HUSSO 2015

The First International Conference on Human and Social Analytics

ISBN: 978-1-61208-447-3

October 11 - 16, 2015

St. Julians, Malta

HUSSO 2015 Editors

Pascal Lorenz, University of Haute-Alsace, France

Bourret Christian, University of Paris East, France

HUSO 2015

Forward

The First International Conference on Human and Social Analytics (HUSO 2015), held between October 11 - 16, 2015 - St. Julians, Malta, was an inaugural event bridging the concepts and the communities dealing with emotion-driven systems, sentiment analysis, personalized analytics, social human analytics, and social computing.

The recent development of social networks, numerous ad hoc interest-based formed virtual communities, and citizen-driven institutional initiatives raise a series of new challenges in considering human behavior, both on personal and collective contexts.

There is a great possibility to capture particular and general public opinions, allowing individual or collective behavioral predictions. This also raises many challenges, on capturing, interpreting and representing such behavioral aspects. While scientific communities face now new paradigms, such as designing emotion-driven systems, dynamicity of social networks, and integrating personalized data with public knowledge bases, the business world looks for marketing and financial prediction.

The conference had the following tracks:

- Social human analytics
- Emotion basics

We take here the opportunity to warmly thank all the members of the HUSO 2015 technical program committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and effort to contribute to HUSO 2015. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the HUSO 2015 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope HUSO 2015 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of human and social analytics. We also hope that St. Julians, Malta provided a pleasant environment during the conference and everyone saved some time to enjoy the beauty of the city.

HUSO 2015 Chairs

HUSO 2015 Advisory Committee

Longbing Cao, UTS Advanced Analytics Institute, Australia

Pascal Lorenz, University of Haute-Alsace, France

HUSO 2015

Committee

HUSO 2015 Advisory Committee

Longbing Cao, UTS Advanced Analytics Institute, Australia
Pascal Lorenz, University of Haute-Alsace, France

HUSO 2015 Technical Program Committee

Raian Ali, Bournemouth University, UK
Laura Alonso Alemany, National University of Córdoba, Argentina
Panagiotis D. Bamidis, Aristotle University of Thessaloniki, Greece
Juan Manuel Belda Lois, Universidad Politécnica de Valencia, Spain
Shreyansh Bhatt, Kno.e.sis Center - Wright State University, USA
Senaka Buthpitiya, Carnegie Mellon University, USA
M. Emre Celebi, Louisiana State University in Shreveport, USA
Wei Cheng, University of North Carolina, USA
Pietro Cipresso, Applied Technology for Neuro-Psychology Lab - Istituto Auxologico Italiano, Italy
Stefano Cresci, IIT-CNR, Italy
Sérgio Deusdado, Polytechnic Institute of Bragança, Portugal
Matjaz Gams, Jozef Stefan Institute, Slovenia
Paolo Garza, Politecnico di Torino, Italy
Leontios Hadjileontiadis, Aristotle University of Thessaloniki, Greece
Yuh-Jong Hu, National Chengchi University, Taiwan
Baden Hughes, Glentworth Consulting, Australia
Abdessamad Imine, LORIA-INRIA, France
Roberto Interdonato, University of Calabria, Italy
Mehmed Kantardzic, University of Louisville, USA
Jonghwa Kim, University of Augsburg, Germany
Andreas Koch, University of Salzburg, Austria
Stefanos Kollias, National Technical University of Athens, Greece
Georgios Lappas, Technological Educational Institute of Western Macedonia, Kastoria, Greece
Carson Leung, University of Manitoba, Canada
Georges Linares, LIA - Avignon University, France
Cheng Long, Hong Kong University of Science and Technology, Hong Kong
Mai S. Mabrouk, Misr University for Science and Technology, Egypt
Sotiris Manitsaris, University of Thessaly, Greece / IRCAM | MINES ParisTech, France
Massimo Mecella, Sapienza Università di Roma, Italy

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Integration of Emotions and Feelings of Patients to Improve their Care - the Case of Healthcare Interface Organizations (HIO) in France

Bourret Christian
DICEN IDF
University Paris East Marne-la-Vallée
Marne-la-Vallée, France
e-mail : christian.bourret@u-pem.fr

***Abstract*— In all the developed countries, Healthcare Systems are in crisis with the central question of costs. They therefore face the key issue of performance. Everywhere (United Kingdom, USA, Spain, France, etc.) solutions are sought in new uses of information and involvement of patients or empowerment. The French Healthcare System has strong characteristics (particularly divisions between primary care and hospitals) that determine its evolution that we will analyze after explaining our research position. Better integrating patients' and in particular their emotions and feelings to improve their involvement in a coproduction of services perspective is one of the main goals of New Healthcare Interface Organizations (HIO) developed recently between primary care and hospital sector. We analyze how in this perspective they may play a key role to improve the French Healthcare System.**

Keywords - Healthcare; Interface Organizations; France; emotions; feelings; services.

I. INTRODUCTION

Healthcare Systems are in crisis with the central question of costs, accentuated during the last years in all the developed countries with key issues of efficiency and performance. Solutions are sought in new uses of information and involvement of patients or empowerment. The case of France is specific: the French Healthcare System has strong characteristics that determine its evolution (Section III). We will analyze these specificities as strong constraints after explaining our research position (Section II).

We will then show the key role of new Interface Organizations (HIO) to promote coordination and traceability and a cooperative process between primary care and hospital sector (Section IV) [1] [2] [3] [4]. Then we will examine the new role that could be given to patients by including their emotions and feelings to improve their involvement in a new coproduction of services perspective (Section V) [5].

II. RESEARCH POSITION

Our research position corresponds to French Academic discipline of Information and Communication Sciences, as proposed by F. Bernard [6] at the convergence of four problems: that of link (relationships, interactions), that of meaning, that of knowledge and that of action. We choose an approach of complexity (global and systemic) and of Action Research: produce usable knowledge for action, validated by all the stakeholders [7].

We refer to the "situational and interactionist semiotic method", in French, "sémiotique situationnelle et interactionniste" (A. Mucchielli [8]). This method tries to access meanings by understanding what things mean for an actor in a specific situation. Its key concept is the contextualization. Different levels of background are built by actor interpretation and define a situation, split into different frameworks: the intentions and issues of the actors, their cultural background, their positions in relation to other actors, the quality of relations, historical and temporal framework. All the meanings for all the actors build a "global sense" of the studied phenomenon.

We propose an extension of this method to integrate the experiences of actors including the emotions and the feelings of patients in a dynamic approach, considering the patient care pathway as a process in quality management. Always in a constructivist approach, we also consider the Activity Theory of Y. Engeström [9].

Our work in progress is based on cooperation established in France (particularly around students' works for Masters' degrees: observation, participation in meetings, interviews of key actors, etc.), but also with comparisons with other countries such as the Spanish Basque Country.

III. THE SPECIFICITIES OF THE FRENCH HEALTHCARE SYSTEM

The French Healthcare System can be regarded as "out of breath" that is at the end of its possibilities (Isaac, [10]) if we consider the results in comparison with the costs around 12 % of GNP. It is strongly affected by divisions (walls) [11], particularly between the primary care and the hospital sector (very dominant: this is called "hospital-centrism"), between physicians (cure) and nurses (care), between medical and social jobs, etc.

In the 1920s, physicians practicing in primary care (outside hospitals) refused the role of Health Insurance Companies (German model) to impose fees directly paid to them by patients who are subsequently reimbursed by Health Insurance Companies. So primary care physicians deny any link with Insurance Companies. For H. Isaac, ICT (Information and Communication Technologies) can afford

 ISBN: 978-1-61208-447-3

to develop another model, promoting prevention and cost savings. These prospects meet issues of "augmented human" or "quantified self " with all the new uses of Open Data or above all of Big Data with the risk of abuses, particularly on privacy, highlighted by the CNIL – National Commission for Informatics and Freedom (in French, Commission Nationale de l'Informatique et des Libertés) [12], and also another risk around ultra-connection and ultra-transparency.

Patient involvement is essential, in reference to the Anglo-Saxon concept of empowerment. In France, we prefer to talk about "responsibility". In this perspective, the law of March 2002 on "The Rights of Sick People and the Quality of the Healthcare System" (in French, Loi relative aux Droits des malades et à la qualité du système de santé) was a major step by proposing the concept of "health democracy" (démocratie sanitaire) and valuing Healthcare Networks (Réseaux de santé). HIV-AIDS networks have played a key role for a new approach of patient role. The law of July 2009 ("Hospital, Patients, Health, Territories", in French : HPST - Hôpital, Patients, Santé, Territoires) proposed a new regionalization of Health management with the creation of ARS (Health Regional Agencies, in French, Agences Régionales de Santé) and two agencies to improve information systems (Agency for Shared Information Systems in Health (in French, Agence pour les Systèmes d'Information Partagés en Santé) and performance in the health and social sector (ANAP: Agence Nationale d'Amélioration de la Performance), particularly in the medico-social sector, also confirming the role of the High Authority for Health (HAS, Haute Autorité de Santé). As an extension of the Companion - Ghali report (2014) [13], a new law voted by Parliament inl 2015 wants to give a new impetus to the concept of "health democracy". It also insists on a better use of information with the repositioning of the Electronic Health Record or DMP (in French, Dossier Médical Personnel), whose management since 2004 is largely a failure. It is now redefined not as personal belonging to the patients (first perspective) but as shared by all the medical actors.

In this context, the development of new Interface Organizations for facilitating cooperations between primary care and hospital sector, using ICT and involving more patients, is a key issue. The Godet - Durance - Mousli report (2010) [14] emphasized the role of the Health field as essential for innovation, focusing in particular on the experience of Healthcare Networks.

IV. HEALTHCARE INTERFACE ORGANIZATIONS (HIO) AS A CHANGE LEVER

New Healthcare Interface Organizations (HIO) developed to overpass divisions between primary care and hospital sector may be considered as spaces for innovation, experimentation and development of trust (Bourret, [2] [3]), both for the human actors (individuals or belonging to organizations) and also in digital tools (digital trust), also with the recognition of different roles: that of doctors, nurses, social workers, patients and their families, etc.

Healthcare Networks (in French, Réseaux de santé) appeared in the 1980s, especially with the AIDS epidemic and the need for coordination between general practitioners (primary care) and hospital sector and between the medical and the social sector. Their role has been enshrined in the Law of March 2002. They were developed for different diseases (diabetes, cardiology, etc.), or for specific situations (perinatal care, addictions, obesity, geriatrics, oncology, etc.) and in some cases their role would be better defined with HAD (Hospitalization at Home, in French, Hospitalisation à Domicile) organizations, with whom they are often in strong competition on the same areas.

First Healthcare Networks are coordination structures of professional activities engaged in different spaces, often with just a phone center to receive patients' calls (Healthcare Network Bronchiolitis in Paris, Gérontopastel Network for old people in Toulouse, etc.).

Their rivals, Multi Medical Homes (in French, Maisons de Santé Pluriprofessions) gather different practitioners on the same site, as it is, for example in Spain for ambulatorios. In the recent years, the authorities seem to prefer them and encourage their development, focusing also on grouping and coordinating activities. Thus the Healthcare Network on Diabetes in Eastern Ile-de-France (Revesdiab) became the main actor in the new structure of GCS (Health Cooperation Group, in French, Groupement de Coopération Sanitaire) Diapason (in southern Seine-et-Marne department, which represents half of the area of the Ile-de-France Region). The GCS is a new key tool to develop cooperations between public and private sector, but also between the primary care and the hospital sector. It connects public and private health institutions, health centers, nursing homes, medical professionals and the actors of medical-social sector, acting individually or collectively.

Other modalities of cooperation have also emerged in recent years as the PAERPA experiments (Elderly People at Risk of Loss of Autonomy, in French, Personnes Agées en Risque de Perte d'Autonomie) that particularly highlight PPS (Personalized Care Plans – in French, Plans Personnalisés de Soins) (Bloch – Hénaut [4]) which, eventually focus on monitoring of individually sustain of patients' care pathway, set up by the Healthcare Networks, applying in this case to in elderly patients. So GCS Diapason applies a PPS to diabetic patients.

MAIA (Houses for Autonomy and Integration of Alzheimer Patients, in French, Maisons pour l'Autonomie et l'Intégration des Malades d'Alzheimer) were mainly intended to be the devoted entry for patients with this disease and their families. They were put in place by the new CNSA (National Solidarity Fund for Autonomy, in French, Caisse Nationale de Solidarité pour l'Autonomie). They have often considered the PAERPA as unfair and unnecessary competition.

These interface organizations are an attempt to answer the central problem of the divisions or walls [11] of the French Healthcare System. But their proliferation, often without real coherence or overall vision, according to local initiatives of different Health Insurance or Social Security

 ISBN: 978-1-61208-447-3

funds or territory collectivities (departments, municipalities, etc.) can lead to new divisions and a loss of efficiency.

V. A NEW PATIENT'S APPROACH INTEGRATING THEIR EMOTIONS AND ENCOURAGING THEIR INVOLVEMENT

Patients gradually asserted their role, particularly in the applications of the 2002 Law (for example the hospitalized patient's charter in 2006) with the key role of associations such as the CISS (Interassociative Collective on Health, in French, Collectif Interassociatif Sur la Santé) and with the "judicialization" of health from the perspective of "perfect health" described by L. Sfez [15]: obligation of results and not only of means.

Next, we will describe our point of view regarding experiments we are associated with.

We will analyze in particular the concept of Personalized Care Plan (PPS) from the case of GPS Diapason whose mission is animation of the territory of Seine-et-Marne and testing of innovative practices involving the patient, especially diabetic, around the telemedicine project "Diabetes 2.0" (including the Healthcare Network Revesdiab). Diabetes 2.0 is a multi-year large-scale innovative project involving both the city and the hospital. It provides various tools of tele-consultation, remote monitoring, tele-expertise, as well as virtual tools of therapeutic education. 2.0 Diabetes wants to use telemedicine as a training tool for health professionals and to promote patients' compliance. The ultimate goal is to improve the management of chronic diseases (diabetes here) and associated complications, optimizing care pathways and controlling health expenditures.

We will offer broader perspectives from "situational and interactionist semiotic method" (described in section II) extended to emotions and feelings of actors (Goleman, [16]), especially for patients and their families. And also extended to a dynamic approach of change (in a process approach regarding quality management), the patient's pathway can be regarded as a process.

We will insist on the concept of emotional skills of patients, central element of emotional intelligence. The human body is the mediator from which the individual can sensitize its affects and communication support them (Martin-Juchat, [17]).

Our goal hypothesis is that managing emotions can become a collective objective of improving the quality of care. Social sharing of emotions is essential to promote the group membership (Bègue-Desrichard, [5]). This integration of emotions and feelings in the patient's personalized care can improve the quality of care, for example through better patient adherence to prescription compliance and taking medication. The notion of recognition is also essential in the discussion groups. So we will evoke the role of "mediator patients" already developed in some Healthcare Networks and that of "case manager" for coordination of care. So we will place ourselves in the perspective outlined by the High Authority of Health (HAS) (2007 [18]) in its recommendations for patient's education, asking to recognize how to act in the patient's positioning to analyze his psychological evolution by developing techniques of patient-centered communication (active listening, empathy, encouragement, etc.).

In an Action Research perspective, we will conduct specific interviews with patients and their families both in individual and group perspectives and then we will try to improve the integration of their ideas by interactive discussions with medical and social actors. We particularly insist on the role of "mediator patients" with an idea to specific training to help to improve the implication of all the other patients. We will particularly work on the notion of "complex patient", especially elderly people with chronic diseases and social problems.

The digital dimension (ICTs) is also essential. The areas of activities of different Interface Organizations can be considered as Digital Care Territories (in French, TSN: Territoire de Soins Numérique), in the same way an individual patient is.

The prospects outlined represent a new approach of services coproduction integrating emotions and feelings of actors (Maman, [19]) applied in this case to the Healthcare sector. More generally, this crossing from the individual dimension to the collective dimension is a major challenge to found a new citizenship, particularly in the Smart Cities project (digital), as outlined by M. Zacklad [20].

With GCS Diapason we begin to study the Diabetes 2.0 Project. In isolated areas, to obtain an appointment with a specialist is very difficult with often very long delays. The main idea is to use telemedicine solutions to develop connection points on still isolated territories. Priority is given to new patients, but also patients who experience complications and elderly patients with reduced autonomy. The project is going to manage three different cohorts. A first cohort corresponds to newly diagnosed diabetic patients without known diabetes complications. A second cohort of already known and treated patients; diabetics aged 65 and older with at least one known diabetes complication. And finally, a third cohort of patients ages 65 years and older with loss of autonomy at home or in nursing homes, in diabetes care with risk of hypoglycemia. Integration of emotions of patients by discussions and use ICTs to develop their implication and their responsibility are key parts of this project.

Loneliness is a huge problem. Discussing with other people, not to be alone and sharing views on illness is the first step to improve the quality of life of these patients and for integrating their emotions and their feelings. It is also a manner of taking into account the different times of all the actors and their different goals: these of patients being very different than those of doctors, nurses or social workers. Trying to integrate patients' knowledge about their disease will be the second step.

VI. CONCLUSION

The role of patients is essential to reinvent the Welfare State and a new local citizenship, with all the importance of the integration of their emotions and feelings. Health Interface Organizations may constitute privileged spaces for

 ISBN: 978-1-61208-447-3

developing new approaches emphasizing emotional intelligence dimensions, building trust, from the recognition of the role of all stakeholders, not just of the medico-social, but also of patients and their families. These approaches correspond to coproduction of services perspectives and "daily innovation", according to N. Alter [21].

We have also begun to address these issues in a comparative European perspective, especially for cases of elderly patients in complex situations for example with the Carewell Project, principally coordinated by Healthcare Organizations of the Spanish Basque Country. Carewell is a project for the development of a new organizational model based on integrated care for chronic and elderly patients through ICTs [22].

The tracks we started to trace also belong to a more comprehensive approach to "sustainable development" of territories, corresponding to that of the Brundtland Report (1987) [23], which cannot be reduced solely to environmental issues but is also based on two other pillars: economic (growth) and social (education, health and respect for freedoms).

REFERENCES

[1] P. Larcher and P. Polomeni, Health in Networks. Goals and Strategy for a Primary Care – Hospital Collaboration / La santé en réseaux. Objectifs et stratégie dans une collaboration ville-hôpital, Paris : Masson, 2001.

[2] C. Bourret, Organizational Dynamics aroud Collective Production of Information and Communicational Process. The cas or healthcare Interface Organizations / Dynamiques organisationnelles autour de la production collective d'information et de processus communicationnels. Le cas des organisations d'interface du secteur santé, HDR (Accreditation to Supervise Research / Habilitation à Diriger des Recherches), dir. A. Mayère, University Paris East, 2 vol., 2010.

[3] C. Bourret, "E-Health and Societal and Territorial Intelligence in France. Collective Knowledge Production Issues and New Networked Interface Organizations", Chapter 12, Competitive Intelligence and Decision Problems, Edited by David A., ISTE – John Wiley Ed., London, 2011, pp. 247 – 268.

[4] A.M. Bloch and L. Henaut., Co-ordination and Pathways. Dynamics of the healthcare, social and medico-social sectors / Coordination et parcours. La dynamique du monde sanitaire, social et médico-social, Paris : Dunod, 2014.

[5] L. Bègue and O. Desrichard, dir., Treaty of Social Psychology. The Human Interactions Science / Traité de psychologie sociale. La science des interactions humaines, Bruxelles : De Boeck, 2013.

[6] F. Bernard, "ICS a Disciplinary of Openness and Decompartmentalization / Les SIC une discipline de l'ouverture et du décloisonnement ", in A. Bouzon, dir., Organizational Communication in Debat / La communication organisationnelle en débat. Champs, concepts, perspectives, Paris : L'Harmattan, 2006, pp. 33 – 46.

[7] V. Meyer, " Utility of Action-Research in ICS / De l'utiilité des recherches-actions en SIC (sciences de l'information et de la communication)", Communication & Organisation, Bordeaux : Michel de Montaigne University, 30 / 6, 2006, pp. 89 – 108.

[8] A. Mucchielli, Situation and communication, Nice : Editions Ovadia, 2010.

[9] Y. Engeström and al. (eds), Perspectives on Activity Theory, Cambridge: Cambridge University Press, 1999.

[10] H. Isaac, dir., From a Curative to a Preventive Health System Thanks Digital Tools / D'un système de santé curatif à un modèle préventif grâce aux outils numériques, Renaissance numérique, Paris, 2014. Available from : http://fr.slideshare.net/RenaissanceNumerique/lb-sante-preventive-renaissance-numerique-1 (August 28th, 2015).

[11] S. Glouberman, H. Mintzberg, "Managing the Care of Health and the Cure of Disease ", Health Care Management Review, - Volume 26 - Issue 1, 2001, *pp*: 5-90.

[12] CNIL (Commission Nationale de l'Informatique et des Libertés), " The Human Body a New Connected Object. From Quantified Self to Heath : New Territories of World Data Layout / Le corps, nouvel objet connecté. Du Quantified self à la santé : les nouveaux territoires de la mise en données du monde ", Cahiers IP Innovation et Prospective, 2014, n° 2, 63 p.

[13] C. Compagnon and V. Ghali, For Year II of Sanitary Democracy / Pour l'An II de la démocratie sanitaire, Report to The Ministry of Health and Social Affairs, Paris, 2014. Available from : http://www.sante.gouv.fr/IMG/pdf/Rapport_DEF-version17-02-14.pdf (August 28th, 2015)

[14] M. Godet., P. Durance, and M. Mousli, Ceativity and Innovation on Territories / Créativité et innovation dans les territoires, Paris, Conseil d'Analyse Economique – La documentation Française, 2010.

[15] L. Sfez, dir., Utopia of Perfect Health / L'utopie de la santé parfaite, Colloque de Cerizy, Paris : PUF, 2001.

[16] D. Goleman. Working with Emotional Intelligence, London: Bloomsbury, 1998.

[17] F. Martin-Juchat, Human Body and Medias. Fresh tried by the Medias and the Social Spaces / *Le corps et les médias. La chair éprouvée par les médias et les espaces sociaux*, Bruxelles : De Boeck, 2008.

[18] Haute Autorité de Santé (HAS). Patients' Therapeutic Education. How to Propose and to Achieve it ? / L'éducation thérapeutique du patient : Comment la proposer et la réaliser ?, 2007, Available from : http://www.has-sante.fr/portail/upload/docs/application/pdf/etp_-comment_la_proposer_et_la_realiser_-_recommandations_juin_2007 (August 28th, 2015).

[19] C. Maman, "Deafness of the Provider in the Service Relationship: When the Recognition Deficit Deteriorate Customer Satisfaction/ La surdité du prestataire dans la relation de service : quand le déficit de reconnaissance dégrade la satisfaction du client", Development and Strengthening of Social Ties / Développement et renforcement du lien social, Management & Sciences Sociales, n° 16, Janvier – Juin, 2014, pp. 32 – 46.

[20] M. Zacklad, "Economies of Conviviality in Information and Services Society / Les économies de la convivialité dans les sociétés de l'information et des services", Inaugural Lecture, Paris, CNAM, 2009.

[21] N. Alter. Daily Innovation / L'innovation ordinaire, Paris : PUF, 2005.

[22] Carewell Project, Available from: http://www.engaged-innovation.eu/discussions/carewell-project-development-new-organizational-model-based-integrated-care-chronic-and (August 28th, 2015).

[23] G.O Brundland, dir., Our Common Future. Report of the World Commission on Envronment and Development, New York, 1987. Available on: http://www.un-documents.net/our-common-future.pdf (August 28th, 2015).

 ISBN: 978-1-61208-447-3

The analysis of the Movement of Experienced and Inexperienced Persons in Japanese Bowing

Tomoko Ota
Chuo Business Group
Osaka, Japan
e-mail:promotl@gold.ocn.ne.jp

Tomoya Takeda
Taste Inc.
Kyoto Japan
e-mail:t.takeda@taste.jp

***Abstract*— In Japanese, to receive a guest with hospitality and assist him in various ways is called "omotenashi". The word "omotenashi" has become internationally recognized as designating a form of welcoming rooted in the traditions and culture of Japan. Greeting is the basis of omotenashi, and one of the ways to greet is to bow. We conducted an experiment on the difference in maneuver between an experienced and inexperienced person and measuring the positions and timing of the bowing motion. Furthermore for the inexperienced persons, we conducted an analysis of the improvement of the motion with presence or absence of instruction as variable. The bow of the experienced person had a stable angle, with the shoulder angle (θ1) being around 180 degrees all three times, and the angle of the waist (θ2) was at around 20 degrees all three times. The transitions from the beginning of the bow until the head was lowered and from the head beginning to rise until the end of the bow were about equal speed, and the speed of the head was relatively slow at 300 (mm/sec). Inexperienced persons can be trained to a certain level by watching footage of model bowing, though there are significant differences according to the individual.**

Keywords-Hospitality; Japanese; Japanese bow; Omotenashi

I. INTRODUCTION

In Japanese, to receive a guest with hospitality and assist him in various ways is called "omotenashi". The Japan Productivity Center defines "omotenashi" as "work to provide special service from the heart while valuing the perspective of customers and/or residents." In foreign countries, the same concept exists and called for example "hospitality (U.S.)," "*dai ke zhi dao* (China), and "*hospitalité* (France)," but recently the word "omotenashi" has become internationally recognized as designating, along with the definition above, a form of welcoming rooted in the traditions and culture of Japan. According to an investigation of the Japan Productivity Center, a majority of people in the U.S., China, and France have heard of the word "Omotenashi" [1].

One of the reasons for the rising awareness of omotenashi is the increase in foreign visitors to Japan. In 2013, the number of foreign visitors passed 10 million for the first time, due to economic growth in Asian countries as well as success in the promotion of travel to Japan by a tourism policy called "Visit Japan" that was devised in 2003. The Japanese government will further devise a plan to increase the number of visitors to 20 million by 2020. In 2020, Olympic and Paralympic Games will be held in Tokyo. As the host country for the Olympic and Paralympic, Japan has an urgent need to convey its culture to the world in a comprehensible way. As stated earlier, omotenashi is a form of hospitality rooted in the culture and traditions of Japan, one that gives importance to touchpoints with the customer, an original way of giving high quality service from the heart based on mutual communication. Omotenashi could be said to be Japan's most important aspect.

Greeting is the basis of omotenashi, and one of the ways to greet is to bow. Bowing has different shades of meaning according to the country or region, but generally speaking it is the action of bending from the waist to greet someone, express thanks, or apologize. In Japan, bowing also is a way to greet, give thanks, and apologize, but its role as the fundamental action of Japanese hospitality and culture known as "omotenashi" and its designation as high quality service from the heart gave it a different significance from that in other countries. Japanese bowing is classified according to the angle at which the bowing is done, the levels being "*eshaku* (greeting bow)", "*keirei* (respect bow)", and "*saikeirei* (highest respect bow)". The classified bowing are in Figure 1. The levels differ in degree of honoring and are used in different circumstances. The maneuver consists in these three stages: to stand tall, to bend from the lower back, and to return from the lower back.

To spread Japanese culture internationally, it is urgent to consider how to convey omotenashi in ways easily understood by foreigners starting from the act of bowing. In the field of traditional Japanese industry and care, there is a prior case studies on the comparison of experienced and inexperienced person by the motion analysis. Based on these previous studies, we conducted an experiment on the difference in maneuver between an experienced and inexperienced person by recording their movements using a video camera and measuring the positions and timing of the bowing motion. Furthermore for the inexperienced persons, we conducted an analysis of the improvement of the motion with presence or absence of instruction as variable [2]-[4].

 ISBN: 978-1-61208-447-3

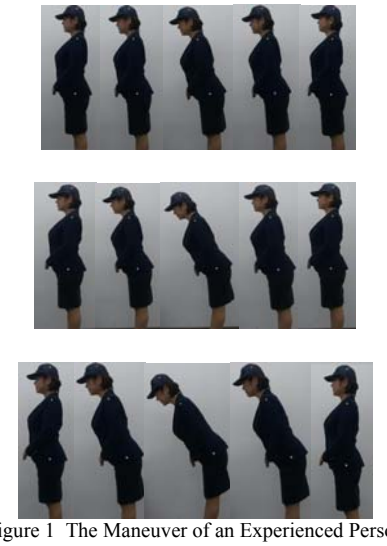


Figure 1 The Maneuver of an Experienced Person
Upper : eshaku (greeting), Middle: keirei (respect),
Lower: saikeirei(highest respect)

By performing this analysis, for the provider of services, it is expected that the point of the teaching becomes clear. On the other hand, for the receiver side of the service, by being able to quickly reach the elementary level that can understand the hospitality, and it is expected to deepen the understanding of the Japanese culture.

The rest of this paper is organized as follows. Section 2 describes experimentation method. Section 3 describes the measured angle, and speed of the bow. Section 4 is discussing this experiment. Section 5 goes into conclusion of this paper.

II. EXPERIMENTATION METHOD

A. *The Test Subject and the Bowing*

The test subjects were one experienced person with a history of teaching omotenashi for 7 years and 4 inexperienced persons. Gender, age and physical condition are shown in TABLE 1. Japanese persons who had never received instruction in the motion of a bow. For the bowing, we used "*keirei*".

TABLE 1 GENDER / AGE / PHYSICAL CONDITION

Test Subject	Experience	Gender	Age	Hight
1	Experienced	female	46	163
2	Inexperienced	female	26	154
3	Inexperienced	female	53	153
4	Inexperienced	male	35	183
5	Inexperienced	male	24	170

B. *Experimentation Method*

As shown in Figure 2, a marker was placed at the head, shoulder, waist, and knees of the test subjects, and the bowing motion of each was recorded by a video camera, and measurements were made for the passing of time and the location of the markers.

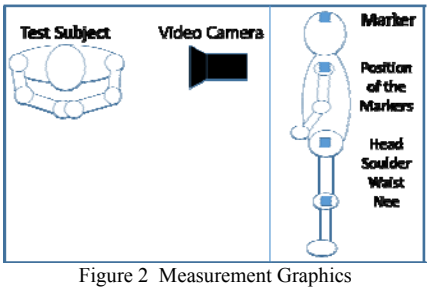


Figure 2 Measurement Graphics

Furthermore, with the inexperienced subjects, measurements were made with three divisions: "bowing without any outside influence", "bowing after looking at the bow of the experienced person", and "bowing after receiving instruction from the experienced person".

1) The bow of the experienced person. With a marker placed on the head, shoulder, waist, and knee of the experienced person, a recording with a video camera was made from the side of the "keirei" bow. The same motion was made three times.

2) Bowing without outside influence. The four inexperienced persons who had never received instruction in bowing each performed a bow as they understood it three times and this was recorded by a video camera.

3) Bowing after looking at the bow of the experienced person. After looking at the footage of the experienced person's bow, the four inexperienced persons performed a bow and this was recorded by video camera.

4) Bowing after receiving instruction from the experienced person. The inexperienced persons were made to study from a video that explains the meaning of bowing, and afterwards they performed a bow and it was recorded by video camera. The explanation video was common for both genders and taught the following 5, items1. The meaning of the act of bowing in Japan. 2. Types of bowing. 3. Speed of bowing. 4. Posture while bowing. 5. Gaze while bowing.

C. *Analysis Method*

With the measured time and information of the markers' locations we discerned the transition of location and speed. We paid special attention to the speed of the head (**V**h). And from the distance between the four points of head, shoulder, waist, and knees, we calculated the angle of head-shoulder-lower back (θ1) and the angle of shoulder-waist-knee (θ2).

III. THE TIME, MEASURED ANGLE, AND SPEED OF THE BOW

A. *Time Needed for Each Test Subject to Bow*

Three measurements were made, and the results are shown in TABLE 2 according to test subject. The three measurements were divided as "from the beginning of the

 ISBN: 978-1-61208-447-3

bow until the head is lowered" as motion 1, "the head lowered and stopped" as motion 2, and "from the raising of the head until the end of the bow" as motion 3.

The total time for the experienced person was 1,122ms, but for the inexperienced persons it ranged from 1,206ms to 1,812ms. The difference between the inexperienced individuals was great, and compared to the experienced person more time was required.

TABLE 2 MOTION TIME OF EACH TEST SUBJECT
(TIME UNIT: MS)

	Number of Motions	1st	2nd	3rd	Average
Test Subject 1	Motion 1	333	400	383	372
	Motion 2	400	350	433	394
	Motion 3	350	350	367	356
	Total	1083	1100	1183	1122
Test Subject 2	Motion 1	600	433	367	467
	Motion 2	183	433	383	333
	Motion 3	433	400	383	406
	Total	1217	1267	1133	1206
Test Subject 3	Motion 1	417	567	650	544
	Motion 2	483	633	517	545
	Motion 3	383	400	467	417
	Total	1283	1600	1633	1506
Test Subject 4	Motion 1	367	583	400	450
	Motion 2	367	317	600	428
	Motion 3	517	533	500	517
	Total	1250	1433	1500	1394
Test Subject 5	Motion 1	533	650	533	572
	Motion 2	700	833	883	806
	Motion 3	503	400	400	434
	Total	1737	1883	1817	1812

B. *The Bow of the Experienced Person*

The transition of the angle of the bow of the experienced person is shown in Figure 3. The angle of the shoulders, θ 1 is about 180 degrees all three times and hardly shows any difference. Also the angle of the lower back, θ 2 is held at about 20 degrees all three times.

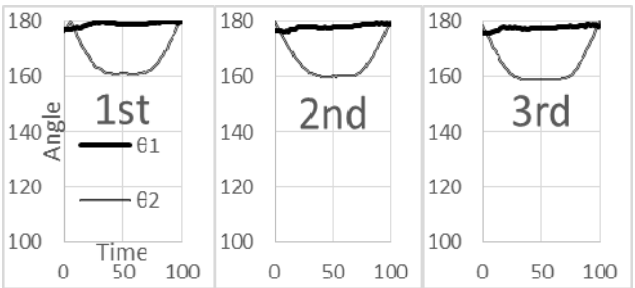


Figure 3 Standardized Transition of Angle of the Bow of the Experienced Person (θ1/θ2)

In Figure 4, the transition of the speed of the experienced person's bow is shown. Motion 1, from the beginning of the bow until the head is lowered, and motion 3, from the raising of the head until the end of the bow, had almost the same speed during the transition.

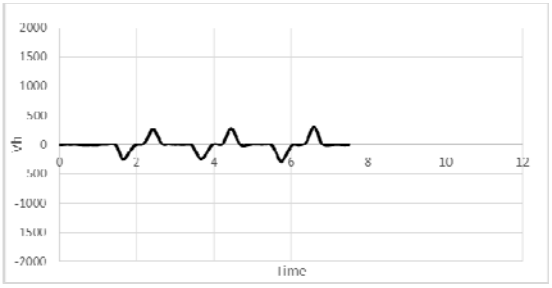


Figure 4 Transition of the Speed of the Head (Vh:mm/sec)

C. *Mastership of Bowing*

Here, we measure for each inexperienced person the influence that the bowing instruction had on their motions.

1) Bowing without any Outside Influence. The test subjects were made to do three bows that they felt were correct, and θ 1 and θ 2 were measured. As a result, a tendency for θ 1 and θ 2 to be synchronized was seen, so bending at the waist occurs simultaneously with the motion of bending the neck. The speed is generally faster than test subject 1 who is experienced, and also there were cases of motion 3 being faster than motion 1 and of speed increasing or decreasing as the three bows were performed (Figures 5~8).

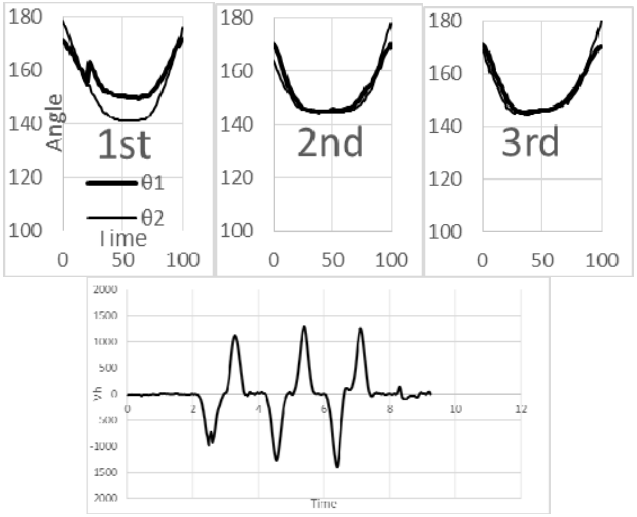


Figure 5 Vh and θ1/ θ2 and Vh of Test Subject 2 (Vh:mm/sec)

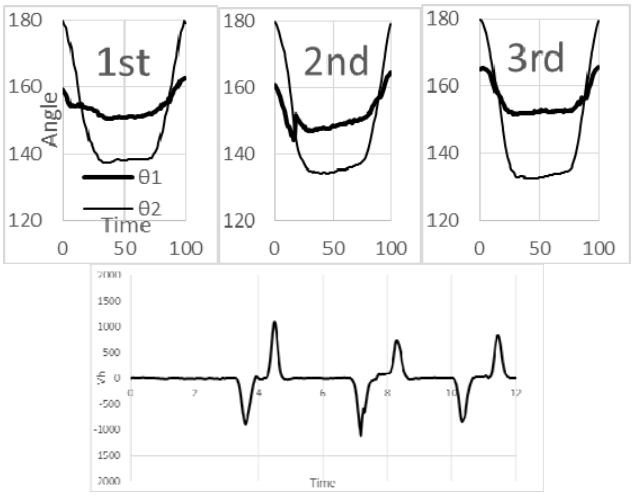


Figure 6 Vh and θ1/ θ2 of Test Subject 3 (Vh:mm/sec)

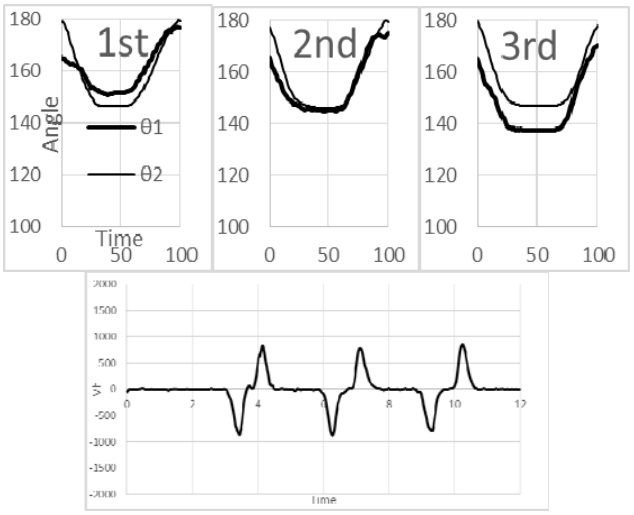


Figure 7 Vh and θ1/ θ2 of Test Subject 4 (Vh:mm/sec)

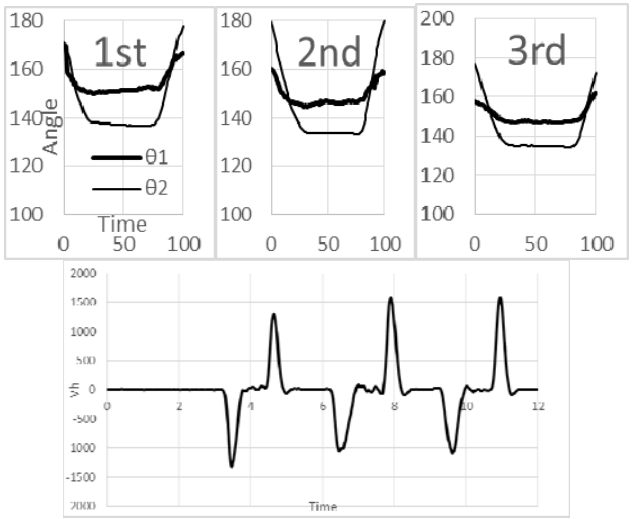


Figure 8 Vh and θ1/ θ2 of Test Subject 5 (Vh:mm/sec)

2) Changes after Observing the Experienced Person's Bow. The results of bowing three times after observing the experienced person's bow are shown in Figures 9~12. There was no change for test subjects 2 and 4 for θ1 and θ2 to synchronize, but with test subjects 3 and 5, θ1 holds a steady angle. This is thought to be because they noticed that the experienced person's bow has no change in the shoulders' angle and the motion is done by bending at the waist. For test subject 5, speed is clearly lowered.

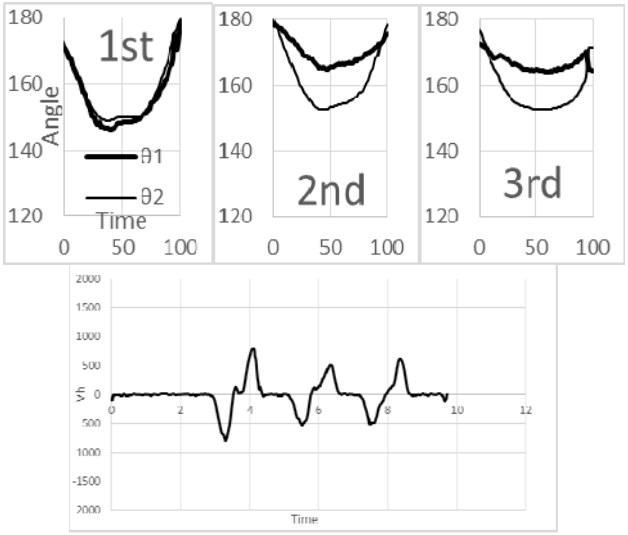


Figure 9 Vh and θ1/ θ2 of Test Subject 2(Vh:mm/sec)

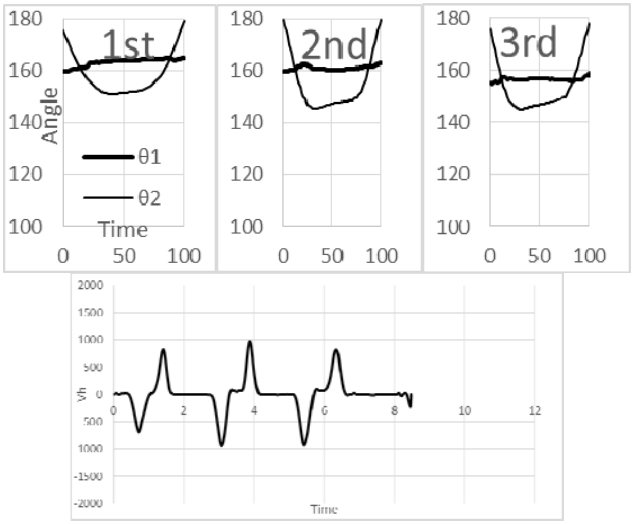


Figure 10 Vh and θ1/ θ2 of Test Subject 3 (Vh:mm/sec)

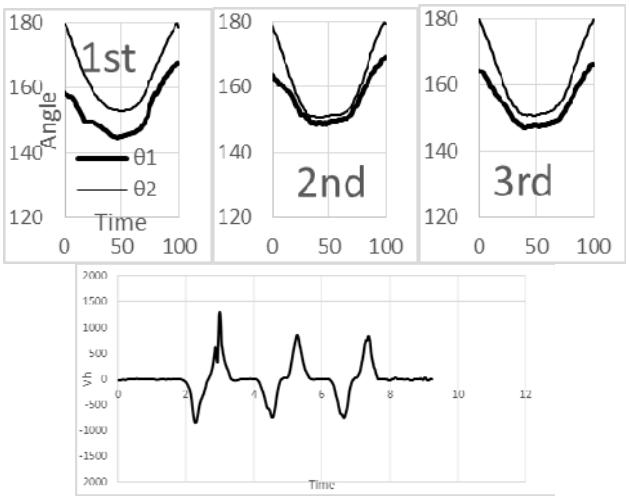


Figure 11 Vh and θ1/ θ2 of Test Subject 4(Vh:mm/sec)

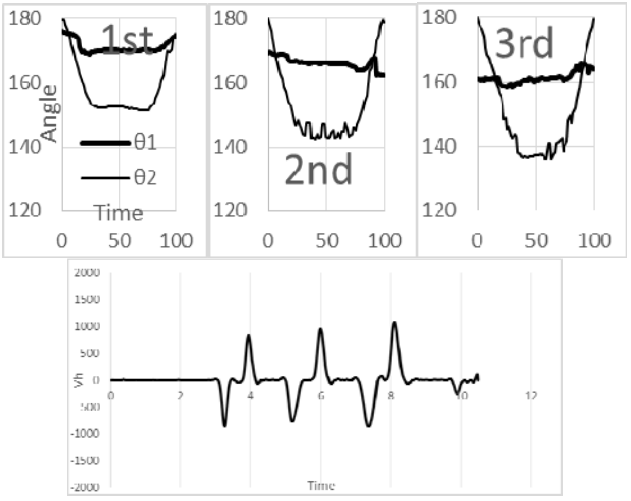
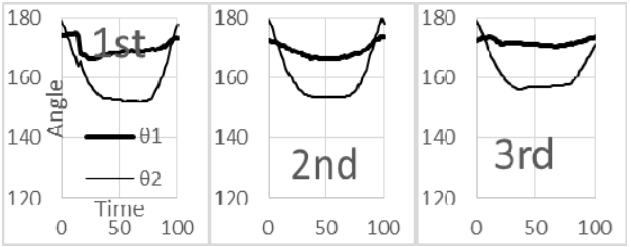


Figure 12 Vh and θ1/ θ2 of Test Subject 5(Vh:mm/sec)

3) Bowing after Receiving Instruction from the Experienced Person. Next, we examined the changes in bowing of the inexperienced persons after receiving concrete instructions of bowing by the instructor.



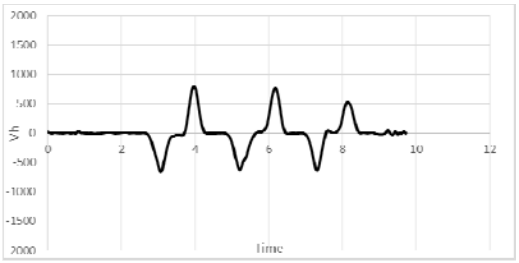


Figure 13 Vhand θ1/ θ2 of Test Subject 2 (Vh:mm/sec)

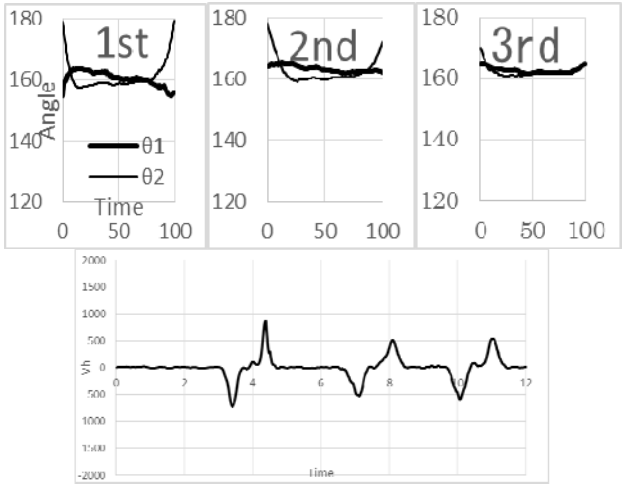


Figure 14 Vh and θ1/ θ2 of Test Subject 3 (Vh:mm/sec)

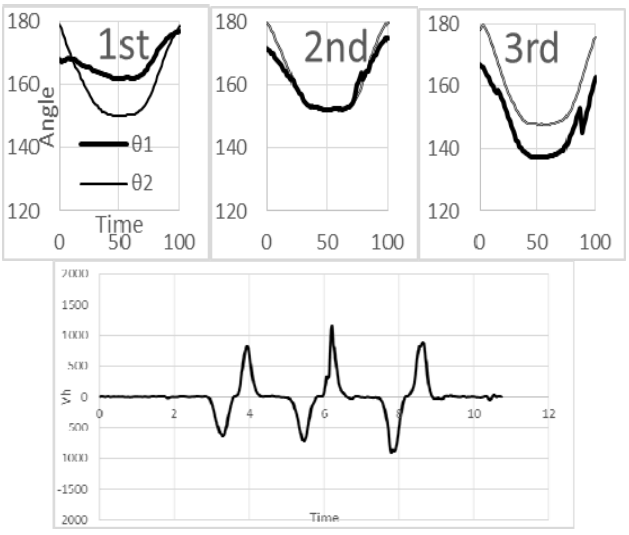


Figure 15 Vh and θ1/ θ2 of Test Subject 4 (Vh:mm/sec)

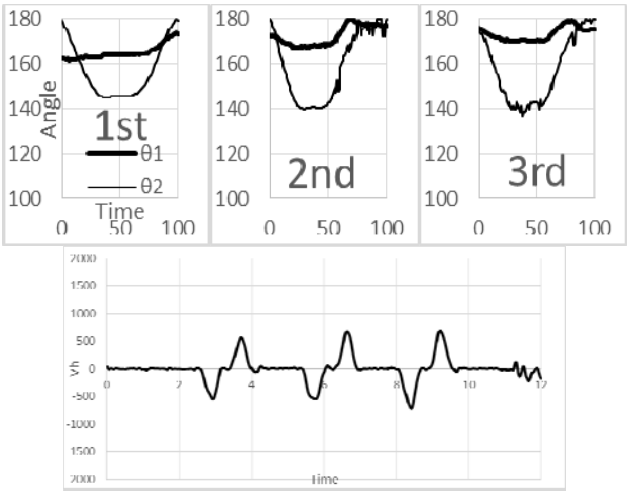


Figure 16 Vh and θ1/ θ2 of Test Subject 5 (Vh:mm/sec)

From the instruction test subjects 2 , 3 and 5 were able to maintain a steady θ1. Speed generally became stable at a low speed (Figures 13~16).

IV. DISCUSSION

First, we will consider angles (θ1/ θ2). The angle of the shoulders, θ1 of the experienced person is about 180 degrees all three times without much change. Also the angle of the waist, θ2 is maintained at about 20 degrees all three times. A bow is done by bending from the waist, and must not be done by bending the angle of the shoulder, or in other words the neck must not be bent. The basics of a Japanese bow are being performed here. For the speed of the head, it could be seen that the speed is relatively low at 300 (mm/sec).

The bows of the four test subjects were standardized from the ones they thought were correct to begin with. As a result, θ1 and θ2 tended to synchronize. This shows that the waist and neck are bent simultaneously. In terms of speed, there were cases of motion 3 being faster than motion 1, the motions becoming faster as the three bows progressed, and the motions becoming slower as the three bows progressed.

For the results of bowing three times after observing the experienced person, θ1 and θ2 being synchronized did not change with test subjects 2 and 4, but with test subjects 3 and 5 θ1 maintained a steady angle. This can be said to be because they noticed that a bow should be done by only bending the waist and not the shoulders. In terms of speed no changes especially were noticed.

In the results of bowing after receiving instruction from the experienced person's bow, subjects 2, 3 and 5 were able to maintain a steady θ1. This is considered to be because they received instruction from the experienced person and understood the meaning of a bow which made them realize they must not change the angle of the shoulders, but go through the bending motion from the waist.

The level of change before and after receiving the instruction differed for each individual test subject, and each motion lacked stability. This is thought to be because there is individual difference in perspective and understanding, even if they all observe the experienced person or receive instruction, and because it is difficult to measure the angle of one's shoulders and waist by oneself. And with speed, regardless of there being instruction or not, the difference between the experienced person and inexperienced persons was great, which shows that verbal instruction is insufficient for learning the "gentleness" of the experienced person's bow. For instruction it is considered that there needs to be more depth in instruction method for stability and maintaining a gentle speed or for skills such as where to look when maintaining the angle of the shoulders.

V. CONCLUSION

In this research, we focused on Japanese bowing, a form of greeting which is fundamental to omotenashi, and conducted experiments involving the angles of bowing of experienced and inexperienced persons, the difference in speed, and the changes that take place when inexperienced persons received instruction. As a result, the following things were understood.

- The bow of the experienced person had a stable angle, with the shoulder angle (θ1) being around 180 degrees all three times. In other words, the head was straight and the angle of the waist (θ2) was at around 20 degrees all three times.
- The transitions from the beginning of the bow until the head was lowered and from the head beginning to rise until the end of the bow were about equal speed, and the speed of the head was relatively slow at 300 (mm/sec).
- Inexperienced persons can be trained to a certain level by watching footage of model bowing, though there are significant differences according to the individual.

REFERENCES

[1] Japan Productivity Center, "Japan's infrastructure in the information economy and society", pp8-11, February 2012.

[2] Zelong Wang, Kenichi Tsuji, Toru Tsuji, Yuka Takai, Akihiko Goto, Hiroyuki Hamada, "Brain activity analysis on "Kana-ami"making process", 17th International Conference on Human-Computer Interaction 2015.

[3] Testuo Kikuchi, Erika Suzuki, Yiyi Zhang, Yuka Takai, Akihiko Goto, Hiroyuki Hamada "Effects of quantified instructional tool on spray-up fabrication method", 17th International Conference on Human-Computer Interaction 2015.

[4] Mengyuan Liao, Takashi Yoshikawa, Akihiko Goto, Yoshihiko Mizutani, Tomoko Ota, Hiroyuki Hamada, " A study of caregiver's waist movement comparison between expert and non-expert during transfer care", 17th International Conference on Human-Computer Interaction 2015.

 ISBN: 978-1-61208-447-3

Effectiveness of Analysis with NIRS for Japanese EFL Learners

Rumi Tobita
Life Systemics
Ashikaga Institute of Technology
Tochigi, Japan
email : rtobita@ashitech.ac.jp

***Abstract*—This study examined the effectiveness of analysis with near-infrared spectroscopy (NIRS) for English as a foreign language (EFL) training from the viewpoint of brain science. The experiment presented in this paper analyzed the amount of blood flow in the brain while learners were training to improve their English conversation skills. The experiment attempts to clarify the preferable combinations of learners' characteristics when teaching English conversation by examining relationships between the brain activities of learners and the different types of training materials. The data suggested that the analysis using NIRS enabled to propose an effective course design for EFL learners.**

Keywords-NIRS; brain activities; EFL; English conversation skill; ATI.

I. INTRODUCTION

In the light of the ever increasing globalization and internationalization of our society, the development of English communication skills is considered crucial in Japan. However, it has been noted Japanese students' English skills were declining [1]. Therefore, designing and developing an effective course design to meet ELF goals for acquiring English communication skills has been a critical need. To solve this concern, the present study examined the effectiveness of analysis using NIRS for EFL listening training from the perspective of brain science to propose a well-matched combination of leaners' characteristics and listening training to create an effective course design for EFL learners.

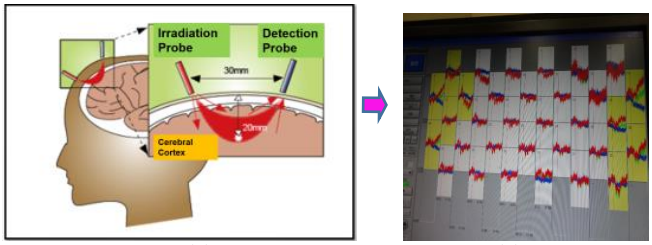
II. INSTRUCTIONAL STRATEGY

In the field of educational technology, Aptitude-Treatment Interaction (ATI) is an important element in planning to develop an effective course design. As ATI's concept and theoretical framework suggest that instructional strategies' effectiveness for individuals depends upon their specific abilities and optimal learning is achieved when the course design matches the learner's aptitude [2]. More suitable training needs to be applied to the less motivated EFL learners [3]. Although the effectiveness of various teaching methods and materials has improved, an assessment based on traditional paper and pencil tests has revealed its limitations [4]. Recently, brain activity has become subject to monitoring by technologically innovative instruments [5]. These technologies provide data that reveals the results of teaching and learning; therefore, these data can be utilized to assess the effectiveness of EFL teaching in Japan.

III. APPLYING NIRS TO COURSE DESIGN

The present study uses NIRS to analyze the amount of blood flow in the brain while learners were learning English. It then examined the relationship between brain activities and learning outcomes to identify the most effective combinations of learners' characteristics and English conversational skills teaching materials.

NIRS is widely recognized as a practical non-invasive optical technique to detect the hemoglobin density dynamics response during functional activation of the cerebral cortex, as shown in Figure 1 (a). The primary application of NIRS to the human body uses the fact that the transmission and absorption of NIR light in human body tissues contains information about changes in hemoglobin concentration. When a specific area of the brain is activated, the localized blood volume in that area quickly changes [6].



(a) (b)

Figure 1. (a) Route of near-infrared (modified from Shimadzu) and (b) Detected Channels : specific area of the brain

The greater the amount of blood flow, the greater the hemoglobin oxygenation increases; measuring the amount of blood can thus indicate the state of brain activation caused by differences among teaching materials, as shown in Figure 1 (b). This experimental technique indicated the well-matched combination of listening materials and training for EFL learners.

 ISBN: 978-1-61208-447-3

IV. ATI BASED EXPEIRMENT USING NIRS

The purpose of this study is to examine the effectiveness of analysis with NIRS by comparing the cerebral parts' activities for the effective course design for EFL learners by proposing the well-matched combination of EFL learners' characteristics and English conversation trainings. To resolve this purpose, ATI based experiment was planned examining the interaction of learners' aptitudes, materials and tasks as Figure 2.

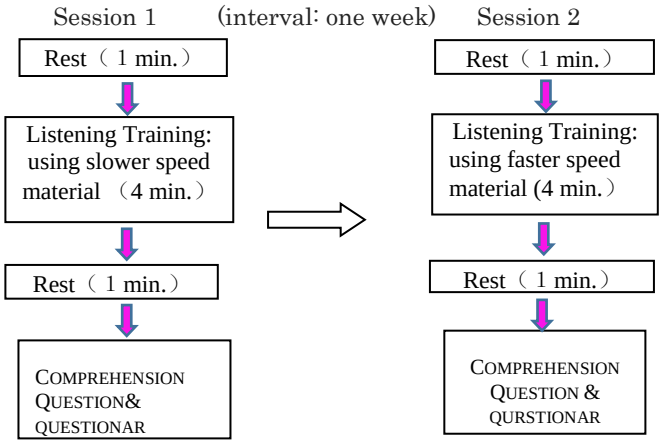


Figure 2 Experimental Protocol

Twelve participants of this experiment were divided into two groups; group A as higher level of proficiency (5 males) and group B as lower level of proficiency (7 males) assessed by Test of English for International Communication (TOEIC) scores. Each participant took part in two sessions and they were given different listening materials and tasks in each session.

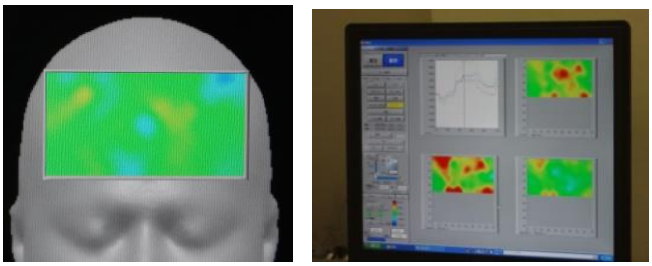
V. RESULTS AND CONCLUDING REMARKS

Comparing the average amount of change per second of Deoxy-Hemoglobin and Oxy-Hemoglobin in each channel during the training (see Figure 3), several significant interactions between learners' characteristics and trainings were found, as shown in TABLE I.

TABLE I. SIGNIFICANT RESULTS

Aptitude: Proficiency	Treatment: Speed of Material	Brain Activation	Average of Quiz Score(/15)
Group A: Higher	Faster	Activate	6
	Slower	Moderate or None	12
Group B: Lower	Faster	Moderate or None	1
	Slower	Moderate	5

In Group A, regardless of their high score of the quiz results, if the material was too easy or not interesting to them, the brain activity was moderate during their listening activity, as shown in Figure 3(a). In contrast, if the material was rather difficult for them, active brain activity was detected, as shown in Figure 3 (b). In Group B, if the task was too difficult, brain activity was similar to Figure 3 (a).



(a) (b)

Figure 3. Brain Activities : (a) Moderate Activity (b) Active Activiry

These results showed that teachers' ideals are not always enough to create effective training program, and it could be said that using analysis of NIRS for effective course design could be a very useful method.

ACKNOWLEDGMENT

This study is supported by a Grant-in-Aid for Scientific Research (C) (No. 26370672), from 2014 to 2016.

REFERENCES

[1] Ministry of Education, Culture, Sports, Science and Technology, Japan. *Survey on the Five Proposals and Specific Measures for Developing Proficiency in English for International Communication*. [Online]. Available from: http://www.mext.go.jp/english/elsec/1319701.htm 2015.08.30

[2] L. J. Cronbach and R. E. Snow, "Aptitudes and Instructional Methods: A Handbook for Research in Interactions". John Wiley & Sons Inc, New Jersey, 1977.

[3] D. Perani and J. Abutalebi, "The neural basis of first and second language processing," Current Opinion in Neurobiology, 15, 2, pp. 202-206, 2005.

[4] R. Tobita, "An Experimental Study on the Use of Metacognitive Learning Strategies of Reading Comprehension in English Learning," Bulletin of Saitama Women's Junior College. 13, pp. 207-234, 2002.

[5] T. Shimura ed., "Prefrontal Lobe Measurement Using Near Infraraed Spectroscopy – Evaluation of Early Detection Methods and Rehabilitation Methods of Dementia －", CORONA Publishing. 2009.

[6] Shimadzu. *LABNIRS*. [Online]. Available from: http://www.an.shimadzu.co.jp/bio/nirs/nirs2.htm 2015.08.30

 ISBN: 978-1-61208-447-3

Automatic Detection and Prevention of Cyberbullying

Cynthia Van Hee*, Els Lefever*, Ben Verhoeven†, Julie Mennes*, Bart Desmet*,
Guy De Pauw†, Walter Daelemans† and Véronique Hoste*
*LT3 - Language and Translation Technology Team
Faculty of Arts and Philosophy, Ghent University, Belgium
Email: firstname.lastname@ugent.be
†CLiPS - Computational Linguistics Group
Faculty of Arts, University of Antwerp, Belgium
Email: firstname.lastname@uantwerpen.be

***Abstract*—The recent development of social media poses new challenges to the research community in analyzing online interactions between people. Social networking sites offer great opportunities for connecting with others, but also increase the vulnerability of young people to undesirable phenomena, such as cybervictimization. Recent research reports that on average, 20% to 40% of all teenagers have been victimized online. In this paper, we focus on cyberbullying as a particular form of cybervictimization. Successful prevention depends on the adequate detection of potentially harmful messages. However, given the massive information overload on the Web, there is a need for intelligent systems to identify potential risks automatically. We present the construction and annotation of a corpus of Dutch social media posts annotated with fine-grained cyberbullying-related text categories, such as insults and threats. Also, the specific participants (harasser, victim or bystander) in a cyberbullying conversation are identified to enhance the analysis of human interactions involving cyberbullying. Apart from describing our dataset construction and annotation, we present proof-of-concept experiments on the automatic identification of cyberbullying events and fine-grained cyberbullying categories.**

Keywords–Cyberbullying prevention; Text classification; Dataset construction.

I. INTRODUCTION

The rise of Web 2.0 applications has substantially affected communication and relationships in today's society. Forums or message boards, blogs and social networking platforms like Facebook, Twitter, Tumblr or WhatsApp have become an important means of communication, especially among teenagers. Although most of the time, a child's Internet use is perfectly safe and enjoyable, there are risks involved in online communication through social media. Like offline communities, online communities can be harmful. Youngsters can be confronted with threatening situations, such as cyberbullying, suicidal behavior or grooming by paedophiles. As a response to those threats, a number of national and cross-national child protective initiatives (e.g., The Suicide Prevention Centre (http://www.preventiezelfdoding.be/), Child Focus (http://www.childfocus.be/)) have been starting projects over the last few years to increase online child safety. In spite of these efforts, much undesirable or even hurtful content remains online.

This research focuses on cyberbullying, one of the problems that emerged with the growing popularity of social media and its rapid adoption into our daily lives. Social media typically possess a number of features that make them a convenient way for cyberbullies to target their victims, including anonymity, lack of supervision and impact [1]. Whereas traditional bullying was originally limited to school yards and youth movements, cyberbullying can continue at home. Cyberbullies can reach their victim through technological devices, such as mobile phones and laptops at any time of the day. Moreover, online content is exposed to a large audience and is difficult to remove. A message can be re-posted, liked or shared, which substantially increases the impact of an offensive or hurtful message, even if it was posted only once [2]. Over the past years, cyberbullying has become an important problem. A recent study among 2,000 Flemish secondary school students revealed that 11% of them had been bullied at least once in the six months preceding the survey [3]. The large-scale EU Kids Online Report [4] revealed that 17% of 9- to 16-year-olds had been bothered or upset by something online in the past year. Juvonen et al. [5] found that no less than 72% of 12- to 17-year-olds encountered cyberbullying at least once within the year preceding the questionnaire. Tokunaga [6] found that cybervictimization rates among teenagers vary between 20% and 40% on average [1], [7], [8], [9]. The figures vary depending on location, interval and the conceptualizations researchers use to describe cyberbullying. All of them demonstrate, however, that online platforms are increasingly used for bullying and that cyberbullying is thus not a rare problem. Moreover, it poses a significant threat to a teenager's mental and physical well-being with studies linking cyberbullying to depression, low self-esteem and school problems [10], [11], [12]. In extreme cases, its effects have even been linked to self-harm [10] and suicide [13]. Successful detection of cyberbullying is therefore of key importance to identify possibly threatening situations online and prevent them from escalating. Given the massive information overload on the Web, it has become unfeasible for humans to keep track of all conversations produced online. In order to manage this amount of information in an efficient way, there is an urgent need for intelligent techniques to signal harmful content automatically. This would allow for large-scale social media monitoring and early detection of harmful situations, such as cyberbullying, suicidality and sexually transgressive behavior (e.g., paedophilia). Recent research on the desirability of such detection systems found that a major part of the respondents favoured automatic monitoring on the condition that effective follow-up strategies are included and that privacy and autonomy are guaranteed [14].

Dadvar [15], Dinakar et al. [16] and Reynolds et al. [17] describe some of the first forays into the automatic detection of cyberbullying. To the best of our knowledge, however, we present the first study on recognizing cyberbullying events

 ISBN: 978-1-61208-447-3

in social media content by means of a fine-grained textual annotation of the corpus, in addition to implementing a binary distinction (cyberbullying versus non-cyberbullying).

The main objective of this research is to gain insight into the linguistic characteristics of cyberbullying by collecting and annotating an adequate dataset. This will allow us to explore text characteristics (or *features*) that are potentially useful in distinguishing between cyberbullying and non-cyberbullying content. For the annotation of the data, we consider fine-grained categories related to cyberbullying, such as insults and threats [18]. Such a fine-grained distinction provides insight into various types of cyberbullying and the degree to which they are alarming (e.g., expressions of a threat are considered more alarming than a single insult). Moreover, typical roles in a cyberbullying event are annotated (i.e., bully, victim, bystander). This way, cyberbullying incidents can be reconstructed through its participants, which may provide clearer insight into the severity of the incident. For instance, cyberbullying incidents where bystanders defend the victim or discourage the bully from continuing might not be as alarming as those where a victim stands alone and feels powerless when faced with a bully. Finally, we investigate the feasibility of automatically recognizing potentially offensive or harmful messages in Dutch user-generated content. Such an automatic system could serve as a first filter that reduces the amount of incoming messages for human moderators. Several users are targeted here: child protection agencies, social care organizations, such as the Suicide Prevention Centre, as well as parents and teachers.

The remainder of the paper is structured as follows: in Section II, a brief literature review of studies that have focused on cyberbullying detection is presented. Our experimental corpus is described in Section III, as well as the data collection and annotation. Section IV gives an overview of the experimental setup and results. Finally, we draw conclusions and formulate directions for future research in Section V.

II. Related Research

Cyberbullying has been a widely covered research topic over the past few years, especially in the realm of social sciences. Studies have focused on the conceptualization of cyberbullying and the occurrence of the phenomenon [19], [20], [21]. Additionally, different types of cyberbullying have been identified [22], [23], [24] and the consequences of cyberbullying have been investigated [9], [10], [25]. More recently, studies have focused on the use of NLP techniques for the detection and prevention of cyberbullying. Yin et al. [26] applied a supervised machine learning approach for the automatic detection of cyberharassment. They combined local tf-idf features with sentiment features and features capturing the similarity between several posts and obtained an F-score of 0.44. Dadvar [15] applied a hybrid approach combining supervised machine learning models with an expert system that incorporates knowledge from a sociological and psychological point of view (e.g., identifying characteristics of potential bullies on social networks) to recognize cyberbullying. They showed that combining user information and expert views with lexical features, yields fairly good results (F = 0.64). Reynolds et al. [17] applied rule-based learning to develop a model for detecting cyberbullying based on textual features (e.g., the number of curse words in a message) and compared its performance to a bag-of-words model (i.e., based on a matrix of all the words that occur in the training corpus). They found that the rule-based method outperformed the bag-of-words model, achieving a recall of 78.5%. Dinakar et al. [27] conducted text classification experiments on a YouTube corpus. Using supervised machine learning and bag-of-words features, they built topic-sensitive classifiers to determine whether the topic of an insulting document is of a sensitive nature (i.e., sexuality, intelligence or race). In all of the aforementioned studies, however, cyberbullying detection is approached as a binary classification task (cyberbullying versus non-cyberbullying) without taking into account specific forms of cyberbullying such as threats, exclusions or insults. Moreover, these studies mainly focused on the detection of offensive posts written by a harasser, without specifying whether and how posts from victims and bystanders were considered. However, recent studies in the domain of automatic role assignment have emphasized the importance of community detection and role identification to enhance the analysis of online conversations [28].

The current research focuses on the detection of cyberbullying *events*, which include posts from harassers, as well as from victims and bystanders. We present two sets of experiments in which we explore 1) the detection of cyberbullying events (i.e., cyberbullying posts irrespective of the author's role) and 2) the classification of more fine-grained categories related to cyberbullying, such as threats and insults.

III. Dataset Construction and Annotation

The availability of suitable data represents an important challenge in research on cyberbullying. However, a suitable dataset is needed for building representative models for cyberbullying detection. This section describes the construction of a Dutch corpus of social media messages containing both cyberbullying and non-cyberbullying content.

A. Data Collection

We constructed a corpus by collecting data from the social networking site Ask.fm (http://ask.fm), by receiving donations and by setting up simulation experiments with volunteer youngsters. In total, 91,370 Dutch posts were collected.

Ask.fm A substantial part of our corpus was collected from the social networking site Ask.fm where users can create profiles and ask questions and answer them, with the option of doing so anonymously. Typically, Ask.fm data consists of question-answer pairs published on a user's profile. The data was retrieved by crawling a number of seed sites using the GNU Wget software (https://www.gnu.org/software/wget). After filtering out non-Dutch content this resulted in 85,462 posts. As the posts containing cyberbullying were underrepresented in the corpus, we started two initiatives to complement the dataset:

Donations Firstly, we launched a media campaign in which people were asked to donate evidence of personal cases of cyberbullying. This resulted in a rather small but highly topical set of messages including Facebook hate pages, message board posts and chat conversations.

Simulations Secondly, a series of simulation experiments were set up in which volunteer teenagers were asked to participate in a cyberbullying simulation on a social network

 ISBN: 978-1-61208-447-3

by means of a role-playing game. A social networking platform was designed that is comparable to Facebook using SocialEngine (http://www.socialengine.com).

TABLE I. DATA DISTRIBUTION FOR THE FINE-GRAINED TEXT CATEGORIES RELATED TO CYBERBULLYING.

Category	Positive Instances	Harmfulness Score		
		0	1	2
Threat/blackmail	204	-	137	67
Insult	4,265	381	3,796	88
Curse/exclusion	1,111	-	1,009	102
Defamation	162	-	160	2
Sexual talk	495	398	4	93
Defense	2,226	-	2,087	139
Encouragements to the harasser	42	-	41	1

B. Data Annotation

In order to keep track of harmful user-generated content, we developed a fine-grained annotation scheme for the analysis of textual cyberbullying which is detailed in Van Hee et al. [18] and applied it to our corpus. To provide the annotators with some context, all posts were presented within their original conversation where possible. The annotation scheme describes two levels of annotation. First, the annotators were asked to indicate, at the post level, whether a post is part of a cyberbullying event. This was done by assigning a harmfulness score to the post on a three-point scale, with 0 signifying that the post does not contain indications of cyberbullying, 1 that the post contains indications of cyberbullying although they are not severe, and 2 that the post contains serious indications of cyberbullying. When a post was considered to be part of a cyberbullying context (i.e., it was given a harmfulness score of 1 or 2), the annotators indicated the author's role in the cyberbullying event. In addition to victim and harasser, two types of bystanders are distinguished in our annotation scheme: 1) bystander-defenders, who help the victim and discourage the harasser from continuing his actions and 2) bystander-assistants, who do not initiate, but take part in the actions of the harasser. Secondly, at the subsentence level, the annotators were tasked with the identification of fine-grained text categories related to cyberbullying, even if the post was not considered harmful. For instance, in the sentence "Hey bitches, zin in een filmpje vanavond?" (*Hi bitches, anyone in for a movie tonight?*), *bitches* should be annotated as an insulting word. More concretely, they identified all text spans corresponding to one of the categories described in the annotation scheme. All annotations were done using the brat rapid annotation tool [29]. Table II presents the fine-grained cyberbullying categories and some example annotations of our dataset in brat.

In total, 85,462 Dutch posts were annotated by two annotators. To demonstrate the validity of our guidelines, inter-annotator agreement scores were calculated using Kappa [30] on a subset of the corpus. The Kappa score for the identification of cyberbullying events is 0.69. Kappa scores for the categories *Threat*, *Insult*, *Defense*, *Sexual Talk* and *Threat* range from moderate to substantial (i.e., from 0.52 to 0.66). They are low, however, for the categories *Defamation*, *Encouragements* and *Curse*, the identification of which seems to be rather difficult.

C. Experimental Corpus

For our preliminary experiments, we focused on the Ask.fm dataset. As shown in Table I, the experimental corpus features a heavily skewed class distribution with the large majority of posts not being part of any cyberbullying event. Regarding the occurrence of the fine-grained categories, we observe that insults are the most frequent type of cyberbullying activity in our corpus, followed by defense statements and curses/exclusions. Encouragements to the harasser is the least represented category. In this respect, it is worth mentioning that in case the annotators had too little context at their disposal to discern encouragements by bystanders from bullying acts by bullies, they annotated the post as a bullying act.

For each category, the number of instances marked with a harmfulness score of 0, 1 and 2 is given. As can be inferred from the table, 381 insults were identified in a non-cyberbullying context (e.g., insults as a 'socially accepted' way of addressing each other among friends). A major part of the category *Sexual talk* received a harmfulness score of zero, which means that these instances contained harmless sexual talk. Utterances considered sexual harassment were assigned a score of 1 or 2. If we consider the different roles in the annotated bullying events, we observe that the role of bully features in more than half of the annotated instances, followed by the victim role in about 30% of the instances. The bystander role in its two different subroles accounts for about 10% of the experimental corpus. These figures show that by focusing only on offensive posts (i.e., typical posts from a bully), as most studies on cyberbullying detection have done, about half of the relevant posts are ignored.

IV. EXPERIMENTS

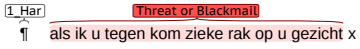
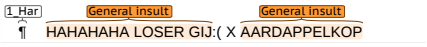
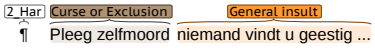
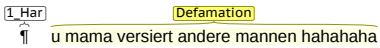
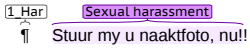
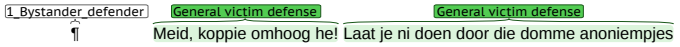
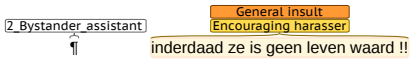
This section describes a set of preliminary experiments that were conducted to gain insight into the detection and fine-grained classification of cyberbullying events.

A. Experimental setup

We explored the feasibility of automatic classification of cyberbullying events (i.e., a binary classifier was developed for distinguishing cyberbullying from non-cyberbullying posts) and more fine-grained text categories related to cyberbullying. To this end, binary classifiers were built for each of these categories (see Table I for an overview of the fine-grained categories). For our experiments, we used Support Vector Machines (SVM) as the classification algorithm, since they have been proven to work well for high-skew text classification tasks similar to the ones under investigation [31]. We used linear kernels and experimentally determined the optimal cost value c to be 1. All experiments were carried out using Pattern [32]. As preprocessing steps, we applied tokenization, PoS-tagging and lemmatization to the data using the LeTs Preprocess Toolkit [33]. In supervised learning, a machine learning algorithm takes a set of training instances (of which the label is known) and seeks to build a model that generates a desired prediction for an unseen instance. To enable the model construction, all instances are represented as a vector of features (i.e., inherent characteristics of the data) that contain information that is potentially useful for distinguishing cyberbullying from non-cyberbullying content. For our experiments, we implemented two types of lexical features: bag-of-word features and polarity features based on

 ISBN: 978-1-61208-447-3

TABLE II. DEFINITIONS AND BRAT ANNOTATION EXAMPLES OF THE FINE-GRAINED TEXT CATEGORIES RELATED TO CYBERBULLYING.

Category	Brat annotation example	Translation
Threat/blackmail Expressions containing physical or psychological threats, or indications of blackmail.		*I'll smash you in the face when I see you x*
Insult Expressions containing abusive, degrading or offensive language that are meant to insult the addressee.		*HAHAHAHA YOU LOSER :(X POTATO HEAD*
Curse/exclusion Expressions of a wish that some form of adversity or misfortune will befall the victim and expressions that exclude the victim from a conversation or a social group.		*Just commit suicide, nobody thinks you're funny...*
Defamation Expressions that reveal confident, embarrassing or defamatory information about the victim to a large public.		*Your mom is flirting with other men hahahaha*
Sexual talk Expressions with a sexual meaning that are possibly harmful.		*Send me a naked picture of yourself, now!!*
Defense Expressions in support of the victim, expressed by the victim himself or by a bystander.		*Cheer up girl, don't let those stupid anons make you feel bad*
Encouragements to the harasser Expressions in support of the harasser.		*Indeed, she shouldn't be alive !!*

existing sentiment lexicons, resulting in a set of ~300.000 features in total. Bag-of-word features represent a corpus as an unordered set (or 'bag') of word or character sequences.

- **Word unigram and bigram bags-of-words:** binary features indicating the presence of word unigrams (i.e., a single word) and bigrams (i.e., a sequence of two words).
- **Character trigram bag-of-words:** binary features indicating the presence of character trigrams (without crossing word boundaries). A character-based bag-of-words representation is useful as it provides some abstraction from the word level and is more robust to variation in spelling or grammar.
- **Sentiment lexicon features:** polarity features that might be useful to provide insight into the polarity orientation of cyberbullying posts. To increase the lexicon coverage, lemmas were taken into account. The features are based on existing sentiment lexicons for Dutch [34], [35]:
  - The number of positive, negative and neutral lexicon words found in the text (averaged over text length).
  - The overall post polarity (i.e., the sum of the values of identified sentiment words, averaged over text length).

B. Results

This section presents the results of our preliminary experiments. Two classification tasks were carried out: cyberbullying event detection and the classification of fine-grained classification text categories related to cyberbullying. Evaluation was done using 10-fold cross-validation. As the evaluation metric we used F-score, which is the weighted average of the classifier's precision (i.e., the fraction of retrieved instances that are relevant) and recall (i.e., the ratio of the number of relevant instances that are retrieved). For the classification of cyberbullying events, our classifier obtains an F-score of 55.39%. F-scores for the fine-grained classification of cyberbullying vary considerably. As shown in Figure 1, the *Insult* classifier yields an F-score of 56.32%, whereas the classification performance for the categories *Encouragement* and *Defamation* is significantly lower with F-scores of 0.12% and 7.41%, respectively. In addition to data scarcity (e.g., only 42 positive instances for the *Encouragement* category), the large discrepancies in performance are presumably due to the extent to which a category is lexicalized. For instance, insults are generally highly lexicalized, whereas threats are often expressed in an implicit way.

As shown in Figure 2, the identification of cyberbullying events performs better in terms of precision than recall. Generally, the fine-grained cyberbullying categories show a good balance between precision and recall. Our experiments show satisfactory preliminary results, especially for the classification of bully events and insults. The best classification performance is obtained for fine-grained categories that are explicitly lexicalized (e.g., insults, sexual talk, defensive statements). This intuitively makes sense as we made use of lexical features to represent the data. The figures also show a correlation between

 ISBN: 978-1-61208-447-3

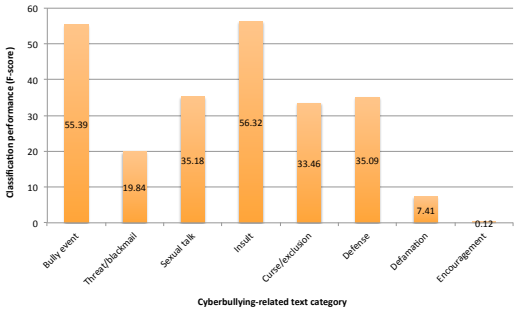


FIGURE 1. CLASSIFICATION RESULTS FOR THE IDENTIFICATION OF CYBERBULLYING EVENTS AND FINE-GRAINED CYBERBULLYING CATEGORIES, REPORTED AS 10-FOLD CROSS-VALIDATED F-SCORE ON THE POSITIVE CLASS (PERCENTAGES).

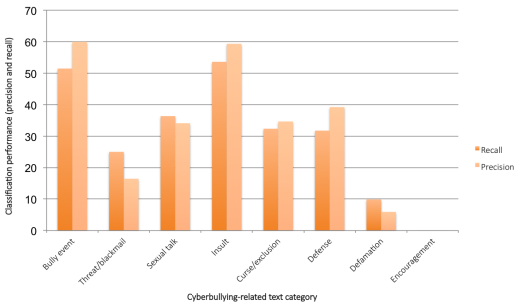


FIGURE 2. CLASSIFICATION RESULTS REPORTED BY MEANS OF PRECISION AND RECALL (PERCENTAGES).

the classification performance and the representation of the fine-grained category in our dataset. We therefore believe that the classification performance might benefit from extending the training corpus. The score obtained for the detection of cyberbullying events is in line with state-of-the-art approaches to automatic cyberbullying detection (e.g. Dadvar et al., 2014; Dinakar et al., 2012). Reynolds et al. [17] worked with data that is similar to ours (i.e., question-answer pairs) and reported an accuracy of 78.5% when the positive posts were overrepresented in the training corpus. However, the classification accuracy was lower (53.82%) when the model was applied to the original corpus where the distribution of the positive posts was left unchanged.

V. CONCLUSIONS AND FUTURE WORK

Web 2.0 offers a multitude of ways to communicate with peers. Both positive and negative experiences are abundant on the Web and children and youngsters are vulnerable groups in harmful online communication. In this paper, we constructed a Dutch dataset of social media messages containing cyberbullying and proposed and evaluated a methodology for adequate annotation of this data. Additionally, we explored the feasibility of automatic cyberbullying detection. Our initial results show that cyberbullying detection is not a trivial task, especially not when focusing on more fine-grained categories.

As the ultimate goal of automatic cyberbullying detection is to reduce manual monitoring efforts on social media, recall optimization will be the prior focus for further research as we want to flag as many online threats as possible for the moderator of a network. We will do a thorough qualitative analysis of the classification results to gain insight into the linguistic realization of cyberbullying and more specifically a series of fine-grained categories related to cyberbullying. We will also explore to what extent author role information can be used to enhance cyberbullying detection. A shallow error analysis revealed that implicit realizations of cyberbullying are fairly hard to recognize, as they are devoid of lexical cues such as profanity. Therefore, we will explore the use of more advanced features (e.g., syntactic patterns, semantic information) in addition to lexical features. Additionally, we will examine feature selection techniques to decrease vector sparseness and hence avoid the introduction of noise. Social media texts tend to deviate from the linguistic norm, which reduces the effectiveness of both lexical and more complex features. Another direction for future work will therefore be orthographic normalization of the data as a preprocessing step [36]. Finally, we will investigate the integration of techniques such as cost-sensitive learning, data resampling or one-class learning to tackle the severe class imbalance.

ACKNOWLEDGMENT

The work presented in this paper was carried out in the framework of the AMiCA (IWT SBO-project 120007) project, funded by the government agency for Innovation by Science and Technology (IWT).

REFERENCES

[1] S. Hinduja and J. W. Patchin, “Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying,” Youth Violence And Juvenile Justice, vol. 4, 2006, pp. 148–169.

[2] J. J. Dooley and D. Cross, “Cyberbullying versus face-to-face bullying: A review of the similarities and differences,” Journal of Psychology, vol. 217, 2010, pp. 182–188, ISSN: 0044-3409.

[3] K. Van Cleemput, S. Bastiaensens, H. Vandebosch, K. Poels, G. Deboutte, A. DeSmet, and I. De Bourdeaudhuij, “Zes jaar onderzoek naar cyberpesten in Vlaanderen, België en daarbuiten: een overzicht van de bevindingen. (Six years of research on cyberbullying in Flanders, Belgium and beyond: an overview of the findings.) (White Paper),” University of Antwerp & Ghent University, Tech. Rep., 2013.

[4] “EU Kids Online: findings, methods, recommendations.” 2014, URL: http://eprints.lse.ac.uk/60512/ [accessed: 2015-07-30].

[5] J. Juvonen and E. F. G, “Extending the school grounds?-Bullying experiences in cyberspace,” Journal of School Health, vol. 78, 2008, pp. 496–505, ISSN: 1746-1561.

[6] R. S. Tokunaga, “Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization,” Computers in Human Behavior, vol. 26, 2010, pp. 277–287, ISSN: 0747-5632.

[7] F. Dehue, C. Bolman, and T. Vollink, “Cyberbullying: Youngster’s Experiences and Parental Perception,” CyberPsychology, vol. 4, 2006, pp. 148–169.

[8] Q. Li, “New Bottle but Old Wine: A Research of Cyberbullying in Schools,” Computers in Human Behavior, vol. 23, 2007, pp. 1777–1791, ISSN: 0747-5632.

[9] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, “Cyberbullying: its nature and impact in secondary school pupils,” Journal of Child Psychology and Psychiatry, vol. 49, 2008, pp. 376–385.

[10] M. Price and J. Dalgleish, “Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People.” Youth Studies Australia, vol. 29, 2010, pp. 51–59, ISSN: 1038-2569.

[11] V. Šléglová and A. Černá, “Cyberbullying in Adolescent Victims: Perception and Coping,” Cyberpsychology: Journal of Psychosocial Research on Cyberspace, vol. 5, 2011, ISSN: 1802-7962. [Online]. Available: http://cyberpsychology.eu/view.php?cisloclanku=2011121901&article=4

 ISBN: 978-1-61208-447-3

[12] H. Vandebosch, K. Van Cleemput, D. Mortelmans, and M. Walrave, "Cyberpesten bij jongeren in Vlaanderen: Een studie in opdracht van het viWTA (Cyberbullying among youngsters in Flanders: a study commissoned by the viWTA). Brussels: viWTA," 2006, URL: http://ist.vito.be/nl/publicaties/rapporten/rapport_cyberpesten.html [accessed: 2015-07-30].

[13] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," Archives of Suicide Research, vol. 14, 2010, pp. 206–221, ISSN: 1381-1118.

[14] K. Van Royen, K. Poels, W. Daelemans, and H. Vandebosch, "Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability," Telematics and Informatics, vol. 32, 2015, pp. 89–97, ISSN: 0736-5853.

[15] M. Dadvar, D. Trieschnigg, and F. de Jong, Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies. Springer International Publishing, Jan. 2014, pp. 275–281, in Sokolova, M. and van Beek, P., Advances in Artificial Intelligence, ISBN: 978-3-319-06482-6.

[16] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying," ACM Transactions on Interactive Intelligent Systems, vol. 2, 2012, pp. 1–30, ISSN: 2160-6455.

[17] K. Reynolds, A. Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyberbullying," in Proceedings of the 10^{th} International Conference on Machine Learning and Applications and Workshops December 18–21, 2011, Honolulu, Hawaii. IEEE Computer Society, Dec. 2011, pp. 241–244, IEEE, ISBN: 978-0-7695-4607-0, URL: http://dx.doi.org/10.1109/ICMLA.2011.152 [accessed: 2015-07-30].

[18] C. Van Hee, B. Verhoeven, E. Lefever, G. De Pauw, W. Daelemans, and V. Hoste, "Guidelines for the Fine-Grained Analysis of Cyberbullying, version 1.0," LT3, Language and Translation Technology Team–Ghent University, Tech. Rep. LT3 15-01, 2015.

[19] S. Hinduja and J. W. Patchin, "Cyberbullying: Neither an epidemic nor a rarity," European Journal of Developmental Psychology, vol. 9, 2012, pp. 539–543.

[20] S. Livingstone, L. Haddon, A. Görzig, and K. Ólafsson, "Risks and safety on the internet: The perspective of European children. Full findings," 2011, URL: http://eprints.lse.ac.uk/33731/ [accessed: 2015-07-30].

[21] R. Slonje and P. K. Smith, "Cyberbullying: Another main type of bullying?" Scandinavian Journal of Psychology, vol. 49, 2008, pp. 147–154.

[22] P. B. O'Sullivan and A. J. Flanagin, "Reconceptualizing 'flaming' and other problematic messages," New Media & Society, vol. 5, 2003, pp. 69–94, ISSN: 14614448.

[23] H. Vandebosch and K. Van Cleemput, "Cyberbullying among youngsters: profiles of bullies and victims," New Media & Society, vol. 11, 2009, pp. 1349–1371.

[24] N. E. Willard, Ed., Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Research Publishers LLC, 2007, ISBN: 978-087822-537-8.

[25] H. Cowie, "Cyberbullying and its impact on young people's emotional health and well-being," The Psychiatrist, vol. 37, 2013, pp. 167–170, ISSN: 1758-3209.

[26] D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0," in Proceedings of the Content Analysis in the Web 2.0 (CAW2.0) April 21, 2009, Madrid, Spain. CAW 2.0, Apr. 2009, pp. 1231–1238, CAW 2.0, URL: http://wbox0.cse.lehigh.edu/~brian/pubs/2009/CAW2/harassment.pdf [accessed: 2015-07-25].

[27] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," in Proceedings of the 5^{th} International AAAI Conference on Weblogs and Social Media, July 17–21, 2011, Barcelona, Spain. AAAI, Jul. 2011, pp. 11–17, AAAI, ISBN: 978-1-57735-505-2, URL: http://dblp.uni-trier.de/db/conf/icwsm/smw2011 [accessed: 2015-07-21].

[28] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email," Journal of Artificial Intelligence Research, vol. 30, 2007, pp. 249–272, ISSN: 1076-9757.

[29] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "BRAT: A Web-based Tool for NLP-assisted Text Annotation," in Proceedings of the Demonstrations at the 13^{th} Conference of the European Chapter of the Association for Computational Linguistics April, 23–27, 2012, Avignon, France. Association for Computational Linguistics, Apr. 2012, pp. 102–107, ACL, ISBN: 978-1-937284-19-0, URL: http://dl.acm.org/citation.cfm?id=2380921.2380942 [accessed: 2015-07-24].

[30] J. Cohen, "A coefficient of agreement for nominal scales," Educational and Psychological Measurement, vol. 20, 1960, pp. 37–46.

[31] B. Desmet and V. Hoste, "Recognising suicidal messages in Dutch social media," in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC) May 26-31, 2014, Reykjavik, Iceland. European Language Resources Association (ELRA), May 2014, pp. 830–835, ELRA, ISBN: 978-2-9517408-8-4, URL: http://www.lrec-conf.org/proceedings/lrec2014/index.html [accessed: 2015-07-26].

[32] T. De Smedt and W. Daelemans, "Pattern for Python," Journal of Machine Learning Research, vol. 13, 2012, pp. 2063–2067, ISSN: 1532-4435.

[33] M. van de Kauter, G. Coorman, E. Lefever, B. Desmet, L. Macken, and V. Hoste, "LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit," Computational Linguistics in the Netherlands Journal, vol. 3, 2013, pp. 103–120, ISSN: 2211-4009.

[34] T. De Smedt and W. Daelemans, "'Vreselijk mooi!' ('Terribly Beautiful!'): A Subjectivity Lexicon for Dutch Adjectives," in Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC) May 23–25, 2012, Istanbul, Turkey. European Language Resources Association (ELRA), May 2012, pp. 3568–3572, ELRA, ISBN: 978-2-9517408-7-7, URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/312_Paper.pdf [accessed: 2015-07-30].

[35] V. Jijkoun and K. Hofmann, "Generating a non-English Subjectivity Lexicon: Relations That Matter," in Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL) March 30–April 3, 2009, Athens, Greece. Association for Computational Linguistics, Apr. 2009, pp. 398–405, ACL, URL: http://www.aclweb.org/anthology/E09-1046 [accessed: 2015-07-27].

[36] S. Schulz, G. De Pauw, O. De Clercq, B. Desmet, V. Hoste, W. Daelemans, and L. Macken, "Multi-Modular Text Normalization of Dutch User-Generated Content," ACM Transactions on Intelligent Systems and Technology, 2015 (in press).

 ISBN: 978-1-61208-447-3

Aircraft in Your Head: How Air Traffic Controllers Mentally Organize Air Traffic

Linda Pfeiffer, Georg Valtin, Nicholas Hugo Müller and Paul Rosenthal

Technische Universität Chemnitz, Visual Computing Laboratory and Institute for Media Research
Chemnitz, Germany
Email: {georg.valtin, nicholas.mueller}@phil.tu-chemnitz.de,
{linda.pfeiffer, paul.rosenthal}@cs.tu-chemnitz.de

***Abstract*—The StayCentered Project at Technische Universität Chemnitz aims for assisting air traffic controllers in stressful traffic situations. Therefore we are seeking to comprehend air traffic controllers' principles of operation within the dyadic team structure. First exploratory research revealed insights into air traffic controllers' practices, their information processing (mental models), potential stressors, and related emotional effects. This paper discusses the results and the implications for air traffic controllers' work in general and the StayCentered project in particular.**

Keywords–Air traffic control; HCI; Decision Support; Mental Models

I. INTRODUCTION

In the state of normal operation, a dyad of two air traffic controllers is responsible for any given airspace. Both have access to task relevant information, such as radar data, weather reports, and flight schedules. An impression of a German air traffic controller's workplace is shown in Figure 1. Within the dyad, the air traffic controllers take different roles: one (executive) is responsible for the communication with the pilots using spoken traffic commands over the radio, while the other one (planner) is coordinating the acceptance or handover of flights from or to other sectors. This is necessary, since each sector has its individual operation of flight-levels and is generally only accepting flights within a certain flight-level threshold in order to keep a smooth vertical alignment between adjacent flights. While arranging the handovers, the planner is also responsible to verify the communication between the executive and the pilots and to intervene, if necessary. Therefore, the division of responsibilities is depending on a good internal communication as well as a transparent work situation. Expediting and maintaining orderly traffic flows can be characterized as the main goal of air traffic controllers' work. However, the adherence to strict separation standards for safety reasons sets nonnegotiable rules that act as constraints [1, p 341]. The combination of these two characteristics results in a demanding work, especially because air traffic controllers have to make most of their decisions in a narrow time frame [2][3]. Due to the characteristics of their work and the general limitations of the human ability to process information, air traffic controllers often experience time pressure [1, p 339] that can lead to a stress response. A stress response is the activation of several physiological systems on the affective, cognitive, neural, endocrinal, and muscular level [4] when individuals are facing a stress inducing stimulus (stressor). However, stress is not per se a negative state, since the evaluation of the stressor depends on the interplay of the situational demands and the abilities of the individual to cope with the situation [5]. Since time pressure is a situational characteristic in the daily work of air traffic controllers, the occurrence of negative stress and its emotional and psychological consequences (short term: anxiety, despondence, anger, cognitive impairments; long term: fatigue, health issues, depression) is likely (see for instance [6][7]). Therefore, the reduction or rather avoidance of stress inducing situations is an important goal in the daily work of air traffic controllers.

Figure 1. An air traffic controller dyad at the German air traffic control center in Munich (Source: DFS Deutsche Flugsicherung GmbH).

Within this paper we will introduce the StayCentered project context that motivates our research. Section III outlines the methods used to gain the findings that will be discussed in Section IV. The following sections relate findings to the project context in terms of the future mental and emotional model as well as future interfaces. Section VII summarizes and concludes the paper.

II. THE STAYCENTERED PROJECT

Typically, the work of an air traffic controller involves managing various flight routes, aircraft, and altitude as well as air speed differences. Additionally, meteorological circumstances, technical maintenance activities or, in rare circumstances, emergencies can occur at any given moment and require swift and correct reactions by the controller. As air traffic controllers often work in dyads, in order to have an inherent corrective at all time and to provide redundancies, the StayCentered project at Technische Universität Chemnitz aims for enhancing the already high security standards of air traffic controllers, and for identifying as well as for offering assistance within cognitive stressful flight situations. Therefore, the dyadic team structure has to be analyzed comprehensively: both their voiced interactions between themselves and with the pilots within

 ISBN: 978-1-61208-447-3

their controlled airspace. The goal is to be able to identify human error potential in voicing commands, interpreting visual data representations and to identify limits in cognitive processing capabilities. The resulting model of a working controller dyad is then used to simulate the emotional and cognitive state of the dyad in regards to upcoming air traffic some hours in advance. For example, planned but delayed flights (e.g., a sandstorm in Dubai and a thunderstorm in Moscow) will lead to an increased number of aircraft in their destination sector. Flight control management would then be able to split sectors and to call in additional controllers in order to keep the workload at a comfortable level. In addition, the controller stations themselves already offer the possibility for the controllers to signal an increased workload. However, the implementation of the projects biophysiological measurements would allow for an objective and immediate feedback to the controllers about their current cognitive state and troubleshooting capabilities [8], as well as for a workload regulation [9]. Therefore, the galvanic skin response, facial action coding, body posture, vocal properties, eye movements and pupil dilation are recorded and used to infer an emotion valence, arousal level, and cognitive load [10].

III. Methodology

To assess whether or not an air traffic controller experiences stress and the associated negative emotions, it is necessary to fully understand how the controller is receiving and processing the crucial information and how this is converted into practical actions. Since it is not possible to gain insight into the information processing objectively from the outside, it is necessary that the air traffic controllers verbalize their cognitive processes. For this purpose, we used semi-structured interviews outside the work situation to gather general information about how air traffic controllers experience work-related stress and how they cope with it. Among others, we let them describe exceptional situations which were especially demanding, how they solved them, and how they felt afterwards. Furthermore, we used the thinking-aloud approach in interviews to get a basic understanding on how air traffic controllers process information. We confronted them with a typical radar screen printout. The sector and scenario were unknown to the participants. It described a situation containing 8 aircraft, a mid term conflict of two aircraft with same heading and differing speed and a lateral conflict of two aircraft with opposite heading, but vertically divided. We asked them to evaluate the given flight situation regarding the salience of important information as well as the order in which critical data are perceived and processed. Additionally, we observed the air traffic controllers during their work at the level of moderate participation, allowing us to ask specific questions. Here, we also used the thinking-aloud approach to get information and explanations about certain actions and events. The observation under real working conditions is especially important since cognitive and emotional reactions are known to be a combination of person and situation, and thus only the inclusion of the given situational characteristics allows for a meaningful interpretation of the data gathered in the interviews. We decided to use this combination of methods in an exploratory approach in order to get the information of the air traffic controllers as authentic and natural as possible. Expressing thoughts, ideas and considerations in their own words in an actual work situation as well as in the reflecting, meta-cognitive form of an interview appears to be the adequate methodical approach for this kind of research problem. The data was collected between February and April 2015 at the facilities of the Deutsche Flugsicherung (DFS) in Langen and Munich. To assure a sufficient variability in the data, we interviewed and observed experienced and novice air traffic controllers likewise. Altogether, we collected data of N=21 air traffic controllers (age: 18 to 57). Since the evaluation of the air traffic controllers' work requires a basic level of expertise regarding the work station, work processes and air traffic, all researchers received an introduction to the air traffic controller's work by an expert of the DFS before data collection. Since recording audiovisual material is problematic due to security reasons, all interviews and observations were recorded by pen and paper. For the purpose of the analysis, all data was coded and categorized. Due to the exploratory nature of the research, we did not follow a standardized coding scheme. Instead, we tried to identify all relevant factors regarding the cognitive and emotional constitution and experiences of the air traffic controllers in relation so the given work situation.

IV. Findings and Discussion

By fulfilling their daily tasks, air traffic controllers face highly demanding situations. They need to process plenty pieces of information simultaneously that are arriving on multimodal channels (primarily auditive and visual). Based on this information, controllers have to make quick and reliable decisions to ensure safety of the aircraft, and thus people, under their control. When thinking of the air traffic controllers task, we first thought about the controller sitting in front of the radar screen and scanning the actual flight situation for potential conflicts all along. But according to our data, the radar screen is most of the time rather a secondary tool that is used to check whether every aircraft behaves the way it should. Usually, the air traffic controller is creating an internal representation of the current flight situation reaching about three minutes into the future. This picture is mainly build upon flight plan data that can be accessed via the (digital) flight strips, the controller's experience, and internalized knowledge about standard routes and so forth. A schematic diagram of the air traffic controllers' mental situation is depicted in Figure 2. According to Mogford [11], such a picture consists of situation awareness that is based upon the controller's mental models. While in literature mental models of air traffic controllers are often described as somewhat three dimensional models [11][12], our controllers explicitly stated that they do not build up a three dimensional model of the situation. They described it as a two dimensional model, similar to the radar display that is expanded by a variable indicating vertical layers. Other studies revealed that air traffic controllers [13] and already controller students [14] do not necessarily build up a three dimensional mental model. They develop an individual mental structure to represent the three dimensional data over time. In standard situations, when a flight strip is appearing on the controller's screen representing an airplane that is about to enter the sector soon, he is first looking for the route it is tending to take and at which flight level. For first conflict detection, the controller is checking overflight times at the fixes. If overflight times are overlapping ten minutes to the ones of another flight he marks a potential conflict. When the involved aircraft appears on the radar screen, the controller is checking a second time for the conflict and then improving gradually the quality of his prediction

 ISBN: 978-1-61208-447-3

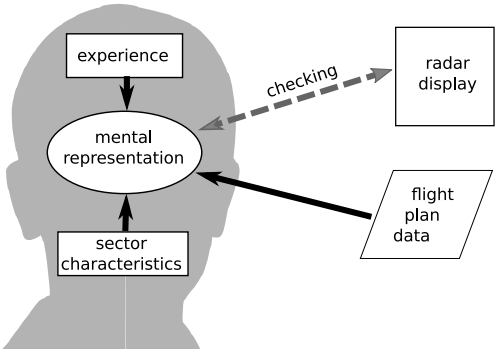


Figure 2. Schematic diagram of the air traffic controllers mental situation representation and its sources.

about a possible problem. First, he is estimating vertical and horizontal separations according to his experience (rule of thumb). He can use distance measuring tools provided by the system on the radar screen, but he is also able to do exact calculations using mental arithmetics, if necessary.

However, there are still situations requiring the air traffic controller to construct the current situation from flight strips and radar screen (e.g., during a hand over, the controller gets a description on the situation while scanning the radar and the strips). The first variable controllers focus on, while scanning the radar display, is altitude information and whether an aircraft is climbing or descending. After this, aicraft's heading and position are considered and lastly, ground speed gives a hint on the existence of a potential conflict. Rantanen et al. identified in their experiments [1] altitude as the information that is processed first for conflict detection. Furthermore, Mogford et al. emphasized altitude and heading as the most important information for air traffic controllers' situation awareness [11]

During interviews the air traffic controllers mentioned three main stressors:

High Traffic Load

The crucial factor for traffic load is the number of aircraft under control. However, the resulting workload goes beyond the sheer number. The structure of the airspace and standard routes as well as directions of the aircraft have an impact on perceived complexity. Plenty of vertical movements, as in approach sectors and sectors in the lower airspace, and lots of crossing trajectories increase the probability for potential conflicts.

Unexpected Events in the Airspace

Since air traffic controllers tend to have a detailed picture of upcoming events, unconsidered events may cause additional load, since they often require a swift reaction while simultaneously adding a unknown variable to their calculations. Usually, these are events that are neither listed in nor logical consequences of flight plan data. Initially, we considered emergency flights as unexpected things causing stress because air traffic controllers have to clear the way for them. However, most of the emergencies will already be marked in the actualized flight plan by the pilots. Thus, they can be regarded as expected traffic, just with a higher priority, making them just another variable in the air traffic controller's mental model. Even closures of single airports aren't surprising, because every flight has an alternative destination stated in its flight plan. However, an unplanned aircraft calling in or flights within their sector boundaries, which are not under their control, are stress inducing factors. Hence, a pilot who forgot next sectors frequency, just asking for it once again, may cause more confusion than emergencies, because the controller already deleted the associated flight strip and thus also removed the flight and callsign already from his mental model.

Malfunction of Equipment

Generally, the air traffic controller is dependent on his equipment. Without radar display the controller has to rely on the pilots following his instructions without any misapprehensions. Without flight plan data, the controller would loose the ability to proactively regulate air traffic. Still, air traffic controllers emphasized especially malfunction of the radio as problematic. Without the ability to communicate with the pilots the air traffic controllers are completely incapable of action. They don't know about pilots' plans and aren't able to forewarn them of an upcoming danger.

Other Things Indirectly Being Relevant

For efficiently building their picture, air traffic controllers rely mostly on their experiences and internalized information, such as standard routes and sector borders. If controllers are returning after a period of absence (e.g., illness or holidays), they perceive their work as more demanding, due to changes in standard routes, sector boundaries, or agreements. Also, other impact factors like general well-being, mood, private problems etc. were mentioned by air traffic controllers to influence the work performance. Therefore, personal factors often change the perceived demands. According to the air traffic controllers' experience, the same workload can be experienced differently.

These results on potential stressors align with the five most stressful items found by Brink [15]. South African air traffic controllers rated the number of aircraft, extraneous traffic, unforeseeable events, peek hour traffic and limitations, and reliability of the equipment to be most stressful factors out of a questionnaire with 20 items.

Let's have a closer look at the term 'team'. When we initially used the word 'team' within the context of air traffic controllers, we had the air traffic controller dyad in mind. However, the air traffic controller's understanding of 'team' covers more than initially assumed. On the one hand, they used the term when speaking about all the air traffic controllers responsible for German airspace and adjacent sectors. When recognizing a potential conflict situation that would happen in the neighboring sector, but could be prevented or already solved within their own sector, they would do so. When recognizing a conflict situation within an other sector, they would warn the responsible controller. When recognizing that controllers responsible for an adjacent sector have high traffic load and they are stressed, they try to keep further traffic away from that sector or try to avoid more stress for their

 ISBN: 978-1-61208-447-3

colleagues by organizing the flights in their own sector in a way that makes them easy to handle in the next one. This understanding of a team is also supported by the fact that air traffic controllers are on a first-name basis with each other. On the other hand, the term 'team' was used while talking about the air traffic controller's organizational entity. In German air traffic control centers, there are groups of air traffic controllers that are responsible for several neighboring sectors. These sectors share borders and in times of low traffic load they can be combined. Each controller out of this group has the admission to work on every position within these sectors. Thus, each of the controllers will somewhen constitute a dyad with every other controller out of this group.

However, good communications and collaboration with the air traffic controller's colleagues, his supervisor, and the pilots is critical to safety and efficiency in air traffic. We determined the following typical forms of communication. Short-term collaboration with pilots consists of speech over radio using predefined terms and routines in order to minimize the number of misunderstandings. The supervisor communicates to the controllers through the display of a duty roster. Change requests, rapid updates, and the like encourage both to move to and talk to each other, although telephone connection is available. Most of the communication is happening within the dyad responsible for a sector. Both, the executive and the planner have to build up a shared picture of the situation. In order to do so and to solve potential conflicts, they communicate using gestures (pointing gestures to guide the others attention onto the screen, sometimes they are also using the other one's mouse), and the flagging of (digital) flight strips (to highlight potential conflicts), but also in this case is speech the dominant communication channel. In times of high traffic load, controllers are sitting up straight, speaking concisely about traffic concerns. In times of low traffic load they are more relaxed and they are chatting with each other and the surrounding controllers. Usually, the planner is talking to other dyads via telephone, except for the ones sitting spatially near to him. Another tool for controller to controller communications is the so called "'Geneva traffic light"'. In German control centers there is a display assigning a color (green, yellow, red) to each sector with green being the default color for normal traffic load. By setting this color, controllers can communicate their actual workload to other controllers and the supervisor. However, if there is really high traffic load controllers are so focused on their work that they often don't think about changing the color.

Emotional aspects within the air traffic controller's work include awareness of own sentiments and awareness of the emotional state of others. Generally, controllers stated that there are no crucial emotional situations. Sometimes private problems cause the controller to "concentrate a little more" but usually they know how to act out of them. After a critical situation at work they are in need of someone to talk to. Often they prefer talking to their colleagues about it. During follow-up discussions some other situations were identified. Air traffic controllers said they are feeling proud, after managing a tricky situation smoothly. They have a sense of delight, when pilots thank them for satisfying their wishes (e.g., a direct). During long periods of low traffic the predominant sense is boredom. The most important indicator for the others emotional state is the sound of their voice and their choice of words, especially during communication using telephone or radio. Succinct answers indicate elevated concentration. During communication with their spatial neighbors, gestures and poses can be accessed additionally for emotional awareness.

V. Implications for the Mental and Emotional Model

From a psychological point of view, it is not surprising that the mental and emotional states of air traffic controllers are influenced by personal as well as situational characteristics. However, without a detailed analysis of the air controllers work, it is impossible to specify the relevant variables and their parameter values. Based on the collected data we are now able to consider precise variables in our model. Regarding the situational aspects, the number of aircraft as well as their flight characteristics are the main aspects for potential workload, and of course the available time is also relevant. Further research is necessary to identify the concrete relationship between those variables. However, it is already clear that there is limit on how many interactions can take place between the air traffic controller and pilots, because every interaction takes several seconds. Considering the well-known relationship between arousal and performance on difficult tasks [16], such as the work of an air traffic controller, we assume that the optimal efficiency lies far below the physical limit of interactions. It has still to be determined how to express the comfort zone of air controllers by an index. One potential solution is the indication of interactions per minute with the option to weight interactions depending on the situation's complexity. The model requires two kinds of critical values for the index that signals a possible overload: One is relating to situational peaks which can be understood as episodes of high workload in a rather short time frame. The other one is applying to longer periods of time with an increased workload that is higher than the optimum but lower than the situational peak. Both kinds of overload can result in mistakes, incorrect decisions or just slower reactions and must be prevented. Even though German air traffic controllers can be considered a homogenous group of specialists who are able to work under pressure, the critical values must be personalized due to differences in personality related factors. Our data suggest that many typical personality variables affect the work of air traffic controllers, such as mood, alertness, work experience, private problems, absence due to vacation, or sickness, etc. The main problem for the consideration of those variables is their problematic measurement: Many of them are only available to the air controllers themselves, and even they are not always able to fully specify all factors that might influence their performance or to quantify them. Furthermore, many of those variables are changing on a daily basis, even though they should not fluctuate that much. The personality factors can therefore be used to improve the index based on the situational variables. Simply put: The critical values can be adjusted depending on how an air traffic controller feels - if this information is available - or based on objective information like the absence of a controller for several weeks which lets him experience the work as more demanding during the first days of his work. For a short-term evaluation of the air controllers state, additional diagnostics will further improve the determination of the personality variables influence. Additionally, a cross-validation and combination, respectively, with psychophysiological parameters, eye-tracking, voice characteristics, facial emotion expression as well as poses and gestures will also

 ISBN: 978-1-61208-447-3

help to classify flight situation regarding their complexity. For instance, a more complex problem will result in longer times of fixation on the involved flights, an increased skin conductance, shorter voice-commands, a straighter body position and a stern facial expression. Our model must therefore take many variables into account, some global and some situational. Things become even more complicated, since the air traffic controllers are usually working together as a dyad. The model has to take into consideration not only the individual parameters, but also the specifics of the team. The same flight situation in a sector might result in excessive demands for one dyad but present an acceptable challenge to a team of veterans. This additional set of team related variables complicated the model, since questions about the structures and relationships between all the variables contained in the model are not fully answered yet.

VI. Implications for the User Interfaces

One of the main goals of the StayCentered project is to identify and to offer assistance within cognitive stressful flight situations. Current interfaces have to be rethought in order to give access to the identified and simulated mental and emotional states. The StayCentered interfaces will be designed to give decision support to the supervisor, to facilitate cooperation, and to adapt with respect to the controller's current state. At the moment the supervisor's decision upon splitting up a sector is done by consulting workload predictions, mainly based on the expected number of aircraft, and controllers demands. The StayCentered supervisor interface will present the simulation's forecasts. Anticipated stressful situations should be visible at a glance and supporting decision making on resolving these situations. As described above, cooperation and communication are crucial elements of the air traffic controllers daily work. These communicative situations shall be supported by the interfaces. There should be adequate ways of communication so that controllers don't have to leave their position for consulting their supervisor. The controllers workspace should be designed in such a way that the actions of one controller are clearly visible to his partner in the dyad. Thus, we are expecting to support the creation of a shared mental model and enhancing communication. The most obvious advantage of the StayCentered system is that the mentioned "Geneva traffic light" can change its color automatically. However, also it's presentation could be enhanced. Currently, each sector is represented by a colored button (green-yellow-red) on a secondary screen. Additional short textual remarks for the sectors in stressful situations are available. A graphical integration of this information into the radar screen would make it accessible at a glance. The interface adaption with respect to the controller's state applies to the interfaces at the controllers workspace. Currently, German air traffic controllers get their information via a plan-view radar screen (aircraft related data), a digital strip board (flight plan data), and several secondary screens (additional information like weather data or the "Geneva traffic light"). For a more detailed description of the controllers workspace see [17]. The information presentation is independent of the emotional state of the controller, the workload, and the complexity of the actual flight situation. However, the importance of information objects differs from situation to situation. StayCentered controller interfaces will consider the identified emotional state of the controller as an indicator for the chosen representation. And it is not just about assisting the controller in stressful situations. In times of low traffic load, controllers often feel bored. Since boredom has a negative impact on their attention, we want to consider these situations within the design of the adaptive interface as well. Good user interfaces support the user's mental models. Recent research on air traffic controllers' interfaces often considers three-dimensional radar displays (for an overview on current research see [17]). According to our data, the controller's mental model of a flight situation is not necessarily three-dimensional. Therefore we would prefer a two-dimensional representation and allow for stepwise adaption to required accuracy. Nevertheless, the aircraft's altitude is still extremely important for detecting potential conflicts and should consequently be considered with high priority within the radar visualization.

VII. Summary and Conclusion

Within this paper we described the StayCentered project at Technische Universität Chemnitz that aims for assisting air traffic controllers' work by identifying and simulating the air traffic controller dyad's mental and emotional states. Within this context we presented the results of our preliminary study and discussed its implications for the mental and emotional models as well as for the user interfaces. We identified high traffic load with plenty of vertical movements, unexpected events and a malfunction of the equipment as the most relevant stressors in air traffic control. Furthermore, stress level is influenced by personal factors. Surprisingly, the controllers stated not to create a three-dimensional mental representation of flight situations. The information used to create the mental representation consists of internal knowledge about the sectors characteristics and standard routes, their experience and flight plan data. For checking the current situation, information is processed in the following order: altitude, climb/descent, horizontal position, heading, and speed on ground.

The order of information processing should be reflected within the user interfaces, as well as the structure of the air controllers mental model. Identified forms of communication should be supported. The automatic recognition of the air traffic controller's workload and emotional state allows for further improvement in the workflow.

Our findings suggest that sufficiently modeling the cognitive and emotional states of air traffic controllers requires the inclusion of many variables regarding the individual controllers as well as the dyad and the current workload. The next steps in the process of model building are the identification of other relevant variables and generally their measurement and further processing. Even though we already know that cognitive and emotional states can be recognized using our multidimensional approach, the relationships between the variables still needs further research. Possible methodological approaches include the recording of actual or simulated work sessions in combination with post-hoc interviews in order to identify critical or demanding situations. By comparing the measurement data with the information given by the controller, we can identify typical patterns that signal stressful episodes which can be used in our model.

References

[1] E. M. Rantanen and A. Nunes, "Hierarchical conflict detection in air traffic control." International Journal of Aviation Psychology, vol. 15, no. 4, 2005, pp. 339 – 362.

 ISBN: 978-1-61208-447-3

[2] V. D. Hopkin, Human Factors in air traffic control. Taylor & Francis, 1995.

[3] R. J. Roske-Hofstrand and E. D. Murphy, "Human information processing in air traffic control," in Human Factors in Air Traffic Control, M. W. Smolensky and E. S. Stein, Eds. Academic Press, 1998, ch. Human information processing in air traffic control, pp. 65–114.

[4] J. Siegrist, "Stress am arbeitsplatz [stress at the workplace]," in Gesundheitspsychologie [Health Psychology], R. Schwarzer, Ed. Göttingen: Hogrefe, 2006, pp. 303–318.

[5] R. S. Lazarus, Emotion and Adaptation. Oxford University Press, 1991.

[6] B. McEwen and J. Morrison, "The brain on stress: Vulnerability and plasticity of the prefrontal cortex over the life course," Neuron, vol. 79, no. 1, 2013, pp. 16 – 29.

[7] D. N. Khansari, A. J. Murgo, and R. E. Faith, "Effects of stress on the immune system," Immunology Today, vol. 11, 1990, pp. 170 – 175.

[8] T. Van Gog, F. Paas, and J. J. G. Van Merriënboer, "Uncovering expertise-related differences in troubleshooting performance: combining eye movement and concurrent verbal protocol data," Applied Cognitive Psychology, vol. 19, no. 2, 2005, pp. 205–221.

[9] S. Crevits, I.and Debernard and P. Denecker, "Model building for air-traffic controllers' workload regulation," European Journal of Operational Research, vol. 136, 2002, pp. 324–332.

[10] H.-C. She and Y.-Z. Chen, "The impact of multimedia effect on science learning: Evidence from eye movements," Computers & Education, vol. 53, no. 4, 2009, pp. 1297–1307.

[11] R. H. Mogford, "Mental models and situation awareness in air traffic control," The International Journal of Aviation Psychology, vol. 7, no. 4, 1997, pp. 331–341.

[12] S. T. Shorrock and A. Isaac, "Mental imagery in air traffic control." International Journal of Aviation Psychology, vol. 20, no. 4, 2010, pp. 309 – 324.

[13] B. Kirwan, L. Donohoe, T. Atkinson, H. MacKendrick, T. Lamoureux, and A. Phillips, "Getting the picture-investigating the mental picture of the air traffic controller," in Contemporary ergonomics 1998, M. A. Hanson, Ed. TAYLOR AND FRANCIS, 1998, pp. 404–408.

[14] M. Tavanti and M. Cooper, "Looking for the 3d picture: The spatio-temporal realm of student controllers," in Human Centered Design, ser. Lecture Notes in Computer Science, M. Kurosu, Ed. Springer Berlin Heidelberg, 2009, vol. 5619, pp. 1070–1079.

[15] E. Brink, "The relationship between occupational stress, emotional intelligence and coping strategies in air traffic controllers," Master's thesis, University of Stellenbosch, 2009.

[16] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," Journal of Comparative Neurology and Psychology, vol. 18, no. 5, 1908, pp. 459–482.

[17] L. Pfeiffer, N. H. Müller, and P. Rosenthal, "A survey of visual and interactive methods for air traffic control data," in Information Visualisation (IV), 2015 19th International Conference on, 2015, pp. 574–577.

 ISBN: 978-1-61208-447-3

Preliminary Study on Bit-String Modelling of Opinion Formation in Complex Networks

Yi Yu, Gaoxi Xiao
School of Electrical and Electronic Engineering,
Nanyang Technological University,
Singapore
e-mail: yyu6@e.ntu.edu.sg; egxxiao@ntu.edu.sg

***Abstract*—Opinion formation has been the topic of increased research interest recently, and various models have been proposed. These models, however, have their limitations, including (i) it is generally assumed that adjacent nodes holding similar opinions will further reduce the difference between them, while adjacent nodes holding significantly different opinions would either do nothing, or cut the link in between them; (ii) opinion mutation, described as "opinion change not due to neighborhood influences" in real life, is typically random. While such models are simple but still help reveal useful insights, they lack the capability of describing many complex behaviors which we may easily observe in real life. In this paper, we propose a new bit-string modeling approach. Preliminary study on the new model demonstrates its great potential in revealing complex behaviors of social opinion evolution and formation.**

Keywords—complex network; opinion formation; bit-string modeling; opinion mutation.

I. INTRODUCTION

Opinion propagation, evolution and formation play a critical role in shaping our society and influences almost every aspect of our life, from as "small" as interpersonal relationship [1] to as big as elections [2][3], etc. There are several works on the propagation of different opinions in social networks [4]-[6] and the impacts of opinion propagation on social structures [7][8], etc. Another important topic is how people's opinions are influenced by each other in their social interactions and how such opinion changes help shape the opinion groups. This is known as *opinion formation* problem.

Extensive studies have been conducted on opinion formation in social population and a few different models have been proposed [9]-[22]. The simplest one among them is probably the voter model [10]-[12]. It assumes that there are only two opinions in the population, representing positive and negative attitudes towards a certain incident, respectively. In every time step, a randomly selected node (or an individual in the network; hereafter "individual" and "node" shall be used interchangeably) may adopt the opinion of its randomly selected neighbor. The voter model has been extended to the case with multiple different opinions [13][14]. Other works typically quantify the opinion as continuous variable [15]-[20]. Two most well-studied models include bounded confidence model [15]-[17] and the Deffuant model [18]-[22]. Both models assume that a node's opinion can be influenced by those neighbors who hold similar, or at least not-so-different, opinions, termed as *similar opinion neighbors* (SONs) hereafter. The only difference is that while Deffuant assumes that a node's opinion may be affected by a randomly selected SON, the bounded confidence model assumes that all the SONs have combined influences on the node. In both models, there is *consensus making*, while the node's opinion and its randomly selected SON (or all SONs) come closer to each other. Note that, in both models two opinions are regarded as similar opinions if the difference between them is smaller than a given *tolerance* value d . For the Defffuant model, existing results show that the network would enter into a final state where several opinion groups are formed and coexist. The number of groups has a linear relationship with $1/d$.

Noise was first introduced in Deffuant model in [20]-[22] to simulate the change of views for any reasons other than a SON's influence. In these studies, it was assumed that all opinions have an equal chance to change to any other opinion (In the rest of this paper, we term such change as *mutation.*). The results showed that the final-state opinion distribution shall resemble a well-defined bell curve [20] and the initial conditions have hardly any effects on the final state, with the only exception of some very special cases (e.g., the initial opinion is of a single value in the whole system) [21].

The imitations of adopting such a simple mutation model in the Deffuant model were revealed in [23]. It was shown that when different opinions have different chances of having mutations, the system dynamics may become rather complex. In fact, for different distributions of the "mutation probability" within the range of opinion, different final steady state may be achieved. In that study, however, it was still assumed that once a mutation happens, the "target" of the mutation is randomly distributed; in other words, the opinion may change to any other opinion with an equal chance.

We may argue that opinion mutation in real life may not have a randomly distributed target in most cases. Everyone is "defined" and "bounded" by his/her current and/or historical states to a certain extent. Some mutations may be relatively easier to happen than the others. In other words, for each opinion the mutation target may also have a non-uniform distribution; and more importantly, different opinions may

 ISBN: 978-1-61208-447-3

have different non-uniform target opinion distributions. In other words, the distribution of mutation target may rely on its current (or even historical) state. A new modeling approach capable of revealing such kind of state dependent mutation is in demand. The random target opinion distribution commonly adopted in current literature shall be viewed as a special case of the requested new modeling approach, where the distribution of the mutation target is independent of a node's current or historical state.

With the understanding of the limitation that the conventional opinion mutation models may have, it would be interesting to also have a look at the conventional consensus making models from this new angle as well. We may realize that the conventional consensus making model is state dependent: whether two neighbors could make consensus depends on the opinions they are holding. While such is applausible, the way that similar opinions come closer to each other may be more complex than what this model can describe. For example, people making consensus may stick to some of their differences, if such differences matter to them: close friends may tend to agree on almost everything, except for one or two "small but important" issues. What may be even more important is that, people with significantly different ideas may have very different chances of cutting the link in between them, depending on what that or those significant differences are.

To make an effort towards tackling the shortcomings of the conventional models as discussed above, in this paper, inspired by the genetic mutation in nature [24], we propose a new bit-string modeling approach. Specifically, we use a string of binary numbers to represent an opinion or a set of opinions. By doing so, we may (i) reflect the importance/relevance of different opinions or different part of an opinion where a higher bit represents a more important/relevant opinion among a set of opinions held by the individual, or a more important part of an opinion held by the individual; and (ii) conveniently reflect the different mutation target distributions of different opinions or different part of an opinion, e.g., by assigning different bits with different probabilities of mutation. It would not be difficult to take one step further by assigning "0" and "1" at different bit positions with different probabilities of mutation, reflecting the case where the probabilities of opinion change in two opposite directions are not symmetric. Our preliminary studies show that such an approach may have great potentials to reveal the complex dynamics of opinion formation in social networks which cannot be conveniently revealed by any of the existing models to the best of our knowledge.

The rest of this paper is organized as follows. Section 2 briefly describes the Deffuant model and then introduces the bit-string opinion model. As a case study, Section 3 discusses on a simple case where the mutation probabilities from 0 to 1 and from 1 to 0 are different on each bit position. We will see that the simple case nevertheless leads to some interesting and complex behaviors. Section 4 concludes the paper.

II. MODEL DESCRIPTION

A. Review of Deffuant Model with Mutation

Deffuant model assumes that opinions are continuously distributed within the interval [0, 1]. At each time step t, a node A is randomly selected together with its random neighbor B. Denote their opinions as $o(t, A)$ and $o(t, B)$, respectively. If the difference between these two opinions is less than a given tolerance d, they make consensus according to the following rules:

$$\begin{cases} o(t+1, A) = o(t, A) - \mu[o(t, A) - o(t, B)]; \\ o(t+1, B) = o(t, B) + \mu[o(t, A) - o(t, B)]. \end{cases} \quad (1)$$

A smaller value of μ may slow down the evolution process while different values of μ, as long as it is within the range of (0, 1/2], is believed to lead to the same final steady state [18]. Hereafter, we use $\mu = 1/2$ as that in most of the existing works.

Noise/mutation was firstly introduced into Deffuant model in [20]. Specifically, in each time step t, a randomly selected node has a probability p to mutate and adopt another randomly chosen opinion.

B. Bit-string modelling approach

The bit-string model is based on a simple idea of describing an opinion or a set of opinions into a string of binary number. For example, an opinion, or a set of opinions, adopted by an individual in a certain circumstance may be written as 01101001. Higher bits may denote something that is more "fundamental" and important to an individual, e.g., whether s/he has any religion belief in a study on "opinion formation of people's interpretation of eternity in a social community", while a lower bit may be generally speaking less significant, e.g., the individual's preference of sport activities in the above study. Certainly a string can also be used to represent a single idea (e.g., the religion belief in the above example), while different bits are of different importance in defining the idea: 01101001 may be regarded as a similar idea to 01101010, but significantly different from 11101001. In the above example, the former case means that two individuals have nearly the same religion belief in almost every detail; while in the latter one, the two individuals are very different in their religion beliefs.

At the first sight, adopting a bit-string model may be of limited benefits: it would be the same thing to write 01101001 as 105 in decimal number, or 105/255 as a real number within the range of [0,1]. The benefits, however, lie in the convenience of defining different "behaviors" on different bits. For example, by defining different bits with different mutation probabilities, we may resemble the fact that changing an individual's religion belief may be easier or more difficult than changing his/her favorite sports activities, both of which may affect, in rather different ways, how likely or unlikely his/her social connections can change his/her interpretation of eternity. Further, for the bit corresponding to religion belief of the individual, assigning

 ISBN: 978-1-61208-447-3

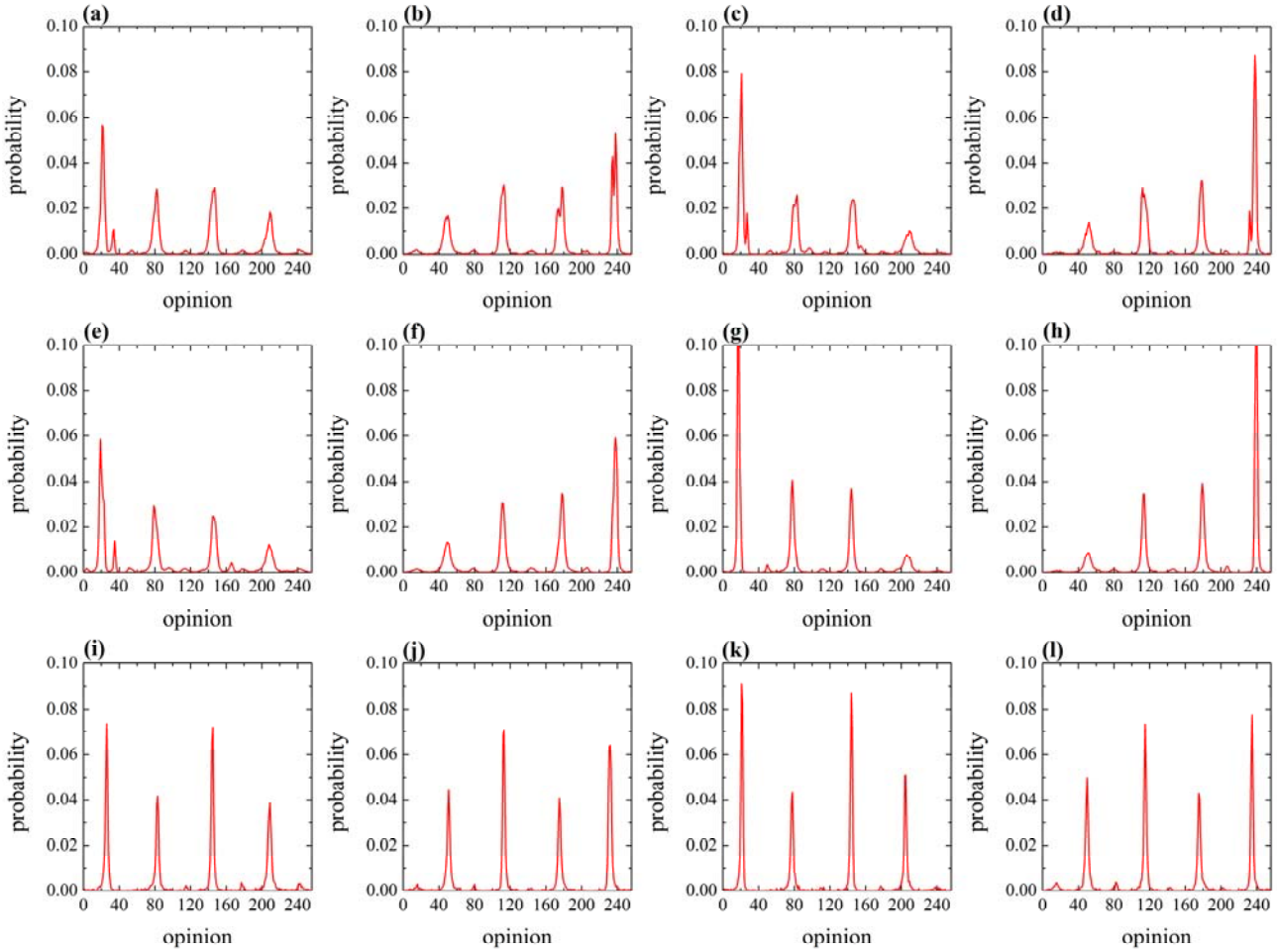


Figure 1. Final opinion distribution at $d = 25$ for different (p_{10}, p_{01}) *i)* when each bit of the string has the same probability of mutation: (a) (0.01, 0.005), (b) (0.005, 0.01), (c) (0.01, 0.003), (d) (0.003, 0.01); *ii)* when $\alpha = 1/28$ hence higher bits have higher probabilities of mutation.: (e) (0.01, 0.005), (f) (0.005, 0.01), (g) (0.01, 0.003), (h) (0.003, 0.01); and *iii)* when $\alpha = -1/28$ hence lower bits have higher probabilities of mutation: (i) (0.01, 0.005), (j) (0.005, 0.01), (k) (0.01, 0.003), (l) (0.003, 0.01).

different probabilities for it to change from 0 to 1 and change from 1 to 0 respectively would resemble the real-life case that it is easier or more difficult to make free thinker become a religion believer, or go through the opposite direction. The potentials of such a new modeling approach are attractive.

There are many different ways to define how different opinions may interact with each other and mutate themselves by using this bit-string model. For example, it would not be difficult to imagine crossover between two bit strings, like that in the genetic algorithm [25]. In this preliminary study, we consider the simple case which essentially is still the well-known Deffuant model with mutation, with the only difference the *i*-th bit has a mutation probability $p(i)$ which may be different for different bit positions (i.e., different values of i). While $p(i)$ may be affected by various combinations of many different current/historical factors as we discussed earlier, we consider the simple case where $p(i)$ is only affected by the current state of the *i*-th bit. Specifically, we consider the case where $p(i)$ is composed of two conditional probabilities: the probability for the *i*-th bit to mutate from 1 to 0 given that its current state is 1, and the probability of mutating from 0 to 1 given that its current state is 0, denoted as $p_{10}(i)$ and $p_{01}(i)$ respectively. Apparently, we have

$$p(i) = p_{01}(i)q_0(i) + p_{10}(i)q_1(i), \quad (2)$$

where $q_0(i)$ and $q_1(i)$ denote the probabilities that the current state of the *i*-th bit is 0 and 1, respectively. Note that, in the above model, since $q_0(i)$ and $q_1(i)$ evolve with the

network system, $p(i)$ is time varying until the system reaches steady state. This is very different from that in the existing studies where the mutation rate is typically a constant throughout the evolution process. We argue, however, that in the real life, mutation rate may be indeed time varying in most cases: a system in transition is expected to witness a relatively higher mutation rate, which may become lower when the system enters into a relatively more stable state.

Also note that (1) does not necessarily lead to an integer value that can be written into a binary bit string, in which case we assign the closest integer opinion to the node, and a tie is broken arbitrarily.

III. SIMULATION RESULTS AND DISCUSSIONS

We simulate the simple case where each opinion is represented by an 8-bit string (or equivalent 0-255 in decimal number). In each time step, in addition to the standard consensus making operation as that in the conventional Deffuant model, a node will be randomly selected as the candidate of opinion mutation. For the selected node, a single bit will be selected as the bit with a non-zero probability of having a mutation, where the *i*-th bit of the opinion is selected at a probability $\rho(i)$, $\sum_{i=1}^{8}\rho(i)=1$. We consider the case where the same set of values of $p_{10}(i)$ and $p_{01}(i)$ apply to all the network nodes and all the 8 bits. Specifically, we examine 4 pairs of different $p_{10}(i)$ and $p_{01}(i)$: (0.01, 0.005), (0.005, 0.01), (0.01, 0.003) and (0.003, 0.01), respectively. Note that by adopting such small values, a bit mutation does not happen more frequently than an average of 1 in every 100 time steps. We present the results in the ER random network [26] with a size of $N=10^4$ and an average nodal degree of $z=10$.

We start by considering the case where $\rho(i)=1/8,\ i=1,2,...8$. Setting the tolerance $d=25$, we perform the simulation for $t=5\times10^7$ time steps for each case and average the opinion distribution of the last 1000 steps as the final-state opinion distribution. Figures 1(a) to 1(d) illustrate the final state for the four different cases respectively. From Figures 1(a) and 1(c), the observation is that when $p_{10}>p_{01}$, the peaks positioned at smaller values would be higher; meanwhile the positions of the four peaks also slightly shift to the left-hand side. When $p_{10}<p_{01}$, the observations we can make from Figures 1(b) and 1(d) go to the opposite: the peaks positioned at larger values are higher and the peak positions shift to the right. The differences between the heights of different peaks become larger when the ratio between p_{10} and p_{01} is larger in the former case (comparing Figures 1(a) and 1(c)) and smaller in the latter case (comparing Figures 1(b) and 1(d)). Such observations match our daily experiences. For example, when the whole society tends to be optimistic (pessimistic), though people may still hold different ideas, different ideas may all tend to be shifted towards the optimistic (pessimistic) side. The more optimistic (pessimistic) the society is, the more people would be found at the optimistic (pessimistic) end, and the peaks of opinions typically also shift to that end. Though such observations are well known in real life, to the best of our knowledge, it is the first time that it is observed in numerical simulation based on a simple mathematical model.

We then consider the slightly different case when

$$\rho(i)=\alpha(4.5-i)+0.125,\quad i=1,2,\ldots8 \qquad (3)$$

where $\alpha\in[-1/28,1/28]$. For this function, a positive α means that higher order digits have higher probabilities to be selected for mutation while a negative α indicts the opposite. We still set $d=25$.

Figures 1(e) to (h) and Figures 1(i) to (l) present the results when $\alpha=1/28$ and $-1/28$, respectively. Note that when $\alpha=1/28$, higher order bits have higher probabilities to be selected for mutation, at a ratio of $\rho(i):\rho(j)=(j-1):(i-1),\ i,j=1,2,\ldots,8;$ while for $\alpha=-1/28$, lower order bits have higher probabilities to be selected, and the ratio becomes $\rho(i):\rho(j)=(i-1):(j-1),\ i,j=1,2,\ldots,8$.

For $\alpha=1/28$, Figures 1(e) to 1(h) respectively present the final opinion distributions corresponding to 4 pairs of different $p_{10}(i)$ and $p_{01}(i)$: (0.01, 0.005), (0.005, 0.01), (0.01, 0.003) and (0.003, 0.01). The observations are almost the same as those in Figures 1(a) to 1(d). The only nontrivial difference is that in Fig. 1(g) (1(h)), the peak at the leftmost (rightmost) side is much higher than the corresponding peak in Fig. 1(c) (1(d)). A rough understanding of the reasons behind is not so difficult to achieve: when $p_{10}(i)$ is much higher $p_{01}(i)$ and higher bits have higher chances of mutation, the chance of having "0" on higher bits becomes higher, making the peaks closer to the left side end higher. This explains the observation in Figures 1(g). Similar reasoning can be adopted to explain the difference between Figures 1(d) and 1(h). Considering that Figures 1(a) and 1(e) however appear to be nearly the same, it remains as a challenge to figure out how big a difference between $p_{10}(i)$ and $p_{01}(i)$ is big enough to lead to nontrivial differences in the final state.

Figures 1(i) to 1(l), however, present very different observations when lower bits have higher probabilities to mutate: while peaks still shift to the left when $p_{10}>p_{01}$ (Figures 1(i) and (k)) and to the right when $p_{10}<p_{01}$ (Figures 1(j) and (l)), the heights of the four peaks do not increase or decrease monotonically from left to right. Rough understanding may still be easily achieved: when higher bits have lower opportunities of mutation and the highest bit has a zero mutation probability (and therefore does not mutate at all), at steady state we shall expect to find half of the nodes holding opinions starting with a bit "0" and the other half a bit "1". Opinion distribution is thus roughly 50-50 on the left and right half of the opinion axis. Mutation of the other bits

 ISBN: 978-1-61208-447-3

(2nd to the 8th bits) can still generate "uneven" distribution in each half of the opinion axis, depending on whether $p_{10} > p_{01}$ or $p_{10} < p_{01}$.

While rough understandings as discussed above are not difficult to achieve, obviously extensive further studies are needed to fully understand the system dynamics.

IV. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new bit-string modeling approach for more efficiently describing the complex dynamics of opinion formation in complex networks. The new approach allows convenient modeling of various non-uniform, state-dependent behaviors of different opinions or different parts of an opinion. Preliminary study on a very simple case reveals the great potentials the new approach may have.

A lot of other interesting observations have been made in our preliminary studies, which have been largely omitted in this paper. These observations shall be carefully sorted into some systematic descriptions and discussions in our future studies. A theoretical framework for analyzing the evolution of the system adopting the new modeling approach will also be developed.

ACKNOWLEDGMENT

This work is partially supported by Ministry of Education (MOE), Singapore, under research grant RG 28/14 and MOE2013-T2-2-006.

REFERENCES

[1] S. L. Parker, G. R. Parker, and J. A. McCann, "Opinion taking within friendship networks," Am. J. Polit. Sci., vol. 52, pp. 412-420, 2008.

[2] B. Norrander, "Measuring state public opinion with the senate national election study," State Polit. Policy Q., vol. 1, pp.111-125, 2001.

[3] B. R. Berelson, P. F. Lazarsfeld, and W. N. McPhee, "Voting: a study of opinion formation in a presidential election," Chicago: University of Chicago Press, 1954.

[4] A. M. Timpanaro and C. P. C. Prado, "Generalized Sznajd model for opinion propagation," Phys. Rev. E, vol. 80, p. 021119, 2009.

[5] F. Amblard and G. Deffuant, "The role of network topology on extremism propagation with the relative agreement opinion dynamics," Physica A, vol. 343, pp. 725-738, 2004.

[6] Y. Wang, G. Xiao, and J. Liu, "Dynamics of competing ideas in complex social networks," New J. Phys., vol. 14, p. 013015, 2012.

[7] F. Wu and B. A. Huberman, "Social structure and opinion formation, arXiv:cond-mat/0407252v3, 2004.

[8] R. S. Burt, "The Social Capital of Opinion Leaders," The Ann. Am. Acad. Polit. Soc. Sci., vol. 566, pp. 37-54, 1999.

[9] M. H. DeGroot, "Reaching a Consensus," J. Am. Statist. Assoc., vol. 69, pp. 118-121, 1974.

[10] R. A. Holley and T. M. Liggett, "Ergodic theorems for weakly interacting infinite systems and the voter mode", Ann. Prob., vol. 3, pp. 643-663, 1975.

[11] V. Sood and S. Redner, "Voter model on heterogeneous graphs," Phys. Rev. Lett., vol. 94, p. 178701, 2005.

[12] C. Castellano, D. Vilone, and A. Vespignani, "Incomplete ordering of the voter model on small-world networks," EPL-Europhys. Lett., vol. 63, p. 153, 2003.

[13] P. Holme and M. E. J. Newman, "Nonequilibrium phase transition in the coevolution of networks and opinions," Phys. Rev. E vol. 74, p. 056108, 2006.

[14] G. A. Böhme and T. Gross, "Fragmentation transitions in multistate voter models," Phys. Rev. E vol. 85, p. 066117, 2012.

[15] R. Hegselmann and U. Krause, "Opinion dynamics and bounded confidence:models, analysis and simulation," JASSS-J. Artif. Soc. S., vol. 5, pp. 1-10, 2002.

[16] J. A. N. Lorenz, "Continuous opinion dynamics under bounded confidence: a survey," Int. J. Mod. Phys. C, vol. 18, pp. 1819-1838, 2007.

[17] M. Pineda, R. Toral, and E. Hernández-García, "Diffusing opinions in bounded confidence processes," Eur. Phys. J. D, vol. 62, pp. 109-117, 2011.

[18] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing beliefs among interacting agents," Adv. Complex Syst., vol. 03, pp. 87-98, 2000.

[19] E. Ben-Naim, P. L. Krapivsky, and S. Redner, "Bifurcations and patterns in compromise processes," Physica D, vol. 183, pp. 190-204, 2003.

[20] M. Pineda, R. Toral and E. Hernández-García, "Noisy continuous-opinion dynamics," J. Stat. Mech. Theory E., vol. 2009, p. 08001, 2009.

[21] A. Carro, R. Toral and M. S. Miguel, "The role of noise and initial conditions in the asymptotic solution of a bounded confidence, continuous-opinion model," J. Stat. Phys., vol. 151, pp. 131-149, 2013.

[22] M. Pineda, R. Toral and E. Hernández-García, "The noisy Hegselmann-Krause model for opinion dynamics," Euro. Phys. J. B, vol. 86, pp. 1-10, 2013.

[23] Y. Yu and G. Xiao, "Influence of random opinion change in complex networks," Proc. IEEE DSP'2015, July 2015, pp. 1-5.

[24] D.N. Cooper, M. Krawczak, "Human gene mutation", Bios Scientific Pub. Ltd, 1993.

[25] J. H. Holland, Adaptation in Natural and Artificial Systems, The MIT Press, 1992 (reprint).

[26] P. Erdos and A. Renyi, "On random graph I," Publ. Math. Debrecen, vol. 6, p. 290, 1959.

 ISBN: 978-1-61208-447-3

Computationally Detecting and Quantifying the Degree of Bias in Sentence-Level Text of News Stories

C.J. Hutto Dennis Folds Scott Appling
Georgia Tech Research Institute (GTRI)
Georgia Institute of Technology
Atlanta, GA U.S.A.
email: {Clayton.Hutto, Dennis.Folds, Scott.Appling}@gtri.gatech.edu

***Abstract*—Fair and impartial reporting is a prerequisite for objective journalism; the public holds faith in the idea that the journalists we look to for insights about the world around us are presenting nothing more than neutral, unprejudiced facts. Most news organizations strictly separate news and editorial staffs. Bias is, unfortunately, ubiquitous nevertheless. It is therefore at once both intellectually fundamental and pragmatically valuable to understand the nature of bias. To this end, we constructed a computational model to detect bias when it is expressed in news reports and to quantify the magnitude of the biased expression. As part of a larger overall effort, we conducted a survey of 91 people to investigate factors that influence the perception of bias in fictitious news stories. During this process, subjects provided ground-truth gold standard ratings for the degree of perceived bias (slightly, moderately, or extremely biased) for every sentence across five separate news articles. In this work-in-progress, we analyze the efficacy of a combination of linguistic and structural information for not only detecting the presence of biased text, but also to construct a model capable of estimating its scale. We compare and contrast 26 common linguistic and structural cues of biased language, incorporating sentiment analysis, subjectivity analysis, modality (expressed certainty), the use of factive verbs, hedge phrases, and many other features. These insights allow us to develop a model with greater than 97% accuracy, and accounts for 85.9% of the variance in human judgements of perceived bias in news-like text. Using 10-fold cross-validation, we verified that the model is able to consistently predict the average bias (mean of 91 human participant judgements) with remarkably good fit.**

Keywords-bias detection; bias quantification; linguistic model; text processing.

I. DATASET OF BIASED AND UNBIASED TEXT

A. Perception of Bias in Unattributed News Stories

In [1], people rated Presidents Bush and Obama on 25 adjectives and were then randomly assigned to read five fictitious news stories about one of them. Three of the stories described positive outcomes, and two described negative outcomes. In every story, one sentence was randomly manipulated to attribute the outcome to either an *internal* trait of the president or to *external* factors in an effort to observe the effects of moderating and mediating aspects of the attribution bias, whereby individuals typically assign greater attribution to internal/personal factors for positive outcomes when the person is someone they like, and to external/situational factors if the outcome is negative.

As part of the initial study, ninety-one people were surveyed. Participant demographics were skewed somewhat toward male (about 60%) and young adults under age 40 (over 50%). The political attitudes of the participants were of primary interest to [1], though, in particular, attitudes toward Presidents George W. Bush and Barack Obama. About two thirds of the sample had positive opinions about Obama and negative opinions about Bush, and one third exhibiting the opposite pattern. Participants were randomly assigned to provide ratings of one president first (Bush or Obama), followed by ratings of the second. Their responses were then used in a stratified sampling strategy to assign participants to read the five fictional news stories using either the name of the president they viewed most positively or most negatively (and 4 individuals who were neutral to both men were randomly assigned). Across the five stories, the story "target" remained the same once the participants were assigned to read about either Bush or Obama. We balanced the presentation order for the five stories to mitigate potential ordering effects. An example news story is presented below:

> *According to Forrester Research, an estimated 200,000 American jobs are lost annually due to offshore outsourcing. While in the past it was predominantly blue-collar jobs and low-level white-collar jobs that were relocated, the data show even mid- to high-level white-collar jobs are now being outsourced. During* ***{Bush/Obama}****'s presidential campaign, he maintained outsourcing is a part of globalization, which will be good for the American people in the long run. High unemployment rates led to growing public condemnation of outsourcing and demand for new regulations to stop or limit outsourcing. In response, corporations increased lobbying efforts to defend their ability to outsource jobs overseas, which they argued is necessary in order to remain competitive with international firms. Ultimately, President* ***{Bush/Obama}*** *rejected the proposal to implement trade protection policies that would discourage outsourcing. The President dismissed the proposal mainly because of…"*
>
> *"…* ***his unwillingness to stand up to corporate special interests****."(internal attribution)*
>
> *OR*
>
> *"…* ***intense pressure from corporations****." (external)*

This first story was about a financial situation where the outcome was negative. The other four stories reported about:

1. The president's decision to eliminate a federal grant program for teachers who would no longer receive incentive grants to work in inner-city school districts due to budget concerns (a negative outcome).
2. The president's promise to seek funding to support better emergency planning efforts, particularly those aimed at assisting with disaster preparedness for individuals with disabilities (a positive outcome).
3. The president's pledge to improve healthcare services to veterans (a positive outcome).
4. A successfully foiled bioterrorism attempt to smuggle aerosolized Ebola virus aboard an airplane in New York City (also a positive outcome).

B. Degree of Bias in News Stories

The current work-in-progress is primarily concerned with automatically characterizing the *intensity* or *degree* of bias perceived to be present in these news stories. In [1], subjects first read an entire story in paragraph form, and then were presented each sentence one a time and asked to rate how biased they believed each statement to be. Response options consisted of a 7-point balanced rating scale, with an option for a neutral rating ([–3] *Extremely* biased AGAINST Bush/Obama, [–2] *Moderately* biased AGAINST Bush/Obama, [–1] *Slightly* biased AGAINST Bush/Obama, [0] Fair and Impartial, [+1] *Slightly* biased IN FAVOR of Bush/Obama, [+2] *Moderately* biased IN FAVOR of Bush/Obama, or [+3] *Extremely* biased IN FAVOR of Bush/Obama). As we are currently interested in quantifying the degree of bias (rather than the polarity), we simplify by using the absolute value of the numerically coded responses.

II. Related Work

There is a rich literature on stance recognition and argument subjectivity that focuses on identifying which side an article takes on a two-sided debate (c.f., [2]), casting the task as a two-way classification of the text as being either for/positive or against/negative (e.g., [3]–[5]) or as one of two opposing views (e.g., [6], [7]). In contrast, our work is primarily interested in estimating the *magnitude*, rather than direction or polarity, of the bias perceived to be present at the sentence level across all five news stories.

Additionally, previous datasets consisted of texts that typically take an overt stance (such as product reviews, debate transcripts, or editorial news); in contrast, we desire the capability to gauge bias even within the much more subtle domain of so-called "objective" news reports. Our work follows in the same vein as [8] who analyze biased language in reference articles using page edits tagged for violating Wikipedia's Neutral Point Of View (NPOV) policy. Again, whereas [8]'s focus is on identifying specific words or phrases that signal bias in reference articles, our work is distinct in that we are interested in characterizing the *degree* of such bias in the context of *news stories*, which – as with reference articles – similarly strive for impartiality.

III. Detecting and Computing Degree of Bias

Using the 7-point balanced rating scale described above (coded as ranging from [–3] to [+3]) and human judgements of perceived bias from 91 participants for each of the 41 sentences from 5 separate news stories, we calculate the mean and distributions of the ratings using the absolute value of the numerically coded responses. As we see from the example text in Table I, some sentences of the news story are clearly perceived by human judges as being somewhat biased (as [1] intended to subtly induce either internal or external attribution biases by manipulating the final two sentence options). Expanding on the insightful work of [8] with additional sentence-level features and a dataset of news stories rather than reference articles, we develop a computational model that reads in a given sentence of text and then extracts and computes the strength of 26 structural and linguistics features present in the text. We next describe these 26 features.

A. Structural Analysis at the Sentence Level

In our sentence level analysis of the text, we observe characteristics of the text statement as a whole, considering syntactical, grammatical, and structural properties captured using the following five features:

1. **Sentiment score**: we use the freely available Python package VADER [9] to compute both the direction and intensity of the sentiment of each sentence (values range continuously from –1.0 [Extremely Negative] to +1.0 [Extremely Positive]). VADER is a highly accurate and well-validated sentiment analysis processing engine that implements numerous empirically derived sentiment processing rules related to textual syntax, grammar, punctuation, capitalization, negation, and other word-order sensitive elements of text [10].
2. **Subjectivity score**: we use Pattern.en [11] to compute the subjectivity of the sentence (values between 0.0 and 1.0). Pattern is a web mining module for Python, and the Pattern.en module is a natural language processing (NLP) toolkit that leverages WordNet to score subjectivity according to the English adjectives used in the text [12].
3. **Modality (certainty) score**: we use Pattern.en to compute the modality, or certainty, of the sentence (values range between –1.0 and +1.0, where values greater than +0.5 represent facts).
4. **Mood**: we use Pattern.en to compute the mood of the sentence. The mood of the sentence can be INDICATIVE (used to express facts, beliefs, e.g., "*It's raining*".), IMPERATIVE (used for commands or warnings, e.g., "*Make it rain!*"), CONDITIONAL (used for conjectures, e.g., "*It might rain today*") or SUBJUNCTIVE (used to express wishes or opinions, e.g., "*I hope it rains today*.").
5. **Readability**: we implement the Flesch-Kincaid Grade Level (FKGL) formula [13] to compute the readability of the sentence and associate it with a typical requisite grade level of reading comprehension. The higher the grade level, the more difficult the text.

 ISBN: 978-1-61208-447-3

TABLE I: MEAN (STANDARD DEVIATION) FOR 91 RATINGS OF PERCEIVED BIAS [SCALE: 0=UNBIASED TO 3=EXTREMELY BIASED]

	Sentence Level Text (for sentences from the first news story)	Mean (SD)
1	According to Forrester Research, an estimated 200,000 American jobs are lost annually due to offshore outsourcing.	0.10 (0.42)
2	While in the past it was predominantly blue-collar jobs and low-level white-collar jobs that were relocated, the data show even mid- to high-level white-collar jobs are now being outsourced.	0.11 (0.46)
3	During Bush/Obama's presidential campaign, he maintained outsourcing is a part of globalization, which will be good for the American people in the long run.	0.71 (1.00)
4	High unemployment rates led to growing public condemnation of outsourcing and demand for new regulations to stop or limit outsourcing.	0.20 (0.64)
5	In response, corporations increased lobbying efforts to defend their ability to outsource jobs overseas, which they argued is necessary in order to remain competitive with international firms.	0.12 (0.51)
6	Ultimately, President Bush/Obama rejected the proposal to implement trade protection policies that would discourage outsourcing.	0.70 (1.04)
7e	The President dismissed the proposal mainly because of intense pressure from corporations.	1.35 (1.22)
7i	The President dismissed the proposal mainly because of his unwillingness to stand up to corporate special interests.	1.90 (1.21)

B. *Linguistic Analysis at the Sentence Level*

Motivated by [8], we implement several linguistic features aimed at detecting either *epistemological* bias (features 6-9) or *framing* bias (features 10-12). To these, we add several additional linguistic features that we hypothesize may effect human perceptions of bias in text (features 13-26). For all of our sentence level linguistic features, we normalized the count of observations of the feature in the sentence by the total number of words in the sentence, producing values between 0.0 and 1.0 for each.

6. **Factive verbs**: are verbs that presuppose the truth of their complement clause (c.f., [8] for use in detecting epistemological bias in reference articles).
7. **Implicative verbs**: implicative verbs imply the truth or untruth of their complement, depending on the polarity of the main predicate (c.f., [8]).
8. **Assertive verbs**: are verbs whose complement clauses assert a proposition. The truth of the proposition is not presupposed, but its level of certainty depends on the asserting verb (c.f., [8]).
9. **Hedges**: used to reduce one's commitment to the truth of a proposition, evading any bold predictions (c.f., [8]).
10. **Strong subjective intensifiers**: are adjectives or adverbs that add (subjective) force to the meaning of a phrase or proposition (c.f., [8] for detecting framing bias in text using [14]'s list of strong subjectives).
11. **Weak subjective intensifiers**: as in [8], we use [14]'s list of weak subjectives.
12. **Bias (one-sided) terms**: One-sided terms reflect only one of the sides of a contentious issue (e.g., *anti-abortion* versus *pro-life*). We use [8]'s lexicon.
13. **Opinion words**: signal the expression of positive or negative attitudes or opinions, which may be biased. We use [10]'s validated opinion lexicon.
14. **Degree Modifiers**: are contextual cues (often adverbs such as *extremely*, or *slightly*) that modify the intensity or degree of an action, an adjective or another adverb. We use [10]'s list of degree modifiers.
15. **Coherence Markers**: are words (*because*, *therefore*, *so*) or lexical phrases (*as a result*, *for that reason*) that may be used to bias a reader towards a particular conclusion. We use [15]'s list of coherence markers.

The Linguistic Inquiry and Word Count (LIWC) [16] is text analysis software designed for studying the various emotional, cognitive, structural, and process components present in text samples [17]. LIWC uses a proprietary dictionary of almost 4,500 words organized into one (or more) of 76 categories, of which we use several for our feature set:

16. **Causation words**: e.g., *create*, *founded*, *generate*
17. **Certainty words**: e.g., *absolutely*, *frankly*, *must*
18. **Tentative words**: e.g., *bets*, *dubious*, *hazy*, *guess*
19. **3rd Person Pronoun**: e.g., *he*, *him*, *she*, *hers*, *they*
20. **Achievement words**: e.g., *accomplished*, *master*, *prized*
21. **Work words**: e.g., *ambitious*, *resourceful*, *hard-work*
22. **Discrepancy words**: e.g., *inadequate*, *mistake*, *liability*
23. **Conjunctions**: e.g., *while*, *although*, *cuz*, *whereas*
24. **Prepositions**: e.g., *within*, *over*, *through*
25. **Adverbs**: e.g., *mostly*, *nearly*, *primarily*
26. **Auxiliary verbs**: e.g., *may*, *oughta*, *should*, *will*

IV. FINAL MODEL FEATURE SELECTION

We next processed the 26-item feature vectors for each sentence through an initial statistical linear regression model using both forward and backwards stepwise Akaike information criterion (AIC) to measure the relative quality of each feature for characterizing the degree of bias in text. Using step-AIC for feature selection in this way helped us restrict the feature space to the most useful and valuable features. For example, in the presence of [14]'s more detailed list of strong and weak subjective linguistic intensifiers, the sentence-level measure of *subjectivity* is less meaningful (we therefore removed it from the model). On the other hand, the sentence-level measure for *modality*

 ISBN: 978-1-61208-447-3

(certainty) is a stronger indicator of bias than the linguistic cues associated with LIWC *certainty words*, so we removed the *certainty words* feature from the model. Unfortunately, there was not enough variation in the sample data to determine whether differences in sentence structure with regards to *mood* affected perceived bias. As one might expect in "objective" news stories, nearly all sentences (85.4%) were computed to be INDICATIVE; so, we removed *mood* as a feature from the model. We found Flesch-Kincaid Grade Level (FKGL) scores for sentence-level *readability* were unrelated to the degree of perceived bias. This might be due to grade-level reading scores being generally high across the sample. The majority of sentences in the news stories ranged from about an 11th grade reading level (high school junior) to an 18th grade reading level (graduate school) (Mean=14.57, Standard Deviation=3.22). We therefore removed *readability* as feature from the model. Finally, we found that measures for *implicative verbs*, *degree modifiers*, *coherence markers*, *causation words*, *conjunctions*, *prepositions*, *adverbs*, and *auxiliary verbs* were all relatively poor indicators of sentence level bias; we therefore removed those features from the final model.

V. PRELIMINARY RESULTS

Table 2 depicts preliminary results of the linear regression analysis for the improved 14-feaure model $F(14,26) = 11.3$, $p = 1.04e\text{-}07$, which accounts for over 85% of the variance in human judgements of bias ($R^2 = 0.859$). Figure 1 depicts the proportion of overall R^2 that each feature accounts for, using the mean of three regression techniques (feature added to model first, feature added to model last, and feature beta squared). We find that a linguistic model motivated by [8]'s list of features for detecting biased language in reference articles is a useful start for determining the intensity (degree) of bias in news stories.

TABLE II: COEFFICIENTS, ERROR, T-VALUES, AND *P*-VALUES FOR THE IMPROVED MODEL. $F(14,26) = 11.3$, $P = 1.04E\text{-}07$.

	b	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.56	0.19	-3.02	0.006
Strong subjective	5.10	1.07	4.74	0.000***
3rd Person Pronoun	8.36	1.95	4.30	0.000***
Weak subjective	4.87	1.19	4.08	0.000***
Modality (certainty)	0.52	0.15	3.42	0.002**
VADER Sentiment	0.35	0.11	3.13	0.004**
Tentative words	4.60	1.65	2.79	0.010**
Opinion words	-2.05	0.95	-2.16	0.040*
Achievement words	5.74	2.66	2.16	0.040*
Factive verbs	-16.64	8.39	-1.98	0.058`
Work words	9.81	5.20	1.89	0.070`
Hedges	3.06	1.75	1.75	0.092`
Assertive verbs	-3.58	2.16	-1.66	0.110
Discrepancy words	5.66	3.62	1.56	0.130
Bias (one-sided) terms	-0.95	0.74	-1.30	0.206

Signif. level codes: $p < 0.001$*** $p < 0.01$** $p < 0.05$* $p < 0.1$`

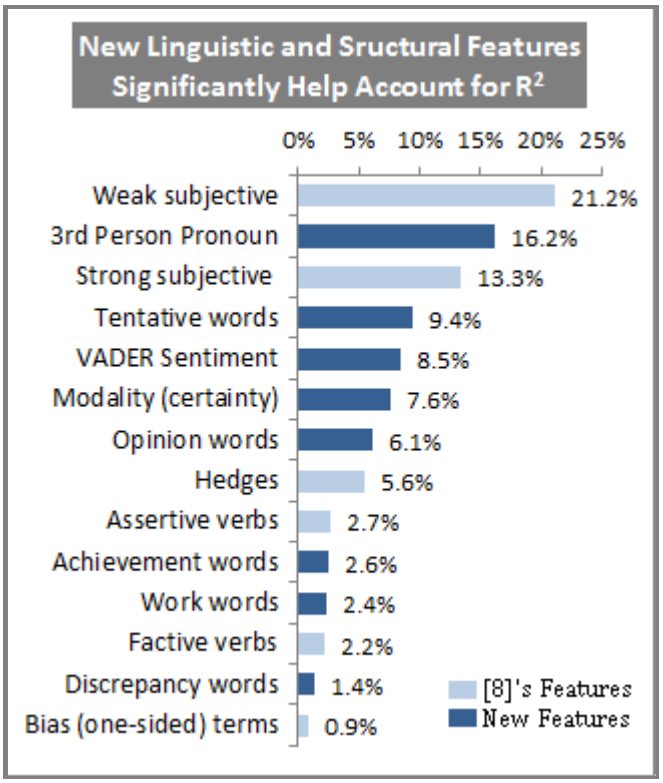


Figure 1. Proportion of variance accounted for by each feature in the improved model using the mean R^2 of three regression techniques (feature added to model first, feature added to model last, and feature beta squared).

Figure 2 shows the match between observed (measured) bias and the degree of bias predicted by the model; the fit is remarkably good. Many of our additional linguistic and structural features help to improve its predictive power:

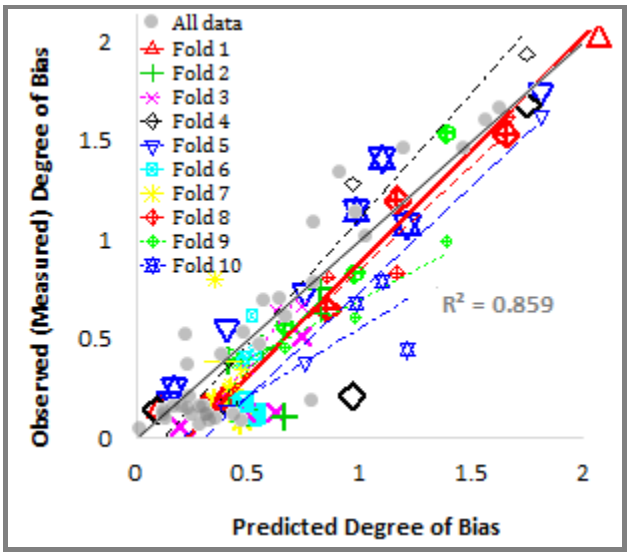


Figure 2. Results of 10-fold cross-validation analysis for fit between observed and predicted values of degree of bias in text.

REFERENCES

[1] D. J. Folds, "Perception of bias in unattributed news stories," in *Procedures of the Annual Meeting of the Association for Psychological Science*, New York, NY, 2015.

[2] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann, "Which side are you on?: identifying perspectives at the document and sentence levels," in *Proceedings of the Tenth Conference on Computational Natural Language Learning*, New York City, New York, 2006, pp. 109–116.

[3] P. Anand, M. Walker, R. Abbott, J. E. F. Tree, R. Bowmani, and M. Minor, "Cats rule and dogs drool!: classifying stance in online debate," in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Portland, Oregon, 2011, pp. 1–9.

[4] A. Conrad, J. Wiebe, and and R. Hwa, "Recognizing arguing subjectivity and argument tags," in *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, Jeju, Republic of Korea, 2012, pp. 80–88.

[5] S. Somasundaran and J. Wiebe, "Recognizing stances in ideological on-line debates," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Los Angeles, California, 2010, pp. 116–124.

[6] T. Yano, P. Resnik, and N. A. Smith, "Shedding (a thousand points of) light on biased language," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Los Angeles, California, 2010, pp. 152–158.

[7] S. Park, K. Lee, and J. Song, "Contrasting opposing views of news articles on contentious issues," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Portland, Oregon, 2011, pp. 340–349.

[8] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "Linguistic Models for Analyzing and Detecting Biased Language," in *Proceedings of the 51st Meeting of the Association for Computational Linguistics*, 2013, pp. 1650–1659.

[9] C. J. Hutto, "VADER Sentiment Analysis Software." [Online]. Available: https://github.com/cjhutto/vaderSentiment. [Accessed: 28-Jul-2015].

[10] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014, pp. 216–255.

[11] CLiPS Research Center, "Pattern.en Software." [Online]. Available: http://www.clips.ua.ac.be/pages/pattern-en. [Accessed: 28-Jul-2015].

[12] T. De Smedt and W. Daelemans, "Pattern for Python," *J. Mach. Learn. Res.*, vol. 13, pp. 2063–2067, 2012.

[13] P. Kincaid, R. Fishburne, R. Rogers, and B. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.," National Technical Information Service, Springfield, Virginia 22151 (AD-A006 655/5GA, MF $2.25, PC $3.75), Feb. 1975.

[14] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 105–112.

[15] A. Knott, "A Data-Driven Methodology for Motivating a Set of Coherence Relations," PhD Thesis, Department of Artificial Intelligence, University of Edinburgh, 1996.

[16] Pennebaker Conglomerates, Inc., "LIWC Text Analysis Software." [Online]. Available: http://www.liwc.net/. [Accessed: 28-Jul-2015].

[17] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, *The development and psychometric properties of LIWC2007*. Austin, TX: LIWC.net, 2007.

 ISBN: 978-1-61208-447-3

Reliability of Physiological Signals induced by Sadness and Disgust

Eun-Hye Jang, Hyo-Young Cho, Sang-Hyeob Kim
Bio-Medical IT Convergence Research Department
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
e-mail: {cleta4u, deardol, shk1028}@etri.re.kr

Youngji Eum, Jin-Hun Sohn
Department of Psychology & Brain Research Institute
Chungnam National University
Daejeon, Republic of Korea
e-mail: petitaudrey@hanmail.net, jhsohn@cnu.ac.kr

***Abstract*—In Human Computer Interaction (HCI), acquisition of physiological signals for emotion recognition is done by emotion researches. Prior to this, one needs to identify the reliability of physiological responses. The aim of this study was to investigate the reliability of physiological responses induced by sadness and disgust using an experiment that was repeated 10 times. Twenty subjects participated in this experiment. For emotion induction, twenty different emotional stimuli were selected in a pilot experiment. Skin Conductance Level (SCL), Skin Conductance Response (SCR), and Heart Rate (HR) were measured before the presentation of stimuli as a baseline and during the presentation of the stimuli as emotional state. The results showed that physiological signals during emotional states for the 10 times the experiment was repeated were stable and reliable compared to the baseline. Our results suggest that physiological signals of sadness and disgust are reliable over time. This means that physiological signals are reliable and useful tools for emotion recognition. These results can be useful in developing an emotion theory, or profiling emotion-specific physiological responses, as well as establishing the basis for an emotion recognition system in HCI.**

Keywords-physiological signals; reliability; sadness; disgust.

I. INTRODUCTION

To recognize human's emotions and feelings, various physiological signals have been widely used in human computer interaction (HCI) [1]. Recently, physiological signals have been applied to continuous ambulatory monitoring of the affective state of individuals. For this, one needs to identify the pattern of physiological responses under specific emotional situations. This is important for basic and applied physiological research [2][3]. Although previous results have shown the temporal stability of physiological response patterns [4]-[11], they are not sufficient to verify whether or not complex patterns are stable [2]. Some results focused on the stability of the physiological responses by introducing different time intervals (e.g., 2 weeks or 4 weeks [9][10]) or using different kinds of biomarkers (e.g., blinking responses [9][11], Respiratory Sinus Arrhythmia (RSA), heart rate, salivary cortisol [12], and startle response [10]). Other studies failed to show consistent findings despite repetitive experiments [13][14]. They were limited to measuring the physiological responses only twice within a relatively short time interval with the same stimuli to examine whether or not the given conditions evoked stable emotions. The results may have possibly been subject to an adaptation effect to stimuli. Also, the methodological issue with these studies is that the stability was determined by physiological measures induced by valence (pleasant and unpleasant) not by a specific emotion. To examine the reliability of physiological responses over a relatively long period of time, we attempted to identify the stability of physiological responses induced by specific emotions (sadness and disgust) using audio-visual film clips in an experiment repeated 10 times. What differs from previous studies is the elimination of possible adaptation and learning effect (e.g., habituation) to the same stimuli by using differential emotional stimuli to effectively provoke each emotion, i.e., sadness and disgust. The rest of the paper is structured as follows. In Section 2, we provide the experimental methods used including description of the subjects, material and procedure. In Section 3, we present the experimental results and we conclude in Section 4.

II. METHODS

The emotional stimuli were 2~4 minutes long film clips, captured originally from a variety of movies and TV shows. The contents of the sadness film clips included scenes to address themes of death of parents or lover, separation, longing for mother, etc., and the contents of disgusting film clips included themes such as mutilation, butchery, and blooding. A total of 20 emotional stimuli were selected (2 emotions repeated 10 times each) by the pilot study. To examine whether the stimuli were suitable for provoking emotion, an appropriateness (the label of the experienced emotion i.e., sadness, disgust, and others) and an effectiveness (the intensity of the emotion in response to the emotional stimulus) of each stimulus were tested by the participants' ratings. The experimental procedures follow. Twenty healthy persons (10 males and 10 females) aged 21.0 (SD 1.9) years old college students participated in this experiment. They filled out a written consent before the beginning of the study and were introduced to the experiment protocols. Then, they were attached electrodes on their wrist, finger, and ankle for measurement of physiological signals, i.e., electrocardiogram (ECG) and electrodermal activity (EDA). Physiological signals were measured for 1 minute during baseline (before presentation of the stimuli) and for 2~ 4 minutes during emotional state (during presentation of stimuli) by using the MP100 (Biopac, USA). Finally, their psychological assessment was rated based on the experienced emotions. The procedures were conducted on each of the two emotions for 10 weeks on a weekly basis. To analyze physiological data, tonic level (skin conductance level, SCL, in uS) and phasic level (skin conductance response, SCR, in uS) were extracted from the

 ISBN: 978-1-61208-447-3

EDA channel and heart rate (HR, in beats per minute) in the ECG. The physiological data were analyzed for each 30 seconds from the baseline and emotional states. Also, Cronbach's alpha [15], which is a measure of internal consistency, was used as a basis to determine the reliability of physiological responses observed during the 10 times.

III. RESULTS

The results of psychological assessment on emotional stimuli showed appropriateness ranging from 83 to 100 percent and effectiveness ranging from 8.7 to 10.4 point (on an 11-point Likert scale [16]). Sadness, in particular, had an average of 96 percent appropriateness and 9.2 point of effectiveness, and disgust showed 94 percent appropriateness and 10.1 point of effectiveness. Also, as Tables 1 and 2 illustrate, SCL and HR showed that Cronbach's alpha was greater than .90 for both sadness and disgust emotions.

TABLE I. RELIABILITY OF PHYSIOLOGICAL RESPONSES DURING BASELINE AND SADNESS

	SCL		SCR		HR	
	Baseline	**Emotion**	**Baseline**	**Emotion**	**Baseline**	**Emotion**
1	4.67±3.18	4.49±3.88	0.02±0.06	1.29±2.86	71.17±11.70	70.07±12.86
2	3.62±2.24	3.37±2.66	0.03±0.07	0.18±0.35	68.81±9.52	66.56±9.69
3	4.10±1.81	3.21±1.72	0.04±0.11	0.29±0.55	69.08±9.32	69.68±9.19
4	3.97±1.84	3.35±1.97	0.04±0.07	0.10±0.14	72.37±10.61	71.70±11.91
5	4.55±2.47	3.91±2.94	0.02±0.05	0.46±0.78	71.43±11.20	68.72±9.99
6	4.79±2.84	5.15±3.30	0.10±0.30	0.27±0.24	72.65±9.20	73.64±14.51
7	3.75±3.06	3.78±3.56	0.00±0.00	0.65±0.96	72.77±8.36	70.26±9.87
8	4.56±3.01	3.74±2.91	0.08±0.20	0.26±0.55	71.83±10.97	69.43±12.85
9	3.38±1.45	3.05±1.70	0.09±0.21	0.31±0.47	73.98±14.37	71.67±12.56
10	4.60±3.13	3.79±3.05	0.13±0.33	0.39±0.59	74.82±12.95	73.31±12.18
M	4.20	3.66	.08	.42	71.90	70.50
α	.96	.97	.58	.79	.96	.97

TABLE II. RELIABILITY OF PHYSIOLOGICAL RESPONSES DURING BASELINE AND DISGUST

	SCL		SCR		HR	
	Baseline	**Emotion**	**Baseline**	**Emotion**	**Baseline**	**Emotion**
1	4.59±3.05	6.45±5.31	0.01±0.02	1.15±1.62	72.83±11.41	68.63±12.19
2	3.58±2.11	4.65±3.28	0.03±0.09	0.62±0.64	70.01±9.41	63.31±11.28
3	3.85±1.69	3.94±2.62	0.07±0.20	0.43±0.59	73.66±7.69	71.31±9.72
4	3.84±1.85	4.53±2.70	0.02±0.07	0.33±0.36	72.02±8.64	66.35±14.84
5	4.60±1.75	5.24±2.61	0.02±0.06	0.98±1.39	71.82±10.06	66.25±10.69
6	4.72±2.53	5.15±3.30	0.13±0.04	0.51±0.31	73.66±10.42	71.52±14.75
7	4.25±3.50	4.29±3.54	0.12±0.25	0.53±0.56	70.07±9.51	70.56±11.63
8	3.95±3.17	5.47±3.73	0.07±0.13	0.77±0.49	71.31±11.14	67.10±12.83
9	3.27±1.54	3.41±1.72	0.12±0.24	0.35±0.36	76.23±12.95	74.32±14.53
10	4.57±3.79	3.06±2.37	0.07±0.22	0.02±0.05	75.60±13.60	69.56±16.13
M	4.12	4.62	.06	.57	72.72	68.89
α	.94	.95	.70	.61	.95	.96

IV. CONCLUSION

This study attempted to verify the reliability of physiological responses associated with specific emotions, namely sadness and disgust, across experiments repeated 10 times. Despite a small sample size, we identified that the physiological responses are stable during the experiment repeated 10 times using different stimuli evoking an emotion (e.g., sadness). In particular, SCL and HR having values higher than .95 indicate stability and consistency. Although the limitations of this study such as small sample size may affect the generalization, the results can be useful in developing an emotion theory, or profiling emotion-specific physiological responses, as well as establishing the basis for emotion recognition system in HCI.

REFERENCES

[1] B. H. Park, E. H. Jang, M. A. Chung, and S. H. Kim, "Design of prototype-based emotion recognizer using physiological signals," ETRI Journal, vol. 35, 2013, pp. 869-879.

[2] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," Biol Psychol, vol. 84, 2010, pp. 394-421.

[3] A. Hinz, B. Hueber, G. Schreinicke, and R. Seibt, "Temporal stability of psychophysiological response patterns: concepts and statistical tools," Int J Psychol, vol. 44, 2002, pp. 57-65.

[4] J. W. Robinson, S. F. Whitsett, and B. J. Kaplan, "The stability of physiological reactivity over multiple sessions," Biol Psychol, vol. 24, 1987, pp. 129-139.

[5] W. F. Waters, D. A. Williamson, B. A. Bernard, D. C. Blouin, and M. E. Faulstich, "Test-retest reliability of psycho-physiological assessment," Behav Res Ther, vol. 25, 1987, pp. 213-221.

[6] J. G. Arena, S. J. Goldberg, D. L. Saul, and S. H. Hobbs, "Temporal stability of psychophysiological response profiles: Analysis of individual response stereotypy and stimulus specificity," Behav Ther, vol. 20, 1989, pp. 609-618.

[7] M. Marwitz, and G. Stemmler, "On the status of individual response specificity," Psychophysiology, vol. 35, 1998, pp. 1-15.

[8] H. Lee, A. J. Shackman, D. C. Jackson, and P. J. Davidson, "Test-retest reliability of voluntary emotion regulation," Psychophysiology, vol. 46, 2009, pp. 874-879.

[9] R. Manber, J. J. B. Allen, K. Burton, and A. W. Kaszniak, "Valence-dependent modulation of psychophysiological measures: Is there consistency across repeated testing?" Psychophysiology, vol. 37, 2000, pp. 683-692.

[10] C. L. Larson, D. Ruffalo, J. Y. Nietert, and R. J. Davidson, "Stability of emotion-modulated startle during short and long picture presentation," Psychophysiology, vol. 42, 2005, pp. 604-610.

[11] M. M. Bradley, P. Gianaros, and P. Lang, "As time goes by: Stability of startle modulation," SPR abstracts, 1995, S21.

[12] J. A. Doussard-Roosevelt, L. A. Montgomery, and S. W. Porges, "Short-term stability of physiological measures in kindergarten children: respiratory sinus arrhythmia, heart period, and cortisol," Dev Psychobiol, vol. 43, 2003, pp. 230-242.

[13] H. Kaviani, J. A. Gray, S. A. Checkley, V. Kumari, and G. D. Wilson, "Modulation of the acoustic startle reflex by emotionally-toned film clips," Int J Psychol, vol. 32, 1999, pp. 47-54.

[14] L. W. Hawk, and E. W. Cook, "Independence of valence modulation and prepulse inhibition of startle," Psychophysiology, vol. 37, 2000, pp. 5-12.

[15] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," Psychometrika, vol. 16, 1951, pp. 297–334.

[16] R. Likert, "A technique for the measurement of attitudes," Arch of Psychol, vol. 140, 1932, pp. 1–55.

 ISBN: 978-1-61208-447-3

Moral Behavior and Empathy Modeling through the Premise of Reciprocity

Fernanda Monteiro Eliott*, Carlos Henrique Costa Ribeiro†
Computer Science Division. Aeronautics Institute of Technology
São José dos Campos, Brazil
Email: fernandaeliott@gmail.com*; carlos@ita.br†

***Abstract*—We may get the opportunity of conceiving modeling artificial moral behavior and empathy if we renounce the perspective of an immaterial soul playing a role in the process of moral behavior. Philosophers such as Michel de Montaigne wrote that the laws of consciousness, supposed to emerge from nature, are essentially born from custom. Hence, we may provide a basis to that modeling if we pore over moral behavior as a form of cooperation built upon customs among emotions and feelings (as part of cognition). With this perspective in mind, we describe herein a bio-inspired computational multiagent architecture composed of artificial emotions, feelings and by an Empathy Module responsible for providing an action selection that rudimentary mimics moral behavior. The Empathy Module follows a reciprocity assumption as its main design concept. As relations between different subjects can be represented by networks, we explore different network topologies that can characterize the agent-agent interactions, by defining the moral agents neighborhood. For assessment of the proposed architecture, we use a version of an evolutionary game that applies the prisoner dilemma paradigm to establish changes over the network topology. Our results indicate the feasibility of artificial moral behavior leading to cooperative selection of action when applied in environments (networks) whose reciprocity assumption works in accordance with the environmental topology: networks with neutral assortativity w.r.t. node degree (*i.e.*, agent neighborhood size) fit more closely with the leading premise of our Empathy Module than those with a disassortative degree correlation.**

Keywords–Artificial moral machine; Empathy; Biologically inspired architecture; Evolutionary game; Assortativity in networks.

I. INTRODUCTION

The complex behavior of living things enlivens research and incongruous reasoning. Despite that, we may get the opportunity of conceiving modeling artificial moral behavior and empathy if we renounce the perspective of an immaterial soul playing a role in the process of moral behavior. The artificial modeling of bio-inspired mechanisms embodies a positioning on the premises and assumptions inherited from the selected and pursued theoretical biological references. The development of a bio-inspired computational multiagent architecture supposed to mimic moral behavior has to be grounded on biological and philosophical investigation to provide a coherent construction and an intuitive working system dynamics. Thinking through the constitution of a group and its members attendance, Tomasello [1] regards cooperation as a sewing up action that connects the members of the group. By using an evolutionary perspective, morality could be conceivable as a form of cooperation: through matching skills and aims for cooperation, morality may emerge [2]. Moral behavior consists of following the set of rules from the group, keeping it cohesive, and a gradual incorporation of new customs can change that set. According to Montaigne [3], when we reiterate a custom and naturally incorporate it among our thoughts and ideas we submit to it and establish it; therefore, the laws of consciousness, supposed to emerge from nature, are essentially born from custom: the common judgments and ideas tacitly respected among our group show themselves as general and natural. To approach the human judgments fallibility and weakness, in Montaigne [4] humans are compared with the other animals and the pyrrhonic suggestions from Empiricus [5] are delineated: our ways of interacting with the environment are fragile. Our bodies, reasoning, interpretation and capabilities are subjected to uncertainty and debate. Supposing we had other sense organs, our apprehension of the world and interaction with it could be different.

Montaigne [3] also addresses the judgments and customs relativity and fallibility. Different groups usually have different customs and follow different laws and rules. Since we are fallible, the groups common behavior provides us with a guidance and, given its continuous application, an indication of the most provable consequences given to its application. Therefore, the modeling of a bio-inspired computational architecture supposed to mimic moral behavior may benefit itself from reflections over moral behavior as a form of cooperation built upon customs among emotions and feelings (as part of cognition) - and it is pertinent to seek the human universal perspective. There are some dilemmas regarding the feelings and emotions participation on judging our actions while interacting with others. Would the human being be naturally sociable or would the sociability have emerged to ensure survival? Would we be the *Zoon Politikon* from Aristotle [6], or the *bon sauvage* from Rousseau [7], or still would our nature be better translated by the fear of all against all [8]? Thus which should be our positioning while designing an artificial empathy module? Should we design an artificial empathy on the "Machiavellian" [9] guidance? Nonetheless premises do have to be assumed.

The division of our paper is built as follows: in Section I, as a preliminary background to think through our computational architecture, we introduce some moral-related philosophical perspectives, as well as our bio-inspired motivation. In Section II, our artificial moral architecture and its Empathy Module are both described. Section III details the experimental setting built to test the feasibility of our artificial moral architecture. In Section IV we analyze the obtained results with the purpose of elucidating the implication of the reciprocity design concept from the Empathy Module. Finally, in Section V, we provide our final remarks.

From a biological standpoint, Damásio [10] highlights the relevance of social emotions and feelings as empathy to the equilibrium of humans homeostatic goals. Moreover, from a cognitive aspect the dynamics involved on the existence of empathy can be approached while holding an emotional background [11] [12]. Truly, emotions and feelings contribution on aiding humans on making faster and more intelligent decisions was already detailed in Damásio [13], inspiring the single agent driven bio-inspired Asynchronous Learning by Emotion and Cognition (ALEC) computational architecture from [14] [15]. Before choosing an action, ALEC is influenced by artificial homeostasis and by a cognitive system motivated by the Clarion Model [16]. With the aim of establishing its internal equilibrium (*i.e.*, holding its internal variables within a threshold), ALEC has to achieve artificial homeostatic goals.

We used the ALEC computational architecture as the outset of

 ISBN: 978-1-61208-447-3

our bio-inspired computational multiagent architecture $MultiA$. The design of $MultiA$ was guided by thoughts on the pertinence of moral behavior to attain a rational *and cooperative* bio-inspired artificial agent. Our leading hypothesis relies on the idea that cooperation can emerge from the assistance of emotions and moral behavior during the process of decision making — even when selfish behavior is rewarded by high reinforcements. The analogy with moral behavior is promoted through simulating the feeling of empathy. The importance of such feeling is its function on regulating $MultiA$ agents priorities making possible the selection of actions that may not be the best selfish selection. Non selfish decision making may be crucial to equalize the interactions among agents and bring up cooperation.

We provided an outline for computational moral modeling in [17], heightened by biological references on basic and social emotions, and mirror neurons (from [18] [19]). Examples of computational moral models were also mentioned. In [20] we detailed our computational architecture based on [17] and discussed some preliminary results. Herein we show the feasibility of our artificial moral architecture and present new results with the purpose of elucidating the implication of the reciprocity design concept from the Empathy Module (EM) of $MultiA$. The EM is a constituent part of the Cognitive System (CS) and determines the intensity of the emotion responsible for feeding the feeling of empathy.

II. The MultiA Architecture and the Empathy Module

Having detailed $MultiA$ in [20], herein we only summarize a few of its key points. As long as our research is grounded on moral behavior, we intend to test and study $MultiA$ agents interacting among themselves. Thus, each $MultiA$ agent i will keep a list of every agent it has interacted with (then neighbors of i). The $MultiA$ architecture consists of three main systems (Figure 1): the Perceptive (PS), the Cognitive (CS) and the Decision Systems (DS). The collaboration between the three systems will result in the selection of actions derived from sensations triggered by the environment while provoking environmental changes that will, in turn, trigger new sensations, and so on. $MultiA$ artificial sensations (all in the range $[0, 1]$) are triggered by reinforcements, and by an identifying index for the neighbor it is interacting with: every $MultiA$ agent has an identifying index $i = \{1, ..., N\}$. Likewise, the neighbors relating to each agent i also have an identifying index $p = \{1, ..., K\}$. A given p value thus refers to a particular neighbor that is interacting with i. There are basic emotions $=\{E^b_{1,i}, E^b_{2,i}, ..., E^b_{d,i}\}$ and social emotions $=\{E^s_{1,i}, E^s_{2,i}, ..., E^s_{y,i}\}$, all normalized to the range $[-1, 1]$. The basic emotions are associated with the general condition of the $MultiA$ agent itself. On the other hand, social emotions are stimulated by neighbors and by the impact of the own agents actions on those neighbors. The artificial feelings $= \{S_{1,i}, S_{2,i}, ..., S_{z,i}\}$ also fall in the range $[-1, 1]$ and are fed by emotions. For a complete description of feelings and emotions, see [20].

The artificial sensations feed emotions, feelings and, afterwards, through a weighted sum on feelings, the general perspective of $MultiA$ (named Well-Being, W_i) about its own performance. $MultiA$ follows its artificial homeostatic goals, which consist of keeping its feelings within a threshold with the aim of achieving high levels of W_i. The feelings maintenance on a threshold relies upon the selection of adequate actions in response to the environment. W_i uses feelings to internally represent the general condition of agent i, and is calculated with normalizing weights, such that the final value will fall in the range $[-1, 1]$. W_i enlightens how suitable has been the action selection (from DS) concerning the reinforcements received by the $MultiA$ agent itself, but also to the remaining feelings, as empathy. The last is represented by $S_{4,i}p$: feeling number 4 of $MultiA$ agent i for neighbor p; in Figure 1, see feeling number 4. We designed the empathy to reflect the impact of the action selection of $MultiA$ on its neighbors. Therefore, the higher the empathy for a specific neighbor p, the lower is W_i, all the remaining variables that feed W_i kept constant. This means that the $MultiA$ agent may not have been selecting its actions appropriately, since it may be affecting negatively on this particular neighbor p, thus high empathy levels are an indication of inadequate action selection. Selected actions are considered adequate when they do generate positive reinforcements while not provoking high empathy levels. If p fires high empathy on i, p may be getting low reinforcements and therefore its neighbors, such as i, should check their actions.

The CS delivers five sets of data to the PS: 1. the current number of neighbors of agent i; 2. the reinforcements history of agent i; 3. the number of times agent i has interacted with each neighbor p; 4. the number of times interacting with p ended up in positive reinforcements; 5. the CS accesses to the current emotions from PS. Then, the EM (from the CS) produces W_pi: an assumption on i related to the current condition of neighbor p. If p is supposed to be facing low reinforcements, $MultiA$ may have its empathy raised to select less selfish actions and try to cooperate with the raise of the reinforcements of p. Regarding W_pi, the CS delivers it to the PS, where it will stimulate the social emotion $E^s_{4,i}p$ (social emotion number 4 of agent i for neighbor p; in Figure 1, see social emotion number 4), then reaching the empathy feeling $S_{4,i}p$. The PS will then calculate its artificial emotions, feelings and W_i. In the PS the emotion $E^s_{4,i}p$ is fed both by W_pi, and by the empathy feeling by p right after the last interaction with p, a residual value from the past influencing the current emotion. Then, right before a new interaction with p, the empathy feeling is fed both by the emotions $E^s_{4,i}p$ and $E^s_{3,i}p$ (social emotion number 3 of agent i for neighbor p; in Figure 1, see social emotion number 3). The last summarizes the utility of neighbor p: the average number of times interacting with neighbor p has resulted in positive reinforcements.

Regarding the EM, a reciprocity assumption works as the main design concept on generating W_pi - and subsequently the empathy feeling. Ergo the EM reproduces a reciprocity assumption: a) due to neighbors mirroring, following a premise of similarity between agents and neighbors current situation. Even though we are aware of the controversy relating to mirror neurons (as in [21]), we used it as motivation on a mechanism for projecting the agents own emotions to mirror other agents' condition - thus we avoid explicit data sharing among local agents. We call that as emotionally reciprocal guidance. Thus, no agent can observe the neighbors actions or reinforcements, but only mirror its own emotions on neighbors to make assumptions about their condition. Therefore, the EM mechanism of generating W_pi was motivated by the notion of mirror-neurons internally mirroring the current condition of another agent, then, a set of the agents own emotions are used to emulate another agent p situation (before interacting with it) and to provide W_pi; b) reciprocity on the way those mirrored emotions are going to be interpreted. We settled the utilitarian calculus from [22] as our guideline on determining how those mirrored emotions would be interpreted on the EM. Thence the $MultiA$ agents have a more sensitive empathy for those agents whose interactions have been resulting in positive reinforcements (it is neighbor reciprocal). Furthermore, the $MultiA$ agent is more likely to cooperate if it has been receiving in general (from its neighborhood) a high number of positive reinforcements. We also apply a reciprocity

 ISBN: 978-1-61208-447-3

assumption through the utilitarian design: the final value of W_pi is motivated by reciprocity. Hence neighbors whose interactions result in positive reinforcements (it only has to be positive; there is never a comparison between positive reinforcements) tend to lead to higher empathy levels.

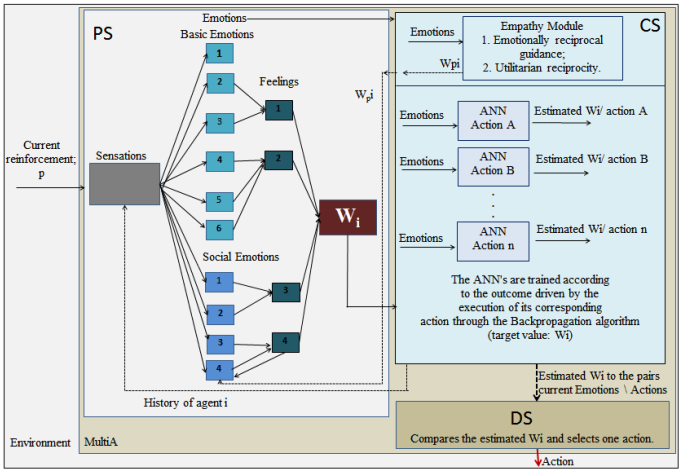


Figure 1. The general scheme of the MultiA Architecture.

In general, those actions related to high empathy are designed to be avoided, since it is considered that when a neighbor rouses high empathy it is because the agent itself may be disturbing the performance of the others. The consequence is that $MultiA$ is designed to seek those actions that will not increase its levels of empathy. The CS applies three-layer feed-forward artificial neural networks (ANNs), one for each action, and the Q-Learning reinforcement learning algorithm [23] to estimate the resulting Well-Being (provided from the single output unit) if, concerning the current emotions (input space from PS) and bias, the equivalent action is to be selected. Each ANN is trained in accordance with the outcome driven by the execution of its corresponding action [24] through the Backpropagation algorithm [25] employing W_i as the target value. The CS will then deliver the outputs from all ANNs to the DS to choose an action with the highest output (in case of existing outputs with the same value, selection will be random), except during the beginning of a simulation, when it will be use a high exploration rate for the state(emotion)-action space.

III. Experimental Setting

A. The Evolutionary Game: Task and Changes on Topology

Relations in natural societies can be analysed through public goods games analogies, with public goods characterized by two main features: they are public and they are not wasted through consumption. It can be shown that those games generalize, to an arbitrary number of individuals, the Prisoner's Dilemma Game (PDG) [26]. In natural societies and games that use them as a metaphor, situations described by unfair relations are common: an agent taking advantage of another agent social commitment. The last may be required to accomplish the best social outcome: for a pack of non-solitary animals, it may be crucial to go hunting together, each one selecting those actions that only as a group will result on the best social outcome. Since cooperating with the group usually inquires a cost to the cooperator and defectors benefit from common resources [27], a dilemma emerges between each one's self-interest and the group's maintenance.

The performance of $MutiA$ agents and the changes on neighborhoods promoted by the agents interactions that we intend to analyze will follow from the generalized PDG model of [28]: it starts with a network where agents are represented by the nodes while the neighborhood by the links between the nodes. Evolutionary games are described in [28] and related to the emergence of cascading failures: agents (nodes) and links being eliminated from a network as the matching result of agents actions. The outcome from a few agents (and its links) elimination may cause another agent elimination. The process can continue until the complete elimination of all links and agents. Wang *et al.* [28] also present a generalized PDG model where connected agents through a link (considered neighbors) choose to defect or to cooperate and the matching strategies will define the nodes reinforcement. Once all agents have interacted with each and every neighbor, a match ends and the individual sum of reinforcements of each agent is calculated, therefore a match is defined by all agents interacting only once with every neighbor. Matches are repeated in sequence until the network topology stops changing as the consequence of agents interactions.

Agents strategies are established before the beginning of the first match, but at the end of each match there is a probability of agents changing their strategy by imitating a neighbor with high final reinforcement. Just before that, the agents that did not get enough cooperative actions from neighbors (then low reinforcements) are eliminated, causing changes on the network topology. If unilateral defection (one agent cooperates and the other defects) renders a higher reinforcement value than the other matching strategies, higher will also be the probability of a cooperating node imitating a defective neighbor. Straightforwardly the defective strategy can spread to the network in such a way that it causes a cascading failure effect: cooperative agents simply being eliminated and their elimination causing neighbors elimination (of both defectors and cooperators).

B. Environment: Networks Initial Topology

As long as networks can be used as metaphors to represent diverse systems [29], the environment where our agents will try to accomplish the task will be delineated by them. In the literature there are different models to construct networks, each one of them ensuring different features emerging from the model application. Indeed the model selection has to fit in the network usage. Insofar as we want to mimic moral behavior (many agents from diverse neighborhoods interacting among themselves), it is relevant to apply a model that provides high clustering and long-range connections.

Our undirected networks were constructed through the growing networks model proposed in [30], supposed to unify certain features of real networks, as a power-law degree distribution ([31] [32]) and the small world effect - high neighborhood clustering and short average distance between the nodes. A power-law degree distribution typically results from a network growing process called preferential attachment, often displayed by real networks [33]: once a growing network is about to receive a new node, the ones that already have more links are more likely to be connected to the new node. Wherefore, the node age in the network is relevant on defining the number of links it will have [34]. According to Klemm and Eguíluz [35] ([30] derived from it), its model of generating scale-free networks presents real-world properties, as a negative correlation between the age of a node and its link attachment rate. On the other hand, the basic reference on growing scale-free networks [36], would present a mean attachment rate positively correlated with age (as the attachment rate is proportional to the degree and the oldest nodes start accumulating links since the beginning of the construction of the network). Additionally, in opposition to Barabási and

 ISBN: 978-1-61208-447-3

Albert [36], the growing networks from Klemm and Eguíluz [35] preserve the degree distribution (still power-law), even if all but the most recently grown part are disregarded. The model [35] produces high clustering scale-free networks and, even though its clustering is higher than in regular lattices, its topology is similar to one-dimensional regular lattices.

The model presented in [35] may include long-range connections (originating from [30]), aiming to obtain small path length [29] while holding the original properties of high clustering and scale-free degree distribution. The guidelines to construct our tested networks are: consider a network that shall end the growing process with N nodes (since we are going to test N agents) and each of them will be taken as active or deactivated. The growing process starts with m active nodes completely connected. Then, until the size of the network grows to N, a new node: 1)is considered in active state and will be connected to m different nodes. For each of the m connections, a decision shall be made: a)the connection will be made to a random active node or b)to a general random node. The probability of connecting to a general random node is μ, that case the random node is chosen following Linear Preferential Attachment; 2)The new node is activated; 3)One of the m nodes is deactivated. The deactivation process was inspired by a memory idea: in general, the newer nodes in the network are more likely to receive links than the older ones - as an example, consider technical papers referencing more recent works rather than older ones.

To examine the degree correlation of our created networks, we used the assortativity coefficient ϱ from Newman [37] with the purpose of studying $MultiA$ agents performance over the influence of a) disassortative degree correlation (*i.e.*, negative values of ϱ, when highly connected nodes have the propensity of being connected to the nodes that are little connected [38]; and b) neutral degree correlation (neutral values of ϱ), when there is not such a propensity, be it to little connected or highly connected nodes. A common condition given by the disassortative degree correlation is the existence of *polarized nodes*: the ones that have just a few links but, most importantly, are connected to highly connected nodes.

IV. RESULTS

The relevant definitions and constraints used to describe our results are:

1. We call game a set of interactions among agents as defined in the PDG by Wang *et al.* [28]).

2. An elimination process will always occur at the end of each match t. Each match will be given by all non-eliminated agents (represented by the corresponding nodes) interacting with those agents they are linked with (neighbors). No agent will interact (choose to defect or cooperate) twice with the same neighbor in the same match.

3. A simulation is a defined number of matches played in sequence.

4. The initial number V_i^1 of neighbors for each agent i will be given by the network original topology.

5. Reinforcements are normalized to $[-1, 1]$. Each agent i has to end up a match t with an individual sum of reinforcements r_i^t at least equals to T_i. The parameter T_i represents a minimal individual survival need (T_i falls in the range $[0; 1]$). The values of reinforcements resulting from the agents interactions will follow as from the following example: suppose that the agent 0 initially has 4 links in the network. Then, it has 4 neighbors ($V_0^1 = 4$). Agent 0 will be represented by the node zero and will receive reinforcement $1/V_i^1 = 0.25$ for mutual cooperation; $2 * 1/V_i^1 = 0.5$ for defecting when a neighbor p cooperates (unilateral defection) and, finally, zero for mutual defection or for cooperation vs. defection. We made $T_i = 0.5$, thus the cooperative agent will need half of its neighbors cooperating to avoid elimination and the defective will only need 25% of it. By receiving a double reinforcement (comparing to mutual cooperation), the defective agent will get the chance of being more resilient in the network (when it has cooperative neighbors) than cooperators. If the agent ends up a match with $r_i^t < T_i$, the network topology changes: the agent itself and all its connections are eliminated. Observe that after elimination, concerning those agents that have fewer neighbors at match t than in the first match: if they follow a cooperative strategy they will never have the chance of getting the full reinforcement of 1 (as V_i^1 will be higher than the current number of neighbors at match t).

6. The results were collected only when the size of the network stopped changing (t^F matches). Aiming to prevent a massive elimination of agents during exploration time (from first match until t^x) and a small number of upcoming matches (from t^{x+1} until t^B), the real elimination process only starts from match t^{B+1}. Despite that, the neighbors of those agents that should have been eliminated during the matches $t < t^{B+1}$ actually do receive information about neighbors elimination. That intervention puts forward some issues, as the mismatch between loss of neighbors and lower reinforcements (as all neighbors of a given agent i will be kept on network, allowing the possibility of the full reinforcement of 1 for cooperation on the next match).

7. In order to present the final results relating to the network topology, we called ρ_d the percentage of defectors in the final network, and ρ_c the percentage of cooperators. The percentage of remaining nodes from the original network (agents that have not been eliminated) is ρ_f.

8. The experimental parameters applied on all simulations are: DS uses a 10% exploration rate, the hidden and output units from DC use the logarithmic activation function and we applied a learning rate 0.07 and a momentum term 0.9.

A. Moral Agents and Degree Correlation

We developed two agent versions. Notice that the well-being W_i (from PS) is calculated with normalizing weights on feelings so that the final value falls in the range $[-1, 1]$. Thus the weights have to be set respecting the relevance of each feeling to the domain. In general, the feeling $S_{1,i}$ is sensitive to the neighbors elimination and $S_{2,i}$ to the agents own reinforcements. The feeling $S_{3,i}$ represents the average number of times agent i has been receiving positive reinforcements and $S_{4,i}$ is the empathy feeling. The agents are:

- The $MultiA$ agent designed to rudimentarily mimic moral agents, with a weight of the empathy feeling over the W_i value supposed to be considerable. The feelings weights used on our experiments are: $S_{1,i} = 0.4; S_{2,i} = 0.05; S_{3,i} = 0.05; S_{4,i} = -0.5$. Thus the empathy feeling ($S_{4,i}$) is responsible for half of the value of W_i. The well-being W_i measures the performance of the $MultiA$ agent in the environment and, if the empathy reaches high levels, W_i will be low. That is an indication that probably the last selected actions may be causing bad outcomes to neighbor p; therefore, the well-being W_i of agent i should be low, even though its reinforcements may be high.
- The $MultiA^A$ agent that rudimentarily mimic amoral agents. We provided theoretical ideas about immoral and amoral agents in [20] [17]. The amoral agent lacks social

 ISBN: 978-1-61208-447-3

emotions and feelings fed by them (it also lacks an EM). Therefore, it has the 6 basic emotions and 2 feelings. The ANNs from CS were adapted accordingly. For this version we tested two different weights set on feelings: $S_{1,i} = 0.3$; $S_{2,i} = 0.7$; and $S_{1,i} = 0.5$; $S_{2,i} = 0.5$. As we had better results on the former setting, we used that.

The agents task is to learn to avoid elimination, and that involves a compromise: learn to accumulate high reinforcements at the end of each match t ($r_i^t >= T_i$) while avoiding neighbors elimination. Notice that the outcome (r_i^t and neighbors elimination or not) of each agent i at match t is due to its own actions, to all its neighbors p actions *and* also to the neighbors actions of its neighbors p. That means the strategies cause-consequence can easily be shadowed: both, defective and cooperative agents may loose a neighbor, influencing both agents PS. If the agents have difficulties learning the task on a given network topology, it may be helpful to increase the exploration rate and make t^x and t^B encompass more matches. However, depending on the network topology, most of the agents will not learn the task at all. Thus, an insufficient number of agents will learn to cooperate and, given to the number and/or position of defectors within the network, a cascading failure effect may occur: all agents will be eliminated. That is an indication that the matching agent-network properties should be reconsidered.

Although we are not going to present it here, we already have preliminary results indicating that $MultiA$ also benefits from highly connected networks as in [30] (increment on the m parameter). By increasing m, we are also increasing the number of links between nodes and likewise the number of completely connected nodes (from the beginning of the network growing process). Those preliminary results also make sense with the EM: a generalized increment on the links number (even if allowing differences on the nodes degree) may produce so highly connected nodes that the actual number of each node's links may loose its importance (allowing similar environmental condition to the $MultiA$ agents through emotions mirroring). Herein through different networks, we explore the relations between small values of m and the effect of degree correlation on $MultiA$s failures. We produced networks with different ϱ by changing the μ value. Thus by varying the value of μ and keeping $m = 1.8\%$ of N, $N = 2000$, we built different networks for the Experiment 1 (Exp.1). For each network, we ran 10 simulations for each of our $MultiA$ and $MultiA^A$ agents. Observe that each network topology is used to define the agents interactions. Therefore, through different network topologies (diverse ϱ given the μ value), in Figure 2 we show both agents simulations that did not lead to a failure (when it occurs a cascading failure effect). Those had $\rho_f > 90\%N$, $\rho_d < 50\%$ and $\rho_c >= 50\%$.

With the purpose of enlightening our results of Figure 2, some observations shall be made: 1. For $\mu > 0$: if we increase the value of μ, nodes with more links will be more likely to be linked to the new nodes in the growing network (making it possible that deactivated nodes with more links, but older in the network, receive the new links). If $\mu = 1$, the model becomes [36], then ϱ tends to a zero value. Regarding the preferential attachment brought through $\mu > 0$: by increasing μ, older deactivated nodes have the chance of receiving new links, then forming new connections between different neighborhoods: older nodes with high number of connections keep receiving new links. The process of avoiding establishing a number of links once the node is deactivated allows a tendency to a neutral assortativeness - as $\mu = 1$ should return to the model from Barabási and Albert [36], [39]; 2. For $\mu < 1$: if we diminish μ, the newest nodes from the list of active ones are given the chance to connect to new nodes (since the selection from the active list is random). When $\mu = 0$, it returns to the original model [35]. The memory process of "forgetting" nodes (deactivating them) promotes networks tending to a disassortative degree correlation (negative values of ϱ). The cross-over ($0 < \mu << 1$) between the two models ([36] and [35]) would reproduce the real networks features.

Considering the reciprocity assumption from the EM, the $MultiA$ agents will achieve better results in networks that provide similar environmental conditions to the neighborhood (as neighbors with similar number of links). By the same reason, its performance is affected by neighborhoods with highly different degree distribution. As shown in Figure 2a), $MultiA$ agents were able to solve the task on networks with a tendency to a neutral assortativeness. Polarized agents (from networks with a negative ϱ) are less impacted by a shadow effect (cooperators loosing neighbors the same way as defectors) than those agents with higher number of links. But given the agents neighborhoods differences (number of neighbors) in those networks described by negative ϱ, polarized and non-polarized $MultiA$ agents EM ends up mirroring neighbors inadequately, leading to a cascading failure effect. On the contrary, a most similar neighborhood benefits the emotional mirroring as well as the EM, preventing from high oscillations in the learning process (see Figure 3a) allowing the agents to solve the task on networks with larger ϱ.

Considering Figure 2b), it is important to note that no agent can observe the neighbors actions or reinforcements. $MultiA^A$ agents are able to solve the task in environments where the matching of strategies in the way of conquering high reinforcements and neighborhood upkeep is easier. Then, as mutual defection renders reinforcement zero, agents learn to match their action selections in a way of avoiding mutual defection (then, cooperating). When different neighborhoods are connected (by long-range connections through high values of μ), more opportunities of matching agents strategies are created. Thus, on the networks with an almost non-negative ϱ, the neighborhoods tending to a cooperative strategy will be affected by connections with unstable or more defective neighborhoods. On those environments described by topologies that repeatedly allow different scenarios (reinforcements and agents elimination) for the same strategy, the agents will be influenced by the strategies combinations possibilities, then the shadow effect will strongly impact on agents DS. That makes it harder for the $MultiA^A$ agents to solve the task. As Figure 2b) shows, $MultiA^A$ agents were able to solve the task on networks with high disassortative degree correlation (negative ϱ). The polarized agents (fewer neighbors equals to fewer matching strategies possibilities) can find more easily the matching strategy that keeps their internal variables balance (feelings). From the polarized nodes strategy establishment, it becomes easier for its neighbors to define their own strategies. Then, a positive cascading effect happens: once the polarized agents have defined their strategies, it is easier for the highly connected agents to find their own strategies.

B. Agents Learning Dynamics

We run two more experiments (using two networks from the first experiment) to present our agents learning dynamics. As $MultiA$ and $MultiA^A$ had different performances for the tested values of μ (given m and N) from Exp.1, we used a network with an almost neutral ϱ on simulating the former (Exp.2, Figure 3) and a network with a negative ϱ on the later (Exp.3, Figure 4). Regarding Exp.2, Figure 3, the parameters used to create the

 ISBN: 978-1-61208-447-3

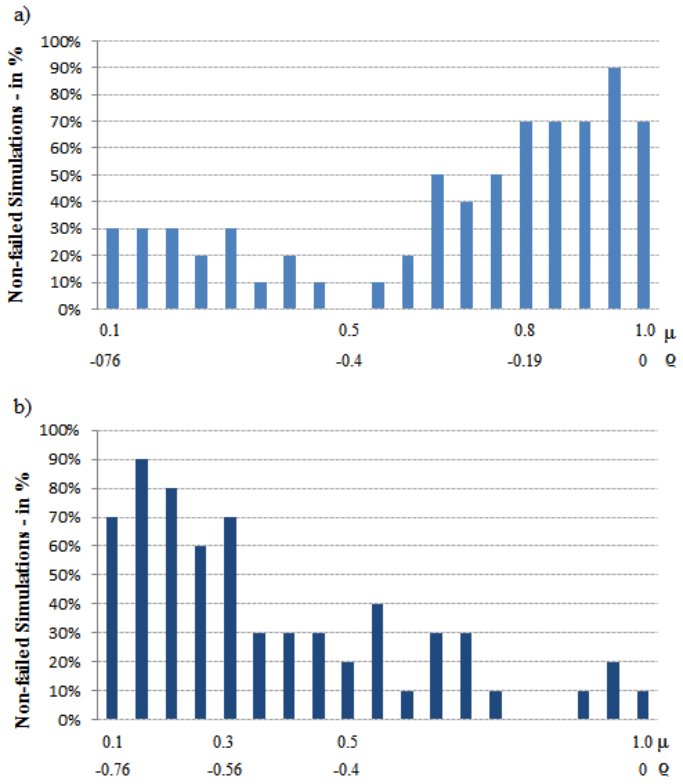


Figure 2. Degree correlation and agents performances: agents simulations that did not lead to a failure. a)$MultiA$ agents; b)$MultiA^A$ agents. The parameters we used: $Ti = 0.5$, $N = 2000$, $m = 1.8\%N$.

network are: $Ti = 0.5$, $N = 2000$, $m = 1.8\%N$, $\mu = 0.95$. Given the μ value, the network had $\varrho = -0.05$. The parameters used to create the network of Exp.3, Figure 4, are: $Ti = 0.5$, $N = 2000$, $m = 1.8\%N$, $\mu = 0.1$. Given the μ value, the network had $\varrho = -0.76$. Observe that in Figures 3 and 4 we do not consider the agents that should have been eliminated from the first match until t^b (real elimination starts at t^{b+1}). Both the mean reinforcement $\bar{R}$ and the mean cooperative neighborhood size $\bar{C}$ were averaged over 20 simulations, and the values of t^x and t^B where set according to the experimental minimum possible values to allow learning while preventing from a cascading failure effect.

The defective strategy showed itself to be a bad decision when there is mutual defection and when it causes neighbor elimination (or only an indication of it, during the first t^B matches). The real elimination (from match t^{B+1}) contributes to the network deterioration, as that may cause a cascading failure effect: the elimination of cooperative nodes causing its neighborhood elimination, and so on. As the outcome of each agent results from its own actions face to the action selection of its neighborhood, the strategy cause-consequence link can easily be shadowed: both defective and cooperative agents may lose neighbors. Another issue is that agents have access to the number of eliminated neighbors only when the match ends. However, once the agents strategies stabilize in such a way to prevent agents elimination (*i.e.*, cooperators and defectors do have a sufficient number of cooperating neighbors), the elimination process ends. The consequence of that is that a bad effect of the defective strategy (neighbors elimination) will stop influencing the agents DS. It is worth considering that there is no local agent (both $MultiA$ and $MultiA^A$) access to neighbor reinforcements — in the case of $MultiA$, the EM tries to mirror the neighbors current state before the DS selects an action. The successive interactions among agents will impact on the PS, *ergo* on the whole architecture, and the action selection (both in the same match and from a match to another) will be influenced by previous interactions.

To better interpret the results in Figures 3 and 4, a general analysis is required: a similar drop on $\bar{C}$ and $\bar{R}$ at the same match indicates mutual defection *or* cooperators elimination. If both $\bar{C}$ and $\bar{R}$ increase, mutual defection is being replaced by unilateral defection or mutual cooperation. If $\bar{C}$ increases and $\bar{R}$ drops, unilateral defection is being replaced by mutual cooperation. When $\bar{C}$ drops and $\bar{R}$ increases, mutual cooperation is being replaced by unilateral defection. Notice that if an agent with fewer neighbors defects and some of its neighbors cooperates, this defector will accumulate high reinforcements more easily than a highly connected defector: *e.g.*, if the defector agent 0 has 2 neighbors (node 0 has two links to other nodes) and one of them cooperates, agent 0 will easily get the full reinforcement of 1 ($V_0^1 = 2$ and unilateral defection for agent 0 will be $2{*}1/V_0^1 = 1$). On the other hand, if the defector agent 5 has 100 neighbors and just one of them cooperates, agent 5 will get the reinforcement of 0.02 ($V_5^1 = 100$ and $2 * 1/V_5^1 = 0.02$).

The $MultiA$ agent performance in Exp.2 is illustrated in Figure 3. Between the first and third matches, once the agents get higher reinforcements for unilateral defection, they start to defect. Then mutual defection results in lower reinforcements and elimination. At the same time, unilateral defection ($\bar{C}$ dropping more heavily than $\bar{R}$) results in information of neighbors elimination also, just on the side of cooperators, lower reinforcements. During exploration (until match t^x), even though the agents are still changing strategies, the EM prevents from high oscillations in the learning process. Right after exploration ends, some agents emphasize the defection strategy, leading to the unilateral defection (increasing $\bar{R}$ and dropping $\bar{C}$), causing neighbors elimination (still not actually applied) and — as a reciprocity utilitarian effect — mutual defection. That causes even more cooperating agents elimination (since $\bar{C}$ becomes smaller than 50% and $\bar{R}$ also drops). Overcoming a shadow effect, from match 44 the defective agents start changing to the cooperative strategy, even before real elimination begins. The elimination induces part of the defective agents to try the cooperative strategy ($\bar{C}$ increases). But at the end (as the elimination process is done since every agent already has $r_i >= T_i$), they return to the defective strategy and all agents end up stabilizing their strategies, with the following final percentages of remaining agents, defectors and cooperators, respectively: $\rho_f = 99\%$, $\rho_d = 36\%$ and $\rho_c = \%63$.

Note that the empathy feeling impacts less on the action selection of those $MultiA$ agents that have been surrounded by agents with which the interactions did not render positive reinforcements. The consequence is that the EM will repeatedly send low levels of empathy and make the agent prioritize its other feelings, thus learning to select actions in accordance to such other feelings (specially the basic ones). Therefore, the utilitarian reciprocity design from the EM will allow a more selfish action selection. On the other hand, $MultiA$ agents that did have enough positive interactions will have a neighborhood-driven empathy feeling, following an utilitarian reciprocity policy on selecting less selfish actions.

To see $MultiA^A$ agents performance in Exp.3, note Figure 4. As the $MultiA^A$ agents are not influenced by the empathy feeling and by the EM, the changing strategies oscillations are very clear by comparing $\bar{C}$ and $\bar{R}$. Driven by reinforcements first and then by keeping neighbors, those agents go by matching strategies. When exploration ends, agents turn to the defective

 ISBN: 978-1-61208-447-3

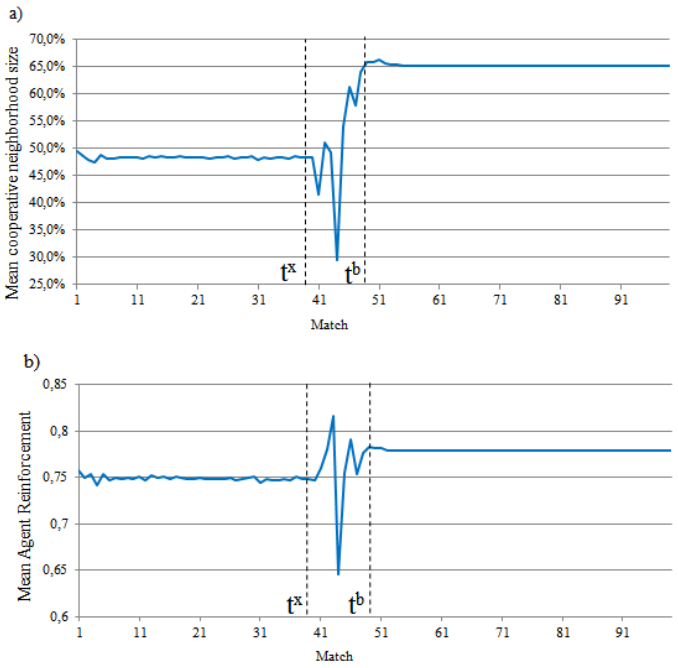


Figure 3. Results for almost neutral degree correlation networks composed of $MultiA$ agents. a)Mean cooperative neighborhood size $\bar{C}$. b)Mean reinforcement $\bar{R}$. $MultiA$ parameters: $t^x = 39$; $t^B = 49$.

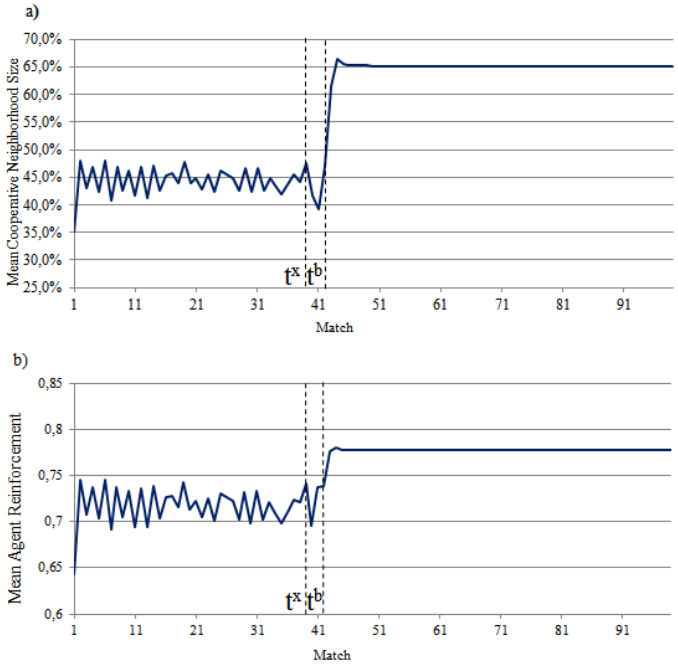


Figure 4. Results for disassortative networks composed of $MultiA^A$ agents. a)Mean cooperative neighborhood size $\bar{C}$. b)Mean reinforcement $\bar{R}$. $MultiA^A$ parameters: $t^x = 39$; $t^B = 42$.

strategy leading to mutual defection and information on neighbors elimination. That makes some agents try the cooperative strategy. Then, the actual elimination causes a high and fast change on defecting agents strategy — before that, observe that $\bar{C}$ was kept below the minimum required value to prevent cooperating agents from elimination (50% of cooperating neighbors). The low value of $\bar{C}$ kept until the real elimination begins does not cause a cascading failure effect, due to the disassortative network topology: polarized agents learn fast and stabilize their strategies and the positive cascading effect, with the following final percentages of remaining agents, defectors and cooperators, respectively: $\rho_f = 98\%$, $\rho_d = 36\%$ and $\rho_c = \%62$.

V. Conclusion

We described a bio-inspired computational multiagent architecture that considers an artificial morality component and presented both its moral ($MultiA$) and amoral ($MultiA^A$) versions. Regarding the reciprocity paradigm over the $MultiA$ design, it prevented a cascading failure effect on networks described by an almost neutral degree correlation, aiding the agents on being more successful on mirroring neighbors condition. The amoral version prevented a cascading failure effect on networks described by a negative degree correlation, but that was due to its reinforcement seeking priorities, game dynamics and network topology. The comparison between both agents versions empirically confirms the influence of the empathy model on $MultiA$ Decision System. As future work, we intend to study the performance of $MultiA$ over the influence of different tasks and positive assortativeness (the tendency of highly connected nodes being also connected among themselves [37]).

Consideration should be given to the fact that technologies are increasingly present in our daily life, progressing to a reality in which our connection with the artificial will be so deep that it will no longer make sense to distinguish between natural and artificial experiences. Thus, we also have the purpose of exploring our hypothesis from [17] regarding a hybrid agent that can trigger both moral and immoral behavior, *e.g.*, autonomously activate moral action policies with biological creatures, and immoral actions otherwise. Hence, it may be relevant an artificial agent able to simulate a moral behavior in general social or domestic assignments, *e.g.*, monitoring highly dangerous criminals, people in quarantine or in other context, where there are social dilemmas to deal with. Furthermore, the artificial morality component could be implemented as a resource in argumentation-based negotiation in multiagent systems.

Acknowledgments

The authors would like to thank CNPQ and FAPESP for the financial support.

References

[1] M. Tomasello, “Human culture in evolutionary perspective,” Advances in culture and psychology, vol. 1, 2011, pp. 5–51.

[2] M. Tomasello and A. Vaish, “Origins of human cooperation and morality,” Annual Rev. of psychology, vol. 64, 2013, pp. 231–255.

[3] M. Montaigne, “Essays: of custom; we should not easily change a law received. (De la coustume et de ne changer aisément une loy receue),” vol. I, XXIII, 1950 (1580).

[4] ——, “Essays: Apology for Raimond Sebond. (Apologie de Raymond Sebond),” vol. II, XII, 1950 (1580).

[5] S. Empiricus, Outlines of scepticism (Pyrrhoniae Hypotyposes, PH). Cambridge U. Press, 2000 (160-210AD).

[6] Aristotle, Aristotle’s Politics. Chicago U., 2013 (350BC).

[7] J. Rousseau, The Social Contract. Penguin, 1971 (1762).

[8] T. Hobbes and E. Curley, Leviathan: with selected variants from the Latin edition of 1668. Hackett Publ., 1994, vol. 2.

[9] N. Machiavelli, The Prince. U. of Chicago, 1985 (1532).

[10] A. Damásio, Looking for Spinoza: Joy, sorrow, and the feeling brain. Random House, 2004.

[11] F. De Waal, The age of empathy: Nature’s lessons for a kinder society. New York: Harmony, 2009.

 ISBN: 978-1-61208-447-3

[12] D. Proctor, S. Brosnan, and F. De Waal, "How fairly do chimpanzees play the ultimatum game?" Communicative & integrative biology, vol. 6, no. 3, 2013, p. e23819.

[13] A. Damásio, "Descartes' error (new york: Putnam)," 1994.

[14] S. Gadanho and L. Custódio, "Asynchronous learning by emotions and cognition," in Procs. of the seventh Int. Conf. on simulation of adaptive behavior on From animals to animats. MIT Press, 2002, pp. 224–225.

[15] S. Gadanho, "Learning behavior-selection by emotions and cognition in a multi-goal robot task," The J. of Machine Learning Research, vol. 4, 2003, pp. 385–412.

[16] R. Sun and T. Peterson, "Autonomous learning of sequential tasks: experiments and analysis." IEEE Transactions on Neural Networks, vol. 9, no. 6, 1998, pp. 1217–1234.

[17] F. Eliott and C. Ribeiro, "A computational model for simulation of moral behavior," in Procs. Of the I. Conf. on Neural Computation Theory and Applications (NCTA-2014). SCITEPRESS (Science and Technology Publications), 2014, pp. 282–287.

[18] G. Di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti, "Understanding motor events: a neurophysiol. study," Experimental brain research, vol. 91, no. 1, 1992, pp. 176–180.

[19] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi, "Premotor cortex and the recognition of motor actions," Cogn. brain research, vol. 3, no. 2, 1996, pp. 131–141.

[20] F. Eliott and C. Ribeiro, "Emergence of cooperation through simulation of moral behavior," in Hybrid Artificial Intelligent Systems. HAIS 2015: 10th I. Conf. on Hybrid Artificial Intelligence Systems, Bilbao, Spain. Lecture Notes in Artificial Intelligence, vol. 9121. Springer International Pub., 2015, pp. 200–212.

[21] G. Hickok, The myth of mirror neurons: the real neuroscience of communication and cognition. WW Norton & Company, 2014.

[22] J. Bentham, An introduction to the principles of morals and legislation. Courier Dover Publications, 2007 (1789).

[23] C. Watkins, "Learning from delayed rewards," Ph.D. dissertation, Kings College, UK, 1989.

[24] L. Lin, "Reinforcement learning for robots using neural networks," DTIC Document, Tech. Rep., 1993.

[25] P. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences." Ph.D. dissertation, Harvard, 1974.

[26] J. Wakano and C. Hauert, "Pattern formation and chaos in spatial ecological public goods games," J. of theoretical biology, vol. 268, no. 1, 2011, pp. 30–38.

[27] L. Wardil and C. Hauert, "Origin and structure of dynamic cooperative networks," Scientific reports, vol. 4, 2014, pp. 5725: 1–6.

[28] W. Wang, Y. Lai, and D. Armbruster, "Cascading failures and the emergence of cooperation in evolutionary-game based models of social and economical networks," Chaos: An Interdisciplinary J. of Nonlinear Science, vol. 21, no. 3, 2011, pp. 033 112: 1–12.

[29] D. Watts and S. Strogatz, "Collective dynamics of 'small-world'networks," nature, vol. 393, no. 6684, 1998, pp. 440–442.

[30] K. Klemm and V. Eguíluz, "Growing scale-free networks with small-world behavior," Physical Rev. E, vol. 65, no. 5, 2002, p. 057102.

[31] D. Price, "Networks of scientific papers." Science (New York, NY), vol. 149, no. 3683, 1965, pp. 510–515.

[32] H. Simon, "On a class of skew distribution functions," Biometrika, 1955, pp. 425–440.

[33] S. Dorogovtsev, J. Mendes, and A. Samukhin, "Structure of growing networks with preferential linking," Phys. Rev. letters, vol. 85, no. 21, 2000, p. 4633.

[34] S. Dorogovtsev and J. Mendes, "Evolution of networks with aging of sites," Phys. Rev. E, vol. 62, no. 2, 2000, p. 1842.

[35] K. Klemm and V. Eguíluz, "Highly clustered scale-free networks," Phys. Rev. E, vol. 65, Feb 2002, p. 036123.

[36] A. Barabási and R. Albert, "Emergence of scaling in random networks," science, vol. 286, no. 5439, 1999, pp. 509–512.

[37] M. Newman, "Mixing patterns in networks," Physical Rev. E, vol. 67, no. 2, 2003, p. 026126.

[38] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," Science, vol. 296, no. 5569, 2002, pp. 910–913.

[39] M. Newman, "Assortative mixing in networks," Phys. Rev. letters, vol. 89, no. 20, 2002, p. 208701.

 ISBN: 978-1-61208-447-3

Loneliness and Relational Biography

Affective communication

Treton Cécile and Bourret Christian
DICEN IDF/ – UPEM (Paris East University Marne-la-Vallée)
Paris, France
e-mail: cecile.treton@paris.fr
e-mail: christian.bourret@u-pem.fr

***Abstract* - This paper addresses the problem of loneliness in the increasing elderly community. This can be considered as a health and social problem. The objective of this research is to identify the relational patterns of the elderly in order to help them recreate a network of relationships suited to their unique needs. The research method examines the dynamics of relationships, which include interactions such alliances and breaks, commonalities and affinities, meeting places, including the use of social networks on the web. Our work is integrated in the field of affective communication as defined by F. Martin-Juchat. The proposed work will rely on observation methods and interviews with elderly and on the notion of situation, described by A. Mucchielli. We want to create a model of intervention which uses narrative methods to understand how a person builds his social network.**

Keywords - loneliness; strategy of relations; emotional social network; age; friendly.

I. BACKGROUND AND PURPOSE

The medico-social sector is subject to numerous changes and problems. In this context, there are two aspects that we wish to explore and connect, even though they seem un-related at first. On one hand, the health sector, led by the goals of streamlining services to the population, explores the development of individualized and modeled benefits, provided remotely through digital technologies in the context of the e-health. On the other hand, the social sector whose demand is increasing in terms of compliance with standards of hygiene, safety and care, is moved by the increasing isolation situations that mainly affect marginalized people, because of their health status, their age, or their social status. It appears that the infrastructure that offers many opportunities for communication and security cannot resolve the appearance of a feeling of loneliness for a growing part of the population. The paradoxical aspect of this situation raises issues pertaining to the humanities. It seems conducive to questions that fall within the field of social innovation in terms of communication sciences. The suffering caused by loneliness disturbs. It emerges in a society characterized by the diversity of its technological possibilities of communication: mobile phones, digital social networking, mails, Web 2.0, etc.

It challenges the institutions. Indeed, E. Durkheim [1] demonstrated the consequences of the transition from a traditional to a modern society. In traditional society, the family ensures the protection of individual. In modern society, the concept of family disappears and the State supports the vulnerable individual's protection according to common standards. According to S. Paugam [2], the institutions contribute in reality to the emergence of this loneliness. They use intervention models that tend to dehumanize the relational practices perceived only as "services".

Our question is how resolve this deficiency of systems and technologies which seems unable to find a solution to those needs expressed through this feeling of loneliness?

Our research is based on three questions.

With the first question we have attempted to theoretically clarify how individuals build their relational environment? For this, we favored specific approaches belonging to sociology, including structural sociology.

Our second question concerns the meaning that the person gives to its communication activities. Indeed, we consider the person as the driving element of the relation that is established. To answer this question, we borrowed our intellectual support from philosophy and from psychology but also from neurobiology which is very active in the field of emotion and cognition.

Our third question deals with the transformation and is particularly relevant with regard to our subject. The question is to identify if there are possibilities of reorganization of the relational environment of a person and, if there are, how these changes of relation can be developed? This third question is nourished by our tools of investigations applied to the qualitative inquiries which we led, but also by the constructivist and phenomenological approaches, belonging to the sciences of communication and to the sciences of education.

 ISBN: 978-1-61208-447-3

II. METHODOLOGICAL FRAMEWORK

In this sense, our work is integrated in the field of affective communication as defined by F. Martin-Juchat [3], who perceives the emotional body as "moved by the emotion" F. Martin-Juchat [3], highlights the lack of studies on this emotional dimension, in the field of science of information and communication which is shared between two conceptual approaches. The first concerns the question of the interpersonal relationship and considers the body in terms of signs verbal and nonverbal (gestures). This approach refers to the work of F. Saussure and of the Invisible College of Palo Alto. The second highlights the manipulative attempts of mass media. In both cases, the receiver is not considered in its ability to act as if he had not feelings. According F. Martin-Juchat, [3], the receiver's action must be considered as a media. She proposes to put the "emotional body" in the heart of the communication device like a media. This approach requires paradigm shifts on two levels. First, as part of our subject, it needs to change the perception of social actors involved in the implementation of services, particularly those integrate technologies. Moreover, it is dependent on a change in posture of the person himself who becomes actor in the established communication system. By an action research, our goal is to bring the people who suffer of a feeling of loneliness to change their behavior. This transformation process must change the perceptions of actors in their relational environment and modify the perception of the way of seeing the technological communication supports. Our approach takes as reference the paradigm of "engaging communication" to the meaning of F. Bernard [4]. The concept of commitment is used to demonstrate the link between the action and the meaning given to action. The commitment depends on the situation. The "engaging communication" is based on the action as a mean of change. The identification is integrated in the processing of change through the action. The role of a mediator is to create situations conducive to changing actions.

III. RELATIONSHIP IN THE CONTEXT OF AGING

Our work integrates knowledge available in social psychology. The main causes of the sense of loneliness are related to a need for security and a lack of recognition. S. Paugam[2] believes that the feeling of vulnerability expresses a research of security. A. Honneth [5], N. Elias [6], D.W. Winnicott [7] highlight the need of recognition of individuals. They are shared between the desire to affirm its uniqueness and to be socially approved. The ageing is a stage of life that seems to reactivate these needs. Building on the contributions of sociologists and gerontologists, we find that the old age - estimated at the retirement age - is a period marked by many transitions and changes. V. Caradec [8] and P. Pitaud [9] observe that the elderly person is faced with a multiplicity of events and ruptures. These changes are due to recurrent affective losses, changes in material and economic living conditions, a decline in physical abilities, often affecting motor skills and the initial conditions of life. The old age is a particularly intense period of identity recomposition. The life changes impact the modes of socialization of the person. C. Bidart [10] emphasizes socialization change with age with a tendency to build proximity links and to keep more distance in the relation. She notes the need to more moments of intimacy. For gerontologist M. Billé [11], these times of intimacy and relational distancing correspond to the need for "interior narrative work." The elderly need to remember the past to ensure a temporal continuity at their life and preserve a consistent and positive image of their life.

IV. RELATIONAL INVESTIGATIONS AND BIOGRAPHIES

From this theoretical knowledges,we studied the relational dynamics of the elderly in order to highlight their mode of action, the uniqueness of their subjective experience and confirm temporal and spatial changes. We chose to proceed by biographical interviews for two reasons. The biographical interview allows the collection of data revealing the subjectivity of relational experience. It can be used as a means to train the person to take action. As stated by C. Delory-Momberger[12], the biography enables a work on the self images which precede action. We agree with this approach that considers the individual as an agent of his own socialization by the action. This method of data collection takes into account the "emotional body". C. Delory-Momberger notes that it is the place of biographical investment.

We conducted fifteen qualitative interviews with women and men, aged 65 to 99 years, on the theme of friendship. To analyze the biographies we used categorization method proposed by A. Mucchielli [13] through situational semiotics. A. Mucchielli raises the question of social identification. He writes: "Identify the other is a judgement for define him in a specific context. Identify others is a means to give a meaning to my "being" situated also in a context." This method allowed us to establish a frame of reference from a categorization that we have defined. These categories form "interpretive frameworks' of intentions and needs of the actor, its reference standards, its positioning in relation to other actors, the quality of relationships in a historic, temporal and sensory setting. The meaning is defined in a constructivist perspective and shows a schematic representation of the operation of the phenomena studied.

The following table shows an extract of the interpretative grid. Our goal is to highlight three aspects common to all the interviews. We find that relational environment evolves according to life events. Relationships are always built on the same pattern. This pattern is established from emotional factors related to values and beliefs. It shows the specific needs of each person and his way of composing relations through the choice of communication spaces and rhythms of interactions.

 ISBN: 978-1-61208-447-3

Table: study of biographies

	Categorizations		
	Biographical frame	***characterization of the relationship***	***Methods activation***
D.	Boarding school Work Marriage Death	Compensatory mode and selective strongly linked to the stages of change and emotional traumas	favors the communal group activity
A.	Childhood Studies Work Travels Marriage Death	Adaptive mode and links marked by complicity siituational Links with family friends Friendship is a family value	friend and ritualized moments : New Year's Day / birthday / holiday
N.	War/Childhood (holocaust) Marriage Accident (vision loss)	Selective mode marked by mistrust, emotional distancing, sharing the difficulty, Classification of friends by period	search for help, support, taking account of disability

The study of intentions of people and their recurrent pattern of actions show that there is a proper relationship strategy for each.

V. CHARACTERISTICS OF THE STRATEGY OF RELATIONSHIP

This work from the interpretations of the actors in their friendly relations helped to highlight the dependence between the events of life and the appearance of new relationships. The person makes a classification and tells for example: "my friends from before." There is therefore a relational biography. Each phase change appears to activate a clean relationship strategy to each person and built on perennial patterns. This strategy is established on the basis of affective elements through values such as the value placed on friendship, beliefs and expectations. They direct the sympathies or antipathies. They provide data on the image that the person has of itself in an idealized form. The relational choices appear homogeneous, standardized, selective based on distinguishing elements of social status, age, etc. These emotional factors determine the level of commitment in the relationship ex. "When they suffer, I suffer." They compose an imaginary of the relationship: "I wanted to have news from him because it was my first love". The person creates, in this case, an ideal relational environment that is considered like a game. This is particularly noticeable when the person uses digital communication media. "I am in relation to a community of artists". The relationship strategy takes the form of ritualized interactions whose rhythm varies among individuals: annual exchange of holidays wishes, weekly meal, evening conversations on Skype, etc. They take singular forms appropriate to the needs of the person such as sharing of activities or friendly moments, or its values: "to give is my life ..." etc. The interactions are located in social contexts and spaces that have a symbolic value as: schools, companies, unions, associations of hikers, etc. They reinforce the value of belonging. These spaces increase connectivity levels cited by G. Simmel [14] and structural sociology (circles, networks). The aging is mentioned systematically in the relational biography. It concerns the disappearance of friendships: "of my generation has disappeared ...", "we were ten friends... we are no more than two". It expresses itself by the regret of the bonds of the past by comparison with those present: "Yes I have had really good friends, it was great ... today, it is different." It indicates a feeling of loneliness evoked on the mode of a battle to fight, "for not to be alone, you have to go out! “
The information collected shows activation of affects in the implementation of the social network and the importance of biographical and identity work that is developed. The experience through the implementation of the relationship strategy own to everyone appears as the essential element of identity’s recomposition.

The role of emotions in action and the impact on the self-image of phases of changes were the subject of many studies in the field of neurosciences and the social sciences more broadly. The psychiatrist L. Ciompi [15] states that pure thought does not exist. Cognition is defined "as part of the emotional logic". It defines the affect as "a psycho-physical condition, with quality, duration and varying degrees of consciousness" and insists on "the energy aspect of affects." For the neurobiologist A. Damasio [16], the affects enable an adaptive learning. The process triggered by the emotion at the level of the body and brain, with the manner whose appears the feeling and his awareness, create new emotional dispositions. He concludes that "emotions are part of bioregulators devices with which we are equipped to survive." Emotion is generated through our experiences. According to P. Ricoeur [17], it allows man to recompose his understanding of his environment through a phenomenology of desire. P. Ricoeur distinguishes a permanent form of personality, "selfhood", around which are developed changing identities. Researches of A. Damasio confirm this hypothesis and distinguish in the constitution of personality a "central self" of an "autobiographical self." P. Ricoeur emphasizes man's anxiety in front of these continual identity changes. He needs to reassure himself by searching for clues of permanence. According to P. Ricoeur, friendship is "a promise in time of a self-preservation."

VI. CONCLUSION

The theories and works highlighted reveal two crucial needs for the human being related to safety and self-awareness. These needs are expressed through communication with the external environment composed of

other people. This constitutes the means for a person to validate their existence through recognition and to strengthen their feeling of protection. The relation which takes shape in the form of mutual identical projections is transformed into attachment when the reliable feeling (implicit protection) evolves and strengthens. This reliable feeling comes along with the certainty of a common cultural belonging which expresses itself through shared activities and symbolic exchanges. This process consolidates the commitment in the relationship.

In this selective configuration, the relation contributes to strengthen a positive self-awareness based on the mutual identification. The aspects of identity are expressed through emotions and feelings. The self-awareness answers a desire of self-idealization through values which direct the emotions. The feelings are the mainspring of the composition of the relational environment because they play beforehand a role in the choice of selected people.

The relation evolves during the relational process between the people and is transformed. It constitutes a story made up of events of which the meeting with the person is a part, and is characterized by the importance of the moments and shared activities. From this point of view, the story of the relation can be the object of a biography characterized by the ascendancy of its emotional contents.

In conclusion, the relationship strategy composed of the elements highlighted in our study in progress shows the importance of the building of the relationship as a sensitive experience. It reveals a space essential to the individual and which seems to correspond to what F. Guattari [18] calls "weaving" about the friendship relation. D.W. Winnicott [19] considers that "the experience of the body" goes through a third space, next to the inner space and the emotional environment of the individual, which he called "cultic area". It offers a creative space game. The condition for the individual to invest in this space depends on his confidence level. The feeling of loneliness that we prefer to call "relational vulnerability" in the sense of insecurity and lack of assertiveness, could come from a relational model unsuitable strategy. Our work opens other perspectives. It seems that the question of the relation of elderly to digital communication tools can be studied through the experience felt. These are the tracks that remain to be explored.

REFERENCES

[1] E. Durkheim, “The division and labour in society”, PUF 1930, Paris 2012, p. 124, new collection « Quadrige ».

[2] S. Paugam, “The social bond”, PUF, Paris, 2013, pp. 7-103.

[3] F. Martin-Juchat, “Think the emotional body as a media, anthropology of affective communication”, Revue Le corps n°4, Paris, 2008, pp. 85-92.

[4] F. Bernard and R. V. Joule, “Link, meaning and action : Towards an engaging communication” Communication and organization, online, URL :

http://communicationorganisation.revues.org/2918, mars 2012 [retrieved : August, 2015].

[5] A. Honneth, “The struggle for recognition”, Folio essais (n° 576), Gallimard, Paris, 2013, pp. 114-235.

[6] N. Elias, “Society of individuals”, Fayard, Paris, 1991, pp. 208-301.

[7] D. W. Winnicott, “Capacity to be alone”, Payot-Rivages, Paris, 2012, pp. 46-108.

[8] V. Caradec, “Sociology of the elderly and aging”, Armand Colin, Paris, 2012, pp. 87-103.

[9] P. Pitaud., “Loneliness and isolation of the elderly”, Editions Erès, Toulouse, 2010, pp. 25-103.

[10] C. Bidart, “Studying networks, contributions and prospects for social sciences”, Informations sociales n° 147, 2008, pp. 100-259.

[11] M. Billé and J. Pollard, “Manifesto for age and life, reenchant old age”, Editions Erès, Paris, 2012, pp. 63-112.

[12] C. Delory-Momberger, “The biographical condition. Essai on self-narrative in the advanced modernity”, Teraèdre, Paris, 2010, pp. 61-75.

[13] A. Mucchielli, “Situation and Communication”, Ed. Ovadia : Nice, 2010.

[14] G. Simmel, “Studies on forms of socialization”, PUF, trad., 1999.

[15] L. Ciompi, “Emotions, affects, and affective logic, their places in our understanding of the world”, Picus, Juillet 2004.

[16] A. Damasio, “The feeling of what happens, body, emotions, conscience”, édition Odile Jacob,: Paris, 2002, pp. 110-299.

[17] P. Ricoeur, “Oneself as another”, édition du Seuil, Paris, 1990, pp. 73-197.

[18] F. Guattari, “Chaosmose”, Revue Chimères, N°77, éditions érès, Paris, 2012, pp. 17-107.

[19] D. W. Winnicott, “Playing and reality”, Gallimard 1971, traduction française 1975, Paris, pp. 17-107.

 ISBN: 978-1-61208-447-3

Automatic Emotion Detection in Social Media for on the fly Organizational Crisis Communication

Karolien Poels
Department of Communication Studies
University of Antwerp
Antwerp, Belgium
email: karolien.poels@uantwerpen.be

Veronique Hoste
LT3 Language and Translation Technology Team
Ghent University
Ghent, Belgium
email: veronique.hoste@ugent.be

***Abstract*—We present early stage research ideas on automatic emotion detection on social media (Twitter) during an organizational crisis and how this can be used for rapid and effective organizational crisis communication.**

Keywords—emotion detection; social media; natural language processing; organisational crisis; crisis communication.

I. INTRODUCTION

Many profit and non-profit organizations use social media like Facebook and Twitter for two-way communication with their stakeholders [1]. It is obvious that social media play an increasing role in organizational crisis communication [2] [3]. A crisis is "the perception of an unpredictable event that threatens important expectancies of stakeholders and can seriously impact an organization's performance and generate negative outcomes" [3]. Examples are product failures, or unethical practices. Effective crisis communication includes identifying stakeholders (i.e., a person or a group that is affected by or can affect an organization – e.g., customers, shareholders, journalists, the general public) and entering in discussion with them [3]. Social media present unique opportunities for crisis communication, such as the absence of journalistic gatekeepers, by which an organization can communicate its own, unbiased account [4]. Also, social media enable informing stakeholders more frequently and faster than traditional media. At the same time, stakeholders respond more directly, open and rapidly and do also expect an evenly fast response from organizations. It is thus utmost important that organizations know how their different stakeholders are feeling when a crisis is unfolding in order to respond efficiently and avoid (further) reputational harm.

In times of crises, stakeholders typically experience various emotions (e.g., anger, fear, worry, relief). Jin et al. [5] have recently classified publics' emotions in response to organizational crises depending on how (much) crisis responsibility is attributed to the organization. An important challenge is that, given the large share of comments posted by stakeholders during a crisis, it is impossible to manually monitor all these posts. A fine-grained automated detection method to tap into emotions communicated through social media on the fly is needed, but currently still underdeveloped.

Our goal is to study how emotions are experienced and expressed through social media in times of crises and how these emotions can be detected automatically with the ultimate aim to adapt corporate crisis response strategies in a rapid and efficient way. This research endeavor requires an interdisciplinary approach. There will be close cooperation between two disciplines: communication sciences (CS), integrating emotion theory and social media dynamics in crisis communication models, and natural language processing (NLP) for text mining of emotions in social media communication. We focus mainly on Twitter as one of the most important social media, especially for crisis communication [2].

In the Section 2 of this ideas paper, we discuss the two core objectives that need to be tackled to realize the future development of automatic emotion detection in organizational crises on social media.

II. RESEARCH OBJECTIVES

A. Objective 1: The Nature of Stakeholder Emotions in Social Media Crisis Communication

There is consensus that emotions expressed by the stakeholders during crises are a crucial component to consider in crisis communication [5]. However, there is reason to assume that some emotions will be more easily expressed through social media than others. Apart from some fragmented studies, a solid integration of emotion theory in social media crisis communication models is still lacking.

Several new steps need to be taken. Emotions are complex and multifaceted phenomena. Therefore, when studying how emotions are expressed through social media in crisis situations, it is crucial to first have an accurate view on *what* emotions are, *why* they are triggered in response to organizational crises, and subsequently *why, when* and *how* they are most likely to be expressed through social media. ***This critical evaluation of emotion theory and the dynamics***

 ISBN: 978-1-61208-447-3

of social media communication and Twitter more concretely is needed to build a conceptual model integrating emotions in social media crisis communication.

B. Objective 2: Automatic Detection of Crisis Related Emotions in Tweets

Until now, most work on subjectivity analysis in the domain of NLP has concentrated on discovering whether a review, tweet, sentence, or specific "object" (person, product, organization, etc.) is regarded in a positive or negative manner by a specific "source". This task has been given many names, from opinion mining, to sentiment analysis, review mining, attitude analysis, appraisal extraction, etc., and exists in different granularities ranging from coarse-grained sentiment analysis at the text level to fine-grained aspect-based sentiment analysis. From a methodological point of view, it is generally assumed that the semantic properties of words are good predictors of the semantic characteristics of the phrase or text that contain them. As a result, a lot of effort has been invested in the manual and (semi-)automatic development of lists of words indicative of sentiment [6] [7], and in the development of systems which both exploit these external lexicons and the lexical information present in the data (e.g., unigrams, bigrams, etc.) to determine for a given text, sentence or phrase whether it is positive, negative or neutral see [8] [9] for an overview. A similar methodology relying on external lexicons, e.g. [10], and automatically derived lexical knowledge has also been used in the more fine-grained approaches aiming at the automatic detection in text of emotions, in most cases restricted to a discrete set of emotions such as the 6 "universal" emotions distinguished by Ekman [11]: anger, disgust, fear, sadness, joy and surprise [12] [13] or the 12 emotions selected in the context of a shared task on emotion classification in suicide notes [14]. The basic Ekman emotions have also recently been investigated in tweets related to natural disasters such as Hurricanes [15].

Fine-grained emotion detection, however, remains a largely understudied research area in the NLP domain. Furthermore, there are no approaches modeling emotions related to organizational crisis in tweets more specifically. ***In order to move beyond superficial emotion modeling and to find a machine learnable operationalization of the types of emotions that are most diagnostic for an organizational crisis, an interdisciplinary approach is required.***

III. RESEARCH CHALLENGES

The integration of emotion theory in social media crisis communication models is crucial to the development of a supervised machine-learning model that can detect crisis-related emotions in tweets. Key to the success of these models is 1) to determine relevant crisis emotions and how they are expressed in social media, 2) to define the optimal set of text features (e.g., detecting words indicative of a given emotion, modeling modality, modeling sarcasm, etc.) for the detection of a given emotion. Important research questions to be investigated are: How well defined are the different emotions or do they represent a continuum? How learnable are the annotated emotions? How to include time series information? Should we weigh twitter users differently depending on their role as a stakeholder (e.g., affected customers, journalists, random twitter users)? How should we handle retweets? How generic are the learned models?, etc.

REFERENCES

[1] J. N. Sutton, "Social media monitoring and the democratic national convention: New tasks and emergent processes," J. of Hom. Sec. and Em. Man., 6(1), 2009, pp. 1-20.

[2] F. Schultz, S. Utz, and A. Göritz, "Is the medium the message? Perceptions of and reactions to crisis communication via twitter, blogs and traditional media," Public Rel. Rev., 37, 2011, pp. 20-27.

[3] T. W. Coombs, "Ongoing crisis communication: planning, managing and responding (4th edition)," Thousand Oaks, California: Sage publications, 2015.

[4] Y. Jin, B. F. Liu, and L. L. Austin, „Examining the role of social media in effective crisis management: the effects of crisis origin, information form, and source on public's crisis responses," Comm. Res., 2011, pp. 1-21.

[5] Y. Jin, B. F. Liu, D. Anagondahalli, and L. L. Austin, "Scale development for measuring publics' emotions in organizational crises," Public Relations Review, 40, 2014, pp. 509-518,.

[6] J. Wiebe, M. Bruce, F. Rebecca, and T. O'Hara, "Development and use of a gold standard data set for subjectivity classifications," Proc. of ACL-99, 1999, pp. 246-253.

[7] V. Jijkoun and K. Hofmann, "Generating a non-English subjectivity lexicon: relations that matter," Proc. of EACL-2009, 2009, pp. 398-405.

[8] A. Balahur, "Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types", PhD thesis, Univ. of Alicante, unpublished, 2011.

[9] B. Liu, "Sentiment Analysis and Opinion Mining," Morgan & Claypool Publishers, 2012.

[10] J. Staiano and M. Guerini, "DepecheMood: a Lexicon for Emotion Analysis from Crowd-Annotated News," Proc. of ACL-2014, 2014, pp. 427-433.

[11] P. Ekman, "An argument for basic emotions," Cognition and emotion, 6(3/4), 1992, pp.169-200.

[12] C. Strapparava and R. Mihalcea, SemEval-2007 Task 14: Affective Text. Proc. of SemEval-2007, 2007, pp.70–74.

[13] J. Bellegarda, "Emotion Analysis Uting Latent Affective Folding and Embedding," Proc. of the NAACL HLT 2010 Worksh. on Computational Approaches to Analysis and Generation of Emotion in Text, 2007, pp. 1-9.

[14] B. Desmet and V. Hoste, "Emotion Detection in Suicide Notes," Expert Systems with Applications, 40 (16), 2004, pp. 6351-6358.

[15] J. Brynielsson, F. Johansson, and A. Westling, A, "Learning to Classify Emotional Content in Crisis-Related Tweets," ISI 2013.

 ISBN: 978-1-61208-447-3