AISB Journal

The Interdisciplinary Journal of
Artificial Intelligence and the Simulation of Behaviour

Volume 1 – Number 6 – June 2005

The Journal of the
Society for the Study of Artificial Intelligence
and the Simulation of Behaviour
http://www.aisb.org.uk

Published by
The Society for the Study of Artificial Intelligence and Simulation of Behaviour

http://www.aisb.org.uk/

ISSN 1476-3036

Contents

A Computational Model of the Trait Impressions of the Face for Agent Perception and Face Synthesis 481
Sheryl Brahnam

Modelling Vocabulary Acquisition: An Explanation of the Link between the Phonological Loop and Long-Term Memory 509
Gary Jones, Fernand Gobet & Julian M. Pine

A Generic Negotiation Model using XML 523
Philippe Mathieu and Marie-Hélène Verrons

Eliciting Test-selection Strategies for a Decision-Support System in Oncology 543
Danielle Sent, Linda C. van der Gaag, Cilia L. M. Witteman, Berthe M. P. Alemany & Babs G. Taaly

AISB Journal **1**(6), © SSAISB, 2005
http://www.aisb.org.uk

A Computational Model of the Trait Impressions of the Face for Agent Perception and Face Synthesis

Sheryl Brahnam

Department of Computer Information Systems, Missouri State University
901 South National, Springfield, MO, 65804, USA
shb757f@smsu.edu

Abstract

This paper reports a first attempt at developing a computational model of the trait impressions of the face for embodied agents that accommodates the social perception and social construction of faces. Holistic face classifiers, based on principle component analysis (PCA), were trained to match the human classification of faces along the bipolar rating extremes of the following trait dimensions: adjustment, dominance, warmth, sociality, and trustworthiness. Although results were marginally better than chance in matching the perception of dominance (64%), classification rates were significantly better than chance for adjustment (71%), sociality (70%), trustworthiness (81%) and warmth (89%). A second exploratory study demonstrates how PCA models of trait classes could be used by agents to generate faces. Novel faces were synthesized by probing specific PCA trait attribution spaces. Human subjects were then asked to rate the synthesized faces along a number of trait dimensions, and it was found that the synthesized faces succeeded in eliciting predicted trait evaluations.

1 Introduction

The semiologist, Magli (1989) has remarked, "upon seeing a face, we immediately produce a symbolic framework that confronts us with a complex and ancient cultural experience" (p. 90). A recurrent theme in the fables, proverbs, and histories, both oral and written, of cultures as diverse as the Egyptian, African, Chinese, and European is that the face is inscribed with signs that reveal the essence of a person's inner soul (Frey, 1993). Although many modern people scoff at such notions and recite such maxims as "Don't judge a book by its cover," there is considerable evidence in the person perception literature that people are predisposed to form impressions of a person's social status, abilities, dispositions, and character traits based on nothing more than that person's facial appearance. Furthermore, there is evidence that these judgments influence and guide people's behavior towards others, especially in situations that are ambiguous or where little information about a person is known (Hochberg and Galper, 1974). As the historian Frey (1993) recently noted, "To this day, the quest to read a person's inner world from her outer appearance has lost nothing of its momentum . . . it seems that the advent of the 'Age of Television' has given additional impetus to the age-old fascination with human appearance" (p. 64).

People are not just caught up in evaluating other people's faces; they are equally preoccupied with managing their own appearances. One would be hard pressed to name one culture that has not required its members to modify their faces in some way. The psychologist Ligget (1974) has observed, "The desire to alter the face is universal; in every culture and in every age examples of facial elaboration can be found" (p. 46). The need to mark a person's social status, to proclaim skill in hunting and in war, and to put one's best face forward at a business meeting are some of the many motives behind facial elaborations. The face, more than any other part of the body, stands for and is identified with the social self, and so important are the social consequences of the appearance of the face that many people are willing to endure enormous pain and expense to manage the messages sent by their faces. It seems that the French poet Henri Michaux may have been right when he wrote, "We lead an excessively facial life" (quoted in Landau, 1989, p. 234).

Once embodied agents enter the social arena, they will be expected to understand the cultural language of the face and not just short-term surface communications and behaviors, such as eye blinking, gazing, head tilting, facial gestures, and emotional expressiveness, which forms the focus of current research into agent faces (Pelachaud and Poggi, 2002). As essential as this research is, research that explores the underlying morphology, or the look of the face, is also important. Sproull et al. (1996) have demonstrated, for instance, that morphological shifts in the facial appearance of embodied agents affect users in ways that mirror findings in the person perception literature, and Donath (2001) and others have cautioned researchers to consider carefully the facial appearance of their agents. Unfortunately, there is no way to predict during design time all the circumstances, tasks, and people the agent will encounter. Thus, there is no way to equip an agent with an embodiment that will function optimally in all situations.

Although people today have recourse to plastic surgery and a host of cosmetic aids, there is a limit to the extent that people can shape their faces for social purposes, but embodied agents do not share this limitation. There is no reason to assume that a particular agent's embodiment must be singular or static or that it must be designed offline by human beings. Facial morphology could be as expressive a channel of communication for embodied agents as are emotional facial displays. Like countless others who each morning prepare their faces to meet the demands of their day, so embodied agents could learn to construct social masks that are appropriate for the situations they encounter, the users they meet, and the tasks they need to accomplish.

To participate in the social world, embodied agents will also need to know how to evaluate the human faces they encounter. Rather than use a predefined set of interaction tactics and practices, for instance, the cultural information found in the user's face could serve the agent as a basis for formulating a more personalized and realistic initial interaction strategy that could then be adjusted as further information about a user is obtained. At the very least, predicting how other human beings would react to a person's facial appearance could produce interaction strategies that mimic the more natural interaction styles of human beings. Understanding the social language of the face will also allow agents to participate in such common activities as commenting on the appearance of others in ways that are realistic and appropriate. This could enhance the agent's believability and acceptability. Being able to perceive faces as people perceive them could also make embodied agents more sensitive in their encounters.

But what is the social language of the face? How can it be modeled for embodied agents? An ongoing area of investigation in social psychology revolves around understanding the facial characteristics that contribute to the formation of impressions about a person's character. As reviewed in section 2, it has been found that large clusters of char-

acter traits are strongly associated with attractiveness judgments, emotional displays, age, and gender (Zebrowitz, 1998). As a result, recent research in this area has focused almost exclusively on investigating the facial characteristics of attractiveness, emotion, age, and gender and the role these characteristics play in the attribution process. Research aimed at directly exploring morphological characteristics that trigger very specific attributions has all but been neglected, primarily because this line of investigation has mostly been feature based, has produced contradictory results, and has not lent itself to theory building.

In section 3, it is argued that a psychologically viable model of the trait attribution process is not necessary for embodied agents. Rather, since agents need to perceive faces in terms of the impressions they produce, it would be best to model specific traits directly using holistic face recognition techniques, such as principle component analysis (PCA), or equivalently, linear autoassociative neural networks. As noted in section 4, these techniques have already proven successful at classifying faces according to identity (Turk and Pentland, 1991b), emotion (Padgett and Cottrell, 1998), gender, and age (Valentin et al., 1994a), characteristics that are strongly correlated with impression formation. Thus, it is reasonable to expect that these classifiers will succeed in modeling the human classification of faces into specific trait attribution classes. Another advantage in using holistic face classification techniques is that they lend themselves to face synthesis (Vetter and Poggio, 1997) and, thus, could be used by agents to generate faces with a high probability of making specific impressions on users.

Two studies are reported in this paper that use PCA to model the trait impressions of the face. The objective of the first study was to model the trait impressions of facial morphology. As described in detail in section 5, PCA classifiers were trained to classify faces that were rated either high or low within the five trait dimensions of adjustment (adjusted/unadjusted), dominance (dominant/submissive), warmth (warm/cold), sociality (social/unsocial), and trustworthiness (trustworthy/untrustworthy). A second exploratory study, presented in section 6, demonstrates how PCA classifiers can be used to create novel faces calibrated to produce specific trait impressions. The results and some limitations of the two studies are discussed in section 7, and the paper is concluded in section 8 by noting some of the contributions of these studies and by offering directions for future research.

2 Person Perception Literature on the Trait Impressions of the Face

As mentioned in the Introduction, psychological research aimed at directly exploring morphological characteristics that trigger very specific trait attributions has virtually been neglected in large part because this approach has not lent itself to theory building. Although several theories have been advanced to explain why it is that certain facial characteristics consistently elicit specific personality impressions, one major theory is that the perception of facial features has adaptive value and that those trait impressions that have the most influence are based on those facial qualities that demand the greatest attention for survival (McArthur and Baron, 1983). Recognizing an angry face, for example, triggers lifesaving fight/flight responses or conciliatory behaviors. It is theorized that faces that are similar in structure to angry faces elicit similar, albeit milder, responses. As Zebrowitz (1998) explains, "We could not function well in this world if we were unable to differentiate men from women, friends from strangers, the angered from the happy, the healthy from the unfit, or children from adults. For this reason, the tendency to respond to the facial quali-

ties that reveal these attributes may be so strong that it is *overgeneralized* [italics mine] to people whose faces merely resemble those who actually have the attribute" (pp. 14–15). Two of the most researched *overgeneralization effects* are the attractiveness halo effect and the facial maturity overgeneralization effect. Two other overgeneralization effects that have received less attention but are nonetheless significant are based on emotion and gender (Alley, 1988; Symons, 1979).

The trait associations and morphological characterizations of each of these overgeneralization effects are summarized below. Included in the summary are descriptions of some of the more important models of facial attractiveness, maturity, gender, and emotion.

2.1 Attractiveness Halo Effect

It is popularly believed that social benefits accrue to those who are most attractive, and current research supports this claim. People respond positively to attractiveness and associate it with positive character traits. Attractive people are considered more socially competent, potent, healthy, intellectually capable, and moral than those less attractive. They are also perceived as being psychologically more adapted (Langlois et al., 2000). Facial abnormalities and unattractiveness, in contrast, elicit negative responses and are associated with negative traits (Langlois et al., 2000). Unattractive people are considered less socially competent and willing to cooperate (Mulford et al., 1998). They are also considered more dishonest, unintelligent, and psychologically unstable and antisocial. Unattractive people are often ignored and, if facially disfigured, avoided (Bull and Rumsey, 1988). The unattractive are also more likely to be objects of aggression (Alcock et al., 1998) and to suffer abuse (Langlois et al., 2000).

What are the morphological characteristics that make a face attractive? To date there is no theory of attractiveness that is generally accepted. Nonetheless, contemporary research into facial attractiveness indicates that straightness of profile (Carello et al., 1989) and closeness to the average (Langlois and Roggman, 1990) are some important factors in attractiveness judgments.

Physical anthropologists have identified three types of facial profiles that depend on measures of straightness: the orthognathic, retrognathic, and prognathic (Enlow and Hans, 1996). The three types can be spotted by determining the position of the chin in terms of a vertical line that drops down along the upper lip and which is perpendicular to a horizontal line that extends outward from the eyeball. A chin that is inside the vertical line produces a retrognathic profile, whereas a chin that extends outside the line, along with the nose, is prognathic. Many studies have demonstrated a preference, even among children, for orthognathic or straight profile shapes (Carello et al., 1989; Magro, 1997; Lucker and Graber, 1980). Least attractive is the prognathic (Carello et al., 1989).

Particularly noteworthy, in terms of its potential for adjusting the impressions of agent faces, is the finding that facial attractiveness increases as faces are moved closer to the average (see Figure 1). One of the first to create and explore average faces was Francis Galton (1878), who did so by ingeniously superimposing photographs of more than one face. His major objective was to obtain a representation of various classes of people: criminals, the healthy, the ill, and the famous. To his surprise, the composites appeared notably more attractive. Little was done with his observation until 1952, when Katz (1952) maintained that composites are more beautiful than the individual faces comprising them by virtue of the fact that they are closer to the average. The first systematic study to lend support to his claim, however, had to wait until 1990, when Langlois and Roggman (1990), using digitized photographs of student faces, demonstrated not only that composites are thought more attractive but also that perceived attractiveness increases as more and more

faces are averaged, with the average being computed arithmetically using the gray scale pixel values of the constituent images. A year later, Langlois, Roggman, Musselman and Acton (1991) produced additional evidence that this preference for the average is exhibited by infants as well as adults. Their findings have more recently been confirmed by Langlois, Roggman, and Rieser-Danner (1990), Rubenstein et al. (1999), and Rhodes and Tremewan (1996).

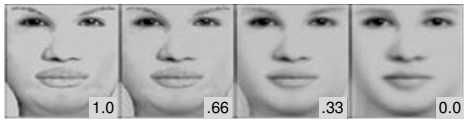


Figure 1: Increased Attractiveness of Averaged Faces. A face (1.0) moved towards the mean (0.0) of 220 randomly generated faces increases in attractiveness. The faces were generated from facial features in the composite program FACES by InterQuest and Micro-Intel, and the average face was computed by averaging the pixel gray scale values of the faces.

2.2 Facial Maturity Overgeneralization Effect

Perhaps no face is more capable of eliciting a favorable response than that of a baby. The favorable response to a baby's face is not just reserved for babies, however, but is generalized to adults whose faces resemble those of babies (Zebrowitz, 1998). Babyfaced people are universally attributed childlike characteristics. They are perceived to be more submissive, naïve, honest, kindhearted, weaker, and warmer than others. They are also perceived as being more helping, caring, and in need of protection (Berry and McArthur, 1986). Mature-faced individuals, in contrast, are more likely to command respect and to be perceived as experts (Zebrowitz, 1998).

The morphological characteristics that mark a baby's face are large eyes relative to the rest of the face, fine, high eyebrows, light skin and hair color, red lips that are proportionally larger, a small, wide nose with a concave bridge, and a small chin. The facial features are also placed lower on the face (Zebrowitz, 1998).

Figure 2: Negative (Left) to Positive (Right) Cardioidal Strain Transformations. Reproduced from Pittenger and Shaw (1975), p. 376. Copyright ©1975 by the American Psychological Association. Reprinted with permission.

Other significant age related differences in faces concern developmental changes in craniofacial profile shape. Of particular note are differences in the relative size of the brain capsule and the slant of the forehead in relation to the chin. The infantile cranium is proportionally much larger than the fully mature cranium, and the infantile forehead protrudes whereas the adult forehead recedes. Another important characteristic is a dramatic increase in jaw size.

Figure 2 illustrates the morphological characteristics of facial maturity. The craniofacial profile shapes were produced using a cardiodial strain transformation developed by Todd and Mark (1980). Applied to standard profile shapes, a positive application of the transformation has been shown to approximate real growth (Todd et al., 1981; Todd and Mark, 1980). As would be expected, studies on the trait attributions of profiles that vary in the degree of cardioidal strain applied are consistent with findings on facial maturity (Zebrowitz, 1998; Alley, 1983). As craniofacial profile maturity decreases, so do perceived alertness, reliability, intelligence, and strength (Berry and McArthur, 1986). Moreover, infantile profile shapes are more loveable, less threatening (Berry and McArthur, 1986), and elicit stronger desires to nurture and protect (Alley, 1983).

Examining Figure 2, it can be observed that an extreme negative cardioidal transformation results not only in the most youthful but also the most retrognathic profile shape. Similarly, an extreme application of a positive cardioidal transformation produces the most mature looking and prognathic profile shape. As noted above, profile shape is related to attractiveness judgments, and there is some evidence that the cardioidal transform influences attractiveness judgments as well as judgments regarding facial maturity (Jones, 1995).

2.3 Gender Overgeneralization Effect

The gender overgeneralization effect is strongly correlated with facial maturity (Zebrowitz, 1998). Female faces, more than male faces, tend to retain into adulthood the morphological characteristics of youth (Enlow and Hans, 1996) and are more likely to be ascribed characteristics associated with babyfacedness: female faces are thought to be more submissive, caring, and in need of protection. Similarly, male faces, tending to be morphologically more mature, are perceived as having the psychological characteristics typically associated with mature-faced individuals: male faces are thought to be more dominant, intelligent, and capable.

2.4 Emotion Overgeneralization Effect

While many social psychologists believe that facial impressions of character are related in part to the morphological configurations that characterize emotional displays, the overgeneralization effect of emotion has not received as much attention as some of the other overgeneralization effects. Nonetheless, there is evidence supporting the idea that morphological configurations suggestive of emotional expressions play a role in the formation of trait impressions. Take smiling for instance. People react positively to smiling faces and find them disarming and thus not very dominant (Keating et al., 1981a). As illustrated in Figure 3, facial dominance significantly declines where even a slight smile is discernible (Mueller and Mazur, 1996). As would be expected, faces where the lips naturally turn upwards are likewise viewed more positively; such faces are considered friendly, kind, easygoing, and nonaggressive (Secord et al., 1954). In a similar vein, faces that have features indicative of anger or hostility, e.g., low-lying eyebrows, thin lips, and

withdrawn corners of the mouth, are perceived to be more threatening, aggressive, and dominant (Keating et al., 1981b).

The morphological characteristics of various emotional displays are well understood due in large part to the facial action coding system (FACS), developed by Ekman and Friesen (1978). FACS describes any facial behavior, including emotion. Recently, a number of emotion recognition systems have been developed that use FACS (Bartlett, 1998; Donato et al., 1999; Essa and Pentland, 1997). There is also a growing body of research concerned with synthesizing emotional displays in artificial faces (Massaro, 1997; Picard, 1997; Waters and Terzopoulos, 1992).

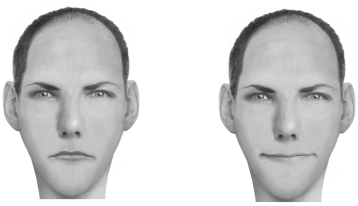


Figure 3: Illustration of the Overgeneralization Effect of Emotion. In the two images, only the lips differ. These faces were generated using FACES by InterQuest and Micro-Intel.

3 Problems with Indirectly Modeling the Trait Impressions of the Face

After reviewing the person perception literature on the overgeneralization effects, it might seem that one effective way to model the trait impressions of the face for agent perception and for face synthesis would be to do so indirectly by modeling facial attractiveness, maturity, gender, and emotion. Certainly an agent could alter the social impact of its face by moving it either further or closer to the average or by applying the cardioidal strain transformation or by *freezing* certain emotional displays. Although building agents that learn best how to adapt their faces using these techniques is a research area worth investigating, there are a number of problems with an indirect approach to modeling the trait impressions of the face.

A major problem concerns the difficulty of using models of attractiveness, facial maturity, gender, and emotion to predict, or to classify, faces in terms of the traits they elicit. In other words, these models would not readily provide embodied agents with *perceptual systems* capable of decoding the impressions faces make on human observers. An exception to this concerns the overgeneralization effect of emotion. As noted above, a number of systems have been developed that recognize and produce emotional facial displays. Most emotion recognition systems, however, utilize FACS, which describes surface facial behaviors more than it describes facial morphology. Recently emotion recognition classifiers have been developed that are based on holistic face recognition techniques (Cottrell and Metcalfe, 1991; Padgett and Cottrell, 1998). Even though these classifiers take into account all the information available in pixel representations of faces, not enough

is known about the relationship of the overgeneralization effect of emotion and the person perception of the face to utilize this technology in this task domain. Furthermore, in terms of production and recognition, it is doubtful that morphological characteristics in common with emotional displays can account for a significant range of traits. What emotional display, for example, best reflects honesty or intellectual competence?

This last point highlights a number of other problems with trait associations and the overgeneralization effects. First, the overgeneralization effects are associated less with individual traits than with clusters of traits. Knowing, therefore, which facial characteristics to alter in order to shift facial impressions along specific trait lines would require a much more refined understanding of the overgeneralization effects. Second, it is possible that some trait impressions may be due to facial configurations that are not accounted for by the overgeneralization effects. Third, no single overgeneralization effect accounts for a comprehensive set of traits. What is a comprehensive set of traits? Rosenberg (1977) has conducted an extensive study of this subject. Employing free-response methods, he has determined seven broad categories that are used to characterize others: intellectual competence, maturity, attractiveness, integrity, sociability, concern for others, and psychological stability. Others have modified his categories to include potency, or dominance (Feingold, 1992; Eagly et al., 1991). To encompass this representative set of traits in developing facial perception systems for embodied agents, all the facial configurations association with the various overgeneralization effects would need to be addressed.

A better approach to take in modeling the trait impressions of the face for embodied agents is to focus directly on the perception of those facial features that give rise to specific trait impressions. It has already been remarked that psychological studies have recently steered away form this line of research because this approach has failed to produce viable psychological theories of the trait attribution process. However, a model of the trait impressions of the face for embodied agents need not be as comprehensive and as capable of explaining the attribution process as psychological models need to be. Focusing on the *perception* of traits in faces using, for example, holistic face classification techniques would allow the classifier to discover the relevant features in trait formations. Other advantages in using holistic face recognition technologies to model the trait impressions of the face are presented in the next section.

4 PCA Face Representation and Classification

Isolating the features that hold the keys to an understanding of how faces can be processed, whether by human beings or by machines, has proven a difficult task. Much of the visual information contained within a face is highly redundant. What varies is but a small set of relations between features and small differences in textures, complexions, and shapes. Historically, the bulk of research has relied on measuring the relative distances between important facial key points: eye corners, mouth corners, nose tip, and chin edge (Brunelli and Poggio, 1993). Although this approach has the advantage of drastically reducing the number of variables, a major drawback is the difficulty in determining the best set of key points to measure (Valentin et al., 1994b; Burton et al., 1993).

An alternative approach is to process faces holistically (Brunelli and Poggio, 1993). Holistic techniques, such as template matching, preserve much of the information contained in the original images and are often preferred because they allow a classifier system to discover the relevant features a posteriori. Furthermore, template approaches have been shown to outperform feature-based systems (Lanitis et al., 1997).

Two related forms of template matching that have achieved considerable success at

classifying faces are linear autoassociative neural networks and a technique based on what is known as the Karhuen-Loève expansion in pattern recognition or PCA in the statistical literature. Since a linear autoassociative neural network is equivalent to finding the principal components of the cross-product matrix of a set of inputs, it is sometimes referred to in the literature as a PCA neural network (Oja, 1992; Diamantaras and Kung, 1996). Kohonen (1977) was one of the first to use a linear autoassociative neural network to store and recall face images. Sirovich and Kirby (1987) were the first to apply PCA to the data compression of faces and succeeded in economically representing faces in terms of an eigenpicture coordinate system. Turk and Pentland (1991a) adapted their techniques into what has now become a popular method of face classification.

The central idea behind PCA is to find an orthonormal set of axes pointing in the direction of maximum covariance in the data. In terms of facial images, the idea is to find the orthonormal basis vectors, or the eigenvectors, of the covariance matrix of a set of images, with each image treated as a single point in a high dimensional space. It is assumed that the facial images form a connected subregion in the image space. The eigenvectors map the most significant variations between faces and are preferred over other correlation techniques that assume every pixel in an image is of equal importance, (see, for instance, Kosugi, 1995).

Since each image contributes to each of the eigenvectors, the eigenvectors resemble ghostlike faces when displayed. For this reason, they are oftentimes referred to in the literature as *holons* (Cottrell and Fleming, 1990), or *eigenfaces* (Turk and Pentland, 1991a), and the new coordinate system is referred to as the *face space* (Turk and Pentland, 1991a). Some examples of eigenfaces are shown in Figure 4.

Figure 4: First 10 Eigenfaces of 220 Randomly Generated Faces. The eigenfaces of 220 randomly generated faces are ordered left to right, top to bottom, by magnitude of the corresponding eigenvalue.

Individual images can be projected onto the face space and represented exactly as weighted combinations of the eigenface components (see Figure 5). The resulting vector of weights that describe each face can be used in data compression and in face classification. Data compression relies on the fact that the eigenfaces are ordered, with each one accounting for a different amount of variation among the faces. Compression is achieved by reconstructing images using only those few eigenfaces that account for the most variability (Sirovich and Kirby, 1987). This results in dramatic reduction of dimensionality. Classification is performed by projecting a new image onto the face space and comparing the resulting weight vector to the weight vectors of a given class (Turk and Pentland, 1991a,b).

To date, no face classification methods have been applied to the task of classifying faces according to the traits they produce. However, because of the structural similarities between female faces and baby faces, the relation of attractiveness to average faces, and emotional expression to morphological facial characteristics that resemble the expressions associated with various emotions (see section 2), it is reasonable to assume that PCA can be employed to this end. Both linear autoassociative neural networks and PCA have successfully been used to classify faces according to gender (Valentin et al., 1997; O'Toole and Deffenbacher, 1997), age (Valentin et al., 1994b), and facial expression (Cottrell and Metcalfe, 1991; Padgett and Cottrell, 1998). What is more, they are simple, well understood, and capable of generating novel images from within the eigenface coordinate system (Turk and Pentland, 1991a; Beymer et al., 1993).

Figure 5: An Illustration of the Linear Combination of Eigenfaces. The face to the left can be represented as a weighted linear combination of eigenfaces.

5 Modeling Trait Impressions of the Face Using PCA

This section describes a model of the perception of traits in faces using PCA. The traits modeled were a modification of Rosenberg's (1977) factor analysis of significant trait descriptors, namely, psychological adjustment (adjusted/unadjusted), dominance (dominant/submissive), sociality (social/unsocial), trustworthiness (trustworthy/untrustworthy), and warmth (warm/cold). For definitions of the traits used in this study, the reader is referred to table 8 in the appendix.

5.1 Overview

As illustrated in Figure 6, modeling the trait impressions of the face using PCA was a two-step process. The objective of step 1, Data Preparation, was to obtain sets of faces clearly representative of ten bipolar trait descriptors (adjusted/unadjusted, dominant/submissive, social/unsocial, trustworthy/untrustworthy, warm/cold) of the five trait dimensions of adjustment, dominance, sociality, trustworthiness, and warmth. In Step 2, PCA Modeling, these attribution class sets were used to train and to test a separate PCA for each trait dimension.

5.2 Data Preparation

The objective of the data preparation process was to prepare faces for PCA classification. As illustrated in Figure 6, this step involved the following: A) generation of the stimulus faces, B) an experiment assessing the trait impressions of the stimulus faces, and C) division of the stimulus faces into trait class sets.

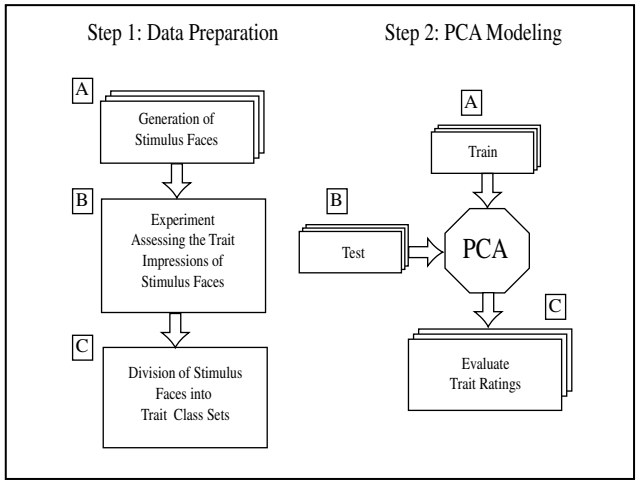


Figure 6: Modeling the Trait impressions of the Face. Note: Although technically PCA is not trained, perhaps because of the equivalency of autoassociative neural networks and PCA, the term *training* is often used in the face recognition literature, (see, for instance, Turk and Pentland, 1991a).

5.2.1 Generation of Stimulus Faces

In order to model the trait impressions of the face, it was necessary to acquire a suitable set of stimulus faces. In the person perception literature, stimulus faces are of three types: photographs of faces, drawings of faces, and faces pieced together using facial composite products such as Identi-Kit (Bruce, 1988). No database of faces known to elicit specific trait impressions has been developed for psychological comparison studies. Researchers are required to develop their own datasets of faces.

In contrast, numerous facial databases have been developed to test face classification algorithms. Wegener-Knudsen et al. (2002) provides a comprehensive review of available face databases. However, because these databases have been developed primarily to evaluate face identification techniques, these databases typically contain numerous photographs of a small set of individuals that vary in pose, lighting conditions, facial expression, and the addition of such occluding accessories as hats and glasses.

To model the trait impressions of the face, it was important to develop a large set of faces that were representative of a broad range of facial types and features. Furthermore, since the objective of this study was to model the trait impressions of *facial morphology*, the faces also needed to be as neutral in facial expressions as possible, and have such incidentals as hairstyle and accessories removed. Developing a proper database of faces for this task is a complex issue and is discussed further in section 8.

For this initial study, permission was obtained to generate faces using the full database of photographs of facial features (eyes, mouths, noses, and so forth) found in the popular composite software program FACES (Freierman, 2000), produced by InterQuest and Micro-Intel. With FACES, it was possible to generate randomly a fairly large number of unique faces by manipulating individual facial features. Moreover, by using specific sets of facial features, faces could be reduced to their basic morphological elements without having to block out features, such as hair, with tape or markers as is typically the case with cognitive and machine recognition studies involving photographs of people's faces.

Two hundred and twenty stimulus faces were generated in step 1.A using FACES. The features selected for constructing the stimulus faces included only the full set of 512 eyes, 541 noses, 570 lips, 423 jaws, 480 eyebrows, and 63 foreheads. Excluded were all sets of facial lines, hair, and accessories.

These images were then cropped (see Figure 7) in such a way that missing hair was less noticeable. This did remove forehead information; but, as these were frontal images of faces, the significance of profile head shape noted in section 2.2 was not relevant here. Care was taken to retain eyebrows, however, because they have been found to contribute to both impressions of gender and of facial maturity (Yamaguchi et al., 1995). A final alteration in the images concerned complexion values, which were set to the value of 190 in a gray scale of 256 values to reduce the effects of race.

Figure 7: Examples of Stimulus Face

5.2.2 Experiment Assessing Trait impressions of Stimulus Faces

Once the stimulus faces were generated, they were evaluated in step 1.B by human subjects as detailed below.

Participants. One hundred ten (54 male, 56 female) upper level undergraduate students were recruited from a large urban university to judge the stimulus faces. Each student received extra credit in a Computer Information Systems (CIS) course for participating in the study.

Dependent Measures. Each subject judged a set of 20 faces, randomly selected from the 220 stimulus faces, along the five trait dimensions, using a 7-point bipolar scale. Each image was judged by 10 subjects. The order of the bipolar trait descriptors (adjusted/unadjusted, warm/cold, social/unsocial, dominant/submissive, trustworthy/untrustworthy) was randomized as were the association of the bipolar descriptors with the anchor values of 1 and 7. Subjects were also given trait definitions and, in some cases, behavioral potential questions modeled after Berry and Brownlow (1989) and Zebrowitz and Montepare (1992). Refer to Table 8 in the appendix for the term definitions and the behavioral potential questions. A 7-point scale, rather than a 3-point scale, was used because the general consensus is that it is better to provide research subjects with a gradient of opinion when conducting surveys (Converse and Presser, 1986; Friedman and Amoo, 1999).

Apparatus. Desktop computers in a lab setting were used both to display the stimulus faces and to administer the questionnaires.

Results. Table 1 presents the mean ratings of the 220 faces for each trait dimension and the standard deviations. In general, the impressions elicited by the stimulus faces were slightly skewed towards low facial warmth and high adjustment, dominance, sociality, and trustworthiness.

As subjects were not required to judge the entire set of 220 faces, a complete analysis of human variance is not presented. It is important to stress that the objective of this ex-

Table 1: Rater Means and Standard Deviations of the Stimulus Faces

Trait Dimension Descriptor	Dimension Means	Standard Deviation
Adjusted	4.03	0.82
Dominant	4.16	0.85
Social	4.07	0.97
Trustworthy	4.00	0.87
Warm	3.94	1.01

periment was not to do yet another psychological study regarding the person perception of faces. This is a topic that has been well researched, and people of different ages, genders, races, and cultures have been shown to be remarkably consistent in their judgments (Albright et al., 1997; Zebrowitz et al., 1993). Rather, the experimental design was geared solely towards obtaining human judgments of the 220 faces in order to extract those few faces that unambiguously elicit the specific traits explored in this study.

5.2.3 Division of Stimulus Faces into Trait Class Sets

In most face classification tasks, such as classifying faces by identity, gender, and race, the division of faces into relevant classes poses few problems, as the classes are clearly definable. In the classification task of matching human impressions of faces, however, the division of faces into relevant trait classes is not a straightforward process. It is complicated by the fact that many faces fail to elicit strong opinions and by the fact that human beings, while consistent in their ratings, are not in total agreement.

In this study, faces were divided in step 1.C, based on their average rating, into three classes: low (with a mean range of 1.0 - 2.9), neutral (with a mean range of 3.0-4.9) and high (with a mean range of 5.0 - 7.0). As a PCA classification of faces with weak attributions is irrelevant for that trait dimension, that is, the classification is not unambiguous, neutral faces were excluded from the PCA training and testing sets. In addition, faces were pruned from the low and high classes that had a standard deviation greater than 1.5 or that had 50% or more ratings marked neutrally or in the opposite class. Thus, only those few faces that elicited strong impressions were used to develop the PCA models.

Table 2 lists the total number of images selected to form the high and low attribution class sets for each of the five trait dimensions. The total number of images is greater than 220 because some images produced significant trait impressions along more than one dimension.

5.3 Step 2: PCA Modeling

Once a suitable dataset was developed, five separate PCAs, one for each of the five trait dimensions, were trained and evaluated using MATLAB (MathWorks, 2000). Outlined below are the operations involved in training and testing the PCAs for each trait dimension. The reader should refer to Turk and Pentland (1991a) for additional details.

5.3.1 Training

Training a PCA, in step 2.A, requires three operations: 1) randomly dividing the trait class sets into separate training and testing sets; 2) calculating the eigenvectors from the

Table 2: Number of Images Selected for the Trait Attribution Classes

Trait Dimension	Attribution Class	Number
Adjustment	Low	11
	High	12
Dominance	Low	11
	High	14
Sociality	Low	12
	High	14
Trustworthiness	Low	10
	High	15
Warmth	Low	14
	High	14

training set; and 3) calculating the distribution of each class within the face space.

Operation 1. The two attribution classes of images (high and low) for each dimension were merged and divided into a training set of images and a testing set of images, with an equal number of images from both classes (high and low) represented in the training and testing sets.

Operation 2. The eigenvectors were computed using the following algorithm:

1. Reshape the training images into column vectors, which together form the matrix $\mathbf{\Gamma}$. Let $\mathbf{\Gamma}_k$ represent the column vector of face *k*.

2. Normalize the column vector for each face *k* in the training set of *M* images:

$$\mathbf{\Phi_k} = \mathbf{\Gamma}_k - \mathbf{\Psi},\ where\ \mathbf{\Psi} = \frac{1}{M}\sum_{k}^{M}\mathbf{\Gamma}_k \tag{1}$$

3. Compute the eigenfaces using singular value decomposition:

$$\mathbf{\Phi} = \mathbf{USV}^T \tag{2}$$

where **S** is a diagonal matrix whose diagonal elements are the singular values, or *eigenvalues*, of $\mathbf{\Phi}$, $\mathbf{V}^T$ is the transpose of **V**, and **U** and **V** are unary matrices. The columns of **U** are the eigenvectors of $\mathbf{\Phi\Phi}^T$, and are referred to as *eigenfaces*, as they are face-like in appearance. The columns of **V** are the eigenvectors of $\mathbf{\Phi}^T\mathbf{\Phi}$ and are not used in this analysis.

Operation 3. The distribution within the face space for each of the classes was computed by projecting each training image $\mathbf{\Gamma}_k$ onto the eigenfaces as follows:

$$\omega_k = \mathbf{U}_k^T(\mathbf{\Gamma}_k - \mathbf{\Psi}) \qquad (\text{for } k = 1, ..., M) \tag{3}$$

Let $\mathbf{\Omega}^T = [\omega_1, \omega_1, ..., \omega_M]$, be the weight vector that describes the contribution of each eigenvector in representing a face. A representative class vector is obtained by averaging the projected vectors, $\mathbf{\Omega}$, for each training class (Turk and Pentland, 1991b).

5.3.2 Testing

Evaluating the system using the testing set of images in step 2.B required two operations: 1) projecting each test image $\mathbf{\Gamma}_k$ onto the face space to obtain $\mathbf{\Omega}_k$ as in Operation 3 above, and 2) determining the best-fit class membership. Best-fit membership was determined by calculating the smallest Euclidian distance, d, of $\mathbf{\Omega}_k$ from $\mathbf{\Omega}_j$, where $\mathbf{\Omega}_j$ represents the average weight vector of the training images in some class *j*. The number of correct classifications was then averaged and used as an index to evaluate the performance of the system.

5.3.3 Model Evaluation

Because of the small number of images in the trait sets, a cross-validation technique was employed in step 2.C such that only two images from each set were selected to form the testing set, and training and testing were performed as outlined in sections 5.3.1 and 5.3.2. This process was repeated twenty times for each trait dimension. The ratio of right to wrong classifications was used as the classification index, and the twenty classification indexes of each trait were averaged to form the final classification score for that trait.

Table 3: Averaged PCA Classification Scores

Trait Dimension	Classification Rate
Adjustment	.71
Dominance	.64
Sociality	.70
Trustworthiness	.81
Warmth	.89

Table 3 displays the classification scores for the five PCAs. All five PCA classification rates were above chance, with trustworthiness and warmth scoring well above chance. Results were not as good for dominance. See section 7 for a more complete discussion of the results of this study.

6 Synthesizing Faces with Predicted Trait Evaluations

Reported in this section is a preliminary study conducted to demonstrate the possibility of using PCA to construct novel faces with a high probability of eliciting specific trait impressions. As described in section 6.1, certain stimulus faces were projected onto PCAs trained with stimulus faces that ranked in the first study as either high or low in a trait dimension; they were then reconstructed. This process generated novel faces. As described in section 6.2, predictions were made regarding the impressions the synthesized faces would make on human observers. These predictions were then compared with the evaluations of human subjects.

6.1 Face Synthesis

Although composite facial systems such as SpotIt! (Brunelli and Mich, 1996) have utilized the PCA face space to organize facial features in terms of their similarity, little work

has been done in synthesizing faces directly from within the PCA face space.

Two notable exceptions are Vetter and Poggio (1997) and a pilot composite system developed by Hancock (2000). In the latter system, shape-free facial information and shape information are extracted and subjected to PCA. Using a genetic algorithm, novel faces are evolved by recombining the eigenfaces of shape-free facial images and then morphing them to any number of shape components.

One of the benefits in using the database of facial features in FACES to produce the stimulus faces used in these studies is that the features were normalized and aligned to facilitate seamless combinations. Since the entire set of stimulus faces were therefore automatically normalized and aligned (the degree of alignment can be seen in the clarity of the eigenfaces in Figure 1), this exploration into face synthesis used a simpler approach to recombine the eigenfaces: face synthesis was performed by probing the appropriate trait attribution space. With PCA, image projection is onto a low-dimensional space (Turk and Pentland, 1991b). For this reason, even images that look nothing like a face, when projected onto a face space, produce face-like reconstructions. In other words, as illustrated in Figure 8, these non-face images serve as a means of probing the face space since the reconstructions combine characteristics of the faces used to define the PCA face space.

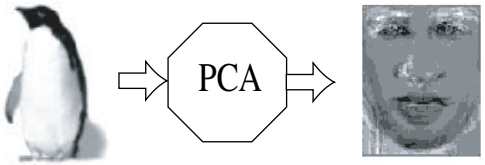


Figure 8: Illustration of Face Space Probing. An Image of a Penguin Projected onto a PCA Face Space Results in a Face-like Reconstruction.

In this study, a subsection of the face space, namely the PCA trait attribution space, was similarly probed. Attribution space probing was accomplished as follows: two PCAs, one for each of the two attribution classes of high and low for each trait dimension, were trained using all images in the appropriate attribution class set. In order to generate novel faces, the PCA attribution spaces needed to be seeded with as many faces as possible. For this reason, all stimulus images in the first experiment with an average rating ≤ 3.0 within each trait dimension were used to train the low PCA attribution spaces, and all stimulus images with an average rating of ≥ 5.0 were used to train the high PCA attribution spaces. See Table 4 for the total number of images used to train the PCAs for each of the eight attribution classes.

Face synthesis was performed by probing the two PCA attribution spaces for each of the five trait dimensions. This was accomplished by taking an image in one attribution class set and projecting and reconstructing it using a PCA trained with the images of the opposite attribution class set.

Figure 9 shows two examples of face synthesis using the cold (low) and the warm (high) PCA attribution spaces. On the right of Figure 9, an image classified as cold (top) and an image classified as warm (bottom) were projected onto the opposite PCA attribution space and reconstructed. This resulted in the new images shown on the left. A total of 340 images (every image in one attribution class was projected onto the opposite

Table 4: Number of Stimulus Faces Used For Face Synthesis.

Trait Dimension Descriptor	Number Rated ≤ 3.0 (Low Attribution Class)	Number Rated ≥ 5.0 (High Attribution Class)
Adjusted	25	26
Dominant	17	37
Social	45	34
Trustworthy	29	34
Warm	51	42

PCA attribution space) were synthesized from the eight PCA attribution spaces using this procedure. As the stimulus faces were closely aligned, few artifacts were introduced in the reconstruction process. Compare, for example, in Figure 9, the artifacts introduced in the synthesized faces (right) to the stimulus faces (left). Although some faces produced more artifacts than others, no attempt was made to enhance the synthesized faces.

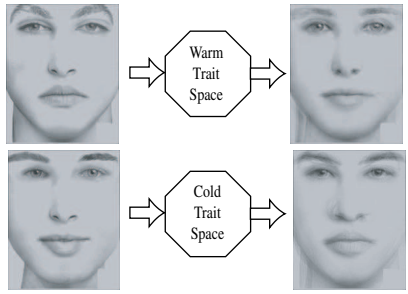


Figure 9: Examples of Faces Synthesized by Probing the PCA Attribution Spaces of Facial Warmth

6.2 Human Assessment of Synthesized Faces

One hundred ten synthesized images were randomly selected and rated by ten human subjects from the same pool of subjects used in the first experiment to assess the stimulus faces. The same procedures were also followed as in the first experiment. It was predicted that faces synthesized by probing the low PCA attribution space of a particular dimension would be ranked by the human subjects at the lower end of that trait dimension's rating scale (i.e., < 3.5) and that faces synthesized by probing the high PCA attribution space of the same dimension would be ranked at the higher end of the rating scale (i.e., > 3.5).

6.3 Results

Table 5 shows the average assessment of the faces synthesized from the two PCA attribution spaces for each of the five trait dimensions. In general, faces synthesized by

probing the low PCA attribution spaces were rated at the lower end of the scale (average low score is 3.11), whereas faces synthesized by probing the high PCA attribution spaces were rated at the higher end of the scale (average high score is 4.82). Only faces synthesized by probing the low PCA attribution space of adjustment failed to be rated as predicted.

Table 5: Averaged Trait Ratings of Synthesized Faces and Standard Deviations.

Trait Dimension	Attribution Space	Average Rating of Synthesized Faces	Standard Deviation
Adjustment	Low	3.86	1.44
	High	5.12	1.38
Dominance	Low	3.34	1.13
	High	5.00	1.30
Sociality	Low	3.17	1.38
	High	5.40	1.53
Trustworthiness	Low	3.21	1.58
	High	4.69	1.42
Warmth	Low	1.98	1.44
	High	3.90	1.35

7 Discussion

Study 1: Modeling the Trait Impressions of Faces Using PCA

The first study modeled the perception of traits in faces using PCA face recognition techniques. Although the results were marginally better than chance in the classification of faces according to the trait of dominance (.64), the PCA classifiers did a fair job matching average high and low human ratings of faces in the traits dimensions of adjustment (.71) and sociality (.70), and a good job matching user ratings of trustworthiness (.81), and warmth (.89).

At present, no hypothesis can be offered to account for the lower dominance classification score. A shortcoming in the first study was the design of the experiment assessing the stimulus faces. Had it been designed to provide a complete statistical analysis of user ratings, such an analysis might have provided some insight into the poorer performance of PCA recognition of high and low facial dominance.

As mentioned in section 5.2.2, extremely high scores were not expected. Unlike the task of classifying faces according to gender, age, and identity, matching human ratings of faces into trait categories is fuzzy. Although there is considerable evidence that people across cultures and age groups are consistent in their ratings of faces, people are not in total agreement. An attempt was made to produce a dataset of faces within each trait attribution class which demonstrated strong consensus ratings, but even so, consensus was not one hundred percent.

Study 2: Face Synthesis Using PCA

To the extent that agents increasingly have simulated faces, it is desirable to have the agents design those faces themselves rather than rely on human designers to do so. The second study explored the possibility of generating novel faces from within the attribution spaces of adjustment, dominance, sociality, trustworthiness, and warmth. It was predicted that faces synthesized by probing the low PCA attribution spaces would be rated at the lower end of the scale and that faces synthesized by probing the high PCA attribution spaces would be rated at the higher end of the scale.

Table 6: Average Ratings of Synthesized Faces and Stimulus Faces.

Trait Dimension	Attribution Space	Average Rating of Stimulus Faces	Average Rating Synthesized Faces
Adjustment	Low	2.65	3.86
	High	5.35	5.12
Dominance	Low	2.67	3.34
	High	5.52	5.00
Sociality	Low	2.64	3.17
	High	5.44	5.40
Trustworthiness	Low	2.60	3.21
	High	5.30	4.69
Warmth	Low	2.60	1.98
	High	5.34	3.90

The average trait ratings of the synthesized faces used to develop the various PCA trait attribution spaces are presented in Tables 5 and 6. Except for the trait dimension of adjustment, human subjects rated the synthesized faces as predicted. Table 6 also presents the average ratings of the stimulus faces used to train the PCAs for the ten attribution classes. From this table, the differences between the average ratings of the synthesized faces and the average ratings of the stimulus faces for each trait dimension can be calculated as 2.06 for warmth, 1.44 for adjustment, 1.19 for dominance, 1.22 for trustworthiness, and 0.57 for sociality. In particular, the faces synthesized from within the high and low sociality attribution spaces closely matched the average ratings of the stimulus faces used to train the PCAs. The largest difference was for warmth and adjustment. Table 7 shows the total average of high and low ratings for both the synthesized faces and the stimulus faces used in training the PCAs. The total difference between the average ratings of the stimulus faces and the average ratings of the synthesized faces in the ten attribution classes is 0.52, nearly half a point in the seven point scale. Clearly the synthesized faces elicited trait impressions that closely matched the trait ratings of stimulus faces used to train the PCAs.

Although the results of the second study indicate that it may be possible to generate faces with a high probability of eliciting specific impressions in users, much more work needs to be done in this area. This was an exploratory study into face synthesis within refined face spaces, and because the stimulus faces were highly processed and aligned, PCA synthesis was limited to recombining *shape-free* facial information within the PCA attribution spaces.

Table 7: Total Average of the Synthesized Faces and the Stimulus Faces

Attribution Class	Total Average of Stimulus Faces	Total Average of Synthesized Faces
Low	2.63	3.11
High	5.39	4.82

8 Conclusion

This paper reports a first attempt at developing a computational model of the trait impressions of the face for embodied agents that accommodates the social perception and social construction of faces. Two studies were presented. In the first study, a standard holistic face recognition technique based on PCA was used to match the human classification of faces at the bipolar rating extremes of the following trait dimensions: adjustment, dominance, warmth, sociality, and trustworthiness. Although results were marginally better than chance in the classification of faces according to the trait of dominance, PCA did a good job matching the average high and low human ratings of faces in the trait dimensions of adjustment, sociality, trustworthiness, and warmth. A second study explored the possibility of synthesizing faces intended to elicit particular trait impressions in observers. Using PCA models, 110 faces were synthesized and assessed by human subjects. The results were promising: the difference between the average ratings of the synthesized faces and the average ratings of the stimulus faces used to train the PCAs was found to be slightly over half a point in a rating scale of seven.

The research reported in this paper makes a number of contributions. It is the first research endeavor that not only suggests that embodied agents learn to design their own *socially intelligent embodiment*, or *smart embodiment*, but also indicates how this might be accomplished. This paper also presents the first computational model of the trait impressions of the face, and is further unique in using face recognition technology to classify social, or cultural, perceptions of faces rather than attributes of faces that are factual, such as identity and gender.

There are a number of directions that offer promising avenues for further exploration, some of which take into consideration limitations in the two studies presented in this paper. Particularly important, for both modeling the trait impressions of the face and for smart face synthesis, is the need to develop a database of faces that exhibit strong human consensus in a comprehensive set of trait categories. The creation of this database could be approached in several ways. Large collections of two-dimensional photographs and three-dimensional scans of actual faces could be evaluated, and those that produce marked attribution effects could be assembled into appropriate trait categories. Datasets of faces could also be generated artificially using either a variety of geometrical transformations, such as the cardioidal transform mentioned in section 2.2, or by simply piecing together facial features, either randomly, as was the case in this project, or with an eye towards eliciting specific trait attributions. These artificially generated faces would also need to be evaluated by human subjects.

Each of these approaches offers some attractive benefits. Two-dimensional photographs and facial composites have been widely studied in the person perception literature and present a simpler approach to modeling faces in terms of the traits they elicit than would be offered by three-dimensional scans. An advantage using facial composite programs,

whether two-dimensional or three-dimensional, is that the contributions of individual facial features in the attribution process could more easily be investigated. Future studies might even investigate the possibility of designing embodied agents that learn to compose faces that are calculated to produce specific impressions in users by manipulating a relatively small set of facial features.

Each of these approaches is also problematical. A danger in using a dataset of faces that have been *artificially* produced is that models developed from these faces might oversimplify the problem too much and not model actual faces. These are criticisms that could also be leveled against many psychological studies that use artificially generated faces. It might be thought that using photographs of actual faces would solve these problems. However, photographs are two-dimensional representations, and it could be argued that people form impressions of faces based on multidimensional views of faces. Three-dimensional scans of actual faces also present representational dilemmas. How faces are seen in space for instance could affect viewer ratings. Will the viewer control how the scans are viewed or will the scan move on their own? Even judging films of faces is problematical as the perspective of the camera is typically artificial and stationary.

As stressed in section 3, a psychologically viable model of the trait attribution process is not essential for embodied agents; rather, the focus should be on selecting a dataset of faces that accommodates the particular tasks and the perception capabilities of the agents. Given the fact that faces, no matter how they are represented, are similar in appearance and, unless highly schematized, produce trait impressions in observers (Brunswik, 1947), it is likely the case that any fairly realistic representation of faces will model the *real faces* the agent will encounter as long as those faces are represented to the agent in the same fashion, e.g., as a set of pixels or geometrical shapes.

In addition to developing appropriate datasets to use in modeling the trait impressions of the face for embodied agents, future research will also need to explore additional face classification techniques. This study used PCA because it is capable of face synthesis as well as face classification. However, other face classification techniques have proven superior to PCA. Two face recognition techniques that should be explored in future studies are independent component analysis (Bartlett, 1998), a generalization of PCA that separates the higher-order moments of the input in addition to the second-order moments, and support vector machines (Vapnik, 1995), learning systems that separate a set of input patterns into two classes with an optimal separating hyperplane. Future studies might also explore classifying faces along a given trait dimension into three classes, i.e., a neutral category as well as the bipolar extremes.

As mentioned in the introduction, one of the most interesting possibilities a model of the trait impressions of the face offers embodied agents is the prospect of designing agents capable of creating an embodiment for themselves that is calculated to produce specific effects on users. Towards this aim, new face synthesis techniques from within these models need to be developed. In this study, trait spaces were probed using images of faces that were perceived to be at the opposite extreme of the trait dimension. Future studies might explore simply perturbing the average weight vector for each trait class. In addition, the synthesized faces in this study were reconstructions of eigenfaces, or *shape-free* image components, a process that introduced artifacts that may have made an impact on impression formation. Future studies in face synthesis will need to consider what Hancock (2000) calls *eigenshapes*, or vectors subjected to PCA that describe the outline of the face and its features. Studies also need to be conducted that appraise the degree of novelty that is exhibited by faces synthesized from within the face class spaces.

Finally, the value and practicality of embedding these models in embodied agents need to be evaluated.

Acknowledgements

Special thanks go to Pierre Cote for granting permission to use the database of facial features found in FACES by Interquest and Micro-Intel.

Appendix: Term Definitions and Behavioral Potential Questions

Table 8 below provides the definitions and behavioral potential questions (some of which were adapted from (Zebrowitz and Montepare, 1992; Berry and Brownlow, 1989)) that were available to subjects filling out the computerized questionnaires (see section 5).

Table 8: Definitions and Behavioral Potential Questions.

Term Definition	Behavioral Potential Questions
Adjusted, Unadjusted, Uncertain Here we are looking at how mentally healthy and adjusted the person is. **Adjusted** Is a person who is fairly happy, mentally healthy, and who feels s/he belongs to society. **Unadjusted** Is a person who is unhappy or discontent, possibly even mentally ill or troubled, and who feels like an outsider.	(None offered)
Dominant, Submissive, Uncertain Here we are looking at how dominating the person is. **Dominant** Is person who is most likely to tell other people what to do. **Submissive** Is a person who usually follows orders, and is not very assertive	A helpful question might be: "Does s/he look like someone who would be the kind of roommate who would comply with most of your wishes about the furniture arrangements, quiet hours, and house rules?"

Table 9: Definitions and Behavioral Potential Questions (Continuted)

Term Definition	Behavioral Potential Questions
Trustworthy, Untrustworthy, Uncertain **Trustworthy** Is a person who is mostly honest and who is not likely to steal, lie, or cheat. **Untrustworthy** Is a person who is often not honest and who possibly steals, lies, and cheats	A helpful question might be: "Does s/he look like someone you would ask to watch your backpack while you made a quick visit to the restroom?"
Social, Unsocial, Uncertain Here we are looking for how social the person is. **Social** Is person who is very outgoing, extroverted, and who enjoys parties and other social activities. **Unsocial** Is a person who introverted, a loner, and who would prefer to stay home rather than go out.	A helpful question might be: "Does s/he look like someone who would attend a school dance or party?"
Warm, Cold, Uncertain Here we are looking for how approachable the person is.	A helpful question might be: "Does s/he look like someone who would turn a cold shoulder to your attempts at friendly conversation?"

References

Albright, L., Malloy, T. E., Dong, Q., Kenny, D. A., and Fang, X. (1997). Cross-cultural consensus in personality judgments. *Journal of Personality and Social Psychology*, **72**(3), 558–69.

Alcock, D., Solano, J., and Kayson, W. A. (1998). How individuals' responses and attractiveness influence aggression. *Psychological Reports*, **82** (June 3/2), 1435–38.

Alley, T. R. (1983). Infantile head shape as an elicitor of adult protection. *Merrill-Palmer Quarterly*, **29**, 411–27.

Alley, T. R. (1988). Social and applied aspects of face perception. In Alley, T. R., editor, *Social and applied aspects of perceiving faces*, pages 1–10. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ

Bartlett, M. S. (1998). *Face image analysis by unsupervised learning and redundancy reduction*. Ph.d. dissertation, University of California.

Berry, D. S. and Brownlow, S. (1989). Were the physiognomists right? *Personality and Social Psychology Bulletin*, **15**(2), 266–79.

Berry, D. S. and McArthur, L. Z. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological Bulletin*, **100**(1), 3–18.

Beymer, D., Shashua, A., and Poggio, T. (1993). Example based image analysis and synthesis. Technical Report A. I. Memo 1431, MIT.

Bruce, V. (1988). *Recognising faces*. Lawrence Erlbaum Associates, Publishers, Hove and London.

Brunelli, R. and Mich, O. (1996). Spotit! an interactive identikit system. *Graphical Models and Image Processing*, **58**(5), 399–404.

Brunelli, R. and Poggio, T. (1993). Face recognition: Features versus templates. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **15**(10), 1042–52.

Brunswik, E. (1947) *Perception and the representative design of psychological experiments* University of California Press, Berkeley and Los Angeles.

Bull, R. and Rumsey, N. (1988). *The social psychology of facial appearance*. Springer-Verlag, New York.

Burton, A. M., Bruce, V., and Dench, N. (1993). What's the difference between men and women? evidence from facial measurement. *Perception*, **22**(2), 153–76.

Converse, J. M., Presser, S. (1986). *Survey Questions: Handcrafting the standardized questionnaire* Sage Publications, London.

Carello, C., Grosofsky, A., Shaw, R. E., Pittenger, J. B., and Mark, L. S. (1989). Attractiveness of facial profiles is a function of distance from archetype. *Ecological Psychology*, **1**(3), 227–51.

Cottrell, G. W. and Fleming, M. K. (1990). Face recognition using unsupervised feature extraction. In *International Conference on Neural Networks*, pages 322–25. Kluwer Academic Publishers, Dordrecht.

Cottrell, G. W. and Metcalfe, J. (1991). Empath: Face, emotion, and gender recognition using holons. In Touretzky, D., editor, *Advances in neural information processing systems*, pages 564–71. Morgan and Kaufman, San Mateo, CA.

Diamantaras, K. I. and Kung, S. Y. (1996). *Principal component neural networks: Theory and applications*. John Wiley and Sons, Inc., New York.

Donath, J. (2001). Mediated faces. In Beynon, M., Nehaniv, C., and Dautenhahn, K., editors, *Cognitive Technology: Instruments of Mind: Proceedings of the 4th International Conference*, *Lecture notes in artificial intelligence*, page 373, Springer, Warwick, United Kingdom.

Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., and Seinowski, T. J. (1999). Classifying facial action. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**(10), 974–89.

Eagly, A. H., Ashmore, R. D., Makhijan, M. G., and Longo, L. C. (1991). What is beautiful is good, but . . .: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, **110**(1), 109–28.

Ekman, P. and Friesen, W. V. (1978). *Facial action coding system*. Consulting Psychologists Press Inc., Palo Alto, CA.

Enlow, D. H. and Hans, M. G. (1996). *Essentials of facial growth*. W. B. Saunders Company, Philadelphia.

Essa, I. A. and Pentland, A. P. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**(7), 757–63.

Feingold, A. (1992). Good-looking people are not what we think. *Psychological Bulletin*, **111**(2), 304–41.

Freierman, S. (2000, Feburary 17). Constructing a real-life mr. potato head. Faces: The ultimate composite picture, *The New York Times*, page 6.

Friedman, H. H. and Amoo, T. (1999). Rating the rating scales. *Journal of Marketing Management*, **9**, 114-23.

Frey, S. (1993). Lavater, lichtenberg, and the suggestive power of the human face. In Shookman, E., editor, *The faces of physiognomy: Interdisciplinary approaches to johann caspar lavater*, *Studies in German literature, linguistics, and Culture*, pages 64–103. Camden House, Columbia, SC.

Galton, F. (1878). Composite portraits. *Nature*, **18**, 97–100.

Hancock, P. (2000). Evolving faces from principal components. *Behavior Research Methods, Instruments and Computers*, **32**(2), 327–33.

Hochberg, J. and Galper, R. E. (1974). Attribution of intention as a function of physiognomy. *Memory and Cognition*, **2**(1A), 39–42.

Jones, D. (1995). Sexual selection, physical attractiveness, and facial neoteny. *Current Anthropology*, **36**(5), 723–48.

Katz, D. (1952). Le portrait composite et la typologie. *Ikon (Revue Internationale de Filmology)*, **2**, 207–14.

Keating, C. F., Mazur, A., and Segall, M. H. (1981a). A cross-cultural exploration of physiognomic traits of dominance and happiness. *Ethology and Sociobiology*, **2**, 41–48.

Keating, C. F., Mazur, A., and Segall, M. H. (1981b). Culture and the perception of social dominance. *Journal of Personality and Social Psychology*, **40**(4), 615–26.

Kohonen, T. (1977). *Associative memory: A system theoretic approach*. Springler-Verlag, Berlin.

Kosugi, M. (1995). Human-face search and location in a scene by multi-pyramid architecture for personal identification. *Systems and Computers in Japan*, **26**(6), 27–38.

Landau, T. (1989). *About faces: The evolution of the human face*. Anchor Books, New York.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., and Smoot, M. (2000). Maxims or myths of beauty? a meta-analytic and theoretical review. *Psychological Bulletin*, **126**(3), 390–423.

Langlois, J. H. and Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, **1**, 115–21.

Langlois, J. H., Roggman, L. A., Musselman, L., and Acton, S. (1991). A picture is worth a thousand words: Reply to 'on the difficulty of averaging faces'. *Psychological Science*, **2**(5), 354–57.

Langlois, J. H., Roggman, L. A., and Rieser-Danner, L. A. (1990). Infants' differential social responses to attractive and unattractive faces. *Developmental Psychology*, **26**(1), 153–59.

Lanitis, A., Taylor, C. J., and Cootes, T. F. (1997). Automatic interpretation and coding of face images using flexible models. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **19**(7), 743–56.

Liggett, J. (1974). *The human face*. Stein and Day, New York.

Lucker, G. W. and Graber, L. W. (1980). Physiognomic features and facial appearance judgements in children. *The Journal of Psychology*, **104**(Second Half), 261–68.

Magli, P. (1989). The face and the soul. In Feher, M., Naddaf, R., and Tazi, N., editors, *Fragments for a history of the human body*, volume 2, pages 86–127. Zone, New York.

Magro, A. M. (1997). Why barbie is perceived as beautiful. *Perceptual and Motor Skills*, **85**(1), 363–74.

Massaro, D. M. (1997). *Perceiving talking faces: From speech perception to a behavioral principle*. MIT Press, Cambridge, MA.

McArthur, L. Z. and Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychology Review*, **90**(3), 215–38.

Mueller, U. and Mazur, A. (1996). Facial dominance of west point cadets as a predictor of later military rank. *Social Forces*, **74**(3), 823–28.

Mulford, M., Orbell, J., Shatto, C., and Stockard, J. (1998). Physical attractiveness, opportunity, and success in everyday exchange. *American Journal of Sociology*, **103**(6), 1565–92.

Oja, E. (1992). Principal components, minor components and linear neural networks. *Neural Networks*, **5**, 927–35.

O'Toole, A. J. and Deffenbacher, K. A. (1997). The perception of face gender: The role of stimulus structure in recognition and classification. *Memory and Cognition*, **26**, 146–60.

Padgett, C. and Cottrell, G. W. (1998). A simple neural network models categorical perception of facial expressions. In *Proceedings of the 20th Annual Cognitive Science Conference*, Lawrence Erlbaum Associates, Publishers, Madison, WI.

Pelachaud, C. and Poggi, I. (2002). Subtleties of facial expressions in embodied agents. *Journal of Visualization and Computer Animation*, **13**(5), 301–12.

Picard, R. W. (1997). *Affective computing*,1st edition. The MIT Press, Cambridge, MA.

Pittenger, J. B. and Shaw, R. G. (1975). Aging faces as viscal-elastic events: Implications for a theory of nonrigid shape perceptions. *Journal of Experimental Psychology: Human Perception and Performance*, **1**, 376.

Rhodes, G. and Tremewan, T. (1996). Averageness exaggeration and facial attractiveness. *Psychological Science*, **7**(2), 105–10.

Rosenberg, S. (1977). New approaches to the analysis of personal constructs in person perception. In Land, A. L. and Cole, J. K., editors, *Nebraska symposium on motivation*, volume 24, pages 179–242. University of Nebraska Press, Lincoln.

Rubenstein, A. J., Kalakanis, L., and Langlois, J. H. (1999). Infant preferences for attractive faces: A cognitive explanation. *Developmental Psychology*, **35**(3), 848–55.

Secord, P. F., Dukes, W. F., and Bevan, W. W. (1954). An experiment in social perceiving. *Genetic Psychology Monographs*, **49**(Second Half), 231–79.

Sirovich, L. and Kirby, M. (1987). Low dimensional procedure for the characterization of human faces. *Journal of Optical Society of America*, **4**(3), 519–24.

Sproull, L., Subramani, R., Kiesler, S., Walker, J. H., and Waters, K. (1996). When the interface is a face. *Human Computer Interaction*, **11**, 97–124.

Symons, D. (1979). *The evolution of human sexuality*. Oxford University Press, New York.

The MathWorks. (2000). *Using MATLAB: The language of technical computing*. The Mathworks, Inc., Natick, MA.

Todd, J. T. and Mark, L. S. (1980). The perception of human growth. *Scientific American*, **242**(2), 132–44.

Todd, J. T., Mark, L. S., Shaw, R. E., and Pittenger, J. B. (1981). Issues related to the prediction of craniofacial growth. *American Journal of Orthodontics*, **79**, 63–80.

Turk, M. A. and Pentland, A. P. (1991a). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, **3**(1), 71–86.

Turk, M. A. and Pentland, A. P. (1991b). Face recognition using eigenfaces. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–91, IEEE Computer Society Press, Silver Spring, MD.

Valentin, D., Abdi, H., Edelman, B. E., and O'Toole, A. J. (1997). Principal component and neural network analyses of face images: What can be generalized in gender classification? *Journal of Mathematical Psychology*, **41**(4), 398–413.

Valentin, D., Abdi, H., and O'Toole, A. J. (1994a). Categorization and identification of human face images by neural networks: A review of the linear autoassociative and principal component approaches. *Journal of Biological Systems*, **2**(3), 413–29.

Valentin, D., Abdi, H., O'Toole, A. J., and Cottrell, G. W. (1994b). Connectionist models of face processing: A survey. *Pattern Recognition*, **27**(9), 1209–30.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Vetter, T. and Poggio, T. (1997). Linear object classes and image synthesis from a single example image. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **19**, 733–42.

Waters, K. and Terzopoulos, D. (1992). The computer synthesis of expressive faces. *Philosophical Transactions of the Royal Society of London*, **B335**, 87–93.

Wegener-Knudsen, M., Martin, J.-C., Dybkjær, L., Berman, S., Bernsen, N. O., Choukri, K., Heid, U., Kita, S., Mapelli, V., Pelachaud, C., Poggi, I., van Elswijk, G., and Wittenburg, P. (2002). Survey of nimm data resources, current and future user profiles, markets and user needs for nimm resources. sle natural interactivity and multimodality. Technical Report Deliverable D8.1, http://isle.nis.sdu.dk/reports/wp8/.

Yamaguchi, M. K., Hirukawa, T., and Kanazawa, S. (1995). Judgment of gender through facial parts. *Perception*, **24**(5), 563–75.

Zebrowitz, L. A. (1998). *Reading faces: Window to the soul?* Westview Press, Boulder, CO.

Zebrowitz, L. A. and Montepare, J. M. (1992). Impressions of babyfaced individuals across the life span. *Developmental Psychology*, **28**(6), 1143–52.

Zebrowitz, L. A., Montepare, J. M., and Lee, H. K. (1993). They don't all look alike: Individuated impressions of other racial groups. *Journal of Personality and Social Psychology*, **65**(1), 85–101.

AISB Journal **1**(6), © SSAISB, 2005
http://www.aisb.org.uk

Modelling Vocabulary Acquisition: An Explanation of the Link between the Phonological Loop and Long-Term Memory

Gary Jones*, Fernand Gobet† and Julian M. Pine‡

* Centre for Psychological Research in Human Behaviour, University of Derby, Western Road, Mickleover, Derby DE3 9GX, *g.jones@derby.ac.uk*
† School of Social Sciences and Law, Brunel University, Uxbridge, Middlesex, UB8 3PH, *fernand.gobet@brunel.ac.uk*
‡ School of Psychology, University of Liverpool, Liverpool, L69 7ZA, *julian.pine@liverpool.ac.uk*

Abstract

The acquisition of vocabulary represents a key phenomenon in language acquisition, but is still poorly understood. Recently, the working memory model (Baddeley & Hitch, 1974) has been adapted to account for vocabulary acquisition (e.g., Gathercole & Baddeley, 1989). It is claimed that the phonological store, one of the components of working memory, offers a critical mechanism for learning new words. However, one of the theoretical weaknesses of this approach is that no account is given for the mechanisms and representations used in long-term memory learning. This paper presents a computer model combining the EPAM/chunking approach (Feigenbaum & Simon, 1984) with the working memory approach. Phonemic learning is simulated as the elaboration of a discrimination net. Naturalistic input, consisting of utterances from nine mothers interacting with their child, is used during the learning phase. Simulations show that the model can account reasonably well for the nonword repetition task described by Gathercole and Baddeley (1989), a task often presented as a powerful diagnostic of vocabulary learning.

1 Introduction

Children are remarkably adept at learning new verbal information. After an initial slow period from about 12 to 16 months when most children learn around 40 words, the learning rate increases such that in the next four months children will have learnt 130 more new words (Bates et al., 1994). By the beginning of their school years children learn up to 3,000 words per year (Nagy & Herman, 1987).

A major part of learning new words is learning the novel sequences of sounds that represent the word. However, it is difficult to directly examine the processes involved in learning the sound patterns of new words because it is impossible to be certain that the new sound pattern has never been encountered before. The use of nonwords which conform to the phonotactic rules of English provides a good test of vocabulary learning because it ensures that the (non)word to be learned is novel.

1.1 The Nonword Repetition Test

The nonword repetition (NWR) test (Gathercole, Willis, Baddeley & Emslie, 1994) was designed to investigate the role of phonological memory in word learning. The test involves the experimenter speaking a nonword and the child attempting to repeat it. Gathercole and Baddeley (1989) found that, compared to the other measures they used, the NWR test was the best predictor of vocabulary size, even after vocabulary scores (as calculated by the British Picture Vocabulary Scale, Dunn & Dunn, 1982) were partialled out of correlations.

Gathercole and Baddeley (1990) used the NWR test to categorise children into two groups, those with low NWR scores, and those with high NWR scores. Children in the high NWR group were better at learning nonword labels than children in the low NWR group. Gathercole, Willis, Emslie, and Baddeley (1991) found better NWR performance on nonwords that were rated high in wordlikeness than nonwords rated low in wordlikeness. These NWR studies show the influence that vocabulary knowledge has upon the learning of new words.

The nonword repetition test involves two sets of nonwords, one set having single consonants (e.g. **r**u**b**i**d**) and one set having clustered consonants (e.g. **gl**i**st**ow). There are twenty nonwords in each set, divided into four groups of five based on the number of syllables in the nonword (one to four). Several studies using these types of nonwords have consistently found that repetition accuracy decreases as the number of syllables in the nonword increases, excepting one-syllable nonwords (e.g., Gathercole & Adams, 1993; Gathercole, Willis, Emslie & Baddeley, 1991), and that accuracy is worse for clustered consonant nonwords.

Based on these findings, NWR ability would seem to provide a good test of phonological memory and is a good indicator of vocabulary size. The NWR findings can be explained within the theoretical framework of the working memory model.

1.2 The Phonological Loop Explanation of NWR Findings

The working memory model (Baddeley & Hitch, 1974) has recently been adapted to account for vocabulary acquisition (e.g., Gathercole & Baddeley, 1993). It is claimed that the phonological loop part of the model is a critical mechanism for learning new words. The phonological loop has two linked components: the phonological short-term store, and the sub-vocal rehearsal mechanism. Items in the store decay over time (around 2,000 ms, Baddeley, Thomson & Buchanan, 1975). The sub-vocal rehearsal mechanism (involving

sub-vocal articulation in real-time) can refresh items in the store in a serial, time-based manner (Gathercole & Martin, 1996). The store is linked to the central executive part of the model, which provides a link to long-term memory (LTM) (Gathercole & Baddeley, 1993). The influence of LTM is acknowledged (e.g., Gathercole, Willis, Baddeley & Emslie, 1994), but the nature and definition of the link is yet to be defined.

The phonological loop hypothesis is able to explain the basic NWR findings involving nonword length because of the decay that takes place in the phonological store (items remain in the store for 2,000 ms unless refreshed). Longer nonwords take longer to rehearse and so their representations in the phonological store are not refreshed as often as shorter nonwords. Repetition ability will therefore decrease for longer nonwords. The poorer repetition performance for clustered consonant nonwords can be explained in a similar way (because although the clustered consonant nonwords contain the same number of syllables as the single consonant nonwords, they contain more phonemes).

Differences in NWR ability between children of the same age were originally attributed to differences in rates of subvocal rehearsal (Gathercole & Baddeley, 1993). However, recent findings show that children do not use subvocal rehearsal until around seven years of age (see Cowan & Kail, 1996, for a review). This has led to the phonological store being assumed to be the primary language learning device, with differences in language learning across children of the same age being explained by the quality of the phonological representation of just-spoken items (Baddeley, Gathercole & Papagno, 1998). The lack of a rehearsal process for children below seven years of age does not appear to hinder the model. Brown and Hulme (1996) show that NWR phenomena can be explained solely by a decay based model (which the phonological store now becomes for children under seven years of age). This model will be discussed later.

The phonological loop is able to explain a lot of the vocabulary acquisition phenomena using a very simple mechanism. However, it fails in two critical areas: there is no explanation of how words are learned, and there is no explanation of how the loop interacts with LTM. Speculative explanations have been given as to how the loop may interact with LTM (e.g., Baddeley, Gathercole & Papagno, 1998; Gathercole & Baddeley, 1989). Gupta and MacWhinney (1997) have proposed a formal specification but their model has not yet been implemented computationally.

The phonological loop hypothesis also lacks precision because it is part of a verbal theory. For example, there is no definition of how rehearsal rate changes based on the length of the sound pattern being rehearsed (except to say that long strings are rehearsed slower than short strings). Implementing the loop within a computational architecture forces precision because the theory is implemented as a running computer program. Several computational implementations of the phonological loop exist.

1.3 Existing Computational Implementations of the Phonological Loop

Burgess and Hitch (1992) detail a connectionist network which was primarily intended to model serial order effects (the recall of a set of items in the correct presentation order). The decay in the phonological store is represented by decay on the weights between layer nodes (decay is proportional to the number of phonemes to be output). Rehearsal is synonymous with articulation: the most active word in the network is selected for output once the phonemic input has been processed. The model can explain word length and articulatory suppression effects, but does not explain any of the NWR findings. In addition, no phoneme or word learning takes place; the model provides no theory as to how phonemes and words are created in LTM.

Brown and Hulme (1996) give an account of a trace-based decay model which bears

resemblance to the phonological store (the model intentionally has no rehearsal process). Time is represented in 0.1-sec slices; input items are split into segments such that each segment corresponds to a time slice. Longer words therefore take up more segments and so occupy more time slices. As each segment of an input item enters the store, it is given a fixed initial strength, which decays over time. This means that the early *segments* of an input item suffer more decay than the later segments, as well as earlier input *items* suffering more decay than later input items.

The probability of recalling an item is the product of the current strength of each of the segments of the item. The probability is increased to reflect the influence of LTM during recall; wordlike nonwords would therefore have their probability increased more than non-wordlike nonwords. Using these mechanisms, the model can account for recall effects for different lengths of both words and nonwords (Brown & Hulme, 1996). However, the model does not account for any learning processes.

There are problems with each of the existing phonological loop models. For the phonological loop to successfully provide an account of vocabulary learning, a precise specification of its interaction with LTM is required. EPAM is a computational modelling architecture which is able to provide such a specification.

2 Implementing the Phonological Loop within the EPAM Architecture

EPAM is a computational modelling approach whereby a discrimination network is built based on the input that the model receives. In terms of Artificial Intelligence (AI), the approach can be seen as being very similar to tries (Fredkin, 1960), and particularly the "suffix" trie whereby input is presented to the trie as a whole and then as every part of the suffix (e.g., W H A T, then H A T, then A T, then T). In much the same way as tries, a hierarchy of the input is built, as will be shown later. Tries have commonly been used in AI to represent dictionaries (e.g., Arslan & Egecioglu, 2004), but have also been used in other domains such as matching for similarity in video databases (e.g., Park & Hyun, 2004). Discrimination networks have also been used in AI (e.g., in expert systems, Gerevini et al., 1992), although the main area for the EPAM form of discrimination networks has been in simulating various areas of human cognition, such as learning, memory, and perception in chess (De Groot & Gobet, 1996; Gobet, 1993; Gobet & Simon, 2000; Simon & Gilmartin, 1973), verbal learning behaviour (Feigenbaum & Simon, 1984), the digit-span task (Richman, Staszewski & Simon, 1995), the context effect in letter perception (Richman & Simon, 1989), and the acquisition of syntactic categories (Freudenthal, Pine & Gobet, 2005; Gobet, Freudenthal & Pine, 2004; Jones, Gobet & Pine, 2000) (see Gobet et al., 2001, for an overview). EPAM provides a modelling environment which is well suited for describing how sound patterns can be learnt. When a sentence is heard, it is heard as a sound pattern in the form of a sequence of phonemes. This sequence of phonemes needs to be processed and stored in a hierarchical fashion (to illustrate the order of the sound patterns).

EPAM provides a simple mechanism by which this goal can be accomplished. Furthermore, there would seem to be an easy method by which the sound patterns in LTM (i.e., the resulting discrimination network) can be linked to the phonological store - when sound patterns come in, they can be matched to those that exist in LTM and thus any sequence of sound patterns that match do not have to be stored individually in the phonological store - a link can be placed there to the relevant item in the discrimination network.

Precisely how EPAM will be linked to the phonological store will be explained later.

2.1 The EPAM Architecture

EPAM learns by building a discrimination network. The discrimination network is a hierarchical representation of the input and consists of nodes connected to one another by links. Nodes contain information and links between nodes contain tests which must be fulfilled before they can be traversed. For the purposes of modelling the learning of sound patterns, EPAM has been simplified and is henceforth referred to as EPAM-VOC.

When an input (e.g., a sequence of phonemes) is given to the network, EPAM-VOC traverses down the hierarchy as far as possible. This is done by starting at the top node (the root node) and selecting the first link whose test is fulfilled by the first part of the input. The node at the end of the link now becomes the top node and the rest of the input is applied to all the links below this node. When a node is reached where no further traversing can be done (e.g., the input fulfils none of the tests of the nodes links, or the node is a leaf node), EPAM-VOC compares the information at the node with the input information. Learning now occurs in two ways.

1. *Discrimination*. When the input information mismatches the information given at the node, a new link (i.e., test) and node are added to the tree below the node that has just been reached. The new test will relate to the mismatched part of the input.

2. *Familiarisation*. When the input information is under-represented by the information at the node (e.g., features from the input are not present in the information at the node), new features (from the input) are added to the information in the node. In EPAM-VOC, the *image* of a node will always consist of all the information in the links that lead to the node.

Discrimination therefore creates nodes and links, and familiarisation creates or modifies the information contained in nodes. Examples of the discrimination and familiarisation learning mechanisms will be given later.

2.2 Learning Sound Patterns in EPAM-VOC

EPAM-VOC provides an efficient method for representing items in LTM. The basic idea is to give as input to the model the utterances from mothers speech so that EPAM-VOC can learn phonemes and combinations of phonemes. Mothers' utterances will be converted into a sequence of phonemes before being used as input. This will be done using the CMU Lexicon database (available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict) which cross-references words with their phonemic representations. The use of phonemic input assumes that some form of phonemic feature primitives already exist to distinguish one phoneme from another.

EPAM-VOC will begin with a null root node. When it receives an input (a sequence of phonemes), new nodes and links will be created. At first, most of the new nodes and links will just be for single phonemes, as EPAM-VOC learns to master individual phonemes. As learning progresses, the information at nodes will become sequences of phonemes and therefore segments of speech (e.g., specific words) rather than just individual sounds (i.e., phonemes). To accomplish this, the EPAM learning mechanism is altered in two ways. First, before a sequence of phonemes can be learnt, the individual phonemes in the sequence must have been learnt. Second, when individual phonemes are learnt, they are

linked to the root node (in this way all sequences of phonemes are below the node which represents the initial phoneme in the sequence).

Let us consider an example of the network learning the utterance ”What?”. Using the CMU Lexicon database, this utterance is converted to the phonemic representation ”W AH1 T” (all of the phonemes used in the database map onto the standard phoneme set for American English). Note that the phonemic input to the model does not specify gaps between words, but does specify the stress of particular phonemes (0=unstressed; 1=primary stress; 2=secondary stress).

The first part of the input (”W”) is applied to all of the root nodes' links in the network. If the network is empty, there will be no links. At this point EPAM-VOC must discriminate because the information ”W” mismatches the information at the root node (the root node information is null). The discrimination process creates a new node, and a link from the root node to the new node with the test ”W”. EPAM-VOC must then familiarise itself with the input, in order to create the ”W” information in the image of the node. EPAM-VOC then moves on to the remainder of the input (i.e., ”AH1 T”) much like a suffix trie. In a similar way as for ”W”, the phoneme ”AH1” will be learnt. EPAM-VOC then moves on to the remainder of the input (i.e., ”T”), and in a similar fashion, learns the phoneme ”T”. Thus when the input is received the first time, the individual phonemes ”W”, ”AH1” and ”T” are learnt.

When encountering the input for the second time, the link ”W” can be taken, and the input can move to the next phoneme, ”AH1”. As node ”W” does not have any links, discrimination occurs below the ”W” node, creating a new node below the ”W” with a link of ”AH1”. Familiarisation then fills this node with the contents ”W AH1”. The remainder of the input (i.e., ”T”) is then examined, but as this has already been learnt, the processing of the input terminates.

The third time the input is received, the ”W” link can be taken, with the input moving on to ”AH1”. As there is an ”AH1” link below the ”W” node, this link can be taken, and the input can move on to the ”T”. As there is no ”T” link below the ”W AH1” node, discrimination occurs. A new node is created below the ”W AH1” node with the link ”T”. Familiarisation will fill in the contents of the new node with ”W AH1 T”. Thus after three presentations of the input, the network is as shown in Figure 1. The simple example serves to illustrate how EPAM-VOC works; in the actual learning phase each utterance line is only used once, encouraging a diverse network of nodes to be built. Note that EPAM-VOC needs to know individual phonemes before they can be learnt as part of a sequence of phonemes. For example, should the network in Figure 1 see the utterance ”Which?” (”W IH1 CH”), it will traverse down the ”W” link, and move on to the next part of the input (i.e., ”IH1 CH”). However, the network does not know the phoneme ”IH1”, and so it needs to discriminate at the root node, learning the individual phoneme ”IH1” before moving on to the remainder of the input ”CH” (and learning this as an individual phoneme also).

2.3 Implementing the Phonological Loop and Linking it to Long-Term Memory

The model now requires a specification of the phonological loop and a method by which the loop interacts with LTM. The findings relating to the NWR test (the standard test of the phonological loop) were all carried out on children below the age of six. As children below the age of seven are believed not to rehearse, the rehearsal part of the loop should not be used to simulate the NWR findings reported above.

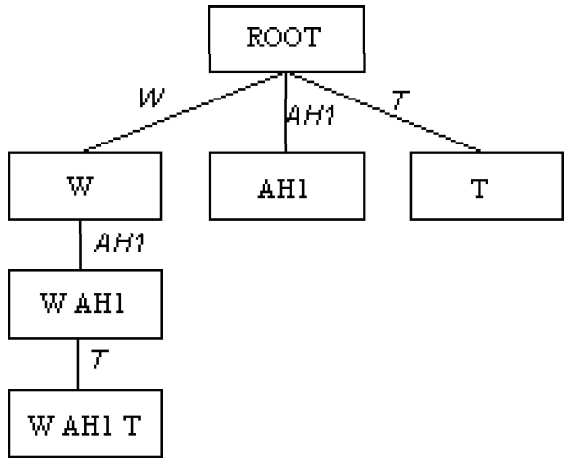


Figure 1: Structure of an EPAM-VOC net after receiving the input ”W AH1 T” three times.

The storage part of the phonological loop is a decay based store which allows items to remain in the store for 2,000 ms. The model will have a time-limited store which will allow 2,000 ms of input (i.e., consistent with the phonological loop estimates). The input will be cut-off as soon as the time limit is reached, because there is no rehearsal to refresh the input representations.

The cumulative time required by the input provides a theory of how the amount of information in the phonological store is mediated by LTM. When an input is heard, LTM (the EPAM-VOC network) is accessed and the input is recoded using the minimum number of nodes possible. Rather than the actual input being placed in the phonological store, the nodes which capture the input are used. The length of time taken to represent the input is therefore calculated on the number of nodes that are required to represent the input. The time allocations are based on values from Zhang and Simon (1985), who estimate 400 ms to match each node, and 84 ms to match each syllable in a node except the first (which takes 0 ms). As the input will be in terms of phonemes, with approximately 2.8 phonemes per syllable (based on estimates from the nonwords in the NWR test), the time to match each phoneme in a node is 30 ms.

Using the example input ”What about that?” (”W AH1 T AH0 B AW1 T DH AE1 T”) and the network as given in Figure 1, the actual input to the model will be ”W AH1 T AH0 B AW1” because this is all that can be represented in the phonological store within the 2,000 ms timescale. The ”W AH1 T” part of the input is represented by a single node, and is allocated a time of 460 ms. Most of the other phonemes are not known to the model and are therefore assumed to take the same time as a full node (400 ms) (the time allocated to each phoneme is assumed to be constant). This means that only three more phonemes can be represented within the phonological store (the actual input to the model having a time allocation of 1,660 ms). When the EPAM-VOC network is small, only a small amount of the input information can be represented in the store, and so new nodes will not contain much information. When the EPAM-VOC network is large, a lot of the input information can be represented in the store and so the model can create new nodes which contain large amounts of information.

3 Simulating the Nonword Repetition Results

There are two main sets of results for the NWR test. One set was tested on children of four and five years of age (Gathercole & Baddeley, 1989; these results were reported in the introduction). The problem with this study is that the children are of an age where they already have a reasonably large vocabulary size. A second set of NWR results is reported by Gathercole and Adams (1993), who used a simpler version of the test on children of two and three years of age. They found the modified test still allowed phonological memory skills to be reliably tested.

The EPAM-VOC model is able to learn sequences of sounds from inputs that are strings of phonemes (converted from speech utterances). It should therefore be capable of simulating both sets of NWR data (2-3 year olds and 4-5 year olds) by modifying the input that is given to the model to reflect the type of input that will be received by these age groups.

In the simulations that will be presented, the model normally learns something for every input it receives. The EPAM parameter of 8 s to learn a single node was dispensed with because of the long time scale involved in the simulations. Given the same input to the model, there should be no significant difference to the results whatever the time taken to learn a node.

The NWR test for the model will be performed by presenting each nonword to the model (as a string of phonemes), and seeing if the components of the nonword can be accessed within the same time limitations that were used for the input (see above). By definition, the information at one node will not be able to represent all of a nonword (because the nonword will never have been received as input, and the presentation time is assumed to be too short to build a new LTM chunk). The information from several different nodes will be required to represent the nonword. If the number of nodes and the phonemes in each node can fit into the time limit, the nonword is repeated accurately, otherwise the nonword is repeated incorrectly. The models' NWR test does not involve articulation because the current EPAM-VOC model does not include a theory of articulation.

3.1 Simulation of Two to Three Year Old Children

For the simulation of the NWR test for children of 2-3 years of age, naturalistic input was used for the model. The input consisted of the mother utterances from nine mothers interacting with their 2-3 year old children, taken from Theakston, Lieven, Pine and Rowland (2000). The average number of utterances for each mother was 25,711 (range 17,474-33,452). The duration of the phonological store was changed from 2,000 ms to 1,750 ms, because there is a high probability that the phonological store of very young children has less duration than adults (existing timing estimates for the phonological store have been based on studies involving adults).

The model was run once for each of the mother's input, resulting in nine different simulations. The NWR test for the model consisted of presenting each nonword as input to the model and seeing if it could represent the nonword within the 1,750 ms time capacity. Note that Gathercole and Adams used a simplified version of the NWR test (using 1-3 syllable nonwords and not distinguishing between single and clustered consonants). They also performed a word repetition test. The results of the children and the model (after seeing 30 percent of the mother's utterances as input) for both the nonword and word repetition tests are shown in Figures 2 and 3.

The model performs at ceiling for the one and two-syllable words and nonwords. A minimum of four phonemes can fit into the 1,750 ms time limit in the phonological store

(one phoneme uses up 400 ms at most). In the words and nonwords used, the average number of phonemes for one-syllable items is 3.2 and for two-syllable items is 5.0. The model has therefore chunked at least one pair of phonemes contained in each of the words and nonwords used in the tests. The children do not perform at ceiling for any of the conditions. Nevertheless, the model still provides a significant correlation with the child data ($r(4)=0.859$, $p<.05$).

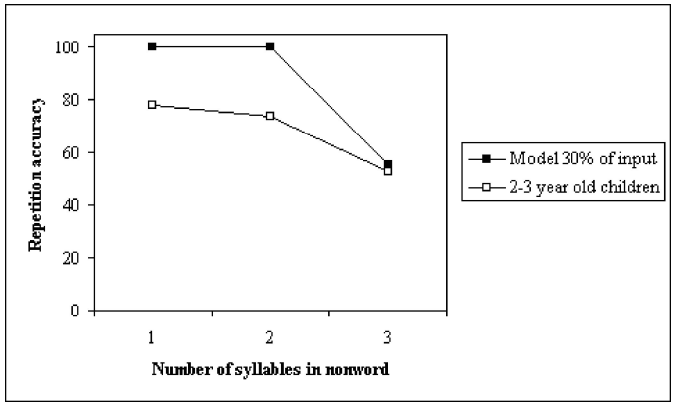


Figure 2: NWR accuracy for 2-3 year old children and the model.

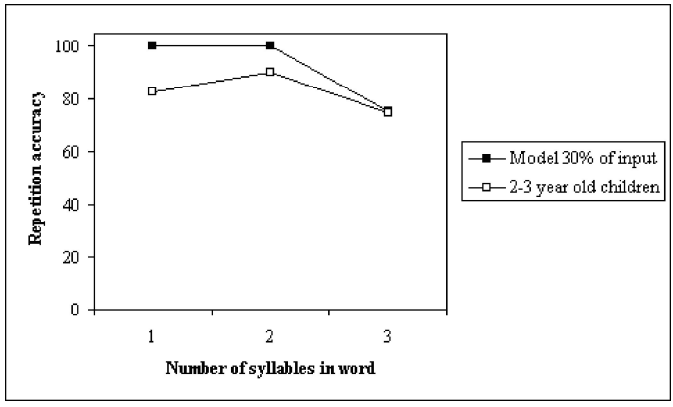


Figure 3: Word repetition accuracy for 2-3 year old children and the model.

The children may not perform at ceiling for one and two-syllable items because of noise during either recognition or articulation of the item. For example, simple nonwords such as *nate* may be perceived by the child to be a real word (e.g., *mate*) and therefore mis-articulated (similarly real words such as *hate* could also be mis-articulated). The studies by Gathercole and colleagues do not perform any analysis of error types so it is difficult to ascertain why children perform poorly for short words and nonwords. However, the model supports the hypothesis of mis-articulation on two counts. First, in performing at ceiling for one and two-syllable items, the model suggests that all 2-3 year old children should be capable of repeating back one and two-syllable words and nonwords. Second, the model provides a very good match for the three-syllable words and nonwords, and

items of this length are very unlikely to be mis-attributed to being other words (because they will share relatively few characteristics with other words of the same length, unlike shorter items).

The results show that the EPAM-VOC model can produce repetition results which are comparable to young children using a simple learning mechanism and naturalistic input. In particular, repetition performance for three syllable items is closely matched by the model. The simulation also raises questions about children's performance on repetition tests for one and two-syllable items.

3.2 Simulation of Four to Five Year Old Children

Carrying out a simulation for each set of mother's utterances is not expected to provide a representative sample of the input that 4 and 5 year olds receive, because by this age the children are beginning school, and beginning to read. Each of the nine mother's utterances were therefore matched to a random selection of words from the CMU Lexicon database on a 50/50 basis for use as input. However, in order to simulate the difference between 4 and 5 year old children in terms of the amount of language they will have heard, 60 percent of the input and 80 percent of the input was seen by the model respectively. For example, for Anne, there were 31,393 mother's utterances. A random sample of half of these (15,696 utterances) were taken together with 15,696 random lexicon words. The simulation of 4 year olds used 60 percent of this resulting file and the simulation of 5 year olds used 80 percent of it. Nine such files were created (one for each mother), resulting in nine simulations. Each simulation presented the model with an equal amount of mother's utterances and lexicon words. The phonological store capacity was reverted back to 2,000 ms based on the assumption that the phonological store reaches full capacity by 4 years of age. Figures 4 and 5 show the comparisons of the results of the simulations and the children, for single and clustered nonwords.

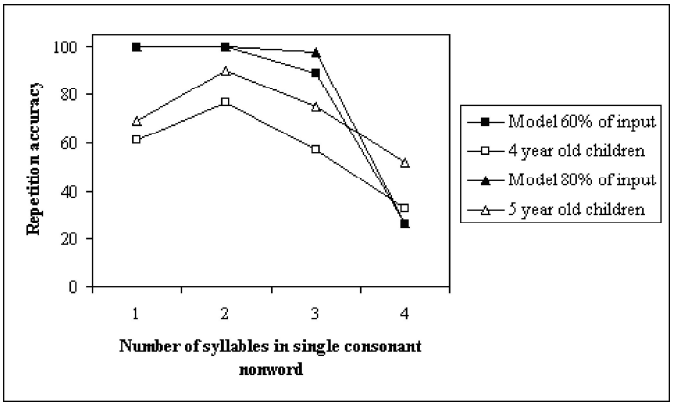


Figure 4: Single consonant NWR accuracy for 4 and 5 year old children, and the model.

The children's performance for one-syllable nonwords is actually worse than for two-syllable nonwords (which should be more difficult). The poor performance for children's repetition of one-syllable nonwords may be due to the acoustic characteristics of their monosyllabic stimuli (Gathercole & Baddeley, 1989), which is consistent with the mis-articulation hypothesis suggested earlier.

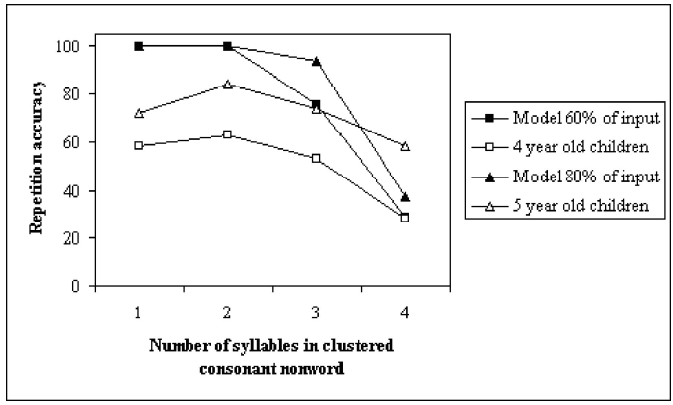


Figure 5: Clustered consonant NWR accuracy for 4 and 5 year old children, and the model.

Although at first glance the fit between the model and the children may not look substantial, there are in fact significant correlations between the 60 percent model and 4 year olds ($r(6)=0.932$, $p<.01$) and between the 80 percent model and 5 year olds ($r(6)=0.847$, $p<.01$).

The model again performs at ceiling for one and two-syllable nonwords. However, the three and four-syllable performance by the model is interesting. The model over-performs for three-syllable nonwords and has a tendency to under-perform for four-syllable nonwords. The three-syllable nonwords average 7.4 phonemes whereas the four-syllable nonwords average 10.1 phonemes. In order to repeat four-syllable nonwords correctly, the model has to chunk up large groups of phonemes that make up the nonword, whereas relatively few have to be chunked to correctly repeat three-syllable nonwords. Building up large chunks in the model is very dependent on the variety of the input that the model sees - the more varied the input, the more rich the chunks in the model. Under-performing on four-syllable nonwords therefore suggests a lack of variation in the input. *This highlights the role of the input as a mediating factor in repetition performance.* The problem for the model, given that the variation of the input is critical, is in determining the type of input that a 4 or 5 year old child will have heard. Clearly, this is an impossible task and any attempt to replicate the input will be a crude approximation. For example, even though the model receives half of the mother's utterances as input, this only constitutes 3,046 different words on average. The lexicon words are used as an attempt to bolster this amount, but clearly they fail to replicate the diversity of input that 4 and 5 year old children receive. The model thus provides a reasonable approximation of repetition performance based on what would seem to be a reasonable, but not perfect, approximation of the input. The results suggest that a more realistic input would produce a good match to the data.

4 Discussion

The simulations have shown that the EPAM-VOC model is able to approximate the NWR performance of both 2-3 and 4-5 year old children. The model accomplishes this by using a combination of a simple learning mechanism, naturalistic input, and a simple implementation of the phonological loop. This represents a parsimonious approach to

learning novel sound patterns. In particular, the model is able to give a specific account of how existing vocabulary knowledge influences the learning of new sound sequences.

The EPAM learning mechanisms are very sensitive to the input that the model receives. This allows the model to make very precise predictions. For example, the sound patterns of the most frequent words in the input will be learnt first. New words which consist of frequent sound patterns will be learned quicker than new words which consist of infrequent sound patterns. The model also allows comparisons of how different approaches to memory can affect learning. A time based store can be compared to a chunk based store (e.g., Miller, 1956) and the effects on learning can be examined.

One aspect of the model that does not correspond to children's vocabulary learning is that the model merely learns sequences of sound patterns - it does not learn words per se. Although the resulting discrimination network (after training) will include a lot of vocabulary, there will also be sound patterns that do not correspond to actual words (for example "W AH1 T AH0" from the beginning of "What about that"). The child must therefore process the input in a more discerning way than the model does, in order to determine word boundaries. This process of "segmentation" is very important and has attracted a great deal of interest in its own right (Brent & Cartwright, 1996; Kazakov & Manandhar, 2001; Perruchet & Vintner, 1998) (see Jusczyk, 1999, for a review). Clearly the model presented here represents first steps in the computational modelling of vocabulary learning, with the next steps involving how to incorporate the processes of segmentation.

This work represents a new modelling research program which aims to examine the extent to which the linguistic input a child receives can account for the child's vocabulary development. While the detail of the simulations could be improved, an important contribution of this paper is to provide mechanisms showing how the phonological store links to LTM. The phonological store was shown to mediate LTM learning by limiting the amount of phonemes that could be learnt in LTM. In turn, LTM mediated how much information could be represented in the phonological store by chunking phonemes such that more information could be stored over time. In addition, the model, which learns from naturalistic input, has been used to explain a variety of other phenomena using very similar mechanisms to those employed here. The use of the same computational approach in various domains such as vocabulary learning, the acquisition of expertise, verbal learning, and the acquisition of syntactic categories, ensure a model that has few degrees of freedom.

5 Acknowledgements

The authors would like to thank two anonymous reviewers for their helpful comments during the preparation of this paper.

6 References

Arslan, A. N., & Egecioglu, O. (2004). Dictionary look-up within small edit distance. *International Journal of Foundations of Computer Science, 15,* 57-71.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158-173.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation*, (Vol. 8), 47-90. New York: Academic Press.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14,* 575-589.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J., & Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language, 21,* 85-123.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61,* 93-125.

Brown, G. D. A., & Hulme, C. (1996). Nonword repetition, STM, and word age-of-acquisition: A computational model. In S. E. Gathercole (Ed.), *Models of short-term memory,* 129-148. Hove, UK: Psychology Press.

Burgess, N., & Hitch, G. J. (1992). Toward a network model of the articulatory loop. *Journal of Memory and Language, 31,* 429-460.

Cowan, N., & Kail, R. (1996). Covert processes and their development in short-term memory. In S. E. Gathercole (Ed.), *Models of short-term memory,* 29-50. Hove, UK: Psychology Press.

De Groot, A. D., & Gobet, F. (1996). *Perception and memory in chess: Studies in the heuristics of the professional eye.* Assen: Van Gorcum.

Dunn, L. M., & Dunn, L. M. (1982). *British Picture Vocabulary Scale.* Windsor: NFER-Nelson.

Feigenbaum, E. A., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science, 8,* 305-336.

Fredkin, E. (1960). Trie memory. *Communications of the ACM, 3,* 490-499.

Freudenthal, D., Pine, J. M., & Gobet, F. (2005). Resolving ambiguities in the extraction of syntactic categories through chunking. *Cognitive Systems Research, 6,* 17-25.

Gathercole, S. E., & Adams, A.-M. (1993). Phonological working memory in very young children. *Developmental Psychology, 29,* 770-778.

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language, 28,* 200-213.

Gathercole, S. E., & Baddeley, A. D. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology, 81,* 439-454.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language.* Hove, UK: Lawrence Erlbaum Associates.

Gathercole, S. E., & Martin, A. J. (1996). Interactive processes in phonological memory. In S. E. Gathercole (Ed.), *Models of short-term memory,* 73-100. Hove, UK: Psychology Press.

Gathercole, S. E., Willis, C., Emslie, H., & Baddeley, A. D. (1991). The influences of number of syllables and wordlikeness on children's repetition of nonwords. *Applied Psycholinguistics, 12,* 349-367.

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory, 2,* 103-127.

Gereveni, A., Perini, A., Ricci, F., Forti, D., Ioriatti, C., Mattedi, L., & Monetti, A. (1992). POMI - An expert system for integrated pest-management of apple orchards. *AI Applications, 6,* 51-62.

Gobet, F. (1993). A computer model of chess memory. In *Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gobet, F., Freudenthal, D., & Pine, J. M. (2004). Modelling syntactic development in

a cross-linguistic context. *Proceedings of the COLING 2004 Workshop "Psychocomputational Models of Human Language Acquisition"* (pp. 53-60). Geneva: COLING.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5,* 236-243.

Gobet, F., & Simon, H. A. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science, 24,* 651-682.

Gupta, P., & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language, 59,* 267-333.

Jones, G., Gobet, F. & Pine, J. M. (2000). A process model of children's early verb use. *Proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society* (pp. 723-728). Mahwah, NJ: Lawrence Erlbaum Associates.

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Sciences, 3,* 323-328.

Kazakov, D., & Manandhar, S. (2001). Unsupervised learning of word segmentation rules with genetic alogorithms and inductive logic programming. *Machine Learning, 43,* 121-162.

Miller, G. A. (1956). The magic number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81-93.

Nagy, W. E., & Herman, P. A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition*, 19-35. Hillsdale, NJ: Lawrence Erlbaum Associates.

Park, S., & Hyun, K. H. (2004). Trie for similarity matching in large video databases. *Information Systems, 29,* 641-652.

Perruchet, P., & Vintner, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*, 246-263.

Richman, H. B., & Simon, H. A. (1989). Context effects in letter perception: Comparison of two theories. *Psychological Review, 96,* 417-432.

Richman, H. B., Staszewski, J., & Simon, H. A. (1995). Simulation of expert memory with EPAM IV. *Psychological Review, 102,* 305-330.

Simon, H. A., & Gilmartin, K. J. (1973). A simulation of memory for chess positions. *Cognitive Psychology, 5,* 29-46.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2000). The role of performance limitations in the acquisition of 'mixed' verb-argument structure at stage 1. In M. Perkins & S. Howard (Eds.), *New directions in language development and disorders.* Plenum.

Zhang, G., & Simon, H. A. (1985). STM capacity for Chinese words and idioms: Chunking and acoustical loop hypotheses. *Memory and Cognition, 13,* 193-201.

AISB Journal **1**(6), © SSAISB, 2005
http://www.aisb.org.uk

A Generic Negotiation Model using XML

Philippe Mathieu and Marie-Hélène Verrons

Equipe SMAC, LIFL, USTL
Cité Scientifique, Bat M3, 59650 Villeneuve d'Ascq, FRANCE
mathieu@lifl.fr ; verrons@lifl.fr

Abstract

In this paper, we present a generic negotiation model for multi-agent systems called GeNCA, built on three levels: a communication level, a negotiation level and a strategic level, which is the only level specific to a particular application. XML files are used to configure both each agent and the global system, freeing the end-user from the need to reconfigure the system each time they want to change a parameter. The aim of this paper is then to show that it is possible to give a precise description of a generic negotiation model that we can use in several real problems. This model has been implemented with a Java API used to build our applications. GeNCA is the only platform which enables the use of different communication systems and of negotiation strategies specific to the applications achieved. These researches on negotiation take place in software engineering works for artificial intelligence and multi-agent systems.

1 Introduction

With the progress of information technology, multi-agent systems and electronic market places, the need of automatic agents able to negotiate with the others on behalf of the user becomes stronger and stronger. Moreover, the utility of using an agent during negotiations is perfectly justified by the explosion of the number of messages exchanged between agents. In certain cases, specially with cascaded renegotiations, the number of messages can be in $O(m^n)$ if n is the depth of the cascaded process and m the number of agents involved in one negotiation.

Since several years, negotiation has been studied by many researchers ((Rosenschein and Zlotkin, 1994; Sykara, 1989; Schwartz and Kraus, 1997)), and many negotiation systems have been achieved in specific domains like auctions or market places often in the aim of electronic commerce, let's cite Zeus (Nwana et al.,) developed by British Telecommunications, Magnet (Collins et al., 1998b) developed by the university of Minnesota, the SilkRoad project (Ströbel, 2001) developed by IBM, the platform GNP (Benyoucef et al., 2000) developed at the Montreal university and works done at HP Laboratories (Bartolini and Preist, 2001). Of course, negotiation can be used in other domains like meeting scheduling or reservation systems, but it seems that these ways have not been really studied. When studying such negotiation problems, we can see that many used notions are the same in many systems. For example, `contracts`, `resources`, `contractors (initiators)`, `participants` have a semantic equivalent in all negotiation systems. Our aim in the software engineering field, is to show that these notions can be reified

in a generic and open negotiation model and to build the corresponding API. The model we propose here is broad enough to allow classical negotiation applications to be covered without an adaptation effort, and has enough parameters to adapt to different negotiation applications, which is a difficult engineering problem.

Although it is difficult to define formally what is negotiation, we will base our arguments on the following consensual definition (Smith, 1980; Jennings et al., 2000; Walton and Krabbe, 1995), which can be applied to many fields such as auctions, appointment taking systems, games or others.

definition : Negotiation is carried out on a *contract* to obtain common *resources* and on the request of an *initiator*. It brings together a set of *participants* and an *initiator* and runs until an agreement satisfying a percentage of participants is reached. Participants equally try to obtain the best possible solution for themselves while giving a minimum set of information to the others.

definition : A contract is the entity which will be negotiated. It contains the *initiator* of the negotiation, the *resources* involved, the *answer delay* and a *default answer* in the case where a participant wouldn't have answered at time.

This definition is of course inspired of the Contract Net Protocol proposed by Smith (Smith, 1980) in 1980, which is a fundamental of many negotiation works (Sandholm, 2000). The main differences with the Contract Net is that negotiation ends with a contract between the initiator and several participants after possible rounds of proposals and counter-proposals. The initiator is the equivalent of the manager of the Contract Net and is in fact the first person who talk in the negotiation process. In the context of our study, we consider that a minimum number of information must be revealed to other agents, because when all information is known, we fall in a problem solver context, where algorithms such as a CSP is more fitted.

To conceive our model and allow a real generality, we have chosen a three-level architecture as a basis. The internal level which contains the management of data structures and speech acts necessary for agents to evolve their knowledge; the communication level allowing agents to send messages in a centralised way if agents are on the same computer, or in a distributed way if they are on different computers; the strategic level allowing agents to reason on the problem and infer on the knowledge obtained from the others. In our work, each level can be changed independently of the others. It is for example possible to use GeNCA in a round robin way with synchronous communication with all agents on the same computer to realise a video game where virtual beings will negotiate turn to turn, and to use it in a distributed way with asynchronous communication for electronic marketplace. In our model, the negotiating agent is composed of reactive micro-agents, where each micro-agent manages a negotiation.

The success of a negotiation depends of course on strategies adapted to the problem processed. We will not discuss here about strategies, which, to be optimal, must be different according to the kind of negotiation done. This is an important field which goes out of this paper. Therefore, we propose simple but generic strategies, which work for several kinds of problems, and that the user can easily refine.

We have identify many criteria to describe a negotiation, where we can find the number of rounds in a negotiation process, the minimum number of agreements needed to confirm the contract, the retraction possibility, or the answer delay. Many of them have been taken into account to build GeNCA.

A human user has two ways to use his agent. Manually, it is then a help-decision tool which shows the state of all the concurrent negotiations. In such case, it is the human user

who agrees a query. Automatically, this time the agent is hidden and proposes or answers queries by itself.

In GeNCA, the general server has an XML configuration file which allows to define the general notions like retraction possibility or the number of rounds in a negotiation process. Each agent can also have his own XML file to define the parameters of his owner (minimum number of agreements needed to confirm the contract, answer delay, etc.). Having XML files to configure the system makes it easier for the user to define a negotiation problem.

In this paper, we will first detail the protocol used (the phases of the protocol, the communication primitives and its properties). Then, we will describe GeNCA and the different ways to use it. After this, we detail two applications realised with GeNCA. Finally, we compare our works to others achieved on the same subject.

2 Proposed protocol

The protocol we propose here aims to define the messages that agents can send to each others with the operational dynamics associated. This negotiation protocol (Figure 1) is characterised by successive messages exchanged between an initiator (the agent who initiates the negotiation) and participants (the agents who participate to the negotiation) as in the Contract Net Protocol framework (Smith, 1980). We first describe the phases that compose our negotiation protocol, and then the communication primitives between agents used in this protocol. Finally we give characteristics of our negotiation protocol.

2.1 Protocol phases

We distinguish three phases for a negotiation process : the first one is the proposal phase which begins the negotiation process. Then, there is an optional phase named conversation phase. This phase consists of rounds of proposals and counter-proposals in order to converge to an acceptable contract for everyone. Finally, there is the final decision phase where the contract is either confirmed, either cancelled.

Proposal phase In this phase, the initiator proposes a contract to participants and waits for their answer. In response to the proposal, each participant answers if he agrees or rejects it.

Conversation phase This phase is necessary if there was not enough participants who agreed the contract proposal. A conversation is then started between the initiator and participants during which modification proposals are exchanged. Following these proposals, the initiator proposes a new contract to participants, and a new proposal phase is entered.

Final decision phase This final decision phase comes to either a confirmation or a cancellation of the contract. This decision is taken by the initiator in response to participants' answers.

2.2 Negotiation primitives

To carry out a negotiation process between agents, it is necessary to define several negotiation primitives between agents. We thus need specific primitives for initiators and

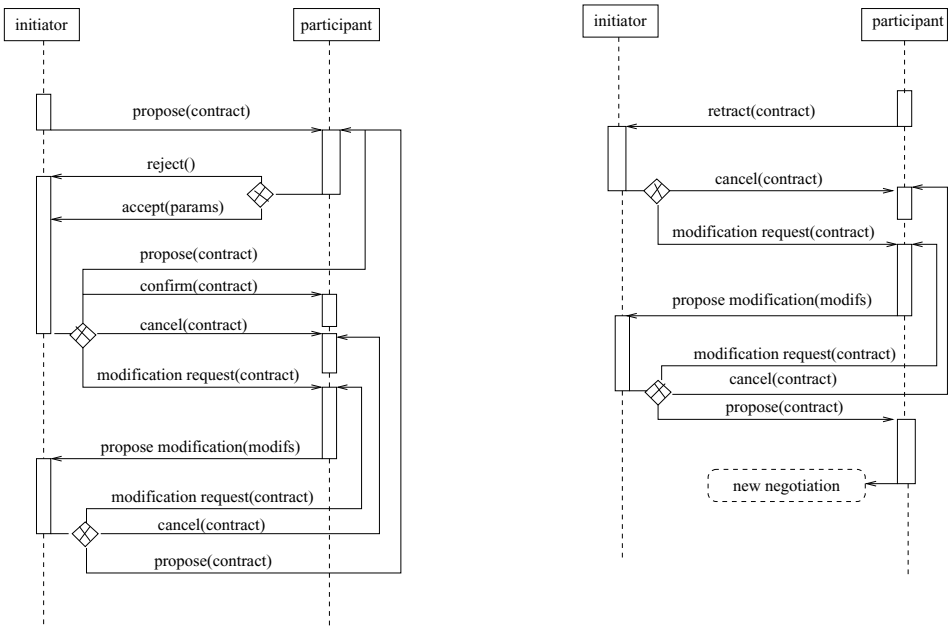


Figure 1: Negotiation protocol of GeNCA. Left : the sequence of messages between the initiator and the participants during a negotiation. For clarity, only one participant is shown here. Right : the sequence of messages for the renegotiation of a contract for which a participant has retracted.

specific primitives for participants. Our aim here is not to ensure communication between one of our agents and any other agent from another different platform (which would require a "FIPA-compliant" platform or more simply agents communicating via ACL), but to facilitate the development of an application with our agents. We don't use FIPA ACL or KQML for our negotiation primitives because they are not adapted to our protocol. The primitives defined by FIPA ACL deal with actions to perform or believes to assert. The specifications of the FIPA ACL primitives include conditions that can't be met with our model, so we can't use them with the meaning we want to give them. For example, the *propose* primitive denotes the intention to perform an action under certain conditions, whereas our meaning of *propose* is a contract offer from the initiator to participants, which will be accepted, rejected or discussed. We are not concerned here with believes. Moreover, FIPA ACL messages seem to be only textual messages, and the negotiation primitives we need for our model can't be used only with textual messages. Because of the content they use, our messages need to contain objects. The sequencing of these primitives is shown in Figure 1. Let us examine these primitives more deeply.

Initiator primitives The initiator begins and leads his negotiation process. He thus has specific primitives to do so. The initiator can send four negotiation primitives to a set of participants :

- *propose(contract)* : this is the first message sent by the initiator. He sends a contract proposal to the participants. The contract contains different resources to negotiate.
- *modification request(contract)* : this message indicates to participants that the con-

tract can't be taken like this and it has to be modified. The initiator asks participants to send him one or several possible modifications of the contract in order to propose a new one, better fitting everyone. This can also be a way to refine the contract.

- *confirm(contract)* : this message indicates participants that the contract is confirmed. The negotiation has been a success.

- *cancel(contract)* : this message indicates participants that the contract is cancelled. The negotiation failed.

Participant primitives Messages sent by a participant are only received by the initiator. It's a choice we made so that other participants don't know about these messages. Moreover, participants don't know the set of participants in the negotiation, they thus cannot form a coalition during negotiation. It is for example useful in Vickrey auctions where bids are private, or in other commercial negotiations where buyers could join their offers in order to have an interesting price as the quantity of goods asked is greater than if each buyer makes an offer for a lower quantity of goods.
Participants have three communication primitives which are answers to the initiator queries.

- *accept(parameters)* : this message replies to a contract proposal from the initiator. By this message, the participant indicates the initiator that he accepts the contract as it is. Parameters can be used in case of a partially instantiated contract. For example, it is the case in Vickrey auctions where participants have to propose a price for the article sold.

- *reject* : this message replies to a contract proposal from the initiator. By this message, the participant indicates the initiator that he refuses the contract.

- *propose modification(modification list)* : this message replies to a modification request from the initiator. The participant sends to the initiator a list of possible modifications for the contract. The number of modifications contained in the list is a negotiation parameter. This list can be empty if there is no possible modification for the contract.

A communication primitive is common to initiators and participants :

- *retract(contract)* : this primitive can be used only for a contract that has been confirmed earlier (after a *confirm* message has been sent for this contract). Both participants and initiators can use it. The agent sends this message to the initiator when he can't meet the contract taken anymore. The initiator can't prevent the agent to retract itself. Whether retraction is allowed or not depends on the application. Typically, retraction is not allowed in auctions, but is for appointment taking. That's why this possibility is a parameter of our negotiation model that is set up by the application designer, and the number of retractions allowed for the same negotiation is also a parameter.

2.3 Protocol characteristics

In this subsection, we present the type of applications achievable with this protocol, as it is aimed to be general, and then we give the complexity in number of messages exchanged during a negotiation process.

2.3.1 Applications achievable with this protocol

As we mentioned before, this protocol is inspired of the Contract-Net, and it adds an optional phase of conversation. As the protocol describes messages exchanged between agents but especially the order of messages and agents' turn to talk, and not what is the content of the message (for example, always a price ...), it allows many different applications to use it, which is not the case of many protocols such as the one used in ZEUS which is dedicated to marketplaces.

For example, you can use it in a "take it or leave it offer" form if you don't use the conversation phase. If you want to make auctions applications, you can implement English auctions as well as Dutch auctions. For English auctions, the initiator proposes his articles and participants answer giving a price as argument of the accept message if they are interested in the article, or rejecting the proposal otherwise. If no participant has proposed a satisfying price for the initiator, a conversation phase is entered where each modification consists of a new bid. The process finishes when a satisfying price has been proposed or when no one rebids or the maximum number of turns predefined by the initiator has been reached.

For Dutch auctions, the initiator proposes an article with a high price, and if no participant accepts the proposal, the initiator proposes again the article with a lower price without asking for a modification from participants. The process finishes when a participant accepts the contract, or when the price reaches the minimum price wished by the initiator, or when the maximum number of rounds defined by the initiator is reached.

This protocol is not adapted to negotiations that have to be processed on several levels, for example, for negotiating to buy a car, you can first negotiate the colour, and then the price .... This protocol is not adapted to combined negotiations (Aknine, 2002), where contracts need to be linked. For example, you can't create two contracts and say both must be taken or none. If you want several resources from the same person, you put them in a single contract, but if you want several resources from several persons, you'll need one contract per person/resource but you can't specify that all contracts must be taken or none. Despite the protocol could fit it, negotiation with argumentation (Parsons et al., 1998) is not included in GeNCA. The protocol could be adapted since the parameters of acceptance or modifications could be arguments.

2.3.2 Complexity

Complexity is an important feature in negotiation. Negotiation complexity is the reason why you can't do without negotiating agents. Let's examine here complexity in number of messages induced by our protocol.

In the worth case, for m participants at a negotiation process, the number of messages to be sent is m^n if n is the depth of cascaded renegotiation process. You imagine easily what could happen to your secretary in such case to organise a meeting with fifty people.

To prove this result, let us look at the different cases that can happen.

Linear order Assume that m persons want to take a contract. Let's call *initiator* the person who wishes to take the contract and *participants* the others. Figure 2 shows five persons, before and after that the contract has been taken (each dot represents one person).

Firstly, let us consider that all participants agree with the proposal. The initiator *proposes* the contract, the participants *agree* and the initiator *confirms* : $3 * (n - 1)$ messages are exchanged.

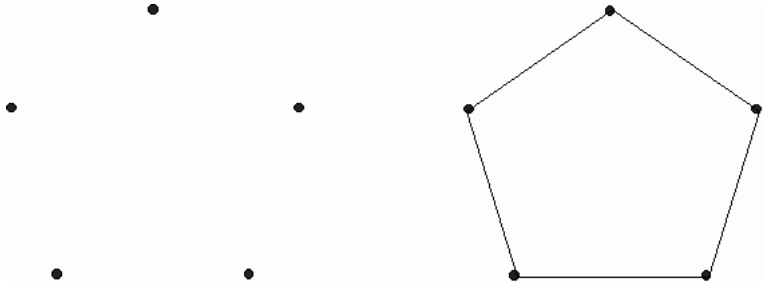


Figure 2: Complexity in linear order

As soon as one participant disagrees, the initiator *requests a modification* from participants who send one to the initiator (*propose modification* message). $2 * (n-1)$ messages are then exchanged. The initiator sends a new *proposal* which will be accepted, adding $3 * (n-1)$ messages. In total, $7 * (n-1)$ are exchanged, taking into account those of the first proposal and answers of participants with at least a negative one. The initiator sends $4 * (n-1)$ messages and receives $3 * (n-1)$. Each participant receives 4 messages and sends 3.

Taking a contract, with or without modification request, without renegotiation of other contracts, has a global complexity in O(n), is linear for the initiator and in O(1) for participants.

Quadratic order

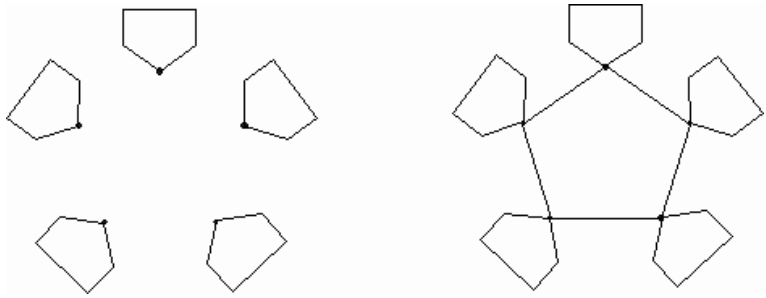


Figure 3: Complexity in quadratic order - first case

First case Let us now assume that taking a contract calls previous contracts already taken with other persons into question (Figure 3).

To simplify, all contracts will involve n persons and will have the same priority.

Participants will modify the contract, $7 * (n-1)$ messages will then be sent. But, at time to confirm the contract, each participant will have to request a modification for the contract he has already taken. Let us assume that modifications are accepted without any problem. The number of exchanged messages in this renegotiation is $5 * (n-1)$. Participants of the first contract, considered as initiators of the second ones, send $3*(n-1)$ and receive $2 * (n-1)$ messages. If all renegotiations are independent, there are $(n-1)$

renegotiations and thus $5 * (n-1)^2$ messages. The total number of exchanged messages for taking the contract is thus $5 * (n-1)^2 + 7 * (n-1)$. The initiator sends $4 * (n-1)$ and receives $3 * (n-1)$ messages. Each participant receives $4 + 2 * (n-1)$ messages and sends $3 + 3 * (n-1)$.

Taking a contract with renegotiation of another one by participant has a global complexity of O(n^2) and is linear for the initiator and participants.

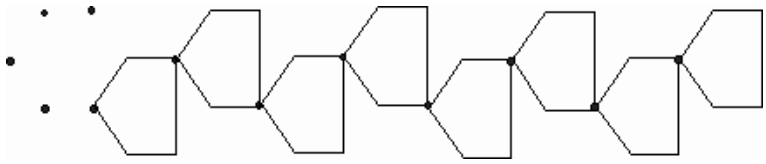


Figure 4: Complexity in quadratic order - second case

Second case Let us now assume that only one participant has to modify a contract already taken with another person (Figure 4). During renegotiation, this person also has to modify another contract and recursively on m persons. The principal negotiation needs $7 * (n-1)$ messages, the others $5 * (n-1)$. The total number of messages is $(2 + 7 * m) * (n-1)$ messages.

Taking a contract with renegotiation of another one by one participant and this recursively at a depth of m, has a global complexity of O($n * m$) and is linear for the initiator and participants.

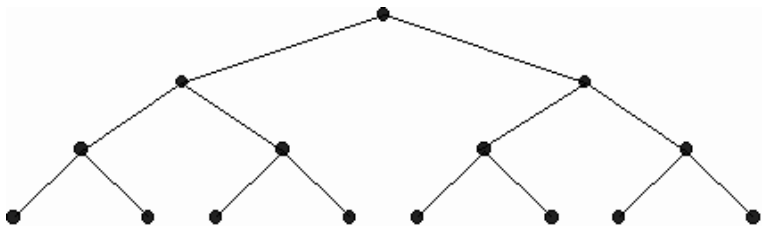


Figure 5: Complexity in exponential order

Exponential order To prove the result given at the beginning of subsection, let us take a formal example. For this example, a contract will always be negotiated between one initiator and two participants. Figure 5 shows a binary tree representing the cascaded renegotiation process. The root of this tree is the initiator of the first contract. He has got two children : the two participants. Each participant is in his turn the initiator of another contract, having also two children etc. We suppose here that there are no other relations between all these agents, ie. they are all different, all nodes represent a different agent.

Having this, we can now compute the number of messages that will be exchanged.

The number of exchanged messages for a modification of a contract which will be immediately accepted is equal to five : modification request, modification from participants, proposal of a new contract (the old one modified), agreement from participants and then confirmation of this new contract. The number of agents at level i equals 2^i and the number of messages exchanged at that level is $5 * 2^i$.

Global complexity is thus O(2^n) and is linear for the initiator and the participants.

If we now suppose that contracts are not independent anymore but that agents at level n ask the initiator of the principal contract to modify another one, the number of asks for renegotiations will be 2^n for the initiator.

Global complexity is still O(2^n) and keeps linear for participants, on the other hand, it becomes O(2^n) for the initiator.

In this section, we presented the negotiation protocol used in GeNCA, let's now see the different use modes of GeNCA.

3 GeNCA

GeNCA is a Java API for negotiation between agents. It is aimed to provide a generic software architecture for contract-based negotiations to applications developers in order to facilitate their work. The internal objects needed to the implementation of GeNCA are described in (Mathieu and Verrons, 2002). The novelty in GeNCA is that the parameters that are needed to configure a negotiation application are set up in XML files, thus avoiding recompilations at each change of a parameter value and facilitating the writing of a new application. Two kinds of files are defined : one for the system parameterisation, one for each agent which is optional. The system file contain common characteristics for all users of the negotiation system. We define them in a DTD file called genca.dtd available at http://www.lifl.fr/SMAC/projects/genca. Common resources, agents initially present in the system, retraction ability are found in it, plus default values for users parameters. Each agent can have its own file to set up its individual resources, its communicator, its strategies and negotiation parameters like default answer and answer delay. Figure 6 shows the system XML file for an appointment taking application.

We discuss here about the different ways to use GeNCA, and its major features.

3.1 GeNCA features

GeNCA major features are its conception in three levels, its negotiation cardinality and the management of deadlocks.

Conception in three levels The first feature of GeNCA is his conception in three levels, in order to separate the implementation of communications between agents, the implementation of negotiations management and the implementation of negotiations strategies. These three levels are presented more deeply in (Mathieu and Verrons, 2003a; Mathieu and Verrons, 2003b). We decided to separate these three levels in order to provide more facilities to adapt the negotiation system to applications as their common need is the negotiation level. As a matter of fact, each application has its own communication system and needs specific strategies of negotiation. For example, communications between distributed agents can be done via e-mail or a MAS platform, while communications between centralised agents can be done in a round-robin way. It is easy to define which communicator or which strategy an agent will use as it is set up in an XML file. This separation of these three levels is a difficult software engineering problem, and from our knowledge, no other platform than GeNCA separates them.

Negotiation cardinality Negotiation cardinality is an important feature for MAS. Its purpose is to know how many agents negotiate together. Different kinds of negotiation cardinality exist (Guttman and Maes, 1998), from one-to-one to many-to-many. Kasbah is an example of one-to-one negotiation : one buyer negotiates an article with one seller at a time. This form of negotiation is useful when only two persons are involved in the negotiation. But when a negotiation involves many participants with an initiator, it is a one-to-many negotiation. Our protocol enables contract-based negotiation between one initiator and several participants. Our implementation of this protocol in GeNCA allows several negotiations to take place simultaneously between one initiator and several participants, that is to say many-to-many negotiation, or more precisely many (one-to-many) negotiation. The advantage provided by many-to-many negotiation is that it enables one-to-many and one-to-one negotiation.

Deadlocks Deadlocks are an important problem in negotiation applications. It can cause many damages if it is not resolved. Deadlocks can appear when two agents propose a contract on the same resource one to the other, and when they chose to negotiate sequentially contracts on same resources. Both are then waiting to the other's answer and the deadlock appears. Deadlocks are avoided in GeNCA thanks to our mechanism of answer delay. As a matter of fact, each initiator defines the delay that have participants to answer. If a participant doesn't answer before this delay, the initiator takes into account a default answer for him and so, negotiation is not blocked.

3.2 GeNCA use modes

GeNCA can be used in different modes, which gives its genericity. Among these ways to use it, we find the kind of resources negotiated, simultaneous management, automatic renegotiation, tools for strategies and agents use modes.

Resources Resources that will be negotiated can be common to all agents or individual. If we take the example of meeting scheduling, each agent has the same agenda, and so the same time slots. Thus, resources (time slots) are common to all agents and any of them can make a proposal on the time slots he wants. On the contrary, auctions applications are typically those where we find individual resources. Agents wishing to sell articles will sell only their own articles, and not the one of its neighbours. So, for this kind of applications, resources are individual, visible to all agents but only the agents that possess them can make a contract proposal. Resources are described in XML files. If they are common to all agents, they are set up in the system file, but if they are individual, they are set up in the agent file.

Simultaneous management The management of negotiations is an important criterion in a negotiation application. Negotiations can be processed sequentially, or in parallel, depending on the constraints of the application. Two managements are possible in GeNCA, immediately or deferred simultaneous management. The user opts for the one he prefers. When he chooses to negotiate immediately all contracts, no restriction is made on the resources, they can already being negotiated for another contract. But if the user chooses to negotiate in a deferred way, the only negotiations that will take place simultaneously are the ones which involves disjoint sets of resources. The other negotiations will wait for their turn. This management of simultaneous negotiations is possible because we have designed a structure to check if all resources needed for a negotiation are free or

yet under negotiation, and so to know if the negotiation process can begin or not. This structure is a Tetris like matrix, which is described in (Mathieu and Verrons, 2002). Simultaneous negotiations are possible because we've chosen to entrust micro-agents with the management of one negotiation. In fact, each time an agent creates or receives a proposal, a micro-agent is created (a goal if the agent is the initiator, an engagement if the agent is a participant) which is responsible for the whole negotiation process of this proposal. It is thus possible to negotiate simultaneously several contracts, and being initiator as well as participant in the same time.

Automatic renegotiation Many times, during negotiations, some contracts can't be met any longer and has to be negotiated again. It is the case when appointments are negotiated. For this purpose, we propose to renegotiate automatically contracts that have to be moved. But you can't always question a contract that has been taken. For example in auctions, when an article is sold, it is definitely sold, you can't retract yourself. That's why we define a parameter called retraction allowed, used to know whether it is possible or not to retract yourself from a contract previously taken. This is a common parameter to all agents which is defined in the system XML file. If retraction is allowed, when an agent retracts itself, the initiator of the contract can automatically renegotiate the contract, and a number of renegotiations is defined by the initiator (in the agent XML file) to know how many times a contract can be negotiated again.

Tools for strategies The success of a negotiation depends of course on strategies adapted to the problem processed. We will not discuss here about strategies, which, to be optimal, must be different according to the kind of negotiation done. This is an important field which goes out of this paper. Therefore, we propose simple but generic strategies, which work for all kinds of problems, and that the user can easily refine. In order to give basis to develop strategies, two priority lists are defined in GeNCA. Each person defines a priority list for resources and a priority list for persons. Thus, each person will be able to give a priority to a contract according to priorities of resources included in the contract, and according to the initiator's priority. For example, if I took an appointment with a colleague and my boss asks me for an appointment at the same time, I will take the appointment with my boss (who has a greater priority) and I will move the appointment with my colleague. These lists can also be used in case that I am initiator of a contract and I requested modifications from participants, I can weight their answer according to the priority I gave them.

GeNCA also provides rates of success or retraction of negotiations that have been done in the past, given a participant and a set of resources. It is thus possible to know if a participant globally accepts proposals he receives, and if he keeps his engagements.

Agents use modes As we mentioned before, a human user has several ways to use its agent. He can use it with a graphical interface to interact with it, in this case, the agent is a help decision tool for the user. The agent manages the negotiations and it is the user who answers contract proposal, and creates contract to negotiate. Through the interface, the user views messages received and sent, contracts taken and being negotiated, and he can create a new contract, cancel a contract he has previously taken and reply to a contract proposal.

Another way to use the agent is the automatic way, in this case, the agent manages the whole negotiation and replies itself to proposals, the graphical interface is not used, and the agent runs like a background task.

GeNCA features and use modes have been applied to several negotiation applications like appointment taking, Dutch and English auctions and timetable creation. These applications can be downloaded at http://www.lifl.fr/SMAC/projects/genca.

In the next section, we present two applications realised with GeNCA.

4 Applications

Our aim is to propose a generic model to negotiate contracts whatever they are. The model, we called GeNCA, has been implemented in the Java language in order to provide an API for the creation of contract negotiation applications. Here we present two applications among those we have developed with GeNCA. One of them is a classical one, it uses participant individual resources, it is an auction application. The other is much less classical, it uses resources common for all participants, it is an appointment-taking system.

4.1 Application with common resources

The first application we describe here is the one which involves common resources for all participants in the negotiation. It's an appointment taking application where resources are time slots. Each agent must be able to negotiate appointments for the user. Each user defines a schedule with time slots which are free or not. In addition, he gives preferences on slots and on persons with whom he prefers to take appointments. As resources are common for all participants, each one is able to create a contract for one or several resources and to propose it to a set of participants. There is no essential need for each user to have his own XML file since resources are defined once for all in the system XML file. We obviously don't let the agents share their schedules in order to find a suitable time slot for an appointment.

This problem is a full-featured one because it needs preferences over persons, for example, the boss has a greater priority than the colleague, but also priorities over resources (here time slots), e.g. if I don't want to have appointments at lunch time or before 8 am, I'll give the corresponding time slots a lower priority. Moreover, appointment taking is an application where there are typically many renegotiations and retractions, because it is difficult to find time slots that fit everyone.

This appointment taking application involves resources of one hour timeslot in one day, and four agents running on the same computer. The system file (Figure 6) contains thus these resources and agents, and defines that retraction is possible, ie an appointment can be moved if it can't be maintained at the time defined. For this application, we used the Magique platform to run our agents, so the Magique communicator is used. Specific strategies have been implemented to fit the application, particularly to group consecutive hours if one hour was too short for the appointment.

Default values for users' parameters are set up like this : each participant has 10 minutes to answer the proposal, and would be considered as rejecting the proposal if he doesn't answer. Everyone must agree for the appointment to be taken. The initiator can request 20 times modifications from participants who can propose 5 modifications at a time. The appointment can be moved 3 times and all negotiations that take place simultaneously must involve different time slots.

This single file is sufficient to launch the application with these four agents. They all have their own GUI to create contracts, answer to proposals, view their messages sent and received and the contracts they've taken.

```
<?xml version="1.0"?>
<!DOCTYPE genca SYSTEM "genca.dtd" >
<genca>
<negotiation-type>rdv</negotiation-type>
<resources-list>
<resource>8h-9h</resource>
<resource>9h-10h</resource>
<resource>10h-11h</resource>
<resource>11h-12h</resource>
<resource>14h-15h</resource>
<resource>15h-16h</resource>
<resource>16h-17h</resource>
<resource>17h-18h</resource>
</resources-list>
<agents-list>
<agent><name>Paul</name>
       <address>localhost</address></agent>
<agent><name>Pierre</name>
       <address>localhost</address></agent>
<agent><name>Jean</name>
       <address>localhost</address></agent>
<agent><name>Jacques</name>
       <address>localhost</address></agent>
</agents-list>
<default-communicator>
fr.lifl.genca.magique.MagiqueCommunicator
</default-communicator>
<default-initiator-strategy>
rdv.RdvInitiatorStrategy
</default-initiator-strategy>
<default-participant-strategy>
rdv.RdvParticipantStrategy
</default-participant-strategy>
<nbRounds>20</nbRounds>
<nbRenegotiations>3</nbRenegotiations>
<minAgreements>100%</minAgreements>
<answer-delay>10</answer-delay>
<default-answer value="refuse"/>
<simultaneity value="deferred"/>
<retraction-allowed value="true"/>
<nb-modifications-by-round>5
</nb-modifications-by-round>
<magique><skill><class>
fr.lifl.genca.magique.NegotiationSkill
</class></skill></magique>
</genca>
```

Figure 6: System XML file for appointment taking application

4.1.1 Initiator behaviour

The initiator first chooses the participants he wants to meet and a time slot for the meeting. He also checks the parameters of the negotiation, such as the default answer, the minimum number of agreements to take the appointment, etc. All this define the contract and its properties. The contract is then proposed to the set of participants. The initiator thus uses the *propose* message of the protocol. Then, he waits for participants answers during the answer delay he has defined.

When the delay is over, the initiator checks participants answers. If there are more agreements than the minimum number of agreements he has chosen, he then *confirms* the contract for the participants who have agreed, and *cancels* the contract for the others. Otherwise, he *requests a modification* to all participants if the maximum number of rounds of negotiation is not reached. In the other case, he *cancels* the contract for everyone.

If the initiator requests a modification, he then waits for propositions from participants during the same answer delay. After this delay, he takes one of the following decisions :

- He *proposes* a new contract based on the propositions of the participants.
- He can't find a new contract proposal, so he *requests* again a modification from participants.
- He *cancels* the contract.

If the initiator receives a *retraction* message, he checks if there are enough participants left. In this case, he only removes the retracting participant from the list of agreed participants. In the other case, he *cancels* the contract for everyone and *requests a modification* from all participants in order to find a new contract that satisfies the participants.

4.1.2 Participant behaviour

When a participant receives a contract proposal, he first checks if the time slots proposed are free in his agenda. If they are, he accepts the proposal, thus sending the *accept* message. If the slots aren't free, he compares the priority of the initiator of the contract taken previously for these slots with the priority of the initiator of the new contract. If the older initiator has a greater priority, he then *rejects* the proposal. Otherwise, he *accepts* it.

When the participant receives a modification requests, he sends to the initiator a list of free time slots in order of preferences according to the priority he has given to the slots via the *propose modification* message.

When a contract is confirmed, the participant adds it in his agenda and *retracts* itself from previous contracts he has taken on the same time slots if they exist.

This application allows agents to negotiate appointments for its user. Contrary to other systems that can be found in shops, users' agendas are private and the problem isn't to find a suitable time slot free and common to all participants, knowing their agendas, but to negotiate the hour of the appointment, taking into account the preferences of the users on hours and persons. Moreover, this system renegotiate automatically an appointment that has to be moved due to participants retractions.

4.2 Application with individual resources

Auction applications are typically applications where resources are individual for participants. The only participants who will create contracts are the ones who possess goods to

sell. We describe here an auction application where some participants want to sell goods they have defined in their own XML file.

In this auction application, each agent must be able to negotiate auctions for the user. For this purpose, each user defines an amount of money (his credit), and a bidding strategy (linear, quadratic,...).

Auctions are defined like this : a seller proposes an article for which he wants to obtain a minimal price that he keeps secret (reservation price). Then, buyers tell him if they are interested (accept) or not (reject) in it, and if they are they propose a price for it. The seller keeps the highest price proposed and the buyer who proposed it. If this price is greater than or equals the reservation price, the buyer wins the auction. Else, the seller proposes again his article to the interested buyers for them to propose a higher price. This process is repeated until a buyer wins the auction or the number of rounds is reached.

For this application, retraction is not allowed, once an article is sold, it is definitely sold.

For this application, there are no common resources in the XML system configuration file and we launch four agents on the same computer. For this application, these agents run on a Magique platform and so they use the Magique communicator to exchange messages. Two strategies have been written to evaluate and propose bids, which are the default strategies set up in the system file. Only one person can buy the article, so the parameter *minAgreements* is set up to 1. *Three minutes* are granted for participants to bid, if they don't, the initiator considers that they *reject* the proposition. If no bid fits the initiator, he can ask a new bid *20 times* to participants, who propose a *single bid* by round. *Retraction* is not allowed. Auctions on same goods are processed sequentially, that's why the parameter simultaneity has the value *deferred*.

If users are satisfied with these parameters and do not have goods to sell, they do not need to have their own XML files. Let us take the example of our agent named Jean who wants to sell goods . Thus, he has his XML file Jean.xml where his goods (a fridge, a table and a chair, for example) are listed in the resources list. The other parameters this agent will use are those defined in the system file.

4.2.1 Initiator behaviour

The initiator first creates a contract containing the article to sell, the reservation price, the other negotiation parameters and the set of participants. The initiator then sends this *proposal* to the participants.

When an agreement is received, the initiator updates the highest bid proposed so far. If the new price tops the highest bid proposed, this new bid becomes the highest and the buyer who proposed it the current winner of the auction. Once all replies have been received, the initiator decides to *confirm* the auction for the current winner if the highest bid tops the reservation price, and thus to *cancel* the auction for the other participants. If neither the reservation price nor the maximum number of rounds are reached, then the initiator *requests a modification* from the interested buyers, in other cases, he *cancels* the auction.

When a modification proposal is received, the initiator proceeds exactly as for an agreement, as a modification proposal is a new price for the article.

4.2.2 Participant behaviour

When a participant receives an auction proposal, he first checks if the article interests him or not. If he is interested in it, he *accepts* the contract and proposes a price. Otherwise, he

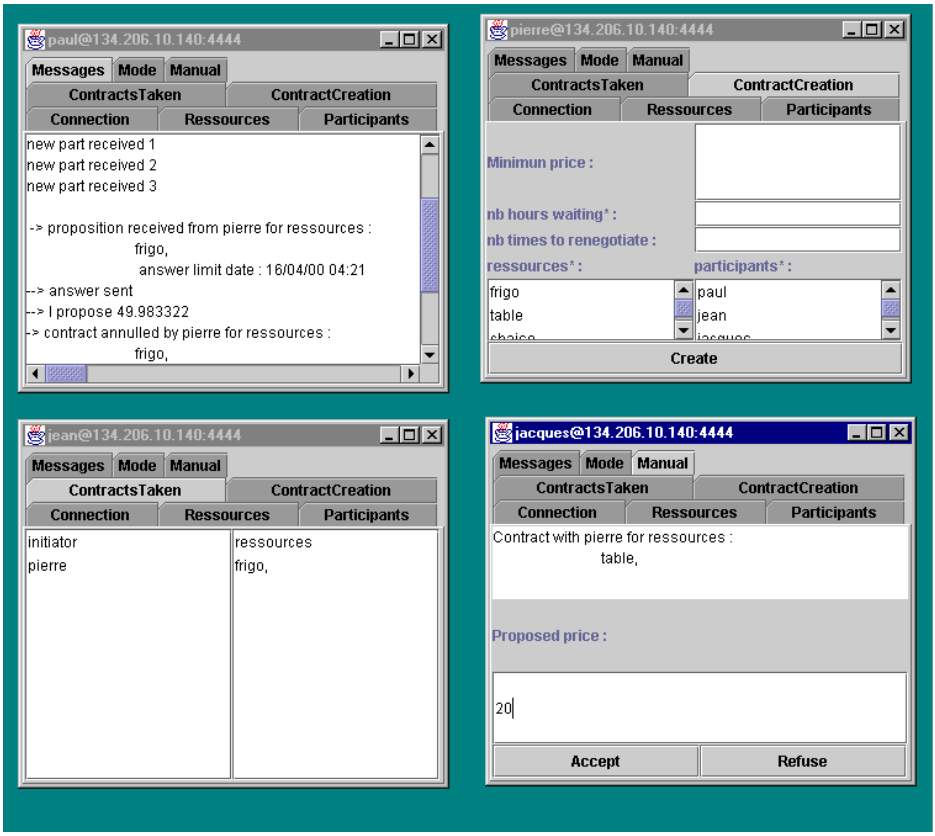


Figure 7: Four agents participating in the auction application

rejects the proposal.

When an auction confirmation is received, the participant adds the article in his bag and virtually pays the price to the seller.

When a modification request is received, the participant checks the amount of money he has and proposes a higher price than in the previous round if he has enough money or a price equal to 0 if he doesn't want to participate further in the auction.

Figure 7 shows the graphic interfaces of four agents negotiating auctions with our API.

The top left-hand screen is an agent showing his window for visualising messages sent and received by him. It permits to see the different proposals received and the proceedings of the negotiation (answer sent, confirm, cancel, modification request,...). The top right-hand screen is an agent showing the new contract input interface, the bottom left-hand one displays contracts chosen with the name of the initiator and the negotiated resources. The last one shows the display of a contract proposal for manual mode.

The advantages of this application are numerous, the most important ones are mentioned here. First, this application helps the user to bid, and bids in his place when he's not there, according to the strategy he has defined. Secondly, this application can easily be extended to other kinds of auctions, like English, Dutch, Vickrey auctions... And thirdly, this application is portable, as a matter of fact, agents can be placed on PDAs or over a

heterogeneous network.

These two examples show that our protocol can be applied to different kinds of negotiation applications such as auctions or appointment-taking. This illustrates our purpose of a generic protocol. In the next section, we compare our protocol with different applications developed by other researchers to show the differences between them.

5 Comparison with other works

We are obviously not the only ones who are interested in negotiation between agents and in proposing a generic architecture to accomplish it. Let's cite the works realised at HP Laboratories by Claudio Bartolini et al. (Bartolini and Preist, 2001; Bartolini et al., 2002b; Bartolini et al., 2002a) who want to create a general framework for automated negotiation dedicated to market mechanisms. In this paper, they define two roles : participant and negotiation host. A participant is an agent who wants to reach an agreement, while the negotiation host is responsible for enforcing the protocol and rules of negotiation. Rules of negotiation include posting rule, visibility rule, termination rule .... It is the negotiation host who is responsible for making agreements. This framework proposes a general negotiation protocol parameterised with rules to implement a variety of negotiation mechanisms. It has common properties with our, like enabling one-to-one, one-to-many and many-to-many negotiations, or like parameterisation.

Another formal work we can cite is the one done by Morad Benyoussef et al. (Benyoucef et al., 2000) who want to create a Generic Negotiation Platform for marketplaces.

A third work is the SilkRoad project (Ströbel, 2001). This project aims to facilitate the design and implementation of negotiation support systems for specific application domains. SilkRoad facilitates multi-attribute negotiations in e-business scenarios through a specific design methodology and a generic system architecture with reusable negotiation support components. A negotiation support system built on the basis of the SilkRoad architecture model acts as an intermediary between the actual negotiating agents (which might be software agents or humans) and thereby provides rule-driven communication and decision support. This project has common points with ours, like the possibility to have either software or human agents and the genericity of the system.

These three works are close to our, but they are more directed to electronic commerce whereas our model aims to fit also other types of automated negotiations.

Let's now examine two platform for negotiation : magnet and zeus. **Multi AGent NEgotiation Testbed** (Collins et al., 1998a) is a testbed for multi-agent negotiation, implemented as a generalised market architecture and developed at the university of Minnesota. It provides a support for a variety of types of transaction, from simple buying and selling of goods to complex multi-agent contract negotiation. A session mechanism enables a customer to issue a call-for-bids and conduct other business. The negotiation protocol for planning by contracting consists of three phases : a call-for-bids, bidding and bid acceptance. In contrast, our protocol enables the initiator of the call-for-bids to make counter-proposals until an agreement is reached. In MAGNET, there is an explicit intermediary into the negotiation process and agents interact with each other through it, whereas all agents directly interact with each other in our negotiation process.

ZEUS (Nwana et al.,) is a generic Java API realised by British Telecom in order to easily conceive cost-based negotiation applications between autonomous agents. Zeus proposes a negotiation protocol between two agents (an initiator and a participant) and on a single resource per contract. The protocol consists of a call-for-bids, and no mechanism

of counter-proposal is provided. Moreover, it is possible to negotiate simultaneously different contracts on the same resource, that we don't allow. Another difference with our protocol is that retraction is not possible with Zeus. Once a contract is taken you can't retract yourself. Moreover, Zeus provides only cost-based strategies, and so is less generic than our protocol which is not dedicated to cost-based contracts. Although it is possible to add an interaction protocol in Zeus, it is a difficult thing to do, as says S. Thompson in the mailing list of Zeus in April 2002. On the other hand, parameters of GeNCA negotiation protocol can be set up in XML files, which simplifies modifications.

These previous works, like our, are based on the general **Contract Net Protocol** model (Smith, 1980) which works on bids invitation between a Manager agent and Contractor agents. From all these works, Magnet is probably the one which is closest to what we present. Nevertheless, none of them takes into account at the same time generic aspects, automatic renegotiations and a mechanism to manage conflicts between simultaneous negotiations, that we propose in GeNCA. Moreover, GeNCA is the only platform which separates the communication level, the negotiation level and the strategic level.

6 Conclusion

In this paper, we have presented a generic protocol for contract-based negotiation and a Java API called GeNCA, which enables many-to-many negotiations, simultaneous negotiation of several contracts, and the management of deadlocks in conversation. Three distinct levels were defined : the knowledge representation level allowing the agent viewing the advancement of his/her negotiations, the communication level which we realised with a multi-agent platform allowing physical distribution, and the strategic level for which we propose generic strategies adaptable to any kind of problem. Each level can be easily extended by the developer as he wants to map with his application, which is a feature that only GeNCA proposes. Moreover, XML files are used to set up parameters and define an application, which facilitates the end-user work, and avoid useless recompilations. These works are a part of software engineering and distributed artificial intelligence works. Many implementation perspectives of these works on different software supports are possible (distributed, centralised, WEB) and strategic level enhancement for different specific problems is considered. This API will now be applied to different problems like distance teaching, network games, workflow systems.

References

Aknine, S. (2002). New Multi-Agent Protocols for M-N-P Negotiations in Electronic Commerce. In *National Conference on Artificial Intelligence, AAAI, Agent-Based Technologies for B2B Workshop*, Edmonton, Canada.

Bartolini, C. and Preist, C. (2001). A framework for automated negotiation. Technical Report HPL-2001-90, HP Laboratories Bristol.

Bartolini, C., Preist, C., and Jennings, N. R. (2002a). Architecting for reuse: A software framework for automated negotiation. In *Proc. 3rd Int Workshop on Agent-Oriented Software Engineering*, pages 87–98, Bologna, Italy.

Bartolini, C., Preist, C., and Jennings, N. R. (2002b). A generic software framework for automated negotiation. Technical Report HPL-2002-2, HP Laboratories Bristol.

Benyoucef, M., Keller, R. K., Lamouroux, S., Robert, J., and Trussart, V. (2000). Towards a Generic E-Negotiation Platform. In *Proceedings of the Sixth International Conference on Re-Technologies for Information Systems*, pages 95–109, Zurich, Switzerland.

Collins, J., Tsvetovatyy, M., Mobasher, B., and Gini, M. (1998a). MAGNET : A Multi-Agent Contracting System for Plan Execution. In *Workshop on Artificial Intelligence and Manufacturing: State of the Art and State of Practice*, pages 63–68, Albuquerque, NM. AAAI Press.

Collins, J., Youngdahl, B., Jamison, S., Mobasher, B., and Gini, M. (1998b). A Market Architecture for Multi-Agent Contracting. In *2nd Int'l Conf on Autonomous Agents*, pages 285–292, Minneapolis.

Guttman, R. H. and Maes, P. (1998). Cooperative vs. Competitive Multi-Agent Negotiations in Retail Electronic Commerce. In *Proceedings of the Second International Workshop on Cooperative Information Agents (CIA'98)*, Paris, France.

Jennings, N. R., Parsons, S., Sierra, C., and Faratin, P. (2000). Automated Negotiation. In *Proc 5th Int. Conf. on the Practical Application of Intelligent Agents and M.A.S., PAAM-2000*, pages 23–30, Manchester, UK.

Mathieu, P. and Verrons, M.-H. (2002). A generic model for contract negotiation. In *Proceedings of the AISB'02 Convention*, London, UK.

Mathieu, P. and Verrons, M.-H. (2003a). ANTS : an API for creating negotiation applications. In *Proceedings of the 10th ISPE International Conference on Concurrent Engineering: Research and Applications (CE2003), track on Agents and Multi-agent systems*, pages 169–176, Madeira Island, Portugal.

Mathieu, P. and Verrons, M.-H. (2003b). A Generic Negotiation Model for MAS using XML. In *Proceedings of the ABA workshop Agent-based Systems for Autonomous Processing, held by the IEEE International Conference on Systems, Man and Cybernetics.*, Washington, USA. IEEE Press.

Nwana, H., D.T. Ndumu, L. L., and Collis, J. ZEUS : A Toolkit for Building Distributed Multi-Agent Systems.

Parsons, S., Sierra, C., and Jennings, N. R. (1998). Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261–292.

Rosenschein, J. and Zlotkin, G. (1994). *Rules of encounter : designing conventions for automated negotiation among computers.* MIT Press, Cambridge, Mass.

Sandholm, T. (2000). *Automated Negotiation*. MIT Press.

Schwartz, R. and Kraus, S. (1997). Negotiation on Data Allocation in Multi-Agent Environments. In *Proc. of the AAAI-97*, pages 29–35.

Smith, R. G. (1980). The Contract Net Protocol : high-level communication and control in a distributed problem solver. *IEEE Transactions on computers*, C-29(12):1104–1113.

Ströbel, M. (2001). Design of Roles and Protocols for Electronic Negotiations. *Electronic Commerce Research, Special Issue on Market Design*, 1(3):335–353.

Sykara, K. (1989). Multiagent compromise via negotiation. In Gasser, L. and Huhns, M., editors, *Distributed Artificial Intelligence*, volume 2, pages 119–137, Los Altos, CA. Morgan Kaufmann Publishers.

Walton, D. and Krabbe, E. (1995). *Commitment in Dialogue*. SUNY Press.

AISB Journal **1**(6), © SSAISB, 2005
http://www.aisb.org.uk

Eliciting Test-selection Strategies for a Decision-Support System in Oncology

Danielle Sent*, Linda C. van der Gaag*, Cilia L.M. Witteman*,
Berthe M.P. Aleman† and Babs G. Taal†

* Institute of Information and Computing Sciences, Utrecht University, P.O. Box 80.089, 3508 TB Utrecht, The Netherlands,
danielle@cs.uu.nl ; linda@cs.uu.nl ; cilia@cs.uu.nl

† The Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, Department of Radiation Oncology and Gastroenterology, Plesmanlaan 121, 1066 CZ Amsterdam, The Netherlands,
b.aleman@nki.nl ; b.taal@nki.nl

Abstract

Decision-support systems often include a strategy for selecting tests in their domain of application. Such a strategy serves to provide support for the reasoning processes in the domain. Generally a test-selection strategy is offered in which tests are selected sequentially. Upon building a system for the domain of oesophageal cancer, however, we felt that a sequential strategy would be an oversimplification of daily practice. To design a test-selection strategy for our system, we decided therefore to acquire knowledge about the actual strategy used by the experts in the domain and, more specifically, about the arguments underlying their strategy. For this purpose, we used an elicitation method that was composed of an unstructured interview to gain general insight in the test-selection strategy used, and a subsequent structured interview, simulating daily practice, in which full details were acquired. We used the method with two experts in our application domain and found that the method closely fitted in with their daily practice and resulted in a large amount of detailed knowledge.

1 Introduction

Decision-support systems are being developed for a wide range of domains. To support the reasoning processes in its domain of application, such a system often includes a strategy for selecting tests. In the medical domain, a decision-support system may, for example, suggest a sequence of diagnostic tests to be performed in order to reduce the uncertainty about a patient's true condition; the test-selection strategy thereby provides support for the task of diagnostic reasoning. In most decision-support systems, a strategy is offered in which tests are suggested sequentially, that is, on a one-by-one basis. The system then suggests a single test to be performed and awaits the user's input; after taking the test's result into account, the system suggests a subsequent test, and so on.

With the help of two experts in gastrointestinal oncology, we have developed, over a period of more than five years, a decision-support system for the domain of oesophageal cancer (Van der Gaag et al., 2000). Our system is based on normative principles and thus has a mathematical foundation in probability and utility theory (Fishburn, 1970; Jensen,

2001). The system includes detailed knowledge about tumour growth and about the processes of invasion and metastasis. It further captures knowledge about the various different diagnostic tests that can be performed to gain insight in the often hidden condition of a patient. The system moreover contains knowledge about the beneficial effects and complications to be expected from the different treatment alternatives. Building upon this knowledge, the system provides for establishing the stage of a patient's cancer and for prognosticating the likely effects of the various therapies, based upon the patient's symptoms, signs, and test results.

Our decision-support system for oesophageal cancer at present does not support the selection of diagnostic tests. For a specific patient, the attending physician orders a number of tests, based upon his or her own judgement, and simply enters the results into the system; the system does not provide the physician with information about which tests would be relevant and should be considered next for the patient under consideration. Building upon the system's mathematical foundation, however, a sequential test-selection strategy could easily be designed. Such a strategy would select, on a one-by-one basis, the test that is the most informative, for example in terms of entropy reduction, given the already available patient specifics (Andreassen, 1992; Doubilet, 1983). Upon working with our system, however, we noticed that our experts do not select tests one after the other, but in packages instead. We felt that a sequential test-selection strategy would be an oversimplification of our experts' problem-solving practice and we decided to design a test-selection strategy for our system that would build upon the arguments used by the experts for deciding whether or not to order specific tests. The resulting strategy would thus more closely fit in with the strategies for test selection currently used in the domain than a standard sequential test-selection strategy.

To acquire knowledge about the actual test-selection strategy employed by our experts and about the arguments underlying their strategy more specifically, we used an elicitation method that combined several different techniques for knowledge elicitation (Evans, 1988). The method consisted of two main interviews. The first of these was an unstructured interview that was aimed at providing insight in the overall strategy used by the experts. The second interview was a structured interview in which further details were acquired. In this latter interview, the experts' problem-solving practice was carefully simulated by means of cards, or vignettes, describing realistic patient cases. By simulating daily routine, we aimed to exclude, as much as possible, the various different biases that could possibly originate from the elicitation method used. We note that the idea of following up an unstructured interview by a structured one has been proposed before, for example in Cognitive Task Analysis (Schraages et al., 2000). We used the elicitation method with the two experts in our domain of application. We found that the method, and the use of carefully designed patient cases more specifically, closely fitted in with the experts' daily practice. The method resulted in a large amount of detailed knowledge, not just about the actual order in which tests are selected but also about the experts' reasons for ordering certain tests and for deciding not to order other ones.

Since a test-selection strategy offered by a decision-support system should support physicians in their daily problem-solving practice, we feel that for the design of such a strategy, knowledge about the actual strategies employed in the domain of application should be elicited from experts; a standard, sequential strategy may then turn out to deviate too much from daily routines to be acceptable. Our experiences in the domain of oesophageal cancer have demonstrated that the knowledge required for the design of a tailored test-selection strategy can be feasibly acquired: with our elicitation method, we were able to elicit the arguments underlying our experts' strategy in little time.

The paper is organised as follows. In Section 2, we provide some preliminaries on

oesophageal cancer and its therapies. In Section 3, we give an overview of the method that we used for eliciting our experts' test-selection strategy. In Section 4, we describe the results that we obtained from the first, unstructured interview. Section 5 reports on the second, structured interview. The paper ends with our concluding observations in Section 6.

2 Preliminaries

Cancer of the oesophagus may develop as a consequence of a lesion of the oesophageal wall, for example associated with smoking habits and alcohol consumption. The primary tumour typically invades the oesophageal wall and may in time invade neighbouring organs beyond the oesophagus. When the tumour has invaded lymphatic vessels and blood vessels, it may give rise to secondary tumours, or metastases, in lymph nodes and in such organs as the liver and lungs. The latter are called haematogenous metastases, while the former are referred to as lymphatic metastases. The depth of invasion of the oesophageal tumour and the extent of its metastases are indicative of the severity of the disease, which is summarised in the cancer's stage.

In order to establish the stage of a patient's oesophageal cancer, generally a number of diagnostic tests are performed. Various different tests are available, giving insight in different aspects of the cancer. A gastroscopic examination, for example, provides information about the presentation characteristics of the primary tumour, which include its length and its location in the oesophagus. A biopsy reveals the histological, or cell, type of the tumour. A laparoscopic examination of the liver, a CT-scan of the liver and of the lungs, as well as an X-ray of the lungs provide evidence about the presence or absence of haematogenous metastases. An endosonographic examination serves to yield information about the depth of invasion of the primary tumour into the oesophageal wall. The available tests differ considerably with respect to their reliability characteristics. Table 1 gives an overview of the tests, along with an indication of their sensitivity and specificity. For example, an X-ray of the lungs is stated to have a sensitivity of 0.85, which indicates that in 85% of the patients with lung metastases, the X-ray indeed reveals them. The specificity of the X-ray is 0.98, which indicates that in 98% of the patients without lung metastases, the X-ray will not show evidence of a secondary tumour in the patient's lungs. The sensitivity and specificity characteristics of a diagnostic test play an important role in the selection of tests in normative decision making, since these characteristics indicate how useful, or how informative, a negative or a positive result of the test actually is (Sox et al., 1988).

For patients suffering from oesophageal cancer, various different treatment alternatives are available. These alternatives include surgical removal of the primary tumour, administering radiotherapy, and positioning a prosthesis. Providing a therapy aims at removal or reduction of the patient's primary tumour to prolong life expectancy and to improve the passage of food through the oesophagus. The therapies differ in the extent to which these effects can be attained, however. The main goal of a surgical procedure is to attain a better life expectancy for a patient, that is, the procedure is curative in nature. Positioning a prosthesis in the oesophagus, on the other hand, is a palliative procedure that cannot improve life expectancy: it is performed merely to relieve the patient's swallowing problems. Radiotherapy can be administered in a curative regime, aimed at prolonging a patient's life, as well as in a palliative regime, aimed just at improving the patient's quality of life. The most preferred treatment in essence is to provide a curative therapy; of these curative therapies, a surgical removal of the primary tumour is preferred to a cura-

Table 1: An overview of the diagnostic tests that give insight in the stage of an oesophageal cancer

Test	*Sensitivity*	*Specificity*
Biopsy	1.00	1.00
Bronchoscopy	0.92	0.96
CT: liver, loco-region, lungs, organs, truncus coeliacus	0.48 – 0.90	0.88 – 0.98
Sonography: neck	0.90	0.95
Endosonography: loco-region, mediastinum, wall, truncus coeliacus	0.51 – 0.78	0.77 – 0.86
Gastroscopy: circumference, length, location, shape, necrosis	0.87 – 0.99	0.89 – 0.99
Laparoscopy: liver, diaphragm, truncus coeliacus	0.25 – 0.85	0.95 – 0.98
Barium swallow	0.87	0.99
X-ray	0.85	0.98
Physical examination	0.75	0.97
Interview: passage, age, weight loss	-	-

tive regime of radiotherapy. Providing a therapy, however, is often accompanied not just by beneficial effects but also by complications. These complications can be quite serious and may even prove to be fatal. The beneficial effects and complications to be expected from the different therapies for a specific patient depend on the general condition or health status of the patient, on the characteristics of his or her primary tumour, on the depth of invasion of the tumour into the oesophageal wall and neighbouring organs, and on the extent of metastasis of the cancer. If serious complications are expected for the patient, the attending physician may decide to abstain from providing a curative therapy and to administer one of the palliative treatment alternatives.

With the help of two experts in gastrointestinal oncology, we have developed, over a period of more than five years, a decision-support system that provides for assessing the stage of a patient's oesophageal cancer and for prognosticating the likely effects of the different treatment alternatives (Van der Gaag et al., 2000). The kernel of our system is a probabilistic network that captures the state-of-the-art knowledge about oesophageal cancer and its treatment. The diagnostic part of the network is reproduced in Figure 1; this part captures the knowledge about the various different diagnostic tests available and is of interest to the present paper. We would like to note that the results of a single test are often represented by a number of statistical variables in the network. For example, while a gastroscopic examination of the oesophagus is a single diagnostic test, its results are modelled by the five variables *Gastro-circumf*, *Gastro-length*, *Gastro-location*, *Gastro-shape* and *Gastro-necrosis*.

3 A Method for Eliciting Test-selection Strategies

Before describing the method that we used for eliciting test-selection strategies and before introducing the setting in which we used it, we briefly review some well-known elicitation methods.

The background

For knowledge acquisition, generally a distinction is made between methods that are aimed at eliciting object knowledge (knowing that) and methods with which to elicit

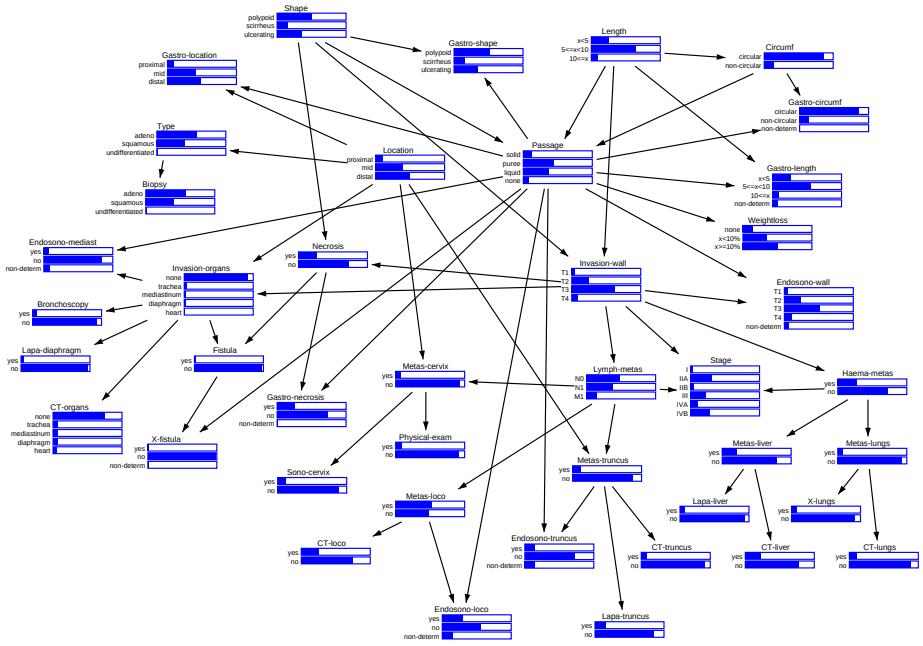


Figure 1: The oesophageal cancer network

problem-solving knowledge (knowing how). For finding out how our domain experts select diagnostic tests, we needed a method that focused on the latter type of knowledge. Several such methods are available, among which observation methods, methods for eliciting verbal reports, and think-aloud methods are the most well known (Schreiber et al., 2000; Van Someren et al., 1994).

Observation amounts to recording the behaviour that is exhibited by experts while solving problem situations in their domain. Studying the observed behaviour results in a so-called protocol of actions. This protocol can be analysed to infer the actual problem-solving strategies used by the experts. Observation methods are most suitable for domains in which problem solving requires objects to be handled or overt actions to be performed.

The different methods in use for eliciting verbal reports are all variants of the interview. An interview may be focused but otherwise unstructured, with general questions; the topics to be addressed during the interview but not the precise questions, are prepared in advance. The interview may also be structured, containing mostly closed questions. An interview may be held orally or presented on paper, in the form of a questionnaire. Interviews are especially appropriate for domains in which it is relatively easy for experts to verbalise their knowledge, for example because their daily routines involve verbalisations of problem-solving behaviour. Unstructured, oral interviews closely resemble normal conversation. The experts are for example asked what they commonly do when they are confronted with a problem situation in their domain of expertise. This type of interview provides the interviewer with a global understanding of the structure of the knowledge domain, and of the type of strategy used for problem solving; more specifically, the interview results in an overview of the order in which the various reasoning steps generally are performed. Subsequent structured interviews are suitable to deepen the understanding, to further clarify the structure of the knowledge domain, and to zoom in on the details of the problem-solving strategies used by the experts.

Thinking aloud is the only method available for eliciting mental processes directly. The experts are presented with a prototypical problem situation and are asked to verbalise their reasoning processes while solving the problem without interruption. Talking out loud concurrently with solving a problem situation leaves the experts no room or time for interpretation. They therefore directly reveal the strategies they use for solving the problem. Thinking aloud appears to be quite easy for most people and has been found not to interfere with their performance (Ericsson and Simon, 1993). Think-aloud sessions are generally tape-recorded and transcribed for further analysis.

To conclude we would like to observe that all knowledge-acquisition methods require careful preparation by the knowledge engineer. While some prior knowledge of the domain is advisable on the one hand, to understand the experts' answers and to put the right questions, too much knowledge on the other hand may cause the engineer not to listen carefully to the expert and to interpret answers in the light of his or her own views. Also, the interview questions and problem situations have to be prepared with care to guarantee that the knowledge aimed at is indeed elicited. Moreover, the setting in which the various sessions are to be held should be carefully selected. It should further be noted that, to the experts, some information or reasoning steps may be so self-evident that they are not verbalised, and consequently not acquired with any of the available methods. When analysing the acquired information, therefore, the knowledge engineer should be attentive to unjustified reasoning steps and prompt the experts for further elaboration.

The method

To acquire knowledge about the test-selection strategy used by our experts in the domain of oesophageal cancer, we employed an elicitation method that combined several of the methods reviewed above. Because the process of selecting diagnostic tests in essence is a type of problem solving in which the experts' behaviour is predominantly mental, the method of observation did not suit our purpose; with observation methods, we would not be able for example to gain insight in the experts' reasons for ordering specific tests. Also, thinking aloud with real, concurrent patients was not feasible, for evident reasons. We thus focused on interviews for eliciting verbal reports of problem-solving behaviour. We decided to conduct two consecutive interviews. The aim of the first, unstructured interview, was to obtain insight into the overall test-selection strategy employed and into the general arguments used by the experts. Since such an unstructured interview would be focused on the strategy and not on real patients, we were aware that we risked acquiring a general, text-book procedure rather than the experts' daily problem-solving routines. We decided therefore to follow up the first interview by a structured interview in which the experts were asked to think aloud while deciding, for a number of patients, which diagnostic tests to order. We felt that working with patient cases in a carefully conducted manner would closely fit in with the experts' problem-solving practice and would thus reduce possible biases from the elicitation method used. The aim of this second interview was to fill in details of the elicited test-selection strategy and, more specifically, of the arguments underlying the strategy. We decided not to work with historical patient cases, since the experts might recall these patients and let the real final outcomes influence their test-selection behaviour. We decided to employ fictitious patient cases instead, that were designed to be as realistic as possible. Working with fictitious patient cases brought the additional advantage that it allowed us to design cases with which we were able to explore the experts' decision boundaries.

The setting

The method for eliciting test-selection strategies outlined above was used with two ex-

perts from the Netherlands Cancer Institute, Antoni van Leeuwenhoekhuis, who are the last two authors of the present paper. These experts had been involved in the construction of the decision-support system for oesophageal cancer from its very inception. Also present during the interviews were three researchers from the Institute of Information and Computing Sciences of Utrecht University; they are the first three authors. The first author has a background in medical computer science, providing her with some knowledge about cancer in general and about the various diagnostic tests involved. The second author has a background in mathematics. She had constructed the decision-support system, for which she had held numerous interviews. The two domain experts and this interviewer thus were very well acquainted with one another. Over the years, moreover, the second author had gained considerable knowledge about oesophageal cancer and its treatment. The third author, to conclude, has a background in cognitive science and knowledge acquisition.

The two interviews were conducted at the Netherlands Cancer Institute, the home institute of the two experts. The first author conducted the interview, asking the questions that had been prepared. We felt that the second author, because of her accumulated knowledge about oesophageal cancer and its treatment, might unknowingly and unwillingly bias the experts in their answers. She therefore did not partake in the main interview and only asked the more elaborate questions about the experts' decision boundaries that emerged during the interviews. The third author recorded the elicited knowledge and monitored the elicitation process. She typed the words from the interviews in a laptop, not just concentrating on relevant knowledge but also on remarkable meta-phrases uttered by the experts. We were aware that typing in a laptop was likely to result in a less accurate recording of the elicited verbalisations than taping with a voice recorder. Still a laptop was used instead of a voice recorder because the experts had previously indicated that they would feel embarrassed by the recording. They did not seem to feel uneasy by the use of the laptop.

4 The First Interview

The first interview conducted with the two domain experts was an unstructured, oral interview. We briefly restate the main goal of the interview and the procedure followed, before presenting the results.

The goal

The goal of the first interview was to elicit general knowledge about the selection of diagnostic tests for patients suffering from oesophageal cancer. The main issues to be addressed during the interview were:

- Are the experts guided by a standard procedure for selecting diagnostic tests, or are they mainly guided by their own experience?
- Are the various tests performed in parallel or sequentially?
- Are some tests always performed together, or one after the other? For instance, is the biopsy always combined with a gastroscopic examination of the oesophagus?
- What are the criteria that the experts use for selecting tests?
- What are the experts' criteria to stop testing?

The procedure
For the interview, we prepared a small number of open questions. The main question was ”Can you describe the way in which you select and order diagnostic tests, starting from the very first consultation with a patient up to and including your final decision about the most suitable therapy?”. Since we wanted to avoid biasing the experts, we let them talk freely and did not interrupt unless it was strictly necessary, for example when further elaboration was desired. These interruptions then only consisted of open questions such as ”Why?” or ”Can you describe what you are thinking right now?”. The interview was conducted in the setting described in the previous section and took some 30 minutes.

The results
We found that upon first seeing a patient, the experts start with

- a physical examination of the patient;
- an interview with the patient, resulting in information about
  - the age of the patient;
  - the amount of weight loss suffered;
  - the patient’s ability to swallow food.

Subsequently, independent of the results of the physical examination and interview, a number of diagnostic tests are ordered simultaneously:

- a gastroscopic examination of the oesophagus, resulting in information about
  - the shape of the primary tumour;
  - the location of the tumour in the oesophagus;
  - the circumference of the tumour;
  - the length of the tumour;
  - the presence of necrosis (substantial decay of tissue);
- a biopsy, mostly performed together with the gastroscopy, revealing
  - the histological type of the primary tumour.

In the sequel, we will refer to the physical examination, the interview, the gastroscopic examination and the biopsy together as the *starting package of tests*. The gastroscopic examination and biopsy serve to give insight in the presentation characteristics of the primary tumour. The physical examination and the interview with the patient result in an assessment of the patient’s physical condition. We would like to note that, because our experts work at a highly specialised centre for cancer treatment, they generally see patients who are referred from regional hospitals where these tests have already been performed. Often, therefore, the test results are available. If the experts feel, however, that the tests were performed too long ago, they will order them to be performed anew.

After the results of the tests from the starting package have become available, the experts decide whether or not further testing is indicated. Patients with a very poor physical condition will now just receive highly palliative treatment, without having to undergo further testing. For all other patients, again a number of tests are ordered simultaneously:

- a CT-scan of the liver, lungs and thorax, resulting in information about

  - the presence of metastases in the loco-regional lymph nodes;
  - the presence of haematogenous metastases;

- an X-ray of the thorax, resulting in information about
  - the presence of haematogenous metastases in the lungs;
- a sonographic examination of the neck, providing information about
  - the presence of metastases in the lymph nodes in the neck;
- an endosonography of the local region of the primary tumour and of the mediastinum, giving insight in
  - the depth of invasion of the primary tumour into the oesophageal wall;
  - the presence of loco-regional lymphatic metastases.

These tests primarily serve to establish the extent of metastasis of the primary tumour. In the sequel we will refer to these four tests together as the *basic package of tests*. The tests from the basic package are again requested in parallel, but only after the results of the tests from the starting package have become available.

The remaining tests constitute the *extensive package of tests*:

- a bronchoscopy, resulting in information about
  - the depth of invasion of the primary tumour into the trachea and bronchi;
- a barium swallow with fluoroscopy, yielding insight in
  - the presence of a fistula (an open connection as a result of decay of tissue) between the oesophagus and the lungs;
- a laparoscopic examination of the liver, diaphragm and abdomen, resulting in information about
  - the depth of invasion of the primary tumour into the diaphragm;
  - the presence of haematogenous metastases in the liver;
  - the presence of metastases in the lymph nodes near the truncus coeliacus.

In contrast with the starting and basic packages of tests, not all tests from the extensive package are ordered just like that: one or more tests may be selected. Whether or not a specific test from the package is performed very much depends on the location of the primary tumour in the patient's oesophagus. If the tumour is located in the upper part of the oesophagus, a bronchoscopy and a barium swallow are performed to investigate whether or not the primary tumour has invaded the lungs. No laparoscopic procedures are performed, however, because the primary tumour cannot have invaded the diaphragm and, moreover, it is very unlikely that lymphatic metastases will be found in the upper abdomen. If the primary tumour is located in the lower part of the oesophagus, on the other hand, no bronchoscopy or barium swallow are performed. Laparoscopic procedures are then ordered, yet only if surgical treatment is considered.

To summarise, we found that diagnostic tests are ordered in three different packages: the starting package, the basic package, and the extensive package of tests. The tests from the starting package will always be ordered, for every patient. These tests are performed

in parallel. If a patient is in a very poor physical condition, the experts will then stop testing and provide highly palliative care. For all other patients, the basic package of tests is ordered as well. In addition, one or more tests may be chosen from the extensive package, dependent on the location of the tumour and on the most preferred therapy at that particular moment in the patient's management. Figure 2 presents a flowchart summarising the knowledge acquired from the first interview. In the flowchart, the grey boxes with rounded corners describe the beginning and end points of the experts' strategy. The white rectangular boxes capture requests to perform specific tests; the diamonds represent alternative choices and capture the appropriate questions to decide upon which tests to order. The arrows in the chart indicate the sequential order in which the various decisions have to be taken and choices have to be made.

A discussion

From the first interview, various distinguishing features of the test-selection strategy employed by our experts emerged. We found that diagnostic tests are not ordered sequentially, where the decision whether or not to order a specific test depends on the result of a previous test. Instead, the tests are ordered simultaneously, in packages. The most important argument underlying the experts' strategy of parallel testing is the loss of time that would be incurred by sequential testing. It may take several weeks before the results of a test become available. The tumour may have progressed within that time and may thereby render the results from earlier tests obsolete. Moreover, patients often are in such a poor physical condition that it is preferable not to have them return to the hospital too often for yet another test. And, even more importantly, the loss of time may make the difference between a curable cancer and an incurable one. As a consequence of ordering diagnostic tests simultaneously, however, more tests are likely to be performed than are strictly necessary. When questioned about such unnecessary testing, our experts indicated they did not see it as a problem, as the tests are not inconvenient for patients and gaining time is of primary importance:

"Ordering tests in packages is time saving. You might perform too many tests, but time is so important that you order all the tests anyway."

Our experts' strategy of ordering diagnostic tests in packages thus is supported by a strong argument. This argument in fact indicates that a sequential test-selection strategy for our decision-support system would be an unacceptable oversimplification of problem-solving practice. Our system should offer a strategy that is able to select tests in packages, based upon the argument reviewed above.

A second feature that we noticed of our experts' test-selection strategy pertains to the role of the stage of a patient's cancer. The experts had indicated before that they first establish the most likely stage for a cancer before deciding upon an appropriate therapy. We found, however, that a cancer's stage only very indirectly plays a role in the selection of diagnostic tests. The decision which tests to order appears in fact to be based upon the experts' current idea about the most suitable therapy for a patient rather than on the uncertainty about the stage of his or her cancer. For example, if the tests from the basic package reveal lymphatic metastases in distant lymph nodes, then surgical removal of the primary tumour is no longer a feasible treatment option for the patient under consideration. Invasive laparoscopic procedures for establishing the exact extent of the cancer's metastasis then are not performed, even though the stage of the patient's cancer is still uncertain. To establish the most appropriate treatment alternative for a patient, the experts appear to gather information that helps them weigh the beneficial effects and complica-

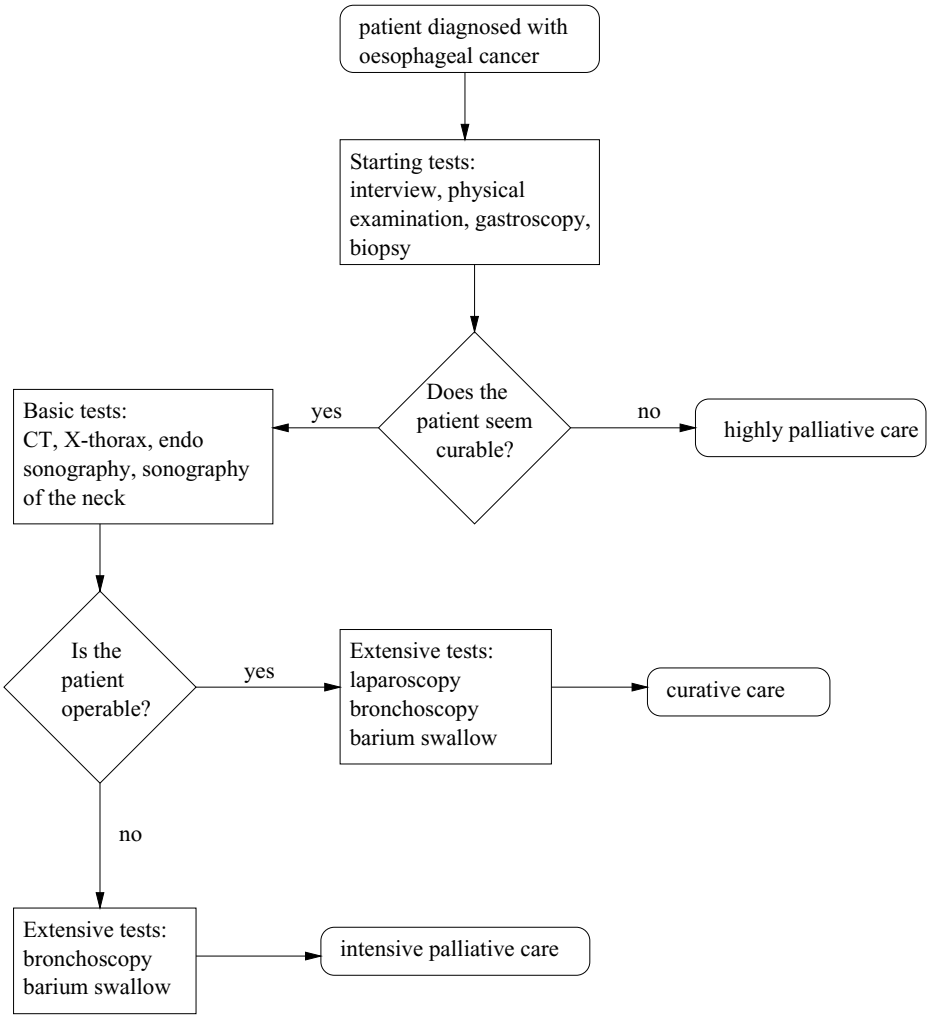


Figure 2: A flowchart summarising the knowledge from the first interview

tions to be expected for each alternative. Starting with the best possible alternative, that is to perform a surgical procedure, they order tests to see whether or not such a procedure is contra-indicated. Only if the test results indicate that a treatment alternative is not feasible, do they focus their attention on the next-best alternative. We note that over-treatment of a patient may easily prove to be fatal; under-treatment on the other hand may cause a patient to die prematurely from a cancer that might have been curable.

"You definitely do not want to miss anything. Missing and over-diagnosing are both bad. For physicians, however, missing feels worse than over-diagnosing as you want to do no harm. [...] When your feeling tells you something, but you cannot find it, you continue to look for it."

The test-selection strategy to be offered by our decision-support system should therefore take into account the impact of the possible results of a diagnostic test on the appropriateness of the available therapies rather than just its influence on the uncertainty of the most likely stage of a patient's cancer.

Closely related to the role of the stage of a patient's cancer is the role of the reliability characteristics of the various diagnostic tests in the test-selection strategy employed by our experts. As we have briefly mentioned in Section 2, the literature on medical decision making generally stresses the importance of taking the sensitivity and specificity characteristics of diagnostic tests into consideration upon test selection (Sox et al., 1988). We found, however, that these characteristics play no role with our experts, since they do not gather information to reduce their uncertainty about the most likely stage of a patient's cancer. They also appear not to take the informativeness of the remaining available tests into account when they decide whether or not to stop testing. Their stopping criterion for ordering tests instead seems to be their certainty about the most appropriate treatment alternative for the patient at hand.

5 The Second Interview

From the first interview we had gained general insight in the test-selection strategy used by our two domain experts. Building upon the acquired knowledge, we followed up on the first interview with a second, more focused interview. Once again we briefly restate the goal of the interview and the procedure followed, before presenting the main results.

The goal
The goal of the second interview was to fill in the details of the general test-selection strategy that had been acquired from the two experts during the first interview. More specifically, we aimed at eliciting the exact arguments used by the experts in their decisions to order a new package of tests or to refrain from further testing.

The procedure
For this second interview, we wanted to walk through the entire process of test selection for a specific patient, from the very first moment the experts see the patient up to and including the selection of the most suitable therapy. To this end, we designed a structured interview in which we carefully simulated the experts' daily problem-solving practice, by means of realistic patient cases. The experts were asked to think aloud while deciding for these patients which tests to order.

For the interview, we created eight fictitious patient cases; the specifics of these cases

Patient 1: An 87-year old male with a very poor physical condition and a large primary tumour.

Patient 2: A male of 76 years old who has a large primary tumour, yet is in good physical condition

Patient 3: A 57-year old male having a very small primary tumour who is in an excellent physical condition. The tumour is located in the upper part of the oesophagus (a proximal tumour).

Patient 4: This 60-year old male is in excellent condition and has a very small primary tumour located in the lower part of the oesophagus (a distal tumour).

Patient 5: A 64-year old male in good physical condition with a moderately-sized primary tumour. The patient has lymph node metastases in his neck.

Patient 6: This 67-year old male has a moderately-sized distal primary tumour and is in a good physical condition. He has both proximal and distal lymphatic metastases.

Patient 7: An 80-year old male in a very poor physical condition, having a very small primary tumour.

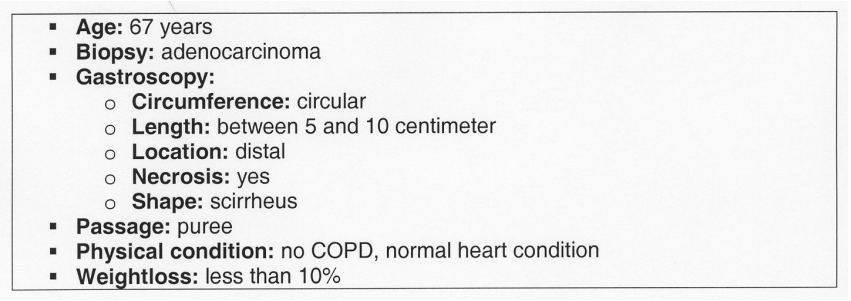
Patient 8: A 59-year old male with a large primary tumour, in a very good physical condition.

Figure 3: The eight patient cases designed for the second interview

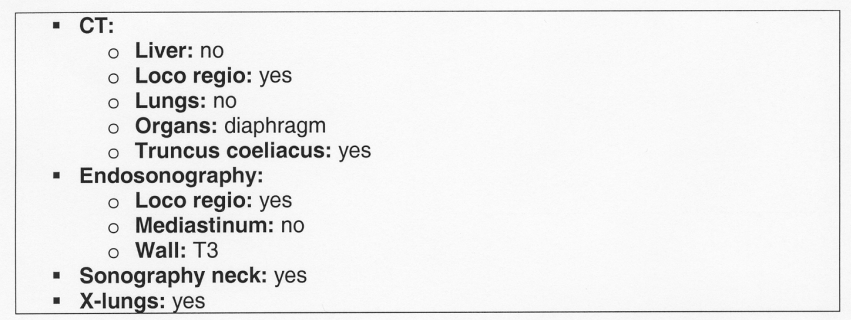
are summarised in Figure 3. Because we wanted to obtain as many details as possible about the test-selection strategy used by our experts, we created both patients for whom the most appropriate therapy seemed obvious and patients for whom the best therapy was not so evident. The first patient case mentioned in Figure 3 is an example of a patient for whom the most suitable therapy is quite evident. We expected that the experts would decide to administer highly palliative care for this patient. We further expected that the experts would not order the basic package of tests, nor any tests from the extensive package. For patient 6, however, it is not so clear what the best treatment alternative would be. The patient has a moderately-sized primary tumour and is in a very good physical condition. In fact, from the results of the tests from the starting package, the patient appears to be curable. We therefore expected that the experts would consider surgery and would order the basic package of tests. The results of the tests from the basic package reveal distant metastases. With this evidence and the primary tumour being distal, we expected that the experts would refrain from further testing and would decide to administer intensive palliative care.

The eight patient cases were carefully designed as illustrated above, by means of the flowchart that had resulted from the first interview. For each patient case, moreover, we prepared a small number of questions that allowed us to more closely investigate the experts' exact decision boundaries. For patient 7, for example, we expected that the experts would pronounce him to be incurable, mainly as a consequence of his poor physical condition. The patient's tumour, however, is very small and seems to be resectable. To investigate under which conditions the patient would no longer be deemed incurable, we prepared various what-if questions, such as "What would you do if the patient were just 58 years of age?" and "What would you do if the patient were in moderate health?".

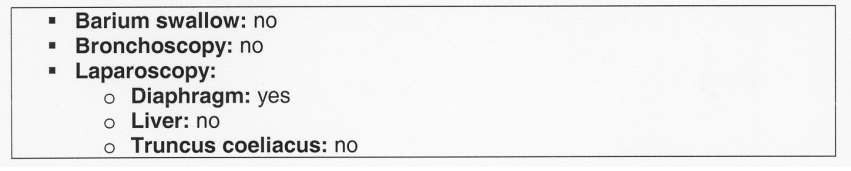
For each patient case we prepared three cards, or vignettes, with the results from the three different packages of tests. We created the three cards for every patient, also if it were very unlikely that even the basic package would be selected. The experts were



(a)



(b)



(c)

Figure 4: The three cards representing the results of the tests from the starting package (a), the basic package of tests (b), and the extensive package (c)

informed that there were three cards per patient irrespective of whether or not we expected them to request the second package of tests or order tests from the extensive package. On each card, a result was indicated for each test from the package under consideration. Since the experts might feel that the first test results on a card are the most important, we decided to simply list the results in alphabetical order. We informed the experts about this alphabetical order to avoid biasing their problem-solving behaviour. Since the tests from a single package would be ordered in parallel, we decided to present the results of these tests simultaneously, and not one by one. As an example, Figure 4 shows the three cards that we prepared for patient 6.

We asked the experts to discuss each patient case aloud. We asked them more specifically to verbalise their subsequent reasoning steps in ordering tests and to conclude the discussion of a patient case with an indication of the most appropriate therapy. We asked them to pretend that they were ordering real tests for real patients. For each patient, the card with the results of the tests from the starting package was presented first. Only when the experts indicated that they would order additional tests, would we show the second card with the test results from the basic package. Upon studying the second card, the experts also had access to the first card; they could thus survey the accumulated patient data. When still further testing was desired, the last card was presented. If the experts did not order any test from the extensive package or even from the basic package, the associated cards were not shown. The interview was conducted in the setting described in Section 3 and took approximately two hours for the eight patient cases.

The results

We present some fragments of the dialogue between the two experts while they were discussing patient 6, for whom the three cards shown in Figure 4 were created. Upon being presented with the first card, with the results of the tests from the starting package, the experts reasoned as follows:

"Oh, an average patient! We regularly see this type of patient. As his physical condition is quite good and the tumour is of moderate size, we might wish to consider surgery, so let's do the basic package of tests."

We presented the second card, with the results of the tests from the basic package:

"Mmm, there is some discrepancy here. Ah, well, we see them like this. Both proximal and distal metastases with a distal tumour. However, his condition is still quite good, so we would prefer to do something. We could consider palliative radiotherapy, also because he is not so old. [...] If the result from the sonography of the neck had been negative, there would only be metastases near the truncus coeliacus, which would make surgery a feasible option. Then, laparoscopy might be interesting. Now that the result of the sonography is positive, laparoscopy is no longer necessary."

The experts indicated that no further tests would be ordered.

The two experts discussed the eight patient cases at length. From the knowledge thus acquired, we constructed a new flowchart to capture the experts' test-selection strategy; the resulting flowchart is shown in Figure 5. To summarise, our observation from the first interview that diagnostic tests are ordered simultaneously in three different packages, was not contradicted by the second interview. Tests from the starting package are always performed. Only if a patient's physical condition is quite poor will the experts refrain from further testing and provide highly palliative care by positioning a prosthesis

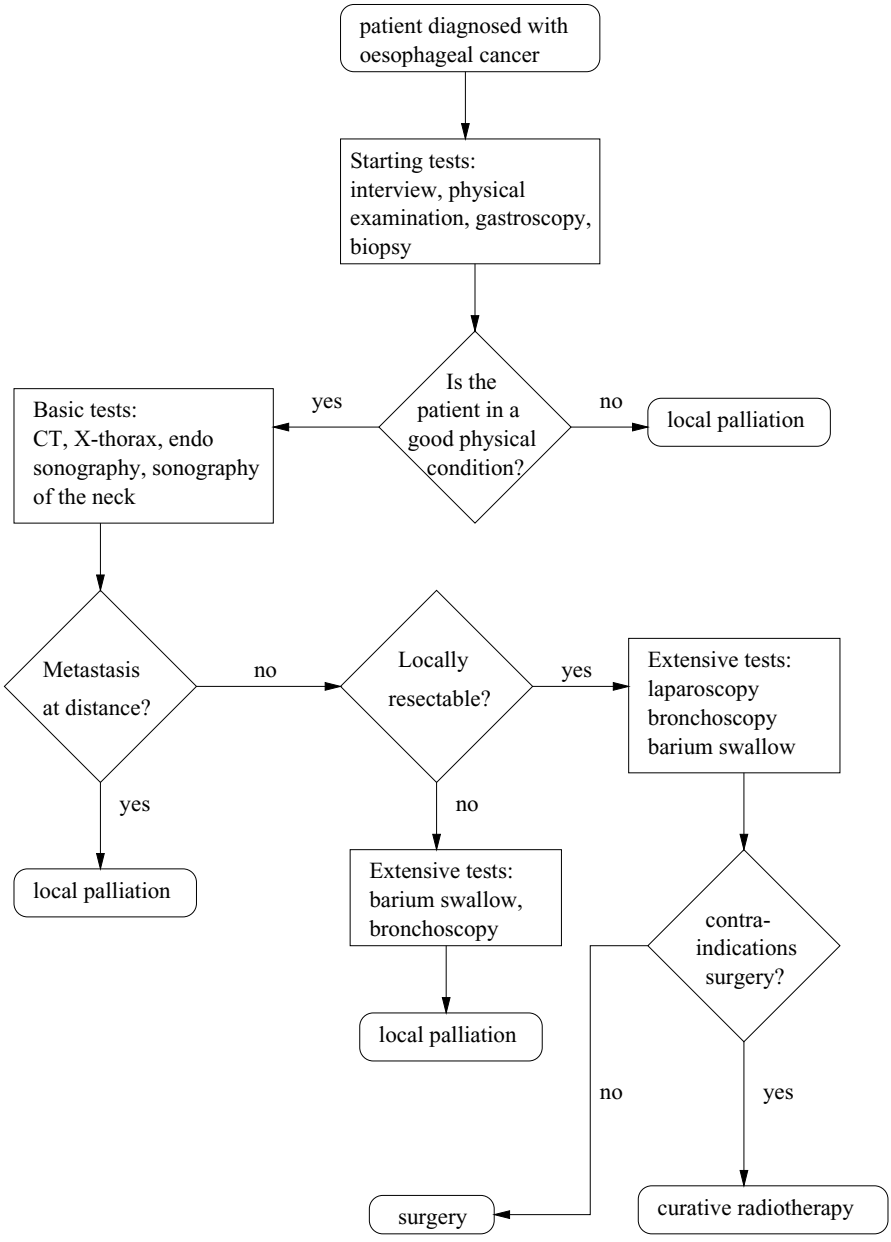


Figure 5: A flowchart summarising the knowledge from the second interview

in the patient's oesophagus. For all other patients, the basic package of tests is ordered as well. If the results of the tests from this package reveal distant metastases, the experts will not order any new tests to be performed and start with local palliative care, that is, either a prosthesis is positioned in the patient's oesophagus or a palliative regime of radiotherapy is administered. If the metastases of the primary tumour are loco-regional, on the other hand, further tests will be selected from the extensive package to investigate whether or not surgical removal of the oesophagus is a feasible treatment alternative. If the tumour appears to be resectable, the appropriate tests from the extensive package will be performed. If it is evident that the tumour is not resectable, or if contra-indications for surgery have been found, such as a relatively poor heart condition, the laparoscopic procedures will not be ordered.

A discussion
From the first interview we had learned various distinguishing features of the test-selection strategy employed by our experts. The second interview served to corroborate and further detail our previous observations. It provided additional insight especially in the way in which the experts used the results from the different tests as arguments for their subsequent decisions. The second interview further most prominently demonstrated that, from the very first moment of seeing a patient, the experts think in terms of appropriate treatment alternatives. In fact, our observations strongly suggest that it is the task of selecting an optimal therapy that drives the ordering of diagnostic tests, rather than the task of establishing the stage of a patient's cancer. For example, when discussing one of the patient cases, the experts mentioned:

"We are thinking of surgery as the best therapy right now. Let's see how deep the tumour has invaded into neighbouring organs to see if surgery really is an option."

6 Conclusions

Upon working with our decision-support system for oesophageal cancer, we felt that using the sequential test-selection strategies commonly proposed in the decision-making literature would be an oversimplification of our experts' daily problem-solving practice. We decided to acquire knowledge about the actual strategy used by the experts to provide for the design of a tailored test-selection strategy. For this purpose we used an elicitation method that was composed of two focused interviews: an unstructured interview, followed up by a structured one. With the first, unstructured interview, we found that tests were ordered not sequentially but in three different packages, with the tests from a single package ordered simultaneously. With the second, highly structured interview, we were able to fill in the details of the general strategy that we had elicited. More specifically, we were able to establish the different arguments underlying the experts' test-selection decisions.

With the structured interview, we carefully simulated the experts' problem-solving practice through the use of fictitious patient cases. Each patient case was captured by three different cards, or vignettes, with the results of the three packages of tests. These cards were presented sequentially to the experts. We found that our approach indeed closely fitted in with our experts' daily practice. In fact, they explicitly mentioned that using the cards was very intuitive:

"These cards and the way we discuss them are very similar to how patients are presented during the sessions we have with colleagues when we discuss patients."

The use of the cards thus worked quite well. The only difference with the experts' daily problem-solving practice may have been that in the current interview setting, not facing a real patient and with less time pressure, the experts were more consistent and more thorough in their decisions than they usually are. One of the experts mentioned:

"Perhaps we are now more consistent than we normally are in practice."

We felt that linking up with practice was highly advantageous for the purpose of acquiring knowledge of the test-selection strategy used by our experts. We would like to note, however, that especially for the set-up of the second, structured interview, prior knowledge appeared to be imperative. Without prior knowledge, we would not have been able to design the fictitious patient cases in a way that allowed us to explore the experts' decision boundaries.

To conclude, we feel that a test-selection strategy offered by a decision-support system should support physicians in their daily problem-solving practice and should therefore be based upon the argument experts use in their decisions to order specific tests. We feel that to design such a strategy, knowledge about the actual test-selection strategy used should be elicited from experts in the domain of application. A standard sequential strategy may then turn out to be unacceptable to the physicians who are the projected users of the system. In this paper, we have demonstrated that eliciting test-selection knowledge from experts can indeed be feasible and is likely to result in a wealth of detailed information that can provide for a carefully tailored test-selection strategy.

Acknowledgements

The research of the first two authors was (partly) supported by the Netherlands Organisation for Scientific Research (NWO). We are most grateful to Eveline Helsper and Silja Renooij for their useful comments on an earlier draft of this paper.

References

Andreassen, S. (1992). Planning of therapy and tests in causal probabilistic networks. *Artificial Intelligence in Medicine*, 4:227 – 241.

Doubilet, P. (1983). A mathematical approach to interpretation and selection of diagnostic tests. *Medical Decision Making*, 3(2):177 – 195.

Ericsson, K. A. and Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data, revised edition*. MIT Press, Cambridge.

Evans, J. S. B. T. (1988). The knowledge elicitation problem: a psychological perspective. *Behaviour and Information Technology*, 7(2):111 – 130.

Fishburn, P. (1970). *Utility Theory for Decision Making*. John-Wily & Sons, New York.

Jensen, F. V. (2001). *Bayesian Networks and Decision Graphs*. Statistics for Engineering and Information Science. Springer, New York.

Schraages, J. M., Chipman, S. F., and Shelin, V. L., editors (2000). *Cognitive Task Analysis*. Lawrence Erlbaum Associates, NJ.

Schreiber, A. T., Akkermans, J. M., Anjewierden, A. A., de Hoog, R., Shadbolt, N. R., Van de Velde, W., and Wielinga, B. J. (2000). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press, Massachusetts.

Sox, H. C., Blatt, M. A., Higgins, M. C., and Marton, K. I. (1988). *Medical Decision Making*. Reed Publishing, USA.

Van der Gaag, L. C., Renooij, S., Aleman, B. M. P., and Taal, B. G. (2000). Evaluation of a probabilistic model for staging of oesophageal carcinoma. In Hasman, A., Blobel, B., Dudeck, J., Engelbrecht, R., Gell, G., and Prokosch, H.-U., editors, *Medical Infobahn for Europe, Proceedings of MIE2000 and GMDS2000*, pages 772 – 776. IOS Press, Amsterdam.

Van Someren, M. W., Barnard, Y., and Sandberg, J. (1994). *The Think Aloud Method – a Practical Approach to Modelling Cognitive Processes*. Academic Press, London.