is the selection process, or what can be viewed as a decompression of the originally stored information. In the context of the imagination, this step largely focuses on the determination of what objects should be included in the scene, on the basis of prior experience, and where they should be spatially situated. The result is that a new scene description, or design, is derived from compressed visual memories.

We will also discuss an analogy step that can derive yet unknown relations from semantically similar and previously known content. This step can best be seen as an additional sub-step of the decompression phase. First we will discuss how information goes from the environment into representations in the memory of the agent (learning), and then how the agent uses these representations to create imagined scenes (imagining).

Learning: Creation of Visual Memories

Many of SOILIE's subsystems model functional aspects of a human mind. The input to this agent is modeled with labeled 2D images of a 3D environment (mostly photographs), English text, and some simple objects and spatial relations that bind these two sources of information together (see Figure 2).

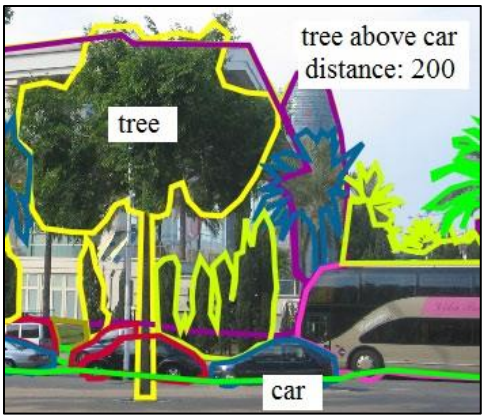

Figure 2: Input to the SOILIE agent

From the labeled images, SOILIE collects the pixel coordinates of objects and the labels of those objects (e.g., man, cat, car). Basic spatial relations between objects (corresponding to English words such as "above") are then manually encoded, which SOILIE decomposes into the component parts (e.g., tree above car, tree, above, car).

All of the content that SOILIE experiences as its external environment is derived from the Peekaboom database of labelled images. For each image, labels are associated with areas of pixels in order to indicate object positions in the image. For example, if the label is "computer," then the database knows the rough outline of the computer's pixel coordinates in the image.

Peekaboom The Peekaboom database is a collection of over fifty thousand labeled images and more than ten thousand object labels. This database is the combination of two online games: the ESP Game and Peekaboom (von Ahn, Liu, & Blum, 2006). In the ESP Game, two players are randomly paired on the internet. They are both shown the same image and try to enter the same words. Because they cannot communicate, these words usually describe an object in the image. When both players enter the same word, the system assumes the label is relevant to the image and stores it.

Peekaboom's strategy was effectively the same as the ESP Game, but it focused on the position of the labels or objects that were determined in the ESP game. One player would use mouse clicks to reveal parts of an image to another player. The second player's goal was to guess the label given to the first player. When the second player guesses the right word, it is assumed that the parts of the image revealed represent where the object is in the image.

The result of these two games is that labels are associated with objects in an image, and point clouds on specific locations in the image represent the location of the objects. An advantage of these games is that they can gather diverse and relatively accurate labeling data on large sets of images in a fast and efficient way (von Ahn & Dabbish, 2004). The fact that this information is derived from human judgment gives it cognitive legitimacy. Thus, the Peekaboom database is not only a collection of object labels and positions bound to images; it is a database that captures many implicit properties of human classification. SOILIE uses this database as a proxy for human visual memory.

Compression and Storage

The process of experiencing can be accurately described as is a special type of input where information is lost as it is converted into the internal structure of the agent (i.e., compressed, see Hutter, 2005; Schmidhuber, 2009; Wolff, 2013). For SOILIE, this structure is made up of exemplars, fuzzy magnitudes, prototypes, and co-occurrence probabilities. Each of these structures is represents a particular encoding of different associations in the world.

Exemplars Exemplars are internalized atoms of experience (Tulving, 1984; Davies & Gagné, 2010). For SOILIE, they map a collection of measurable properties, mainly angle and distance in the current implementation, to English words using a simple, context-free, recursive grammar. The result of this grammar is sentences like: `bird above field`, for example. The grammar allows such a sentence to be broken down into its component parts (e.g., bird, above, field, bird above field). Each part is then associated with the values of the corresponding angles and distances. In the current implementation, we use a single placeholder to encode all distance and angle relations (e.g., above, below) for simplicity.

A number of assumptions with important cognitive implications are related to this basic layout of the exemplar. First, some type of internalized grammar must be present in

the agent. Since associating measurement relations with places and objects is of such a low-level functionality, one would expect that many simpler organisms than humans possess this ability to some degree (Dehaene, 2009). To impose a lexical grammar on these species is harder to justify. Thus, it is better to think of this grammar as a type of world logic whose instantiation in a lexical syntax is not intended to be a model of human thinking.

Another assumption is that there must be at least some underlying conceptual equivalents to the categorical distinctions implied by the adjectives and prepositions in the sentence structure. We are agnostic as to whether these are instantiated in words, *per se*. The exemplar as a structure only requires that they be associated with some form of mental unit of measure that can be used to gauge relative difference between the various objects in their internalized representations.

A final assumption is that this unit of measure is only secondarily associated with a culturally invested measure (e.g., feet, liters, ounces). Preceding this is an internal representation of relative scale. Research in cognitive neuroscience supports this notion through a parallel concept: an analogical representation of quantities or magnitudes (for an extensive review, see Dehaene, 1992). This numerical representation in the brain is used for abstract calculations and comparisons of weight, area, size, etc. It is an inherently relative process, where magnitudes are compressive and follow a logarithmic distribution (i.e., 1 is 'far' relative to 10, but 100,000 is 'close' relative to 100,010). The neural representation is also equally relevant for visual and symbolic (e.g., Arabic numbers) domains (Buckley and Gillman, 1974). Thus, this assumption of a mental unit of measure is reasonable given this support in the literature.

Fuzzy magnitudes The next internal structure is fuzzy set membership for magnitudes (Zadeh, 1965). In order to better characterize the analogical representation of quantity previously described, SOILIE stores magnitudes as fuzzy set memberships in a range of fuzzy number values. This means that any given magnitude (e.g., 6) is stored a membership value between [0.0, 1.0], where 0 indicates a value is not in the set and 1 indicates that the value is clearly in the set. Thus, 6 would have a fuzzy set membership in the sets 5, 10, and so on. In this example, 5 and 10 are fuzzy numbers, and 6 has a partial membership in both of them. The membership value is determined by the equation for linear interpolation:

$$m_i = m_{i-1} + (m_{i+1} - m_{i-1})\frac{N_i - N_{i-1}}{N_{i+1} - N_{i-1}}$$

where m is a membership value at a given index i, N is the numeric value of the set at index i, and the intermediary values are calculated on the basis of the original number (e.g., $N = 6, m = 1.0$) and the outer bounds (i.e., $N_{i\pm3}$, $m = 0.0$; Davies & Gagné, 2010; Gagné & Davies, 2013; Somers, Gagné, Astudillo & Davies, 2011).

For distances and equivalent magnitudes, the range of sets is logarithmic as per the neuroscientific model (e.g., 0, 2, 5, 10, 20, 35, 60, 100, 160, 250, 400, 600, 900, 1350, 1800; Dehaene, Izard, Spelke, & Pica, 2008). For angles, the range is linear (e.g., -157.5, -135, -112.5, -90, -67.5, -45, -22.5, 0, 22.5, 45, 67.5, 90, 112.5, 135, 157.5, 180). Because angles are bounded or restricted to a 360 degree range, or 180 degrees to the left and right of the current position, we do not use logarithmic magnitudes for them. In summary, a particular magnitude is represented as an array of numbers indicating membership in the sets of the fuzzy numbers listed above.

Prototypes Following the work of Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976), prototypes are abstract generalizations of the redundancy structure of a given class of objects (i.e., of the properties that are the most characteristic of the class as a whole). They are instantiated as a synthesis of the specific experiential units of the exemplars with the fuzzy magnitudes previously described. Thus, each of the lexical component parts of the exemplar are bound with the quantity of experiences and the average fuzzy set membership for every possible measure (e.g., size). The average is calculated with this formula:

$$\bar{m}_{i,n} = \frac{n\bar{m}_{i,n-1} + m_{i,n}}{n+1}$$

where $\bar{m}$ is the average value at fuzzy set i for the n experience, and m is the membership value at set i for the n particular experience. An example prototype might be stored as:

```
Prototype: bird above field
Property: distance
Fuzzy set for distance property:
(… 0.0, 0.2, 0.75, 0.8, 0.25, 0.0 …)
Number of experiences: 1
Property: angle
Fuzzy set for angle property:
(… 0.0, 0.2, 0.5, 0.2, 0.0, 0.0 …)
…
```

Adjectives and prepositions get individual prototypes of their most characteristic use of each measure (e.g., weight, distance, angle) despite that "above" does not intelligibly have an obvious value for these measures. These scores are used to better represent higher-order correlations between the various measures and component parts (e.g., older birds might fly higher than younger birds so birds with a greater distance measurement above something often have a greater weight measurement).

The use of English words might be more confusing theoretically, but it lends itself to the determination of the prototype map for humans in a given domain. That is, in as much as the use of "duck" and "bird" in the database is typical of human usage and the real world, the abstract prototypes created from these words should also be typical

of human usage or the real world. Since SOILIE seeks to capture the visual equivalents of this typicality, this situation is desirable. In addition, the use of a more abstract set of prototypes and features, as per some holographic models (e.g., BEAGLE, see Jones & Mewhort, 2007), might better capture world relations outside of their human associations and conceptualizations.

As a final caveat, despite that SOILIE's prototypes are effectively groups of fuzzy sets, fuzzy logic is never used by the model nor should it be implied. Thus, the model is still consistent with the psychological literature that recognizes fuzzy logic as inconsistent with human, categorical reasoning (see Rosch, 2013).

Co-occurrence probability The final internal structure that will be covered in this section is co-occurrence probability (Vertolli & Davies, 2013), which is used to determine which labels get put in the image with the trigger label. These probabilities are derived from the presence of one object in the same prototype with another object bound by a preposition (e.g., `bird` and `field` in `bird above field`). The number of experiences (n) of each prototype is used to derive the probability P as:

$$P(i,j) = \frac{\sum n_{i,j}}{n_i}$$

where i and j are objects and n is the number of exemplars experienced for a prototype containing the indexed object(s).

It is worth noting that this is functionally isomorphic with taking the number of images containing both objects and dividing it by the total number of images containing object i so long as the prepositions do not have inverses (e.g., above *and* below). In the current implementation of SOILIE, they do not. Thus, the isomorphism holds.

Recent research in cognitive neuroscience supports the use of co-occurrence, if not co-occurrence probabilities, specifically. In the "memory space" hypothesis neuronal firing in the hippocampus encodes the elements of a given experience based on their spatiotemporal associations (Konkel & Cohen, 2009). Parallel research supporting the memory space hypothesis suggests that the hippocampus is also involved in the construction of conceptual knowledge and generalization (Kumaran, Summerfield, Hassabis, & Maguire, 2009). Thus, this research lends support to the use of co-occurrence probabilities, the abstractive capacities of the prototype structures, and their association in the constellation of neural processes in the hippocampus.

Learning Each of SOILIE's internal data structures features a lossy compression from the original image data. Thus, for this part of the processing, the exemplar loses all the details outside of the component features and simple sentences; the fuzzy magnitudes of angles and distances lose the original point-clouds of the objects in the image; the prototypes lose the differentiated experience events; and, the co-occurrence probabilities similarly lose the original experience instances. Though these processes are rather simplistic, the averaging present in the development of the prototypes and co-occurrence probabilities, for example, is reminiscent of learning in more complex computational models. Individual experiences are input into the model, processing occurs which integrates these experiences, and the internal representation changes as a consequence. Thus, this sequence of successive compression steps can be seen as a simple form of learning for the SOILIE model.

Imagining: Decompression and Selection

Once the original information is stored in memory, the agent must be able to use it in future circumstances. Any information that was lost from the original experiences (e.g., relative spatial positioning of the objects in a scene, what objects were in the scene) must be re-generated on the basis of internal procedures and the compressed data. In the context of the current task in the visual imagination, most of this regeneration is related to the reconstruction of the scene given a particular, one-word trigger or query (e.g., "mouse"). Thus, once more, it is consistent with contemporary research on the hippocampus, specifically scene construction theory (Hassabis & Maguire, 2007; Maguire & Mullally, 2013).

Two primary processes and one associated process will be outlined in this section. All of these processes seek to answer the questions "what" and "where." Through these processes the overall layout of a newly imagined scene is coordinated in working memory.

Recent research suggests that visual working memory can hold approximately three to five objects of average complexity (Cowan, 2001; Edin, et al., 2009). Thus, we chose to constrain SOILIE to this cognitive limitation. Although, there is the possibility of building more complex scenes through the use of chunking, the current model does not implement this functionality. An imagined scene contains the triggering object and four associated objects in a given spatial configuration.

Coherence and selection The first selection procedure determines what other objects would be present in a scene with a given source object or trigger. At present, the model currently only uses co-occurrence probabilities to regenerate a coherent selection of objects from memory. The process proceeds as follows.

First, a top-4 search gathers the pool of four objects with the highest co-occurrence with the query given to SOILIE. Because these returned objects often would form an incoherent scene (e.g., a bank with money and a river), an associative search assesses the co-occurrence probability between every pair of objects, including the query. It is worth noting that co-occurrence probability values are not commutative (i.e., $P(i, j) \neq P(j, i)$) so each pair of labels is given two values, one for each ordering of the pair. The average co-occurrence of the network of co-occurrence

relations that results is tested against a selection threshold (λ) as per this equation:

$$\frac{1}{20}\sum_{i=1}^{5}\sum_{j=1}^{5} P(o_i, o_j) > \lambda$$ [1]

Labels with low co-occurrence in the network are swapped out and new labels that co-occur with the query are randomly swapped in until the threshold for the network as a whole is exceeded. Once the threshold is exceeded, the set that remains is returned for further processing.

Vertolli and Davies (2013) used a train-test design with two random samples of the Peekaboom database to assess the efficacy of this process. The results suggest that it is a significant improvement in coherence over selecting the top-4 objects for a given trigger and a random selection. More recent research has shown that this approach is a significant improvement over a connectionist algorithm that Thagard (2000) argues is the best in the literature (Vertolli & Davies, 2014).

Spatial positioning Once the collection of objects are selected, the spatial configuration of these objects must be determined. The first step is the determination and selection of the corresponding object prototypes in memory. Should a higher-order prototype (e.g., a particular preposition or adjective combination) be missing or not yet exist, the underlying architecture does possess the ability to generalize and use analogies of the prototypes it has present in memory. However, only the latter, inference procedure is currently used in this instantiation of the model. Due to space constraints, we leave the rather complex discussion to Gagné and Davies (2010).

Once the prototypes are selected, the fuzzy magnitudes for the angles and distances are de-fuzzified according to this formula:

$$N = \frac{\sum_i m_i N_i}{\sum_i m_i}$$

where N is the crisp number, N_i is the ith number of the fuzzy set, and m_i is membership for ith number. Objects are then placed entirely relative to the triggering object. Following de-fuzzification of each element, SOILIE has determines what elements will be in the image and where they should appear. This scene description is then returned for processing by some future cognitive architecture.

Prototype analogy SOILIE uses the WordNet database (Fellbaum, 1998), specifically semantic distance, to determine the meaning of unrecognized words. For example, if there is no prototype for "mac above tiles" it might return "computer above floor," if the latter prototypes is present in memory and has the highest similarity index (see Wu & Palmer, 1994). This allows the program to make semantic inferences on the basis of approximate information for objects it does not yet know (Gagné and Davies, 2010). Currently, this functionality is restricted to the spatial properties, but future instantiations of the project plan to generalize for coherence probabilities, as well.

Final Generation of an Image With the objects chosen, and their locations determined, all that is left is for the model to actually place pixels on a canvas. In humans, this might be visual imagery. Our model's method of generating imagery is intended for demonstration purposes only. We do not propose SOILIE's method of imagery as a model of human imagination.

SOILIE chooses a random instance of the label from the LabelMe database (e.g., a "computer") and places the pixels on a canvas in the correct place. This results in images such as that depicted in Figure 1.

Conclusion

In summary, SOILIE is a model of the visual processes of the human imagination consistent with empirical findings in cognitive science. We have outlined the step-by-step processes as they might occur in the model in its simulated world. Throughout the discussion, we describe the many assumptions and implications that are distributed through the current implementation of the model. The exposition thus provides another step in the integration of the cognitive, computational, and neuroscientific domains implicated in our approached.

References

Buckley, P.B., & Gillman, C.B. (1974). Comparison of digits and dot patterns. *Journal of Experimental Psychology, 103*, 1131-1136.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, Mass.: MIT Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87-114.

Davies, J., Atance, C. & Martin Ordas, G. (2011). A framework and open questions on imagination in adults and children. *Imagination, Cognition, and Personality, Special issue on mental imagery in children*. 31:1-2, 143-157.

Davies, J. & Gagné, J. (2010). Estimating quantitative magnitudes using semantic similarity. *The American Association for Artificial Intelligence workshop on Visual Representations and Reasoning (AAAI-10-VRR)*, 14-19.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*(1), 1-42.

Dehaene, S. (2009). Origins of mathematical intuitions. *Annals of the New York Academy of Sciences, 1156*(1), 232-259.

[1] The diagonal, where $i = j$ or the co-occurrence of an object with itself, is ignored. Thus, the denominator of the average has to be decremented by the cardinality of this diagonal (i.e., by 5).

Dehaene, S., Izard, V., Spelke, E., & Pica, P. 2008 Log or Linear? Distinct Intuitions of the Number Scale in Western and Amazonian Indigene Cultures. *Science, 320*(5880), 1217-1220.

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J., & Compte, A. (2009). Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences, 106*(16), 6802-6807.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Gagne, J. & Davies, J. (2013). Visuo: A model of visuospatial instantiation of quantitative magnitudes. *Knowledge Engineering Review: Special Issue on Visual Reasoning*, 1-20.

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in cognitive sciences*, *11*(7), 299-306.

Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer.

Jones, M. N., & Mewhort,D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 1-37.

Konkel, A., & Cohen, N. J. (2009). Relational memory and the hippocampus: Representations and methods. *Frontiers in Neuroscience, 3*, 166–174.

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the emergence of conceptual knowledge during human decision making. *Neuron, 63*(6), 889-901.

Maguire, E. A., & Mullally, S. L. (2013). The hippocampus: A manifesto for change. *Journal of Experimental Psychology: General*, *142*(4), 1180.

Rosch, E. (2013). Neither Concepts Nor Lotfi Zadeh are Fuzzy Sets. *On Fuzziness* (pp. 591-596). Springer Berlin Heidelberg.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382-439.

Schmidhuber, J. (2009). Driven by compression progress: a simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In G. Pezzulo, M. V. Butz, O. Sigaud, and G. Baldassarre (eds.), *Anticipatory Behavior in Adaptive Learning Systems, from Sensorimotor to Higher-level Cognitive Capabilities*, Springer.

Somers, S., Gagné, J., Astudillo, C., & Davies, J. (2011). Using semantic similarity to predict angle and distance of objects in images. *Proceedings of the 8th ACM Conference on Creativity & Cognition* (pp. 217-222). Atlanta, GA.

Vertolli, M. O. & Davies, J. (2013). Visual imagination in context: Retrieving a coherent set of labels with Coherencer. In R. West & T. Stewart (eds.), *Proceedings of the 12th International Conference on Cognitive Modeling*, Ottawa: Carleton University.

Vertolli, M. O. & Davies, J. (2014). Coherence in the visual imagination: Local hill search outperforms Thagard's connectionist model, *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Quebec City, QC: Cognitive Science Society.

von Ahn, L., and Dabbish, L. (2004) Labeling Images with a Computer Game. *ACM Conference on Human Factors in Computing Systems* (CHI)

von Ahn, L., Lui. R., & Blum. M. (2006) Peekaboom: A game for locating objects in images. *Proceedings of the SIGCHI conference on Human Factors in computing system* (pp. 55-64).

Wolff, J. G. (2013). The SP theory of intelligence: an overview. *Information*, *4*(3), 283-341.

Wu, Z. & Palmer, M. (1994). Verb semantics and lexical selection. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics* (pp 133-138).

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338- 353.