Virtual Facework Trainer: Use of Offendable Bots for Learning Cross-Cultural (Im)Politeness

Ronald M. Lee
Florida International University
r.lee@fiu.edu

Elizabeth Dominguez Campillo
Universidad de la Habana
lisydc69@yahoo.es

Gregory Diaz
Florida International University
gregdiaz11@gmail.com

Abstract

This project focuses on artificial social interactions where things get nasty and mean. The purpose is training in social 'facework' -- managing the situation so that participants maintain their social dignity or 'face'. This can be especially delicate in cross-cultural contexts, where assumptions about social protocols and the emotional associations of utterances and gestures may differ. The purpose of this project is two-fold. First, it is intended as a training system, so that users might learn the do's and don'ts of social interactions in different cultures and different situations. The knowledge base draws from existing theories of diplomacy, facework, and (im)politeness theory. The other goal is to provide a platform for observation and experimentation of social interaction in an artificial, virtual setting in order to improve these theories.

Modeling Emotional Behavior

Artificial intelligence, along with economics, has mainly focused on modeling, and perhaps improving, rational thinking. Yet, increasingly, emotions are recognized as an important aspect of human cognition. Emotions provide an important link between thought and action, what Frijda (1986) calls action tendencies. Emotions propel us to action, and can only be restrained with effort.

In psychology, despite decades of research about emotions, there is still little agreement about what, exactly, an emotion is; how many different emotions there are; and whether some cultures have different emotions than others (e.g. the Japanese "amae", indicating a special affection for caregivers). One theory of emotion, the dual process model, has gained acceptance, and has certain intuitive appeal. Essentially, this proposes dual emotional functionality between a low level ('reptile') brain, and higher cognition, which appraises emotions (Scherer, Shorr, Johnstone, 2001). LeDoux (1998) has demonstrated dual emotional processing in the case of fear[1]. We can react to a fear stimulus (e.g. snakelike movement in the grass), before we become consciously aware of the stimulus. This rapid response mechanism has obvious survival advantages. We jump away first, and then later realize it is only a garden hose.

In this research, we are especially concerned about the emotions that relate to maintaining 'face' or respect in social interactions. These aspects are perhaps more consequential in their breach: an act of disrespect can have negative effects on the relationship lasting years. An important element of disrespect is the anger that it can cause. Verbal acts of disrespect are captured in the notion of insult.

Anger

In taxonomies of 'basic' emotions, the second most popular (after fear) is anger. Anger seems to be primitive emotion that we share with many animals. However, in civilized society, anger is usually regarded as dysfunctional. It is an intense emotion that can overpower rational control of behavior (Tavris, 1989). Whereas many emotions, such as being sad or happy, can be felt in solitude, anger is social: one is (usually) angry *at* someone else. Indeed, being angry at an object, such as a car, is considered rather irrational. However, anger at animals, such as a dog or horse, is plausible, assuming they have misbehaved.

This work was partially supported by a grant from CIBER, FIU.

[1] see his clever animation at www.cns.nyu.edu/ledoux/

This suggests that anger is somehow caused by the behavior of others, somehow in violation of our expectations.

Insults

Insults are a truly remarkable kind of speech act. By simply uttering a word or two (or perhaps just raising a finger), one can almost instantly transform the emotional state of another person. Speech acts are usually analyzed for their illocutionary effects -- the change in social state brought about by their utterance (Kimbro, Lee, Ness, 1984). Insults are noteworthy for their perlocutionary effect, the emotional effect they stimulate in others. Also different from many other kinds of speech acts, insults seem to have a greater potential for accidental performance. For instance, just using the wrong form of address (tu versus vous; mister versus doctor) may be an unintended insult. When communicating across cultural boundaries, the potential for unintended insults is greatly increased. In one of the very few works specifically about insults, Jerome Neu in his book *Sticks and Stones* (2007) observes that there is a difference between being insulted and feeling insulted. This is essentially the distinction between the illocutionary and perlocutionary effects of insults. To be insulted is socially defined, whereas to feel insulted is the anger resulting from a perceived slight. Here we are primarily interested in better understanding the perlocutionary effects of insults -- what it is about the situational context, roles and personal attributes of the parties, and the linguistic form of the utterances, that are so emotionally provocative.

For example, social interactions and situations have an implicit level of formality, sometimes called the linguistic register (Joos, 1967). For example, utterances acceptable in a bar might be shocking in a college classroom. This also applies to relationships: we have certain ways of speaking with certain classes of people. Downward violations of register sound crass; upward violations might sound arrogant.

Generalizing, people seem to project expectations about how they should be respected by certain other people. When the expectations are not met, they may feel insulted. These can range from broad, cultural rules (e.g. proxemics or body distance during conversation); institutional (e.g. military protocols of saluting); or idiosyncratic (e.g. my boss is very touchy about being overweight).

Aim of Virtual Facework Trainer Project

The aim of this project is to develop a simulator for bots that are culturally 'fragile' in the sense that they take offense easily. Furthermore, their anger can escalate, if they are not treated properly. However, they have one important redeeming grace: when asked, bots will explain to you why they are angry.

The theoretical objective of this project is essentially a situational theory of facework (or politeness / impoliteness). Every culture has certain do's and don'ts, some of which can be learned from books such as the series *Kiss, Bow and Shake Hands* (Morrison and Conaway 2006). But this is essentially rote learning. What we are after is deeper understanding of perlocutionary effects of cross-cultural communications. Our method relies on a simulation platform to test theories about anger escalation and other aspects of cross-cultural facework.

The practical importance of this project is as a training tool for in cross-cultural facework -- especially in situations where there is distrust, potentially even hostility. An example might be a business person venturing into foreign circumstances where there has been little or no cultural contact. Another example might be emergency aid workers suddenly entering from abroad, e.g. the tsunami in Indonesia, or the earthquake in Haiti. Another context is military interventions where soldiers need to interact with distrustful civilians.

Related Work

Emotions are of tremendous importance for video games. Indeed, the success of the game largely depends on the nature, intensity, and duration of the emotions it creates in the players. Freeman (2003) calls game design for emotional response 'emotioneering'. Bateman (2011) ranks the top ten emotions experienced by game players as bliss; relief; 'naches' (pride of accomplishment of one's children or students); surprise; 'fiero' (triumph over adversity); curiosity; excitement; contentment; amusement. Three emotions least often experienced in gaming were the negative emotions of sadness; guilt; and embarrassment. From a sales/marketing standpoint, it is not surprising that games avoid promoting negative emotions. It is also interesting that several of the emotions in this list are not normally found in psychological taxonomies of emotion. Perhaps games succeed, at least partly, by producing new kinds of emotional experiences.

Our interest is somewhat different. Our primary goal is not so much to provoke emotional reactions in the user, but rather to develop a social interaction simulator to help users to better understand emotional responses from other people, and in particular, how to avoid angry, aggressive responses. The purpose of the system is to facilitate understanding of how one's social interactions affect the emotional responses of others. Terms for this include diplomacy; facework; politeness and impoliteness. We call this kind of system a *facework trainer*.

In our search of the literature, we have found only one other project that has a similar objective: the system is called "FearNot!" (VICTEC 2005). The purpose of

FearNot! is interactive training for children (8-12) about how to deal with school bullys: "to enable children to explore physical and relational bullying issues and coping strategies through empathic interactions with synthetic characters." The technical details of this project are well described in the MS thesis of João Dias (2005). Emotional modeling in FearNot! in turn is based on earlier developments, especially by Gratch and Marsella (2003), who provide a model for continuous emotional planning, which differentiates behaviors according to the character's personality.

Facework, Politeness, Insults

A central concept for the study of conflict and aggression is the notion of face. Face is essentially one's self-esteem or dignity. It is internal emotional support that protects oneself in a social situation. Normal, civilized behavior seeks to avoid conflict stemming from acts of disrespect. Our efforts to avoid threatening the face of others have been dubbed "facework", originally by Erving Goffman (1963). Facework is a definite social skill (Cupach and Metts, 1994). Cultures have specific rituals for maintaining face, e.g. protocols of offering, acceptance and thanks. Tact consists of ways to avoid embarrassing moments of losing face. "Trouble is caused by a person who cannot be relied upon to play the face-saving game" (Goffman, 1963).

In the lingo of street talk in the USA, to "dis" someone is to disrespect them. This is a face-threatening act. Preserving face is important in practically all cultures and sub-cultures. Stella Ting-Toomey has studied how facework strategies differ between national cultures (Ting-Toomey, 2005). According to her face negotiation theory, an important cultural variable is individualist vs collectivist. In individualist cultures, such as the United States, Germany, and Great Britain, there is great value on personal rights, so that face issues are primarily about the individual. Collectivist cultures such as Japan, Saudi Arabia, and Colombia, place more value on the face of the group.

Another important cultural variable, according to Ting-Toomey, is power distance. This refers to the degree and importance of hierarchy in the culture. Low power distance cultures such as the USA value equality, whereas high power distance cultures, such as Japan, place high value and respect on authority. In the latter case, maintaining the face of a superior is potentially more important than maintaining one's own face.

Closely related to theories of facework are theories of politeness and impoliteness. The book *Politeness* by Brown and Lewinson, (1978/1987) established politeness theory as an important branch of socio-linguistics (Leech, 1983). Impoliteness theory is an offshoot from politeness theory, but it is not simply a negation of the politeness rules. According to Culpeper, Bousfield, and Wichmann (2003, p 1546) "impoliteness fulfills a function that the speaker intended, and is not simply failed politeness". To give a flavor for impoliteness analysis, these authors remark (p. 1549):

> *You have shit for brains* could be interpreted as very impolite for several reasons: (1) the criticism is personalized through the use of you (this is, of course, also true of You fool), (2) shit is a taboo word, and (3) the speaker flouts Grice's maxim of quality, in order to implicate the impolite belief that the target has absolutely no intelligence.

Rules that capture these aspects allow bots to formulate their own impolite responses.

Modeling Anger and Aggression

As per the relevant literature (e.g. *Handbook of the Affective Sciences,* Davidson, Scherer, Goldsmith 2003), anger is an internal emotion which may or may not be displayed in observable behavior. Following Frijda (1986), emotions have 'action tendencies' -- they are primers for certain kinds of actions. For instance, the action tendency of fear is to flee. The action tendency of anger is aggression. As the intensity of the emotion increases, ever greater effort is required to suppress the action tendency.

In addition to such action tendencies, human emotions differ from animal emotions in that they have a strong cognitive component, called *appraisals* (Scherer, et al. 2001). For instance, that a certain utterance is construed as foul and offensive is something that is socially learned. (Your cat is not offended.) One might appraise the smell of gas during the night as something fearful. Again, the cat does not make the connection, though it certainly detects the smell.

Dynamics: Escalation vs. Self-Control

An intuition of folk psychology and a theme in many movies is that anger intensifies, until it overwhelms self-control and violent rage ensues. However, this intensification process can also be interrupted, sometimes successfully (e.g. the police arriving). There is also a problem of so-called displaced aggression, where the person interrupting becomes a surrogate target. Thus, domestic violence situations may be quite dangerous for the police that intervene.

Escalation of Anger

While there is a great deal written about anger, there is surprisingly little to be found about how anger escalates. Zillmann (1994) is one of the few social psychologists who have considered it with some detail. He comments (p. 50):

> Escalating conflict can be conceptualized as a sequence of provocations, each triggering an excitatory reaction that materializes quickly and that dissipates slowly. If a second sympathetic reaction oc-

curs before the first has dissipated, the second reaction combines with the tail end of the first. Moreover, if a third reaction occurs before the second and first reactions have dissipated, this third reaction combines with the tail ends of both earlier reactions. In general, any excitatory reaction to provocation late in the escalation process rides the tails of all earlier excitatory reactions.

Self-Control

Several researchers, notably Baumeister and colleagues (e.g. Baumeister and Heartherton, 1996), have done research into the mechanisms of self-control. Relating back to the emotion-appraisal/aggression-response model suggested earlier, self-control is exercised when one or more provocations lead to a strong sense of anger, but the individual resists the action-tendency towards aggression. Various empirical studies by Baumeister and colleagues have shown that such self-control is in limited supply, and behaves something like muscular energy. Like a muscle, self-control becomes depleted with repeated use in a brief period. Replenishment of self-control is accomplished either by rest or by a strong dose of positive affect (e.g. laughter).

Emotional Interaction Model

Figure 1 reflects the cognitive/emotional process model of a single bot. We want to emphasize that this is one possible model -- the framework will allow simulation of alternative models as well. An emotional interaction takes place between two or more bots. In a training situation, one of these roles might be taken by a human-operated avatar. But for the sake of discussion, let us assume two bots. Thus there would be two emotional process models that exchange speech acts.

By speech act we mean either something that is said or a gesture that has symbolic importance, such as slamming a door in a way that expresses anger. Typically, a conversation is a series of 'turns' where one party says or does something, and followed by the next party saying or doing something. The schematic in Figure 1 is of a single turn in an emotional conversation. The symbols in this figure use a kind of Petri net notation, where the boxes are transitions, and the ovals are places. The hexagons are also places, but are regarded as fixed, for the scope of the interaction. Thus, culture, one's own personality, the other's personality, one's values, etc. are regarded as fixed during the interchange, though they might of course change over a longer period.

The process begins as an utterance is received as a speech act. This is appraised according to culture rules and aspects of one's own personality. The result is an emotion. Emotions may decay over time. The emotion is input to an impulse generator, which incorporates action tendencies associated with emotions (e.g. anger creates the action tendency to attack). These action tendencies may be tempered by one's values, and politeness principles that apply in the current situation, as well as consideration of the personality of the other (e.g. will they return the attack? maybe pull a gun?). Another input to the action regulator is self-control, which may deplete. The result is a response speech act (or possibly a physical action).

In face-to-face exchanges between humans, non-verbal aspects may be very important, especially with regard to impoliteness and the determination if a speech act is insulting (Culpeper, et al. 2003). These include facial expressions, tone of voice, gestures, and other forms of body language. Givens (2012) provides a marvelous inventory of these non-verbal aspects. In a computational setting, a partial incorporation of such non-verbal aspects could be included in a dialogue setting between a bot and a user, with a video camera, using facial expression recognition software (Visual Recognition, 2012). Indeed, this software has already been applied in SecondLife, whereby the avatar imitates the user's facial expressions.

Target Functionality

Offendable Bots

Our target functionality for this project is what we call 'offendable' bots. We assume the bot engages the user in a conversation, about some potentially provocative subject. As is usual in conversations, there is a sequence of turn taking -- first one party speaks then the other party speaks. The view of a single round in this turn from the bot perspective is shown in Figure 1. As indicated, the bot's reaction to a speech act from the Other involves two sub-processes. The first is the appraisal, which results in an emotion. The second process is the bot's response. Both appraisal and response have inputs of culture and personality. These inputs are parametric in that they do not change during the course of the conversation.

In refining this model, we want to show how the intensity of the bot's angry emotional response can increase with each iteration. (Or maybe not, if the Other does certain anger reducing moves such as apologies.) Self-control is exercised in the behavioral response module. As discussed earlier, this is a depletable resource. If self-control is diminished and anger continues to escalate, there will be a certain threshold when the bot explodes into a violent reaction. Thus, a number of gaming scenarios are possible to try and achieve some kind of diplomatic objective without the bot exploding in rage.

Explaining Anger

Artificial emotion, like artificial intelligence more generally, is based on rules. As noted above, the variables in these rules may include aspects of culture and personality, as well as emotion-provoking attributes of the situation and history of the interactional dialogue. The rule struc-

ture of these bots affords a feature not often found in actual angry aggression: the bot can be asked to *explain* why it has a certain angry emotion. This explanation feature is extremely important for training as well as for theoretical tuning of the anger model.

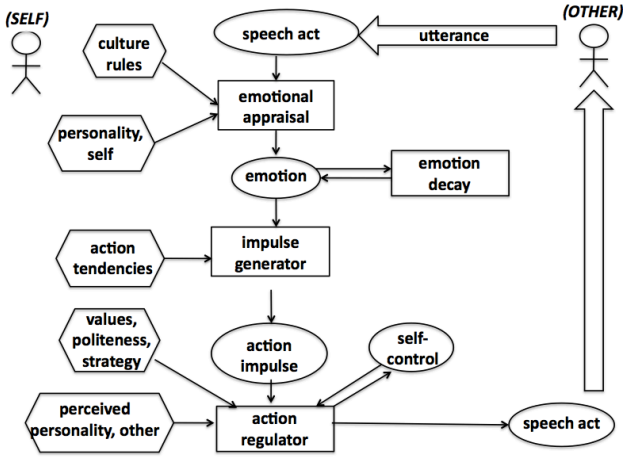

Figure 1. Emotional Interaction Process

Language Interface

We have not yet described how the user will interact with these bots. Actually this aspect has been addressed in a previous project, developed as a system called ChitChat[2]. ChitChat provides a grammar driven interface for users to compose sentences in a foreign language. It is designed to communicate with language bots that also use the same grammar as the composition interface. Thus, ChitChat supports semi-structured natural language interaction in specific problem domains. In the design of offendable bots, we also assume that the user will interact with the bot in the target foreign language. This offers the possibility that words and gestures that are innocent in one's home culture (e.g. the American OK sign gesture) may be quite offensive in the host culture (such as Brazil).

Another approach would be to parse the user's written input and identify emotional words. This strategy has been explored for English by Neviarouskaya, Prendinger, and Ishizuka (2010). This strategy offers generality where in cases where parsers for the language exist. However, the ChitChat approach serves well for special purpose dialogues, and other languages where parsers are not available.

Implementation Aspects

The prototype is still in development. In its ultimate form, we want it to be a kind of 'shell' for representing different socio-emotional process models. The input would be a graphics interface something like that in Figure 1, with input as a stylized Petri net with primitives appropriate to emotion process modeling[3]. The output would be a set of Prolog routines, that would control the behavior of the bot. Prolog is especially useful for customizing syntax for rule-based shell languages. Additionally, it also includes rule representation for definite clause grammars, which facilitates the modeling the conversational interactions among the bots.

We are experimenting with two virtual world platforms for this project: SecondLife[4] and OpenSim[5]. SecondLife is quite popular, with millions of residents. However, with SecondLife, one does not have direct access to the server except via their rather limited scripting language. With OpenSim, the server is also open source, so deeper modifications are possible.

As noted earlier, our special interest is in dialogues that affect the emotions of the participants. One model we developed focuses on just-in-time language and culture training for emergency response workers for the Haiti earthquake. In Haiti, while French is usually understood in the major cities, people in villages and rural areas often only understand Haitian Creole. Furthermore, the foreign medics represent a medical paradigm that may be foreign and even frightening to some locals, who may be accustomed to vodou (voodoo) medicine. In Figure 2, the little girl (Mai) is actually a bot, who interacts based on a grammar fragment for Haitian Creole (see Lee, 2012, for more details about grammar modeling).

Figure 2. Avatars and Bot in Virtual Haiti

Comparison to Other Systems

There are a couple of 'serious games' applications that focus on foreign language and culture training. One of

[2] http://web.me.com/ronald_m_lee/ChitChat/

[3] We have prior experience developing a graphics tool for Petri net process modeling, but in a quite different application area: international trade procedures, e.g. Lee (1999).

[4] www.secondlife.com

[5] http://opensimulator.org

these is Alelo's Tactical Iraqi[6], which is used for Arabic language training for the military, and highlights cultural sensitivities that may arise in soldiers' encounters with Iraqi locals. Another interesting application, by Serious Games Interactive[7], is their Global Conflicts games: Palestine and Latin America, which uses the scenario is of a journalist trying to uncover ethical abuses. Both of these systems provide excellent graphic representations of the local environment and especially the appearance and mannerisms of local people. However, in neither case is the underlying socio-emotional process model open to inspection or able to provide explanation of behavior. Moreover, these are sealed gaming applications -- the process model that drives character behavior is not modifiable.

Remaining Mysteries

George Box (1979) remarked "All models are wrong; some are useful." The process model presented here is also wrong. There are certainly other aspects of insult / disrespect that it does not (yet) include. At present, we can model various speech act behaviors such as name calling, lowered language register, obscene gestures, etc. that provoke anger in the hearer. This by itself has value for training purposes, especially for cross-cultural applications where learning facework symbolism of behaviors is extremely important (Morrison, Conaway, 2006).

Beyond this, targeting anger as the key emotion triggered by disrespect seems inadequate. Anger by itself dissipates, yet certain forms of disrespect can endure years[8]. This suggests an interplay between social cognition and emotion, leading perhaps to such social emotions as guilt, embarrassment and shame, which are not yet addressed in our current model. For this reason, we are attempting to make the system as open as possible for future modifications and refinements.

References

Bateman, C. 2011. Only a Game: Top Ten Videogame Emotions. (viewed Nov, 2011) [http://onlyagame.typepad.com/only_a_game/2008/04/top-ten-videoga.html]

Baumeister, R. and Heatherton, T. 1996. Self-Regulation Failure: An Overview. *Psychological Inquiry* (7:1) pp 1-15.

Box, G E.P. 1979. Robustness in the strategy of scientific model building, in Robustness in Statistics, R.L. Launer and G.N. Wilkinson, Editors. Academic Press.

Brown, P. and Lewinson, S. 1978/1987. *Politeness -- Some Universals in Language Usage*, Cambridge University Press.

Conley, T. 2010. *Toward a Rhetoric of Insul*t. University of Chicago Press.

Culpeper, J., Bousfield, D., Wichmann, A. 2003. Impoliteness Revisited: with Special Reference to Dynamic and Prosodic Aspects. *Journal of Pragmatics* 35: 1545-1579.

Cupach, W., Metts, S. 1994. *Facework*. Sage Publishing.

Davidson, R. Scherer, K., Goldsmith. H.H. 2003. *Handbook of the Affective Sciences*, Oxford University Press.

Dias, J. 2005. FearNot!: Creating Emotional Autonomous Synthetic Characters for Emphatic Interactions. MS Thesis, Universidad Técnica de Lisboa.

Elliot, C. (1992). The Affective Reasoner: A process model of emotions in a multi- agent system. Northwestern University, PhD Thesis.

Freeman, D. 2003. *Creating Emotion in Games The Craft and Art of Emotioneering*, New Riders Games Publishing.

Frijda, N. 1986. *The Emotions*. Cambridge University Press.

Givens, D. 2012. NonVerbal Dictionary of Gestures, Signs and Body Language Cues. [http://center-for-nonverbal-studies.org/6101.html]

Goffman, E. 1963. "On Face-Work." *Interaction Ritual* New York: Anchor Books.

Gratch J., Marsella S. 2003. Modeling Coping Behavior in Virtual Humans: Don't Worry, Be Happy. In Proceedings of Second International Joint Conference on Autonomous Agents and Multiagent Systems, Australia, 2003.

Gudykunst, W. 2005. *Theorizing About InterCultural Communication*. Sage Publications.

Joos, M. 1967. *The Five Clocks*. Harcourt.

LeDoux, J. 1998. *The Emotional Brain*, Touchstone.

Lee R. "Distributed Electronic Trade Scenarios: Representation, Design, Prototyping", *International Journal on Electronic Commerce* Special Issue on Formal Aspects of Digital Commerce, eds. S. O. Kimbough and RM Lee, Vol 3, No 1, 1999, pp. 23-26.

Lee, R. 2012. ChitChat. Guided Grammatical Conversation in a Foreign Language. (viewed Feb, 2012) [http://web.me.com/ronald_m_lee/ChitChat/index.html]

Leech, G. 1983. *Principles of Pragmatics*. Longman Group Publishing.

Martinho C. (1999). *Emotions in Motion.* Universidade Técnica de Lisboa, MSC Thesis.

Morrison, T. and Conaway, W. 2006. *Kiss, Bow, or Shake Hands*, Adams Media.

Neviarouskaya, A., Prendinger, H., Ishizuka, M. 2010. Affect Analysis Model. *Natural Language Engineering*, 17(1): 95-135.

Neu, J. 2007. *Sticks and Stones The Philosophy of Insults*, Oxford University Press.

Ortony A., Clore, G., Collins, A. (1988). The Cognitive Structure of Emotions. Cambridge University Press, 1988.

Pool, J. 2011. C#Prolog. [http://sourceforge.net/projects/cs-prolog/]. (accessed November, 2011).

Scherer, K. R., Shorr, A., and Johnstone, T. (eds.). 2001. *Appraisal Processes in Emotion Theory, Methods, Research*. Oxford University Press.

Tavris, C. 1989. *Anger The Misunderstood Emotion*. Touchstone Press.

Ting-Toomey, S. 2005. The Matrix of Face: An Updated Face-Negotiation Theory. In W.B. Gudykunst (ed.), *Theorizing About Intercultural Communication* (pp. 71–92). Sage.

VICTEC. 2005. Virtual ICT with Empathic Characters: FearNot! (viewed November, 2011) [www.macs.hw.ac.uk/EcircusWeb]

Visual Recognition. 2012. eMotion Software Now Integrates with Second Life (viewed February, 2012) [http://staff.science.uva.nl/~aldersho/isis/VisualRecognition]

Zillman, D. 1994. Cognition-Excitation Interdependencies in the Escalation of Anger and Angry Aggression. in Potegal, Michael and Knutson, John (eds) *The Dynamics of Aggression Biological and Social Processes in Dyads and Groups*, Psychology Press.

[6] www.alelo.com/language_culture.html
[7] www.globalconflicts.eu
[8] There is a Vietnamese saying: "For revenge, ten years is a short time."