Automated Inference of Socio-Cultural Information From Natural Language Conversations

Richard Scherl
Department of Computer Science
and Software Engineering
Monmouth University
West Long Branch, New Jersey 07764
Email: rscherl@monmouth.edu

Daniela Inclezan
Department of Computer Science
Texas Tech University
Lubbock, Texas 79409
Email: daniela.inclezan@ttu.edu

Michael Gelfond
Department of Computer Science
Texas Tech University
Lubbock, Texas 79409
Email: michael.gelfond@ttu.edu

***Abstract*—We discuss a methodology for extracting socio-cultural information from transcripts of natural language conversations. The methodology is applicable to a wide variety of languages. We use Russian and Tamil for illustration. The extracted socio-cultural information pertains to the nature of the relationship between the participants in the interaction. We concentrate on the information implicit in the use of terms that refer to people (pronouns, terms of address etc.). We have constructed an AnsProlog theory of the use of these language indicators in Russian and also in Tamil. It is this theory that looks at the usages of these indicators in the conversation in question and produces information about the relationships of the participants in the conversation.**

I. Introduction

We discuss a methodology for extracting socio-cultural information from transcripts of natural language conversations. The methodology is applicable to a wide variety of languages. We use Russian and Tamil for illustration. The extracted socio-cultural information pertains to the nature of the relationship between the participants in the interaction. We concentrate on the information implicit in the use of terms that refer to people (pronouns, terms of address etc.). We have utilized a knowledge representation language, AnsProlog, to build a theory of the use of these language indicators in Russian and also in Tamil. It is this theory that looks at the usages of these indicators in the conversation in question and produces information about the relationship of the participants in the conversation.

This work is part of a larger project to develop methods to discover the social goals and organization of a group through the group's use of language. The work reported here aims to support these aims by looking at the language usage of a group (as evidenced by transcripts of conversations) to provide insight into and evidence concerning the nature of the roles and relationships of its members.

We have built an AnsProlog theory for Tamil and also for Russian that describes the usages of the various pronouns and address terms in the language. The theory is used in a protytpe system that takes an AnsProlog annotation of the text and outputs claims about the relationship between the participants supported by the evidence in the text. Currently the input annotation (represented as a set of AnsProlog facts) is done manually. We are working on tools to automatically extract input representation from the text[1].

The main contribution of this work is the exploration of the use of AnsProlog as a language (and associated inference mechanism) to represent the background knowledge needed to infer socio-cultural information from the information contained in natural language conversations. We make two main points:

- Background knowledge and default resoning[2] are important for drawing conclusions from conversational data. AnsProlog is able to represent and reason with these defaults.
- The declarative nature of AnsProlog enables us to extract common pieces of code that will be needed to construct this type of software for most natural languages.

These points will be developed in the course of the paper.

Section II describes the sociolinguistic background for this study. The needed background on AnsProlog is given in Section III. We are limiting our conclusions to claims about deference and closeness. This is discussed in Section IV. Section V describes the data from Tamil and Russian with which we are concerned. Additionally, the input AnsProlog annotation format is described. Section VI gives an overview of the AnsProlog theories of deference and closeness that we have developed. We illustrate a couple of sample results in Section VII. Finally, in the Conclusion (Section VIII) we discuss our current and future work.

II. Sociolinguistic Background

Language and culture interact in a wide variety of ways. For example, the choice of lexical items, linguistic constructions as well as pronunciation can indicate something about the region of the speaker or the social group (class, caste, religious group, tribe) to which the speaker belongs. In the now classic study [1], it is shown how the realization of the postvocalic /r/ in

[1]A prototype extractor has been built for Russian by Greg Gelfond. A very preliminary extractor for Tamil has been built by Anubhav Gupta and Shubhendra Singh.

[2]Defaults are statements of the type "Normally (typically) elements of class **C** have property **P**." We may also define when the normal situation does not hold.

New York City English varies with class (and also context). The issue of context raises the culture/language specific defining of speech acts, speech events, and genres. All of these are indicated or correctly carried out through the appropriate use of linguistic constructions. In [2], the special speech styles used among the Ilongot people (Philippines) when settling disputes is outlined and can be contrasted with the special language used in our own legal settings [3]. Furthermore, the nature of the relationship between speaker/hearer/referent can also be indicated and altered by the choice of linguistic forms. For example, Javanese [4], [5] has different speech levels *ngoko*, *madya*, and *krama* (consisting of different lexical items of basic vocabulary) that are chosen in accordance with the relative rank difference of the speaker and hearer as well as the nature of their relationship (e.g., intimate, distant). As is well known (and will be discussed in more detail below), in many languages one may choose to use a plural 2nd person pronoun to indicate politeness or deference to the addressee. All of these topics and more have been studied in the fields of sociolinguistics and linguistic anthropology. For some surveys of work in this field see: [3], [6], [7], [8], [9].

We propose to concentrate on the use of terms that refer to people (e.g., choice of pronoun, address terms, kinship terms) and infer information about the nature of the relationships between the people referred to in the text at hand. Are they equal in status? Is one in some sense higher or superior to another? Is their relationship one of solidarity, intimacy, distance or a particular culturally defined type of relationship? Are they newly acquainted or have they been interacting over a long period of time. This topic has been studied extensively in linguistic anthropology. It can not be separated from the other concerns listed in the paragraph above. For example, different social groups may use different terms of address or have different patterns of pronominal usage. Additionally, the nature of the circumstances (e.g., public, formal, private, or a particular culturally defined event) will also influence the particular forms used. Finally, one can not ignore the wider choice of linguistic constructions as these will alter the interpretation of the referring words (pronouns, address terms).

The classic article on the topic of the indexing of relative social differences by choice of linguistic forms is certainly that of Brown and Gilman [10]. They uncover a pan-European pattern in which an asymmetric power relationship is indexed by the non-reciprocal use of the referentially second person plural pronoun (V) by the lower person, who receives the singular form (T) from the higher. Additionally, the singular pronoun (T) is used reciprocally in a solidary equal dyad while the plural (V) is used reciprocally in the non-solidary equal dyad. In a later work, Brown suggests that the pattern may be universal [11]. Numerous studies have explored these patterns in a variety of European languages, and non-European languages as well. These include Russian [12], Italian [13], Swedish [14], Persian [15] and Tamil [16]. Finally, it is shown in another article [17] that even though modern English has lost the 2nd person singular/plural distinction, similar patterns exist in the choice between use of the first name (FN) and title with the last name (TLN).

It is worth citing an example, since these will make clear to English speakers the social force of these usages which are expressed differently and often more elaborately in other languages. We all are most comfortable with the collegial and symmetric use of first names. Yet in ordinary usage there are people to whom we have to stick to Mr. Jones, or Mrs. Jones, or Dr. Jones etc. Brown and Ford cite the following illustrative story:

> The day after a convivial office party a breezy young clerk calls out to the president: Morning, Jack! and in icy tones the president replies: Good morning, Mr. Jones. [17]

The person of lower status must never use a higher status form than the higher status person uses. So the president puts the subordinate in that position to indicate that he should not use the first name. Not only does this story illustrate the rules delineated in [17] but it also illustrates a conversational violation of norms designed to achieve a particular effect. In the languages discussed above with a T/V distinction, these features of interaction are more salient than in English. Yet more elaborate pronominal systems are found in East Asian languages such as Thai, Burmese and Vietnamese [18]. Other languages have distinct speech levels [5], [4].

From a semiotic perspective, the meaning of these forms (in particular, the socio-cultural meaning) is generally understood as an instance of the indexical mode of signification [19], [20], [21]. Indexical signs obtain their meaning through a physical (perhaps spatial contiguity) connection with what they refer to. Thus the use of a particular pronoun indicates something about the relationship between the speaker of the utterance containing the pronoun and the addressee of that utterance. One of the characteristics of indexical meaning is that it can be used in a relatively presupposing or a relatively creative fashion [21]. A relatively presupposing use merely reinforces the understood socio-cultural context while a creative use establishes a new context (relationship between the speaker and hearer) that may be transitory or may persist over time. By its very nature indexical meaning is defeasible [22]. New features of the context (socio-cultural context) may be made evident from further conversational interaction that will force a reinterpretation of the earlier occurrence[3]. So, default reasoning as provided by AnsProlog is needed.

Furthermore, the literature [23], [13], [14], [24], [25], [26] shows that patterns of use of pronominal forms index a vector of aspects of the socio-cultural context, not just a single feature. These additional features include for example, the age of the speaker and hearer, the location of the speech (home or a public location), etc. For example, in both Tamil and Russian, the use of the singular pronoun between two educated adult males is highly indicative of a close relationship. But

[3]Certainly we can not at this point capture much of the subtlety that occurs in the use of these forms, e.g., the interpretation of very creative usages that violate established norms between the speaker and hearer. But our experience with AnsProlog so far shows that it may be suitable to extend further to capture this sort of reasoning as well.

if we have the knowledge that the speakers are children or are people who are less educated and from a rural area, this conclusion can no longer be drawn. So, default reasoning is needed along with background knowledge. We now turn to AnsProlog which provides us with the default reasoning capabilities. In the terminology of our AnsProlog theory, the default is that a symmetric use of the singular pronoun is indicative of closeness. But an exception to the default would be information that the speaker and hearer are members of categories of people for which this default does not hold.

III. Syntax and Semantics of AnsProlog

An AnsProlog[27], [28] knowledge base consists of rules of the form:

$$l_0 \leftarrow l_1, \ldots, l_m, \mathit{not}\ l_{m+1}, \ldots, \mathit{not}\ l_n \qquad (*)$$

where each of the l_is is a literal, i.e. an atom, a, or its classical negation, -a and *not* is a logical connective called *negation as failure* or *default negation*. While -a states that a is false, an expression *not* l says that there is no reason to believe in l.

The answer set semantics of a logic program Π assigns to Π a collection of *answer sets* – consistent sets of ground literals corresponding to beliefs which can be built by a rational reasoner on the basis of rules of Π. In the construction of these beliefs the reasoner is guided by the following informal principles:

- He should satisfy the rules of Π, understood as constraints of the form: *If one believes in the body of a rule one must believe in its head.*
- He should adhere to the *rationality principle* which says that *one shall not believe anything he is not forced to believe*.

The precise definition of answer sets is first given for programs whose rules do not contain default negation. Let Π be such a program and X a consistent set of ground literals. Set X is *closed* under Π if, for every rule (*) of Π, $l_0 \in X$ whenever for every $1 \leq i \leq m$, $l_i \in X$.

Definition 1: (Answer set – part one)
A state X of $\sigma(\Pi)$ is an *answer set* for Π if X is minimal (in the sense of set-theoretic inclusion) among the sets closed under Π.
To extend this definition to arbitrary programs, take any program Π, and consistent set X of ground literals. The *reduct*, Π^X, of Π relative to X is the set of rules

$$l_0 \leftarrow l_1, \ldots, l_m$$

for all rules (*) in Π such that $l_{m+1}, \ldots, l_n \notin X$. Thus Π^X is a program without default negation.

Definition 2: (Answer set – part two)
X is an answer set for Π if X is an answer set for Π^X.

Given, the translation of a text and a background theory, the initial task of inferencing in AnsProlog is to compute all of the answer sets or models of the text and background theory. To determine whether a fact follows from our text and background theory, it is necessary to have a definition of entailment.

Definition 3: (Entailment)
A program Π entails a literal l ($\Pi \models l$) if l belongs to all answer sets of Π.
The program Π's answer to a query l is *yes* if $\Pi \models l$, *no* if $\Pi \models \bar{l}$, and *unknown* otherwise.

Given an AnsProlog program Π, the output of an implementation of AnsProlog such as SMODELS[29] is a set of the answer sets of the program Π. Each of the answer sets is represented by a listing of the ground literals true in that answer set. The output can be quite large, but the user can restrict the output to see the specific results that are relevant to his/her purposes.

Note that the language of AnsProlog includes both negation as failure (not), and logical negation (-). This is important as the two together are used to represent defaults. Many examples of this combination occur in the next sections.

The goal in the rest of this paper is to first encode a background theory[4] concerning the use of pronouns and address terms in AnsProlog. The answer sets of the combination of the background theory and the conversation (a set of AnsProlog facts representing the occurrences of the salient forms) can be generated with Smodels[5].

IV. Claims

We make claims about *language uses*; delimited spaces of socio-cultural meaning, about which language use allows us to draw conclusions. One is *deference* and the other is *closeness*. We try to define these in such a way that they can be applied across multiple languages, although we do certainly realize that this ignores the culture-specific definition of these concepts.

A. Deference

Deference is a recognition (esteem, honor, etc.) given to a superior. It is an indication of the higher status of the receiver. We propose the following scale:

(2) X defers greatly to Y.
(1) X defers to Y.
(0) X and Y have the same social status
(-1) X expresses superiority to Y.
(-2) X expresses great superiority to Y.

For Tamil, we use all five positions on the scale, while for Russian we do not use the (2) and (-2). Note that if X defers to Y then Y expresses superiority to X. We can not have one without the other since for example an isolated occurrence of a singular pronoun by X addressing Y does not tell us if X is expressing superiority to Y or if X and Y are close.

Having drawn the inference of X showing deference to Y does demand further culture-specifc information in order to

[4]AnsProlog had been utilized in a wide variety of knowledge representation and reasoning tasks [30], [31], [32], [33].

[5]Smodels is one of a number of reasoners that implement the stable model semantics for AnsProlog. It can be downloaded from **http://www.tcs.hut.fi/Software/smodels/**.

determine what this shows about the nature of the relationship of X and Y. The information concerns what particular categories of people show deference (or a particular form of deference) to what other categories in the society/culture under consideration. The inference of deference is thus an intermediate step which when combined with background information can be used to possibly infer more specific features of the relationship of X and Y by our AnsProlog reasoning system.

B. Closeness

Closeness is the personal distance between two people. We have settled on the following scale:

(-2) X and Y are strangers.
(-1) X and Y barely know each other.
(0) X and Y have close contact at work or in some other comparatively small circle, but do not have close emotional ties.
(1) X and Y are in close contact and emotionally connected.
(2) X and Y are close friends or family members.

We use all five positions for Tamil and Russian. But we also use a coarser scale containing only **distant** and **close** when the available data does not allow us to make a finer distinction. Typically closeness will be indicated by a symmetric use of forms.

V. The Language Data

We have considered both Tamil and Russian. The forms that are being used are discussed for each language in turn below. Then we describe the AnsProlog based representation of the information provided by the conversation.

A. Tamil

We have the distinction between the 2nd person singular and second person plural pronoun. Following the common pattern, the second person pronoun is commonly used as an honorific form to refer to a single person. Tamil requires a verbal suffix that agrees with the pronoun. Therefore the pronoun is often dropped. We do not discuss the suffixes here as in our AnsProlog annotation we count an occurrence of the pronoun and/or the suffix as a single occurrence.

The second person pronouns are as follows:

nii	2nd person singular
niiŋgaḷ	2nd person plural/honorific
niir	2nd person archaic plural/middle honorific rarely used

The third person pronouns are as follows:

avan	Masculine 3rd person pronoun
avaḷ	Feminine 3rd person pronoun
atu	Neuter 3rd person pronoun (also used for people)
avar	Honorific 3rd person pronoun
avarkaḷ	Plural/honorific 3rd person pronoun

There are also a variety of different forms that can be used to refer to people (in either a 2nd person or 3rd person fashion). Some of these possibilities are listed below:

- **proper name**: The use of a proper name (in address or 3rd person reference) is similar in effect to singular pronoun in that a symmetric use is indicative of closeness and it is used asymmetrically from a higher status to a lower status person.
- **address term**: Examples are *ayyaa*, *caar*, *caami*, and *esamaan*. They can substitute for a personal name in a deferential fashion.
- **kinship terms** These also commonly are used in address or reference instead of the personal name. They are often used to non-relatives. Examples are *aṇṇan* ‘elder brother, *ammaa* ‘mother, *akkaa* ‘elder sister, *appaa* ‘father.

Particles are also suffixed at various positions and indicate something about the relationship between the speaker and addressee.

-ŋgaḷ	honorific particle
ṭaa	masculine intimate/dishonorific particle
ṭiii	feminine intimate/dishonorific particle
-ppaa	intimate/dishonorific/honorific usages
-mmaa	intimate/honorific usages
-yyaa	intimate/dishonorific/honorific usages

The honorific particle is clearly related to the plural marker. Additionally a number of the particles are related to kinship terms, but have a much wider sphere of use than the kinship terms themselves.

B. Russian

The second person pronouns are as follows:

ty	2nd person singular
vy	2nd person plural/honorific

As with Tamil, the pronoun may be dropped in the presence of the verb ending. Even if only the ending is present, we count the occurrence as an occurrence of the pronoun.

Additionally, there are a variety of options for terms of address. Each option indicates something about the nature of the relationship between the speaker and hearer.

- **Diminutive of Intimacy**: e.g. *Vanushka*
- **Diminutive of Derogation**: e.g. *Vanka*
- **Casual diminutive** e.g. *Vanya*
- **Forename + patronymic**: e.g. *Ivan Grigorevich*
- **Title + name**: e.g. *Dr. Ivanov*

Additionally, we make use of introductions; sentences of the type “Let me introduce Dr. Ivanov to you” and “Hello. My name is X.”

With Russian an asymmetric use of the singular 2nd person pronoun *ty* by a person A to person B, when A receives the 2nd person plural pronoun *vy* in return, is a typical indicator of B giving deference to A. But our data[6] calls for an exception

[6]From before the fall of the Soviet Union.

to this default behavior. Among Communist Party members use of the singular 2nd person pronoun *ty* was the accepted norm. A speaker is resisting inclusion among party members by using the 2nd person plural pronoun *vy* instead. So here the user of *vy* is not giving deference to the user of *ty*. The knowledge that one participant is a party member and the other is not only not a member, but is negatively predisposed towards membership is used to cancel the default conclusion[7].

C. AnsProlog Input

For our sample data we need written representations of conversations. Our Russian conversational data is taken from novels and short stories. The Tamil data comes from transcripts of movies and radio plays.

We number the utterances in the dialog. Each utterance is spoken by a particular speaker. A single speaker may speak many utterances in succession. We do have to give as part of the AnsProlog input a specification of the number of utterances in the dialogue. For example:

```
#const num_of_utterances = 95.
```

We also have to specify each of the participants in the conversation. For example:

```
participant(c1).
participant(c2).
```

Then we need to represent the various usages of the forms. The general format for representing the use of a form of address (i.e., 2nd person usage) is as follows:

```
address(n, X, Y, F).
```

In this case participant **X** uses form **F** to address participant **Y** in utterance **n**. The general format for representing the use of a form of 3rd person reference is:

```
refer(n, X, Y, F).
```

Here in utterance **n**, participant **X** uses form **F** to refer to person **Y**. Our Tamil data is particularly rich in the use of the 3rd person forms[8].

As mentioned earlier, the testing of our AnsProlog theories presented here is done with a manually prepared set of AnsProlog facts representing the occurrences of the salient forms in our conversational data. Development of the programs that extract these facts from the raw conversational data is in progress. Note that the extraction of the relevant forms, although not trivial since we have to do a certain amount of morphological and syntactic analysis, is not difficult to do with a high degree (not perfect) of accuracy[9]. But it is more difficult to determine who is being addressed or referred to in each occurrence of the forms. Address is easy if there are only two conversational participants, much more difficult with three or more[10]

VI. AnsProlog Theories

We have constructed AnsProlog theories to compute information about closeness and also deference giving between participants from the AnsProlog input representation of the conversations in both Tamil and Russian. The theories for the two languages are completely separate, but share some code.

The goal of the theories is to deduce predicates of the form:

```
defers(X,Y,L).
close(X,Y,L).
```

Clearly the first indicates that participant `X` defers to participant `Y` with level `L` deference and the second indicates that participants exhibit level `L` closeness. As mentioned earlier, the levels used for the two languages are different[11].

Given that these predicates are inferred for particular participants and a level (i.e., occur in the produced answer set[12]). We also provide rules to ensure that the claims about deference and closeness are appropriately displayed along with explanations of how these claims were inferred from the conversational data. These explanations are illustrated in the examples in the next section.

A rule for `defers(X,Y,L)` is as follows:

```
defers(X,Y,1) :-
        participant(X),
        participant(Y),
        form_of_address(FA1),
        indicates_deference(FA1, 1),
        form_of_address(FA2),
        indicates_deference(FA2, -1),
        addresses(X,Y,FA1),
        addresses(Y,X,FA2),
        not -defers(X,Y,1).
```

Here addresses indicates "addresses consistently." This will be defined below. The rule above defines level 1 deference as being an asymmetrical pattern of use of deference indicating forms. The predicate `defers(X,Y,1)` holds for participants `X` and `Y` as long as the asymmetrical pattern holds and we can not show that the predicate is false. Note that this rule abstracts out the language independent features and therfore can be reused in theories for different languages.

We introduce the notation `addresses` for addresses consistently:

[7]Another direction of reasoning would be from the pronominal usages and the background knowledge that the participant using *vy* is not of lower status than the participant using *ty*, to conclude that the user of *vy* is negatively predisposed to party membership.

[8]Currently, the AnsProlog theory does not make use of the information from 3rd person reference.

[9]We assume the written representation of the conversation includes a specification of who the speaker is for each utterance along with the linear ordering of the utterances that make up the conversation.

[10]Third person reference is much more difficult. It demands more background and/or much more natural language understanding.

[11]This is not surprising. One would expect there to be many more levels in a theory of deference for many East Asian languages. For example, see the descriptions of Thai, Burmese and Vietnamese found in [18].

[12]We do not discuss here the possibility of producing multiple answer sets. This corresponds to there being more than one plausible conclusion given the input data.

```
addresses(X, Y, P1) :-
        participant(X),
        participant(Y),
        X != Y,
        pronoun(P1),
        pronoun(P2),
        P1 != P2,
        utterance(N1),
        utterance(N2),
        num_of_addr(X, Y, P1, N1),
        num_of_addr(X, Y, P2, N2),
        N1 > N2 * c.
```

The constant *c* can be set to different numbers with AnsProlog statements such as `#const c = 2`. We do make different choices for different forms (e.g., pronouns vs address terms). There is often some variation in pronominal use and so one needs to decide on a cut off point. Note again that the form of the rule is language independent.

Closeness can be handled as indicated below:

```
close(X, Y, Level) :-
    participant(X),
    participant(Y),
    level_of_closeness(Level, Sc),
    scale(Sc),
    evidence(X, Y, Level),
    not
     contrary_evidence_exist(X, Y, Level),
    not finer_evidence_exist(X, Y, Level).
```

We conclude that two participants are at a particular level of closeness if there is evidence that they are at that level and there is not evidence contradicting that level and one can not find evidence for a finer level of granularity[13].

The evidence that two participants exhibit a particular level of closeness comes from their consistent use of a form indicating that level of closeness.

```
evidence(X, Y, Level) :-
        participant(X),
        participant(Y),
        level_of_closeness(Level, Sc),
        scale(Sc),
        form_of_address(FA1),
        form_of_address(FA2),
        uses_indicator(X, Y, Level, FA1),
        uses_indicator(Y, X, Level, FA2),
        not
         abnormal(X, Y, FA1, FA2, Level).
```

The predicate `uses_indicator(X,Y,Level,FA)` is defined in terms of `addresses(X,Y,F)` and therfore incorporates the notion of being consistently used. Note that the above rule is written in a general fashion and therefore can be shared across code for the theories for different languages.

[13]Recall that for closeness we have a coarse and a fine scale.

The definition of `abnormal(X,Y,F1,F2,L)` needs to be defined for each specific language, though. These are the abnormal conditions that will overide a default. The examples below are for Russian. The first example represents the abnormal condition of the speaker and hearer being children. This overides the default that the symmetric exchange of *ty* is evidence of a close relationship.

```
abnormal(X, Y, ty, ty, Level) :-
        child(X),
        child(Y),
        indicates_closeness(ty, Level).
```

The following rule specifies that the situation being "formal" overides the default that symmetric exchange of distance indicating forms (e.g., titles) is evidence for a distant relationship.

```
abnormal(X, Y, FA1, FA2, distant) :-
        participant(X),
        participant(Y),
        situation(formal),
        form_of_address(FA1),
        form_of_address(FA2),
        indicates_closeness(FA1, distant),
        indicates_closeness(FA2, distant).
```

The form of the theories rely upon an initial counting of the number of times a particular participant uses a particular form in addressing (or referring to) a particular person. The following predicate does this counting.

```
num_of_addr(X, Y, FA, N) :-
    N{utterance(U) :
      addresses(U, X, Y, FA)}N,
    utterance(N),
    participant(X),
    participant(Y),
    form_of_address(FA).
```

VII. Results and Sample Output

We have tested the systems on a variety of conversations in both languages. Here we give two examples (one from Russian and one from Tamil) to illustrate our approach.

A short Russian conversation from a novel was represented in the AnsProlog input mode as follows:

```
INPUT (to the system): LP-form:
#const num_of_utterances = 3.
participant(a).
participant(b).
addresses(1, a, b, ty).
addresses(2, b, a, ty).
addresses(3, a, b, ty).
```

The output was as follows:

```
Answer Set:
claim(a,"and",b,"are close")
evidence(a,"uses",ty,"when addressing",b)
evidence(b,"uses",ty,"when addressing",a)
```

```
support("Symmetric use of ty indicates
     that the speakers are close")
claim(a,"and",b,"have the same status")
evidence(a,"uses the informal form
   of address",ty,"when addressing",b)
evidence(b,"uses the informal form
   of address",ty,"when addressing",a)
support("Symmetric use of informal forms
  of address by both speakers indicates
  that the speakers have the same status")
```

The speakers symmetrically exchange the second person singular pronoun. Therefore they are judged to be close and of the same status.

One scene from a Tamil radio play was represented in the AnsProlog input mode. An abbreviated form is as follows:

```
#const num_of_utterances = 100.
participant(mank).
participant(kann). %
participant(kanak).
participant(nitt). % non present
address(1, mank, kanak, ngal).
address(3, mank, kanak, niingal).
address(5, kanak, mank, nii).
        .
        .
address(27, mank, kanak, nka).
address(29, mank, kanak, nka).
address(30, mank, kanak, niingal).
address(31, mank, kanak, niingal).
        .
        .
address(56, kanak, mank, nii).
refer(58, kanak, nitt, avan).
refer(58, kanak, nitt, name).
address(61, kanak, kann, name).
address(61, kanak, kann, taa).
refer(64, knak, nitt, avan).
address(65, kanak, kann, name).
address(65, kanak, kann, taa).
address(66, kann, kanak, part_ppa).
        .
        .
address(76, kanak, kann, taa).
address(79, kann, kanak, part_ppa).
address(80, kanak, kann, insult_term).
address(80, kanak, kann, taa).
address(81, kanak, kann, nii).
         .
         .
address(92, kann, kanak, part_ppa).
address(93, kanak, kann, taa).
address(94, kann, kanak, part_ppa).
address(95, kanak, kann, nii).
         .
         .
```

The output is given below:

```
assertion(kann,"shows deference to",kanak,
              "with level 1 deference")
assertion(mank,"shows high deference",
            kanak,"with level 2 deference")
assertion(kanak,"shows negative deference
          to",kann,"with level -2 deference")
Support("Asymmetric use of pronouns nii
  and niingal indicates that the speaker
  using niingal is showing deference
                          to the hearer")
support("Asymmetric use of particle taa
  indicates that the speaker using taa is
  giving high negative deference to
                          the hearer")
support("Asymmetric use of particle
 ngal is an  indication that the speaker
   is giving a high level of deference to
                               the hearer")
support("Asymmetric use of personal
    name indicates that the speaker
    avoiding the name of the hearer
    is showing deference to the hearer")
justification("Participant ",kann,"
      uses niingal to ",kanak," ")
justification("Participant ",mank,"
       uses niingal to ",kanak," ")
justification("Participant ",kanak,"
             uses nii to ",kann," ")
justification("Participant ",kanak,"
         uses nii to ",mank," ")
justification(kanak,"uses taa to",
                          kann,"","")
justification(kann,"does not use taa
                    to ",kanak,"","")
justification(mank,"uses ngal to",
                          kanak,"","")
justification(kanak,"does not use
                  ngal to",mank,"","")
justification(kanak,"addresses ",kann,
            "with the personal name","")
```

The system correctly concludes that both Kann and Mank give deference to Kanak. Kann shows level 1 deference but receives level -2 deference from Kanak, while Mank exhibits level 2 deference (and receives level -2 deference from Kanak). The detailed explanations and justifications for these claims are given as well.

VIII. Conclusion and Future Work

We have discussed a methodology for extracting information about the relationship of people from transcripts of their conversations. The methodology is applicable to a wide variety of languages. We use Russian and Tamil for illustration. We have concentrated on the information implicit in the use of terms that refer to people (pronouns, terms of address etc.).

The core of our method is based on the use of AnsProlog as a representation and reasoning language. We have constructed an AnsProlog theory of the use of the relevant language indicators in Russian and also in Tamil. Our work has shown that AnsProlog is useful for representing the default reasoning needed to draw reasonable conclusions from conversational data. With additional background information, AnsProlog will draw the appropriate different conclusion. Additionally, the declarative nature of AnsProlog enables us to extract common pieces of code that will be needed to construct this type of software for many different natural languages.

Currently we are building and testing the software to extract the information from the natural language conversation and produce the input to the AnsProlog theory. We expect to have reasonable (although not perfect) performance. In the future, we will consider utilizing PLog[34] (a version of AnsProlog that incorporates probabilistic information) to make better use of the frequencies of the occurrences of forms. Additionally, we will extend both our theories and natural language extraction tools to produce and utilize more information[14] that can be used to help draw conclusions about the nature of the relationship between the participants and their roles in the social group of which they are a part.

ACKNOWLEDGMENT

We thank Gregory Aist, Chitta Baral, and Joohyung Lee for useful discussions related to the work reported here. We thank Greg Gelfond for building the Russian extractor and also Anubhav Gupta and Shubhendra Singh for their work on the Tamil extractor.

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the US Army. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

REFERENCES

[1] W. Labov, *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press, 1972.

[2] M. Rosaldo, "I have nothing to hide: The language of ilongot oratory," *Language in Society*, vol. 2, pp. 193–223, 1973.

[3] W. Foley, *Anthropological Linguistics: An Introduction*. Malden, Mass: Blackwell Publishers, 1997.

[4] S. Poedjosoedarmo, "Javanese speech levels," *Indonesia*, vol. 6, pp. 54–81, 1968.

[5] C. Geertz, *The Religion of Java*. Chicago: The University of Chicago Press, 1960.

[6] A. Duranti, *Linguistic Anthropology*. Cambridge, UK: Cambridge University Press, 1997.

[7] W. Downes, *Language and Society*. Cambridge, UK: Cambridge University Press, 1998.

[8] R. Fasold, *The Sociolinguistics of Society: Introduction to Sociolinguistics Volume I*. Oxford, UK: Basil Blackwell, 1984.

[9] ——, *The Sociolinguistics of Language: Introduction to Sociolinguistics Volume II*. Oxford, UK: Blackwell Publishers, 1990.

[10] R. Brown and A. Gilman, "The pronouns of power and solidarity," in *Style in Language*, T. A. Sebeok, Ed. Cambridge, MA: The MIT Press, 1960, pp. 253–276.

[11] R. Brown, *Social Psychology*. New York: The Free Press, 1965.

[12] P. Friedrich, "Structural implications of russian pronominal usage," in *Language, Context, and the Imagination: Essays by Paul Friedrich*, A. S. Dill, Ed. Stanford, California: Stanford University Press, 1979, pp. 126–167.

[13] E. Bates and L. Benigni, "Rules of address in italy: A sociological survey," *Language in Society*, vol. 4, no. 3, pp. 271–288, 1975.

[14] C. B. Paulston, "Pronouns of address in swedish: Social class semantics and a changing system," *Language in Society*, vol. 5, no. 3, pp. 359–386, 1976.

[15] W. O. Beeman, *Language, Status, and Power in Iran*. Bloomington: Indiana University Press, 1986.

[16] S. Levinson, "Social deixis in a tamil village," Ph.D. dissertation, University of California, Berkeley, 1977.

[17] R. Brown and M. Ford, "Address in american english," in *Style in Language*, D. Hymes, Ed. New York: Harper and Row, 1964, pp. 234–244.

[18] J. Cooke, *The Pronominal Systems of Thai, Burmese and Vietnamese*. Berkeley and Los Angeles: The University of California Press, 1970.

[19] C. Peirce, *Philosophical Writings of Peirce*. New York: Dover Publications, 1955.

[20] R. Jakobson, *On Language*. Cambridge, Massachusetts: Harvard University Press, 1990.

[21] M. Silverstein, "Shifters, linguistic categories, and cultural description," in *Meaning in Anthropology*. Albuquerque, New Mexico: University of New Mexico Press, 1976, pp. 11–55.

[22] ——, "The indeterminacy of contextualization: When is enough enough?" in *The Contextualization of language*, P. Auer and A. DiLuzio, Eds. Amsterdam: John Benjamins, 1992.

[23] A. Agha, *Language and Social Relations*. Cambridge, UK: Cambridge University Press, 2007.

[24] W. Geoghean, *Natural Information Processing Rules: Formal Theory and Applications to Ethnography*. Berkeley, California: Language-Behavior Research Laboratory, 1973.

[25] S. Ervin-Tripp, "On sociolinguistic rules: Alternation and co-occurrence," in *Directions in Sociolinguistics: The Ethnography of Communication*, J. Gumperz and D. Hymes, Eds. New York: Holt, Rinehart and Winston, Inc., 1972, pp. 213–250.

[26] M. Silverstein, "Indexical order and the dialectics of sociolinguistic life," *Language & Communication*, vol. 23, pp. 193–229, 2003.

[27] M. Gelfond and V. Lifschitz, "The stable model semantics for logic programming," in *Logic Programming: Proc. of the Fifth Int'l Conf. and Symp.*, R. Kowalski and K. Bowen, Eds. MIT Press, 1988, pp. 1070–1080.

[28] C. Baral, *Knowledge representation, reasoning and declarative problem solving*. Cambridge, UK: Cambridge University Press, 2003.

[29] I. Niemela and P. Simons, "Smodels – an implementation of the stable model and well-founded semantics for normal logic programs," in *Proc. 4th international conference on Logic programming and non-monotonic reasoning*, 1997, pp. 420–429.

[30] C. Baral and M. Gelfond, "Reasoning agents in dynamic domains," in *Logic Based AI*, J. Minker, Ed. Kluwer, 2000.

[31] C. Baral, M. Gelfond, and R. Scherl, "Using answer set programming to answer complex queries," in *Workshop on Pragmatics of Question Answering at HLT-NAACL 2004*, Boston, Mass., May 2004.

[32] M. Gelfond, "Going places -notes on a modular development of knowledge about travel," in *Proceedings of the AAAI 2006 Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, 2006.

[33] ——, "Answer sets," in *Handbook of Knowledge Representation*, F. van harmelen, V. Lifschitz, and B. Porter, Eds. Amsterdam: Elsevier, 2008, ch. 7, pp. 285–316.

[34] C. Baral, M. Gelfond, and N. Rushton, "Probabilistic reasoning with anwer sets," *Theory and Practice of Logic Programming*, vol. 9, pp. 57–144, 2009.

[35] P. Brown and S. Levinson, *Politeness: Some Universals in Language Usage*. Cambridge, UK: Cambridge University Press, 1987.

[14] For example, we can make use of forms that are usually thought of as indicators of politeness phenomenon [35]. Additionally, we can perhaps find information about the dialect of the participants [1], [8], [9] that will give us some information about things like social group or class, or level of education.