



ARL-TR-7465 • SEP 2015

**ARL**

US Army Research Laboratory

# Exploring Social Meaning in Online Bilingual Text through Social Network Analysis

by Elizabeth K Bowman, Nkonko Kamwangamalu, Heather Roy, Alla Tovaes, Sue Kase, Michelle Vanni, Mugizi R Rwebangira, and Mohamed Chouikha

Approved for public release; distribution is unlimited.

## **NOTICES**

### **Disclaimers**

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.



# **Exploring Social Meaning in Online Bilingual Text through Social Network Analysis**

**by Elizabeth K Bowman, Heather Roy, Sue Kase, and Michelle Vanni**

*Computational and Information Sciences Directorate, ARL*

**Nkonko Kamwangamalu, Alla Tovaes, Mugizi R Rwebangira, and Mohamed Chouikha**

*Howard University, Washington, DC*

**REPORT DOCUMENTATION PAGE**

Form Approved  
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

**PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

<b>1. REPORT DATE (DD-MM-YYYY)</b> September 2015		<b>2. REPORT TYPE</b> Final		<b>3. DATES COVERED (From - To)</b> 1 October 2012–30 September 2014	
<b>4. TITLE AND SUBTITLE</b> Exploring Social Meaning in Online Bilingual Text through Social Network Analysis				<b>5a. CONTRACT NUMBER</b>	
				<b>5b. GRANT NUMBER</b>	
				<b>5c. PROGRAM ELEMENT NUMBER</b>	
<b>6. AUTHOR(S)</b> Elizabeth K Bowman, Nkonko Kamwangamalu, Heather Roy, Alla Tovares, Sue Kase, Michelle Vanni <sup>a</sup> , Mugizi R Rwebangira, and Mohamed Chouikha				<b>5d. PROJECT NUMBER</b>	
				<b>5e. TASK NUMBER</b>	
				<b>5f. WORK UNIT NUMBER</b>	
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b> US Army Research Laboratory ATTN: RDRL-CII-C Aberdeen Proving Ground, MD 21005-5067				<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b> ARL-TR-7465	
<b>9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b>				<b>10. SPONSOR/MONITOR'S ACRONYM(S)</b>	
				<b>11. SPONSOR/MONITOR'S REPORT NUMBER(S)</b>	
<b>12. DISTRIBUTION/AVAILABILITY STATEMENT</b> Approved for public release; distribution is unlimited.					
<b>13. SUPPLEMENTARY NOTES</b> <sup>a</sup> US Army Research Laboratory, Computational and Information Sciences Directorate <sup>b</sup> Howard University, Washington DC					
<b>14. ABSTRACT</b> This report documents the intersection of computational social network analysis and sociolinguistic research aimed at discovering how social intent is communicated through online bilingual speech acts in African cultures. Researchers from the US Army Research Lab (ARL) and Howard University (HU) exchanged information, data, and analyses to examine the feasibility of using automated text analytics software to provide contextual understanding within a text corpus. This effort extends the Army Research Office Partners in Research Transition program titled “Extracting Social Meaning from Linguistic Structures Involving Code-Switching in English (and French) with Selected African Languages” led by HU. It also provided test and evaluation opportunities for ARL prototype software designed to extract relational networks and sentiment from unstructured Tweets. This collaboration was driven by the realization that more social input is needed to refine context for sociolinguistic analysis and also by the increasing importance of modeling social issues for military decision making. To address social communication acts, we focus on using Twitter for sharing individual and collective opinions. Social media services in general have gained popularity in recent years and are frequently used for discovery and analysis of social intent. We examine the sociolinguistic features that can be used to discover social intent, discuss how social network analysis can be used to inform contextual nuances in which that intent is communicated, and describe how automated tools can be used to support sociolinguistic analysis. We conclude with future research directions that can extend the rich connections between computational social network analysis and the study of sociolinguistics.					
<b>15. SUBJECT TERMS</b> social network analysis, sociolinguistics, code-switching, social intent, text analytics					
<b>16. SECURITY CLASSIFICATION OF:</b>			<b>17. LIMITATION OF ABSTRACT</b>	<b>18. NUMBER OF PAGES</b>	<b>19a. NAME OF RESPONSIBLE PERSON</b>
<b>a. REPORT</b>	<b>b. ABSTRACT</b>	<b>c. THIS PAGE</b>			Elizabeth K Bowman
Unclassified	Unclassified	Unclassified	UU	72	<b>19b. TELEPHONE NUMBER (Include area code)</b> 410-278-5924

## Contents

---

---

<b>List of Figures</b>	<b>v</b>
<b>List of Tables</b>	<b>vi</b>
<b>Acknowledgments</b>	<b>vii</b>
<b>1. Introduction</b>	<b>1</b>
1.1 Sociolinguistic Theories and Social Meaning of Linguistic Structures in Communication	2
1.2 Contextual Factors and Social Motivations for CS	2
<b>2. Computational Approaches to Defining Social Meaning</b>	<b>7</b>
2.1 Statistical Approach	7
2.2 Fuzzy Logic Approach	8
2.3 Feature Extraction	9
2.4 Relationship Determination	11
2.5 Fuzzification	12
2.6 Statistical Identification of Hierarchy	13
2.7 Grouping Results of the 2 Methods	15
<b>3. Informing Context through Social Network Analysis</b>	<b>23</b>
<b>4. Challenges in Extracting Meaning from Social Media and Networks</b>	<b>27</b>
<b>5. Automated Text Analytics for Social Network and Sociolinguistic Analysis</b>	<b>29</b>
5.1 GATE	29
5.2 Contour	32
5.2.1 Contour Activity Analysis	32
5.2.2 Contour Topic Table	33
5.2.3 Contour Entity Relationship Analysis	33
5.2.4 Contour Media View	34

<b>6. Method</b>	<b>34</b>
6.1 Kenya as a Sociolinguistic Research Site	35
6.1.1 Ethnic and Gender Stereotypes	36
6.1.2 Political Affiliations and Ethnic Groups	37
6.1.3 Class Issues and Stereotypes	37
<b>7. Data Collection</b>	<b>37</b>
<b>8. Results</b>	<b>40</b>
8.1 Contour Detection of Code-Switching	42
<b>9. Discussion</b>	<b>43</b>
<b>10. Conclusions and Future Research</b>	<b>44</b>
<b>11. References</b>	<b>45</b>
<b>Appendix. R-Code</b>	<b>51</b>
<b>List of Symbols, Abbreviations, and Acronyms</b>	<b>61</b>
<b>Distribution List</b>	<b>62</b>

## List of Figures

---

Fig. 1	Vector representation of each person who participates in a conversation .....	7
Fig. 2	Representation of each person in a subgroup using binary vectors .....	8
Fig. 3	Part of a conversation.....	12
Fig. 4	Flow diagram of the method .....	17
Fig. 5	GATE Developer GUI .....	30
Fig. 6	Activity word distributions .....	33
Fig. 7	Topic table .....	33
Fig. 8	Media .....	34
Fig. 9	Media message with person highlighted.....	34
Fig. 10	Map of Kenya .....	38
Fig. 11	Example of a formatted Kenyan Tweet .....	39
Fig. 12	Automatically extracted organizations for Mombasa by Contour .....	41
Fig. 13	Example of CS Tweets.....	43

## List of Tables

---

Table 1	Relation matrix of conversation in Fig. 3 .....	12
Table 2	Classification of the selected members of a conversation into 2 groups.....	15
Table 3	Confusion matrix for clustering by parts of speech .....	15
Table 4	Closest relation of 2 persons .....	16
Table 5	Comparison between the results of first and second fuzzy logic box..	18
Table 6	Comparison between the results of statistical-based method and fuzzy logic-based method .....	18
Table 7	Hierarchies of the 3 existing groups .....	19
Table 8	Computed ranking of Season 2 of <i>The Wire</i> by using coordination ....	20
Table 9	Computed ranking of Season 2 of <i>The Wire</i> by using asked question.	20
Table 10	Computed ranking of Season 2 of <i>The Wire</i> by using deontic modal verbs .....	20
Table 11	Computed ranking of Season 2 of <i>The Wire</i> by using epistemic modal verbs .....	21
Table 12	Computed ranking of Season 2 of <i>The Wire</i> by using Hedges .....	21
Table 13	Computed ranking of Season 2 of <i>The Wire</i> based on use of profanity .....	21
Table 14	Computed ranking of Season 2 of <i>The Wire</i> by using the terms of address.....	22
Table 15	Overall computed ranking of Season 2 of <i>The Wire</i> .....	22
Table 16	Errors computed by the hierarchy methodology.....	22
Table 17	Demographic spread of social network site users.....	24



## **Acknowledgments**

---

The authors also would like to acknowledge the contributions of William Mckinnon for developing the R code used to structure the data for analysis. We would also like to thank Andrew Hunyh who participated in the data collection effort.

INTENTIONALLY LEFT BLANK.

## 1. Introduction

---

This report documents the use of automated social network analysis to provide context for the sociolinguistic study of communication acts. Working with open source text documents, researchers from the US Army Research Laboratory (ARL) and Howard University (HU) explored how automated text analytics tools can accelerate the study of language. This collaboration leverages separate initiatives aimed at improving the interpretation of social intent as expressed in communication venues. The HU research is supported by the Army Research Office Partners in Research Transition Extracting Social Meaning from Linguistic Structures involving code-switching (CS) in English (and French) with Selected African Languages (ESCALES). The computational social network analysis software used for this analysis was developed under the Office of the Secretary of Defense (OSD) Small Business Innovative Research (SBIR) program in support of the ARL Tactical Information Fusion group. The goals of this collaboration revolve around the 3 following guiding questions that address sociolinguistics, social network analysis, and the use of automated tools for rapid text processing in support of military decision making:

- What do multilingual Africans communicate on a “social level” when they engage in CS as a linguistic strategy, and how can such knowledge aid military leaders as they interact with African leaders to support military goals?
- What are the linguistic features of code-switched language that operate as reliable predictors of social meaning, and how can these features be used as a conceptual construct to frame knowledge extraction for guiding course of action selection in military decision making?
- What computational techniques best leverage the predictive value of code-switched language for social meaning, and what decision support technologies can support rapid situational understanding for military leaders who are dealing with an immense variety of languages and diversity of social meaning dispersed across Africa?

These questions will be addressed from a theoretical and computational standpoint, in turn.

## **1.1 Sociolinguistic Theories and Social Meaning of Linguistic Structures in Communication**

---

Sociolinguistics, defined as the study of the relationship between language and society—how society influences language and how language practices influence society—has been of particular interests to linguists over the years (Fishman and Garcia 1972; Hymes 1974; Gumperz 1982). However, it is only recently, especially in the post-9/11 era, that the field, particularly language practices in multilingual communities, has attracted the attention of the military and intelligence agencies both at home and abroad. This report will focus on language practices involving 2 or more languages, broadly known in sociolinguistics as CS, in the African context. As indicated in the introduction, we want to know why multilingual Africans engage in CS, and how can such knowledge inform military policy-decision making, enhance military capabilities and situational awareness in humanitarian, peacekeeping, and military operations in the African context. To address these and the related questions raised in the introduction, we argue that it is extremely important that we understand the social context in which Africans use CS in the first place. We then discuss 2 computational tools, General Architecture for Text Engineering (GATE) and Contour and their applicability toward sociolinguistics. Furthermore, we describe how Contour was used to provide the context needed to understand the social meaning of CS in online text. We discuss the analytics provided by Contour and the process by which such information products can inform rapid situational understanding for military leaders.

## **1.2 Contextual Factors and Social Motivations for CS**

---

CS, the alternating use of 2 or more languages or varieties of a language in the same speech situation, has been one of the most researched topics in sociolinguistics over the past 50 years (Oksaar 1974; Blom and Gumperz 1972; Gumperz 1976; Jacobson 1990; Myers-Scotton 1993; Heller 2009; Wei 2013). Kamwangamalu (1999) offers a state-of-the-art report on this research at the dawn of the new millennium. He remarks that in the first known study of CS, based on Spanish-English CS in the United States, Espinosa (1917) claimed that there was no rationale for CS; that is, CS was just a random mixture of the languages available to a bilingual speaker. This claim has since been debunked (Gumperz 1982; Kamwangamalu 1990; Myers-Scotton 2002). There is overwhelming evidence that CS is socially meaningful. To understand the social meaning of CS, such as connotation of social hierarchy, subgroup structure, or power relationship in interaction, one must understand not only the linguistic structures of the languages involved, but also and more importantly, the social context in which CS is used. In this regard, John Gumperz, one of the pioneers in CS research, views CS as one kind of

“contextualization cue”. He says that contextualization cues are “constellations of surface features of message form ... by which speakers signal and listeners interpret what the activity is, how semantic content is to be understood and how each sentence relates to what precedes or follows [it]” (Gumperz 1982, p. 131). As a contextualization cue, CS “signals contextual information equivalent to what in monolingual settings is conveyed through prosody or other syntactic or lexical processes. It generates the presuppositions in terms of which the context of what is said is decoded” (Gumperz 1982, p. 98).

Gumperz's theoretical approach to CS—the interactional model—is mostly known for the distinction it makes between “situational CS” and “metaphorical CS”. A parallel distinction can be found in Oksaar (1974, p. 492), who uses the terms “external CS” and “internal CS”; or in Jacobson (1978), who distinguishes between “sociologically conditioned CS” and “psychologically conditioned CS”. Situational CS (i.e., external or sociologically conditioned CS) has to do with the social variables that may trigger CS, among them the setting where the interaction takes place; the participants in the interaction and the relationship they have with one another; and the topic of the interaction. A change in any of these variables will trigger CS. The bilingual's code choice is partly dependent on them. The demographic information we have collected on language practices in Kenya and South Africa, for instance, reveals a wide range of contextual variables around which the social meaning of CS can be interpreted. These include age, gender, socioeconomic class, social status, education level, profession, language attitude, social attributes and stereotypes, etc.

Metaphorical CS (i.e., internal or psychologically conditioned CS) concerns language factors, including the bilingual's attitude towards the codes at his/her disposal, his/her fluency in the languages involved, and his/her ability to use various emotive devices. These include what is commonly referred to as “discourse markers” (DMs). DMs (e.g., oh, well, but, or, so, because, now, then, you know, I mean) may coincide in form with adverbs, conjunctions, and phrases, but they differ in function. Specifically, DMs are syntactically independent: that is, removing a marker from a sentence still leaves the sentence structure intact. In other words, DMs are supplementary mechanisms for signaling social meaning, identity construction and display, and relational network. The meaning of DMs is situational; that is, the intentions of the speakers are situated “within an integral framework of interactionally emergent structures, meanings, and actions” (Schiffrin 1987, p. 22). For instance, in the following exchange, by using “well” as a DM, Sandy politely signals to Bob that her answer is negative.

Bob: Let's watch a movie tonight.

Sandy: Well, I have a Spanish test tomorrow.

DMs are present in all languages and occur in different modes and mediums of communication, including computer-mediated communication: text messages, Tweets, emails, and blogs. DMs are of particular importance/interest in bilingual communication because “one language can serve as ‘commentary’ on the other” (Maschler 1994, p. 325). For instance, as de Rooij (2000) shows, French-Swahili bilinguals use nearly exclusively French discourse markers (*donc*, *puisque*, *alors*, *et puis*) as a type of metaphorical CS. Bilinguals tend to use metaphorical CS when there is no change in the variables of the context of situation (e.g., setting, topic, participants). Gumperz (1982, p. 75–84) has identified the following among the communicative functions of metaphorical CS: quotation, addressee specification, interjection, and reiteration, to list some. (For a critique of the concepts of metaphorical versus situational CS, see Myers-Scotton [1993].)

Myers-Scotton (1993) has proposed an integrative model, the Markedness model, for the social interpretation of CS. Wei (2013) says this model has attempted to bring together the external and internal factors proposed by Gumperz (1982). Within the Markedness model, it is argued that bilinguals have an innate theory of socially relevant markedness and indexicality of the different languages and language varieties spoken in their community. This theory enables them to make rational choices in social interactions by calculating the costs and benefits of their actions. The main claim of this model is that all linguistic choices, including CS, are indices of social negotiations of rights and obligations existing between participants in a conversational exchange (e.g., Myers-Scotton 1993, p. 152–153). These rights and obligations are said to derive from whatever situational features are salient to the exchange, such as the status of the participants, the topic, and the setting. It is the interplay between these features and more dynamic, individual considerations that determines the linguistic choices that individuals make about media for conversational exchanges.

Within the Markedness Model, CS fulfils at least 2 main functions: it can be a “marked choice” or an “unmarked choice”. CS is an unmarked choice in the sense that it is “the expected code” in a given linguistic exchange. In this case, its use signals solidarity and in-group identity among the participants in a conversational exchange. In other words, CS as unmarked choice reflects the norms for language use in the given community, as illustrated in the following Swahili-English example, where the speaker describes how a car accident happened. In this particular example, the speaker uses CS not to convey any specific meaning, but because it is the medium through which educated speakers in particular communicate in everyday interaction.

He accident *ilitokea alipolose control na aka* overturn and landed into a ditch.

“The accident occurred when he lost control and overturned and landed into a ditch” (Mkilifi 1978).

On the other hand, CS can be used as a marked choice if it is the unexpected code choice in a given speech situation, as illustrated in the following interaction between a lecturer (Kamwangamalu) and a student in a lecture room at the University of Swaziland, Southern Africa. In this example, the lecturer negotiates a date for a test with his students, but the students are reluctant to write the test because they have other tests to study for. Not to challenge the lecturer openly, one student switches to siSwati in appealing to fellow students for support against writing the test. The student uses siSwati so that the lecturer, not a siSwati speaker, would not understand what the student is saying. Following the Markedness Model, the switch from English to siSwati is a marked choice, for siSwati is the least expected medium of communication in a University lecture, especially if the parties involved all do not share this language (siSwati).

**siSwati-English CS** (Kamwangamalu 1996, p. 299)

Lecturer: What if I gave you a short test tomorrow.

Students: No, sir, tomorrow we are writing a test in another course.

Lecturer: When do you think we can write it? We should definitely have one this week.

One student (turning to fellow classmates):

*Yeyi nine ningadli nivune kutsi siyibuale le-TEST. Onkhe maviki sibhala iTEST yakhe ingatsi ngiyo yodwa iCOURSE lesiyentakiko*  
(Translation: Hey, you! Never agree to write the test! Every week we write his test as if his is the only course we are taking this term.)

Lecturer (to the student who was addressing his classmates): What are you saying?

The Student: I'm saying, Sir, what if we write it next week. (The rest of the class laughs.)

The siSwati-English example supports the view, expressed in Heller (2009), that sociocultural values, power relations, and identity projection are all factors that can motivate CS as a marked choice. However, that in terms of the dichotomy it makes between marked choice and unmarked choice, the Markedness Model appears to be too static in its account of the pragmatic functions of CS in multilingual communities. Kamwangamalu (2000, p. 62) explains that in the above example, for instance, CS may agreeably be characterized as a marked choice in the sense that it

creates distance between the speaker and the lecturer, but at the same time, however, from the speaker's perspective CS also qualifies as an unmarked choice; it is intended to create solidarity between the speaker and his fellow students. What this means is that CS as a marked choice can be a double-edged sword: it can simultaneously exclude and include; it can create rapprochement and distance; much as it can reinforce the *we-ness* vs the *other-ness* among the participants in a linguistic exchange. Put differently and contrary to the Markedness Model, CS as a marked choice does not necessarily or always entail social distance among the participants. For instance, when US politicians use CS (as marked choice; switch from the expected code, English, to the unexpected one, Spanish) at public rallies or in formal meetings, their aim is not to distance themselves from their audiences or addressees. Rather, they use CS to create an opposite, no matter how symbolic, effect: rapprochement, oneness, and solidarity with their audiences. (For further critique of the Markedness Model, see Meeuwis and Blommaert [1994].)

Thus, following Bourdieu (1991), we argue that CS, especially when it is used as a marked choice, is both a communicative resource and linguistic capital. As such, CS is one of the powerful and potentially effective strategies that multilingual speakers have at their disposal and which they use to achieve a wide range of social goals. Some of these may be predetermined, such as exercising power over others and/or identifying with certain groups for political gains (e.g., votes). Others, such as confidentiality or secrecy, accommodation, or exclusion of others from a conversation, may not be predetermined; rather, multilingual speakers deploy CS to achieve them depending on the context of situation. The social context, then, undergirds the social interpretation of CS. It allows for identification of relationships such as the following among participants in a conversation: social hierarchy (who is the leader and who is the follower?); speech accommodation (who accommodates to whom and why and through what linguistic means?); social attitudes (how do participants—through their language practices—perceive one another socially?), and enables the speakers, as Myers-Scotton (2002, p. 44) puts it, to express themselves through a hybrid medium according to the sociolinguistic norms of their communities. The challenge for language engineering experts, therefore, is to design automated analysis procedures to extract the social meaning of CS, such as those just listed, from online text involving English and African languages (e.g., Swahili, Zulu). The following section describes some candidate approaches for addressing this challenge.

A team at HU has recently been experimenting with computational approaches involving statistical methods and fuzzy logic. The initial results appear promising and indicate that these methods are suitable for analyzing social meaning. The results that will be shown were obtained using transcripts of the TV show *The Wire*.



## 2. Computational Approaches to Defining Social Meaning

The 2 approaches that will be shown are a statistical method based on linguistic structures and a fuzzy logic method. The statistical method is based on counting linguistic features such as parts of speech and clustering similar features together. The fuzzy logic approach explores assigning numerical values to ambiguous linguistic features and determining membership. The last portion uses the statistical method to identify hierarchy between members in a text conversation.

### 2.1 Statistical Approach

The statistical approach to identification of large subgroups using parts of speech (POS) computes a histogram of the usage of POS for each speaker. To do this, Stanford Log-linear POS Tagger developed by Kristina Toutanova for syntax analysis was used. This gives a vector of 37 features—the number parts of speech recognized by the tagger. Each position of the vector represents the number of times that a person uses that particular POS element. Fig. 1 shows the representation of each person in the subgroup of people.

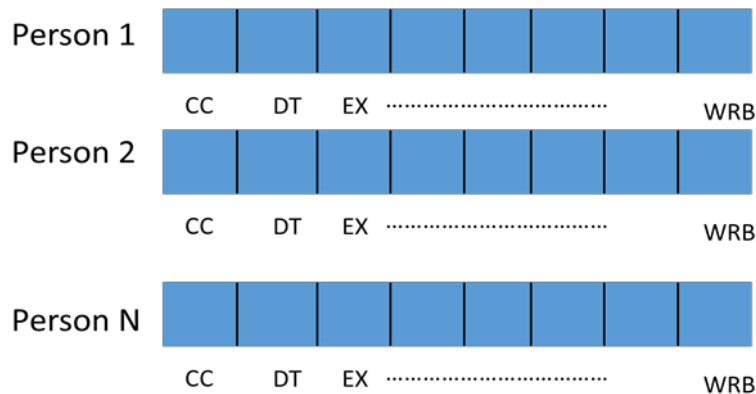
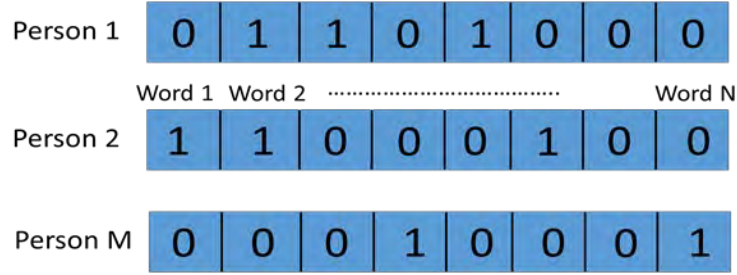


Fig. 1 Vector representation of each person who participates in a conversation

A clustering algorithm was then used to group similar vectors together. For the experiments the K-means method was used, which meant the number of clusters is specified for the desired output.

The approach that worked best for identifying small subgroups was to compare the vocabulary used by the speakers. A vector for each speaker was formed with equal length to the total vocabulary used by all the speakers. Each position of the vector corresponds to a word used in the text. The vector is binary so with a “1” in position “i” of speaker X’s vector if speaker X uses the word represented by position i at some point in his speech. Fig. 2 illustrates this approach.



**Fig. 2 Representation of each person in a subgroup using binary vectors**

To measure the closeness of several vectors, we simply add them up and divide by their dimension. We compute our similarity function  $D(.)$  in the following way:

$$D(X_1, X_2, \dots, X_N) = \frac{1}{K} * \sum_{j=1}^K X_1 + X_2 + \dots + X_N, \quad (1)$$

where  $N$  is the number of people in a subgroup and  $K$  is the dimensionality of the binary vector.  $D(.)$  takes a value from 0 to  $K$ , where 0 means no relation among the  $N$  members, and  $K$  means a maximum relation. Finally, the values of the function  $D(.)$  for all the combinations of  $N$  persons are sorted in descending order to identify the combinations of individuals who have the tightest relations or highest similarity.

## 2.2 Fuzzy Logic Approach

---

The fuzzy logic approach to determine groupings of individuals in written conversation is an underdeveloped area of study. This approach extracts features from conversations and determines, through fuzzy logic, the likelihood that individuals are in the same group. The approach then considers those who communicate with each other coupled with the previous results to increase the accuracy of the grouping. The fuzzy logic method allows for weight and values to be assigned to features displayed through written speech and who is in conversation. The approach relies on empirical data that is extracted from the written conversations then the empirical data is used to determine the grouping of each individual. The frequent use of features is considered amongst many individuals to pair them into groups. The counting of features provides a way to transition from qualitative space to quantitative space, which enables the measurement of the distance between characters and allows us to put them into different groups. Features are used as input into the fuzzy logic algorithm, which groups individuals in conversation based on the empirically used features. After the initial grouping is complete, the relationship between characters is considered to add

members to the initial grouping. Next, the relationships between characters are assigned values. The relationship values are combined with the feature values as inputs into the fuzzy logic algorithm which produces the grouping.

Numerical results are based on the data from the Home Box Office (HBO) television show, *The Wire*. The method analyzes several episodes of the Baltimore, MD-based series that consists of 3 main groups: the Baltimore police department and 2 organized crime groups. We will show that the approach results in a high classification accuracy. The proposed work seeks to identify groupings of individuals who are displaying similar features and their relationship with members of the group. The core members of the group are assumed to display the most features and have the highest frequencies during the extraction phase. Individuals who converse with members of a group are then considered for membership based on features that they individually display and their relationship between the individual and members of the particular group. A key point to get to the correct classification of members of different groups is a good set of features.

### **2.3 Feature Extraction**

---

Extracting features in this context amounts to identifying linguistic characteristics of individuals being examined. This is performed in 2 steps using the LightSide tool (Yen and Langari 1999): identifying the common characteristics of a group with feature vectors and then reducing the feature vectors to a minimum of independent characteristics. LightSide is an open-source text mining and machine learning tool that can extract frequency of word usage and parts of speech in order to predict membership in certain groups (Mayfield and Rose 2014). LightSide is used in this case to extract features that are frequently displayed by multiple individuals within the text. *The Wire* text data for 10 episodes were used and the goal is to classify the characters into gang, police, or informant. By informant, we mean that the person is contacting both the gang and the police, but he/she is a gang member or used to be a gang member.

To do that, we need to have some initial knowledge about each group and then extend our knowledge using our learning algorithm. This is similar to the way human beings learn; they try to find some connection between known and unknown. Toward that end, we marked 4 characters in each group with group affiliation to create the initial knowledge base. To extract those features, LightSide was used. It scans the speech of the characters that are labeled (as police or gang members) and extracts the corresponding POS. As a review, each part of speech refers to a category to which a word is assigned in accordance

with its syntactic function. In English, the main parts of speech are noun, pronoun, adjective, determiner, verb, adverb, preposition, conjunction, and interjection. For instance, the POS distribution in the sentence below is as follows:

This (determiner) student (noun) is (auxiliary verb) working (verb) on (preposition) an (determiner) interesting (adjective) project (noun).

The way the features are dealt with is general and they can be extended to more general cases because the method is not based on specific words but is based on the structure of the text. Now 2 categories of feature are gathered, the first one is the initial police POS features ( $F_{1P}$ ) and the second one is the initial gang POS features ( $F_{1G}$ ). Both groups can have some features in common, but filters are applied to reduce the features to the ones that are specific to each group making the features independent. Equation 2 shows that  $F_{1P}$  is the initial police POS features, and  $F_{1G}$  is the initial gang POS features, after reduction the feature space will change to  $F_P$  and  $F_G$ . The 2 feature spaces,  $F_{1P}$  and  $F_{1G}$ , may have some features in common, but  $F_P$  and  $F_G$  are fully independent feature vector space and orthogonal to each other.

$$\exists f_{1P} \in F_{1P}: f_{1P} \in F_{1G} \text{ or } F_{1P} \cap F_{1G} \neq \emptyset. \quad (2)$$

$$\exists f_{1G} \in F_{1G}: f_{1G} \in F_{1P} \text{ or } F_{1P} \cap F_{1G} \neq \emptyset. \quad (3)$$

$$\text{If } f_G \in F_G \rightarrow f_G \notin F_P \text{ or } F_P \cap F_G = \emptyset. \quad (4)$$

$$\text{If } f_P \in F_P \rightarrow f_P \notin F_G \text{ or } F_P \cap F_G = \emptyset. \quad (5)$$

The police POS features and gang member POS features are characterized by separate groups. Next we look at those characters whose affiliation is unknown. In other words, extend the algorithm throughout the whole text and get some information about the unknown characters. The unknown characters are the ones with undetermined group affiliation. The labels are removed from the characters that were labeled previously, and the text is run through LightSide, which will extract all possible POS feature for each characters. Each character might have thousands of POS features, but among them only the ones common with  $F_P$  and  $F_G$  are needed. As a result, the feature space is reduced.

At the next step, we assigned 2 values for each character; one for the number of times they show  $F_P$  and the other for the number of times they show  $F_G$ . Each person is assigned 2 values of  $A$  and  $B$  in which,

$A$  = Number of times character shows  $F_G$

$B$  = Number of times character shows  $F_P$

Equations 6 and 7 show the ratio that each character shows  $F_G$  and  $F_P$ .

$$a = \frac{A}{A+B}. \quad (6)$$

$$b = \frac{B}{A+B}. \quad (7)$$

## 2.4 Relationship Determination

---

The objective of our approach is to explore the relationships of individuals in conversation and use this as another of other factors in finding characters of the same group. The approach used to examine relationships is developed out of the need to determine who is communicating with whom and assign values to those in conversations together. Values are assigned to each individual in a conversation by taking the previous individuals in the conversation and those who follow. There are 2 methods to determine the values: assign a true or false value to those in conversation or to simply count the number of times that there is a back and forth in the conversation. A vector was created for each individual in the conversations. The vector consists of  $N$  columns for each of the  $N$  individuals in the total text. The approach takes the individual currently talking and assigns a 1 to the individuals directly before and after. This would mean that the characters are in direct conversation according to the text. The algorithm also assigns the value 0.5 to the person who precedes and follows the individual within a distance of 2 (indirect contact).

Suppose the part of the conversation is in the order of Figure 3. Consider Person 1, as connected to person 2 directly one time and indirectly one time so a value of 1.5 is assigned for the relation between Person 1 and Person 2. The same can be done for the relation between Person 1 and 3; they are connected 3 times directly, which means they are repeated right after each other 3 times and they are connected indirectly one time which will assign the value of 0.5 to their connection. The total value of relation between Person 1 and Person 3 would be 3.5. Without seeing the individuals who are in conversation this algorithm yields insight as to those who are related simply by when they speak. The next step is to aggregate the results of the line by line conversation and create a vector for each person that represents the relationships developed throughout the text. In Table 1 the relation matrix related to Figure 3 is shown; a higher value in the matrix is considered as a higher weight in the relationship between the 2 individuals. The relation matrix is named  $R$  and is an  $N \times N$  matrix, which is a symmetric matrix.

Conversation
Person 1
Person 2
Person 3
Person 1
Person 3
Person 1
Person 4
Person 1

**Fig. 3 Part of a conversation**

**Table 1 Relation matrix of conversation in Fig. 3**

Conversation	Person 1	Person 2	Person 3	Person 4
Person 1	...	$1 + 0.5 = 1.5$	$0.5 + 1 + 1 + 1 = 3.5$	$1 + 1 = 1$
Person 2	$1 + 0.5 = 1.5$	...	1	...
Person 3	$0.5 + 1 + 1 + 1 = 3.5$	1	...	0.5
Person 4	$1 + 1 = 2$	...	0.5	...

## 2.5 Fuzzification

Fuzzy logic is the tool used to deal with the crisp numbers of Eqs. 6 and 7 so that the numbers are made into some interpretable values. The fuzzy logic algorithm used is based on C-mean fuzzy logic clustering algorithm (Mayfield et al., 2014). Using 3 different groups (police, gang, and informant) and fuzzy logic approach helps to assign membership values for each person for each group. After implementing the first fuzzy logic box, a fuzzy box is used with the aid of relationship matrix  $R$ .

This report attempts to use the fuzzy C-Means algorithm, but it is a little different with C-Means method. The difference is that it is not iterative and the center of each cluster (group) is known, so there is no need to iterate. The C-Means algorithm is a method to calculate the degree of membership for each person for each group. The basic part in fuzzy logic is the rules and this algorithm helps us to reduce the rules in a great deal, which leads to a faster process. We have 3 major rules in our 2 fuzzy logic boxes that are as follows;

- 1) If  $a = 1$  and  $b = 0$  and  $d = A - B = 15$  character is gang.
- 2) If  $a = 0$  and  $b = 1$  and  $d = A - B = -15$  character is police.
- 3) If  $a = 0.5$  and  $b=0.5$  and  $d = A - B = 0$  character is informant.

Different characters have different  $a$ ,  $b$ , and  $d$  values. If their values are closer to each of the above rules, they will have a higher membership value of that group.

As previously stated, individuals in this case study (*The Wire*) are grouped into 3 categories: the police, the gang, and the informant. Further, according to Eq. 8, the algorithm developed allows for the degree of membership of each character to be calculated for the represented groups (see Fig. 1).

$$\mu_{G_i}(x) = \frac{1}{\sum_{j=1}^3 \frac{\|x-F_i\|^2}{\|x-F_j\|^2}} \quad 1 \leq i \leq 3, x \in X, \quad (8)$$

$G_i$  is one of the 3 groups, in this study, where  $G_1$  is for Gang,  $G_2$  is for police, and  $G_3$  is for the informant case.  $F_i$  is the center of each group in which the values are assigned based on empirical observation from dataset. Knowing that there are 3 groups implies that there will be 3 centers of the groups  $F_1$ ,  $F_2$ , and  $F_3$ . From the values extracted from the labeled characters feature, the appropriate values for  $F_1$ ,  $F_2$ , and  $F_3$  were identified.  $X$  is the dataset, and  $x$  is one member in the dataset. In other words,  $x$  is the normalized feature vector that yields some values for each character to compare them together and see how close they are to each other or how far they are from the center of each group ( $F_1$ ,  $F_2$ ,  $F_3$ ). Each character has a normalized vector  $x$ . By this equation the membership values of each character are determined for each of the 3 groups. Indeed, 3 numbers are determined for each character, which shows the membership of the character in the gang, police, and informant group. Since we performed our fuzzy logic algorithm in 2 steps, we have different membership parameters. For the first fuzzy logic box the membership values are named  $\mu_{G_i1}$ , and for the second fuzzy logic box the membership values are named  $\mu_{G_i2}$ , in which  $i = 1,2,3$  refers to gang, police, and informant, respectively. This algorithm reduces the overlearning process and central processing unit processing time since this clustering (grouping) method will reduce our rules drastically.

## 2.6 Statistical Identification of Hierarchy

---

To determine the existing hierarchy in a group of persons, the coordination concept is used, which is based on the observation that when conversing, people unconsciously adapt to one another's communicative behaviors. Additionally, power differentials in a conversation are revealed by how much one individual immediately echoes the linguistic style of the person they are responding to. That is, the individuals with less power in a conversation tend to adapt their linguistic style of those with higher rank. An example of coordination is shown in the following dialogue:

Person A: At what time does your store close?

Person B: At 5 o'clock.

Coordination occurs in the previous conversation when “B” adapts the preposition “At” to answer the question formulated by “A”. Based on the work of Danescu-Niculescu-Mizil et al. (2002), we derive the coordination value by calculating 2 terms: the probability of person B using a marker in response to person A when person A uses the marker, and the probability of person B using a marker in response to person A. The mathematical expression of coordination is given by the following equation:

$$C(B,A) = P(\epsilon_{u2 \rightarrow u1}^m | \epsilon_{u1}^m) - P(\epsilon_{u2 \rightarrow u1}^m). \quad (9)$$

The greater value of coordination, the more powerful a person is. The following discourse markers were used in the calculation: “I mean”, “you know”, “well”, “like”, “so”, and took the average coordination from all the markers as the measure of power.

Moreover, our analysis also considered other linguistic strategies that are typically associated with power in conversation, for example, less powerful people tend to use more hedges, such as “sort of”, “kind of”. Additionally, these persons may use less profanity, which is often dubbed as “strong language”. Conversely, more powerful people formulate more questions, while those less powerful are obligated to answer the questions. Also, more powerful people use deontic and epistemic modal verbs that require action (“must” and “have”), while less powerful use modal verbs that suggest action (“should”, “would”, “could”). Another characteristic is that more powerful people use informal address forms to address others but are typically addressed in a more respectful/polite way. In the languages with “tu/vous” distinction (such as French), the more powerful use “tu” to address others, while less powerful use “vous” to address the more powerful. In English, this distinction can be manifested in a nonreciprocal use of the titles and first names: the powerful can address less powerful using their first names, while the less powerful use titles (Mr, Sir, Ma’am, etc.) and last names to address those in power.

These 7 characteristics (i.e., average value of coordination, number of formulated questions, use of deontic and epistemic modal verbs, number of hedge, use of profanity, and number of terms of address [Mr, Miss, Sir, Ma’am]), are used to create 7 lists to sort in descending order all the participants in a conversation. Then, we calculated each person’s average ranking on each characteristic to create an overall ranking.



## 2.7 Grouping Results of the 2 Methods

---

We tested this method on data collected from episodes of the HBO television show, *The Wire*. Concretely, we used 3 episodes from season 1 and 7 episodes from season 2. There are 2 major groups of characters: the Baltimore police department and a drug dealing organization run by the Barksdale family.

Table 2 shows the results of clustering the characters into 2 groups based on POS.

**Table 2** Classification of the selected members of a conversation into 2 groups

First Group	Second Group
McNulty	Nick
Bunk	Stringer
Sobotka	Ziggy
Freamon	Avon
Daniels	Greggs
Russell	Dee
Herc	Omar
Prez	Carver
Landsman	Valchek
Levy	Spiros
Pearlman	Elena
Rawls	Bodie
	Horseface

The first group consists mainly of police officers with the only character classified in this group who is not a police officer is Mr Sobotka. During the TV show, he makes arrangements with European gangsters to smuggle illegal goods through Baltimore's port. The second group consists mostly of persons involved with criminal activities, (drug dealers, smugglers, etc.). However, 4 characters who are police officers were misclassified in the second group. The confusion matrix is shown below in Table 3.

**Table 3** Confusion matrix for clustering by parts of speech

	First Class	Second Class	Accuracy
First Class	11	1	91.7%
Second Class	4	9	69.2%
	73.3%	90.0%	...

The confusion matrix shows that the proposed methodology has an accuracy of 91.7% for the first class and 69.2% for the second class. Moreover, the overall accuracy is 80%. On the other hand, the results show a reliability of 73.3% and 90.0% for the first and second class, respectively. This indicates that this method is effective in distinguishing between these 2 major groups of characters purely based on their used of different parts of speech.

On the other hand, the results showing the 5 closest pairs in each of the 2 detected subgroups are shown in the Table 4.

**Table 4** Closest relation of 2 persons

<b>Rank</b>	<b>First Class</b>		<b>Second Class</b>	
1	McNulty	Bunk	Nick	Ziggy
2	McNulty	Sobotka	Nick	Stringer
3	McNulty	Freamon	Stringer	Avon
4	McNulty	Daniels	Nick	Avon
5	Bunk	Sobotka	Nick	Dee

These results were obtained by the methodology described earlier. This methodology allows us to identify the closest relations among members of small homogeneous subgroups. For example, in Table 3 the most significant relation in the first detected group is the one between Mr McNulty and Mr Bunk, 2 detectives who are close friends and partners in the Baltimore Police Department. On the other hand, the closest relation in the second class detected the familiar link between Nick and Ziggy, who are cousins. Nick shows considerable patience to his cousin, whom he often has to keep out of trouble. However, Nick is much more cautious and level-headed than Ziggy. Other important relations shown in Table 3 are the ones between Mr McNulty with his fellow police officers, Freamon and Daniels. Likewise, among the identified relations in the second group is the one between Avon and his childhood friend Stringer. Avon runs a criminal organization located in West Baltimore with total autonomy, and he is assisted by Stringer who is responsible for the economics of the criminal organization as Avon's second-in-command.

The overall flow diagram of the fuzzy logic method is shown in Fig. 4. Elements that are inside the dashed line make up our black box, which processes the input data and gives the output as the degree of membership of being gang, police, and informant for each character.

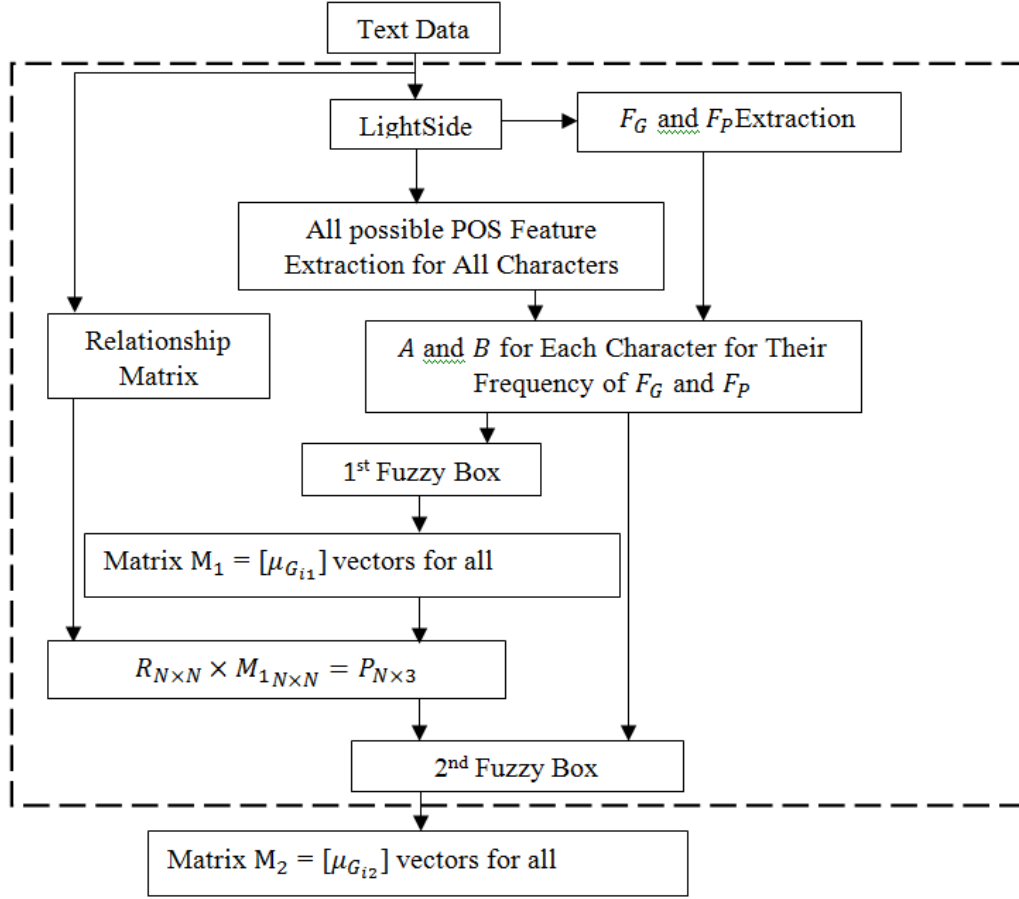


Fig. 4 Flow diagram of the method

As was noted in Section 2.5, for each character there will be 3 values for  $\mu_{Gi1}$  and 3 values for  $\mu_{Gi2}$  in which

$$\mu_{G_{11}} + \mu_{G_{21}} + \mu_{G_{31}} = 1 \quad \text{and} \quad \mu_{G_{12}} + \mu_{G_{22}} + \mu_{G_{32}} = 1. \quad (10)$$

The  $\mu_{Gi1}$  values for all characters are collected in one matrix and called  $M_1$ , which is the same for  $\mu_{Gi2}$  and called  $M_2$ . By considering the total number of characters as  $N$ ,  $M_1$  and  $M_2$  are shown in Eq. 11.

$$M_1 = \begin{bmatrix} \mu_{G_{11}}^1 & \mu_{G_{21}}^1 & \mu_{G_{31}}^1 \\ \mu_{G_{11}}^2 & \mu_{G_{21}}^2 & \mu_{G_{31}}^2 \\ \vdots & \vdots & \vdots \\ \mu_{G_{11}}^N & \mu_{G_{21}}^N & \mu_{G_{31}}^N \end{bmatrix}, \quad M_2 = \begin{bmatrix} \mu_{G_{12}}^1 & \mu_{G_{22}}^1 & \mu_{G_{32}}^1 \\ \mu_{G_{12}}^2 & \mu_{G_{22}}^2 & \mu_{G_{32}}^2 \\ \vdots & \vdots & \vdots \\ \mu_{G_{12}}^N & \mu_{G_{22}}^N & \mu_{G_{32}}^N \end{bmatrix}. \quad (11)$$

The characters that are going to be dealt with are the main characters in the second 10 episodes of *The Wire*. The output result after applying the first and second fuzzy box to the characters and their text are according to Table 2. As can be seen in

Table 5, the first fuzzy box was not successful in extracting the exact rule for some characters. It was after using the relation matrix and the second fuzzy box that we were able to classify each character as gang, police, or informant (Ramirez et al. forthcoming). It can be seen that the K-means method is able to extract rules with 85% accuracy. Table 6 compares our output results with the K-mean statistical based method.

**Table 5 Comparison between the results of first and second fuzzy logic box**

Characters	Results After First Fuzzy Box				Results After Second Fuzzy Box			
	$\mu_{G11}$	$\mu_{G21}$	$\mu_{G31}$	Accuracy	$\mu_{G12}$	$\mu_{G22}$	$\mu_{G32}$	Accuracy
Bay	0.14	0.03	0.81	✓	0.81	0.07	0.10	✓
Avon	1	0	0	✓	0.98	0.00	0.00	✓
Stringer	1	0	0	✓	0.96	0.01	0.01	✓
Phelan	0.00	0.99	0.00	✓	0.02	0.82	0.15	✓
McNulty	0	1	0	✓	0.01	0.96	0.02	✓
Pearlman	0.00	0.99	0.00	✓	0.00	0.96	0.02	✓
Carver	0.00	0.99	0.00	✓	0.02	0.82	0.15	✓
Freamon	0.00	0.99	0.00	✓	0.01	0.94	0.04	✓
Greggs	0.00	0.99	0.00	✓	0.02	0.83	0.14	✓
Dee	0.17	0.17	0.65	✓	0.00	0.00	0.99	✓
Omar	0.00	0.76	0.22	✗	0.03	0.33	0.63	✓
Bunk	0	1	0	✓	0.00	0.96	0.02	✓
Norris	0.00	0.01	0.98	✗	0.00	0.98	0.01	✓
Daniels	0	1	0	✓	0.00	0.96	0.02	✓
Landsman	0.00	0.99	0.00	✓	0.01	0.94	0.03	✓
Prez	0.00	0.99	0.00	✓	0.01	0.90	0.07	✓
Burrel	0.00	0.99	0.00	✓	0.01	0.90	0.07	✓
Russel	0	1	0	✓	0.00	0.98	0.00	✓
Nick	1	0	0	✓	0.97	0.01	0.01	✓
Sobotka	1	0	0	✓	0.89	0.04	0.06	✓

**Table 6 Comparison between the results of statistical-based method and fuzzy logic-based method**

Character	K-Mean Statistical Based Method		Fuzzy Logic Based Method	
	Accuracy	Overall accuracy	Accuracy	Overall accuracy
Avon	✓	85%	✓	100%
Stringer	✓			
McNulty	✓			
Carver	✓			
Freamon	✓			
Greggs	✗			
Dee	✓			
Omar	✓			
Bunk	✓			
Daniels	✓			
Russel	✓			
Nick	✓			
Sobotka	✗			
Ziggy	✓			

To identify the hierarchy, we analyze 3 distinct groups of people present in *The Wire*: the police officers, and 2 criminal organizations. One of the criminal groups is the Barksdale organization, which is led by Avon Barksdale and Stringer Bell. The other criminal group is the Sobotka family headed by Frank Sobotka, a

treasurer for the local union at the Baltimore docks who is also involved along with his family in smuggling illegal goods through the port. Thus, the Sobotka family not only has extensive connections to the Baltimore port, but also links to the criminal underworld. The hierarchies of the main members that constitute the 3 existing groups are shown in Table 7.

**Table 7 Hierarchies of the 3 existing groups**

<b>Police Officers</b>	<b>Barksdale Organization (Criminals)</b>	<b>Sobotka Family (Docks)</b>
1. Daniels (Deputy Commissioner)	1. Avon (Kingpin)	1. Sobotka (Head of the family)
2. Freamon (Detective)	2. Stringer (Kingpin)	2. Nick (Sobotka’s Nephew)
3. McNulty (Detective)	3. Bey (Soldier)	3. Ziggy (Sobotka’s Son)
4. Bunk (Detective)	4. D’angelo (Dealer)	...
5. Greggs (Detective)	5. Bodie (Dealer)	...
6. Carver (Detective)	6. Poot (Dealer)	...
7. Russell (Port Authority Police Officer)	...	...

Table 7 shows the main members of the distinct groups. To automatically detect the social hierarchy we use the methodology described previously that takes into account the following features: average value of coordination, number of formulated questions, use of modal verbs, number of hedges, use of profanity, and number of terms of address. To determine the quality of our ranking we calculate the squared difference of the ranking and the actual ranking as shown in Table 3 using the Spearman ranking correlation.

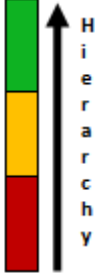
$$e = \frac{3}{n^3 - n} \sum_{k=1}^n (R_k - E_k)^2, n > 1. \quad (12)$$

Where  $n$  is the number of members of the group being analyzed and the parameters,  $R_k$  and  $E_k$  are the actual ranking and the computed ranking, respectively. The parameter  $e$ , called ranking error, takes values that range from 0 to 1, where a value of 0 means no error in the ranking, and a value of 1 is the maximum error.

Tables 8–14 show the computed rankings for each one of the 7 implemented strategies. Additionally, Table 15 shows the overall ranking of the members in the 3 existing groups after averaging the positions of each member from Tables 8–14. The ranking errors registered for the 7 implemented strategies and the overall ranking are shown in Table 16.

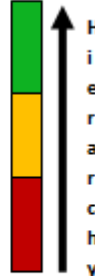
**Table 8** Computed ranking of Season 2 of *The Wire* by using coordination

Police	Criminals	Docks
Freamon	Avon	Ziggy
Daniels	Stringer	Sobotka
McNulty	D'angelo	Nick
Bunk	Bodie	
Russell	Bey	
Carver	Poot	
Greggs		




**Table 9** Computed ranking of Season 2 of *The Wire* by using asked question

Police	Criminals	Docks
Russell	Bey	Ziggy
McNulty	Bodie	Nick
Bunk	Avon	Sobotka
Freamon	D'angelo	
Greggs	Stringer	
Daniels	Poot	
Carver		



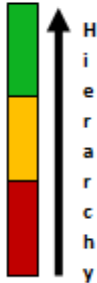
**Table 10** Computed ranking of Season 2 of *The Wire* by using deontic modal verbs

Police	Criminals	Docks
Russell	Stringer	Sobotka
McNulty	Bey	Ziggy
Daniels	D'angelo	Nick
Freamon	Avon	
Bunk	Bodie	
Carver	Poot	
Greggs		



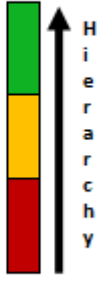
**Table 11** Computed ranking of Season 2 of *The Wire* by using epistemic modal verbs

Police	Criminals	Docks
Greggs	Poot	Nick
Bunk	Stringer	Ziggy
Freamon	Avon	Sobotka
McNulty	Bodie	
Carver	D'angelo	
Russell	Bey	
Daniels		



**Table 12** Computed ranking of Season 2 of *The Wire* by using Hedges

Police	Criminals	Docks
Russell	Bodie	Nick
Greggs	D'angelo	Sobotka
McNulty	Bey	Ziggy
Daniels	Poot	
Bunk	Stringer	
Freamon	Avon	
Carver		



**Table 13** Computed ranking of Season 2 of *The Wire* based on use of profanity

Police	Criminals	Docks
Bunk	Bey	Ziggy
Carver	Poot	Sobotka
Freamon	Bodie	Nick
McNulty	Avon	
Greggs	D'angelo	
Daniels	Stringer	
Russell		



**Table 14** Computed ranking of Season 2 of *The Wire* by using the terms of address

Police	Criminals	Docks	
Greggs	Poot	Nick	
Daniels	Bey	Ziggy	
McNulty	D'angelo	Sobotka	
Russell	Bodie		
Bunk	Avon		
Freamon	Stringer		
Carver			

**Table 15** Overall computed ranking of Season 2 of *The Wire*

Police	Criminals	Docks	
Freamon	Avon	Ziggy	
Daniels	Stringer	Sobotka	
McNulty	Bodie	Nick	
Bunk	D'angelo		
Russell	Bey		
Carver	Poot		
Greggs			

**Table 16** Errors computed by the hierarchy methodology

Strategy	Coordination	Questions	Deontic Modal Verbs	Epistemic Modal Verbs	Hedges	Profanity	Terms of Address	Overall Ranking
Police	0.089	0.607	0.536	0.536	0.643	0.464	0.393	0.089
Error Criminals	0.157	0.45	0.171	0.443	0.829	0.371	0.857	0.229
Docks	0.75	1	0.25	0.625	0.25	0.75	0.75	0.75
Average Error	0.332	0.687	0.319	0.535	0.574	0.528	0.667	0.355

The results in Table 16 show that the use of deontic modal verbs and coordination give the lowest errors in the hierarchy of the 3 existing groups of persons in the dataset. In contrast, the use of terms of address and number of asked question give the poorest results. On the other hand, the overall ranking achieved by averaging the positions from the 7 strategies (Tables 8–14) gives the third lowest error. Additionally, the hierarchy obtained by the overall ranking registered in Table 15



shows that the heads of the criminal organization run by the Barksdale family were successfully identified. Also, the hierarchy of the police officer presents a low error and good results. Conversely, the high error obtained for identifying the hierarchy of the Sobotka organization (docks) using the overall ranking strategy is because of the small number of members that constitute this organization (3 in total), and so even just one mistake in the classification considerably increases the total error.

### **3. Informing Context through Social Network Analysis**

---

Understanding the context in which CS is occurring is essential to interpreting the social meaning in the communication act under study. Social network analysis is one computational method with utility for that purpose. Contextual understanding includes the ability to perceive individuals who are engaged in shared communications, topics of discussion, sentiment expressed on those topics, objects generated through exchanges (ideas, plans, concepts, etc.), geographic locations of importance to the discussions, and dynamic changes in the network over time. Together, these factors help to define the “who, what, where, when, why, and how” for a time period under study.

For situations that are not familiar to the person trying to determine the context in which discussions are occurring, text extraction and analysis tools can provide deep insight into attitudes, opinions, behaviors, and actions of relevant actors in a social network. While these analytics have been developing over time, the recent burst of activity in social networking sites around the globe has provided an opportunity to leverage these communications in ways not imagined even 5 years ago. With the pervasiveness of mobile computing technologies, people are Tweeting, facebooking, you-tubing, yelping (and more) a variety of ordinary and not-so-ordinary messages. Just 4 years ago, in January 2011, protests organized through social networking sites toppled governments in Tunisia (14 January 2011), Egypt (11 February 2011), and Libya (protests began on 15 February 2011, and led to civil war and the overthrow of the government on 23 August 2011). In Syria, protests began on 26 January 2011 and continue today. In Yemen, protests began on 3 February 2011 and the government transferred power on 23 November of the same year. Around the Middle East, protests and uprisings have used social networking sites to organize protesters, notably in Algeria, Iraq, Jordan, Kuwait, Morocco, and Oman (National Public Radio 2011). The utility of social networking sites for organizing political protests was not limited to that region of the globe as evidenced by the rapid spread of the Occupy Wall Street movement. These protests against social and economic inequality began in New York City on 17 September 2011 and quickly spread to over 951 cities across 82 countries and over 600 communities in the United States (Thompson 2011).

While early social networking adopters may have been young and affluent, current demographics for users of social networking sites, reported by the Pew Research Center (2014), show solid market penetration in most age, gender, educational, and income groups. As shown in Table 17, men and women are nearly equal in their use of social networking sites, and while users younger than 50 are represented in higher numbers, older adults are quite actively represented. Of note, educational and income differences do not appear among users, demonstrating the ubiquity of the media application to everyday life.

**Table 17 Demographic spread of social network site users**

<b>Who uses social networking sites</b>	
<i>% of internet users within each group who use social networking sites</i>	
All internet users	74%
a Men	72
b Women	76
a 18-29	89 <sup>cd</sup>
b 30-49	82 <sup>cd</sup>
c 50-64	65 <sup>d</sup>
d 65+	49
a High school grad or less	72
b Some college	78
c College+	73
a Less than \$30,000/yr	79
b \$30,000-\$49,999	73
c \$50,000-\$74,999	70
d \$75,000+	78

Pew Research Center's Internet Project January Omnibus Survey, January 23-26, 2014.  
 Note: Percentages marked with a superscript letter (e.g., <sup>a</sup>) indicate a statistically significant difference between that row and the row designated by that superscript letter, among categories of each demographic characteristic (e.g., age).

**PEW RESEARCH CENTER**

The growth and popularity of online social media applications has become a rich data collection and testing ground for researchers interested in a wide variety of social network topics. Social media is defined as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content” (Kaplan and Haenlein 2009). Social media platforms are also being used across mainstream business, government, and military organizations to maintain strong connections with, and build, a customer base. In military settings, social media can be used to interact with community groups, influence crowds, deliver information, and increase unity of effort in operations (Serena et al. in press).

These applications are increasingly accessed via mobile technologies and are changing the way individuals and groups communicate. The variety of communication venues includes magazines, Internet forums, weblogs, social blogs, microblogging, wikis, podcasts, photographs or pictures, video, rating, and social bookmarking.

Social network analysis is an important methodology for military tasks; approaches are being developed for automatic network extraction, topic identification, topic-author and topic-person pairing, sentiment analysis, and identification of cultural artifacts and cultural norms (Bowman 2012a, 2012b). These decision support tools can provide rapid understanding of complex sociocultural environments for military leaders in uncertain conditions.

A further component of social network analysis closely related to sentiment analysis is topic or concept extraction. This approach can be used to cluster actors around shared interests or to discover relationships among individuals and topics. As an example of how topic extraction and co-referencing can be useful in sensing environmental and social changes, we might consider the case of an increasingly unstable nation. One might expect fluctuations in political, social, religious, and economic conditions, leaders, influencers, and protagonists. Without a cadre of experts on the ground, such intelligence would be difficult to gather. Topic extraction and relationship linking tools can provide a monitoring capability to detect changes in established trends based on newspaper reports and text blogs. Such a capability could be used to monitor the population's expressed support for one leader or another, or for proposed military or diplomatic actions. The Arab Spring protests provide a good case for content and topic extraction as a useful adjunct to existing persistent sensing assets. The ability to identify major topics motivating human action from the "first person" perspective is important. The cultural and social components of action are so complex that it is impossible for an outsider to fully comprehend the range of motivations behind an action (Bowman 2012a). Indeed, culture is so implicit that insiders often have difficulty perceiving, understanding, and articulating the motivation in full. It is in this regard that CS is potentially so vital to the deep understanding of communication acts for situational understanding and accurate forecasting of future events.

Social network analysis applications have dramatically increased with the rapid development of Internet sites that allow information sharing of many types. Twitter is a popular social networking and microblogging service that enables users to send and read short 140-character messages, called 'Tweets' (Stieglitz and Dang-Xuan, 2012). Though Twitter is considered a social networking site, it is more of an information network where people express their opinions, provide brief thoughts on items that are of interest to them, and share information for a variety of reasons

(Bravo-Marquez et al. 2012). Tweets may cover a wide range of topics with obvious utility (e.g., the sharing of news or conversations) or with little obvious value. In that latter category, researchers estimate that up to 40% of Twitter users post messages that are ‘pointless babble’ (e.g., “I am eating a sandwich now.”) (Pear Analytics, 2009). Of particular interest to sociolinguistics and sociologists, some researchers have argued that Tweets appearing to be pointless babble are better characterized as “social grooming and/or peripheral awareness” (Holton et al. 2014), where individuals gauge opinions, feelings, and activities of the wider social group when they are not physically present with their peers. Following Bell’s (1984) theory of “Audience Design”, Kamwangamalu (2001) says that individuals who are not physically present when their peers engage in communication acts can be described as “referees”. Bell explains that while in every interaction there is a second person whom the speaker directly addresses—the addressee—in some instances, there may also be third parties who, though not physically present, are actually ratified participants of the interaction. These third parties are what Bell terms referee (Kamwangamalu 2001, p. 90). They sometimes possess such salience for a speaker that they influence his/her linguistic behaviors or practices even in their absence. This influence can be so great that the speaker diverges from the addressee and converges toward these third parties. As would be expected in a social network application, Twitter users can be clustered into 3 groups: organizations, journalists/media bloggers, and ordinary individuals (De Choudhury, et al. 2012). Each of these user types can contribute to a societal need; organizations may coordinate social welfare services, journalists/media bloggers may synthesize and amplify information, and individuals may report views, observations, and opinions on any topic of interest.

Though Twitter and other microblogging sites serve a social networking function, Twitter user demographics have a somewhat different demographic profile than the social network sites as described previously. For the 2014 time period, the Pew Research Center survey reported that 55% of respondents under the age of 50 were Twitter users, compared with only 16% over 50 (Pew Research Center 2014). This should signal a cautionary note to researchers who increasingly use Twitter as a source of text for contextual understanding; sampled populations are not representative to the wider population as in the case of some social networking sites. Indeed, some have estimated that 5% percent of Twitter users account for 75% of all activity and that New York City has more Twitter users than other metropolitan areas (Cheng and Evans, 2009).

Perhaps due to the ease by which Tweets can be collected and processed, research using Tweet collections proliferate in the scholarly press, online journals, and open source Internet. The global prevalence of social media services such as Twitter and

their growing significance to the evolution of events has attracted the attention of many agencies from humanitarian nongovernment organizations and disaster response agencies to homeland security and counter-terrorism. Tweet collections have been used to study critical events such as the Arab Spring and the Syrian civil war (Comminos 2011; Lynch et al. 2014).

In general, social media services have changed the way information is communicated to the public by providing local community and international coverage of events, and, assisting governments in optimizing execution of missions in response to political and social change. For example, Kase et al. (2014) utilized a dataset of Tweets collected 2 days following the Syrian sarin gas attack (22–23 August 2013) and a maximum likelihood estimation approach to identify which Tweet sources propagated specific claims. Sentiment-specific probability distributions and networks of opposing polarities were extracted from the top source/claim Tweet cascades. Tyshchuk et al. (2014) used this same Tweet collection to investigate the formation of Twitter communities and their characteristics such as types of leadership and their roles, most prominent event and sub-event mentions, and top 10 verbs used in Tweet content.

#### **4. Challenges in Extracting Meaning from Social Media and Networks**

---

---

While the acceptance of social media mining for social network analysis may have some benefits to informing diplomatic and military applications, there are many challenges yet to overcome. These problems are compounded when we attempt to detect, translate, and comprehend CS events in social media. We now consider a range of technical issues for both social network analysis and CS.

In a United States Institute of Peace seminar to address the emerging role of social media in times of conflict, several challenges were identified (Omestad 2011). First, it was acknowledged that the social media networks produce extremely large amounts of data generated across a range of applications. Collecting, processing, and exploiting the relevant factors of information is a significant hurdle in the best of computing environments. Doing this in a tactical and limited bandwidth environment is exponentially complicated. Second, the complexity of the data presents analytic challenges. Social media applications use a variety of terms or features to indicate support of an idea, and aggregating across these will be difficult. Sentiment analysis can be particularly difficult, for example, when irony is used (e.g., “Well”, I like “that!”). Added to this complexity, Lakshmanan and Oberhofer (2010) suggest that the unpredictability and frequency of change in social media content will pose difficulty for data mining efforts. They note that some blogs

change very rapidly, both in terms of quantity of content but also in context, as contributors' opinions, personal tastes, and beliefs change. A third challenge is the propensity of social media users to employ language shortcuts, idioms, dialects, and alphabets (e.g., using 'k' instead of 'ok') that confuse machine processing algorithms. Extracting meaning from text shorthand using automated methods presents a problem when using existing ontology frameworks but could lead to simplified dictionaries and processing methods. Similarly, because social media are designed for specific purposes, off-the-shelf solutions for observation, analysis, and understanding may be insufficient for military applications (Serena et al. in press). These challenges have potential impact for CS analysis primarily in the ways in which language is used in social media; both in using deconstructed words and phrases and in idiomatic language use. Developing automated tools that can detect code-switched language use is problematic due to the exponential variety that such communication employs. Unpredictability, uncertainty, and frequency in communication style shifts will be a continuing barrier to automatic detection and interpretation of CS in text. For non-CS text, however, there are several approaches demonstrating potential. We describe some of those in the following paragraphs, especially Contour and GATE, and explore how they might be used to further our exploration of CS in text.

To address these challenges, we are engaged in developmental efforts to produce extraction techniques and processing algorithms to map the change in a social network over time with respect to important concepts discussed in the network. We are also developing approaches to identify dynamic shifts in leaders and followers and subgroup formation. We are supporting small business research to develop semantic representations of text analysis to discover topic formation and identify polarity of associated text. In each of these endeavors, we are working across scientific disciplines to forge partnerships with computer scientists, psychologists, linguists, social scientists, and network/electronics engineers to develop a common understanding of how a comprehensive persistent surveillance system can utilize findings from online social media. One exemplary tool that we used in this proof of concept exploration is a software application developed under a SBIR effort (Decisive Analytics Corporation developed the Social Network Analysis Realization and Exploitation component of their larger text analytics software, Contour). The Social Network Analysis Realization and Exploitation is a network extraction tool that works in a larger text analytics software package known as Contour. The full range of Contour applications was deployed in this analysis, thus the name of the larger toolsuite, Contour, will be used hereafter in this report.

## **5. Automated Text Analytics for Social Network and Sociolinguistic Analysis**

---

As asymmetric military operations have increased over recent years, it has become increasingly important for the military to develop capabilities for rapid information fusion to facilitate and maintain accurate situational awareness and understanding in dynamic and often foreign territories. The ability to extract meaning in relationships between people, organizations, events, and locations from a variety of text and multi-source data sets is critical to proactive decision making during humanitarian and tactical military efforts. The following section describes 2 automated social network extraction and topic modelling tools GATE and Contour which can be applied towards this problem. For the purposes of this report, Contour was used for the analysis discussed in later sections and GATE was included as an alternative and potential analytical tool for future sociolinguistic investigation and analysis of bilingual texts.

### **5.1 GATE**

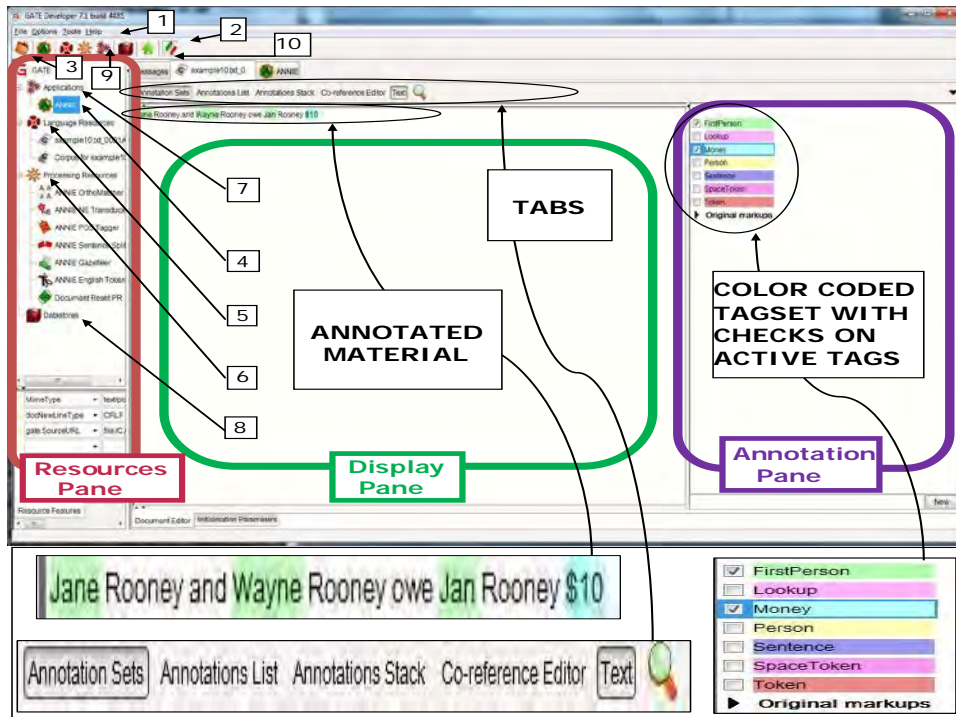
---

The SBIR tools are designed to handle the linguistic diversity of various domain and genre text types that serve as input. When considering social meaning and the constellation of features which constitute a given socio-cultural context, such as group membership and power hierarchy, standardization is scarce. This is unfortunate because standard feature types are what is required for development of annotation schemes and mark-up languages on which automation resulting from computational sociolinguistic analysis relies. For now, developers working with expert analysts will have to customize mark-up schemes that suit the individual purposes of the systems. Clearly, a standardized sociolinguistic mark-up language is soon to appear on the technological horizon. The present state-of-the-practice, however, is to account for contextualization cues with an annotation feature set that is characterized by precise definitions structural refinements, and, more important, that is informed by the analytic processes of sociolinguistic experts.

To aid in this process, familiarization with software developed at the University of Sheffield, UK, and known as GATE may be helpful (Cunningham et al. 2011). This tool's flexibility and ease of use make it accessible to linguistics and language specialists with little programming experience. It proved effective for Named Entity Extraction (NEE) on the synthetic messages contrived by military scenario developers to have been sent and received by individuals living in and around the Baghdad area (Vanni and Neiderer 2014, p. 1).

GATE development began in 1995. As techniques for natural language processing (NLP) are investigated by the research community and become part of the NLP repertoire, developers incorporate them with wrappers, which allow the output from GATE processes to be recognized as input by the new process and the output from the new process to be recognized as input by the existing GATE processes. It is the ongoing incorporation of emerging techniques to process text content for multiple purposes, formats and languages that accounts not only for GATE's longevity but also for its widespread use. Its Wiki site provides tutorials and documentation.

GATE's Developer Graphical Users Interface (GUI) in Fig. 5 displays 1) Menu and 2) Icon Bars horizontally. Action icons are 3) Restore Application from File, 4) Load ANNIE System, 5) New Language Resource, 6) New Processing Resource, 7) New Application, 8) Data Stores, 9) Manage CREOLE Plug-Ins and 10) Annotation Differences. Resources Pane Actions apply to every application and Icon Bar Actions are project-specific. Tabs indicate possible Display Pane and Annotation Pane contents (i.e., Annotation Sets, Annotations List, Annotations Stack, Co-Reference Editor and Text). Here, active tabs are Annotation Sets and Text. (For a detailed discussion of each of the tabs, consult the GATE manual site at <http://gate.ac.uk/sale/tao/>.)



**Fig. 5 GATE Developer GUI**

The set of categories defined for mark-up of this text are listed in the Annotation Pane and include a person's first name (FirstPerson), a money reference (Money)



and a person's full name (Person). One category could just as easily have been "expert you", for uses of "you" text strings by persons presenting themselves as experts. Current practice in information extraction system development is to draw up a set of guidelines that define category membership. Considering the sentence, "Greg Miller is here because Rosa Parks High School is closed for break", strings "Greg Miller" and "Rosa Parks", are candidates for membership in the "person's full name" category because names, "Greg Miller" and "Rosa Parks" both refer to persons.

Annotation guidelines tease apart category-relevant features of references. The features then become the criteria for category membership. In the early stages, a project will iterate on multiple-annotator trials for refinement of guidelines with feature details and distinguishing examples, to prevent misinterpretation. Measuring inter-annotator agreement soon follows and, depending on results, there is further revision of guidelines and previously annotated data, to maintain annotation consistency. Ideally, projects create the full complement of annotated data in a double-blind manner, take inter-annotator agreement measures, and judiciously resolve annotation conflicts to produce 'Gold Standard' datasets.

In the present example, meaning-based guidelines are likely to reserve the 'person's full name' category for full name person references that stand alone in both semantic and syntactic or phrase structure. The 'Greg Miller' string, which, alone, both refers to a person and constitutes grammatical subject constituent, would meet those criteria. By contrast, the "Rosa Parks" string, which also, alone, refers to a person, fails to meet the second criterion. Alone, it can function only within its subordinating phrase, "Rosa Parks High School," which constitutes the dependent clause subject. This is an example and different projects may define the bounds of a similar category in varying ways. What is important for the purposes of creating high-quality annotated data is that annotators converge on the category concept, the bounds of its expression and the project's conventions.

For sociolinguistic and social network analysis, the category of 'expert you' is certainly in scope. Social meanings with a constellation of linguistic cues are also plausibly represented with annotations. An increase in the number of constituting cues, however, may result in an exponential increase in the amount of data required for development of automated systems. Systems such as GATE, which speed up the human annotation process with capabilities for user-defined pattern-matching rules of any manner of complexity, can serve to produce seed data for lightly supervised approaches, results of which can be recycled through the human loop.

## 5.2 Contour

---

ARL sponsored an OSD funded SBIR project award to the Decisive Analytics Corporation for the development of a visual analytic and computational social network analysis tool, named Contour. Contour is a media analysis and management platform that provides several features designed to facilitate the search, analysis, and exploitation of text and intelligence documents. An analyst may select to conduct a word search or semantic search from Contour's home screen. Contour uses semantic analysis to extract persons, organizations, concepts, roles, and locations from text messages. Contour uses frame-based semantic modeling software to automatically build detailed network models from unstructured text. Contour imports unstructured text and then maps the text onto an existing ontology of frames at the sentence level, using FrameNet—a structured language model—and through Semantic Role Labeling (SRL). SRL automatically identifies the fundamental concepts expressed by text and maps these ideas into semantic roles (Kase and Dumer, 2013). This extracted information is displayed alongside the main screen of Contour along with the number of messages associated with each term listed under those headings. Analysts may click on a term to view the specified messages associated with it. Contour provides 4 main advanced analysis capabilities: activities, topic table, entity relationships, and media.

### 5.2.1 Contour Activity Analysis

The activities analysis capability displays the extracted information by grouping it together by related themes in word distributions (Fig. 6). Larger font is used to represent more prevalent words within that theme. Analysts may select a theme or a specific word to bring up the specific media associated with either. The themes are displayed in gradients of red indicating negative sentiment, or green indicating more positive sentiment is being expressed in relation to those words. Analysts may sort the activities by topic number, sentiment, or document count. The screen shots in this section demonstrate the Contour capabilities using Tweets collected from Eldoret, Kenya. Twitter user names have been redacted for the purposes of this report.



Fig. 6 Activity word distributions

### 5.2.2 Contour Topic Table

The topic table provides the analyst a quick visual overview of topics, themes, and sentiment. Similar to the activities analysis, the topic table may also be sorted by topic number, sentiment, or document count. Figure 7 shows the 6 topic results for a search organized in ascending order by sentiment. As with the activities analysis, red indicates negative sentiment while green indicates positive sentiment. The topic table also provides the media (e.g., messages) count associated with each theme and the topic number assigned to each theme. For example, in Fig. 7, topic “9”, contains 32 media that express approximately twice as much negative sentiment as positive sentiment in messages where the prevalent words include “na, players, and bad”. Analysts may click on the document icon next to the topic number to bring up the specific media associated with that topic. The word distribution similar to that shown in the activities analysis is also displayed allowing the analyst to click on a specific word within the theme to further specify a search.

#	Count	Sentiment	Words
📄 9	32		na, players, lots, long, kwa, bad, RT, bado, nmeanz, pole, eldy, il, gud, words, poa
📄 17	16		Stoic, Regret, geeksquad,error, boss, yangu, Probably, Ultimate, ya, DREAM, Baaqay, black, jams, 27.72
📄 8	13		DM, occasions, citizens, migrant, multi, o, acc, insured, Brains, khedira, child, leader, fruits, disclosing, JustAsking
📄 12	20		companies, Monaco, lead, run, talk, body, God, monacoDL, rights, Mogotio, dwelling, Group, claiming, 45.91, player
📄 7	39		shicoshix, hot_, kenya, dropzone, iko, plz, play, RICHIE SPICE, eldy, MOTHER, bro, kwane, massive, tukisonga, ikwom
📄 27	21		dog, BridesmaidProblems, idiot, supper, AfricaFactsZone, Real, followers, number, subscription, opening, shikungigi, Students, JKylux, Picture

Fig. 7 Topic table

### 5.2.3 Contour Entity Relationship Analysis

The entity relationships analysis capability extracts entities such as people, organization, and places from the text and uses this information to construct a social network. Analysts may adjust the connection strength to display the social network as it changes by introducing weaker connections or only showing the strongest connections between entities. Contour also allows analysts to select to view only

the persons, places, or organizations contained in a social network displayed. An analyst may also select to have the network organized and color-coded by facets so that smaller networks contained in the graph may be more easily viewed. The facet view also displays the major themes expressed in each facet.

### 5.2.4 Contour Media View

Media is the fourth primary capability in Contour. The media view displays all the media or messages contained within the data set being explored in Contour at that time (Fig. 8). Analysts may read the messages from the media page or they may select an individual message to view further detail such as the metadata associated with that message (e.g., asset type, created date, source date, last viewed). When viewing an individual message, the analyst is also provided the capability to download the original file and/or to create a clip of the file. An analyst may also select to have any persons, locations, or organizations that were mentioned in a message and automatically extracted as such to be highlighted. Figure 9 demonstrates a message where an analyst has selected to view any persons mentioned in the file and Contour has correctly identified and highlighted the person entity of a famous athlete. Semantic frames contained in a message may also be viewed such that the frame, frame element, and both are color-coded in the message.



Fig. 8 Media



Fig. 9 Media message with person highlighted

## 6. Method

In this section we describe the methods we used for data collection and the sites where the data were collected. The main target for data collection in this report is Kenya. We provide the rationale for choosing this site, describe language practices in the region, and argue that these practices are influenced by ethnic and gender stereotypes, social class issues, and political affiliation. We then describe how the Tweets for this report were collected, how they were filtered, how we analyzed the

Tweets, and how we formatted the Tweets for analysis by Contour. We work through the stages of Contour, provide screenshots to support each stage, and identify what we have learned.

## **6.1 Kenya as a Sociolinguistic Research Site**

---

Tweets were collected from Kenya because the ESCALES research team had previously collected a diverse data set (sociolinguistics interviews, observations, and focus group discussions) in Eldoret specifically, which provided ground truth to the collected Tweets. Eldoret, with a population of 289,380 inhabitants, is a city in the northwestern Kenyan county of Uasin Gishu. Moi University, a public university with a student of population of just over 16,000 students, is one of the main employers in Eldoret. The area is known for its agribusiness, with wheat, maize, and tea as the main agricultural products. It is also famous for producing some of the most successful long-distance runners; many long-distance runners from around the world come to this area for training.

Kenya is officially a bilingual state, with English and Kiswahili as its 2 official languages. At the same time, Kenya is a multilingual country. According to *Ethnologue: Languages of the World* (Ethnologue 2014), a comprehensive web-based reference catalogue of the world's known living languages, there are 69 languages spoken in Kenya. The local population of Eldoret consists of the members of Kalenjin tribes, yet other ethnic (tribal) communities are also widely represented (e.g., Kikuyu, Luo, Maasai). Each ethnicity has its own language.

The ESCALES team of Drs Kamwangamalu, Tovaes (HU), and Rosé (Carnegie Mellon University) conducted field research at Moi University from 30 July–3 August 2013. Prior to the research trip, the research partners at Moi University, who share the participants' cultural background and linguistic repertoire, conducted sociolinguistic interviews with a group of 20 students. The goal of the interviews was to understand language attitudes and practices and various issues that young educated Kenyans are concerned with and talk about. In addition to Swahili and English, the students' linguistic repertoire included a variety of ethnic languages: Giriama, Kikuyu, Kisii, Luhya, Luo, Kalenjin, Kuria, Kimeru, Kipsigis, and Sabaot. Furthermore, most of the students report using Sheng, a mixed Swahili-English language.

Having reviewed a few sample recordings, the team identified class, ethnicity and gender as the main social variables that influence linguistic and social behavior. Each of the variables has some cultural beliefs (stereotypes) attached to them and young educated people at Moi University are navigating and negotiating their

identities against the backdrop of those beliefs. The following paragraph is a summary of the beliefs (stereotypes) as reported by the students.

### **6.1.1 Ethnic and Gender Stereotypes**

The Kisii and the Merus are said to have anger issues and to be arrogant. The Kissis talk fast and believe that women should be subservient to men. The choice of a wife in this tribe is based on how good the future bride will be at house chores, especially cooking. If a woman is not a good cook, she is likely to be returned to her parents. Kisii men are stereotyped as unemotional; they do not express their feelings. The Merus are described as reclusive and primitive. They have difficulty pronouncing the sound ‘sh’ and render it as ‘s’ (e.g. fish [fiʃ] → [fis]). The Sabaot are known for their alleged loose morals. The Kalenjins are described as tall, athletic, not flashy, ready to fight, introverted, uncivilized, primitive, illiterate, slow, satisfied with themselves, not business-oriented, do not invest in luxury or education but rather treasure investing into the land for farming purposes. They strongly believe that wives should be submissive to their husbands. The Nandis, one subgroup of the Kalenjin tribe, are said to be generous, warriors, very well-organized, united, and humble. They own lots of land and are mostly farmers.

The Luhya group (one of the major ethnic groups in Kenya) are known to be hard workers and tea growers. There are about 18 tribes under the umbrella of what is generally known as the Luhyas. They practice circumcision and entertain themselves with cock fighting. They are also described as big eaters and are known to be generous and talkative. When they call in to send greetings to relatives via a radio talk show, the list of relatives goes on and on. In Luhya community, it is traditionally the role of the men to provide education for children and welfare of the family, but nowadays it is the women who are responsible for these duties. The Luos live in geographical proximity with the Luhyas. They are said to be extravagant, lazy, loud, clever, smart, egocentric, hate defeat and complain when they are handed one, like fighting, believe in the best of themselves, are very proud, and pretend to be wealthy even if they have no wealth at all. They are in love with their language, are responsible but cunning, and believe that they own the country and the government. They consider English as their first language, tend to buy European-made luxury items and are thus known as ‘Luoceans’—people who identify themselves with everything European, be it language or clothing. They tend to use big words and push everything to the extreme. If a Luo is a window cleaner, for instance, he will describe himself as ‘a transparent wall technician.’ When a Luo man dies, his relatives claim all the property because he was “one of us, our brother”. In so doing, they leave a widow to fend for herself.

The Kikuyu compete with the Luos for political supremacy over the country. They want to dominate others, especially their rivals, the Luos. The Kikuyu are business-oriented. They love money more than they do human beings, can kill over money, and are known as capitalists or money-making professionals. They value land but not to the same extent as the Kalenjins. Unlike the Kalenjins, the Kikuyu invest in the future. They are known to be the most wanted robbers in the country. If there is a robbery somewhere, the finger is first pointed at them as the culprits. The Kikuyu are considered more matriarchal than patriarchal. Kikuyu women are known to be very aggressive and to beat up their men.

### **6.1.2 Political Affiliations and Ethnic Groups**

The students reported that Kenyan political parties are either ethnically based or ethnically oriented (Elischer 2008). For instance, the Luos are expected to belong to and support the Coalition for Reforms and Democracy. Kikuyus are expected to be affiliated with Jubilee and the National Alliance. People from the Rift Valley tend to belong to the United Republican Party.

### **6.1.3 Class Issues and Stereotypes**

Class issues are closely connected with a) level of education and b) rural versus urban distinction. Because education is linked primarily to the urban areas, people from rural areas tend to believe that all urban dwellers are rich. Students report that public universities are more prestigious (remember Moi is a public university); admission standards are higher and the students are not pampered. The graduates from public institutions are more marketable because in private schools students do not speak Swahili (English only). Many of the interviewed students grew up in rural areas and then moved to the Eldoret area either during their secondary education or when they started the university. Those who come from the rural areas had a hard time adjusting to their new life in Eldoret. The causes range from switching from more traditional to more modern outfits to encountering people from different linguistic backgrounds. Also, some of them did not have computer skills.

Students report that rich people try to sound American, and to do so they talk through their nose. Rich people talk about experiences that poor (rural) people never had, like going swimming or going to the mall, but poor (rural) people talk about going to the market to buy vegetables.

## **7. Data Collection**

---

The first step in collecting Tweets was to choose a location and define the search parameters. The search parameters were first set to Eldoret, Kenya but later

expanded to include 4 of the largest cities or towns in Kenya: Mombasa, Nairobi, Nakuru, and Eldoret. To search for Tweets in a particular area, the following code had to be entered into the Twitter search bar: “geocode: Latitude, Longitude, Radius (50 km).” Latitude and longitude coordinates were obtained from Google Maps. With that code, Twitter was able to search for and produce the Tweets in that area. The 4 regions are shown in Fig. 10 (Bus-Africa, 2014).

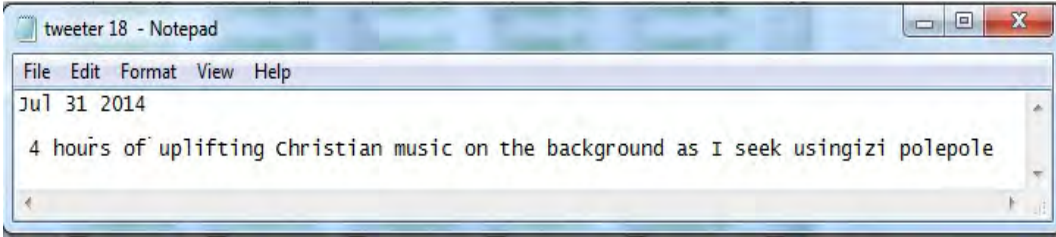


**Fig. 10 Map of Kenya**

The Tweets were collected objectively, meaning that all the Tweets were collected regardless of the content or language they contained. A total of 1,407 Tweets were collected from the 4 cities. The breakdown is as follows: Eldoret, 365 Tweets; Nairobi, 339 Tweets; Mumbasa, 318 Tweets; and Nakuru, 385 Tweets.

Tweets from each city were initially collected manually by copying and pasting the individual Tweets into a separate text document (Fig. 11). However, the Tweets needed to be formatted in a certain way for Contour to process them. Therefore, the first line only contained the date that the Tweet was made. The second line contained the username and Twitter handle. For the purposes of this report, the username has been redacted from the example in Fig. 11. Any subsequent line contained the content of the actual Tweet. The contents of the Tweets were unedited for composition or format.





**Fig. 11 Example of a formatted Kenyan Tweet**

To facilitate the collection process, a computational program was written using R. R is a free software programming language for statistical computing and graphics (Jaffar et al. 1991). The R program enhanced the speed and quantity by which Tweets were collected. R was also used to format the Tweets for ingestion into the visual analytic and computational network analysis tool, Contour. To make the Tweets ingestible for Contour, the Tweets were copied into a text document to detect and remove any hyperlinks. Several words such as “ReTweet,” “Favorite,” “Expand,” and the hashtag symbol (#), were also deleted. Once the Tweets were made compatible, the messages were placed into an Excel document and the R program was able to ingest the Tweets into Contour for analysis. The R code developed for this project is available in the Appendix of this report.

The first programming element of the R code split each line in the comma separated file into an individual file. Once this was completed, the program needed to go through each file and check to see if the file contained a date, username or Twitter handle, Twitter syntax such as Tweet, ReTweet, Favorite, etc., or if the file contained an actual Tweet. The program would then save the date file and save the username and Twitter handle, or ignore the Twitter syntax files, or merge the Tweet with the date file and the saved username and Twitter handle. The date file needed to determine the date of the Tweet with a dd/month/yyyy format. This was complicated for Tweets made within the last 60 min or within the last 24 h as they would not have a date but simply 5m or 4h (designating how much time had passed). To tackle this problem we checked to see if the Tweet had a minute 1–59 or an hour 1–24. If it did it would set the date in the date file to the current day’s date. If the program did not remove the Twitter syntax, it would associate the date with the wrong Tweets and it would confuse the Contour analysis because words like “favorite” or “view media” would come up and affect the analysis Contour does to find the frequency of key terms. To merge the date file and usernames with the Tweet file to be ingestible into Contour the date needed to be the first line of the new ingestible file followed by the next line containing the username and Twitter handle while the rest of the file contained the actual Tweet text. Once it had merged the date, usernames, and the Tweet and removed all the Twitter syntax it

would save the new file and continue until it had finished running through all the Tweets collected. The formatted files would then be loaded into a zip file which was then loaded into Contour.

## **8. Results**

---

As described earlier, Contour offers several advanced analysis capabilities to support the analysis workflow through facilitating the search, analysis, and exploitation of text documents. It supports automated extraction and analysis tools designed to assist an analyst with searching large data sets to build an understanding of the context in an area of interest through the extraction of sentiment across topics and themes. In a military context, Soldiers are often deployed to foreign countries where the language and culture may be unfamiliar to them and vastly different from their own. Analysis tools such as Contour may provide a critical link to assist analysts that are operating in an unfamiliar territory with building an understanding of important cultural issues. To explore this application of Contour, an analysis was conducted of the 4 datasets representing the following 4 regions in Kenya: Eldoret, Mombasa, Nairobi, and Nakuru. An author unfamiliar with these regions of Kenya was designated to conduct the analysis so that knowledge gained would be the result of using the tool. Once the initial analysis was concluded, data were shared with the authors with knowledge of Kenya and African languages to document ground truth. Contour's 4 primary analysis platforms (i.e., Activities, Topic Table, Entity Relationships, and Media) were used to conduct the analysis. The following section describes this analysis process and the social context gleaned by the analyst.

Individual studies were conducted for each of the 4 regions and across regions to demonstrate Contour's capabilities. To guide the analysis, the user was instructed to apply Contour for developing a social context for the 4 regions. Social context includes economic, political, institutional, historical, cultural, and ideological factors that can be decisive for particular decision-making processes (Kallis and Coccossis 2003). Relevant examples of the analyses conducted across the 4 datasets were selected for presentation in this report.

Looking for key institutions, the analyst reviewed the organizations automatically extracted from the messages and displayed by Contour. As can be seen from the Fig. 12, Moi University School was the most referenced institution occurring in 14 Tweets and again as Moi University in 9 Tweets. However, upon closer examination of the 14 Tweets referencing Moi University School, due to an error in Contour, several Tweets had been repeated so that in actuality, there were 6 unique Tweets referencing Moi University School. Contour developers are addressing this error presently. From these Tweets, it could be gleaned that there

are several separate colleges within Moi University including a School of Law and a School of Medicine. Furthermore, the authors of the Tweets appeared to be students as several Tweeted that they were waiting for class to begin. When comparing extracted organizations across the regions, Moi University also appeared in several Tweets from Nairobi. As the Tweets were collected by geolocation, and these cities are relatively far apart from each other, this suggests that there are several campuses for Moi University that are extended across Kenya. This institutional social context was confirmed by ground truth, through conducting a Google search on Moi University, which revealed several addresses for the university including branches in Mombassa and Nairobi (Kenya Postel Directories 2014). Additionally, this ground truth was confirmed by several ESCALES research assistants, some of whom are natives of Kenya, who confirmed that Moi University has several campuses located across Kenya.



**Fig. 12 Automatically extracted organizations for Mombasa by Contour**

In addition to identifying an institution, social context regarding ideology was also evident in the Mombasa Tweets. For instance, several of the Tweets referencing Moi University also referenced music, with some referencing a song “Call on Jesus”. Not only does this indicate that music may be relevant for college-aged persons in Kenya, but it also indicates that religion may also be an important part

of the social context, specifically a Christian religion as references were made to Jesus, the Lord, prayer, and God. When comparing the extracted persons referenced in the Tweets across the 4 regions, Tweets in each region mentioned “God,” with 6 Tweets in Eldoret, 14 in Mombasa, 20 in Nairobi, and 4 in Nakuru. Messages in Mombasa and Nairobi also referenced “Jesus” with 6 and 7 Tweets, respectively. From Mombasa, 6 Tweets referenced the “Lord.” When reviewing the organizations across the 4 regions, several references were made towards Christian institutions: Reimagining Church (Eldoret, 5 Tweets); Catholic Church and Catholic University of East Africa-Eldoret (Mombasa, 2 Tweets each); and Heart of Jesus Cathedral (Mombasa and Nairobi, 3 Tweets each). Contour did not seem to extract any persons or organizations associated with another religion besides Christianity. With references to Christianity occurring across the cities, this suggests that Christianity may be an important influence in Kenya. This inferred social context was verified by ground truth by conducting a Google search of religion in Kenya, which revealed more than two-thirds of the population in Kenya, is in fact, Christian (Encyclopedia Britannica 2014). Further bolstering the ground truth, the ESCALES research assistants, as mentioned previously, some of who are natives of Kenya also verified that the majority of Kenya’s population are Christians.

## **8.1 Contour Detection of Code-Switching**

---

At this time, Contour is limited as a tool for analyzing CS. It was hypothesized that if a Twitter author was found to have employed CS in a message, that author may be more likely to do so in the future. Once a message was manually identified as an example of CS, a search was conducted on the author’s username. Since the datasets were collected during a short timeframe, the datasets may not have provided a long enough period for individual Tweepsters to send multiple messages. However, searching usernames proved somewhat successful. For instance, a Tweet featuring CS posted by one Twitter user was manually located and a search was subsequently run on that username. The search revealed 20 Tweets by that user and, of those Tweets; 3 used CS (Fig. 13). User names have been redacted from the Tweets shown in Fig. 13 and replaced with asterisks. These Tweets were also not revealed by a manual search of the raw Tweets prior to ingestion by separate analysts, indicating some value to searching usernames when looking for examples of CS using Contour versus manual analysis of all messages. Moreover, when using Contour’s search capability, an analyst can select messages, such as those demonstrating CS, and add them to a folder they create. This would allow an analyst to extract CS messages in the future and to then use Contour’s analysis tools such as the Topic Table and Entity Relationships on only that data.

```

tweeter 249 .txt
Jul 31 2014
PREACH the good news QT @***** Manchester utd tunarudi na ubaya this season,come here august..EPL finally!!

tweeter 256 .txt
Jul 31 2014
Big up tew sana @***** kazi yako naipenda @***** @***** @*****

tweeter 302 .txt
Jul 31 2014
haha kwanza inafaa anitolee supper @*****

```

**Fig. 13 Example of CS Tweets**

## 9. Discussion

---

This report documents a proof of concept exploration of how a computational social network analysis tool can be used to provide context for sociolinguistic research, notably in CS communication acts. The goal is to extend understanding of how social intent is communicated through CS speech acts with automated processing capabilities. The analysis used a convenience sample of Tweets collected in a short period from 4 cities in Kenya where research associates had expertise and could interpret ground truth. The analysis of Tweets in Contour demonstrated the feasibility of using automated tools for understanding context in which a text corpus is situated. In particular, the network representation of the entities extracted from the Tweets allowed iterative exploration of the datasets and helped to establish improved awareness of the messages contained in the sets. Further, the clustering of Tweets by Twitter Handles, organizations, and content were useful and insightful.

This analysis also demonstrates how a software extraction capability can support iterative analysis of meso- and micro-level features in social dialogue. Examining individual messages allows the interpreter to understand how contextual features in the environment are affecting one person; studying the larger corpus (which include re-tweets and mentions of others' messages) provides insight into how a social group is reacting to environmental stimuli.

This type of contextual discovery tool has relevance to military applications in its ability to inform leaders of sociocultural norms; these have great impact on attitudes and behaviors of social groups. Automated analysis is required in today's world of increased information availability and shortened decision cycles.

The analysis of Twitter messages with the Contour software did provide a detection mechanism for CS within Tweets. These were manually translated by HU students

with expertise in African languages. This was feasible due to the small number of CS messages and the short length of the CS messages. The Contour software also was instrumental in providing relevant context for the regions of Kenya from which the Tweets were extracted. As noted earlier, context is indispensable for the social interpretation of such linguistic phenomena as CS. The capability of Contour to provide contextual information demonstrates how a leader could rapidly develop sociocultural knowledge of an unfamiliar environment.

## **10. Conclusions and Future Research**

---

We have identified the importance that computational social network analysis and sociolinguistic analysis—especially of such linguistic phenomena as CS—play in identifying the context in which social communication acts occur. These methods provide complementary awareness of social intent and can be used to forecast future actions. A prototype text analysis software application was used to demonstrate and used for informing contextual understanding. This software was used for meso and micro analysis of communication acts.

We presented results from applying 2 different methods of extracting social meaning from text. We found that fuzzy logic is particularly effective at clustering and subgroup determination. Examining speech accommodation was effective for the problem of hierarchy determination. Of course *The Wire* is a work of fiction, but since it was constructed to closely follow “real world” language usage there is good reason to expect that these techniques will be effective in analyzing real world data.

Future research efforts will continue to explore the common boundaries of computational social network analysis and sociolinguistic communication acts involving CS. These activities will revolve around the question of understanding how social intent is communicated in the African languages. We will also explore how social intent and CS acts can be detected and interpreted by text analysis algorithms.

## 11. References

---

- Bell A. Language style as audience design. *Language in Society*. 1984;13(2):145–204.
- Blom JP, Gumperz JJ. Social meaning in linguistic structures: CS in Norway. In: Gumperz JJ and Hymes D, editors. *Directions in sociolinguistics*. New York (NY): Holt, Rinehart and Winston; 1972.
- Bourdieu P. In: Thompson JB, editor. *Language and symbolic power*. Cambridge (UK): Polity Press, 1991.
- Bowman EK. Persistent ISR: The social network analysis connection. Proceedings from SPIE 8389: Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR III, 83891F; 2012 Apr 23; Baltimore, MD: SPIE. doi:10.1117/12.923021, 2012a.
- Bowman EK. Cultural situational awareness for counterinsurgency operations. Proceedings from: NATO Joint Symposium on Persistent Surveillance: Networks, Sensors, Architectures; 2012 Apr 3—May 1, Quebec, Canada, 2012b. [accessed 2014 Sep 16]. <https://www.cso.nato.int/abstracts.aspx>.
- Bravo-Marquez F, Gayo-Avello D, Mendoza M, Poblete B. Opinion dynamics of elections in Twitter. Proceedings from LA-WEB: 2012 Eighth Latin American Web Conference. 2012 Oct 25–27; Cartagena de Indias, Colombia. New York (NY): IEEE. doi:10.1109/LA-WEB.2012.11; 2012.
- Bus-Africa. Map of Kenya. Buses in Kenya. 2014 [accessed 2014 Sep 16]. <http://www.bus-planet.com/bus/bus-africa/Kenya/index.html>.
- Cheng A, Evans M. Inside Twitter: an in-depth look inside the Twitter world. 2009 [accessed 2014 Aug 20]. <http://www.sysomos.com/insidetwitter/appendix/>.
- Comminos A. Twitter revolutions and cyber crackdowns: User-generated content and social networking in the Arab spring and beyond. 2011 [accessed 2014 Sep 12]. [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCMQFjAA&url=https%3A%2F%2Fwww.apc.org%2Fen%2Fsystem%2Ffiles%2FAlexComminos\\_MobileInternet.pdf&ei=dKT8U4HYEObIsAT8roCgCg&usg=AFQjCNEOeI4kZkDvu-ez1nwsCgkWnZtM5w&bvm=bv.73612305,d.cWc&cad=rja](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCMQFjAA&url=https%3A%2F%2Fwww.apc.org%2Fen%2Fsystem%2Ffiles%2FAlexComminos_MobileInternet.pdf&ei=dKT8U4HYEObIsAT8roCgCg&usg=AFQjCNEOeI4kZkDvu-ez1nwsCgkWnZtM5w&bvm=bv.73612305,d.cWc&cad=rja).

- Cunningham H, Maynard D, Bontcheva K, Tablan V, Aswani N, Roberts I, Peters W. Text processing with GATE (Version 6). Sheffield (UK): University of Sheffield Department of Computer Science; 2011.
- Danescu-Niculescu-Mizil C, Lee L, Pang B, Kleinberg J. 2012. Echoes of power: language effects and power differences in social interaction. In: Proceedings of the 21st International Conference on World Wide Web (WWW '12). ACM, New York, NY, USA, 699–708. doi=10.1145/2187836.2187931 <http://doi.acm.org/10.1145/2187836.2187931>.
- De Choudhury M, Diakopoulos N, Naaman M. Unfolding the event landscape on Twitter: classification and exploration of user categories. Proceedings from CSCW'12: ACM 2012 Conference on Computer Supported Cooperative Work; 2012 11–15 Feb. Seattle, WA. New York (NY): ACM; 2012 doi:10.1145/2145204.2145242.
- de Rooij VA. French discourse markers in Shaba Swahili conversations. *The International Journal of Bilingualism*. 2000;4(4):447–467. doi:10.1177/13670069000040040401.
- Elischer S. Ethnic coalition of convenience and commitment: political parties and party systems in Kenya. No. 68. Hamburg (Germany): German Institute of Global and Area Studies; 2008 Feb [accessed 2014 Aug 22]. <http://www.giga-hamburg.de/workingpapers>.
- Espinosa AM. Speech mixture in New Mexico: the influence of the English language on New Mexican Spanish. In: Stephens HM, Bolton HE, editors. *The Pacific ocean in history*. New York (NY): Palgrave Macmillan; 1917. p. 408–428.
- Ethnologue Languages of the World: about the Ethnologue. 18th edition. Dallas (TX): SIL International Publications; 2014 [accessed 2014 Aug 22]. <http://www.ethnologue.com/about>.
- Fishman JA, Garcia O. (editors). *Contributions to the sociology of language* (Vols. 1–2). De Gruyter Mouton. [accessed 2014 Sep 16]. Available from <http://www.degruyter.com/view/serial/16644>. 1976/1972.
- Gumperz JJ. The sociolinguistics significance of conversational code-switching. In Cook J, Gumperz JJ (editors.), *Papers in Language and Context*, Working Paper no. 46 (pp. 1–26). Language Behavior Research Laboratory. Berkeley (CA): University of California; 1976.
- Gumperz JJ. *Discourse strategies*. Cambridge (UK): Cambridge University Press; 1982.



- Heller M. *Bilingualism: a social approach*. Basingtoke (UK): Palgrave Macmillan; 2009.
- Holton AE, Baek K, Coddington M, Yaschur C. Seeking and sharing: motivations for linking on Twitter. *Communication Research Reports*. 2014;31(1):33–40. doi:10.1080/08824096.2013.843165.
- Hymes D. *Foundations in sociolinguistics: an ethnographic approach*. Philadelphia (PA): University of Pennsylvania Press; 1974.
- Jacobson R. *Codeswitching as a worldwide phenomenon*. New York (NY): Peter Lang; 1990.
- Jaffar J, Michaylov S, Stuckey PJ, Yap RHC. The CLP(R) language and system: an overview. *Proceedings from Comcon Spring '91 Digest of Papers: IEEE Computer Conference; 1991 Feb 25–Mar 1; San Francisco, CA*. New York (NY): IEEE; 1991. p. 339–395.
- Kallis G, Coccossis, H. Integrated evaluation of sustainable river basin governance: comparison of the institutional context of the five case-studies. [accessed 2014 Aug 21] <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.518.6597&rep=rep1&type=pdf> 2003.
- Kamwangamalu NM. *Code-mixing across languages: functions, structure, and constraints [doctoral dissertation]*. [Champaign (IL)]: University of Illinois, Urbana-Champaign; 1989.
- Kamwangamalu NM. Sociolinguistic aspects of siSwati-English bilingualism. *World Englishes*. 1996; 15(3):295–306. doi:10.1111/j.1467-971X.1996.tb00116.x.
- Kamwangamalu NM. The state of codeswitching research at the dawn of the new millennium (1): Focus on the global context. *South African Journal of Linguistics and Applied Language Studies*. 1999;17(4):256–277.
- Kamwangamalu NM. The state of codeswitching research at the dawn of the new millennium (2): Focus on Africa. *Southern African Journal of Linguistics and Applied Language Studies*. 2000;18(1–4):59–71.
- Kamwangamalu NM. Ethnicity and language crossing in post-apartheid South Africa. *International Journal of the Sociology of Language*. 2001;152:75–95.
- Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*. 2009; 53(1):59–68. doi:10.1016/j.bushor.2009.09.003.

- Kase SE, Bowman EK, Al Amin T, Abdelzaher T. Exploiting social media for Army operations: Syrian crisis use case. *Proceedings from SPIE 9122: Next Generation Analyst II*, 91220D; 2014 May 5–9. Baltimore, MD: SPIE, 2014. doi:10.1117/12.2049701
- Kase S, Dumer J. Accelerating exploitation of low-grade intelligence through semantic text processing of social media. *Proceedings from 18th ICCRTS: International Command and Control Research and Technology Symposium*; 2013 Jun 19–21; Alexandria, VA: ICCRTS. [accessed 2014 Aug 20]. [http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCQQFjAA&url=http%3A%2F%2Fwww.dtic.mil%2Fget-tr-doc%2Fpdf%3FAD%3DADA587022&ei=Gmz\\_U8pOhqeCBP2bgtgN&usg=AFQjCNEJ6o0mFgQJitJcTZ-CembhI3jt3Q&bvm=bv.74035653,d.eXY&cad=rja](http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCQQFjAA&url=http%3A%2F%2Fwww.dtic.mil%2Fget-tr-doc%2Fpdf%3FAD%3DADA587022&ei=Gmz_U8pOhqeCBP2bgtgN&usg=AFQjCNEJ6o0mFgQJitJcTZ-CembhI3jt3Q&bvm=bv.74035653,d.eXY&cad=rja).
- Encyclopædia Britannica: Kenya. Chicago (IL) [accessed 2014 Aug 20]. <http://www.britannica.com/EBchecked/topic/315078/Kenya/259732/Religion2014>.
- Kenya Postel Directories. The official yellow pages of Kenya. [accessed 2014 Sep 23]. <http://www.yellowpageskenya.com/search/?business=Moi+University+Coast+Campus&reflocal=Mombasa>.
- Lakshmanan GT, Oberhofer MA. Knowledge discovery in the blogosphere: approaches and challenges. *IEEE Internet Computing*. 2010;14(2):24–32.
- Lynch M, Freelon D, Aday S. Syria's socially mediated civil war. *United States Institute of Peace*, 91. [accessed 2014 Sep 16]. <http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCMQFjAA&url=http%3A%2F%2Fwww.usip.org%2Fsites%2Fdefault%2Ffiles%2FPW91-Syrias%2520Socially%2520Mediated%2520Civil%2520War.pdf&ei=jsv8U7K1MITgsASktlLgAw&usg=AFQjCNFp4CqsGzZrFhy4mAnW00rma7CDAg&bvm=bv.73612305,d.cWc&cad=rja>, 2014.
- Maschler Y. Metalanguaging and discourse markers in bilingual conversation. *Language in Society*. 1994; 23:325–366.
- Mayfield E, Adamson D, Enand R. *Computational Linguistics*. Available from [accessed 2014 Sep 16]. <http://www.lighsidelabs.com>. 2014.
- Mayfield E, Rose C. *LightSide Researcher's Workbench*. [accessed 2014 Sep 10]. <http://ankara.lti.cs.cmu.edu/side/download.html>.

- Meeuwis M, Blommaert J. The 'markedness model' and the absence of society: remarks on codeswitching (a review article). *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*. 1994; 13(4):387–423. doi:10.1515/milt.1994.13.4.387.
- Mkilifi AM. Triglossia and Swahili-English bilingualism in Tanzania. In: Fishman JA, editor. *Advances in the study of societal multilingualism*. The Hague (The Netherlands): Mouton; 1978. p. 129–149.
- Myers-Scotton C. *Social motivations for codeswitching*. Oxford (UK): Clarendon Press; 1993.
- Myers-Scotton C. *Contact linguistics: bilingual encounters and grammatical outcomes*. Oxford (UK): Oxford University Press; 2002.
- National Public Radio. The Arab spring: a year of revolution. 2011 Dec 17 [accessed 2014 Aug 21]. <http://www.npr.org/2011/12/17/143897126/the-arab-spring-a-year-of-revolution>.
- Oksaar Els. On codeswitching: an analysis of bilingual norm. In: Qvistgaard J, Schwarz H, Spang-Hanssem H, editors. *The Proceedings of the Third AILA World Congress*; 1972 Aug 21–26; Copenhagen, Denmark. Heidelberg (Germany): Julius Groos Verlag; 1974. p. 491–500.
- Omestad T. USIP conference assesses social media's role in conflict. 2011 [accessed 2014 Aug 13]. <http://www.usip.org/publications/usip-conference-assesses-social-media-s-role-in-conflict>.
- Pear Analytics. Twitter Study – August 2009. [accessed 2014 Aug 19]. [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCAQFjAA&url=https%3A%2F%2Fwww.pearanalytics.com%2Fwp-content%2Fuploads%2F2012%2F12%2FTwitter-Study-August-2009.pdf&ei=\\_N78U\\_vQGabNsQTyroCYDg&usg=AFQjCNGgCaB8yGNuDbZyvV7ThJgHROalsw&bvm=bv.73612305,d.cWc&cad=rja](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCAQFjAA&url=https%3A%2F%2Fwww.pearanalytics.com%2Fwp-content%2Fuploads%2F2012%2F12%2FTwitter-Study-August-2009.pdf&ei=_N78U_vQGabNsQTyroCYDg&usg=AFQjCNGgCaB8yGNuDbZyvV7ThJgHROalsw&bvm=bv.73612305,d.cWc&cad=rja). 2009.
- Pew Research Center. Social networking fact sheet. [accessed 2014 Aug 21]. <http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/>.
- Ramirez AA, Taylor A, Rwebangira MR, Chouikha M, Levin L. Automated detection of social structure from transcripts of conversations. (forthcoming).
- Schiffrin D. *Discourse markers*. New York (NY): Cambridge University Press; 1987.

- Serena C, Clarke CP, Marcellino W, Winkelman Z, Tingstad A, Baim M, Paul C. Leveraging social media in army operations. RAND Arroyo Center Force Development and Technology Program (in press).
- Stieglitz S, Dang-Xuan L. Political communication and influence through microblogging: An empirical analysis of sentiment in Twitter messages and retweet behavior. Proceedings from HICSS 2012: 45th Hawaii International Conference on System Science; 2012 Jan 4–7. Maui, HI. New York (NY): IEEE; 2012. doi:10.1109/HICSS.2012.476.
- Thompson, D. Occupy the world: The '99 percent' movement goes global. The Atlantic. 2011 [accessed 2014 Sep 16] <http://www.theatlantic.com/business/archive/2011/10/occupy-the-world-the-99-percent-movement-goes-global/246757/>.
- Tyshchuk Y, Wallace W, Li H, Ji H, Kase SE. The nature of communications and emerging communities on Twitter following the 2013 Syria sarin gas attacks. Proceedings from JISIC: IEEE Joint Intelligence and Security Informatics Conference; 2014 Sep 24–26. The Hague, Netherlands. New York (NY): IEEE; 2014. p. 41–47.
- Vanni M, Neiderer A. General architecture for text engineering (GATE) developer for entity extraction: overview for SYNCOIN. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2014. Report No.: ARL-TR-7000. Also available at [http://www.arl.army.mil/www/default.cfm?technical\\_report=7142](http://www.arl.army.mil/www/default.cfm?technical_report=7142).
- Wei Li. Code-switching. In: Bayley R, Cameron R, Lucas C, editors. The Oxford handbook of sociolinguistics. Oxford (UK): Oxford University Press; 2013. p. 360-378.
- Yen J, Langari R. Fuzzy logic: intelligence, control, and information. center for fuzzy logic, robotics, and intelligent systems. Upper Saddle River (NJ): Prentice-Hall, Inc.; 1999.

## **Appendix. R-Code**

---

---

---

This appendix appears in its original form, without editorial change.

## R Code Written by William Mckinnon

```
library ('tm')
Tweet <- read.csv ("csv.csv",strip.white = TRUE,)
dbk_corpus <-Corpus(DataframeSource(Tweet), readerControl =
list(language="eng"), sep="\t")
corp <- Corpus(DataframeSource(Tweet))
dtm <- DocumentTermMatrix(corp)
writeCorpus(corp, path = ".", filenames = NULL)
elem1 %in% Tweet
### MAIN HUGE LOOP WHICH OPENS ALL MY FILES
year <- "2014"
day <- "1"
count = 1
for (i in 1:30000){
a <- (paste(i,"txt",sep="."))
aa <- (paste(i-1,"txt",sep="."))
dontcopy = 0
dontcopy2 = 0
#NEW STUFF START###
test <- read.table (a)
limit <- (length(test) + 1)
x <- ncol (test)
for (z in 1:limit){
qip = 1

if (x > 1) {qip = (x-1)}
if (test[,x] %in% ("ReTweet") | test[,x] %in% ("Reply") | test[,x]
%in% ("Favorite")
| test[,x] %in% ("Expand") | test[,x] %in% ("View") | test[,x] %in%
("permalink")
```

```

| test[,x] %in% ("Kenya") | test[,qip] %in% ("View") | test[,x] %in%
("Embedded")
| test[,x] %in% ("Jan") | test[,x] %in% ("Feb") | test[,x] %in% ("Mar")
| test[,x] %in% ("Apr") | test[,x] %in% ("May") | test[,x] %in%
("Jun")
| test[,x] %in% ("Jul") | test[,x] %in% ("Aug") | test[,x] %in% ("Sep")
| test[,x] %in% ("Nov") | test[,x] %in% ("Dec")

){dontcopy = 1}
for (v in 1:60) {if (test[,x] %in% (c[,v]) | test[,x] %in% (e[,v]))
{dontcopy = 1}}

```

```

if (x > 1) {(x = (x-1))}
}
if (dontcopy == 0) {
filebefore <- test}

```

```
#NEW STUFF END#
```

```

file <- read.table (a)
##### Minutes setup##### you need to edit the output
file#####
b <- cat (1:30,31:60, sep = "m ",file="output.txt")
c <- read.table ("output.txt")
##### Minutes
for (j in 1:60) {

```

```

# Need to change Print 5 to set the date of the file to the current date
#
if (file[(length(file))] %in% (c[,j])) {

  ## Here Goes the Tweet username without date#####
  dontcopy2 <- 0
  file2 <- read.table (a)
  mom <- (ncol(file2) + 1)
  x <- ncol (file2)
  for (u in 1:mom){
    dontcopy2 <- 0

    if (x > 1) {qip = (x-1)}
    if (file2[,x] %in% ("ReTweet") | file2[,x] %in% ("Reply") | file2[,x]
%in% ("Favorite")
      | file2[,x] %in% ("Expand") | file2[,x] %in% ("View") | file2[,x]
%in% ("View")
      | file2[,x] %in% ("Kenya") | file2[,x] %in% ("View")
      | file2[,x] %in% ("Jan")| file2[,x] %in% ("Feb")| file2[,x] %in%
("Mar")
      | file2[,x] %in% ("Apr")| file2[,x] %in% ("May")| file2[,x] %in%
("Jun")
      | file2[,x] %in% ("Jul")| file2[,x] %in% ("Aug")| file2[,x] %in%
("Sep")
      | file2[,x] %in% ("Nov")| file2[,x] %in% ("Dec")| file2[,x] %in%
("2014")
      | file2[,x] %in% ("2013")| file2[,x] %in% ("2012")| file2[,x] %in%
("2011")
    )

    {dontcopy2 = 1}
  }
}

```



```
for (v in 1:60) {if (test[,x] %in% (c[,v])| test[,x] %in% (e[,v]) |
test[,x] %in% (v)) {dontcopy2 = 1}}
```

```
if (dontcopy2 == 1) {
file2[,x] <- ""}
```

```
if (x > 1) {(x = (x-1))}
```

```
}
```

```
#End of newest stuff#####
```

```
rr <- sapply(filebefore,as.character)
```

```
rrr <- sapply(file2,as.character)
```

```
ll <- ("Jul 31 2014")
```

```
ff <- (paste ("Tweeter",count,".txt"))
```

```
#Changed for username BEFORE CHANGE #hobbes <- cat("Jul 31
2014", "\n",rr,file =(ff))#
```

```
hobbes <- cat("Jul 31 2014", "\n",rrr, "\n",rr,file =(ff))
```

```
#End changed
```

```
count = (count+1)
```

```
}
```

```
}
```

```
#Safty line
```

```
dontcopy2 <- 0
```

```

##### hours setup##### Edit this outputfile as well add a space
d <- cat (1:30,31:60, sep = "h ",file="output2.txt")
e <- read.table ("output2.txt")
#Extra sheets start
daysheet <- cat(1:32,file="output3.txt")
daysheetoutput <- read.table ("output3.txt")
yearsheetsheet <- cat(2000:2015,file="output4.txt")
yearsheetsheetoutput <- read.table ("output4.txt")
for (j in 1:60) {
# Need to change Print 5 to set the date of the file to the current date
#
if (file[(length(file))] %in% (e[j])) {
## Here Goes the Tweet username without date#####
dontcopy2 <- 0
file2 <- read.table (a)
mom <- (ncol(file2) + 1)
x <- ncol (file2)
for (u in 1:mom){
dontcopy2 <- 0

if (x > 1) {qip = (x-1)}
if (file2[,x] %in% ("ReTweet") | file2[,x] %in% ("Reply") | file2[,x]
%in% ("Favorite")
| file2[,x] %in% ("Expand") | file2[,x] %in% ("View") | file2[,x]
%in% ("View")
| file2[,x] %in% ("Kenya") | file2[,x] %in% ("View")
| file2[,x] %in% ("Jan")| file2[,x] %in% ("Feb")| file2[,x] %in%
("Mar")
| file2[,x] %in% ("Apr")| file2[,x] %in% ("May")| file2[,x] %in%
("Jun")

```

```

    | file2[,x] %in% ("Jul")| file2[,x] %in% ("Aug")| file2[,x] %in%
("Sep")
    | file2[,x] %in% ("Nov")| file2[,x] %in% ("Dec")| file2[,x] %in%
("2014")
    | file2[,x] %in% ("2013")| file2[,x] %in% ("2012")| file2[,x] %in%
("2011")
  )

```

```

{dontcopy2 = 1}

```

```

for (v in 1:60) {if (test[,x] %in% (c[,v])| test[,x] %in% (e[,v]) |
test[,x] %in% (v)) {dontcopy2 = 1}}

```

```

if (dontcopy2 == 1) {
file2[,x] <- ""}

```

```

if (x > 1) {(x = (x-1))}

```

```

}

```

```

#End of newest stuff#####

```

```

rrr <- sapply(file2,as.character)

```

```

##ANOTHER CHANGE ABOVE####

```

```

r <- sapply(filebefore,as.character)

```

```

l <- ("Jul 31 2014")

```

```

f <- (paste ("Tweeter",count,".txt"))

```

```

hobbes <- cat("Jul 31 2014","\n",rrr,"\n",r,file =(f))

```

```

count = (count+1)

```

```

}

```

```

}
#Safty line
dontcopy2 <- 0
#####For Jan#####
month <- ncol(file)
if (month > 1) {(month = (month-1))}
if (file[,length(file)] %in% ("Jul")) {print ("JanFUCKME")}
if (file[,month] %in% ("Jan")| file[,month] %in% ("Feb")|
file[,month] %in% ("Mar")
| file[,month] %in% ("Apr")| file[,month] %in% ("May")|
file[,month] %in% ("Jun")
| file[,month] %in% ("Jul")| file[,month] %in% ("Aug")|
file[,month] %in% ("Sep")
| file[,month] %in% ("Nov")| file[,month] %in% ("Dec")

) {
rrrr <- sapply(filebefore,as.character)
## Here Goes the Tweet username without date#####
dontcopy2 <- 0
file2 <- read.table (a)
mom <- (ncol(file2) + 1)
x <- length(file2)
for (u in 1:mom){
dontcopy2 <- 0

if (x > 1) {qip = (x-1)}
if (file2[,x] %in% ("Jan")| file2[,x] %in% ("Feb")| file2[,x] %in%
("Mar")
| file2[,x] %in% ("Apr")| file2[,x] %in% ("May")| file2[,x] %in%
("Jun")

```

```

| file2[,x] %in% ("Jul")| file2[,x] %in% ("Aug")| file2[,x] %in%
("Sep")
| file2[,x] %in% ("Nov")| file2[,x] %in% ("Dec")| file2[,x] %in%
("2014")
| file2[,x] %in% ("2013")| file2[,x] %in% ("2012")| file2[,x] %in%
("2011")
)

```

```
{dontcopy2 = 1}
```

```
for (v in 1:60) {if (test[,x] %in% (c[,v])| test[,x] %in% (e[,v]) |
test[,x] %in% (v)) {dontcopy2 = 1}}
```

```
if (dontcopy2 == 1) {
file2[,x] <- ""}
```

```
if (x > 1) {(x = (x-1))}
```

```
}
```

```
#End of newest stuff#####
```

```
rrr <- sapply(file2,as.character)
```

```
##ANOTHER CHANGE ABOVE####
```

```
ff <- (paste ("Tweeter",count,".txt"))
```

```
###Change to fix day and year pasting correctly####
```

```
for (loopall in 1:(month + 1)){
```

```
for (loopday in 1:31)
```

```
if (file[, (loopall)] %in% daysheetoutput[, (loopday)])
```

```

    {day = loopday}
  for (loopyear in 1:15)
  if (file[,loopall] %in% yearsheetoutput[, (loopyear)])
  {year = yearsheetoutput[, (loopyear)]}
  year2 <- sapply(year, as.character)
  }

  army <- file[, (month)]
  army2 <- sapply(army, as.character)
  hobbes <- cat(army2, day, year2, "\n", rrr, "\n", rrrr, file = (ff))
  ####End change to fix day and year pasting correctly#####

  count = (count+1)
  }
  #Safty line
  dontcopy2 <- 0
  print(i)
}

```

## List of Symbols, Abbreviations, and Acronyms

---

ARL	US Army Research Laboratory
CS	code-switching
DM	discourse marker
ESCALES	English (and French) with Selected African Languages
GATE	General Architecture for Text Engineering
GUI	Graphical Users Interface
HBO	Home Box Office
HU	Howard University
NEE	Named Entity Extraction
NLP	natural language processing
OSD	Office of the Secretary of Defense
POS	parts of speech
SBIR	Small Business Innovative Research
SRL	Semantic Role Labeling

1 DEFENSE TECHNICAL  
(PDF) INFORMATION CTR  
DTIC OCA

2 DIRECTOR  
(PDF) US ARMY RESEARCH LAB  
RDRL CIO LL  
IMAL HRA MAIL & RECORDS  
MGMT

1 GOVT PRINTG OFC  
(PDF) A MALHOTRA

1 DIR USARL  
(PDF) RDRL CII T  
L BOWMAN