Computational Linguistics in the Netherlands Journal 3 (2013) 193-216 Submitted 06/2013; Published 12/2013

Interpersonal stance in police interviews: content analysis

Rieks op den Akker H.J.A.OPDENAKKER@UTWENTE.NL
Merijn Bruijnes M.BRUIJNES@UTWENTE.NL
Rifca Peters R.M.PETERS@STUDENT.UTWENTE.NL
Teun Krikke T.F.KRIKKE@STUDENT.UTWENTE.NL

Human Media Interaction, University of Twente, Enschede, The Netherlands

Abstract

A serious game for learning the social skills required for effective police interviewing is a challenging idea. Building artificial conversational characters that play the role of a suspect in a police interrogation game requires computational models of police interviews as well as of the internal psychological mechanisms that determine the behaviour of suspects in this special type of dialogues. Leary's interactional circumplex is used in police interview training as a theoretical framework to understand how suspects take stance during an interview and how this is related to the stance and the strategy that the interviewer takes. Interactional stance is a fuzzy notion. The question that we consider here is whether different observers of police interviews agree on the type of stance that suspect and policemen take and express in a face-to-face interview. We analyzed police interviews and report about a stance annotation exercise. We conclude that although inter-annotator agreement on stance labeling on the level of speech segments is low, a majority voting "meta-annotator" is able to reveal the important dynamics in stance taking in a police interview. Then we explore the relation between the stance taken by the suspect and turn-taking behaviour, overlaps, interruptions, pauses and silences. Our findings contribute to building computational models of non-player characters that allow more natural turn-taking behaviour in serious games instead of the one-at-a-time regime in interview training games.

1. Introduction

Despite the growing importance of forensic research and the implementations of new laws that improved the protection of the crime suspect, interviews are still one of the most important means employed in crime investigations (Nierop 2005, Holmberg and Christianson 2002, Holmberg 2004, Snook et al. 2012). This holds for interviewing witnesses and victims as well as for interviewing suspects. Extensive research has resulted in a broad consensus and agreement about how police interviews should be conducted. Research has shown that intensive training can change interviewing behaviour, but also that benefits are obtained only when extensive efforts are made to enhance the maintenance of the learned interview practices (Lamb et al. 2002). Interview training, often with actors playing the role of a suspect, is expensive and time consuming. In a Dutch COMMIT project[1] the authors investigate in co-operation with the Dutch police how artificial intelligence based on conversational behaviour modeling can enhance police interviewing by building systems that can support the interviewer during the interview or that can be used in training interview skills. Serious games with virtual suspects have potential in training interview tactics (Luciew et al. 2011). Serious games are one of the possible tools we envision.

In a serious game a police trainee can interrogate a conversational character that plays the role of a suspect. Building these tools requires valid analysis of police interviews. Interpersonal stance is a key construct that is used in training suspect interviews. In The Netherlands at the Police Academy Leary's two-dimensional circumplex (Leary 1957), also known as Leary's Rose, is used as a framework to describe and understand how the interlocutors respond to each others' stances (Amelsvoort et al.

1. http://www.commit-nl.nl/

2010). The relation between stances taken in a police interview not only shows in the words being spoken, in postures and facial expressions, but also in the timing of speech, in interruptions and silences. We are interested in building computational suspect models that underlie turn-taking behaviour, so that the artificial suspect shows believable and natural turn-taking behaviour that fits the stances taken in the course of the interview (Jonsdottir et al. 2008, Thórisson 2002). In order to analyse the relation between suspects' behaviours, stance taking and turn-taking we transcribed speech and annotated stance of interlocutors in a number of video recorded police interviews collected at the Dutch Police Academy.

This paper reports about this annotation work where a number of annotators labeled the interlocutors' turns with categorical stance labels. We discuss the results in terms of inter-rater agreement. What are we annotating when we annotate stance in police interviews? In Section 2 we explain the model that we use as basis for our stance annotation scheme. Section 3 reviews related work in stance annotation as well as in discourse analytical studies of various styles of police interviews that are employed. Since stance is a fuzzy notion we can expect that annotators will disagree quite often if we force them to make a single choice from a fixed set of stance labels. In Section 4 we present our annotation results. We argue that, despite a low inter-rater agreement, using a majority voting system with a number of annotators we are able to identify the agreed global dynamics in stance taking over the course of an interview. Fuzzyness of labels is one of the causes of a low inter-rater agreement. In Section 5 we present a computer simulation to get an idea how the fuzziness of the stance labels used in the annotation influences our reliability measure. Then, in Section 6 we discuss some fragments from our corpus focusing on turn-taking and we formulate and present hypotheses about the relation between strategies, stance and turn-taking behaviour in police interviews. In Section 7 we conclude with a reflection on our findings and we formulate some challenges ahead.

2. Interpersonal stance

Stance is according to the English dictionary either posture or attitude, "the way in which someone stands especially when deliberately adopted". Stance and stance taking is subject of research in social psychology, in social linguistics and, more recently, in technology oriented social signal processing circles. "One of the most important things we do with words is take a stance." (DuBois 2007, p.139). For DuBois stance is realized usually by a linguistic, i.e., a social act. People take stance interactively. Stance has three aspects: evaluation, positioning and alignment. The stance taker *evaluates* (assesses or appraises) something or someone ("that's horrible"), he *positions* himself towards something, a situation or someone ("I don't like that"), in alignment or dis-alignment with others (agreement, disagreement). Stance can be affective or epistemic or both. DuBois proposes a model of stance with three components: the stance taker, the object of stance taking, and the stance that the stance taker is responding to. DuBois as well as Karkkainen consider inter-subjectivity an essential ingredient of stance taking. "Stance is not primarily situated within the minds of individual speakers, but rather emerges from dialogic interaction between interlocutors in particular dialogic and sequential contexts." (Karkkainen 2006, p.700). The notion of stance is explored by Chindamo et al. (2012) through a review and discussion of some of the relevant literature. In line with the studies mentioned above the conclusion drawn by Chindamo et al. is that: "studies of stance and stance-taking should (therefore) focus both on the expression of a speaker's stance and the reaction it leads to in his/her interlocutors." Scherer analyses interpersonal stance as a particular affect category and provides the following characterization (Scherer 2005, p.705-706): "The specificity of this category (of stance) is that it is characteristic of an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, coloring the interpersonal exchange in that situation (e.g. being polite, distant, cold, warm, supportive, contemptuous). Interpersonal stances are often triggered by events, such as encountering a certain person, but they are less shaped by

spontaneous appraisal than by affect dispositions, interpersonal attitudes, and, most importantly, strategic intention."

We are particularly interested in the interpersonal stances that the interlocutors deliberately or "automatically" take in the encounter of a police interview. Holmberg (2004) analyses the function of stance in police interviews. He actually uses the term "attitude", "the psychological tendency to evaluate and express a positive or a negative value with regard to a certain attitude object" (Holmberg 2004, p.37). Negative attitude generates avoidance, positive attitude serves an approaching function. Many police officers have been exposed to stressful events that may cause a negative attitude towards serious crime suspects, causing interview practices that are characterized by dominance and hostility. In training conversational skills and strategies police trainees in The Netherlands use Leary's theory of interpersonal relations as a framework for analyzing their own behaviour. They learn to understand the suspect's behaviour as a response to their own behaviour as expression of stance taking. Leary's model is known as the interpersonal circumplex or under the more popular name *Leary's Rose* (Leary 1957). It is presented by a circular ordering of eight categories of interpersonal behaviour, situated in a two-dimensional space spanned by two orthogonal axes, representing the two "basic dimensions of interpersonal behavior" (Kiesler 1996, p.5): affiliation (friendliness versus hostility), the horizontal axis, and power (dominance versus submission), the vertical axis. Accordingly, every form of interpersonal behavior is determined by the amount of affiliation and by the amount of dominance towards the other (see Figure 1A).

Leary (1957) formulated the principle of "reciprocal interpersonal relations": "any interactional act is designed to elicit from a respondent reactions that confirm, reinforce, or validate the actor's self-presentation and that make it more likely that the actor will continue to emit similar interpersonal acts." (Kiesler 1996, p.6). Two conversational partners are influencing each other with their stance during a dialog ('interpersonal reflexes'). Acts on the dominance dimension are complementary and acts on the affect dimension are symmetric. This means that a dominant act (e.g. power display) will elicit submissive acts, whereas an act with positive affect (e.g. cooperative) elicits another positive affect act (see Figure 1B, where the two leftmost arrows form an action-reaction pair as well as the two rightmost arrows). However, Orford (1986) pointed out that empirical evidence shows a slightly more complicated picture. He showed in a meta study that there is empirical evidence that friendly-dominant and friendly-submissive behaviour are complementary, but that hostile-dominant behaviour leads to more hostile-dominant behaviour, and hostile-submissive behaviour often leads to dominant-friendly behaviour (see Figure 1C).

Psychological research with the interpersonal circumplex model has demonstrated the value of that model for integrating a broad range of psychological topics. Rouckhout and Schacht (2000) present the results of a Dutch study with the purpose (1) to find out whether there is a circumplex structure underlying a comprehensive set of Dutch interpersonal adjectives, and (2) to construct a set of Dutch interpersonal circumplex scales. They found for each of the eight categories of Leary's rose a set of Dutch adjectives. These frequently recur when people describe the interpersonal stance of actors in a personal encounter. Table 1 (top) shows the Dutch adjectives scale from Rouckhout and Schacht (2000). Table 1 (bottom) shows a similar set of English adjective scales from Wiggins (2003). Understanding of interpersonal behaviour requires study of both the linguistic and the nonverbal levels of human communication. When annotating police interviews we used both scales for deciding what category labels best fit the stances we observed (see Figure 2).

3. Related work

3.1 Interpersonal stance annotation

Compiling a reliable corpus of emotion annotated dialogues is hard. Many emotion researchers have discussed the problems involved, such as the choice between a categorical forced choice annotation scheme versus a one or multi-dimensional continuous scheme, e.g., Craggs and McGee Wood (2004),

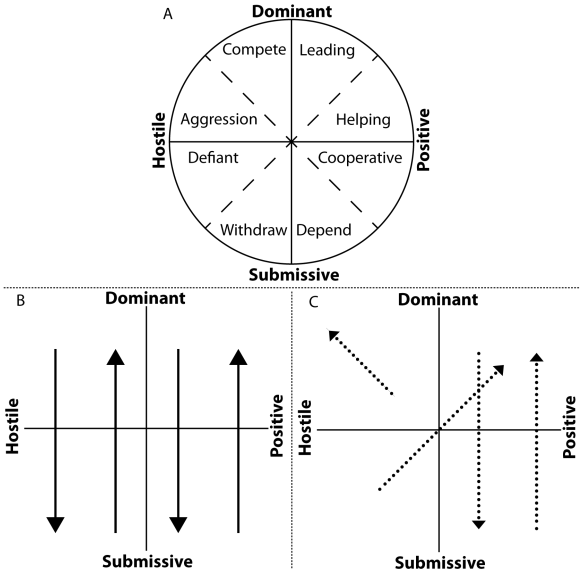

Figure 1: Leary's Rose with some adjectives used for the segments

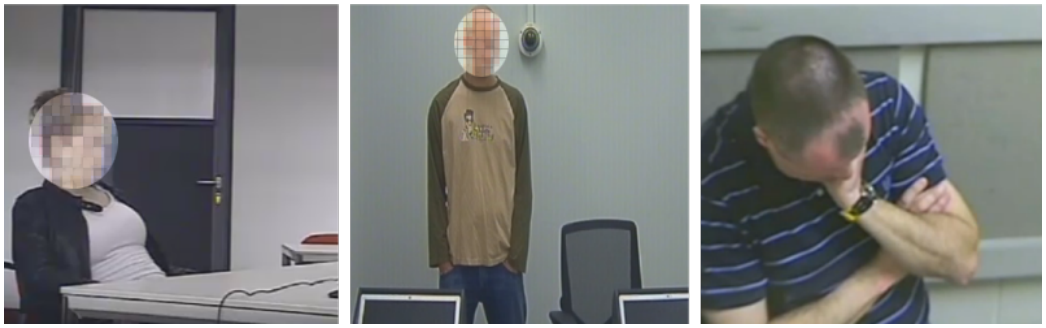

Figure 2: Stances taken by suspects during a police interview. (Faces blurred for privacy).

Compete	**Aggression**	**Defiant**	**Withdrawn**	**Depend**	**Cooperative**	**Helping**	**Leading**
Eigenwijs	Onbarmhartig	Afhankelijk	Verlegen	Pretentieloos	Minzaam	Ongedwongen	Onwankelbaar
Cynisch	Geniepig	Onbeholpen	Onderdanig	Eenvoudig	Lief	Charmant	Extravert
Gehaaid	Pretentieus	Gezagsloos	Schuchter	Moederlijk	Tolerant	Mededeelzaam	Praatgraag
Aanvallend	Wreedaardig	Tactloos	Stil	Braaf	Liefdevol	Vrolijk	Krachtig
Vrijpostig	Grof	Onpersoonlijk	Beschaamd	Doodernstig	Inschikkelijk	Spontaan	Geestdriftig
Autoritair	Bevooroordeeld	Kortaf	Gereserveerd	Nederig	Zachtmoedig	Opgewekt	Geraffineerd
Dominant	Doortrapt	Asociaal	Twijfelend	Onschuldig	Barmhartig	Galant	Hardnekkig
Vechtlustig	Brutaal	Egocentrisch	Timide	Overbezorgd	Bedaard	Attent	Fier
Impulsief	Achterdochtig	Liefdeloos	Introvert	Buigzaam	Gewillig	Voorkomend	Resoluut
Onstuimig	Tegendraads	Onoprecht	Terughoudend	Weekhartig	Onbevooroordeeld	Sociaal	Ambitieus
Ongegeneerd	Sluw	Bedrieglijk	Gesloten	Alledaags	Bescheiden	Vriendelijk	Spraakzaam
Dikhuidig	Geslepen	Oneerbiedig	Schuw	Onopvallend	Teerhartig	Loyaal	Kordaat
Onbedeesd	Listig	Arrogant	Gemaakt	Bedeesd	Zachtzinnig	Tactvol	Vol vuur
Onverlegen	Despotisch	Afgunstig	Afzijdig	Volgzaam	Plooibaar	Menslievend	Vastbesloten
Onbeschroomd	Schaamteloos	Intolerant	Naief	Meegaand	Teergevoelig	Behulpzaam	Levendig

Compete	**Aggression**	**Defiant**	**Withdrawn**	**Depend**	**Cooperative**	**Helping**	**Leading**
Firm	Crafty	Coldhearted	Unsparkling	Forceless	Unwily	Charitable	Perky
Dominant	Cunning	Cruel	Introverted	Unauthoritative	Uncunning	Tender	Enthusiastic
Forceful	Boastful	Unsympathetic	Timid	Unbold	Unsly	Sympathetic	Outgoing
Domineering	Wily	Warmthless	Bashful	Unaggressive	Softhearted	Kind	Extraverted
Cocky	Calculating	Uncheery	Shy	Unargumentative	Accommodating	Cheerful	Self-assured
	Tricky	Unneighbourly	Meek	Undemanding	Gentlehearted	Friendly	Self-confident
	Sly	Distant		Uncalculating	Tenderhearted	Neighbourly	Assertive
	Ruthless	Dissocial		Uncrafty		Jovial	Persistent
	Ironheaded	Unsociable		Boastless			
	Hardhearted	Antisocial					
	Uncharitable						

Table 1: Top: Dutch adjectives scales for the categories of the interpersonal circumplex (Rouckhout and Schacht 2000). Bottom: English adjectives scales for the categories of the interpersonal circumplex (Wiggins et al. 1988).

Busso and Narayanan (2008). One of the problems is the low inter-rater agreement due to the subjectivity of the perceptions of emotions. Similar issues arise if we want to study and annotate interpersonal stance in dialogues. The "emotion classification task" based on Leary's Rose was introduced in Vaassen et al. (2011) and executed in the Belgian project deLearyous. Annotation work can have two different goals: (1) content analysis, aiming at finding correlations between various aspects of the content, for example to study dependencies between stance taken by the police interviewer and the response stance taken by the suspect or (2) to build a train and test corpus for machine classification. The aim of Vaassen et al. was the latter. They report about the performance of a number of machine classifiers for the task of classifying the stance expressed in the words spoken by the human interlocutor when interacting with the virtual non-player character in a serious game. Our aim is of the first type and related to building a computational model of the virtual suspect and his verbal and nonverbal behaviours when being interviewed.

In the deLearyous project the focus was on the machine classification of stance in written dialog. To investigate the quality of their stance annotations, four annotators labelled a small subset of sentences from the corpus. The inter-annotator agreement was calculated using Fleiss' kappa and was found $\kappa = 0.29$ over the eight stances, and $\kappa = 0.37$ on a quarterly metric ("Leading equals Helping", etc.) (Vaassen and Daelemans 2011). Their low kappa scores are similar to the kappa scores we found in the current study and a further indication that identifying the position of a speaker on the interpersonal circumplex is a difficult task. However, they state "since the goal of the application is to simulate human behaviour, these results also imply that it is not critical for the final application to reach a perfect level of prediction. In fact, due to the subjective nature of the annotation process, an objectively 'correct' does not exist." (Vaassen et al. 2012, p.5). Using machine learning techniques to perform the stance classification task, Vaassen et al. (2010) managed to achieve an accuracy of 52.5% on the classification of the stance quadrants. This score means that their classifier can correctly label one out of two sentences into the correct quadrant in Leary's Rose, a result that might not be sufficient for a social communication training tool. However, by using more context information the classifier's accuracy could be increased sufficiently to have a convincing artificial conversational agent (Wauters et al. 2011).

Burkett et al. (2012) describe the results of stance annotation of textual chat interactions in an educational game setting using Leary's Rose. The goal is to see if personality traits can be automatically detected from dialogues and what personality traits are most prevalent over the course of the game. Six categories of Leary's Rose are used for the coding scheme: the Helping and Co-operative categories and also the Aggressive and Defiant were categorized into one. Statements that didn't fit into any of the categories were coded as neutral, indicating that there is no evidence of any of the six categories present. Two independent raters annotated a corpus of 1,000 excerpts with an average kappa of 0.65.

Allwood et al. (2012) analysed stance taking and its relation with conflict in political television debates. They define stance as "an attitude which for some time is expressed and sustained in communication, in a unimodal or multimodal manner", where attitude is taken as "a complex cognitive, emotive and conative orientation towards something or somebody"(p.1). Their qualitative analysis of a number of conflict episodes shows that some clusters of stances co-occur. Three stances are found to be characteristic for conflict episodes: aggressive, provocative, resignation. The latter differs from the first two in a number of expressive features, quiet voice and non-focused gaze. Other behaviours such as overlap, interruption and raised voice are less unique for types of conflict related stances.

3.2 Analysis of police interviews

Police interviews are analysed by psychologists and sociolinguists interested in the effectiveness and characteristics of various interview styles and interrogation tactics. Special attention is paid to the interaction between interview style and the suspect's willingness to talk freely, to admit or to deny.

Benneworth (2009) performed a discourse analytical study of UK police interviews of suspected paedophiles. Interruptions by the interviewer are a prominent phenomenon of the commonly used interrogative and accusatory interview style. Benneworth highlights the importance of encouraging uninterrupted narratives from suspects.

Jones (2008) studied differences between Afro-Caribbean and White British suspect interviews in the UK. She focussed on overlapping talk and found differences in the uptake of the interrupted talk; the Afro-Caribbean suspects' propositions were taken up to a lesser degree than any other group. This clearly shows, according to Jones, "that the police officers had more power and control than the Afro-Caribbean suspects in these interviews and potentially has something to do with race and suspect status".

Holmberg et al. (2002) reports about an explorative study among 83 criminals into the relationship between police interviewers' behaviour and suspects' inclination to admit or deny crimes. From the perpetrators' point of view they found two basic interview styles: one characterized by dominance, the other by humanity. In response to these styles the suspect will experience being respected or worried. Dominance is related to the perceptions of interviewers as aggressive, brusque and impatient. It also relates to hostility, dissociation and nonchalance. Humanity showed a positive correlation to feelings of respect, and a negative correlation with feelings of being condemned and anxiety (Holmberg and Christianson 2002, p.93).

Snook et al. (2012) examined questioning practices of Canadian police officers. Transcripts of police interviews with suspects and accused persons were coded for the type of questions asked, the length of interviewees' responses to each question, the proportion of words spoken by interviewer and interviewee, and whether or not a free narrative was requested. Results showed that, on average, less than 1% of the questions asked in an interview were open-ended, and that closed *yes or no* questions and probing questions composed approximately 40% and 30% of the questions asked, respectively. Free narratives were requested in approximately 14% of the interviews. The limited knowledge about the current questioning practices being utilized in interrogation rooms in North America provided the impetus for their study.

Beune et al. (2009) analysed sequences of dialogue acts in police interviews with suspects from different cultures. The police acts were coded using the "Table of Ten" strategies, a list of ten tactics for hostage negotiations proposed by Giebels (2002). Strategies are among others: "Emotional Appeal", "Rational Convincing" and "Direct Pressure". The suspects' acts were coded by three different content categories of inform acts. The aim of the study was to see if cultural factors (in particular the difference between high- and low context communicators) mediate the effect of the interview strategy and the responsiveness of the suspect in terms of the willingness to provide information about his own involvement, or about other's involved in the case at hand (Beune et al. 2010, Taylor et al. 2008).

All in all we can conclude that a variety of interview strategies have been identified and described in the literature; from different perspectives and with different aims. The focus is mostly on the interviewer, the police officer. The suspect's behaviours and stance are seen as dependent, in response to the interviewer's strategy and stance towards the suspect. Focal issue is the relation between the strategy the police officer follows (whether deliberately chosen or not) and the suspect's denial or admission. Although none of the studies we have seen explicitly use Leary's interactional circumplex as to describe the stances taken by the interviewer it will be clear from the above that dominance and hostility recur as important factors in studies that describe the characteristics of the predominant interview styles. Interview styles followed are related to stance taken towards the suspect and his (criminal) acts. Styles differ in the types of questions used by the police as well as in interrupting behaviour. Police officers are trained in applying various strategies and in monitoring the influence that their behaviour has on the suspect. We expect that Leary's Rose is a valid framework to describe what is going on in terms of stances taken. But do different annotators see the same things happen regarding the stances taken by the interlocutors?

4. Annotating stance in police interviews

We performed an annotation task in which annotators independently annotated police interviews with labels for the stance categories of Leary's Rose. The question is if different annotators see the same stances taken by the interlocutors. In this section we explain the corpus, the annotation effort and we present statistics on the inter-rater agreement. We argue that a majority voting "meta-classifier" is able to give a reliable picture of the essential changes in stance over the course of a police interview.

4.1 Annotation material and task

The corpus consists of video recordings of training sessions in which a police trainee interviews a suspect played by a professional actor. The interviews were recorded at the Dutch Police Academy. The trainees are introduced to a documented case that is based on a real-life case before they start the interview. Sessions have a length of 20-30 minutes. The language spoken is Dutch.

For our annotation task we selected an interview from the Wassink case (more about this case in Section 6). Speech was pre-segmented into speaker turns and transcribed. Annotators, students that were introduced to Leary's Rose and to the annotation task, independently labeled speaker turns with one of the 8 stance labels of the Rose. Annotators used the tables of Dutch adjectives (Table 1, top) to help them find the best fitting stance label. If there was no clear stance expressed a *neutral* stance category was chosen, label: Neutral. ELAN was used as annotation tool (Sloetjes and Wittenburg 2008). Inter-annotator agreement was measured using Krippendorff's alpha, a very general method for comparing an arbitrary number of annotators allowing different distance metrics on the label set (Krippendorff 2004). Labels next to each other on the Rose can be considered more similar than labels of more distant categories of the Rose. When two annotators label one and the same speaker turn with labels A and B respectively, the distance between A and B as defined by the metrics used in the alpha statistics specifies how much we penalize for this disagreement.

4.2 Annotation results

A fragment of the Wassink interview with a length of 148 speaker turns was labeled by 9 independent annotators. The total number of labeled items produced by the 9 annotators on the 148 speaker turns shows the label distribution: Neutral: 215 Leading: 309 Helping: 269 Cooperative: 218 Depend: 75 Withdraw: 70 Defiant: 89 Aggression: 23 Compete: 64.

Table 2 shows the values of the alpha statistics for the leave-one-out groups of annotators. We computed alpha with the following distance metrics (see columns 2-5 in Table 2):

- Boolean metric - two labels are equal (distance is 0) or not (distance is 1). For all annotators $\alpha = 0.24$;
- Quarterly metric - two labels in the same quarter of Leary's Rose are considered equal (Leading equals Helping, etc.). For all annotators: $\alpha = 0.42$;
- Quarterly metric with neutral - same as quarterly but Neutral is now considered equal to all other labels. For all annotators: $\alpha = 0.44$;
- Quarterly metric diagonal wise - two labels in the same quarter of Leary's Rose, where quarters are the adjacent octants separated by the two diagonal lines, are considered equal (Compete equals Leading, etc.). Neutral is considered equal to all other labels. For all annotators: $\alpha = 0.22$.

The table shows that the results are quite similar for all leave-one-out groups. Moreover, even using the tolerant penalty system for disagreement defined by the Quarterly metric with neutral the

left-out	α (bool)	α (quart)	α (quart+N)	α (diag+N)	MajVote(f)	MajVote(l)
MB	25	43	46	23	43	38
TK	27	43	43	23	25	24
DAV	23	40	42	21	60	61
MER	24	41	43	22	49	48
NIE	25	41	43	23	45	43
SJO	23	40	42	21	57	61
SOF	27	40	51	24	29	29
STE	23	41	42	20	60	59
JAN	23	40	42	21	51	55

Table 2: Krippendorff α values with different distance metrics for the 9 leave-one-out groups of annotators. The last two columns give the Cohen κ values for the annotator and the two Majority Vote "meta-annotators".

alpha values are rather low with a maximum of 0.44 for the whole group and of 0.51 when we leave one out.

From Table 2 we can draw the following conclusions. If two annotators disagree about the stance label and one chooses label A and the other label B, then it is more often the case that these two labels are in the same quarter of Leary's Rose (as for example "Leading" and "Helping", see column quart+N) than that they are neighbouring but not in the same quarter (as for example "Compete" and "Leading", see column diag+N). The difference between the values for the two distance metrics is significant (paired t-test, $p<0.01$).

What leads judges to disagree about the stance label? Formally there are two types of disagreements: noisy-like and systematic. It makes sense to analyse annotations to see what type of disagreements causes the low alpha values (Reidsma and Carletta 2008). This is not only relevant when the aim is machine classification (machine classifiers are able to learn despite noisy disagreements, see also Reidsma and op den Akker (2008)), but also when the aim is to find correlations between different phenomena in conversations, such as between stance and turn taking behaviour. Low inter-rater agreement may reveal problems judges have with the semantics of the stance labels. In the next section we will analyse the effect that the vagueness of the stance labels has on the alpha statistics. Differences in annotators' personal bias for one label can be shown as follows (Reidsma and Carletta 2008). Compare pairs of annotators of the same data. Filter out all pairs of labels where the judges agree. Perform a correlation test on the disagreed pairs. When one of the judges has a bias towards one particular label we will find a correlation. We performed this test for two judges: SJO and STE. In particular we looked at the use of labels Neutral and Helping. In 96 unequal pairs they used 4 vs. 44 times Neutral and 41 vs. 24 times Helping. A $\chi^2(1)$ one-tailed Fisher test shows that there is a very significant difference in the use of Neutral ($p = 0.007$) between the two annotators. Further analysis of this particular case shows that the difference is mainly in the labeling of the stance taken during backchannels (Yngve 1970) and short feedbacks. STE labeled them Neutral where SJO chooses a label that depends on whether the feedback is cooperative or opposing. Explicit instructions in the annotation procedure can easily avoid this type of disagreement.

4.3 Groupwise annotation by Majority Voting

From the group of 9 annotators we construct two Majority Vote "annotators" (MVA), a "meta-annotator" that assigns the label that has the majority vote of the group. Since more than one label can have the maximum number of votes, we construct two Majority Vote annotators or MVAs. Given a fixed label order, the MVA1 takes the first label that has the majority vote, where MVA2

takes the last label that has the majority vote. Since the order is chosen at random this amounts to constructing two MV annotators with random choice in case of a tie. The last two columns of Table 2 contain the κ values for the inter-rater agreement between the annotators and the two MVAs. The maximum value obtained is $\kappa = 0.61$.

Figure 3 shows the annotation of an MVA (based on a group of 5 annotators) for the Wassink interview of about 10 minutes and 300 speaker segments, the initial part of which was annotated by 9 annotators. The graph shows that the stance of the suspect changes over time from Defiant/Withdrawn to Cooperative and to Defiant again at the end of the interview. This pattern confirms the findings among the annotators when reviewing and discussing the interview. The clearly visible pattern in the majority votes indicates that using them seems like a good way to measure what is going on regarding stance and stance changes in a police interview.

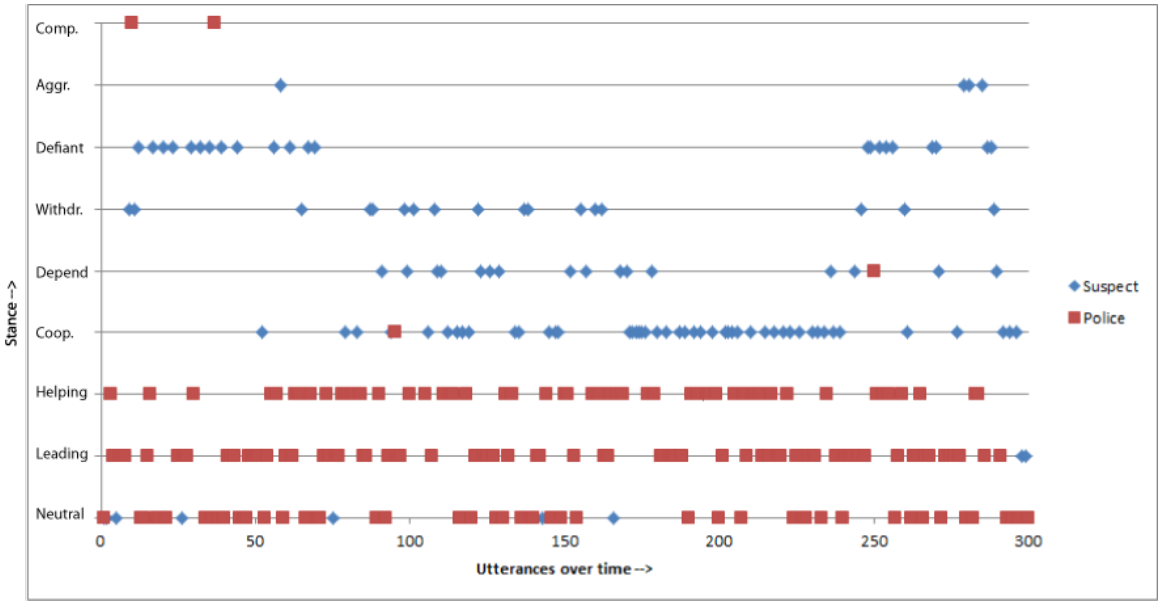

Figure 3: Majority voting applied to stance annotations of a police interview (group of 5 annotators). X-axis: the items/turns ordered along the time axes. Y-axis: the discrete stance label according to Leary's Rose (see Figure 1).

We performed a computer simulation to get an idea of a) when a majority voting system is able to detect the changes in stance over the course of an interview, and b) how the inter-rater agreement between two majority voting "meta-annotators" is related to the within groups agreement. Suppose we have two groups of annotators, each with k members. Each of the members labels the same t items. Both groups follow the Majority Voting Protocol and assign the label (there are 8 labels) that has the maximum number of votes in the group. Then we compute the inter-rater agreement within each of the two groups as well as the inter-rater agreement between the two majority voting groups. We simulated the annotations by a Gaussian distribution around mean values (in degrees on a circle). We did this with mean values 120, 240 and 180 degrees, for the first, second and the third 100 items, respectively and with standard deviations 30, 60 and 100 degrees, respectively. Note that 90 degrees corresponds to a whole quarter of the rose. The higher the standard deviation the more the majority vote will fluctuate around the truth value (the mean) and the less easy it will be to detect a real change in the stance taken. Figure 4 shows the result for a majority voting system

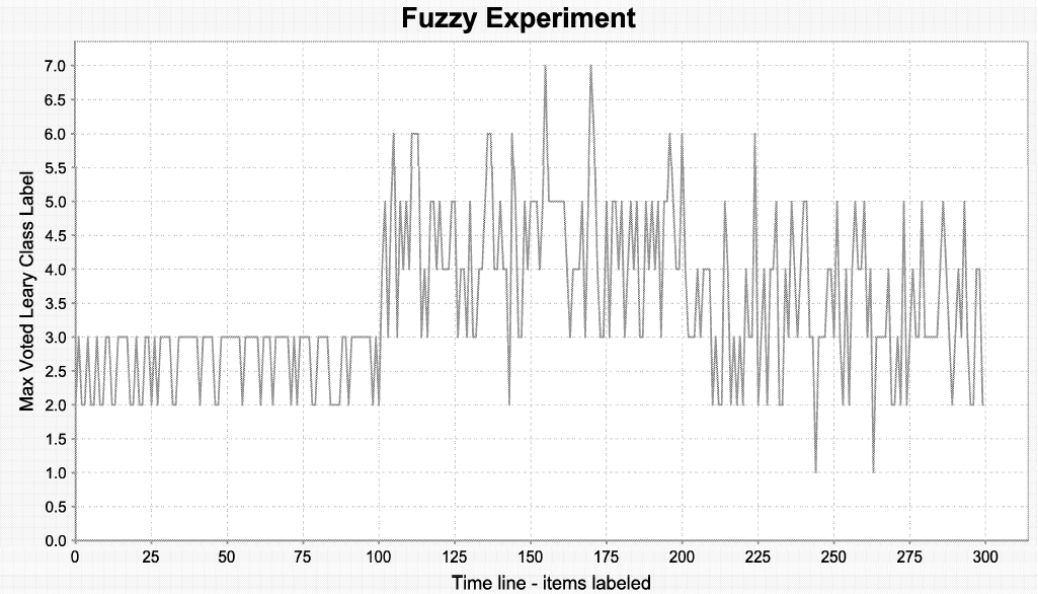

Figure 4: Majority vote annotation (simulated with a Gaussian distribution mean values 120,240 and 180 degrees and standard deviation 30, 60 and 100 degrees): 10 annotators simulated; 300 items; 8 labels. X-axis: the items/turns ordered along the time axes. Y-axes: the whole numbers correspond to the discrete stance label numbers according to Leary's Rose(1=Leading, etc.).

k annos	t items	sdev	$\alpha(1)$	$\alpha(2)$	κ(betw.)
100	300	60	0.05	0.05	0.42
100	300	45	0.11	0.11	0.80
50	300	60	0.06	0.06	0.59
50	300	45	0.11	0.11	0.66
50	300	30	0.23	0.22	0.67
10	300	60	0.06	0.05	0.15
10	300	45	0.10	0.11	0.19
10	300	30	0.23	0.22	0.42
10	300	20	0.40	0.39	0.63

Table 3: α values for two groups of annotators that label items following the majority voting procedure. Columns (from left to right): number of annotators per group; number of items labelled by each of the annotators and each of the groups; the standard deviation of the Gaussian distribution; the α values of each of the groups internally; the κ value between the two max voting groups. Normal (i.e. boolean) distance metric is used.

based on 10 simulated Gaussian annotators. It shows a clear change of stance between the first and second part, but the change is already less clear between the second and third part.

Table 3 shows how the α statistics computed between the two groups depends on the number of annotators k, the number of items t and the standard deviation (degrees of the circle) of the Gaussian. The table shows that even when the within group agreement is low, still for the highest values of α (last row: $\alpha(1) = 0.40$ and $\alpha(2) = 0.39$) the agreement between the two groups that follow the Majority Voting Protocol has a moderate Cohen κ value of 0.63.

We draw the following tentative conclusions from our findings. If we annotate stance on the level of speaker segments and we force judges to choose one of a fixed number of stance labels, we find low inter-rater agreement. Nevertheless, if we take into account the fuzzy character of the meaning of stance labels and we take the most commonly assigned label, we see a moderate agreement. This agreement seems good enough to see the global stance changes over the course of an interview.

5. Simulating annotation with fuzzy labels

We have seen that when we ask annotators to annotate speaker turns with one of 8 stance labels corresponding with the 8 octants of Leary's Rose, the inter-rater agreement is rather low. One of the causes of a low inter-rater agreement is the fuzziness of the stance labels. When do we call the stance that someone takes "leading" rather than "helping"? Or, "competing" rather than "aggressive"? We forced annotators to make a choice for one of the labels. Most of the time this will be a choice between labels of adjacent octants of the circumplex. In this section we present a computer simulation to see what the effect of the fuzziness of the adjectives is on Krippendorff's alpha measure for inter-rater agreement. In the previous section we simulated an annotator with a Gaussian distribution. Here, we provide the background for this choice.

Zadek (1965) modelled vague predicates by means of the mathematical notion of a fuzzy set. A fuzzy set F is defined by a membership function defined on a universe U of objects. $\mu_F(u \epsilon U)$ is a real number in $[0, 1]$, the grade of membership of u in the fuzzy set F. If u is a certain stance and F is for example "helping" then $\mu_F(u)$ is the grade of helpingness of the stance u. In a computational model of stance u will be a sequence of feature-values, a point in a multidimensional space. Since the introduction of fuzzy sets and fuzzy logic there has been a discussion about the interpretation of this notion of fuzziness. One of the issues was the relation between the concepts of probability and fuzziness and the question if fuzziness requires a formal logic of uncertainty that is different from the classical theory of probability. Cheeseman (1985) argues that fuzziness is uncertainty about meaning and he interpreted the membership function of a fuzzy set as a likelihood function. The idea comes from Loginov (1966) and was the basis for constructing membership functions. Given a population of individuals (our annotators) and a fuzzy concept F each individual is asked whether a given object u can be called F or not. The likelihood $P(F|u)$ is the proportion of individuals that answered "yes" to the question (Dubois and Prade 1993).

$$\mu_F(u) = P(F|u)$$

We use this interpretation of fuzziness in our simulation experiment. We assume that the points in the circumplex are generated according to a Gaussian distribution

$$N(\mu, \sigma)$$

with $\mu = u$, a real number in $[0, 360]$, representing a point on the circle, a certain "objective" stance value. If this is a reasonable model we may expect that the majority vote of a sufficient large number of annotators will coincide with the mean of the Gaussian, the "real" stance.

The more fuzzy a concept is, the larger the standard deviation σ and the more "confused" the (simulated) annotator is about the stance. The points generated are mapped on the vague labels

"Helping", etc. For example when the random value u' is in $[0, 45]$, the label generated is "Helping". This way we generate annotations controlled by a selected stance $\mu = u$ and a chosen σ. When σ grows the inter-rater agreement will be less; there will be more confusion. In that case it will be harder to see the differences between two different stances. Consequently, it will be harder to identify changes in stance taken by people over time. Our fuzzy simulation experiment give us some insight in how the vagueness of labels contributes to the α values for the inter-annotator agreement between the members of a group of annotators.

Figure 5 shows how the statistics α depends on the standard deviation. Clearly, the larger the standard deviation, the lower the inter-rater agreement. In this simulation we simulated 10 annotators each annotating 300 items with 8 different labels. The graph shows that for example when $sdev = 35$ degrees $\alpha = 0.25$. When $sdev = 45$ α becomes about 0.18.

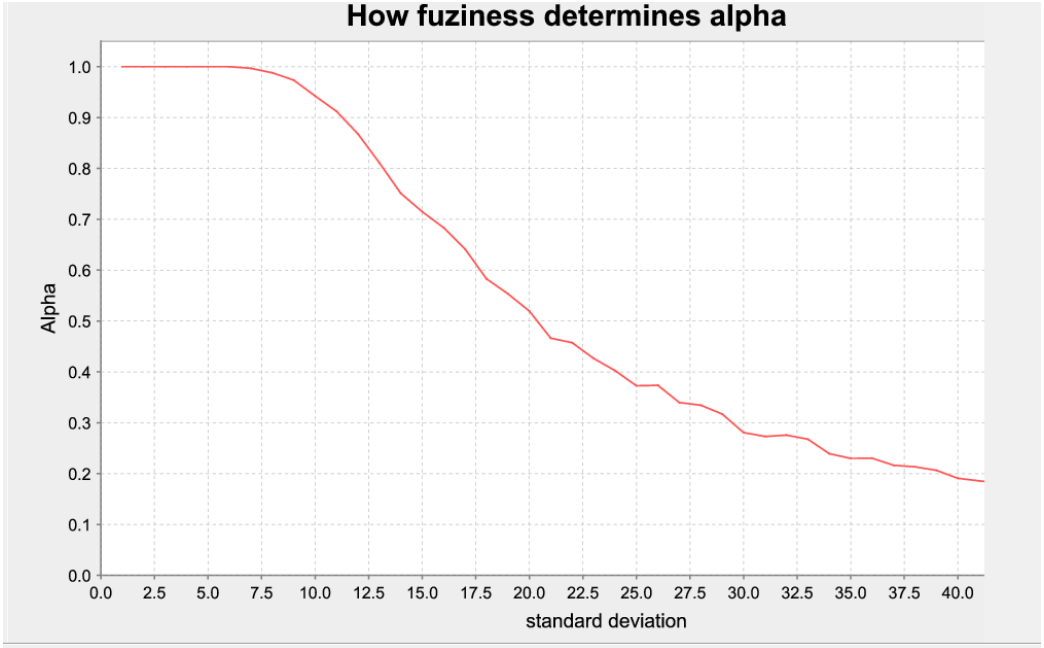

Figure 5: Krippendorff α values for the inter-rater agreement of 10 simulated annotators. (300 items; 8 labels). X-axis: the standard deviation of the Gaussian distribution that models the fuzziness of the labels. The larger the more fuzzy the labels are.

A typical random sequence of 10 annotations generated with $sdev = 45$ and mean 157 degrees is: [Aggressive, Aggressive, Aggressive, Aggressive, Competing, Defiant, Competing, Withdrawn, Aggressive, Aggressive]. The category that has the majority vote is Aggressive, which in this case coincides with the category of the mean of the distribution. Clearly, the more annotators we have the better the majority vote will equal the mean. A final caveat is in order. The use of a normal distribution as a model for the vague meanings of the stance labels seems too simple given the outcome of our analysis in Section 4. We found (see Table 2) that the distance between labels in the same quarter of the circumplex is shorter than the distance between adjacent labels not in the same quarter but in neighbouring quarters, where our model implies that they are similar.

6. Stance and turn-taking

Interviews are a special type of "talk-in-interaction" (Schegloff 2000a) in which turn-taking rules differ from those that Sacks et al. (1974) formulated for the conversation, the type of "speech exchange system" we could see as "normal". In conversations turn-taking is an interactional achievement between interlocutors that are basically operating on the same level. In the emerging conversation, speaker overlaps are rare and if they occur they are short (apart from short backchannels and listener feedbacks (Yngve 1970)). Gaps in between two speakers are also short. Moreover, exceptions are marked and need a sort of repair work. However, "normal" conversations are quite rare. In a survey interview interlocutors have distinguished roles. Basically, the interviewer is asking the questions, the interviewee answers. Role and status determine to a great extent who gets the floor (op den Akker et al. 2010). Police interviews and in particular *suspect* interviews are a special type of interviews and differ from survey interviews in that the interviewee is often not very willing to cooperate. Indeed, the various "interview strategies" (empathic, investigative, dominant) that the police officer employs result in a variety of dialogue types some of which hardly deserve the name "interview". Each type has its own turn-taking style. Here we explore how stances taken by the interlocutors are related to the turn-taking phenomena, in particular to the two observable phenomena, overlapping talk and silences. Previous studies that consider the perception of a person's turn-taking behavior and personality traits that are attributed to him are not univocal.

Robinson and Reis conclude from a perception study that interruptors are seen as less sociable and more assertive than individuals who did not interrupt (Robinson and Reis 1989). Goldberg differentiates between power and non-power interruptions and argues that some interruptions are a display of rapport, others of power (Goldberg 1990). This parallels the distinction between cooperative and competitive speech overlap (Gravano and Hirschberg 2012). But a generally cooperative stance does not exclude a competitive interruption. Interruptions by police interviewers are subject of studies because of the impact they could have on the experience of the suspect or witness (Jones 2008). We are not aware of studies that focus on the suspect's turn-taking behavior related to stance taking and interview strategy. In this section we show results of our explorative study about how turn-taking behaviour in police interviews is related to the suspect's stance.

6.1 Classification of turn-taking behaviour

Patterns in turn-taking become visible when we look at speaker transitions through vocal analysis: an acoustic silence paradigm analysing quantitative chronometrical data on something (speech) and nothing (silence between speech) (Ephratt 2008). In two party conversation, variations in the vocal activity (speech or silence) of both speakers result in four possible dialogue states: self-speaking, other speaking, none speaking and both speaking (Heldner and Edlund 2010). Transitions between dialogue states create an interaction pattern. Heldner and Edlund (2010) distinguish two different classes of silence: *gap*, a silence in which a speaker transition occurs and *pause*, a silence between two consecutive utterances of one and the same speaker. If more than one speaker is speaking there is overlapping speech, distinguishable in the different classes: $boundary - overlap$, an occurrence of overlapping speech where a speaker transition takes place and $within - overlap$, an occurrence of overlapping speech present during one continuous speech activity of one speaker.

The definitions described above are comparable to, though slightly different from, the definitions in Sacks et al. (1974). There a pause is a hypernym for silence, silence after a possible point of completion is a gap, and an extended silence at a transition relevant place is a lapse. We adhere to the terminology used by Heldner and Edlund (2010). However, because of the clear difference in and influence of the role of the interlocutors we look at interactions from a third person view and use *police* and *suspect* to refer to the active speakers (see Figure 6).

Statistics about the occurrences of interaction patterns are useful to find some global characteristics of the type of verbal interaction, but the pattern does not say much about the meaning. We miss the words and the non-verbal communicative signals that contribute to understanding the

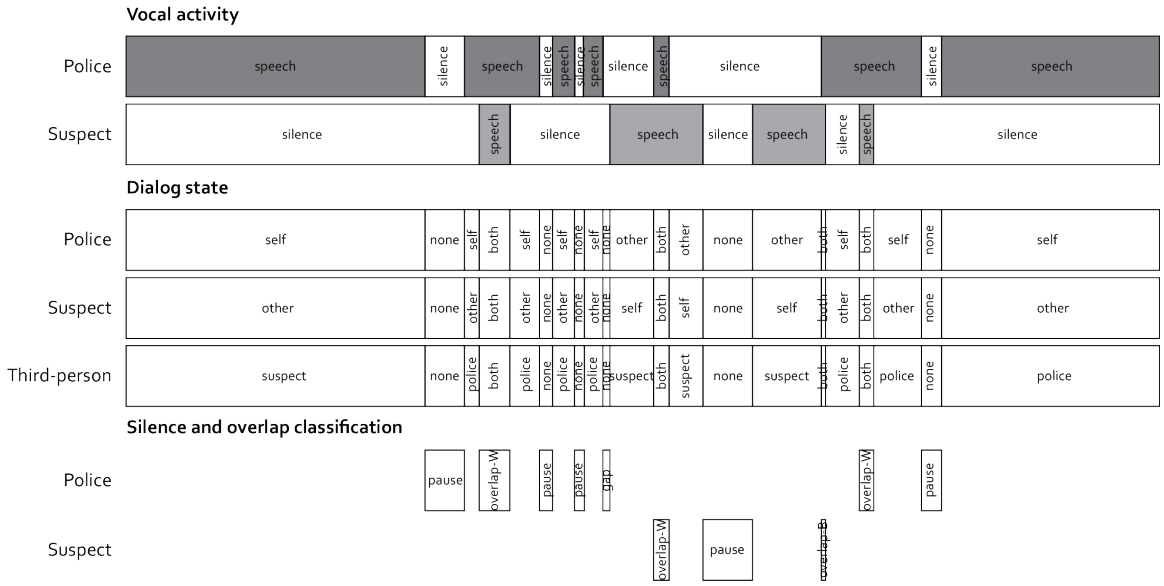

Figure 6: Illustration of original terminology by Heldner and Edlund (2010) and how gaps, pauses, between-speaker overlaps and within-speaker overlaps are classified using observable vocal activity and the dialogue state of the two speakers (police officer and suspect) in a conversation. Additionally depicted is the dialogue state from the third-person view.

meaning carried within these interactions. Occurrences of similar vocal activity patterns may carry different meanings. Also, overlapping talk can have different flavors. And, instead of considering silence as just the absence of talk we can look at silence from a socio-pragmatic point of view.

Overlapping talk is either competitive, neutral or collaborative. A competitive overlap is considered indicative for power, control or dominance or an expression of indifference, aggressiveness or hostility and manifested with high pitch and intensity. A collaborative overlap conveys rapport and is an indicator for coordination and alignment (Gravano and Hirschberg 2012). Schegloff (2000b) defines four classes of overlapping speech: 1) cooperative overlap such as assisting by completion, 2) non-problematic overlapping speech such as chorus, 3) interrupt where the speaker did not finish the utterance and did not yield the floor and 4) backchannel and short feedback, not intended to gain the floor. Overlapping speech of the interrupt class is considered problematic and needs resolution. The overlap can occur after a gap when it is not clear who has the turn, and both want to take the floor or when the listener wants to take over the speaker role of the active speaker. The question is then who gives up the fight for the floor? We suggest that stances of interlocutors are a mediating factor here.

Silence conveys meaning and is considered communicative when silence occurs where the rules dictate to speak and the silence is by choice of the speaker. Ephratt (2008) defines these silences as eloquent silence: a silence as a means chosen by the speaker with a significant communicative meaning. Verschueren (1985) distinguishes several causes of a participant remaining silent; causes that we can categorize into two groups:

1 speaker is temporally disinclined to speak; speaker is concealing something; speaker does not have anything to say;

2 speaker is unable to decide what to say next; speaker is unable to speak because of strong emotions such as amazement or grief; speaker has forgotten what he was going to say; speaker is silent because others are talking;

The causes in the first group are intentional according to Ephratt (2008). Those in the second group are considered causes of non-intentional silences from psychological inhibition (Kurzon 1995). We suggest that, as for overlapping talk, stance is a mediating factor for interpreting the semantics of the silence.

6.2 Suspect's stances in the example conversation

The interpersonal stance of the suspect, as annotated by multiple annotators, changes during the course of the conversation. A global pattern of five segments is visible in Figure 7: A) predominantly Defiant, B) variation between Defiant, Dependent and Cooperative, C) mainly Cooperative, D) predominantly Defiant and E) a final Cooperative moment. These segments are marked in Figure 7 showing the suspect's stance as it is annotated by a majority of the annotators. The occurrences of silences and overlaps *gap*, *pause*, $overlap_W$ and $overlap_B$ are shown in Figure 7 as well.

To discuss our findings we collected from the video recordings of our interview a number of samples showing silence and overlapping speech. They are transcribed according to the Jefferson convention[2]. We will provide English translations when we discuss the fragments (in the next subsection). Non-verbal behaviour essential for understanding what is going on is marked in the sample transcriptions. The samples also give the interpersonal stances of the speakers. Analysis of these samples show that the decision for certain turn-taking behaviour is related to the stances.

The suspect in our interview is Ms. Wassink. She is brought in because her neighbour filed a criminal complaint for assault. Apparently Ms. Wassink became physical after she and the neighbour got into an argument in front of their houses. The police officer read the files and invited her for an interview.

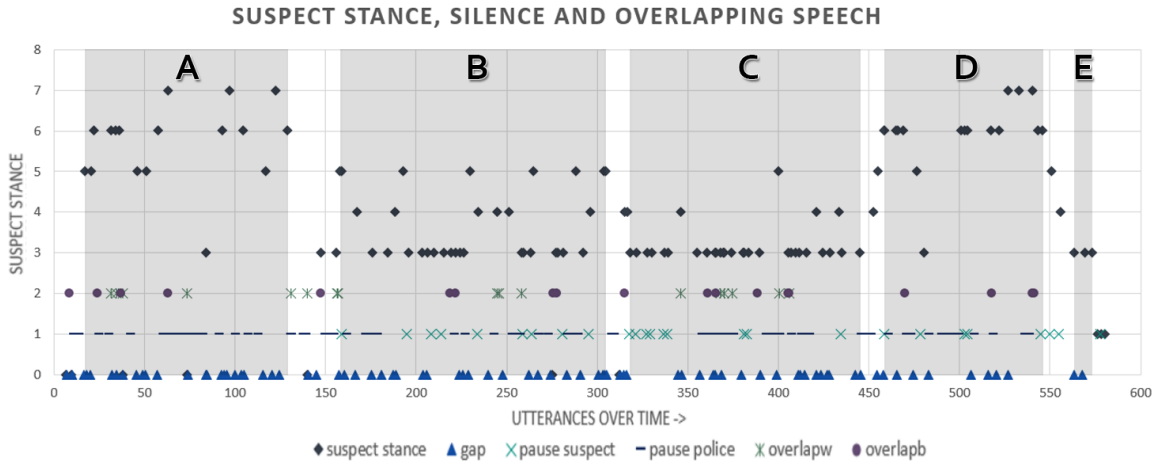

Figure 7: Segmentation of the conversation based on suspect's stance and the occurrences of silence and overlapping speech.

2. See `http://homepages.lboro.ac.uk/~ssjap/transcription/transcription.htm`, last visited 04-10-2013 (Jefferson 2004, Mazeland 2003)

After he has welcomed the suspect the police officer explains the goal of the interview in part A. The suspect is quiet and withdrawn, resulting in a monologue of the police officer with a number of *pauses* between consecutive utterances.

In part B the police officer asks concrete questions about the suspect's home situation; the suspect provides (minimal) responses. *gap* is frequently observed and *pauses* between suspect utterances appear. The silences are strategically used by the police officer to encourage the suspect to speak.

When discussing the topic of the neighbourhood and the suspect's relation with the neighbours in part C, the suspect is more talkative and *pauses* between consecutive utterances of the suspect occur frequently. Turn-taking seems to proceed without many problems. During the course of part C the police officer introduces a new topic, pets. The conversation degrades to a casual conversation style; suspect and police officer share their personal attitude towards pets. Both interlocutors are talkative, speak without being addressed by questioning, resulting in an increase of overlapping speech. Both interlocutors do understand what is said by the other. The overlapping speech does not hinder the conversation. The conversation evolves back to an interview type after the police officer initiates the new topic of yesterday's events. The suspect's willingness to talk decreases and *gaps* occur more frequently.

At the start of part D the suspect explicitly questions the relevance of the proposed question and topic. The suspect provides minimal responses and the frequency of occurrences of *pause* increases. At the very end the new topic of the argument with the neighbour is initiated by the officer. *Boundary-overlaps* occur followed by sequential *pauses* in the speech of the suspect possibly indicating the suspect claims the floor.

In part E the suspect is talkative and re-selects self as next speaker repeatedly causing *pauses* between consecutive utterances of the suspect. The contributions of the police officer are all *backchannels*.

We see that the topic of the conversation is a factor that influences the stance of the suspect, in particular, if the topic is related to the case at hand. The police officer initiates new topics. The topic influences the talkativity of the suspect and the turn-taking behaviour of both interlocutors. If the suspect is less talkative silences are more frequent. Overlapping speech is mainly present during part C in which the conversation turns into casual talk. The topic is innocent, and the strategy employed is relational, of the sort "being kind", see Giebels and Taylor (2009) and Giebels (2002), stressing shared experiences between police and suspect.

6.3 How stance mediates the meaning of silence and overlapping speech

The relation between power and the decision to start speaking and continue speaking is visible in samples 1, 2 and 3. Samples 1 and 2 take place at the boundary of parts B and C. In both fragments the stance of the suspect is Positive and Submissive while discussing the topic neighbourhood.

```
660  Police:   SB  wat voor buurt is het?
                   (0.9)
665  Suspect:  NN  nj[aa ]
669  Police:   SB     ge zellige buurt, of juist niet?
                   (.)
665  Suspect:  SO  ja vink wel
```

Sample 1: occurrence of post-continue overlap where the initiator (police) wins the floor resulting in boundary-overlap—sample taken from Wassink starting at 00:04:33.540.

In Sample 1 the officer asks a question addressing the suspect (line 660: *What type of neighbourhood is it?*). After a moment of silence (duration 0.9) the suspect initiates a response (line 665: *yeah*). The officer decides to rephrase a more concrete question and re-selects self, initiating

overlapping speech by a short delay in onset (line 669: *A cosy neighbourhood, or not?*). The suspect immediately stops speaking and yields the floor to the officer, resulting in a *boundary-overlap* interaction.

In Sample 2 the suspect has the turn but decides to stop speaking before sentence completion (line 816: *Yeah, well then they stay in their houses alone with their dog but na yeah then eh*). After a fairly long silence (duration 1.18) the officer selects self as next speaker. Shortly after onset the suspect continues the previous speech act initiating overlapping speech (line 825: *those kinds of things*). The suspect stops speaking causing *within-overlap* interaction. The police officer extends the turn with a second sentence but stops speaking before sentence completion (end of line 821: *that you like to ...*). After a short *pause* (duration 0.27) the officer re-selects self (line 830: *...do things together with other people*). The officer stops speaking at a point of possible completion where the suspect selects self and almost seamlessly starts speaking, taking up the officer's suggestion (line 835: *yeah I like that together with people, yeah*).

816	Suspect:	SO	ja gewoon helemaal eeh dan in hun huis blijven zitten enzo
			met hun hond alleen maar ja naja dan eh
			(1.18)
821	Police:	SB	ieder voor zi[ch, god voor ons allen]en jij zegt van, geeft
825	Suspect:	SO	van die dingen
	Police:		eigenlijk een beetje aan van dat je toch wel een
			gemeenschapsmens bent. dat je dat graag
			(0.27)
830	Police:	SB	met andere mensen iets samen wilt doen
			(.)
835	Suspect:	SO	ja ik vind wel leuk memensen samen, ja

Sample 2: occurrence of post-continue overlap where the initiator (suspect) loses the battle for the floor resulting in within-overlap—sample taken from Wassink starting at 00:05:39.150.

Sample 3 takes place in part C where the stance of the suspect is fully *cooperative*. While discussing the topic pets, the conversation type devolved to casual conversation —losing the interviewer and interviewee roles. After a short comment by the officer (line 888: *How nice*) the suspect selects self as next speaker (line 893: *Yes, my mum in particular*). The police officer re-selects self as next speaker initiating overlapping speech a brief moment after onset. The suspect continues speaking resulting in a *within-overlap*. However, the suspect does not complete the sentence, pauses and takes up what was said by the officer during the overlapping speech.

888	Police:	SB	wat leuk
			(.)
893	Suspect:	SO	ja[ah mn moeder hoor vooral]ik dr
897	Police:	SB	ik heb dr vier dr vier
			(0.64)
902	Suspect:	SO	ja echt?

Sample 3: occurrence of overlapping speech—sample taken from Wassink starting at 00:06:07.500.

Samples 1 and 2 show that the interlocutor with higher power is more likely to win a battle for the floor. This results in a *within-overlap* when the suspect initiates the overlapping speech and in a *boundary-overlap* when the police officer initiates the overlapping speech. The examples also

illustrate (our hypothesis) that the interlocutor with higher power is more likely to self-select as next speaker during a silence after an incomplete utterance of the other.

Sample 3 illustrates that even a slight change in power—resulting from a change in conversation type—increases the likelihood for a suspect to self-select as next speaker and continue speaking during a battle for the floor.

The relation between affiliation and the pragmatical interpretation of silent responses is illustrated by samples 4, 5 and 6.

Sample 4 takes place in part B where the stance of the suspect is submissive and tends to be positive. The officer asks a polar question addressing the suspect (line 347: *I've understood that you live in the Broekstreet*). The question-answer adjacency pair is taken up by a non-verbal affirmation in the form of a head nod.

```
347  Police:  SB  ik heb begrepen dat jij aan de Broekstraat woont
                  ((looks up at suspect))
                  (0.73) ((suspect headnod))
352           SB  ja?
                  (.)
357           SB  woon je daar alleen of met iemand anders?=
```

Sample 4: occurrence of pause where participation of the other interlocutor takes place by non-verbal behaviour—fragment taken from Wassink starting at 00:02:23.406.

Sample 5 takes place in part A when the stance of the suspect is submissive and hostile. *Pauses* in the speech of the police officer occur frequently and sequentially between incomplete utterances of the police officer. In lines 206-216 (*Your name is Sabrina. Sabrina Wassink I've gathered. I don't know you. You don't know me either*) the officer wants to check some data from his sheet, keeping an eye on the suspect to read her response when he pauses waiting for confirmation. The suspect's responses are non-verbal and minimal. She doesn't feel much like getting to know each other as the officer proposes.

```
191  Police:  NN  ehm
                  (0.69)
196           NN  maar
                  (0.4) ((suspect looks away))
201           NN  goed, ik eheh h
                  (0.63)
206           SB  je heet Sabrina
                  (0.77) ((suspect head nod))
211           SB  Sabrina Wassink heb ik begrepen
                  (0.46) ((suspect head nod))
216           SB  ik ken jou verder niet
                  (0.58) ((suspect head shake))
221           SB  jij kent mij ook niet
```

Sample 5: occurrence of sequential pauses filled with a non-verbal response if the utterance was a polar question—sample taken from Wassink starting at 00:01:14.913.

Sample 6 takes place in part D when the global stance of the suspect is hostile and predominantly submissive. The police officer asks a polar question (line 1235: *Did anything out of the ordinary happen after that?*). The suspect provides a non-verbal negative response. The police officer elaborates

on the question several times (line 1243: *or after you returned from Zwolle?*; line 1253: *also not in the neighbourhood?*; line 1258: *Did you bump into someone with a dog?*) thereby indicating that he has some specific information he wants to check. The suspect responds repeatedly with head shakes and incertitude facial expressions but after some time she starts speaking, admitting in a reluctant way that she may have seen the dog (line 1263: *yeah, eh might be but hhh*).

```
1235  Police:   SB  is dr nog iets bijzonders gebeurd daarna?=
1239            SB  =############= ((suspect head shake))
1243            SB  =of nadat je van Zwolle teruggekomen bent?
                    ((suspect head shake))
                    (3.12) ((suspect head shake))
1248            NN  hmm
                    (2.2)
1253            SB  ook nie in de buurt?
                    (1.51) ((suspect incertitude facial expression))
1258            SB  nog iemand tegen gekomen met een hondje?
                    (0.96)
1263  Suspect:  TO  ja eh vast wel maa hhh
```

Sample 6: occurrence of sequential pauses while the suspect aims to conceal information—sample taken from Wassink starting at 00:08:05.616.

Sample 4 shows an example of a cooperative suspect who provides a response in a non-verbal way nodding her head. This silence could be interpreted as *have nothing to say.* In Sample 5 the suspect provides a complete non-verbal response but given the more hostile stance of the suspect and other non-verbal behaviour (i.e. looking away from the speaker) the silence can be interpreted as the suspect feeling *disinclined to speak.* In Sample 6 the withdrawn stance of the suspect indicates that the absence of speech is aimed at *concealing* possibly incriminating information.

We have seen that stance of the suspect on the affiliation axis shows a correlation with interpretation of silence by the suspect. A silent but contributing response is related to either timidness (positive stance) or withdrawal (hostile stance). A silent response intended to withhold information is only observed in relation to a hostile stance. The global stance on the power axis shows a correlation with talkativeness. The global stance of the suspect is predominantly submissive (low power). This submissiveness is related to: 1) the decision to speak only when selected as next speaker by the other interlocutor and 2) the decision to yield the floor during overlapping speech independently of the initiator and onset of overlap. This difference in speaker activity reduces when the power levels of the suspect and the police officer become more equal.

7. Conclusions

Based on the outcome of our reliability analysis we are convinced that Leary's theoretical model makes sense as a framework for analyzing and describing the interactional stance that people take towards each other in a social encounter. Leary's Rose provides terminology to come to a reasonable agreement between subjects about "what is going on" in a police interview in terms of stance taking and the dynamics of the behaviours and the effects they have on turn taking by the interlocutors. Sometimes, it turns out to be hard for outside observers to tell what the stance of the participant is. The multi-flavored expression of stance in general seems to be an important cause of disagreement between judges. Another cause is the forced choice annotation procedure and the fuzzy character of the meaning of the words used. Judges were forced to make a choice where it is often hard to make a choice. Analysis of the annotated corpus has shown that it is indeed the case that when annotators

disagree in their choice they choose labels that are next to each other in Leary's Rose. Moreover, they agree about the direction in the two main dimensions. If we use a majority voting meta-annotator the inter-subjective content of Leary's view on stance is clearly revealed in the changes in stance. Although judges do often not agree about the exact label on the level of speaker segments they do agree on the global dynamics of the stance changes during a police interview.

The general lesson we learn from this is something we already knew before but sometimes forget: do not use too precise measures for fuzzy phenomena. Based on our analysis of the annotations the annotation instruction can be improved, in particular regarding backchannels and short feedbacks.

An explorative study into the relation between suspect's stance and the types of overlaps, interruptions, and silences indicates that the interview topic and in particular how the topic is related to the case at hand is an important factor that influences the stances taken by the subject. Stances and roles seem to be mediating factors for the meaning of overlaps and silences in suspect interviews. The challenge ahead now is to incorporate these findings into a computational model of a suspect character so that it simulates believable turn-taking behaviour that expresses the suspect's stance as a response to the learner's strategy and stance taking. More interviews need to be analysed to substantiate our preliminary findings and to build better models for different suspect characters. This way we expect to contribute to improving the learner's experience while playing the serious game of interviewing crime suspects.

Acknowledgments

We are grateful to Hans Weijkamp, Arend de Vries and Ronald Waanders, lecturers police interview at the Police Academy, for introducing us to the fascinating world of police interviewing. We thank the Master of Science students for contributing to the annotation exercise. The critical and constructive comments we received from the reviewers on earlier versions of this paper have been a great help to improve the paper. This publication was supported by the Dutch national program COMMIT.

References

Allwood, J., M. Chindamo, and E. Ahlsen (2012), On identifying conflict related stances in political debates, *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Conference on Social Computing (SocialCom)*, pp. 918–925.

Amelsvoort, A., I. Rispens, and H. Grolman (2010), *Handleiding verhoor*, Stapel & De Koning.

Benneworth, K. (2009), Police interviews with suspected paedophiles: A discourse analysis, *Discourse & Society* **20** (5), pp. 555–569.

Beune, K., E. Giebels, and K. Sanders (2009), Are you talking to me? Influencing behaviour and culture in police interviews, *Psychology, Crime & Law* **15** (7), pp. 597–617.

Beune, K., E. Giebels, and P.J. Taylor (2010), Patterns of interaction in police interviews: The role of cultural dependency, *Criminal Justice and Behavior* **37** (8), pp. 904–925.

Burkett, C., F. Keshtkar, A.C. Graesser, and H. Li (2012), Constructing a personality-annotated corpus for educational game based on Leary's Rose framework, *FLAIRS Conference*, pp. 147–150.

Busso, C. and S.S. Narayanan (2008), Recording audio-visual emotional databases from actors: A closer look, *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, pp. 17–22.

Cheeseman, P. (1985), Probabilistic versus fuzzy reasoning, *Uncertainty in Artificial Intelligence Annual Conference on Uncertainty in Artificial Intelligence (UAI-85)*, Elsevier Science, Amsterdam, NL, pp. 85–102.

Chindamo, M., J. Allwood, and E. Ahlsen (2012), Some suggestions for the study of stance in communication, *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Conference on Social Computing (SocialCom)*, pp. 617–622.

Craggs, R. and M. McGee Wood (2004), A two dimensional annotation scheme for emotion in dialogue, *Proceedings of AAAI Spring Symposioum on Exploring Attitude and Affect in Text: Theories and Applications*.

Dubois, D. and H. Prade (1993), Fuzzy sets and probability: Misunderstandings, bridges and gaps, *Second IEEE International Conference on Fuzzy Systems*, pp. 1059–1068 vol.2.

DuBois, J. W. (2007), The stance triangle, *in* Englebretson, R., editor, *Stancetaking in Discourse: Subjectivity, evaluation, interaction*, John Benjamins Publishing Company, pp. 139–182.

Ephratt, M. (2008), The functions of silence, *Journal of Pragmatics* **40** (11), pp. 1909 – 1938.

Giebels, E. (2002), Beïnvloeding in gijzelingsonderhandelingen: De tafel van tien, *Nederlands Tijdschrift voor de Psychologie* **57**, pp. 145–154, Bohn Stafleu Van Loghum.

Giebels, E. and P.J. Taylor (2009), Interaction patterns in crisis negotiations: Persuasive arguments and cultural differences, *Journal of Applied Psychology* **94**, pp. 5–19.

Goldberg, J.A. (1990), Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport-oriented acts, *Journal of Pragmatics* **14**, pp. 883–903.

Gravano, A. and J. Hirschberg (2012), A corpus-based study of interruptions in spoken dialogue, *Interspeech 2012*.

Heldner, M. and J. Edlund (2010), Pauses, gaps and overlaps in conversations, *Journal of Phonetics* **38** (4), pp. 555–568.

Holmberg, U. (2004), *Police Interviews with Victims and Suspects of Violent and Sexual Crimes: Interviewees' Experiences and Interview Outcomes*, PhD thesis, Department of Psychology, Stockholm University.

Holmberg, U. and S.-A. Christianson (2002), Murderers' and sexual offenders' experiences of police interviews and their inclination to admit or deny crimes, *Behavioral Sciences & the Law* **20** (1-2), pp. 31–45, John Wiley & Sons, Ltd.

Jefferson, G. (2004), Glossary of transcript symbols with an introduction, *in* Lerner, G.H., editor, *Conversation Analysis: Studies from the first generation*, John Benjamins, pp. 13–31.

Jones, C. (2008), UK police interviews: A linguistic analysis of Afro-Caribbean and white British suspect interviews, *International Journal of Speech Language and the Law*.

Jonsdottir, G.R., K.R. Thórisson, and E. Nivel (2008), Learning smooth, human-like turntaking in realtime dialogue, *Intelligent Virtual Agents*, Vol. 5208/2008, Springer Berlin / Heidelberg, pp. 162–175.

Karkkainen, E. (2006), Stance taking in conversation: From subjectivity to intersubjectivity, *Text & Talk* **26** (6), pp. 699–731.

Kiesler, D. J. (1996), *Contemporary Interpersonal Theory and Research, Personality, Psychopathology and Psychotherapy*, Wiley.

Krippendorff, K. (2004), Reliability in content analysis: Some common misconceptions and recommendations, *Human Communication Research* **30(3)**, pp. 411–433.

Kurzon, D. (1995), The right of silence: A socio-pragmatic model of interpretation, *Journal of Pragmatics* **23** (1), pp. 55–69.

Lamb, M.E., K.J. Sternberg, Y. Orbach, I. Hershkowitz, D. Horowitz, and P.W. Esplin (2002), The effects of intensive training and ongoing supervision on the quality of investigative interviews with alleged sex abuse victims, *Applied Development Science* **6:3**, pp. 114–125.

Leary, T. (1957), *Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation*, Ronald Press, New York.

Loginov, V.I. (1966), Probability treatment of Zadeh membership functions and their use in pattern recognition, *Engineering Cybernetics* pp. 68–69.

Luciew, D., J. Mulkern, and R. Punako Jr. (2011), Finding the truth: Interview and interrogation training simulation, *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*.

Mazeland, H. (2003), *Inleiding in de conversatieanalyse*, Uitgeverij Coutinho.

Nierop, N.M. (2005), Het verdachtenverhoor in Nederland: Wat wordt verhoorders geleerd, *Nederlands Juristenblad* **17**, pp. 887–890.

op den Akker, R., M. Theune, K. Truong, and I. de Kok (2010), The organisation of floor in meetings and the relation with speaker addressee patterns, *Proceedings of the 2nd International Workshop on Social Signal Processing, SSPW '10*, ACM, New York, pp. 35–40.

Orford, J. (1986), The rules of interpersonal complementarity: Does hostility beget hostility and dominance, submission?, *Psychological Review* **93** (3), pp. 365–377.

Reidsma, D. and J. Carletta (2008), Reliability measurement without limits, *Computational Linguistics* **34** (3), pp. 319–326.

Reidsma, D. and R. op den Akker (2008), Exploiting 'subjective' annotations, *in* Artstein, R., G. Boleda, F. Keller, and S. Schulte im Walde, editors, *Proceedings of the COLING Workshop on Human Judgments in Computational Linguistics*.

Robinson, L.F. and H.T. Reis (1989), The effects of interruption, gender, and status on interpersonal perceptions, *Journal of Nonverbal Behavior* **13** (3), pp. 141–153.

Rouckhout, D. and R. Schacht (2000), Ontwikkeling van een Nederlandstalig Interpersoonlijk Circumplex, *Diagnostiekwijzer* **4**, pp. 96–118.

Sacks, H., E.A. Schegloff, and G. Jefferson (1974), A simplest systematics for the organization of turn-taking for conversation, *Language* **50**, pp. 696–735.

Schegloff, E.A. (2000a), *Accounts of Conduct in Interaction: Interruption, Overlap and Turn-Taking*, New York: Plenum.

Schegloff, E.A. (2000b), Overlapping talk and the organization of turn-taking for conversation, *Language in Society* **29** (01), pp. 1–63.

Scherer, K.R. (2005), What are emotions? And how can they be measured?, *Social Science Information* **44** (4), pp. 695–729.

Sloetjes, H. and P. Wittenburg (2008), Annotation by category: ELAN and ISO DCR, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Snook, B., K. Luther, and H. Quinlan (2012), Let 'em talk!: A field study of police questioning practices of suspects and accused persons, *Criminal Justice and Behavior* **39**, pp. 1328–1339.

Taylor, P.J., K. Jacques, E. Giebels, M. Levine, R. Best, J. Winter, and G. Rossi (2008), Analysing forensic processes: Taking time into account, *Issues in Forensic Psychology* **8**, pp. 45–57.

Thórisson, K.R. (2002), Natural turn-taking needs no manual: Computational theory and model, from perception to action, *Multimodality in Language and Speech Systems*, Kluwer Academic Publishers, pp. 173–207.

Vaassen, F. and W. Daelemans (2010), Emotion classification in a serious game for training communication skills, *Computational Linguistics in the Netherlands 2010: Selected papers from the 20th CLIN meeting*, LOT.

Vaassen, F. and W. Daelemans (2011), Automatic emotion classification for interpersonal communication, *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 104–110.

Vaassen, F., J. Wauters, F. Van Broeckhoven, M. Van Overveldt, W. Daelemans, and K. Eneman (2012), DeLearyous: Training interpersonal communication skills using unconstrained text input, *Proceedings of ECGBL 2012, The 6th European Conference on Games Based Learning*.

Verschueren, J. (1985), *What People Say and Do with Words*, Norwood, N.J.: Ablex.

Wauters, J., F. Van Broeckhoven, M. Van Overveldt, K. Eneman, F. Vaassen, and W. Daelemans (2011), DeLearyous: An interactive application for interpersonal communication training, *Proceedings of CCIS Serious Games: The Challenge*.

Wiggins, J.S. (2003), *Paradigms of Personality Assessment*, Guilford Press.

Yngve, V. (1970), On getting a word in edgewise, *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567–77.

Zadeh, L.A. (1965), Fuzzy sets, *Information and Control* **8**, pp. 333–353.