Article:
Thakker, D, Karanasios, S, Blanchard, E et al. (2 more authors) (2017) Ontology for Cultural Variations in Interpersonal Communication: Building on Theoretical Models and Crowdsourced Knowledge. Journal of the Association for Information Science and Technology, 68 (6). pp. 1411-1428. ISSN 2330-1635

https://doi.org/10.1002/asi.23824

Ontology for Cultural Variations in Interpersonal Communication: Building on Theoretical Models and Crowdsourced Knowledge

Dr Dhaval Thakkar: School of Electrical Engineering and Computer Science, Bradford University. Richmond Rd, Bradford BD7 1DP, United Kingdom. d.thakker@bradford.ac.uk

Dr Stan Karanasios*: Business IT and Logistics, RMIT University. 445 Swanston St Melbourne, 3000, Australia. stan.karanasios@rmit.edu.au

Dr Emmanuel Blanchard: IDÛ Interactive Inc. 4385 Fabre Street Montreal QC, Quebec H2J3V2 Canada. ebl@idu-interactive.com

Dr Lydia Lau: School of Computing, University of Leeds. Faculty of Engineering, Leeds LS2 9JT, United Kingdom. l.m.s.lau@leeds.ac.uk

Dr Vania Dimitrova: School of Computing, University of Leeds. Faculty of Engineering, Leeds LS2 9JT, United Kingdom. v.g.dimitrova@leeds.ac.uk

*corresponding author

ABSTRACT

The domain of cultural variations in interpersonal communication is becoming increasingly important in various areas, including human-human interaction (e.g. business settings) and human-computer interaction (e.g. during simulations, or with social robots). User generated content (UGC) in social media can provide an invaluable source of culturally diverse viewpoints for supporting the understanding of cultural variations. However, discovering and organizing UGC is notoriously challenging and laborious for humans, especially in ill-defined domains such as culture. This calls for computational approaches to automate the UGC sensemaking process by using tagging, linking and exploring. Semantic technologies allow automated structuring and qualitative analysis of UGC, but are dependent on the availability of an ontology representing the main concepts in a specific domain. For the domain of cultural variations in interpersonal communication, no ontological model exists. This paper presents the first such ontological model, called AMOn+, which defines cultural variations and enables tagging culture-related mentions in textual content. AMOn+ is designed based on a novel interdisciplinary approach that combines theoretical models of culture with crowdsourced knowledge (DBpedia). An evaluation of AMOn+ demonstrated its fitness-for-purpose regarding domain coverage for annotating culture-related concepts mentioned in text corpora. This ontology can underpin computational models for making sense of UGC.

Keywords: Ontology, knowledge engineering, culture, crowdsourced knowledge, semantic tagging

Introduction

Awareness of the role of cultural variations in interpersonal communication is critical in the 21st century, characterized by rising globalization and human mobility. A large body of research has demonstrated the influences of cultural variations on interpersonal communication and the effect on the behavior of individuals and groups (Hofstede, 1991). The domain of interpersonal

communication and culture falls within the framing of 'ill-defined' because such interactions are disposed to variations and acceptability (Henrich et al., 2010; Mesquita et al., 1997). For example, different interpretations of non-verbal behavior and emotions, amongst others, have been observed in different cultures (Matsumoto & Hwang, 2013). Studies of cultural variations face the challenge of obtaining authentic and rich data which captures the complexity and breadth of culture (Jones & Alony, 2007). To address this issue, user generated content (UGC) on social media can be exploited as it offers multiple and divergent perspectives on cultural variations. That is, rather than arguing that there are 'correct' or 'incorrect' cultural variations, UGC provides a vast, authentic, rich and evolving corpus of data for both social scientists and computer scientists to understand, learn about and model cultural variations in new ways.

UGC, such as online articles on how to perform a task, YouTube videos suggesting how to behave in certain settings together with the comments viewers have made, blogs describing a blogger's cultural encounters, and so on; are increasingly perceived as a source of "collective knowledge" (Thomas & Sheth, 2011). The ubiquity of social media means that UGC reflects everyday human activities capturing the views of a broad range of people from various cultural backgrounds.

While UGC can provide a rich source for deeper understanding of cultural variations that can be useful in a range of interpersonal communication contexts e.g. businesses (dealing with customers in different countries), leisure (interacting with people from various nationalities), study (advising students from different cultures), or with social robots (autonomous systems that communicate with humans); this potential has not been exploited to date.

To illustrate, let us consider a business training scenario. Ian is a new business development manager at a British firm that has recently developed business interests in Brazil. Ian understands that interpersonal communication and cultural variations play a crucial role in business dealings with the clients from Brazil. For instance, when taking clients to dinner for the first time, Ian would need to understand the cultural nuances and differences between British and Brazilian cultures. Being a savvy social media user, Ian is accustomed to checking for tips shared by others, e.g. he regularly

checks blogs, YouTube videos and comments, and Wikipedia pages. He could review these sources to see others' experiences and gain an understanding of the Brazilian culture. However, finding relevant information in UGC is notoriously difficult and time consuming. Ian may be fortunate and may find a valuable piece of content quickly, but it is most likely that he may encounter various pieces that are disconnected, irrelevant and difficult to relate to his needs. In other words, he would have to invest significant time to make sense of UGC (i.e., creating meaningful associations between the content and drawing useful inferences) in order to discover relevant content.

To facilitate this complex sensemaking process, we propose a computational approach where various pieces of textual UGC (comments, blog posts, Wikipedia pages) are tagged with concepts referring to culture-related aspects (e.g. gestures) and organized in a coherent way that allows user exploration of the content through an interactive tool. This is enabled by a computational model, in the form of an ontology, which defines main concepts and relationships in the domain of cultural variations in interpersonal communication. Such a model, including the methodology for its creation and an illustrating application, is presented in this paper.

Computational models and tools are emerging in the fields of data mining and semantic web in order to assist people to make sense of UGC (Leginus et al., 2015; Syn & Spring, 2013). Data mining theories and tools (Fayyad & Piatetsky-Shapiro, 1996) exploit the volumes of data to derive interesting patterns; for instance, sentiment analysis (Nasukawa & Yi, 2003), also called opinion mining, is used to analyze people's views towards entities such as products, organizations or news events. However, it has been argued that these quantitative approaches may be problematic due to the shallow understanding they afford; and hence, complementary qualitative methods could be exploited to fully understand the meaning embodied in UGC (Thelwall, 2006). A possible way to combine quantitative and qualitative models for more meaningful text analytics is to use semantic technologies underpinned by ontological models. The areas related to semantic technologies influencing the approach in this paper include semantic tagging, ontology and linked data.

Semantics techniques have been used to link social tags to a controlled vocabulary (Yi, 2010) and deploy tag analysis to uncover semantic relations (Yoon, 2012). A prominent theme is semantic tagging for aggregation and analysis. Semantic tagging (also called “augmentation” or “annotation”) is a process of attaching semantics in the form of "ontology concepts" to a selected part of a text (e.g. a YouTube comment) to assist automatic interpretation of the meaning conveyed by the text. It is increasingly adopted as the main mechanism to aggregate and organize Web content by linking it with ontology concepts to provide meaning. Semantic units (i.e. ontology concepts) are extracted mainly by identifying named entities (such as people, organizations and locations). This has become a powerful way of organizing, browsing and publicly sharing personal collections of resources on the Web (Berners-Lee et al., 2001; Kim et al., 2010). Semantic tagging is a type of Named Entity Recognition (NER) technique. NER is the process of locating a word or phrase that references a particular entity within a text. The NER task first appeared in the Sixth Message Understanding Conference (Nadeau & Sekine, 2007; Sundheim & Chinchor, 1995) and involved recognition of entity names (people and organizations), place names, temporal expressions and numerical expressions. This led to a variety of processing approaches such as supervised (learning model by looking at annotated examples) (Witten & Frank, 2005) unsupervised (using extraction patterns) (Etzioni et al., 2005; Munro & Manning, 2012) and semi-supervised (using labelled and unlabeled corpuses to create model) (Etzioni et al., 2005).

Ontologies that define the main concepts and relationships in the domains of investigation underpin the above semantic processing approaches. These approaches involve processing of text to extract particular types of information (information extraction) related to a domain. This information is then connected with entities and properties from one or more ontologies which represent knowledge about the domain (Wimalasuriya & Dou, 2010).

Linked data, an area of research and application with tremendous recent growth, has contributed to the uptake of semantic tagging tools and processes (Bontcheva & Rout, 2012). The term “linked data” refers to “*a set of best practices for publishing and connecting structured data on the Web*”

(Bizer et al., 2009). These best practices have been adopted by an increasing number of organizations, leading to the creation of a global knowledge space containing billions of facts and concepts in a variety of domains (Heath & Bizer, 2011). However, even in this gradually expanding global knowledge space, there are no ontological models that represent cultural variations. This hinders the adoption of semantic approaches for qualitative analysis of UGC for understanding variation of communication due to cultural factors.

This paper argues that there is an opportunity to address this gap, which will enable a step change in gaining insights into UGC that captured views of culturally-diverse users. The following questions framed our research:

RQ1: How to develop an ontology of cultural variations in interpersonal communication?

RQ2: How well does such an ontology support tagging of UGC containing cultural variations in interpersonal communication?

To answer our research questions, a hybrid approach was developed. It combined social science models for understanding culture together with computer science approaches for knowledge engineering. This combination was necessary to bring together the understanding of the 'messy' and fluid nature of culture, as well as to address the need of formulating logical constructs for ontological representations. An automated process was devised to extend the ontology by extracting culture-related facts from DBpedia - a widely used multi-purpose, crowdsourced knowledge base using linked data (Auer et al., 2007). The resultant ontology was termed "Extended **A**ctivity **M**odel **On**tology" (hereafter, referred to as AMOn+). An evaluation was conducted to investigate the effectiveness of AMOn+ for automatic tagging of UGC on cultural variation related to interpersonal communication.

Culture has been an area of abundant academic interest across a range of fields. In the information related fields, the most common focus of enquiry has been on information behavior related to culture, cultural content and the dimensions of the user culture on technology use (Leidner & Kayworth, 2006; Nicholas et al., 2013). For instance, Europeana (http://europeana.eu), a digital

library project (Purday, 2009), enables search and discovery of more than 17 million items by collecting metadata from approximately 1,500 cultural institutions (data providers) across Europe. Europeana utilizes linked data technologies to allow data providers to opt for their data to become linked data and converts their metadata to EDM (Europeana Data Model), hence benefiting from pooling semantically related resources on the Web via Europeana (Isaac & Haslhofer, 2013). Our work differs from these studies as we are concerned with using UGC to better understand cultural variations in interpersonal communication. Hence, the prime contribution of this paper is to illustrate the role of ontology to empower semantic analysis of a vast amount of UGC related to cultural variations.

The remainder of this paper is structured as follows. The next section reviews and synthesizes the literature in the fields of information and computer science on the influence of culture on interpretation of content, adaptation using cultural background and the existing approaches to model culture and interpersonal communication. Following this we present our ontology development approach utilizing theoretical models and linked data. We then provide details of the two-phase ontology development process and presents an evaluation study of AMOn+ for tagging UGC on interpersonal communication. We then illustrate an application of AMOn+. The paper concludes by outlining the main contributions.

Relevant Work

Examining Cultural Variations in Social Media: The Need for an Ontological Model

This section surveys the literature on cultural variation in social media and highlights the need for an ontological model. Research in social media analytics has examined cultural variations in user behavior

In the special issue on "semantics of people and culture" in the International Journal on Semantic Web and Information Systems, Liu and Maes (2007) highlighted the importance of culture in the context of community-produced semantics, such as tagging of content. It is noted that the quality

and legibility of community-produced semantics are varied, and are greatly affected by factors such as tastes, personality and culture. Liu and Maes (2007) stressed the need for modeling culture for interpreting and qualifying knowledge presented by different community participants. It made a strong argument pointing to the pressing need for computer-processable models of culture (which in our case is in the form of an ontology) that capture the shared common sense and sensibilities within 'cultural modules'[1] (Stuckenschmidt et al., 2009). It was envisaged that such a model could afford improvisational manipulations of the tags and ratings corpora - such as normalizing away subjectivity, or translating tags, assertions and ratings from one cultural context to another.

Research to date has highlighted that an individuals' culture can play a crucial role in annotation of content and subsequently content interpretation. However, there has not been analysis of automated tagging of culture-related aspects in UGC. For this, a crucial component is missing - a computer-processable model that can aid content annotation and interpretation. Our work is the first attempt at building such a computer-processable model for cultural variations in interpersonal communication. The AMOn+ ontology is implemented with the state-of-the-art tools such as ROO (Denaux et al., 2011) and Protégé (Knublauch et al., 2004) and published with well-established W3C recommendations such as Web Ontology Language (OWL) (McGuinness & Van Harmelen, 2004). Encoding of ontologies using OWL allows knowledge engineers to create extensible ontologies that are easy to integrate with different applications, primarily due to its XML-like syntax, and web-centric approach.

Modeling Interpersonal Communication and Cultural Variations

There is a limited amount of work on modeling cultural variations and interpersonal communication. Authors et al., (2010) represent an Upper Ontology of Culture (UOC), a formal conceptualization of the culture domain by identifying the common backbone of culture-related disciplines and

[1] Module refers to a part of ontology model that is conceptually grouped together.

guidelines. Concepts from three domains are combined into UOC. Firstly, the 'cognitive domain' as cultural experiences are strongly mind related (Nisbett & Norenzayan, 2002). Secondly, the 'affective domain' since cultural and affective experiences are strongly intertwined (Mesquita et al., 1997). Finally, the 'context', as cultural experiences must be understood as being context-sensitive in order to be correctly address. To place AMOn+ in the context of this existing work on modeling culture domain, it is important to differentiate between the perceptions of formal and semantic ontologies, also sometimes referred to as lightweight (semantic) versus heavyweight (formal) ontologies (Mizoguchi, 2003). Despite sharing the term "ontology", both approaches follow different construction principles and objectives, which consequently lead to very different products. Semantic ontology adopts a pragmatic and application-oriented approach with the main focus on rapid operational capability. On the other hand, formal ontology engineering follows a near-philosophical approach. It tries to obtain a realistic (i.e. non-deformed or simplified) abstract representation of the reality by being application and discipline independent. In this context, we consider AMOn+ as a lightweight ontology which can be utilized in different applications. Specifically, its frame of reference is cultural variations in the context of interpersonal communication. In this respect, AMOn+ is the first semantic ontology designed for representing cultural variations in interpersonal communication domain. The ontology produced and the socio-technical approach we have followed is an important first step to build a computational model of cultural variations. This work opens up possibilities on which we and others can build. Our work also shows that it is possible to extract valuable culturally-relevant facts from a crowdsourced knowledge base (DBpedia) – opening up a new avenue for research in combining theoretical models and crowdsourced knowledge.

Broader Application: Culturally-aware Intelligent Systems

A computer-processable model of cultural variations in interpersonal communication has broad applications. The prime purpose of our ontology is to support semantic processing of UGC, and this is the main area where we have evaluated AMON+. Culturally adaptive user interfaces are another area of application for an ontological model of culture. Adaptation of systems based on a user's

cultural characteristics is an area that is gaining interest. The central technique is the extraction of cultural profile of the user (Reinecke & Bernstein, 2013) in order to guide the adaptation processes. Such systems aim to mimic culturally-intelligent people when interacting with people from different cultures than their own (Earley & Mosakowski, 2004). Reinecke et al., (2010) present a list of aspects that influence cultural background. Aspects of culture that impact system interface preferences are extracted from literature in areas such as cultural anthropology, cognitive science, and human–computer interaction. A knowledge base was developed which formed the basis of a user model represented in web ontology language (OWL) (McGuinness & Van Harmelen, 2004). A rule base system is then used to transform a user's cultural model into a personalized user interface (UI) automatically (Reinecke & Bernstein, 2013). In order to reduce the time needed for the initial information acquisition for the user model, a small number of initial questions have been shown sufficient to predict user preferences and provide a suitable first adaptation of the UI. Hofstede's (1991) cultural dimensions have been a useful source of concepts for representation. For example, Marcus and Gould (2000) and Dormann and Chisalita (2002) apply different Hofstede's dimensions to map cultural characteristics to certain aspects of UIs, and these dimensions have been proven useful for predictive purposes (Reinecke & Bernstein, 2008). Many of these researchers base their work on the fact that, to some extent, design preferences of different cultural groups are generalizable for the people within one group (Ford & Gelderblom, 2003; Sheppard & Scholtz, 1999). Other research has demonstrated that people within the same cultural group even show similar navigation and search behavior (Kralisch & Berendt, 2004). One of the important lessons from these works is the consideration of "groups" while building a model of culture. In our work, we consider "cultural groups" for scoping the culture model.

Further implication of the existing work is that it is important to have a reliable model for describing the aspects where cultural variations may occur. This can enable a broader application, especially in systems that adapt their interaction to the cultural background of their users. For example, situational simulators for learning can generate situations where the user is required to take into

consideration aspects which may be prone to cultural variations (e.g. greetings); recognizing such aspects is important in many modern contexts (e.g. medicine, business, military). An ontological model that shows which communication aspects may have cultural variations can inform the design of social robots (e.g. virtual companions can be tailored to interact with people from different cultures). Ontological models such as AMOn+ provide a starting point, and an illustrative example, for knowledge underpinning future culturally-aware user adaptive systems.

Ontology Development Utilizing Theoretical Models and Linked Data

The foregoing discussion highlighted three key arguments: Firstly, culture is an emerging area of research in the computer science community. Secondly, deep understanding of cultural variation in interpersonal communication has many potential applications. Thirdly, semantic tagging can be used to make sense of such cultural variations in the collected UGC. Finally, a relevant semantic knowledge base, in the form of an ontology, is required to inform the automated semantic annotation process when there is a huge volume of UGC. Following these arguments, we have developed an ontological model (AMOn+) for cultural variation with concepts that indicate aspects related to interpersonal communication.

A two-phase hybrid approach was used for the development of the required ontology. A distinctive characteristic of our approach is the combination of theoretical and empirical work: theoretical in terms of grounding our models on activity theory as well as more expansive cultural theories; and, empirical in terms of using a popular crowdsourced knowledge base from Linked Data (i.e. DBpedia (Auer et al., 2007), for instantiation). For the first phase, we followed the METHONTOLOGY (López et al., 1999) methodology and utilized theories on culture and interpersonal communication to build the core ontology. For the second phase, we extended the core ontology to provide a more concrete conceptualization of cultural related instances. Concrete cultural variations and nuances are necessary in order to semantically annotate UGC on its relevance to cultural aspects in interpersonal

communication. DBpedia (Auer et al., 2007) provided suitable source for this purpose. The following sections provide more detail on each phase.

Phase 1: Building the Core AMOn+ Ontology

The process of building the core AMOn+ ontology started with the identification of a base structure of the ontology. This included a list of widely accepted concepts and their definitions along with supporting references and storing them as part of a glossary of terms (López et al., 1999). The glossary of terms gathers potentially useful domain knowledge and its meaning, following the conceptualization phase in the METHONTOLOGY (López et al., 1999). To enable reuse, the core AMOn+ ontology was developed in a modular fashion by separating several abstraction layers. The modular design allows each of the modules to be used or extended independently and does not necessitate an application to use the whole ontology.

Activity Layer

The base layer of the ontology, called "Activity layer", includes the core concepts from activity theory. Activity theory provides a conceptual toolkit for framing and analyzing object-oriented human activity – for this work, interpersonal communication. Engeström's (1999) third-generation activity theory was used, which is largely based on the notion of activity systems and the expansion of the notion of activity to include the community, rules, and norms, and division of labor as key elements (in addition to subject, tools and object). See Authors (2013) for an overview of how activity theory informs ontology development and Authors (2011) for explanation and discussion of activity theory in information studies. The similarity between the activity system structure and the ontological structure (i.e. "concepts" and "relationships between concepts") bridges the transition between the human based and the machine based models. Demonstrating its usefulness in ontology development, Kuhn (2001) used some of the hierarchical concepts available from activity theory, namely activity-actions-operations, to structure an ontology. O'Leary (2010) used activity theory as a template to capture organizational activity (an "enterprise ontology"). We argued that (1) this socio-

technical approach is useful for developing more insightful conceptual models of ill-defined activities; (2) that this theoretical scaffolding can be used to inform the development of an ontology in ill-defined domains; and, (3) that this ontology can be used to guide the semantic tagging of digital traces for making sense of phenomena (Authors, 2013). However, one of the noted areas where the initial ontology was lacking was in the domain of culture. While activity theory provides a structure for activity, it is beyond its exposition to provide concepts and a vocabulary on detailed aspects of culture. In this work we engage the concepts of activity theory useful for modeling, rather its broader theoretical contributions.

When populating the glossary of terms, concepts from the activity model were used as the starting point. These concepts were converted into an upper layer ontology (see Figure 1). It includes the base concepts provided by the activity theory such as: “Activity”, “Tools”, “Action”, “Operation”, “Motivation”, “Outcome”, “Subject” and “Rules and Norms”. Relationships amongst these concepts are also defined. For example, the concept of “Tool” is defined as “something that is used by Subject in an Activity to achieve an Object”. This layer is considered the ‘root model’ of human activity and the foundation of our understanding of interpersonal communication and culture related activities.

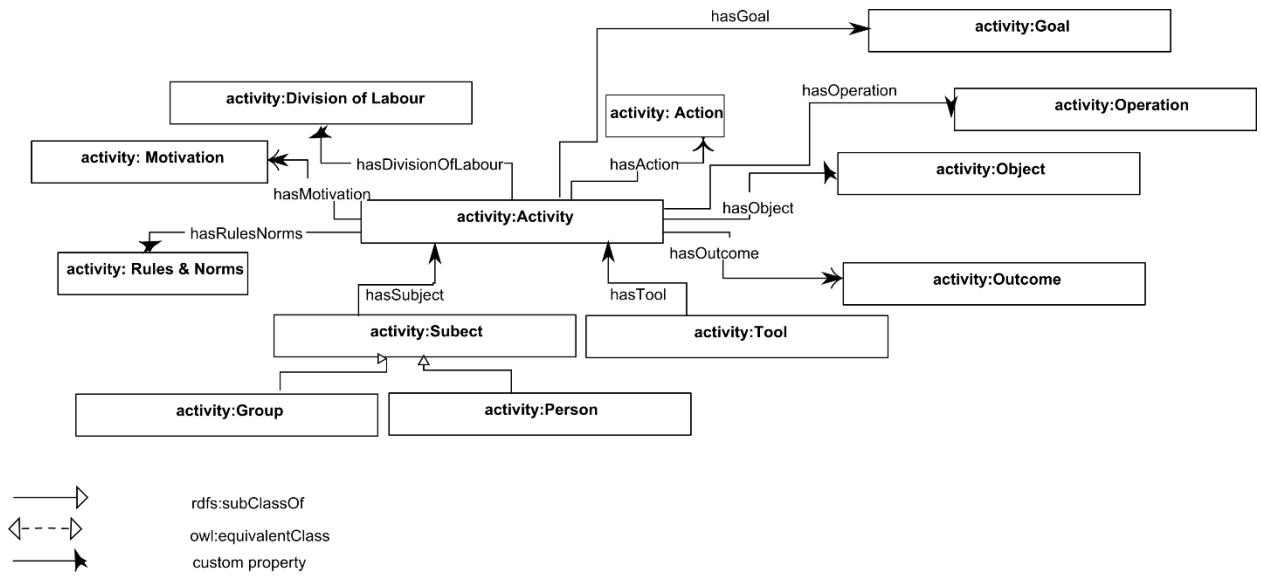

Figure 1: UML representation of the AMOn+ Activity Layer (activity: denotes activity layer concepts)

Interpersonal Communication Layer

One of the starting points for extending the activity model for interpersonal communication activity was to expand the concept of "Tool" in two directions (see Figure 2). Firstly, we expanded "Mental Tools", described by as "tools used by subject of activity for communication and representation" (Authors, 2011; Leiman, 1999). Secondly, we expanded "Physical Tools", described as "*any material objects that has been modified by human beings as a means of regulating their interactions with the world and each other*" (Cole, 1999, p. 90).

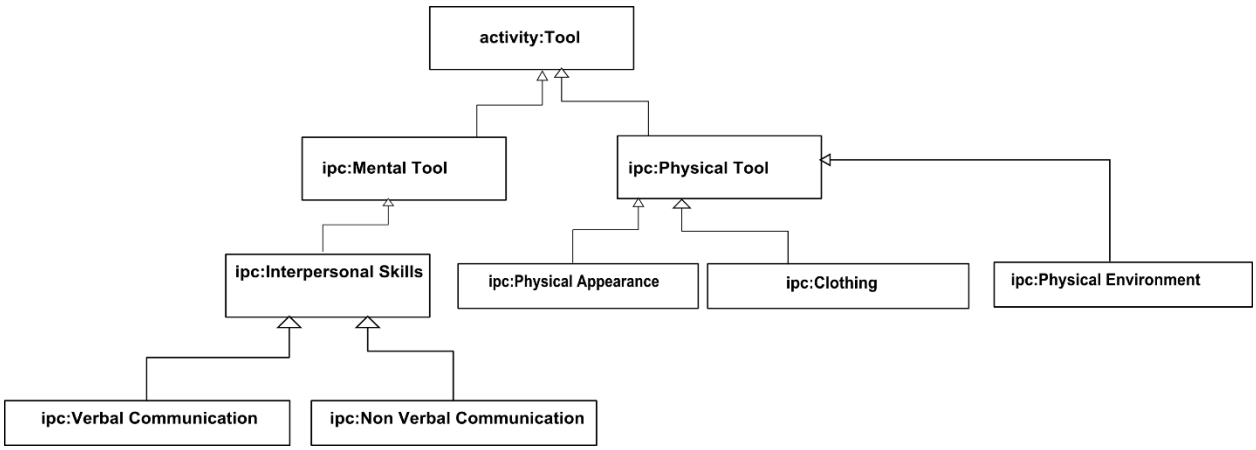

Figure 2: Expansion of the concept "Tool" from the AMOn+ Activity Layer (ipc: denotes interpersonal communication layer concepts)

These two types of tools were then further expanded in the context of interpersonal communication. For example, "Interpersonal Skill" was conceptualized as a category of "Mental Tools" while "Clothing", "Physical appearance", and "Physical Environment" as sub-categories of "Physical Tools". "Interpersonal Skills", which are "mental and communicative algorithms applied during social communications and interaction to reach certain effects or results", was further divided into two classes: "Verbal Communication" and "Non-verbal Communication" (see Figure 2). As the focus of this work is on Non-verbal Communication, it was further expanded into: "Body Language Act", "Paralanguage", "Haptic", "Chronemic", and "Proxemic" (see Figure 3). The concept of "Body Language" was further expanded with non-verbal communication concepts that are related to

movements. Theoretical frameworks provided classification in the form of "Agreement/Disagreement Acts", "Conversation Adaptor", "Behavior Primitives", "Greeting Act", "Conversation Regulator", "Disengagement" and "Demonstrate Affection". (Glossary of these terms is available from http://imash.leeds.ac.uk/ontologies/amon/dataset/Glossary.pdf).

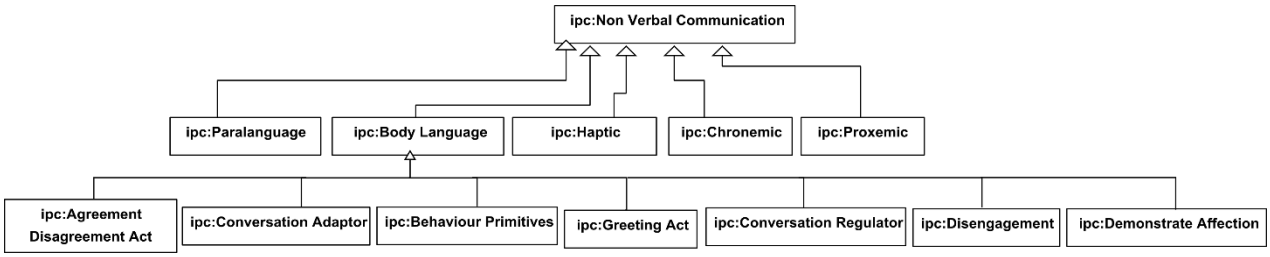

Figure 3: Concepts relating to "Non-verbal Communication"

Further classification was conceptualized for the subclasses of "Body Language" such as "Behavior Primitives" and "Greeting Act". "Posture", "Eye Gaze", "Facial Expression" and "Gestures" were identified as "Behavior Primitives". Various types of greetings were identified with variations in the way greetings are expressed, such as "Greeting with Handshake", "Greeting with Head Nodding", "Greeting with Hug" and "Greeting with Kiss" (see Figure 4).

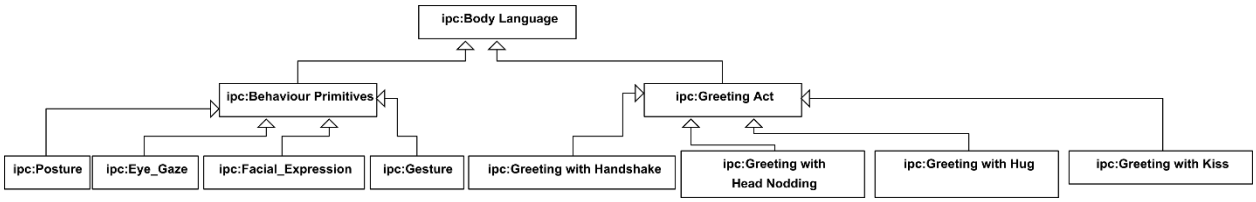

Figure 4: Further classification of the concept of "Body Language"

Cultural Variations Layer

The theoretical foundations that form the basis for grounding our ontology does not rely on any one single theory of culture. Rather a pluralistic approach was adopted that drew on a range of prominent theories; focusing on different aspects such as cognitive and value systems, and contexts such as business settings (the essential theories are listed in Table 1). Such an approach was needed because few theories of culture focus specifically on the context of interpersonal interaction and

most theories focus on the high-level abstractions. Some of the theories further propose strategies to address the risk of relying on cultural stereotypes, another central challenge in cultural research.

Table 1: Theoretical groundings utilized for cultural variations layer

Main Theory/Reference	**Source**
Memetic Theory	(Dawkins, 2006)
Dual Inheritance Theory	(Henrich & McElreath, 2007)
Sperber's Epidemiology of Representation	(Sperber, 1996)
Distribution of Cultural Conceptualizations	(Scharifian, 2003)
Culture and Cognition	(Nisbett & Norenzayan, 2002)
System of Values of Hofstede	(Hofstede, 1991)
GLOBE System of Values	(House et al., 2004)
Schwartz Value Inventory	(Schwartz, 1994)
Cultural Intelligence	(Earley & Mosakowski, 2004)
Cultural Framework of Alwood	(Allwood, 1985)
Framework for Intercultural Training	(Bennett, 1983)
Research on Specific Cultural Variations	(Matsumoto & Hwang, 2013)
Cultural Framework of Hall	(Hall, 1983)
Politeness Theory	(Brown & Levinson, 1987)

Based on syntheses of the theoretical knowledge relevant to our work we extracted several core concepts which formed the basis of the cultural variations layer of our ontology (see Figure 5). In doing so, we had to compromise between the fluid, and often elaborate, terminology used in social theory with the bounded and precise language needed for ontology development. This is the challenge of transposing real-world phenomenon into computational form (Authors et al., 2013). "Culture" is seen as a cognitive phenomenon that emerges at the group level. A "Cultural Group" is a coherent and stable ensemble of individuals, "Enculturated Individual", to which a culture can be associated. Each individual can be a member of several cultural groups. Consequently, each enculturated individual is associated with a specific set of socio-cultural influences. Dual Inheritance Theory provided concept of "Cultural Element" linked directly with the concept of Culture. "Descriptor" is a concept that helps in characterizing the nature of something related to culture in the form of quality, property, condition, function, or situation.

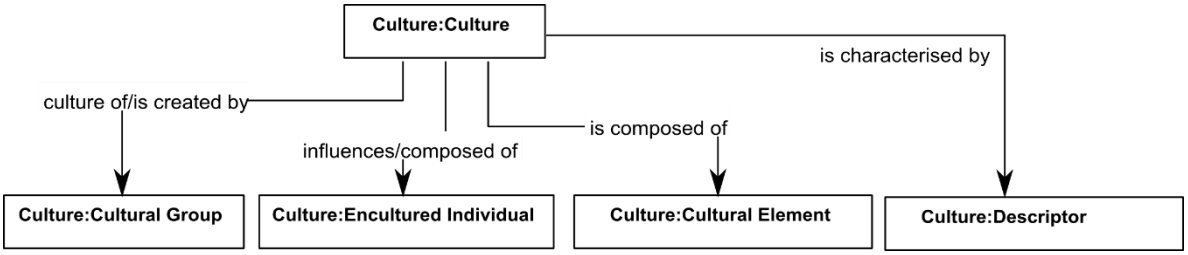

Figure 5: Core concepts related to culture forming basis for cultural variations layer

The cultural theories informed cultural variations and their linkage to the interpersonal communication layer. Our conceptualization was based on expanding base concepts from the activity layer for Interpersonal Communication Activity first and then identifying those concepts from the interpersonal communication activity that have cultural variations. Figure 6 illustrates how the core cultural concepts - "Cultural Group", "Enculturated Individual", "Cultural Element" and "Descriptor" - are linked to the concepts from the Activity Layer and the Interpersonal Communication Layer.

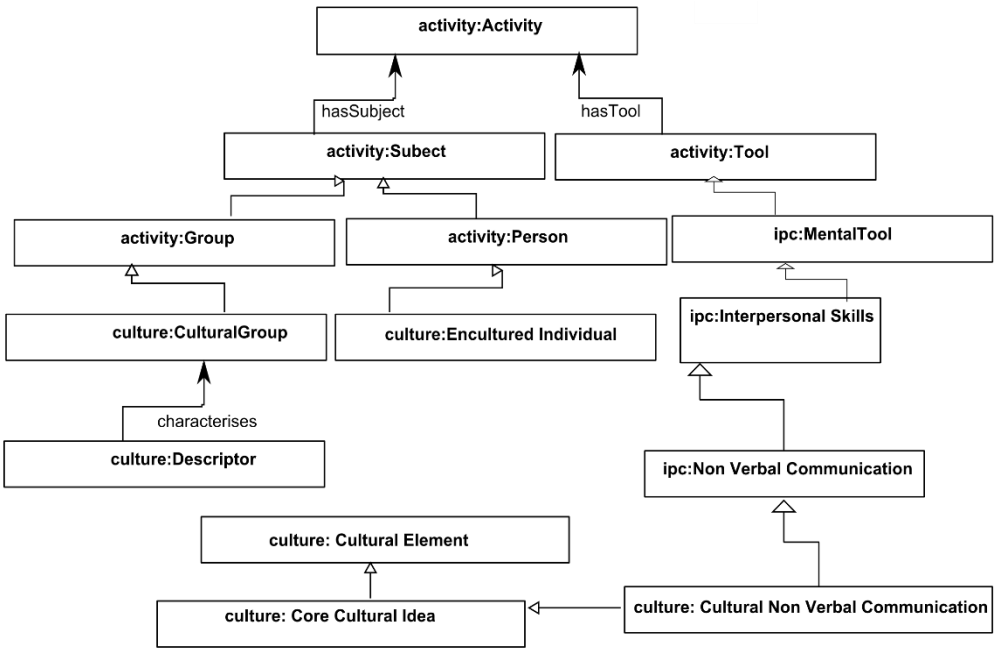

Figure 6: Expanding on concepts from the Activity Layer (denoted with activity:) and Interpersonal Communication Layer (denoted with ipc:) to form the Cultural Variations Layer (denoted with culture:).

"Cultural Group" is an essential concept that is further classified with the help of the literature (see Figure 7). In particular, Cultural Group can be defined on the basis of one or more different shared

criteria that 'unite' individuals, such as nationality, religion, occupation, geographical localization, shared interest, and ethnicity (Castano et al., 2003; Jost & Hamilton, 2005).

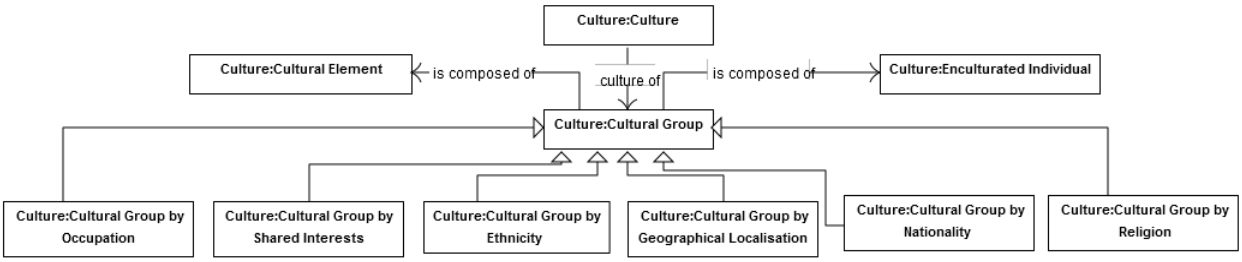

Figure 7: Expanding the concept of "Cultural Group"

The concept "Cultural Element" was further classified in "Core Cultural Idea", as a specific type of cultural conceptualization that refers to values, norms, beliefs, stereotypes, expected situational and social structures, etc.; in other words, ideas that are likely to be endorsed by a large portion of a cultural group (see Figure 8). This conceptualization resulted in further classification of "Core Cultural Idea" in: "Socio Cultural Script", "Beliefs", "Values", "Stereotype", "Social Structure", "Socio Cultural Norm" and "Cultural Non-Verbal Communication". In particular, the rich hierarchy of the concept of "Non-Verbal Communication" from the interpersonal communication formalization was categorized as having cultural variation.

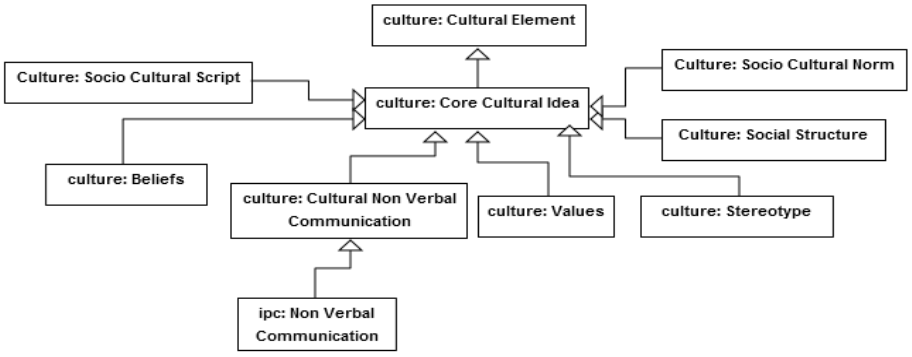

Figure 8: Expanding the concept of "Core Cultural Idea"

Finally, to include an easy mechanism for capturing new context, the concept of "Descriptor" was created. It was identified as quality, property, condition, function, or situation to characterize the nature of "Culture", "Cultural Element" or "Cultural Group". Various rich descriptors were identified

under the concept of “Descriptor”, for example “Gender”, “Eating Norms”, “Paying Norms”, “Hygiene”, “Directness”, “Privacy”, “Indirectness”, “Individualism”, “Collectivism”, “Affective Display”, “Artefact Usage”, “Hosting”, “Gift”, “Social Settings” and “Role” (see Figure 9). (Glossary of these terms is available from http://purl.org/amon#Axioms).

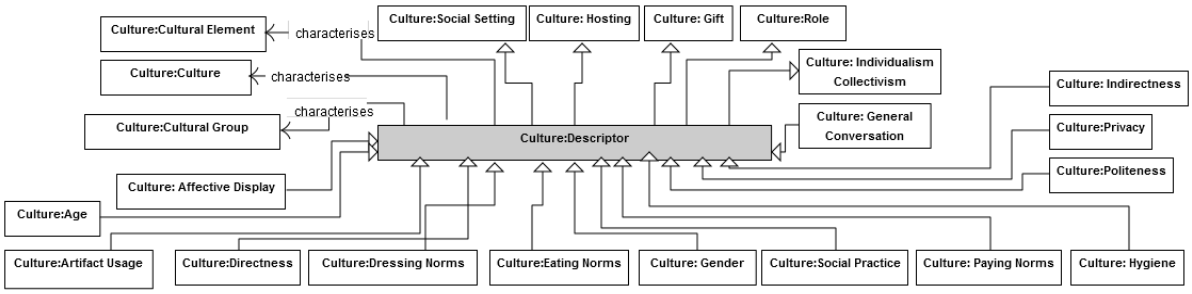

Figure 9: Expanding the concept of “Descriptor”

Core Ontology Implementation

The process of transformation from conceptual form (glossary of terms) to logical form (computer-processable model) involved knowledge engineers. Concepts and relationships identified in the glossary of terms are transposed into a Controlled Natural Language (CNL) using an appropriate authoring tool (cf. Bao et al., 2009). ROO (cf. Denaux et al., 2011) was used to translate the CNL constructs into their corresponding Description Logic (DL) statements (cf. Baader & Nutt, 2003) that enable machine interpretation and reasoning. Table 2 presents a summary of the ontological features of the core ontology implemented in terms of size (number of classes, properties and instances/objects), expressivity (conforming to SHOIQ in descriptive logic (Horrocks & Sattler, 2005)) and complexity of the core knowledge (captured by axioms).

Table 2: AMOn+ features

Feature	Value
No of Classes	125
No of Properties	49
No of Individuals	62
No of Axioms	1148
DL Expressivity	ALCHIQ

Phase 2: Extension of the Core AMOn+ Ontology with DBpedia

It is crucial for semantic tagging that the ontology covers concepts in sufficient details, including concrete instantiations of the main concepts. While the literature and theories provided foundational concepts to model culture, the core AMOn+ ontology describes cultural variations only at an abstract level. For example, only 23 classes out of possible 125 (18.4%) had concrete instances.

In order to extend the core ontology with more concrete instances, DBpedia (Auer et al., 2007) was selected as the knowledge source. DBpedia is one of the largest multi-domain ontologies that currently exist. The DBpedia release used in this work (DBpedia version 3.9) consists of 2.46 billion statements; out of which 470 million are extracted from the English edition of Wikipedia, 1.98 billion are extracted from other language editions, and about 45 million are links to external data sets[2,3]. Previous research has carried out technical evaluation of DBpedia that focuses on the knowledge quality (e.g. Lehmann et al., 2015). However, the general purpose knowledge pool in DBpedia is too huge for a specific domain. Hence, we have devised a mechanism for extracting a domain-specific pool, using the core AMOn+ as an indicator for scope of the domain.

A two-step DBpedia extraction mechanism was followed: (1) identification of resources from DBpedia by mapping the core ontology to the DBpedia dataset, i.e. interlinking concepts from the core AMOn+ ontology to relevant DBpedia concepts; and (2) extraction of concrete instances based on the interlinking conducted in (1).

Interlinking Concepts from the Core AMOn+ Ontology with DBpedia

The interlinking of two ontologies, where one ontology is matched with another, is a common practice in the Web of Data (Bizer et al., 2009). Interlinking is a pre-requisite for extraction from a data set which links to an item in another data set. In order to align concepts from the core ontology

[2] http://blog.dbpedia.org/

[3] Wikipedia, the source for DBpedia, currently supports 280 languages (source: https://en.wikipedia.org/wiki/List_of_Wikipedias, accessed on 03/12/2015)

to DBpedia, a widely known interlinking procedure was used which focused on looking up matches based on lexical forms of entities in two ontologies (Raimond et al., 2007). For this, WordNet[4] was utilized to create enriched surface forms (labels) for all entities from the core AMOn+ ontology. The enriched surface forms were matched with DBpedia entities using the DBpedia Lookup services[5]. As a result of this interlinking process, 76 DBpedia resources matches were found linking with 125 Classes of the core AMOn+ ontology (60.8% match). At the instance level, 40 out of a possible 62 matches were found (64.5%). Hence in total, 116 matches (62%) were found out of 187 entities.

Extraction of Concrete Instances from DBpedia

The interlinking process was followed by a workflow to extract instances from DBpedia, utilizing DBpedia categorization to retrieve relevant resources. Many of the mapped resources from step 1 (described earlier) had links to different DBpedia categories. These category resources, in turn are linked to other categories and page resources by the SKOS vocabulary (Miles & Bechhofer, 2009), creating category chains. This meant that for each category, it was possible to find both broader and narrower categories, as well as page resources which had been placed under a particular category. Figure 10 illustrates the richness in the DBpedia category chains which provide the opportunity to extract DBpedia concepts and instances in an iterative manner. The DBpedia extension in AMOn+ goes up to level 3. For example, in Figure 10, level 3 extraction will include concepts such as “Body Language”, “Human Positions”, “Bowing” and “Gestures”, and “Cheek Kissing”, whereas the concept “Greetings” will not be part of the level 3 extraction (it will be retrieved at level 4 extension).

[4] http://wordnet.princeton.edu/wordnet/related-projects/#REST

[5] http://wiki.dbpedia.org/lookup/

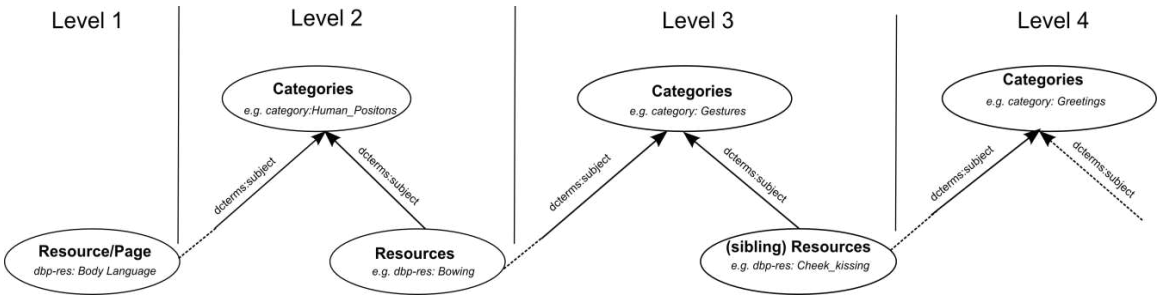

Figure 10: Illustration of the richness in DBpedia category chains. Many resources (e.g. "Body Language" in the figure) in DBpedia are linked to DBpedia categories. These category resources, in turn, are interlinked with other categories (e.g. Category: "Human Positions") and page resources (e.g. "Bowing") through the use of a corresponding SKOS vocabulary.

Figure 11 illustrates the enrichment DBpedia brings to the core AMOn+ ontology. By mapping the concept "Non-verbal communication" from the core AMOn+ ontology to DBpedia, it was possible to enrich the AMOn+ core with new instances such as "Bowing" and "Cheek Kissing". (The AMOn+ ontologies are available from the AMOn+ website[6]).

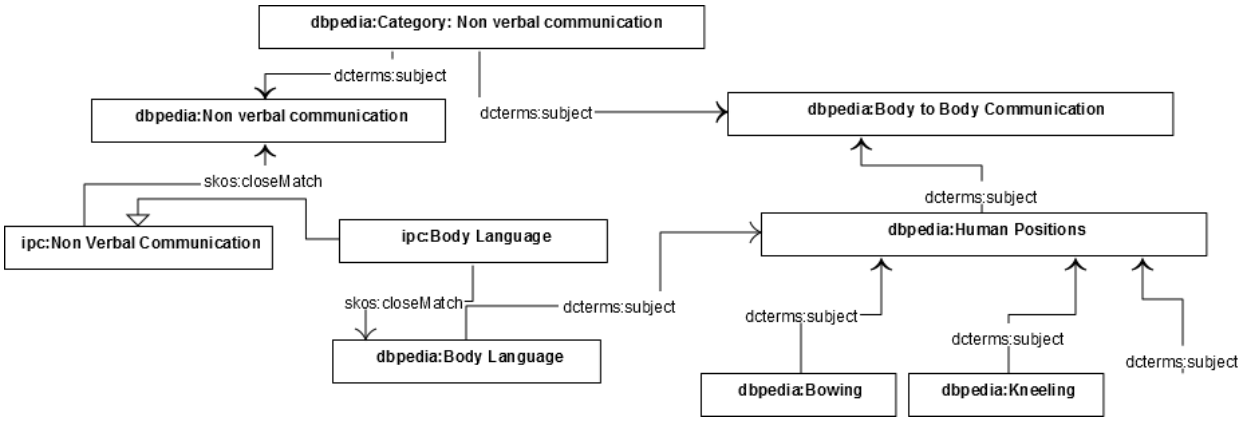

Figure 11: Extending the core AMOn+ ontology with DBpedia, illustrated with the concept "Body Language".

Evaluation

Evaluation Methodology

The AMOn+ evaluation follows the methodological guidelines from the EU project NeOn (Suárez-Figueroa, 2010). The main goal for the AMOn+ evaluation was to check the ontology's fitness-for-

[6] http://purl.org/amon

purpose with regard to automated semantic tagging of content that contains cultural and interpersonal communication aspects. This evaluation examined both the core AMOn+ ontology and its extension with DBpedia. Since we are using a crowdsourced knowledge base (DBpedia) where we have no control over the presence and validity of domain facts, it was important for the evaluation to validate the domain coverage of AMOn+. It is also acknowledged in other evaluation studies (Suárez-Figueroa et al., 2012) that for applications using ontologies for semantic tagging and natural language processing tasks, domain coverage (which measures the extent to which an ontology covers a considered domain, in our case 'cultural variations in interpersonal communication') is a reliable measure for the ontology's fitness-for-purpose. A standard practice for measuring the domain coverage of an ontology is to compare automatic annotations using the ontology with annotations provided by humans (often called gold standard (Brewster et al., 2004). This gold standard is essentially an annotated textual corpus that can be used as a benchmark in the chosen domain (Brewster et al., 2004). Because there is no annotated textual corpus in the domain of cultural variations in interpersonal communication, the production of such a corpus was a key step in the AMOn+ evaluation.

The General Architecture for Text Engineering (GATE) toolkit (Cunningham et al., 2002), a popular platform for manual and automatic annotation, was chosen as the annotation platform. A comparison was made on the domain coverage by using (1) different levels of AMOn+ to inform the automatic annotation of the gold standard using the annotation tool and (2) human experts manually annotating as the gold standard. To complete the evaluation, the results were analyzed using the conventional metrics of precision, recall and F-measure (van Rijsbergen, 1986). Further details about the construction of the gold standard and the annotation results are provided in the following subsections.

Construction of Gold Standard

The gold standard for the AMOn+ evaluation was created in two steps: (1) collection of textual corpus on interpersonal communication rich in cultural variations; and (2) manual annotation of the corpus by experts to identify key terms relevant to the domain.

Collection of corpus

Four corpora, with different degrees of curation, on interpersonal communication including cultural variations were collected. The first corpus - crowdsourced content with no curation - included user comments on YouTube videos with interpersonal communication examples referring to cultural variations. The second corpus – crowdsourced content with some curation – included Wikipedia pages about cultural rules and norms for different nations. The third and fourth corpora – proprietary and are properly curated - included descriptions of various cultural rules and norms collected from newspapers and a web resources for cultural awareness training.

YouTube comments. This corpus was collected within the EU ImREAL project, and was produced by university students from various national cultures. 29 participants (referred here as 'commentators') from 23 different countries were invited to comment on 7 YouTube videos. 16 of them had at least visited more than 5 countries hence can be assumed to have a good level of cultural exposure (Crowne, 2008). The videos were selected based on the following criteria: (1) easy to follow; (2) engaging and provocative to stimulate a wide range of contributions; and (3) referring to relevant cultural aspects. In particular, the videos considered aspects such as gestures, greetings, personal space and use of artefacts. The commentators were encouraged to compare and contrast the episodes in the videos with their experiences from their own culture and other cultures that they had encounters with. The corpus consisted of 102 user comments, including different views on interpersonal communication examples and comments about personal experiences.

Wikipedia Articles. Wikipedia contains articles on etiquette in various cultures including aspects on gifting norms, clothing norms, eating norms, greeting and gestures. All articles on etiquette from

Wikipedia were included (10 Wikipedia pages). This covered Africa, Asia, Australia and New Zealand, Indonesia, Japan, Latin America, North America, Europe, and the Middle East. This corpus and the DBpedia dataset have no overlaps with each other. DBpedia is a semantic version of Wikipedia but solely created using the Wikipedia infobox templates and category information. The Wikipedia corpus uses textual content from these pages and excludes infobox templates and category information.

News Articles and a Specialized Web Resource. A collection of news articles was obtained using a search engine. The corpus consisted of 17 articles from Guardian, Daily mail, Daily telegraph, and Reuters (e.g. Guardian carries an article series[7] on gestures in various countries). In addition, the proprietary corpus included articles from a leading global portal, called eDiplomat[8], which comprises web pages with advice to diplomats on various cultural nuances and cultural etiquette in different countries. This included aspects such as meeting and greeting people, body language, dining norms, dressing norms, and gifting norms. The corpus included 263 articles about 44 countries covering most parts of the globe.

Manual annotation of the corpus by experts

This step included: (1) selection of a suitable expert for the annotation; (2) obtaining a representative sample from the corpus for annotation by the expert (this is a standard practice as it is not possible for an expert to annotate the whole corpus); (3) conducting the manual corpus annotation using an annotation tool; and finally (4) a second expert acted as curator validating the annotations by the first expert.

Expert selection. An expert[9] with cultural intelligence (CQ) (Ang & Van Dyne, 2008; Earley & Mosakowski, 2004) score of 6.5 (out of possible 7) was selected to perform the manual annotation.

Sampling the corpus. To sample the corpus to a size that is manageable for manual annotation, content was selected from each of the four corpora covering at the least one country from each of

[7] http://www.theguardian.com/travel/gallery/2010/feb/05/launguage-gestures-japanese-spanish-arabic

[8] http://www.ediplomat.com/np/cultural_etiquette/cultural_etiquette.htm

[9] The expert was not part of the ontology design process.

the GLOBE culture clusters (House et al., 2004). The sample included a total of 40 items consisting of 10 items from each corpus for the 10 Globe clusters. These 40 items represent 41.8% of the total number of words present in the corpus.

Annotation process. The expert was asked to annotate any terms from the text that would enable search on cultural content collected from the web (from newspapers, web sites, Wikipedia, UGC). In a typical search scenario, a user generally selects keywords to search for; the expert was asked to pick in the text any words/phrases that could relate to keywords that may be used for search.

Validation of the expert's annotation. To ensure confidence in the annotation results, another expert was recruited to annotate a random 10% sample from the corpus annotated by the first expert, following the same annotation conventions. There was 92.7% Inter Annotator Agreement (IAA) (Cohen, 1968) between both experts, which indicated significant level of agreement (Cohen, 1968). This places confidence in the first expert's annotation; hence, this expert's annotations of the sample corpus were used as a gold standard to compare with the automatic semantic annotation (tagging) using AMOn+.

Automatic annotation using AMOn+

The GATE (Cunningham et al., 2002) toolkit was also used to automatically annotate the textual corpus with the AMOn+ ontology. The criteria for selection of an automatic annotation toolkit were: (1) it should be provided as open source so that our experiment could be replicated and the results could be reproduced and verified, and (2) it should allow selection of ontologies for semantic tagging (AMOn+ in this case). GATE is the biggest open source project for textual annotation that allows the use of custom ontologies (Cunningham et al., 2002). GATE is also used in all types of human computational tasks ranging from detecting events from financial text (Hogenboom et al., 2013) to detecting opinions from social media (Maynard et al., 2014). It has one of the largest and most diverse user community for any text engineering tools (Cunningham et al., 2002). The GATE toolkit consists of several processing resources that help produce semantic annotations. First, GATE identifies word and sentence boundaries. Each token (word) is then assigned a grammatical

category and a linguistically true base-form. In our setup, GATE uses gazetteers created from the AMOn+ ontology that allow annotating documents with concepts from AMOn+. Consequently, the GATE annotation relies solely on the AMOn+ ontology to provide all concepts and necessary surface forms to annotate the documents. Therefore, the performance of the GATE annotation, which is dependent only on the AMOn+ ontology, gives an indication of the fitness-for-purpose of AMOn+.

Table 3 depicts the information on the size of the corpus used in the evaluation. The specialized web resource is the largest in size, both in terms of number of articles and words in the corpus. The system also found the largest number of annotations for this corpus. The Wikipedia corpus contains the least number of articles but second highest number of words. In terms of number of words, the news article corpus was the smallest.

Table 3: Details about the sample corpora size and number of annotations

Corpus	Total No. of words	Total No. of articles	No. of AMOn+ annotations
YouTube Comments	9203	102	5012
Wikipedia Articles	18893	10	10188
News Articles	5482	17	3481
Specialized Web Resources	25819	263	17525

Results and Analysis

We compare Inter Annotator Agreement (IAA) between automatic and manual annotations using the standard information retrieval metrics of Precision, Recall and F-measure (van Rijsbergen, 1986). Other IAA measures, e.g. kappa (Carletta, 1996), are seen as unsuitable for text mark-up tasks such as named entity recognition and information extraction (Hripcsak & Rothschild, 2005). Instead, Precision, Recall, and F-measure have been widely used for measuring IAA in relevant information extraction evaluations such as MUC[10] and ACE[11]. These metrics are also commonly used for evaluating information extraction systems, allowing comparison of the IAA results with results from other systems published in the literature. However, defining thresholds where agreement is deemed

[10] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html

[11] http://projects.ldc.upenn.edu/ace/

'good' is subject to debate, including the choice of measures (e.g. Artstein & Poesio, 2008; Eugenio, 2000; Landis & Koch, 1977). This is particularly applicable in cases where F-Measure is the choice of IAA since there is no widely accepted research that define such thresholds. Hence, for the work presented here, no specific thresholds are set, although the aim is, naturally, to reach as high as possible in agreement (F-Measure).

The automatic semantic annotations using AMOn+ were compared against the expert annotations from the gold standard. Figure 12 shows the annotation results using a sample text from the corpus. The top half of the screen shows the expert's annotations, while the bottom half shows the system's annotations. Figure 13 shows the Inter Annotator Agreement between the expert and the AMOn+ annotations. In the given example, the average F-Measure is 68%.

Figure 12: (top) GATE interface showing an expert's annotation on a comment from the YouTube corpus; (bottom) The GATE annotations on the same text using AMOn+.

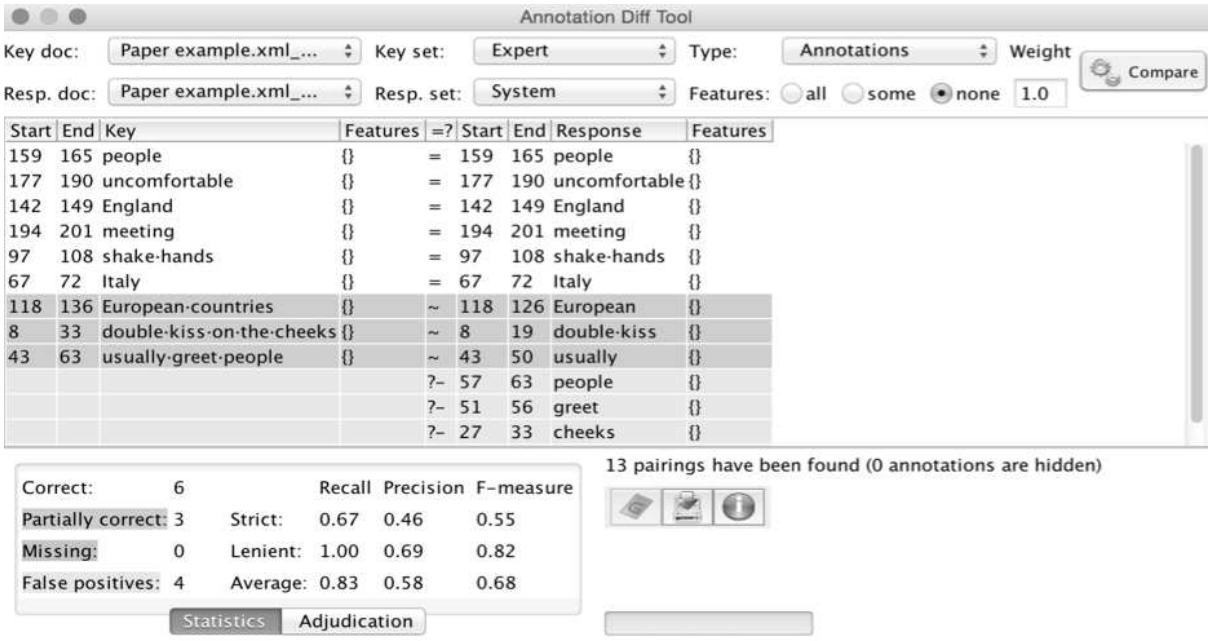

Figure 13: GATE interface showing Precision, Recall and F-Measure between the system and expert annotations for the comment shown in the Figure 12. There is a match between six cases, while the system annotated three annotations that the expert did not. Furthermore, for three AMOn+ annotations, there was just a partial match with the corresponding expert's annotations.

Figure 14 outlines the results using different levels of AMOn+. On average, the core level of AMOn+, which contains no DBpedia enrichment, achieved the highest level of precision (0.93, σ= 0.03), but the lowest Recall (0.09, σ= 0.02) and the lowest F-Measure (0.17, σ= 0.04). This shows that core AMOn+ ontology, which was underpinned only by theoretical models, provided only abstract conceptualization and on its own was inadequate to support a semantic annotation process of UGC.

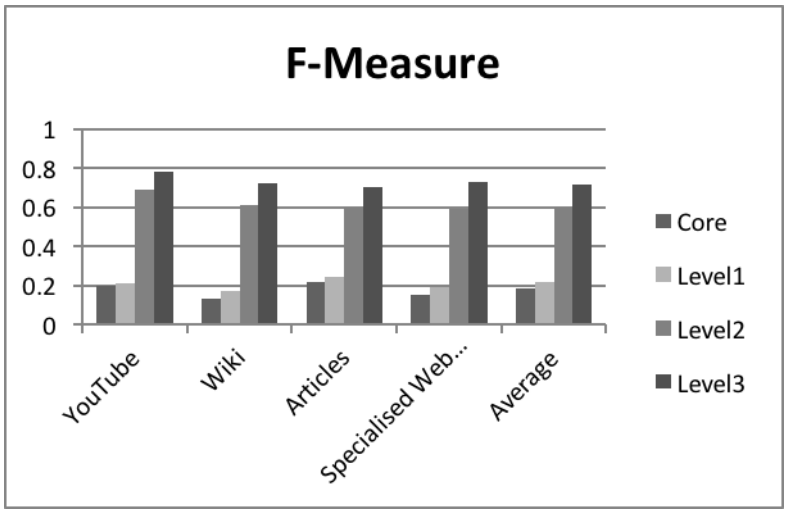

Figure 14: Performance (F-Measure) of the AMOn+ based automatic annotations against the gold standard. On average, Level 3 DBpedia extension performs best.

The level 1 enrichment of the core AMOn+ ontology, where the concepts from core ontology are directly mapped to concepts and relevant lexical forms from DBpedia, achieved slightly better F-Measure (0.2, σ=0.02) than the core level. In this case, the enrichment provides minor boost in recall (0.11, σ=0.02) and has no effect on Precision (0.93, $\sigma = 0.03$).

The level 3 of AMOn+, the highest level of enrichment of the core ontology using DBpedia category chains, achieved highest overall score for the F-Measure (0.73, σ=0.03). Level 3 also achieved the highest overall score of Recall (0.79, σ=0.07) which is an improvement of 0.7 over the performance achieved with core ontology.

The core AMOn+ ontology (based solely on the theoretical models) achieved the highest level of precision (0.93). Therefore, core AMOn+ ontology can be a reliable source when one is looking for a conservative knowledge base that will give a smaller amount of concepts which are guaranteed to be part of the domain of interest.

The core ontology also provides a suitable level of abstract concepts that can be extended by crowdsourced knowledge bases such as DBpedia. Indeed, extending the core ontology with DBpedia is justified in our work as there is improvement in F-Measure for all the extension levels. The evaluation revealed that AMOn+ was representative of the cultural variations in interpersonal communication domain and is fit-for-purpose with regard to semantic tagging of relevant text. In other words, given a corpus with text related to interpersonal communication, AMOn+ will allow identifying the mentions in the text that refer to interpersonal communication concepts which can have differences across different national cultures (e.g. gestures, greetings, clothing). The automatic identification of such concepts (and the corresponding mentions in the text that refer to these concepts), allow the implementation of intelligent features that facilitate search, information exploration, data analytics, etc. In the following section we illustrate one such application.

Application: Semantic Tagging and Exploration with AMOn+

Various studies have shown an increase in the use of the social Web by different demographic groups to contribute content influenced by their cultural background (Thomas & Sheth, 2011). There are potential benefits in mining such content and providing them as exemplar of different cultural perspectives. A semantic exploration system, called Pinta, was developed to support informal learning in the area of cultural variations in interpersonal communications (Authors, 2015). Pinta extensively utilizes AMOn+ ontologies to semantically augment UGC together with proprietary corpus. The current version of Pinta includes the text corpora listed above. The semantic tagging component of Pinta is implemented using the GATE toolkit. An instance of Pinta was created by semantically annotating the corpus content we used in the evaluation presented in the previous section.

To illustrate how Pinta can be used to support informal learning for cultural variations in interpersonal communication, let us return to the scenario described in the introduction. Ian is a new business development manager at a firm that has recently developed business interests in Brazil. Ian was advised by some colleagues that interpersonal communication and cultural aspects play a crucial role in business dealings with the clients from Brazil. Ian plans to take a few clients to dinner, and he wants to learn more about these aspects. Ian now can use Pinta to learn about interpersonal communication and cultural aspects. He starts off by selecting the term “Greetings” and is offered various content on greetings in the form of videos from the YouTube (and comments) and blogs on greetings in different cultures. By reading these comments and blogs, Ian realizes that greeting have cultural variations, and that can be several possible interpretations of different greetings. In addition, he can see concepts related to greetings from AMOn+ (see Figure 15). He is particularly interested in the link pointing to greeting with “kissing” (derived from AMOn+ by Pinta), which leads to various examples of greeting with “kissing”. For example, Figure 16 shows two example stories from Pinta. First, a YouTube video comment, discusses the variations in cheek

kissing in Brazil. The other example is a generic article about etiquette in Latin America focusing on generalities across various countries.

Figure 15: Pinta interface showing AMOn+ concepts related to "Greetings". Each of these concepts is clickable and can offer a different exploration path.

Figure 16: Pinta interface showing example content annotated with the concept of "Kissing". The first story with the title "In Brazil…" is a comment on a YouTube video and refers to the number of kisses appropriate in different occasions in Brazil. The second story is a Wikipedia entry about etiquette in Latin America and discusses generalization across Latin America.

Conclusion

Qualitative analysis of UGC, a technique that relies on semantic tagging using an ontology, offers exciting opportunities for gaining deeper insights in social web content. This approach is dependent on the availability and the quality of the ontology in the specific domain. For the domain of cultural variation in interpersonal communication such an ontological model does not exist. Computational models of ill-defined domains, such as culture, require interdisciplinary approaches. Such an

approach is presented in this paper. The primary contribution of this paper is a hybrid approach to ontology development, combining social science and computer science approaches to capture cultural variations in interpersonal communication. The ontology developed can be used to support semantic tagging of UGC for further exploration. The paper sought to answer two interrelated questions: RQ1: How to develop an ontology of cultural variations in interpersonal communication? RQ2: How well does such an ontology support tagging UGC containing cultural variations in interpersonal communication?

We answered the first question by devising a novel two-phase methodology to build a complex ontology using theoretical models and crowdsourced knowledge from linked data. In the first phase, we used well-established cultural theories and frameworks to extend an ontology for interpersonal communication built in earlier research, resulting in a core ontology with top level concepts that refer to cultural variations. In the second phase, the core ontology was further extended through an automatic enrichment process that allowed further specialization with ontology facts derived from a crowdsourced knowledge base, namely DBpedia. The automated extension added extra layers of entities and relationships to the core ontology, resulting in AMOn+. Detailed discussion of the answer to RQ1 is included in outline of the methodology section, and the sections presenting the core AMOn+ and the DBpedia extension process that produced levels 1, 2 and 3 of AMOn+).

We answered the second question by evaluating the fit-for-purpose of AMOn+, focusing on domain coverage. Human experts were asked to annotate cultural concepts that may have variations in a selected corpora. The outcome was compared with the annotations produced by an automated tagging process, using a well-established semantic tagging tool (GATE) and AMOn+ as the underpinning ontology. Analysis was also conducted on the effectiveness of the different levels of AMOn+.

The core AMOn+ ontology (based solely on the theoretical models) achieved the highest level of precision (0.93). The results indicated that the core AMOn+ ontology is a reliable source when one is looking for a high level picture on the extent of cultural content that may have cultural variations in

a piece of text. For example, tagging educational content (lectures, tutorials, presentations) with domain concepts, broad detection of cultural variations for simulations of culture-aware interaction agents, or social robots that emulate cultural behavior.

The core AMOn+ ontology alone will not be suitable for tagging UGC (as indicated by its poor recall value). When social web content is considered, there is a vast amount of variations and it will be more useful to be specific about the range of concepts and instances related to cultural variations (e.g. various greeting gestures across the globe). For this, the DBpedia extension to the core AMOn+ ontology is more appropriate: the recall increased noticeably when adding knowledge from DBpedia (reaching the value of 0.79), while precision decreased but still the overall F-measure was the highest of 0.73. The benefit of using the AMOn+ extended with DBpedia knowledge facts has been illustrated in the semantic exploration system Pinta that uses heterogeneous UGC coming from different sources to support informal learning in the domain of interpersonal communication.

This research provides a step change for gaining deeper insights into a large volume of UGC on cultural aspects of interpersonal communication by exploiting ontological models. It has had relatively little prior research to build on and has provided the first ontological model for capturing cultural variations. It opens new avenues that can be explored as future work. For example, the automated approach for extracting a culture-related knowledge pool from the crowdsourced knowledge base DBpedia presented in this paper can potentially be followed for extracting knowledge from other linked data sources. Moreover, the hybrid ontology engineering approach presented here can be applied in other ill-defined domains (e.g. sensemaking, decision making, mentoring) where UGC can provide rich source of personal experiences.

Our research suggests that social theory can inform ontology development. Rather than presenting a straightforward process and transition from contemporary social theory to ontology, our work demonstrates that some theoretical concepts suffer in their transposition into logical form (Authors, 2013; Baker, 2000). Future research could examine this process in reverse in order to understand how the ontology upper level can unify and inform the specific social theories.

References

Allwood, J. (1985). Intercultural communication. *Papers in Anthropological Linguistics, 12*, 1-25.

Ang, S., & Van Dyne, L. (2008). *Handbook of cultural intelligence*: ME Sharpe.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555-596.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data*: Springer.

Authors. (2011). *Journal of the American Society for Information Science and Technology*.

Authors. (2013). *Journal of the American Society for Information Science and Technology*.

Authors. (2015). *International Journal of Distributed Systems and Technologies*.

Baader, F., & Nutt, W. (2003). Basic description logics. In B. Franz, C. Diego, L. M. Deborah, N. Daniele, & F. P.-S. Peter (Eds.), *The description logic handbook* (pp. 43-95): Cambridge University Press.

Baker, M. (2000). International Journal of Artificial Intelligence in Education. *The roles of models in Artificial Intelligence and Education research: a prospective view, 11*(2), 122-143.

Bao, J., Smart, P., Braines, D., & Shadbolt, N. (2009). *A Controlled Natural Language Interface for Semantic Media Wiki Using the Rabbit Language*. Paper presented at the Workshop on Controlled Natural Language (CNL'09), Marettimo Island.

Bennett, M. (1983). A developmental model of intercultural sensitivity *Working Paper*.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*(5), 34-43.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems, 5*(3), 1-22.

Bontcheva, K., & Rout, D. (2012). Making Sense of Social Media Streams through Semantics: a Survey. *Semantic Web Journal*.

Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). Data driven ontology evaluation.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. New York: Cambridge University Press.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics, 22*(2), 249-254.

Castano, E., Yzerbyt, V., & Bourguignon, D. (2003). We are one and I like it: The impact of ingroup entitativity on ingroup identification. *European Journal of Social Psychology, 33*(6), 735-754.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70*(4), 213.

Cole, M. (1999). Cultural psychology: Some general principles and a concrete example. In Y. Engeström, R. Miettinen, & R.-L. Punamäki-Gitai (Eds.), *Perspectives on activity theory* (pp. 87-106). Cambridge: Cambridge University Press.

Crowne, K. A. (2008). What leads to cultural intelligence? *Business Horizons, 51*(5), 391-399.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). *GATE: an architecture for development of robust HLT applications.* Paper presented at the Proceedings of the 40th annual meeting on association for computational linguistics.

Dawkins, R. (2006). *The selfish gene* (3rd ed.). Oxford: Oxford University Press.

Denaux, R., Dolbear, C., Hart, G., Dimitrova, V., & Cohn, A. G. (2011). Supporting Domain Experts to Construct Conceptual Ontologies: A Holistic Approach. *Journal of Web Semantics, 9*(2), 113–127.

Dormann, C., & Chisalita, C. (2002). *Cultural values in web site design.* Paper presented at the Proceedings of the 11th European Conference on Cognitive Ergonomics ECCE11.

Earley, P. C., & Mosakowski, E. (2004). Cultural Intelligence. *Harvard Business Review, October*, 139-146.

Engeström, Y. (1999). Innovative Learning in Work Teams: Analyzing Cycles of Knowledge Creation in Practice. In Y. Engestrom, R. Miettinen, & R.-L. Punamaki (Eds.), *Perspectives on Activity Theory* (pp. 377 - 406). Cambridge: Cambridge University Press.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., . . . Yates, A. (2005). Unsupervised named-entity extraction from the web: an experimental study. *Artif. Intell., 165*(1), 91-134. doi: 10.1016/j.artint.2005.03.001

Eugenio, B. D. (2000). *On the Usage of Kappa to Evaluate Agreement on Coding Tasks.* Paper presented at the Second International Conference on Language Resources and Evaluation.

Fayyad, U. M., & Piatetsky-Shapiro, G. (1996). Advances in Knowledge Discovery and Data Mining. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Data Mining and Knowledge Discovery*.

Ford, G., & Gelderblom, H. (2003). *The effects of culture on performance achieved through the use of human computer interaction.* Paper presented at the Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology.

Hall, E. T. (1983). *The Silent Language*. New York: Doubleday.

Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology, 1*(1), 1-136.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61-83.

Hofstede, G. H. (1991). *Cultures and organizations : software of the mind*. London: McGraw-Hill.

Hogenboom, A., Hogenboom, F., Frasincar, F., Schouten, K., & van der Meer, O. (2013). Semantics-based information extraction for detecting economic events. *Multimedia Tools and Applications, 64*(1), 27-52.

Horrocks, I., & Sattler, U. (2005). *A tableaux decision procedure for SHOIQ.* Paper presented at the Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI).

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. J., Gupta, V., & Associates, G. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies.* Thousand Oaks, CA: Sage.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical Informatics Association, 12*(3), 296-298.

Isaac, A., & Haslhofer, B. (2013). Europeana linked open data–data. europeana. eu. *Semantic Web, 4*(3), 291-297.

Jones, M., & Alony, I. (2007). The Cultural Impact of Information Systems – Through the Eyes of Hofstede – A Critical Journey. *Issues in Informing Science and Information Technology, 4*, 408-419.

Jost, J. T., & Hamilton, D. L. (2005). Stereotypes in our culture. In J. Dovidio, P. Glick, & L. Rudman (Eds.), *On the Nature of Prejudice: Fifty years after Allport* (pp. 208-224). Oxford: Blackwell.

Kim, H.-L., Breslin, J., Kim, H.-G., & Choi, J.-H. (2010). Social semantic cloud of tags: semantic model for folksonomies. *Knowledge Management Research & Practice, 8*, 193-202.

Knublauch, H., Fergerson, R. W., Noy, N. F., & Musen, M. A. (2004). The Protégé OWL plugin: An open development environment for semantic web applications *The Semantic Web–ISWC 2004* (pp. 229-243): Springer.

Kralisch, A., & Berendt, B. (2004). *Cultural Determinants of Search Behaviour on Websites.* Paper presented at the IWIPS.

Kuhn, W. (2001). Ontologies in support of activities in geographical space. *International Journal of Geographical Information Science, 15*(7), 613-631. doi: 10.1080/13658810110061180

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Leginus, M., Zhai, C., & Dolog, P. (2015). Personalized generation of word clouds from tweets. *Journal of the Association for Information Science and Technology*, n/a-n/a. doi: 10.1002/asi.23494

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., . . . Auer, S. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web, 6*(2), 167-195.

Leidner, D. E., & Kayworth, T. (2006). Review: A Review of Culture in Information Systems Research: Toward a Theory of Information Technology Culture Conflict. *MIS Quarterly, 30*(2), 357-399.

Leiman, M. (1999). The concept of sign in the work of Vygotsky, Winnicott, and Bakhtin: Further integration of object relations theory and activity theory. In Y. Engeström, R. Miettinen, & R.-L. Punamäki-Gitai (Eds.), *Perspectives on activity theory. Learning in doing : social, cognitive and computational perspectives* (pp. 419-434). Cambridge: Cambridge University Press.

Liu, H., & Maes, P. (2007). Introduction to the semantics of people & culture: IGI PUBLISHING 701 E CHOCOLATE AVE, STE 200, HERSHEY, PA 17033-1240 USA.

López, M. F., Gómez-Pérez, A., Sierra, J. P., & Sierra, A. P. (1999). Building a Chemical Ontology Using Methontology and the Ontology Design Environment. *Ontologies*(January/February), 37-46.

Marcus, A., & Gould, E. W. (2000). Crosscurrents: cultural dimensions and global Web user-interface design. *interactions, 7*(4), 32-46.

Matsumoto, D., & Hwang, C. H. (2013). Cultural Similarities and Differences in Emblematic Gestures. *Journal of Nonverbal Behaviors, 37*, 1-27.

Maynard, D., Gossen, G., Funk, A., & Fisichella, M. (2014). Should i care about your opinion? detection of opinion interestingness and dynamics in social media. *Future Internet, 6*(3), 457-481.

McGuinness, D. L., & Van Harmelen, F. (2004). OWL web ontology language overview. *W3C recommendation, 10*(10), 2004.

Mesquita, B., Frijda, N. H., & Scherer, K. R. (1997). Culture and emotion. In P. Dasen & T. S. Saraswathi (Eds.), *Handbook of cross-cultural psychology:* (Vol. 2). Boston: Allyn & Bacon.

Miles, A., & Bechhofer, S. (2009). SKOS simple knowledge organization system reference. *W3C recommendation, 18*, W3C.

Mizoguchi, R. (2003). Tutorial on ontological engineering. *NEW GENERATION COMPUTING-TOKYO-, 21*(4), 363-364.

Munro, R., & Manning, C. D. (2012). *Accurate unsupervised joint named-entity extraction from unaligned parallel text*. Paper presented at the Proceedings of the 4th Named Entity Workshop, Jeju, Republic of Korea.

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes, 30*(1), 3-26.

Nasukawa, T., & Yi, J. (2003). *Sentiment analysis: capturing favorability using natural language.* Paper presented at the 2nd International Conference on Knowledge Capture.

Nicholas, D., Clark, D., Rowlands, I., & Jamali, H. R. (2013). Information on the go: A case study of Europeana mobile users. *Journal of the American Society for Information Science and Technology*, n/a-n/a. doi: 10.1002/asi.22838

Nisbett, R. E., & Norenzayan, A. (2002). Culture and cognition. In D. Medin & H. Pashler (Eds.), *Stevens' Handbook of Experimental Psychology* (3rd ed., Vol. 2). New York: John Wiley & Sons.

O'Leary, D. E. (2010). Enterprise ontologies: Review and an activity theory approach. *International Journal of Accounting Information Systems, 11*(4), 336-352. doi: DOI: 10.1016/j.accinf.2010.09.006

Purday, J. (2009). Think culture: Europeana.eu from concept to construction. *The Electronic Library, 27*(6), 919-937. doi: doi:10.1108/02640470911004039

Raimond, Y., Abdallah, S. A., Sandler, M. B., & Giasson, F. (2007). *The Music Ontology.* Paper presented at the ISMIR.

Reinecke, K., & Bernstein, A. (2008). *Predicting user interface preferences of culturally ambiguous users.* Paper presented at the CHI'08 extended abstracts on Human factors in computing systems.

Reinecke, K., & Bernstein, A. (2013). Knowing what a user likes: A design science approach to interfaces that automatically adapt to culture. *Mis Quarterly, 37*(2), 427-453.

Reinecke, K., Schenkel, S., & Bernstein, A. (2010). Modeling a user's culture. *Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models*, 242-264.

Scharifian, F. (2003). On cultural conceptualizations. *Journal of Cognition and Culture, 3*(3), 187-207.

Schwartz, S. H. (1994). Beyond individualism/collectivism: New dimensions of values. In U. Kim, H. C. Triandis, C. Kagitcibasi, S. C. Choi, & G. Yoon (Eds.), *Individualism and Collectivism: Theory Application and Methods*. Newbury Park, CA: Sage.

Sheppard, C., & Scholtz, J. (1999). *The effects of cultural markers on web site use.* Paper presented at the Proceedings of the 5th Conference on Human Factors and the Web.

Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Oxford: Blackwell.

Stuckenschmidt, H., Parent, C., & Spaccapietra, S. (2009). *Modular ontologies: concepts, theories and techniques for knowledge modularization* (Vol. 5445): Springer.

Suárez-Figueroa, M. C. (2010). *NeOn Methodology for building ontology networks: specification, scheduling and reuse.* Informatica.

Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E., & Gangemi, A. (2012). *Ontology engineering in a networked world*: Springer Science & Business Media.

Sundheim, B. M., & Chinchor, N. (1995, November 6-8). *Named entity task definition, version 2.1.* Paper presented at the Sixth message understanding conference, Columbia.

Syn, S. Y., & Spring, M. B. (2013). Finding subject terms for classificatory metadata from user-generated social tags. *Journal of the American Society for Information Science and Technology, 64*(5), 964-980. doi: 10.1002/asi.22804

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology, 57*(1), 60-68.

Thomas, C., & Sheth, A. (2011). Web Wisdom: An essay on how Web 2.0 and Semantic Web can foster a global knowledge society. *Computers in Human Behavior, 27*(4), 1285-1293.

van Rijsbergen, C. J. (1986). *(invited paper) A new theoretical framework for information retrieval.* Paper presented at the Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval.

Wimalasuriya, D., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science, 36*(3), 306-323. doi: citeulike-article-id:7291004

Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques.

Yi, K. (2010). A Semantic Similarity Approach to Predicting Library of Congress Subject Headings for Social Tags. *Journal of the American Society for Information Science and Technology, 61*(8), 1658-1672.

Yoon, K. (2012). Conceptual syntagmatic associations in user tagging. *Journal of the American Society for Information Science and Technology, 63*(5), 923-935.