Annotations
and
Subjective Machines

Of Annotators, Embodied Agents, Users, and Other Humans

DENNIS REIDSMA

PhD dissertation committee:

Chairman:
Prof. dr. ir. A. J. Mouthaan, Universiteit Twente, NL
Promotor:
Prof. dr. ir. A. Nijholt, Universiteit Twente, NL
Assistant-promotor:
Dr. ir. H. J. A. op den Akker, Universiteit Twente, NL
Members:
Prof. dr. N. Campbell, ATR SLC Labs, Japan
Prof. dr. J. Carletta, University of Edinburgh, UK
Prof. dr. F. M. G. de Jong, Universiteit Twente, NL
Dr. A. Popescu-Belis, IDIAP, Martigny, CH
Dr. Z. M. Ruttkay, Universiteit Twente, NL
Prof. dr. M. F. Steehouder, Universiteit Twente, NL
Prof. dr. D. R. Traum, University of Southern California, USA
Paranymphs:
J. D. van Belle
J. A. J. Brand

CTIT Dissertation Series No. 08-121
Center for Telematics and Information Technology (CTIT)
P.O. Box 217 – 7500AE Enschede – the Netherlands
ISSN: 1381-3617

AMIDA Publication
The research reported in this thesis has been supported by the European IST Programme Projects AMI (Augmented Multi-party Interaction) and AMIDA (Augmented Multi-party Interaction with Distant Access), FP6-506811 and FP6-033812. This thesis only reflects the author's views and funding agencies are not liable for any use that may be made of the information contained herein.

Human Media Interaction
The research reported in this thesis has been carried out at the Human Media Interaction research group of the University of Twente.

SIKS Dissertation Series No. 2008-29
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN: 978-90-365-2726-2
ISSN: 1381-3617, No. 08-121

ANNOTATIONS AND SUBJECTIVE MACHINES

OF ANNOTATORS, EMBODIED AGENTS, USERS, AND OTHER HUMANS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. W.H.M. Zijm,
on account of the decision of the graduation committee
to be publicly defended
on Thursday, October 9, 2008 at 13.15

by

Dennis Reidsma

born on January 15

in Amersfoort, The Netherlands

Promotor: Prof. dr. ir. A. Nijholt

Assistant-promotor: Dr. ir. H. J. A. op den Akker

ISBN: 978-90-365-2726-2

Acknowledgements

> *“ ‘It’s a dangerous business, Frodo, going out of your door,’ he used to say. ‘You step onto the Road, and if you don’t keep your feet, there is no knowing where you might be swept off to.’ ”*
> **J.R.R. Tolkien**

Several years ago, I stepped outside that door, onto the road of research. I never knew where I would be swept off to next — so many things happened in those years. Luckily, as soon as you step outside that door, onto that road, there you find other people, walking along the same road. One usually does not write a PhD thesis alone. I have had the great luck to be indebted to so many people that I cannot even begin to thank you all, individually, for what you have meant to me these years. For encouragement and support, knowledge and wisdom, for helping me with my work or providing much-needed distraction. Nevertheless, I will make an attempt.

There exist no supervisors who fit my style of working better than Anton and Rieks. They gave me the freedom to go where I liked when I couldn’t settle down on a topic, and challenges, active help, encouragement and inspiration for the topics that I did work on. I certainly hope our work together does not end here! Jean Carletta suggested the joint work that finally led to this thesis, and we held endless discussions to get things just right. I consider myself lucky for having collaborated with her, and for having her in my committee. I also want to thank my whole committee, for their willingness to participate in my defense. Several of them additionally went way beyond the call of duty in providing me with feedback to improve my thesis. Andrei, David, Jean (again), Franciska, and Zsofi, thank you for your extensive reviews. I hope I managed to answer some of your questions.

The Human Media Interaction group has been an extremely supportive environment during the past years. The secretaries, Charlotte and Alice, were always helpful. The staff members were always willing to talk when I had a specific problem related to their work. Most especially, I would like to mention Dirk for his support at the end of my PhD work, who read everything and assured me it would work out, at a time I couldn’t yet believe it, and Jan (now at CAES) and Franciska for their help at the beginning, when they got me started on research in the first place. Lynn, without your correction work throughout the years, it would have taken so much longer for me to learn to write adequately in English! All language errors remaining in this thesis are, of course, entirely mine.

Throughout the years that I worked there, the HMI group has always had a large and lively tribe of PhD students. They made for a stimulating environment, where few things are taken for granted and boredom is a referent-free concept. A few of the PhD students should be mentioned explicitly. First and foremost, Rutger, for four very good years together, and Nataša, for inspiration and data: without their work, friendship, and collaboration, there would have been nothing for me to write about. Arthur (at the Database group) and Ronald provided places to flee to when I was stuck and needed an intelligent and slightly skewed outlook on science and on life to help me further. Herwin gave me just the right demo at the right time. My new roommates Boris and Christian showed unlimited patience in suffering my moods during the hard months of writing this thesis. Wietse (at Bosch GmbH), finally, has helped me tremendously in getting my bibliography in order.

Several BSc and MSc students, in passing through our group, worked on projects that inspired me and helped me further in my work. Of these, three should be mentioned here. Pieter, Mark, and Rob were responsible for one of the most inspiring, non-thesis-related projects that I worked on. The Virtual Orchestra Conductor was absolutely the best work-related distraction I could have wished for!

I had the good fortune to be part of several large international collaboration efforts. MUMIS saw me started; AMI and AMIDA offered an environment with broad opportunities for developing myself further. Those people from AMI who served duty as annotators for the huge amounts of material, and Jonathan Kilgour, who kept our data safe and our programs stable, deserve a special mention in any AMI-related thesis. We owe them so much.

My family deserves special thanks, both for their honest interest in the process and content of my PhD work and for their patience during the times that I was not there. I missed too many birthdays and other events during the last half year.

It seems I am not really one for doing only a single thing at a time. Throughout the years, there have been an amazing number of special groups and people who filled my daily life. Thanks to André, there are the fencers of Gascogne and of Agilité, who kept me fit, helped me empty my mind, and taught me at least as much as I taught them when I became a teacher. The anarchistic and unorganized group of musicians calling themselves the 'Gonnagles' made me go places that I had never before imagined going — both figuratively and literally. The 'Parkweg en uitpandige inboedel' were my family in Enschede throughout the years, and meant safety and friendship to me. Individually and collectively, they deserve thanks for more than I can ever explain. And among those, finally, we find my house mates, friends, and paranymphs, Jenno and Jaap: thank you for everything.

Dennis Reidsma
Enschede, October 2008

Contents

1 Introduction 1
1.1 Corpus Based Research and the AMI Project 3
1.2 Research Questions and Thesis Structure 4

I Theory 7

2 Data and Inter-annotator Agreement 9
2.1 Methods from Content Analysis 9
2.2 Reliability Metrics 12
2.3 Fitting Data to Metric 15
2.3.1 Unitizing 15
2.3.2 Annotations with Graph Structure 18
2.4 Sources of Disagreement 19
2.5 Types of Content 22
2.5.1 Manifest Content 22
2.5.2 Pattern Latent Content 23
2.5.3 Projective Latent Content 23
2.5.4 Choosing Between the Types of Content 24
2.5.5 Inter-Annotator Agreement for the Types of Content 26
2.6 From Data Quality to Data Use 26
2.6.1 Constraining Possible Use of the Data 27
2.6.2 Evaluating and Explaining Machine-Learning Results 27

3 Some Limits of Reliability Measurement 29
3.1 The Problem 30
3.2 Method 31
3.3 Results 33
3.3.1 The Case of Noise 33
3.3.2 The Case of Over-Using a Label 35
3.4 Discussion 36
3.5 Implications 37

II Praxis 39

4 Moving to the AMI Corpus 41
4.1 The AMI Corpus: an Overview 41
4.1.1 Visions of AMI 41
4.1.2 Recorded Scenario Meetings 42
4.2 The AMI Dialog Act Annotations 44
4.3 The AMI Addressee Annotations 47
4.4 The AMI Focus of Attention Annotations 48
4.5 Summary 48

5 Contextual Agreement 49
5.1 Basic Agreement and Class Maps for Addressee 50
5.1.1 Reliability for the Addressee Label UNKNOWN 51
5.1.2 Class Map: Group/A/B/C/D vs Group/Single 52
5.2 The Multimodal Context of Utterances 53
5.3 Finding More Reliable Subsets 54
5.3.1 Context: Focus of Attention 54
5.3.2 Context: Elicit Dialog Acts 55
5.4 Discussion and Summary of Addressing Agreement 56
5.5 Contextual Performance of Classification 56
5.5.1 Approach 57
5.5.2 Results 57
5.6 Summary and Discussion 57

6 Explicitly Modeling (Inter)Subjectivity 59
6.1 Modeling Subjectivity for 'Yeah' Utterances 60
6.1.1 'Yeah' Utterances 60
6.1.2 Division of the Data into Training and Test Sets 61
6.1.3 Approach 61
6.1.4 Results 62
6.1.5 Discussion 64
6.2 Exploring More Aspects of Subjective and Voting Classifiers 64
6.2.1 Generalisation to Other Data 65
6.2.2 Performance Improvement or Context Selection? 66
6.2.3 Precision and Recall 67
6.2.4 Ensemble Learning 68
6.3 More Questions 71
6.3.1 Questions About Improving the Results 73
6.3.2 Questions About Underlying Fundamentals 73
6.3.3 Final Thoughts 74

III Reflection 75

7 Designing for Interaction 77
7.1 Application Contexts for Recognition Modules 77
7.2 Requirements for Subjectively Annotated Data 78
7.3 Two New Types of Classifiers 79
7.4 Subjective Machines? 79

8 Conclusions 81

Bibliography 85

Abstract 91

Samenvatting 93

SIKS Dissertation Series 95

Chapter 1

Introduction

People interact. They enter into polite conversations, participate passionately in debates, hold meetings for work and for volunteer committees; they dance, teach, sell apples, and compete in sports; and all these activities involve communication with other people. A certain type of researcher spends his time recording these activities as video and audio. Subsequently he annotates them manually — or has other people do this — describing in varying levels of detail what happened during the interaction. The result is an *annotated corpus of recorded human interactions*. Corpus based research may be carried out for several reasons. Human interactions are analysed in order to learn more about how people interact, to develop automatic detection and recognition systems for interaction behavior, to develop technology designed to support humans in their daily activities (ambient intelligence), or to build animated lifelike Virtual Humans that interact in a human-like way with computer users. These topics will all be elaborated further in Section 1.1.

Researchers who make use of multimodal annotated corpora are always presented with something of a dilemma. On the one hand, one would prefer to have research results that are reproducible and independent of the particular annotators who produced the corpus that was used to obtain the results. This issue can be characterized in the way that Bakeman and Gottman [1986] and Krippendorff [1980] described it: they stated that annotators should be interchangeable and the individuality of annotators should influence the data produced by them as little as possible. They said that one needs data which is annotated without disagreement between annotators, because disagreement is, among other things, a sign of errors and lack of reproducibility. Researchers spend a lot of effort on developing ways to determine the amount of disagreement and finding out how much disagreement is 'too much'. On the other hand, labeling a corpus is a task which involves a judgement by the annotator and is therefore, in a sense, always a subjective task. This subjectivity is, of course, a matter of degree. For some annotation tasks different annotators can realistically be expected to always have the same judgements, such as when annotators are asked to label episodes where somebody raised his hand in a recorded interaction. In that case, disagreement between annotators may indeed be caused by one or both of the annotators having made errors. Other annotation tasks are more strongly subjective. They require an annotator to *interpret* the communicative behavior being annotated. Expressing yourself in communication is a very personal activity, full of human variability; everybody behaves in his or her own unique way, and *therefore, as an observer, will also judge the communicative behavior of others in his or her own unique manner*. Throughout this thesis, the terms 'subjective annotation task' and 'subjective annotation' refer to annotation tasks in which the judgements made by the annotators are strongly dependent on the personal way in which the annotator interprets communicative behavior. Different annotators will, for such tasks, produce different annotations of the same recorded interaction. For example, two people who are asked to point out all episodes where people are 'being ironic' in the same recorded interaction will surely produce different annotations. This is not just because irony is sometimes difficult to see, but also because everyone to a certain extent has his own idea about what irony is. Such personal views of what irony is may not overlap perfectly. In that case, the amount of similarity or agreement between the annotations will not just depend on how many errors the annotators made. It will also be influenced by the amount of *intersubjectivity* in the judgements of the annotators.

The difference between an annotation task being subjective or not has clear relevance for the use that is subsequently made of the annotated data. Conclusions drawn from annotations that are *not* subjective relate to the (communicative) behavior that was observed in the recordings, possibly somewhat contaminated by errors in the annotations. One consequence of a certain annotation being subjective is that conclusions drawn from the data may reflect not just the human behavior observed in the recorded interaction but also the world view of the annotator. The subjective aspects of an annotation task also carry over into the development and use of technological applications that are built using the annotations. One can, for example, think about an automatic summarization system that is supposed to indicate the fragments in a meeting where the participants had a disagreement about the issue being discussed, or where they were particularly enthusiastic about a certain proposal. The system has been trained to make judgements similar to those of the annotator whose data was used to train the system. Since the judgements of this annotator were subjective, it is not certain that the system will judge in a way that fits also with what the participants in the meeting thought of their own interactions. Consequently, the resulting summarizations are at risk of seeming irrelevant to the participants for whom they were made.

This thesis is an investigation into the diverse relations that exist between the *agreement* in judgements of different annotators, the *quality* of the annotated corpus (number of errors in the data, and reproducibility of the annotations) and the *use* of the data for developing interactive applications. The data can be used in different ways. Development and evaluation of machine-learning modules for automatic recognition of the annotated behavior are strongly represented in this thesis, but other applications such as building Virtual Humans or the design of interaction patterns are also affected by the quality and nature of the data. Throughout the thesis there is a special focus on behavior for which annotation is a subjective task. The rest of this introductory chapter first describes the area of corpus based research in more detail as the background against which this thesis originated. After that the research questions underlying the thesis will be presented and an overview of the structure of thesis will be given.

1.1 Corpus Based Research and the AMI Project

This thesis has been written in the context of the AMI (Augmented Multi-party Interaction) and AMIDA (Augmented Multi-party Interaction with Distant Access) projects.[1] These two projects were concerned with developing technology to support meetings, leading to increased effectiveness and efficiency. They have been carried out by the 15-member multi-disciplinary AMI Consortium, a large collaboration of (mostly European) academic and industrial partners. The project members have very broad interests, ranging from hard core signal processing algorithms to organizational psychology and from developing marketable applications to doing pure research. One of the most visible results of this collaboration has been the development of the AMI Meeting Corpus, a large corpus with annotated synchronized audio and video recordings of 100 hours of meetings. The corpus contains many different layers of annotation describing the communicative behavior of the participants in the meetings [Carletta, 2007].

As was already mentioned at the start of this introduction, such corpora can be used for many different goals. In areas such as psychology, content analysis, and linguistics, a carefully annotated corpus can help in understanding verbal and non-verbal language use, communication, social interaction processes and the myriad relations between all the different elements that make up human behavior. While such research is concerned with describing and understanding human behavior, there is also a lot of work on *generating* human-like interactive behavior. In that research area, people build animated lifelike Virtual Humans that interact with humans and each other in a human-like way. In order to make the Virtual Human display the right behavior, people turn to corpus analysis to find out what kinds of behavior are best used in which situations. This concerns not only speech and the accompanying gestures and facial expressions, but also many completely different activities, as can be seen in the Virtual Conductor project at the Human Media Interaction group [Bos et al., 2006; Reidsma et al., 2008b]. In that project, an artificial orchestra conductor was built that not only displays the appropriate conducting behavior for a piece of music, but also interactively leads and corrects an ensemble of human musicians based on audio analysis of the music being played. That system depends on knowledge about the interactive behavior of human conductors that has been extracted from an analysis of a corpus of recorded conducting sessions [ter Maat et al., 2008].

Finally, an annotated corpus is often used to develop detection and recognition technology for the range of human behavior present in the corpus. This technology is to be used in smart environments to support humans in their activities. In the AMI and AMIDA project the goal is to help users have efficient and effective meetings. Smart environments, or Ambient Intelligent environments, need to be able to *perceive* and *interpret* all kinds of subtle human behavior in order to support the

[1] `http://www.amiproject.org/`

inhabitants of the environment in their daily activities. A high-quality annotated corpus can play a central role in Ambient Intelligence research [Reidsma et al., 2005c]. Aspects such as attitude and mood, the flow of conversations taking place in the environment, and the intentions underlying the actions of the inhabitants cannot be interpreted automatically without knowledge of how people behave in daily interaction. Within the AMI Consortium, a vast range of recognition technologies are being developed [Al-Hames et al., 2007]. There is work on dialog act segmentation and labeling [Dielmann and Renals, 2007; op den Akker and Schulz, 2008; Verbree et al., 2006], addressee detection [Jovanović, 2007], recognition of visual focus of attention [Ba and Odobez, 2006], decision tracking [Hsueh and Moore, 2007], emotion recognition [Müller et al., 2004], summarization [Kleinbauer et al., 2007], and many other subjects. These technologies have been developed for systems that support easy browsing and retrieval of information from past meetings *after* the meeting took place [Moran et al., 1997; Whittaker et al., 2008], but also for systems that assist the users in many ways *during* the meeting [Rienks et al., 2006, 2007]. Such technologies and applications form the background against which this thesis should be read.

1.2 Research Questions and Thesis Structure

The start of the introduction presented two topics that are present throughout this thesis. On the one hand there is the method of corpus based research and the application contexts that go with it, discussed in Section 1.1. On the other hand there are the subjective aspects inherent to elements of this method: in the judgements made by the annotators producing the corpus and in the perception that end users of applications have of their own and others' (inter)actions.

As said before, in the method of corpus based research, the common way of assessing whether the corpus is fit for purpose centers on the level of inter-annotator agreement that can be achieved on the annotation task. One problem in this context is that it is not well understood how the relation between inter-annotator agreement and 'being fit for purpose' works out if the errors that led to reduced inter-annotator agreement are not homogeneous. A second problem is that there is a tension here between *errors* and *subjectivity*: both can lead to reduced inter-annotator agreement, but they can have quite a different impact with respect to the question of the corpus being fit for purpose. The research questions of this thesis are defined against the background of these two problems related to corpus based research, inter-annotator agreement and subjectivity.

The main, abstract, research question is as follows.

Main RQ — What are the relations between inter-annotator agreement, subjective judgements in annotation, and whether a corpus is fit for the purpose for which it was constructed?

This question is addressed in this thesis by answering three concrete questions, that are explained in detail below.

One of the most prevalent ways of assessing the quality of an annotated corpus using inter-annotator agreement analysis is to compare the level of agreement to a certain fixed threshold. If the threshold is exceeded, the corpus annotations are considered to be of sufficient quality for any purpose. Usually, no additional quality analysis is carried out. This practice is addressed in Chapter 3 by the following research question.

RQ 1 — What is the relevance of placing a threshold on the level of inter-annotator agreement for assessing the reliability of a corpus, especially if the errors that caused reduced agreement may not have been homogeneous?

There is a growing interest in, and amount of work with, corpora with annotations that are inherently subjective. These annotations often have a quite low overall level of inter-annotator agreement. The next research question, answered in Chapter 5, concerns possible ways of making such annotations more useful for machine learning or other purposes by removing the less reliable parts of the data.

RQ 2 — Given annotation with a low inter-annotator agreement, how can one pinpoint more reliable subsets of the annotated data, for which a higher agreement was achieved?

Another way of making subjective annotations more useful when the different personal points of view from the annotators have led to low level of inter-annotator agreement is to find out which information and relationships that can be derived from the annotations are common sense and which stem from idiosyncrasies of annotators. This approach is addressed in Chapter 6 by the last research question.

RQ 3 — Is it possible to find out how subjective the annotations are, and to model the subjectivity explicitly as it relates to the overlap and disjunctions between the personal points of view of the annotators, using machine-learning methods?

Answering the research questions involves both an analysis and evaluation of some existing methods for determining the quality of an annotated corpus as well as the development of new methods on real data.

The rest of this thesis is structured as follows. Chapter 2 discusses past related work on data inspection, reliability analysis, and agreement metrics; furthermore the chapter presents relevant related work on subjectivity in annotation tasks and on the relation between inter-annotator agreement/data quality on the one hand and data use on the other. Chapter 3 looks closer into the most used method for determining the level of inter-annotator agreement in annotated data: calculating a reliability metric such as Krippendorff's α [1980]. Some shortcomings of that method are discussed that are related to the use of the annotated data for machine-learning purposes, concluding that additional methods for analysing annotated data are needed. Chapter 4 presents the corpus data that will be used throughout this thesis. Chapter 5 concerns *contextual agreement*. The inter-annotator agreement of the addressee annotations from the AMI corpus is analysed in detail, showing that it is possible to define subsets of the data, on the basis of the multimodal context, that have a higher level of inter-annotator agreement than the overall annotation. In Chapter 6, subjective annotations are the main topic of investigation. Inter-annotator disagreement for subjective annotation tasks is likely to be caused by *systematic* differences in the way different annotators interpret human interaction, more so than for other types of content. In this chapter, a method is presented that explicitly models overlap and divergence in the subjective judgements of different annotators using machine learning. This leads to the discussion of two new concepts in Chapter 7, namely classifiers as subjective entities and classifiers as embodiment of consensus objectivity. Chapter 7 also contains some generalized ideas and recommendations for designing interactive systems. The thesis ends with conclusions in Chapter 8.

Part I

Theory

Chapter 2

Data and Inter-annotator Agreement

In order to understand how people have been analysing the quality of annotated corpora it is necessary to see the place of annotation within the larger methodological context of corpus based research. There are many types of research fields where annotated corpora play a major role. In **Corpus Linguistics** and **Computational Linguistics**, annotated data is collected and analysed in order to increase our understanding of human language use and for developing automatic recognition or processing algorithms. For example, researchers use an annotated corpus to develop automatic dialog act classification algorithms or to find possible correlations between topic changes and posture shifts. In **Ambient Intelligence research**, annotated data is collected and analysed in order to find out how the daily activities of people are structured, how they can be supported, and, again, to serve as training material for developing recognition algorithms. Besides verbal and nonverbal language use, this involves such various issues as attitude detection, action recognition, intention recognition, and many other topics. The methodology of **Content Analysis** can be thought of as concerning the labeling of material (texts and other content) with certain concepts (see below) present in the material in order to be able to answer research questions using the annotated data [Krippendorff, 1980]. Although this could equally be a description of any of the above fields, the term is usually associated with the analysis of content from television shows, newspaper articles, teaching materials, and similar material, for concepts (often socially inspired), that may be present in the data or not, such as violence, qualitative evaluations of politicians, opinions about commercial products in advertisements, and morally tinted messages.

2.1 Methods from Content Analysis

As the central tenets of Content Analysis have been thought about for a long time and are generally cast in relatively generic terms, Content Analysis literature sources will be used to introduce the main issues for corpus based research. Other researchers, working in the fields of (Computational) Linguistics or Ambient Intelligence have reached for Content Analysis literature as well, for the same reasons.

Note that it is not the intention to give a complete survey or exposé of the field of Content Analysis. This introduction merely presents the main issues to give the reader a background for the later sections of this chapter. The material in this section is freely summarized from the works of Krippendorff [1980] (and the second edition [2004a]); Poole and Folger [1981]; Weber [1983]; Bakeman and Gottman [1986]; and Potter and Levine-Donnerstein [1999], in which the interested reader can find more information about the material introduced here.

Consider a research project in the field of Content Analysis aimed at investigating the relation between the presence of elements targeting trust, quality of living, and cost in advertisements on the one hand and changes in sales of a product on the other hand. The first step in such a project would be to define a **research question**. In this case, the question may be 'how can one improve sales by targeting trust, quality of living and low cost in advertisement campaigns?' The next step is for the researcher to **determine which concepts are relevant** when answering the question. This concerns both the broad categories of concepts that should be taken into account when answering the research question, but also the specific classes that relate to those categories. In the advertisement example, it is at least necessary to have an idea of what the global concepts 'trust' and 'quality of living' are taken to mean. But it is also necessary to specify in detail different types of trust (for example, 'trust in the competence of the producer' and 'trust in the benevolence of the producer') and the ways in which trust can be a relevant factor in an advertisement. It can be present or not, it can be addressed directly (using a phrase such as "We want the best for you") or indirectly (displaying serious, knowledgeable people in white coats), and so forth. One also needs to think about the relation one expects to hold between the concepts.

The researchers could decide to answer their research question by Content Analytic means. That would involve investigating the content of a collection of advertisement campaigns on the presence and absence of the identified concepts, and correlating the results with, for example, information about sales fluctuations for the different advertised products. For this, the collection of advertisement campaigns (the content) must be turned into an annotated corpus. First, an **annotation schema** is defined that specifies how the content is to be annotated with the presence of the concepts. This involves many decisions about, for example, the granularity with which the annotation is performed, whether an annotator can assign multiple labels to one item, whether a 'bucket class' can be used to indicate that an item cannot be labeled with any of the concepts, and many other decisions. Defining an annotation schema is like the operationalization of the concepts under investigation. In the advertisement example, the schema may, for example, define that every advertisement will be assigned exactly one label specifying the type of trust alluded to in the advertisement. In contrast, the schema may define that every advertisement can be assigned several of the six possible types of trust when more than one is relevant for a single advertisement. An **instruction manual** is written for training the observers who are going to annotate the material. It is important to make sure that the annotators properly understand the annotation schema and are able to recognize the concepts accurately. The manual explains the relevant concepts ("Someone is benevolent when (s)he wants the best for others. In an advertisement, trust in the benevolence of the producer may be evoked or referred to in the following ways: ..."), but may also give very specific practical advice ("With respect to the cost factor, every advertisement should be labeled with one label from the following: NoCostFactor, Cheap, NotExpensive, ExpensiveLuxury"). Using the annotation schema, the annotators will then **annotate a selection of content** that was chosen for being representative and sufficient for answering the research question. Table 2.1 displays an example of the (fictitious) result of having all advertisements in a television show labeled by two different annotators.

Advertisement	Annotator 1	Annotator 2
1	NoCostFactor	NoCostFactor
2	Cheap	NotExpensive
3	ExpensiveLuxury	Cheap
4	Cheap	Cheap
5	NoCostFactor	NoCostFactor
6	NoCostFactor	NoCostFactor
7	NotExpensive	NotExpensive
8	Cheap	Cheap
9	Cheap	NoCostFactor
10	ExpensiveLuxury	ExpensiveLuxury

Table 2.1: Fictitious labeling of advertisements with respect to how the advertisement refers to the cost factor, as produced by two (fictitious) annotators.

Finally, the complete set of annotated data can be used to **make quantitative and qualitative inferences** that help answering the research question(s). The researchers in the advertisement example might combine the annotations with quantitative data on sales fluctuations in the periods in which the annotated advertisements ran. They may, for example, conclude that "Advertisements for expensive luxury items that refer to the benevolence of the producer and that stress the improved quality of living that comes from buying the product have a more positive impact on sales than those that mostly stress the price aspect." Researchers in other fields than Content Analysis may build annotated corpora for other reasons. One frequently occurring task in corpus based research is **machine learning**. To stay close to the example above, someone might annotate a corpus of advertisements with an annotation schema concerning trust and quality of living in order to train a classifier to automatically classify advertisements with respect to those concepts. The researcher then also needs to extract features from the advertisements on the basis of which they can be classified. Nevertheless, the process described above still applies to such cases.

All possible uses of annotated corpora require the annotations to be of high enough **quality**. This means that, firstly, the annotators should all have the same understanding of the concepts they are annotating. If they do not, for example because they were not well trained or because the concepts have not been clearly defined, conclusions drawn from one annotator's data may not be valid for another annotator, or for the intended consumer of the research findings. Secondly, the annotators should not make too many errors. Fatigue, or problems with understanding a language foreign to the annotator, or simply careless work may cause the annotator to make real errors in the annotation task. Such errors may introduce noise in the data, making it harder to draw conclusions or to train classifiers. Issues leading to low quality annotations are elaborated in Section 2.4. For *socially oriented* annotation schemas, the risk of problems occurring is usually higher than for *physically oriented* annotation schemas [Bakeman and Gottman, 1986, page 17] (see Section 2.5).

To assess the risk of the annotations being of too low a quality, one can try to quantify the number of errors made by the annotators. One method to do that works from the assumption that annotators do not usually make the *same* errors. To assess the quality of the annotations, a certain (limited) amount of data is annotated several times by multiple annotators. This set of data is called the 'reliability data' [Krippendorff, 2004a, page 219]. Given the assumption above, errors made by the annotators would show up in this data as a lack of **inter-annotator agreement**: the same item is assigned different labels by the different annotators, not only for one item, but many times (in Table 2.1 there are 3 such points of disagreement). Disagreement may also be caused by other problems such as the annotation schema requiring annotators to make distinctions that make no sense and cannot be applied to the actual content. If there is too much disagreement in the data, the annotation schema and/or the particular annotators are then said to be **unreliable**. If an annotation schema is not reliable, chances are that the results of the research are also not **reproducible** or **valid**, meaning one cannot trust the conclusions drawn from the data to be really true. Inter-annotator agreement analysis is a very powerful tool for assessing the reliability and validity of an annotation schema. Note, though, that the basic assumption that annotators do not make the same errors depends on other aspects such as the source and kind of errors (see Section 2.4). The relation between agreement and reliability is not without complications, as the discussions in Chapter 3 show. Furthermore, disagreement may also stem from differences in the subjective interpretations of annotators rather than from 'errors' that are a deviation from some hypothetical objective ground truth. This topic is taken up in Chapter 6.

2.2 Reliability Metrics

One of the major techniques for determining the level of inter-annotator agreement achieved for an annotation task is to calculate a chance-corrected reliability metric. It was introduced in computational linguistics by Carletta [1996], having been in use in other fields such as content analysis [Cohen, 1960; Krippendorff, 1980] and medicine (see the survey of Fleiss [1975] for an overview of early work on this topic) for a very long time before that.

When one needs to know the level of agreement between two annotations of the same content produced by two different annotators, a naive way is to count the number of instances where two annotators agree on the assigned label compared to the total number of instances that the annotators had judged. This results in an observed agreement percentage. For the sake of argument, let us assume that this observed agreement is 70%, as is the case in the example in Table 2.1. Note, next, that had the annotators been assigning labels blindly and randomly, without actually looking at the content, one should have expected a certain amount of agreement by chance as well. Consider the following quotation from Carletta:

"Taking just the two-coder case, the amount of agreement we would expect coders to reach by chance depends on the number and relative proportions of the categories used by the coders. For instance, consider what happens when the coders randomly place units into categories instead of using an established coding scheme. If there are two categories occurring in equal proportions, on average the coders would agree with each other half of the time: each time the second coder makes a choice, there is a fifty/fifty chance of coming up with the same category as the first coder. If, instead, the two coders were to use four categories in equal proportions, we would expect them to agree 25% of the time (since no matter what the first coder chooses, there is a 25% chance that the second coder will agree.) And if both coders were to use one of two categories, but use one of the categories 95% of the time, we would expect them to agree 90.5% of the time ($.95^2 + .05^2$ *, or, in words, 95% of the time the first coder chooses the first category, with a .95 chance of the second coder also choosing that category, and 5% of the time the first coder chooses the second category, with a .05 chance of the second coder also doing so). This makes it impossible to interpret raw agreement figures [...]"* [Carletta, 1996]

The same observed agreement of 70% has an entirely different meaning for each of Carletta's three examples. In the last case, the annotators have even achieved *less* agreement than one would have achieved by flipping coins with the same relative frequencies. This makes raw agreement percentages impossible to compare with each other. Chance-corrected reliability metrics make levels of agreement obtained on different annotation tasks comparable by normalizing them with respect to chance-expected agreement for the task. For example, Cohen [1960]'s κ is defined as $\kappa = \frac{P(A)-P(E)}{1-P(E)}$, where P(A) is the observed agreement among annotators and P(E) is the agreement expected by chance. When the achieved agreement is exactly the same as what would be expected by chance, $\kappa = 0$; when achieved agreement is perfect $\kappa = 1$ (perfect disagreement might lead to $\kappa = -1$, but that is extremely unlikely to occur). This holds for every annotation, no matter what the expected level of agreement is (50%, 25%, or 90.5%, as in the examples above, or any other value) or what the observed agreement is. A value of $\kappa = 0.5$ can in all those cases be interpreted as 'the level of agreement for this annotation is exactly midway between perfect agreement and the level that would be expected by chance.'

Throughout the years there have been many proposals for different reliability metrics as well as many publications dealing with the differences, similarities, advantages and drawbacks for all those metrics. Two very good starting points for information about these subjects are the works of Krippendorff [2004b] and Artstein and Poesio [to appear], both of which contain excellent reviews of the relevant discussions as well as extensive pointers to other literature. The general goal of the different metrics is always as described above: expressing a chance-corrected level of agreement. Another similarity is that they all operate on annotations defined as *labels assigned to units*. The units may be gesture episodes, text fragments, interview sessions, single television programs or fragments thereof. The labels may be anything from a perceived level of antagonism present in the unit to the number of references to technical devices. If an annotation is not defined as a labeling of units, these reliability metrics cannot deal with the annotation directly. Given an annotation that *does* satisfy this condition, one commonly encodes the points of agreement and disagreement in the form of a *confusion matrix* or *coincidence matrix* from the set of units labeled by two or more annotators. Both types of matrices encode information about how many times a unit labeled with class label C_i by one annotator was labeled with class label C_j by another annotator (agreed cases are counted in cells with $i = j$; disagreed cases in cells with $i <> j$). From these matrices, observed agreement and expected agreement are derived and used to calculate the chance-corrected agreement metric. Table 2.2 displays the confusion matrix for the example annotations from Table 2.1.

	Annotator 2			
Annotator 1	NoCostFactor	Cheap	NotExp	ExpLux
NoCostFactor	3	0	0	0
Cheap	1	2	1	0
NotExp	0	1	0	0
ExpLux	0	1	0	1

Table 2.2: Confusion matrix for the two annotations from Table 2.1.

Not every disagreement between different labels assigned to the same unit by two annotators has necessarily the same impact. For example, turning to the advertisements one last time, one can imagine that it is worse when one annotator assigns the label ExpensiveLuxury to an advertisement and another annotator chose the label Cheap for the same advertisement, than when the annotators assigned Cheap and NotExpensive respectively in the same annotation schema. In both cases the annotators disagreed, but the first pair of labels differ more than the second pair of labels. A distance metric is sometimes used to determine for how much 'agreement' or 'disagreement' any combination of two classes counts. The exact conceptual and formal definition of the calculations — Cohen's κ as defined above is only one possible example — separates the different existing reliability metrics from each other, and even now discussions about which metric is more appropriate abound [Di Eugenio and Glass, 2004; Krippendorff, 2004b; Craggs and McGee Wood, 2005; Stegmann and Lücking, 2005]. For the two most commonly used metrics, κ and Krippendorff [1980]'s α, the differences tend, in most real data sets, to lie in the third decimal place, though [Artstein and Poesio, to appear].

The availability of a chance-corrected agreement metric allows one to compare the levels of agreement obtained on two different annotation tasks, or on two variations of the same task or annotation schema. It also allows one to define a threshold value and define an obtained level of agreement higher than the threshold as 'good enough' and one lower than the threshold as 'not good enough.' Krippendorff [1980] very tentatively suggested a threshold that has subsequently been quoted extensively. Nowadays many researchers simply assume that this level of agreement of $\alpha > 0.8$ indicates that the annotated data is good enough to use. The drawbacks of that particular way of using reliability metrics will come up in the next chapter.

2.3 Fitting Data to Metric

Annotation tasks in which the annotator is asked to label predefined units with labels from a discrete set — such as labeling pre-segmented sentences with dialog act classes — lend themselves well to the calculation of a reliability metric. However, not all annotation tasks concern only labeling predefined units. Sometimes the annotator is first asked to *identify* the units to be labeled in the content, and possibly also to *assign start and end boundaries* to them. Subsequently, the units may be *labeled*. Finally, some tasks require an annotator to *link units to each other* in (labeled or unlabeled) relations. If the first or last steps mentioned here are part of an annotation task, such as for segmenting and labeling continuous video data, annotating discourse relations, and many other tasks, the different annotators do not necessarily label exactly the same units. This makes it more complicated to construct a coincidence matrix from which to calculate the reliability metric. In Figure 2.1 a fictitious example is visualized of a segmentation-and-labeling task where annotators are requested to mark periods in a recorded meeting where the participants were laughing, as well as to label each period with the type of laughter. Some laughing events were clearly similarly identified and can be seen as "the same unit annotated by both annotators," whereas other events were identified by only one annotator. For the second segment of annotator B it is even not at all clear how to relate it to the segments of annotator A. It is not clear what values should be put in the confusion matrix. The data needs to be transformed into 'labeled units' before α can be calculated. In this section examples of data transformations used to fit the data for calculating α or κ reliability are discussed for two types of data: one for analysing the segmentation of data along a time line and one for analysing annotations with a graph structure.

2.3.1 Unitizing

The reliability metrics discussed in Section 2.2 can be used to assess the level of agreement of labels assigned to units. Many annotation tasks start one step earlier: annotators are first required to *identify* the units, as fragments in a text or episodes in a video recording. In the construction of the AMI corpus, which plays a central role later in this thesis, annotators were asked to identify and label communicative gestures in the recordings in one annotation task, together with start and end times. The result looks somewhat similar to the example in Figure 2.1. The agreement ana-

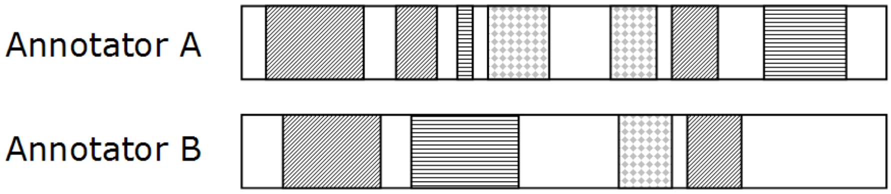


Figure 2.1: A fictitious example annotation for "types of laughter during a meeting" for two annotators. Each type of shading stands for a different label (for example, the labels SarcasticLaughter, SocialSmile, and AmusedLaughter), with the white, unlabeled areas standing for periods where no laughter occurred.

lysis for such tasks ideally includes an additional step, namely determining whether the annotators identified the same units or episodes to be labeled.

However, one often used agreement analysis for such segmentation-and-labeling tasks leaves out this step. Frame level agreement calculations are based on discretizing the time line in an annotation into small equal-sized windows (often single video frames). Figure 2.2 shows the resulting transformed annotation for the laughter example. Each window defines one separate labeled unit. From this 'labeling of units' one can calculate a measure of inter-annotator agreement as described above. Many researchers report agreement using this method. Quek et al. [2005] used frame level percent agreement for gaze and gesture annotation. Ciceri et al. [2006] used it for behavior annotations including gaze direction, FACS units, posture and vocal behavior. Falcon et al. [2005] presented frame level reliability for an annotation of (social) group behavior.

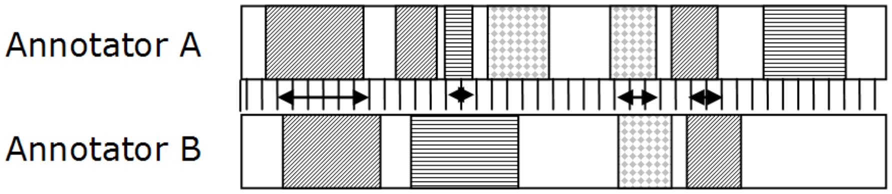


Figure 2.2: The annotation from Figure 2.1 discretized into windows. The arrows indicate the windows that count as units with the same label.

A clear drawback to the method is that it does not give any insight into how well annotators identify the same episodes or segments, or how well they assign the same timing to them. This problem is illustrated in Figure 2.3, which visualizes part of the real gesture annotations from the AMI corpus. It can be seen that it is impossible to distinguish with this method disagreement that occurs because people *do not detect the same episodes* (2), because people *label them differently* (3) and because people *assign different timing* (1,4) to the episodes. All of these different types of disagreement have exactly the same impact on the agreement analysis: they end up as frame level disagreement in the same confusion matrix, losing important information about the (dis)agreement between annotators. It is also unclear how appropriate this method is for annotations that are more easily interpreted as event-like segments than as frame-by-frame occurrences or states.

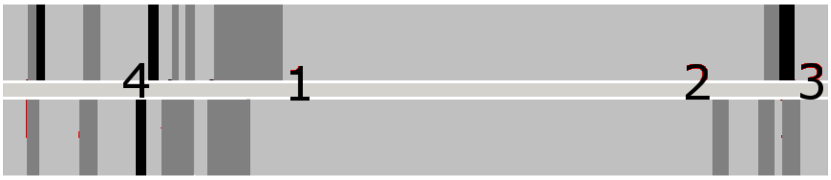


Figure 2.3: A fragment of gesture annotations from the AMI corpus for two annotators in which (1) different timing was assigned by the two annotators (2) only one annotator identified a segment (3) different timing and different label assigned by the two annotators (4) different timing was assigned by the two annotators (non-overlapping)

A better approach to this agreement analysis is to first identify which episodes have been found by both annotators, before looking at the agreement on the assigned labels and the exact timing of the boundaries for the agreed episodes. One can, for example, consider two segments identified and labeled by different annotators to be the same unit for the purpose of calculating α when their respective start and end times differ by at most some threshold value θ (see Figure 2.4). Allwood et al. [2006] ($\theta = 0.25s$), Jovanović et al. [2005] ($\theta = 0.8s$), and Martell and Kroll [2006] ($\theta = 0.25s$ or $0.5s$) present such an agreement analysis using different (fixed) values for θ. Kita et al. [1998] manually determine commonly identified episodes during a discussion between annotators, and unlike the previous three report separately the *variation* in the exact timing of the assigned boundaries. Reidsma et al. [2006] present an automatic method — developed for analysis of an emotion annotation but also applied for FOA reliability [Jovanović, 2007, page 80] — in which the threshold value θ varies depending on the length of the segments being compared. In the method of Reidsma et al. [2006], very long segments need to have enough overlap to be considered the same segment, whereas short segments do not need to have overlap but need to conform to a smaller threshold difference in timing. This makes the method more robust for annotations that exhibit a large variation in size of episodes.

Almost all of the above mentioned works proceed to assess labeling agreement on the commonly identified episodes, that is, the episodes identified by both annotators, using a chance-corrected reliability metric. However, none of them consider chance-correction in their treatment of the *segmentation agreement*. Krippendorff [1995] treats exactly that problem when he presents a chance-corrected reliability metric for the unitizing/segmentation of continuous data. His method derives chance-corrected agreement from the relative amount of overlap and non overlap for all (partly) overlapping segments. A drawback of his method is that it does not distinguish the identification of episodes from the timing assigned to them. Krippendorff also completely separated the inter-annotator agreement analysis of segmentation from the analysis of labeling, in contrast to the threshold based method pre-

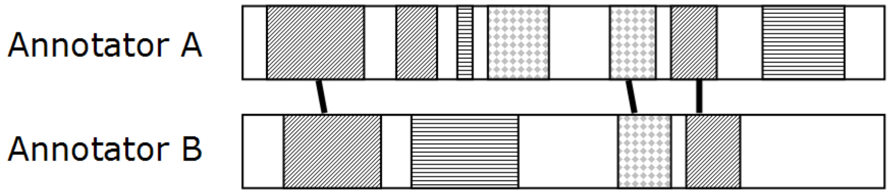


Figure 2.4: The annotation from Figure 2.1 aligned on a threshold distance: two segments identified by the different annotators are considered to be an identification of the same segment if the start and end times differ by at most θ seconds.

sented above. He does not present advice on how to move beyond agreement analysis for unitizing to determine which units from different annotators are 'the same unit', which is necessary for tackling the analysis of agreement on the *labeling* of episodes. This decoupling of segmentation and labeling, where each is investigated in complete isolation, is actually not very appropriate for most segmentation-and-labeling tasks, because usually annotation is a holistic task, in which segmentation and labeling are closely entwined and mutually influential. Finally, Krippendorff's method requires identified episodes to have at least some overlap before they are compared as to their potential agreement. This is not suitable for all unitizing tasks, as can be seen in Figure 2.3: case 4 would count as two points of disagreement for α_u.

In conclusion, it can be said that the threshold-based method for identifying the commonly found segments in two annotations yields the most information and that there is as yet no perfect method for determining *chance-corrected* inter-annotator agreement for segmentation tasks.

2.3.2 Annotations with Graph Structure

Graph structure annotation tasks form another group of tasks for which it is difficult to transform the data into a format defined as a labeling of units in order to calculate α. An annotation has a graph structure when the task involves creating (labeled or unlabeled) *links* between units. Under this heading one finds annotations such as anaphoric reference markup, rhetorical structures, discourse relations, and the like. For annotations that are structured as trees or graphs there are no obvious units. There have been many publications discussing the interpretation of discourse relations as labelings of units or the definition of good distance metric on the resulting labels in order to calculate α or κ [Carlson et al., 2001; Marcu et al., 1999; Passonneau, 2006; Jovanović, 2007]. The survey of reliability metrics by Artstein and Poesio [to appear] contains a clear discussion of the main problems with such data transformations. In the first place, it is not obvious in individual cases that the chosen transformation is the best representation of the (dis)agreement between annotators. In the second place, both the choice of transformation and the definition of the distance metric can have a great effect on the outcome of the reliability metric. According to Artstein and Poesio the effect is so great that for some types of annotations it defeats the purpose of assessing overall quality of the annotation using a threshold on the value of the reliability metric.

2.4 Sources of Disagreement

After the chance-corrected agreement metrics have been calculated, the results must be interpreted. In order to understand how the annotated data can be used it is important to find out *how* and *why* annotators disagree, instead of just *how much*. To be sure that data is fit for the intended purpose, Krippendorff [1980] advised the analyst to look for structure in the disagreement and consider how it might affect data use. Others have reiterated this advice [Carletta, 1996; Craggs and McGee Wood, 2005; Passonneau et al., 2006; Artstein and Poesio, to appear], although concrete guidelines for how to do this are few.

It is important to note that some kinds of disagreement are more systematic and other types are more noise-like. Systematic disagreement is particularly problematic for subsequent use of the data, more so than noise-like disagreement (this will be discussed further in Chapter 3). Many different sources of low agreement, and many different solutions, are discussed in the literature. The main sources of disagreement are given below, freely summarized from the same literature used in the introduction of this chapter [Krippendorff, 1980; Poole and Folger, 1981; Weber, 1983; Bakeman and Gottman, 1986; Potter and Levine-Donnerstein, 1999].

(1) *'Inadequate selection of relevant concepts for inclusion in the annotation scheme'.* One of the first steps in setting up corpus based research is to select relevant concepts to take into account. A lack of insight into the theoretical background of the subject matter may lead the researcher to select concepts for inclusion in the research that are not valid, bearing no relation to 'reality'.

(2) *'Invalid or imprecise annotation schemas'.* The theoretical ideas underlying the research may have been badly operationalized into an annotation schema. The schema may contain class labels that are not relevant or may lack certain relevant class labels, or may force the annotator to make choices that are not appropriate to the data (e.g. to choose one label for a unit where several labels are applicable). Solutions usually concern redesigning the annotation schema, for example by merging difference classes, allowing annotators to use multiple labels, removing classes, adding new classes, and so on.

(3) *'Insufficient training of the annotators'.* If the instruction manual has been badly written, or the annotators are not trained well enough, they may not be able to properly apply the annotation schema to the data. They may assign the wrong labels to units because they do not understand their task well enough. Solutions are to provide better instructions and training and to use only the annotators who perform well on the training task.

(4) *'Clerical errors'.* Such errors may be caused by a limited view of the interactions being annotated (low quality video, no audio, occlusions, etc) or by careless work of the annotator. Some solutions are, again, providing better instructions and training, having the annotators take enough rest breaks, and using high quality recordings of the interaction being annotated.

(5) *'Genuinely ambiguous expressions'.* Poesio and Artstein [2005] discussed how some annotation tasks can involve instances of genuinely ambiguous language use. Sometimes language expressions, such as some anaphoric relations, are ambiguous in themselves. They argued that disagreement caused by this cannot simply be counted as errors. One solution might be to introduce the label AMBIGUOUS as an extra class.

(6) *'A low level of intersubjectivity'.* Some annotation tasks require a lot of interpretation from the annotators. This interpretation may differ for annotators due to differences in personality, culture, age, gender, profession, and all the other elements that make up the individuality of a person. As already discussed in the introduction, people generally have different views of the world, and often interpret the meaning of verbal and nonverbal communicative behavior different than other people do. These individual differences determine the degree of subjectivity in an annotation task. They lead to a certain amount of disagreement in the annotations. For many people this is seen as a reason to exclude subjective annotation tasks in research. Other people see this as a reason for allowing such annotation tasks to have a low reliability. The problem is a central topic in this thesis. It will be explained more elaborately in Section 2.5, and Chapter 6 is dedicated to exploring the question as to how such data can be used sensibly, even if it exhibits a low reliability.

In general, it is well understood what kind of problems contribute to disagreement in annotated data. However, there are surprisingly few examples of actual corpora for which an in-depth analysis of the sources of disagreement has been published in the fields of computational linguistics and corpus based computer science. By far the most common approach to reporting reliability of an annotation is to only calculate the value of an agreement metric on the subset of multiply annotated data and compare it to some threshold. The most prominent works in which the (dis)agreement of a corpus is investigated in more depth are discussed here.

Carletta et al. [1997] said about their reliability analysis of a dialog annotation: "Reliability in essence measures the amount of noise in the data; whether or not that will interfere with results depends on where the noise is and the strength of the relationship being measured." Subsequently, they focussed on the use of confusion matrices as an important source of information for the type of mistakes that annotators make. They noted, for example, that annotators had difficulty distinguishing between different types of moves that all contribute new, unelicited information (INSTRUCT, EXPLAIN, and CLARIFY), or that annotators had problems distinguishing between QUERY-YN and CHECK. Also, they noted that some of the disagreement stemmed from differences in the *granularity* with which annotators marked up the dialogs rather than fundamental differences in how they interpreted the content.

Kita et al. [1998] presented a reliability analysis of movement phases in signs and co-speech gestures. Because their corpus contained relatively few gesture episodes (about 25 instances of gestures and 25 instances of signs identified by each annotator) they could analyse (dis)agreement manually. After performing the task, two annotators looked at the annotations together and discussed the gross segmentation. Two episodes were considered to match on gross level segmentation when the annotators saw roughly the same stretch of movement as a phase with the same directionality, regardless of exact boundaries and identification of phase-type (that is, precise timing and labeling were not yet considered). Given this alignment, the authors presented a careful analysis of the disagreement on identification of episodes (some annotators missed small movements, some annotators used a different granularity), the timing of boundaries assigned to episodes (the vast majority of agreed boundaries differed by at most 100 msec between annotators) and labeling of the episodes with phase types.

Wiebe et al. [1999] analyzed "*patterns of agreement* in a data set of subjectivity annotations to identify systematic disagreements that result from relative bias among judges." They used several statistical analyses to show that there was systematic relative bias between annotators for certain classes. They used this information to (a) revise the annotation manual and (b) produce bias-corrected tags. The bias-corrected are produced using the latent class model of Lazarsfeld [1966]. Assuming that an underlying 'correct' label exists for each unit, which is imperfectly observed by the different annotators, the systematic relative bias between annotators can be used to calculate the conditional probabilities for the value of the underlying correct label, given the labels assigned by the annotators.

Bayerl and Paul [2007] analyzed data provided by Shriberg and Lof [1991] in which four different facets of the annotation task (annotation schema, granularity, material and annotation team) had been varied. Using Generalizability Theory, they were able to show that disagreement in the corpus stemmed from problems with granularity and the annotation schema rather than from idiosyncrasies of the individual annotators.

Beigman Klebanov et al. [2008] reported an inter-annotator agreement analysis on a collection of newspaper texts annotated with occurrences of metaphors by nine annotators. The data was annotated with an inter-annotator agreement between $\kappa = 0.39$ and $\kappa = 0.66$ for four metaphor types. They set two goals for their subsequent analysis. Firstly, they wanted to find a *subset* of the annotations that was more reliable. Because all data was annotated nine times, they could do this using the procedure developed by Beigman Klebanov and Shamir [2006]. Statistical analysis showed that the "deliberately reliable subset" consisted of all occurrences of metaphors marked by at least four out of nine annotators. Secondly, they wanted to distinguish between two sources of disagreement in the annotations, namely slip of attention and genuine subjectivity. These sources relate to the 'low level of intersubjectivity' and the 'clerical errors' mentioned above. They distinguished between the two sources of disagreement using a validation procedure in which annotators were asked to assess whether they thought annotations of others were correct. They found that there was a clear separation of all metaphors into those where most annotators accepted the judgements of others *even when they themselves had not marked that particular metaphor* versus those with which other annotators often disagreed in the validation experiment. They also showed that the metaphors in the deliberately reliable subset, which each were potentially *not* identified by five out of the nine annotators, were almost always validated by the other annotators (in about 95% of the cases). They concluded that the disagreement in the deliberately reliable subset was mostly caused by a slip of attention from annotators, and therefore that all metaphors in that subset (even those marked by only four annotators) can be used as training and testing material for machine-learning classifiers.

2.5 Types of Content

In the introduction to this thesis it was mentioned that annotation tasks can be subjective to a certain degree, depending on the type of content that needs to be annotated and the amount of personal interpretation required from the annotators. It was also briefly remarked, in the previous section, that this subjectivity can have an impact on the amount of (dis)agreement exhibited by the annotations of two different annotators. There is much work with annotations that require subjective judgements from the annotators. A small illustrative selection of topics includes Human Computer Interaction work in areas such as affective computing [Paiva et al., 2007] and the development of Embodied Conversational Agents that behave in human-like ways [Pelachaud et al., 2007], and work in (Computational) Linguistics on topics such as emotion [Craggs and McGee Wood, 2005], subjectivity [Wiebe et al., 1999; Wilson, 2008] and agreement and disagreement [Galley et al., 2004].

The 'spectrum of subjectivity' in annotations relates to the spectrum of content types discussed extensively by Potter and Levine-Donnerstein [1999]. They distinguish the annotation of *manifest content* (directly observable events), *pattern latent content* (events that need to be inferred indirectly from the observations), and *projective latent content* (loosely said, events that require a subjective interpretation from the annotator). These types of content are presented in more detail below.

2.5.1 Manifest Content

Manifest content is "that which is on the surface and is easily observable" [Potter and Levine-Donnerstein, 1999]. Some examples are annotation of instances where somebody raises his hand or raises an eyebrow, annotation of the words being spoken and indicating whether there is a person in the view of the camera. Annotating manifest content can be a relatively easy task. Although the annotation task involves a judgement by the annotator, those judgements will not diverge much for different annotators. Very early work in Content Analysis followed the principle that this type of content was the only possible subject matter. Interpretation of the observed data by the annotators was something to be avoided: "[this] requirement literally excludes ‘reading between the lines,’ which is what experts do, often with remarkable intersubjective agreement [...]” (Krippendorff [2004a, page 20], about the work of Berelson [1952]).

2.5.2 Pattern Latent Content

Pattern latent content concerns classes that are not directly observable as such. Their presence must instead be inferred from the presence of a pattern of other things that *are* directly observable. A clear example can be seen in the medical domain: although the judgement as to whether a patient has a certain physical illness is not subjective, a doctor cannot directly observe the presence of the illness. Rather, it must be inferred from the presence of a number of symptoms in certain combinations. In this approach, interpretation by the annotator plays a role, but the observer’s own point of view is still explicitly excluded, as can be seen in the pioneer work by Bales [1950] on his method of Interaction Process Analysis: “The classification which [the observer] makes is clearly and unequivocally a matter of interpretation [...] Strenuous efforts are made to [...] cancel out the effects of value judgements from the observer’s own particular point of view.” Another example of a pattern latent content approach can be found in the segmentation instructions of the MRDA dialog act annotation guidelines [Dhillon et al., 2004]. The segmentation rules in these guidelines contain many detailed references to surface forms on syntactic and prosodic level that determine segment boundaries.

2.5.3 Projective Latent Content

At the other end of the spectrum Potter and Levine-Donnerstein [1999] placed projective latent content. They described this as a type of content for which the annotation schema does not specify in extremely precise detail the rules and surface forms that determine the applicability of classes, but in which the annotation relies on the annotators’ existing “mental conception”[1] of the classes. This type of content firstly may concern everyday concepts that most people understand and to a certain extent share a common meaning for, but for which it is almost impossible to provide adequately complete definitions. Potter and Levine-Donnerstein used the concept ‘chair’ — in a way reminiscent of Wittgenstein [1953, PI 80] — as an example of an everyday concept that is difficult to define exhaustively.

“The concept of chair is a common example of a primitive concept. Most people have a clear schema for chair, and this schema is shared by almost all people. If you brought a variety of different people into a house and asked them to point out the chairs, there would be very high consistency among all people in identifying those objects. However, it is a surprisingly difficult task for anyone to write a good definition of “chair”. Any effort by a content analyst to provide coding rules for identifying chairs would only serve to

[1] Potter and Levine-Donnerstein use the word “mental scheme” for this. In this thesis the term ‘mental conception’ will be used because that is less prone to confusion with the term ‘annotation schema’.

confuse coders and thus reduce their confidence in using their perfectly adequate schema in successfully fulfilling the coding challenge." [Potter and Levine-Donnerstein, 1999, page 260]

Secondly, the concept of projective latent content is very relevant in an application context that *requires the end consumer of the data to agree with the distinctions being made*. Potter and Levine-Donnerstein illustrated this mainly with examples from the television domain. In studies about the presence of sexual content in a television show where "the general public [is regarded] to be the primary consumer of their findings", the annotation task should focus on assessing the presence of sexual content in terms that this so-called "general public" would agree with. As another example, they discussed how annotation of violence on television should result in shows and movies being ranked as 'more violent' or 'less violent' similar to how the average TV audience would rank them. Defining violence in movies purely using surface forms such as 'the number of gunshots heard', 'the number of deaths portrayed' or 'the number of visible physical acts of violence' might be inadequate, as one can imagine a TV audience assessing some movies as being very violent even though not a single instance of physical violence is ever displayed on the screen.

The concept of "requiring annotations to be such that the end consumer of the data can agree with the distinctions being made" can also be seen in the context of corpus based research in which machine-learning classifiers are developed to be used in everyday applications. The user of the application takes in this context the role of the end consumer of the data. One can make, for example, a highly circumscribed, ethologically founded definition of the concept 'dominant' to guide annotation. This is useful for research into social processes in multiparty conversations. However, in a scenario where an automatic classifier, trained to recognize this class, is to be used in an application that gives a participant in a meeting a quiet warning when he is being too dominant [Rienks, 2007] one would rather prefer the same concept 'dominant' to fit the mental conception of dominance that an ethologically naive meeting participant may have.

2.5.4 Choosing Between the Types of Content

The choice to approach the concepts one wants to annotate as either pattern latent content or projective latent content is a fundamental one that should reflect the goals for which the corpus is constructed and the status that the researcher accords to the mental conceptions that an annotator may have about what is being annotated. The same annotation task may be presented to the annotator in different ways, depending on whether one defines the content as pattern or projective latent content. One can 'ask the annotators to count the gunshots', specifying the patterns associated with each class in detail, or one can 'rely on the feelings of the annotators' to get the annotations right. Both approaches have actually been used in the past, for a variety of annotation tasks, although usually other terminology was used to explain the decision. The annotation guidelines for the well-known Interaction Process Analysis (IPA) method of Bales [1950] showed the pattern latent content approach, providing detailed rules and descriptions for all classes. The dialog act segmentation rules for the MRDA scheme mentioned above [Dhillon et al., 2004] were also clearly defined in pattern content terms. Conversely, Greene and Cappella [1986] relied on the mental conceptions of the annotators for segmentation of a dialog into 'idea units': "The concept of an 'idea' was not defined, and subjects were left to their own devices in ascertaining such units." The AMI dialog act annotation schema was defined somewhere between pattern and projective latent content: the annotators were instructed to use 'single speaker intentions' as they perceived them as the main unit of segmentation, but subsequently a number of surface form rules were given to support the annotator. The annotation schema for dominance (or influence) in meetings by Hung and Gatica-Perez [2008] is defined in a projective latent content manner ("Annotators were not given any initial definition of dominance [...]"), whereas Otsuka et al. [2006] defined participant influence in meetings as pattern latent content to an extent that they did not manually annotate influence, but rather detected the patterns automatically, and implicitly assumed that those patterns adequately model interpersonal influence. For emotion, the work by Laskowski and Burger [2005] is a very good example of a pure pattern content approach. The authors defined an annotation schema for "emotionally relevant behavior" completely in terms of patterns of manifest content. This was a conscious decision, and they even replaced in the manual all meaningful class labels with random letters to prevent the mental conceptions of the annotators from playing a role in the labeling decisions. The emotion annotation schema by Batliner et al. [2004], on the other hand, is clearly an example of the projective latent content approach. They described how they used for their annotators "[...] less experienced labelers – who so to speak represent 'the man on the street' [...]," and the meaning of the classes was not defined exhaustively.

Both choices make sense in their own context. Projective latent content is defined more intuitively, which can be an advantage when it comes to communicating the results of a study to 'the general public', or training a machine-learning classifier that is to be used by 'naive users'. The intuitiveness of projective latent content can improve usefulness of the data, but it can also lead to a more subjective annotation. It may make the machine-learning task harder and possibly less well defined compared to pattern latent content. When there is no risk of the intended consumer of the findings or data interpreting the class labels very different from their intended meaning, using a pattern latent content approach makes for much more reliable data, possibly more structure in the data, and therefore also an easier learning task. Note, though, that defining an annotation task as one of pattern latent content is not a guarantee that the annotator will not interpret the class labels using his or her own mental conceptions. The annotation manual needs to be carefully designed to avoid this happening. This is also why Laskowski and Burger [2005] decided to replace the meaningful class labels with nonsensical letters.

When one designs an annotation schema for projective latent content, Potter and Levine-Donnerstein [1999] argued, the focus of the annotation guidelines is on instructions that trigger the appropriate existing mental conceptions of the annotators rather than on writing exhaustive descriptions of how classes can be distinguished from one another.

2.5.5 Inter-Annotator Agreement for the Types of Content

Inter-annotator agreement takes on different roles for the two ends of the spectrum. For manifest content and pattern latent content, the level of agreement tells you something about how accurate the measurement instrument (schema plus annotators) is. Bakeman and Gottman, in their text book *Observing Interaction: Introduction to Sequential Analysis* [1986, p 57], said about this type of reliability measurement that it is a matter of "calibrating your observers". For projective latent content, inter-annotator (dis)agreement is a source of other information, too. Although the level of agreement still gives information about errors made by annotators, the level of agreement may be influenced by the level of intersubjectivity, too. Whereas Krippendorff [1980] describes that annotators should be interchangeable, annotations of projective latent content can sometimes say a lot about the mental conceptions of the particular annotator as well as about the person whose interactions are being annotated. Personal interpretations of the data by the annotator should not necessarily be seen as 'errors', even if those interpretations lead to low inter-annotator agreement: they may simply be an unavoidable aspect of the interesting types of data one works with, representing the range of interpretations by different observers. This issue was recently the topic of a workshop organized by Artstein et al. [2008]. One expects disagreement between annotators to be more *systematic* for projective latent content, because it is caused by the differences in the mental conceptions or personal ways of interpreting multimodal interaction.

2.6 From Data Quality to Data Use

When inter-annotator disagreement is caused by the (partly) subjective nature of the annotation task, there are clear consequences for generalizability of conclusions drawn from the data and for generalizability of the performance of machine-learning modules trained on the data. A recognition module trained on data embodying the mental conception of one annotator will not necessarily perform well in the eyes of another observer, be it annotator or application user. Conclusions about conversational behavior drawn from subjectively annotated data may make sense only to the person who produced the annotations. But even when the annotation task is not particularly subjective and disagreement stems from annotators' errors, disagreement can cause problems for using the data. Noise in the data makes it harder to draw significant conclusions and to accurately train machine classifiers. Systematic errors in the data may lead to incorrect conclusions being drawn about correlations and may be picked up by machine-learning modules that then learn to emulate the bias. This section concerns related work in which data quality and data use are discussed in relation. Note that the work discussed in Section 2.4, which reports analyses of the sources of disagreement, are, in a sense, also about the relation between data quality and data use, among other things because they indicate which category distinctions one can rely upon and which one cannot.

2.6.1 Constraining Possible Use of the Data

Data with a low inter-annotator agreement may be difficult to use, but there are related fields of research where the use of data with low agreement has been discussed before. One of those fields is that of information retrieval evaluation, the topic of the Text Retrieval Evaluation Conferences (TREC). Relevance judgements in TREC assessments (and document relevance in general) are quite subjective and it is well known that agreement for relevance judgements is not very high (Voorhees and Harman report 70% three-way percent agreement on 15,000 documents for three assessors in a binary annotation task [1997] in which roughly two thirds of the data was non-relevant). Quite early in the history of TREC, Voorhees investigated the consequences of this level of agreement for the usefulness of results obtained on the TREC collection. It turns out that specifying a few constraints is enough to be able to use the TREC assessments to obtain meaningful evaluation results [Voorhees, 2000]. In their case, the constraints can be summarized as: 'one should only report and discuss *relative* performance differences on different (variations of) algorithms/systems run on *exactly the same set of assessments* using the *same set of topics*.' Finding, and making explicit, such constraints should be part of a good inter-annotator agreement analysis.

2.6.2 Evaluating and Explaining Machine-Learning Results

It may be obvious that a low level of agreement in the annotations will have an impact on the performance of machine learning. If the disagreement is caused by errors, the amount of information available in the annotations from which a classifier can learn the right model goes down. Some researchers take this to mean that the agreement between human annotators defines the maximum that the machine classifier can achieve. Some of the results presented in Chapter 3 show that a classifier may be able to disregard noise caused by random errors in annotations. This would mean that the classifier can achieve a performance that is not completely bounded by the level of agreement between annotators, although it is a different matter if the errors are systematic. Nevertheless, several researchers have proposed to judge the performance of different algorithms relative to this maximum. Vieira [2002] remarked that it makes no sense to penalize machine-learning performance for errors made in situations where humans would not agree either. Vieira only looked at the *amount* of disagreement and did not explicitly relate the classes where the system and annotators disagreed to the classes where the annotators disagreed with each other. Rienks [2007, page 128] argued that the inter-annotator agreement defines an upper bound on the machine-learning performance for classifying relational labels in argument structures found in discourse.

Other researchers have additionally looked at the *kind of errors* that the machine classifier made, in comparison to the kind of disagreement between annotators. Op den Akker and Schulz [2008] presented performance results relative to the maximum defined by inter-annotator agreement, but also made a detailed manual comparison of the distinctions where the machine classifier performed badly with the disagreement typically expected from two human annotators. Steidl et al. [2005] presented an automatic method for comparing machine classifier performance with inter-annotator agreement both with respect to the amount and the type of errors. They used a corpus that was completely annotated by all annotators to determine so-called “soft labels,” vectors that express the distribution of labels assigned to one item by different annotators. The machine classifier was trained to output single class labels. Their performance metric is based on the distance between the soft labels in the annotation and the single label output by the machine classifier. This method has two important drawbacks. It only works if the whole corpus is multiply annotated, which can be a very expensive endeavor. Furthermore, it only makes sense for annotations where a soft label can sensibly be used. Emotions as annotated by Steidl et al. [2005], can be mixed, but if the different possible judgements are mutually exclusive, soft labels have no interpretation.

Finally, Passonneau et al. [2008] present an extensive analysis of the relation between per-class machine-learning performance and inter-annotator agreement obtained on the task of labeling text fragments with their function in the larger text. They confirm some of the results presented in Chapter 3 of this thesis by showing that overall high agreement can indicate a high learnability of a class in a multiply annotated corpus, but that a low inter-annotator agreement for a certain annotator is not necessarily predictive of the learnability of the classification from that single annotator’s data, especially in the context of projective latent content. This may be because the disagreement no longer results from random errors. The machine classifier may then learn to emulate the idiosyncrasies of the annotator. That particular problem is discussed further in the next chapter.

Chapter 3

Some Limits of Reliability Measurement

In the previous chapter, the most common approach to assessing the quality of a corpus was presented. One core element of that approach involves calculating a chance-corrected agreement metric. The result can be used to compare the relative quality of data obtained from different annotation tasks and to determine whether inter-annotator agreement is high enough, that is, whether the outcome exceeds a certain threshold. In the same chapter it was also mentioned that one can perform several additional analyses to find out more about *how* and *why* annotators disagree instead of just *how much*, as well as to find out what the impact of the disagreement is on the use one can make of the data. However, the common practice — contrary to the suggestions of Krippendorff and others — seems to have become to regard 0.8 as some kind of magical reliability cut-off guaranteeing the quality of hand-annotated data, necessitating no further analysis of the data. For example, consider the following examples of reliability analysis of actual data: the dialog act annotation for the VerbMobil project by Reithinger and Kipp [1998]; dialog act annotations on the ICSI corpus using the SWBD-DAMSL schema [Shriberg et al., 1998] and the MRDA schema [Shriberg et al., 2004]; the annotation of dialogs with emotion-related categories by Craggs and McGee Wood [2004]; the annotation of social group behavior in meeting data [Falcon et al., 2005]; annotation based on Rhetorical Structure Theory by Marcu et al. [1999] and Carlson et al. [2001] for data from the Penn Treebank; and annotation of agreement and disagreement on the ICSI corpus [Galley et al., 2004]. For each of those well known example data sets, and for many others, reliability analysis simply consisted of calculating κ or α, in most cases comparing it to a cut-off value, without saying anything about what kind of disagreement was found between the annotators. After having determined that the annotation is ‘good enough’, each of the data sets was then used for machine-learning purposes — in the same paper, or in other papers — without referring back to the agreement analysis during the discussion of performance results.

In this chapter, published earlier in collaboration with Jean Carletta [Reidsma and Carletta, 2008], the widespread focus on thresholding an agreement metric to the exclusion of an analysis of the source and structure of the disagreement is addressed. It is shown that this practice is less than adequate. One of the main ways of using such data, machine learning, might tolerate data with low reliability as long as any disagreement among human annotators looks like random noise. When the disagreement is more systematic and introduces patterns, however, the machine learner can pick these up just as it picks up the real patterns in the data, making the performance figures look better than they really are. For the range of reliability measures that the field currently accepts, disagreement can appreciably inflate performance figures, and even a measure of 0.8 does not guarantee that what looks like good performance really is. Although the problems discussed in this chapter will certainly be exacerbated when one works with 'subjective' annotations in which disagreement is likely to be more systematic, the experiments and conclusions in this chapter are relevant for any type of annotation task.

3.1 The Problem

In corpus based research, 0.8 is often regarded as some kind of magical reliability cut-off guaranteeing the quality of hand-annotated data with 0.67 to 0.8 tolerable — although it is as often honored in the breech as in the observance. The argument for the meaning of 0.8 arises originally from Krippendorff [1980, page 147], in a comment about practice in the field of content analysis. He stated that correlations found between two variables using their hand-annotated values "tend to be insignificant" when the annotations have a reliability below 0.8. He uses a specific reliability statistic, α, for his measurements, but Carletta [1996] implicitly assumes that kappa-like metrics are similar enough in practice for the rule of thumb to apply to them as well. A detailed discussion on the differences and similarities of these, and other, measures was provided by Krippendorff [2004b]. In this chapter Cohen's κ [1960] will be used to investigate the value of the 0.8 reliability cut-off for corpus based research.

Nowadays, researchers working with corpora often use data in a completely different way from 1970's content analysts. Rather than correlating two variables, hand-annotated data is used as training and test material for automatic classifiers. The 0.8 rule of thumb is irrelevant for this purpose, since classifiers will be affected by disagreement differently than correlations. Furthermore, Krippendorff's argument comes with a caveat: the disagreement must be due to random noise. For his case of correlations, any patterns in the disagreement could accidentally bolster the relationship perceived in the data, leading to false results. To be sure that data is fit for the intended purpose, Krippendorff advised the analyst to look for structure in the disagreement and consider how it might affect data use. Although computational linguists and computer scientists using annotated corpora have rarely followed this advice, it is just as relevant. Machine-learning algorithms are designed specifically to look for, and predict, patterns in noisy data. In theory, this makes random disagreement unimportant. More data will yield more signal and the learner will ignore the noise. However, as Craggs and McGee Wood [2005] suggested, this also makes systematic disagreement dangerous, since it provides an unwanted pattern for the learner to detect. Using a simulation of the annotation process, it is demonstrated in this chapter that machine learning can tolerate data with a low reliability measurement as long as the disagreement looks like random noise, and that when it does not, data can have a reliability measure commonly held to be acceptable but produce misleading results.

3.2 Method

To explain what is wrong with using $\kappa = 0.8$ as a cut-off value, one needs to think about how data is used for classification tasks. Consider Figure 3.1, which shows a relation between some features A and a class label B. Learning labels from a set of features is a common task in corpus based research; for instance, in Shriberg et al. [1998], which assumes a pre-existing dialog act segmentation, the labels are dialog act types, and they are learned from automatically-derived prosodic features. In this relatively simple way of using annotated data, only one of the variables — the output dialog act label — is hand-annotated. In the figure, the real relationship between prosody and dialog act label is shown on the left; R relates the prosodic features A to the output act B.

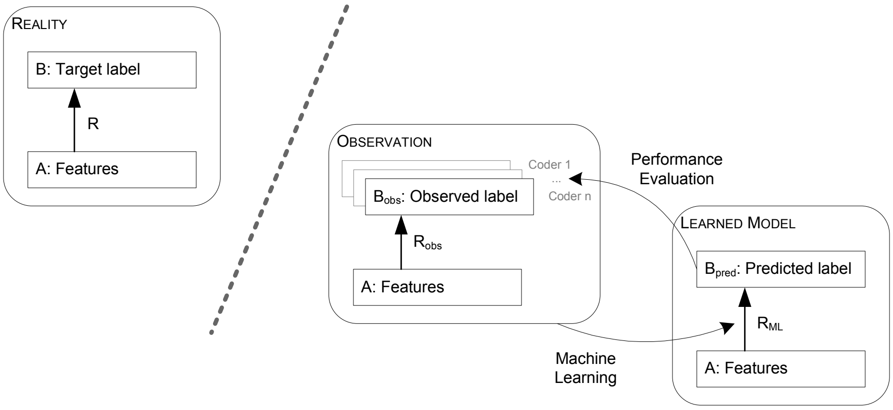


Figure 3.1: Hand-annotated target labels are used to train classifiers to automatically predict those labels from features.

In theory, it is often assumed that there is one correct label for any given act. However, in practice, human annotators disagree (see Section 2.4), choosing different labels for the same act. Sometimes the divergences are large enough to that make one question whether there is only one unique correct label, or any correct label at all. The data actually available for analysis is shown in the center of Figure 3.1. Here, the automatic features, A, are the same as before, but there are multiple, possibly differing labels for the same act, B_{obs}, coming from different human annotators. Finally, on the right the figure shows the classifier, which takes the same prosodic features A and uses them to predict a dialog act label B_{pred} on new data, using the relationship learned from the observed data, R_{ML}. Researchers vary in how they choose data from which to build the classifier when annotators disagree, but whatever they do is restricted by the observations they have available to them. Although one can describe reliability assessment as telling how much disagreement there is among the annotators, a better question is how their individual interpretations of the annotation schema make R_{ML} differ from R.

There is a problem that arises for anyone using this methodology. The 'real' data itself, which is only imperfectly observed by the annotators, is not accessible to the researcher. Even assuming that a correct label exists for any given act, this means that one cannot directly evaluate the quality of the annotations with respect to those correct labels, but rather must rely on an agreement analysis in which annotations produced by different annotators are compared to each other. Furthermore, without the 'real' data, it is impossible to judge how well the learned relationship reflects the real one. Classification performance for B_{pred} can only be calculated with respect to the 'observed' data B_{obs}. In this chapter, this problem is surmounted by *simulating* the real world so that the differences between this 'observed' performance and the 'real' performance can be measured. The simulation uses a Bayesian network [Pearl, 1988] to create an initial, 'real' data set with 3,000 samples of features (A) and their corresponding target labels (B). For simplicity, a single five-valued feature and five possible labels are used. The relative label frequencies vary between 17% and 25%. This gives a small amount of variation around what is essentially equally distributed data. The labels (B) are corrupted to simulate the 'hand-annotated' observed data (B_{obs}) created by a simulated annotator X, corresponding to the output of a hypothetical human annotator. A neural network is constructed and trained using the WEKA toolkit [Witten and Frank, 2005] on 2,000 samples from B_{obs}. Finally, the neural network's performance is calculated twice, using as test data either the remaining 1,000 samples from B_{obs} (the only test data that is normally available) or the initial, 'real' versions of those same 1,000 samples (that is, the samples with the normally inaccessible 'real' correct labels).

There are three ways in which the simulation needs to be varied in order to be systematic. The first is in the strength of the relationship between the features the machine learner takes as input and the target labels, which is achieved simply by changing the probabilities in the Bayesian network that creates the data set. In the simulation, the strength of the relationship is varied in eight graded steps.[1] The

[1] Cramer's phi is used to measure the strength of a relationship. Cramer's phi is defined as
$\phi_c = \sqrt{\frac{\chi^2}{(N)*df_{smaller}}}$
with N the number of samples and $df_{smaller}$ the smallest degree of freedom of the two involved variables, and is a measure of association for nominal variables with more than two values. It can be "considered like a correlation coefficient" [Aron and Aron, 2003] that takes data set size into account and can easily be derived for a Bayesian network from the priors and the conditional probability tables. The strength of the network is varied between $\phi_c = 0.06$ and $\phi_c = 0.45$. Following Cohen [1988], for a five-way distinction Aron and Aron [2003, page 527] would consider 0.06 to represent a small real relationship — that is, one with not much effect — and 0.3, a large one. Thus here 0.06

second is in the *type* of disagreement with which the real data is degraded to create the observed data (B_{obs}), representing the types of annotation errors the human annotators make. Again for simplicity, this chapter only describes the effects of both random errors and the over-use of a single label. The third is in the *amount* of disagreement introduced when creating the observed data (B_{obs}). This variation is expressed in terms of the obtained level of κ agreement in the simulation, in the following way. The simulated annotator X makes errors with varying frequencies for varying runs of the simulation. To calculate κ for a specific simulated annotation, an additional 1,000 samples of 'real' data are generated as 'reliability data' for each strength of relationship and type of disagreement, that have not been used for training or testing. These samples are annotated twice, once by the same simulated annotator X and once by a second simulated annotator Y, who makes the same type of mistakes with the same frequency. The resulting two sets of annotations of the reliability data ($B_{reliabX}$ and $B_{reliabY}$) can be used to calculate κ for the variation of the simulation for which they were generated. This way of calculating the obtained level of agreement mimics the common practice of having one annotator produce the annotations for the actual research, with a second annotator annotating just enough to test the reliability. By introducing a varying number of observation errors the simulated annotations of X, 200 different versions of hand-annotated data are created tho cover a range of values from $\kappa = 0$ to $\kappa = 1$, calculated as described above. These 200 versions, generated for each strength of relationship and each type of disagreement, are used in the next section to graphically show the dependencies between κ and machine-learning performance for the 'real' data B and for the simulated annotation B_{obs}.

3.3 Results

3.3.1 The Case of Noise

Figure 3.2 shows how a neural network, trained on the 2,000 samples of training data from B_{obs}, performs when the simulated annotator X makes random mistakes, that is, for noise-like disagreement, for the cases of (a) weak, (b) moderate, (c) strong, and (d) very strong relationships between the features (A) and labels (B). Here, the y axis shows 'accuracy', or the percentage of samples in the test data for which the network chooses the correct label. The x axis varies the amount of annotator's errors in the data to correspond to different κ values as calculated on the simulated reliability data, with the two black lines marking the values of $\kappa = 0.67$ and $\kappa = 0.8$.

The series depicted as a line, firstly, shows accuracy measured by using the 'observed' version of the test data, which is how testing is normally done. For each

is described as "weak", 0.45 as "very strong", and intermediate points as "moderate" and "strong". It is an open question what strengths of relationships actually occur in the type of data occurring in computational linguistics and computer science, although there may be no point in learning a relationship that is too strong.

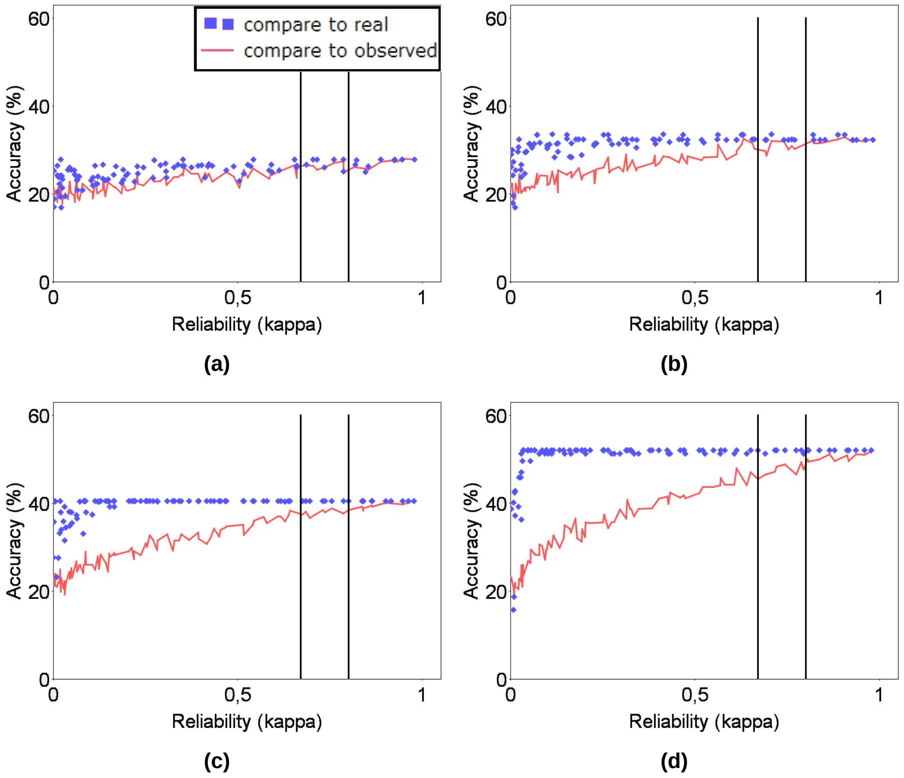


Figure 3.2: Machine-learning performance obtained on simulated annotations with noise-like disagreement for (a) weak ($\phi_c = 0.06$), (b) moderate ($\phi_c = 0.20$), (c) strong ($\phi_c = 0.32$), and (d) very strong ($\phi_c = 0.45$) relationships between the features and labels.

relationship strength, as κ increases, so does accuracy. In all cases, at $\kappa = 0$ — that is, when the annotators fail to agree beyond what one would expect if they were all choosing their labels randomly — accuracy is at 20%, which is what one would expect if the classifier were choosing randomly as well. For any given κ value, the stronger the underlying relationship, the more benefit the neural network can derive from the data. The other of the two series, secondly, depicted as small squares, shows accuracy measured by using the ‘real’ version of the data. Interestingly, the ‘real’ performance, that is, the power of the learned model to predict reality, is higher than performance as measured against the observed data. This is because for some samples, the classifier’s predictions are correct, but because the observations contain errors, the test data actually gets them wrong. The stronger the relationship in the real data, the more marked this effect becomes. The neural network is able to disregard noise-like annotation errors at very low κ values simply because the errors contain no patterns for it to learn.

3.3.2 The Case of Over-Using a Label

Now consider the case where instead of random annotation errors, the simulated annotator who produces the training data over-uses the least frequent one of the five labels for B. Figure 3.3 shows the results for this kind of annotation error. Remember that in the graphs, the series depicted as a line shows the *observed* performance of the classifier — that is, performance as it is usually measured. The two black lines again mark the κ values of interest ($\kappa = 0.67$ and $\kappa = 0.8$).

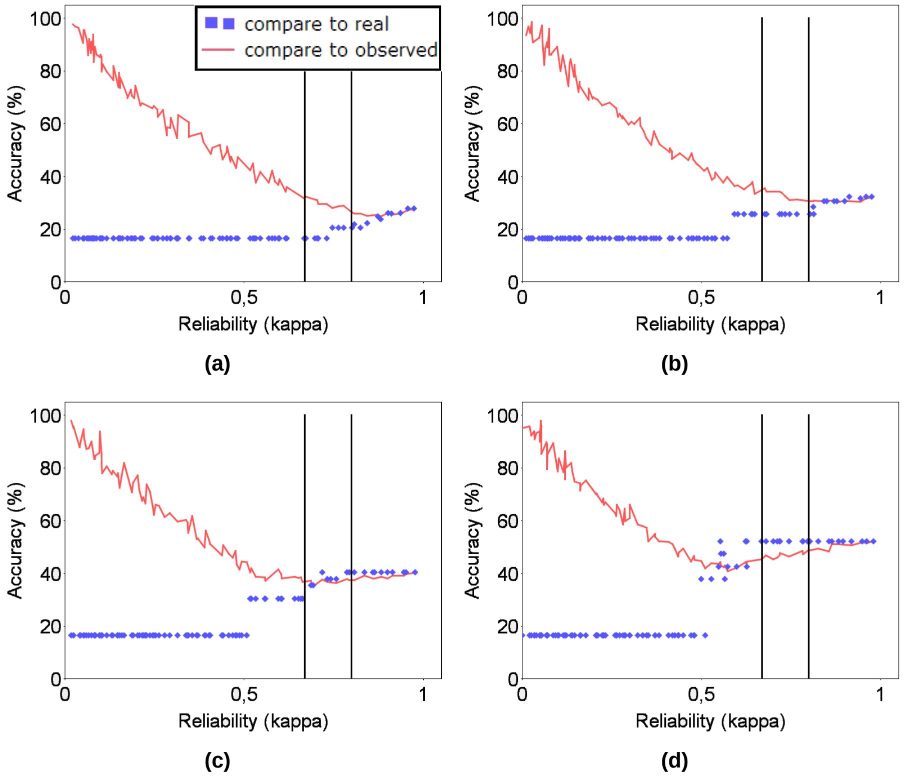


Figure 3.3: Machine-learning performance obtained on simulated annotations that suffered from over-annotation for (a) weak ($\phi_c = 0.06$), (b) moderate ($\phi_c = 0.20$), (c) strong ($\phi_c = 0.32$), and (d) very strong ($\phi_c = 0.45$) relationships between the features and labels.

The graphs show an entirely different effect from the one obtained for noise-like annotation errors: for lower values of κ, the observed performance is spuriously high. This makes perfect sense — κ is low when the pattern of label over-use by the simulated annotator is strong, and the neural network picks it up. When the observed data is used to test performance, some of the samples match not because the classifier gets the label right, but because it over-uses the same label as the human annotator. For data with a very strong correlation between the input features

A and the output labels B, the turning point below which performance is spuriously high occurs at around $\kappa = 0.55$ (Figure 3.3(d)), a value the community holds to be pretty low but which is not unknown in published work. However, when the underlying relationship to be learned is moderate or strong (Figures 3.3(b) and 3.3(c)), the spuriously high results already occur for κ values commonly held to be tolerable. With a weak relationship, the turning point can occur at $\kappa > 0.8$ (Figure 3.3(a)).

3.4 Discussion

The simulation results presented in this chapter highlight a danger for current practice in corpus based research. Over-use of a label is a realistic type of error for human annotators to make. For instance, imagine a annotation schema for dialog acts that distinguishes backchannel utterances from utterances which indicate agreement. In data containing many utterances where the speech consists of "Yeah", individual annotators can easily have a marked bias for either one of these two categories. Clearly, in actual annotation, not all disagreement will be of one type, but will contain a mix of different systematic and noise-like errors. In addition, the underlying relationships that systems attempt to learn vary in strength. This makes discerning the degree of danger more difficult, but does not change the substance of the argument set out in this chapter.

Although the graphs that were shown are for a specific simulation, the general pattern described is robust. The simulations have been done with a number of variations. Those variations are not presented in this chapter because they do not add new insights, but using other machine-learning algorithms, A and B with other cardinalities and with different prior and conditional probabilities, and other training set sizes, all exhibit similar effects. In particular, using α in place of κ does not markedly change the results; neither does increasing or decreasing the data set size. The simulations and results are presented for a machine-learning context. However, that does not mean that other ways of using annotated data are immune to the fundamental problems described here. Any statistical use of data will be affected in its own way by the difference between systematic and noise-like disagreement.

Very recent work by Passonneau et al. [2008], done on an actual corpus annotated by human annotators, ties in very well with the simulation results discussed in this chapter. Their paper concerns a task of assigning semantic relations between art images and sentences in the accompanying text. The authors looked at the level of inter-annotator agreement and its relation to overall learnability of the classification on all data on the one hand, and learnability of the classification on the data of one single annotator on the other hand. Several of their main conclusions are very relevant here. In the first place, they showed that data with a higher inter-annotator agreement overall is easier to learn. This is a reassuring result, confirming that good inter-annotator agreement is still a desirable property of data. In the second place, they showed that in some cases data from an individual annotator who had a high inter-annotator agreement with the other annotators was actually harder to learn from — that is, testing on test data from other annotators yielded a lower performance — than data from another annotator who had lower agreement with the others. They suggest that this relates to the effect discussed in Section 3.3.2: though the annotator had a high average inter-annotator agreement, his errors contributing to the disagreement may have been more systematic. This causes the errors to be picked up as patterns during training, confusing the machine-learning algorithm. They conclude with the observation that when there is variation, and it reflects the inherent variability of language use, machine-learning approaches should be modified to learn from multiple labelings, not from a single labeling, and thus learn where the variation lies. This topic is taken up again in Chapter 6.

3.5 Implications

At the moment, much of the effort the community devotes to reliability measurement is used to establish one or more overall reliability statistics for data sets and to argue about which reliability statistic is most appropriate. Methodological discussions focus on questions such as how to force annotated data structures into the mathematical form necessary to calculate α (see also Section 2.3), or what effects certain aspects of the annotation have on the *values of some metric* rather than on *possible use of the resulting data* [Marcu et al., 1999; Di Eugenio and Glass, 2004; Artstein and Poesio, 2005; Bhowmick et al., 2008]. Investigations of the reliability of actual corpus annotations most often do not move beyond calculating a metric and comparing it to a threshold. Researchers are of course aware that no overall reliability measure can give a complete story, but often fail to spend time analysing annotator disagreement further. Unfortunately, the results presented in this chapter suggest that current practice is insufficient, at least where the data is destined to be input for a machine-learning process and quite possibly for other data uses as well. This complements observations of Artstein and Poesio: apart from the fact that many different ways of calculating reliability metrics lead to different values, which makes comparing them to a threshold difficult [Artstein and Poesio, to appear], the very idea of having any such single threshold in the first place turns out to be impossible to hold. Instead of worrying about exactly how much disagreement there is in a data set and how to measure it, researchers should be looking at the form the disagreement takes. A headline measurement, no matter how it is expressed, will not show the difference between noise-like and systematic disagreement, but this difference can be critical for establishing whether or not a data set is fit for the purpose for which it is intended.

To tease out what sort of disagreement a data set contains, Krippendorff suggests calculating odd-man-out and per-class reliability to find out which class distinctions are problematic [1980, page 150]. Bayerl and Paul [2007] discuss methods for determining which factors (such as schema changes, annotation team changes, etc.) were involved in causing poor annotation quality. Wiebe et al. [1999] suggest looking at the marginals and how they differ between annotators to find indications of whether disagreement is caused by systematic bias (as opposed to being random) and in which classes they occur. Although clearly useful techniques, none of these diagnostics is specifically designed to address the needs of machine learners which are designed to recognize patterns. Over-using a label is just one simple example of a type of systematic disagreement that adds unwanted patterns that a machine learner can find. Any spurious pattern could be a problem. For this reason, one should be looking specifically for patterns in the disagreement itself. This will be a central topic in the next part of this thesis.

It should go without saying that the analyst will benefit from keeping how they intend to use the data firmly in mind at all times. As Krippendorff [2004b, page 429] recommends, one should test reliability for the "distinctions that matter" and perform "suitable experiments of the effects of unreliable data on the conclusions." Patterns found for an overall annotation schema will not always affect every possible data use. For instance, often classifiers are build not for complete annotation schemas, but for some subset of the labels or some 'class map' that transforms the schema into a smaller set of classes. In these cases, what is important is disagreement for the subset or transformation, not the entire schema. Similarly, where classifier performance is reported per class, the reliability for that particular label will be the most important. Finally, different machine-learning algorithms may react differently to different kinds of patterns in the data and to combinations of patterns in different relative strengths. In complicated cases, perhaps the safest way to assess whether or not there is a problem with systematic disagreement is to run a simulation like the one reported here but with the kind and scale of disagreement suspected of the data, and to use that to estimate the possible effects of unreliable data on the performance of machine-learning algorithms.

Part II

Praxis

Chapter 4

Moving to the AMI Corpus

The previous part of this thesis highlighted the three-way relation between (1) subjectivity, or the human variability in interpretation of communicative behavior, (2) inter-annotator agreement and (3) the design and evaluation of machine-learning modules. The final conclusion was that for a good inter-annotator agreement analysis one should, besides calculating the level of inter-annotator agreement, at least pay some attention to the questions *how* and *why* annotators disagree and to the impact of the disagreement on machine learning and other use of the data. These questions are the topic of the second part of the thesis. They have been addressed using several annotations from the AMI Corpus, namely the dialog act annotations, the addressee annotations and the visual Focus of Attention annotations (FOA). In the first place, the annotations have been used as material for novel analyses regarding inter-annotator agreement, and in the second place, machine-learning experiments were conducted specifically to explore the relation between inter-annotator agreement, machine learning and subjectivity.

In this chapter the relevant parts of the corpus are introduced. After a short overview of the setup of the AMI corpus, the three annotation layers are presented. For each annotation, a summary of the annotation scheme is presented, followed by some information about the amount of data that was annotated, the number of annotators involved and the distributions of classes over the data. Finally, information about inter-annotator agreement is given. This concerns only the amount of reliability data that was collected and the results of a basic inter-annotator agreement metric calculation, as more extensive analyses are discussed in the next two chapters. The divisions into training and test sets for machine-learning purposes will be discussed in the relevant sections in later chapters, as they depend heavily on the goals of the various experiments.

4.1 The AMI Corpus: an Overview

4.1.1 Visions of AMI

The AMI Consortium is a multi-disciplinary collaboration of 15 academic and industrial partners. The consortium is dedicated to research and development of technology that will support effectiveness and efficiency of meetings. The following quote can be found on the consortium web site: "Business productivity, by way of individual and group activities between and during meetings, can be dramatically enhanced with the use of advanced signal processing and knowledge management."[1] The technologies fall in two categories. In the first place the consortium dedicates a lot of effort to developing technology that can help make the content of past meetings accessible for later browsing and retrieval [Whittaker et al., 2008]. The possibility to easily review past decisions and the argumentation that led to the decisions or to look up the information from past presentations will enhance group memory, improve long-term consistency in the decisions and overall lead to a better information awareness in the organization [Moran et al., 1997]. In the second place there is an increasing focus within the consortium on systems that assist the users in many ways *during* the meeting. Some of this effort concerns supporting remote participants in distributed meetings to increased engagement and participation [Wrigley et al., 2008; Vyas and Bajart, 2007]. Other work targets proactive support by intelligent systems that, for example, provide information to the meeting or suggest to the chairman that certain actions should be taken [Rienks et al., 2006, 2007; Rienks, 2007]. The AMI Meeting Corpus, described in the next section, plays a central role in the AMI and AMIDA projects.

4.1.2 Recorded Scenario Meetings

The experiments and investigations described in the next two chapters of this thesis are based on the hand annotated face-to-face conversations from the 100 hour AMI meeting corpus. This corpus has been described before in other publications [Carletta, 2007; Carletta et al., 2006]. In this section a brief overview is given, after which the remainder of this chapter will describe three annotations from the corpus in more detail.

The corpus consists of 100 hours of recorded meetings. Of these recordings, 65 hours are of meetings that follow a guided scenario [Post et al., 2004]. In the scenario-based meetings, design project groups of four players have the task to design a new TV remote control. Group members have roles: project manager (PM), industrial designer (ID), user interface design (UD), and marketing expert (ME). Every group has four meetings (20-40 minutes each), dedicated to a sub-task. Most of the time the participants sit around a table. During the meetings, as well as between the meetings, participants will get new information about things such as market trends or changed design requirements, via mail. This process is coordinated by a *scenario controller* program. The whole scenario setup was designed to provide an optimal balance between *control* over the meeting variables and the *freedom* to have natural meetings with realistic behavior from the participants [Post et al., 2004]. In order to make sure that the scenario does not constrain the contents of the corpus too much, 35 hours of corpus data *"is made up of real meetings which progressively push out first from the remote control design scenario into other types of*

[1] http://www.amiproject.org/

new design teams, non-design teams, and finally a few other types of meetings entirely. Most of these meetings are 'real' (that is, they would have occurred whether or not we had been recording) but a few are controlled, more loosely and in different domains than the bulk of the data" [Carletta, 2007].

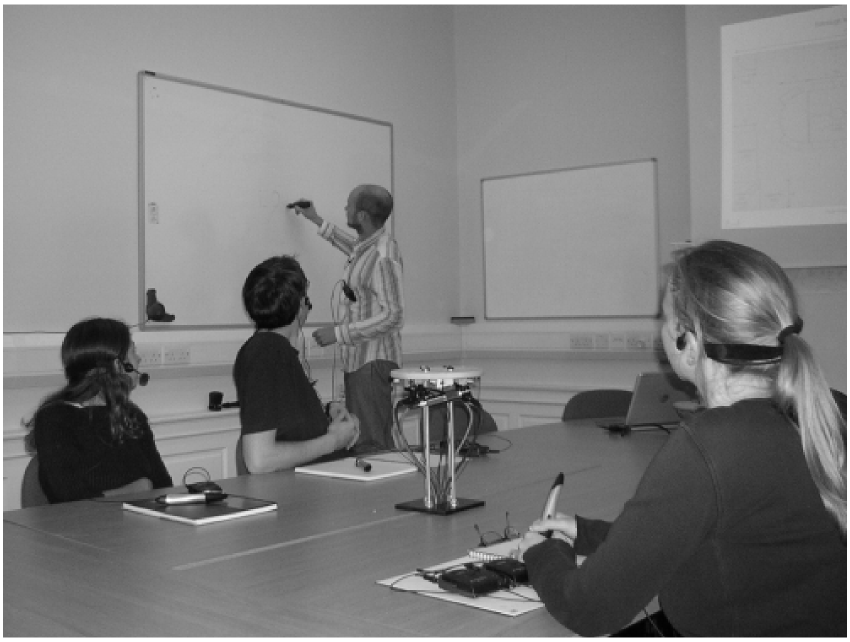


Figure 4.1: A still image of the meeting recording room in Edinburgh.

All meetings were recorded in meeting rooms full of audio and video recording devices (see Figure 4.1) so that close facial views and overview video, as well as high quality audio, is available. Speech was transcribed manually, and words were time-aligned. The corpus has several layers of annotation for a large number of modalities, among which dialog acts, topics, hand gestures, head gestures, subjectivity, visual focus of attention (FOA), decision points, and summaries. The corpus uses the Nite XML Toolkit (NXT) data format as reference storage format, making it very easy to extend the corpus with new annotations either by importing data created in other formats or by using one of the many flexible annotation tools that it comes with [Carletta et al., 2005, 2003; Reidsma et al., 2005a,b]. In the rest of this chapter, the dialog act, addressee and Focus of Attention annotations are presented in more detail.

4.2 The AMI Dialog Act Annotations

The AMI dialog act annotation schema concerns the segmenting and labeling of the transcripts into dialog acts. Part of the AMI corpus was also labeled with relations between dialog acts [Jovanović, 2007, page 76], but these annotations are not considered here. The *segmentation* guidelines are centered around the speaker's intention, with a few rules that describe how the annotators should deal with the different situations they are likely to encounter. The rules are summarized below; more details can be found in the annotation manual [AMI Consortium, 2005b].

- The *first* rule is: *each segment should contain a single speaker intention*.
  - If a speaker, for instance, asks two different questions in a row, without anyone else speaking, each of them is a separate segment.
  - If someone says "No, it's not", the "it's not" is not a separate segment, since it rephrases the same information as the "No".
  - Lengthy pauses or conjunctions that introduce whole new clauses such as "so", "because", and some uses of "and", "but", or "or" can be hints that a new segment is starting.
  - In the case of a (sub-ordinate) conjunction where the first half requires the second half to be complete — neither part expresses a complete intention — the parts should be combined into one segment.
  - If the speaker changes from talking to one person to talking to someone else or the whole group, or the other way around, there would be two intentions, and therefore two segments, although the speaker's intention is the deciding factor.
- The *second* rule is that *all segments only contain transcription from a single speaker*. This rule allows dialog act segmentation to be carried out on the speech of one speaker.
- The *third* rule is that *everything in the transcription is covered in a dialog act segment, with nothing left over*.
- Finally, in case of doubt, annotators were instructed *to use two segments, instead of one*.

The guidelines for *labeling* dialog acts again center around the speaker's intention — as expressed in an utterance — to exchange information, contribute to the social structure of the group, carry out an action, get something clarified, or express an attitude towards something or someone. The schema contains fifteen types of dialog acts, twelve of which are proper dialog acts and three 'quasi-acts'. The full set of class labels in the schema is given in Table 4.1. The labels INFORM and SUGGEST concern giving information and giving suggestions. The label ASSESS concerns evaluating or commenting on something, such as expressing agreement with the opinion of someone else. The dialog acts in the group 'Elicits' are those that explicitly elicit a response from one of the other participants (which may or may not be given). They have a strong forward-looking aspect, something which plays a distinct role in the detailed analysis of the addressee annotations in Chapter 5.

Label	**Number of utterances**	**Frequency**
Task related		
INFORM	28891	28.3%
SUGGEST	8114	7.9%
ASSESS	19020	18.6%
Elicits		
ELICIT-INFORM	3703	3.6%
ELICIT-OFFER-OR-SUGGESTION	602	0.6%
ELICIT-ASSESSMENT	1942	1.9%
ELICIT-COMMENT-UNDERSTANDING	169	0.2%
Quasi-acts		
BACKCHANNEL	11251	11.0%
STALL	6933	6.8%
FRAGMENT	14348	14.0%
Other		
OFFER	1288	1.3%
COMMENT-ABOUT-UNDERSTANDING	1931	1.9%
BE-POSITIVE	1936	1.9%
BE-NEGATIVE	77	0.1%
OTHER	1993	2.0%
Total	102198	100.0%

Table 4.1: The list of all dialog act class labels that can be found in the AMI dialog annotation guidelines, and the distribution of the class labels over the data [AMI Consortium, 2005a].

The dialog acts in the group 'Quasi-acts' are not proper dialog acts at all, but are present in the annotation schema to account for something in the transcript that does not really convey a speaker's intention. Throughout the rest of this thesis, dialog acts from this group will be referred to as 'quasi-acts'. Furthermore, although the class OTHER does actually represent a speaker intention, it is present as a 'bucket' class rather than a real part of the label set, and therefore it has also been included in the group 'quasi acts' for all analyses presented in this thesis. The term 'proper dialog act' will apply to the labels not taken as 'quasi-acts'. The remainder of the dialog acts in the schema concern the speaker expressing an offer, commenting about his or her understanding (such as requests for clarification), and utterances that are mainly intended to affect the social processes of the group.

Most of the scenario data in the AMI corpus has been annotated for dialog acts, resulting in over 100,000 utterances. As can be seen in Table 4.1, which also shows the distributions of the labels over the data, about one third of those are quasi-acts. Of the rest, by far the most frequently occurring labels are INFORM, SUGGEST, and ASSESS. The annotations were produced by three annotators (DHA, S95 and VKAR) who annotated non-overlapping parts of the corpus.

Additionally, one meeting was annotated by all three annotators, plus one extra annotator (MA). This meeting (named IS1003d) served as reliability data, to ascertain the inter-annotator agreement of the schema and annotators. The dialog act annotations were compared[2] on the agreed segments, that is, the utterances on which two annotators agreed on the start and end boundaries of the segment (in terms of word positions). Table 4.2 shows, for each pair of annotators, the number of segments that the two annotators identified in common. Table 4.3 presents Krippendorff [1980]'s α for multiple annotators, for all dialog acts and the proper acts only, and once for each of the single annotators left out of the calculation. All values are between 0.55 and 0.61.

	VKAR	DHA	S95	MA
VKAR	1563 (838)	735 (344)	756 (412)	955 (479)
DHA		1234 (764)	795 (430)	829 (405)
S95			1287 (823)	855 (483)
MA				1504 (897)

Table 4.2: The number of segments that each pair of annotators identified in common in meeting IS1003d. The numbers between parenthesis are for the proper dialog acts only. On the diagonal the total number of segments identified by each single annotator are given.

Group	All dialog acts		Proper dialog acts only	
	N	α	N	α
All	503	0.57	226	0.58
Without VKAR	637	0.55	315	0.56
Without MA	563	0.58	266	0.57
Without S95	640	0.60	271	0.58
Without DHA	643	0.59	322	0.61

Table 4.3: Overview of the multi-annotator α values for dialog act annotation, for the group of all four annotators and for each of the single annotators left out of the group once. The number of agreed segments for each group is given as N; the α values are given for all dialog acts as well as for the subset of the proper dialog acts only.

[2] The results presented in Tables 4.2, 4.3, 4.4, and 4.5 were collected from unpublished work by Rieks op den Akker.

4.3 The AMI Addressee Annotations

A part of the AMI corpus is also annotated with addressee information [Jovanović et al., 2006; Jovanović, 2007]. All proper dialog acts were assigned a label indicating who the speaker addressed his speech to (was talking to). In the type of meetings considered in the AMI project, most of the time the speaker addresses the whole group, but sometimes his dialog act is addressed to some particular individual. This can be, for example, because he wants to know that individual's opinion, or is presenting information that is particularly relevant for that individual. The basis of the concept of addressing underlying the AMI addressee annotation schema originates from Goffman [Goffman, 1981]. The addressee is the participant *"oriented to by the speaker in a manner to suggest that his words are particularly for them, and that some answer is therefore anticipated from them, more so than from the other ratified participants"*. Sub-group addressing hardly occurs, at least in the meetings that make up the AMI corpus, and was not included in the schema. Thus, dialog acts are either addressed to the group (*G-addressed*) or to an individual (*I-addressed*). Annotators could also use the label UNKNOWN when they were unsure about the intended addressee of an utterance.

The AMI addressee annotation schema was applied to a subset of 14 meetings from the corpus[3], containing 9987 dialog acts in total. Annotation of addressee was done by the same annotator who produced the dialog act annotations for a particular meeting. Table 4.4 shows the label distribution in the meetings annotated with addressee. For reliability data, the same meeting was used as for the dialog act annotation (meeting IS1003d). Table 4.5 presents Krippendorff's α for multiple annotators for the dialog acts annotated with addressee, for all annotators and once for each of the single annotators left out of the calculation. A more detailed analysis of these annotations is presented in Chapter 5.

Type		Number of utterances	Frequency
Quasi-acts (no addressee)		3397	34.0%
I-addressed			
	A	804	8.1
	B	598	6.0
	C	638	6.4
	D	703	7.0
	Total	2743	27.5%
G-addressed		3104	31.1%
Unknown		743	7.4%
Total		9987	100.0%

Table 4.4: The distribution of labels in the part of the AMI corpus annotated with the addressee annotation schema.

[3] This concerns the meetings ES2008a, TS3500a, IS1000a, IS1001a, IS1001b, IS1001c, IS1003b, IS1003d, IS1006b, IS1006d, IS1008a, IS1008b, IS1008c, and IS1008d

Group	N	α
All	120	0.38
Without VKAR	213	0.36
Without MA	157	0.39
Without S95	162	0.37
Without DHA	198	0.53

Table 4.5: Overview of the multi-annotator α values for addressee annotation, for the group of all four annotators and for each of the single annotators left out of the group once. The number of agreed segments for each group is given as N.

4.4 The AMI Focus of Attention Annotations

A subset of meetings in the AMI corpus were also annotated with visual Focus of Attention (FOA) information derived from head, body and gaze observations [Ba and Odobez, 2006]. FOA forms an important cue for, among other things, addressing behavior. The FOA annotation contains, for every participant in the meeting, at all times throughout the meeting, whom or what he is looking at. This annotation schema was applied to the same subset of 14 meetings that was used for addressee annotation (but by other annotators). The FOA annotation was done with a very high level of agreement and with very high precision: changes are marked in the middle of eye movement between old and new target with α agreement between annotators ranging from 0.84 to 0.95 [Jovanović, 2007, page 80].

4.5 Summary

This chapter presented three layers of annotation from the AMI corpus. The dialog act annotations were produced for almost the whole corpus. Part of the dialog act annotations are used in Chapter 6 for experiments in explicitly modeling the (inter)subjectivity in annotations. The addressee annotations were produced for 14 meetings, by the same annotators who produced the dialog act annotations for those particular meetings. The Focus of Attention annotations were produced for those same 14 meetings, by other annotators. The next chapter addresses the inter-annotator agreement of the addressee annotations in more detail. Among other things, the FOA annotations are used to define a more reliable subset of the addressee annotations, and machine-learning performance for addressee classification is explored for this more reliable subset. The addressee annotations are also used for the experiments in modeling the (inter)subjectivity in annotations in Chapter 6.

Chapter 5

Contextual Agreement

This second part of the thesis explores the relation between agreement, data quality and machine learning, using the AMI corpus introduced in the Chapter 4. This chapter describes a novel approach that uses contextual information from other modalities to determine a *more reliable subset* of data, for annotations that have a low overall agreement. Furthermore, a preliminary analysis is presented on machine-learning performance within that more reliable subset as opposed to the performance on the overall data set.

Beigman Klebanov and Shamir [2006] argued that, if data has been annotated with a very low level of inter-annotator agreement, one way of making (parts of) the data more usable may be to find out whether one can pinpoint a subset of the data that has been annotated with a higher level of inter-annotator agreement. This more reliable subset can then be used for training and testing of machine learning, with a higher confidence in the validity of the results. In order to find this subset, they proposed an approach in which all data is annotated multiple times. They used annotations from 20 separate annotators on a data set annotated for lexical cohesion. Given these annotations they induced random *pseudo-annotators* from each annotator. Each pseudo-annotator marked up the data with the same distributions as the actual annotator, but chose the items at random. Given these pseudo-annotators, they calculated the probabilities that a certain item would be marked with a certain label by more than N of the random pseudo-annotators. They found that, for items that were marked with a specific label by at least 13 out of 20 human annotators, the label could not have been the result of random annotation processes, with an overall confidence of 99%. For a different data set, concerning markup of metaphors in text, they showed that an item needed to be marked by at least 4 out of 9 annotators to make it sufficiently improbable that the label was the result of random coding behavior [Beigman Klebanov et al., 2008].

A major drawback of the method described above is, of course, that it requires all training and test data to be multiply annotated — without exception. This requires an investment that otherwise might be spent on annotating more content, or different content, or on feature selection and classification experiments, and so forth. A second important drawback to this approach has to do with deploying the classifiers, trained on such a subset, in practice. Classifiers are often intended to ultimately serve as a replacement for human annotation effort. They are to be applied to new, unseen data. This data has not been annotated by humans, so it is unknown *a priori* whether specific new instances would belong to the domain in which the classifier is qualified to render a judgement, that is, the reliable subset of data for which the classifier was trained and tested. The problem is, in other words, that the performance of the classifier as observed *on the reliable subset* in the testing phase is not necessarily a valid indicator of the performance of the classifier on the new, unseen data, as the classifier will assign a label to *all instances* in the new data. The problem would be solved if it were possible to deduce for new, unseen instances whether they belong to this more reliable subset, *without* having to resort again to a multitude of human annotators.

This chapter investigates whether this solution can be achieved, for the case of addressee detection on dialog acts, by taking the *multimodal context* of utterances into account. Naturally, the approach still relies on a certain amount of multiply annotated data. However, in contrast to the method described above, only a limited part of the data needs to be annotated more than once.

The chapter is structured as follows. Section 5.1 concerns the basic inter-annotator agreement for the addressee annotations. Section 5.2 considers the relevance of the multimodal context of utterances to the level of inter-annotator agreement with which they are annotated. In Section 5.3 it is shown that the multimodal context of utterances can indeed be used to determine a more reliable subset of the annotations. Finally, the chapter ends with a preliminary exploration of machine-learning performance within that more reliable subset as opposed to the performance on the overall data set, in Section 5.5, followed by discussion and conclusions. The work presented in this chapter is based on earlier publications with Dirk Heylen and Rieks op den Akker [Reidsma et al., 2008a; Reidsma and op den Akker, 2008].

5.1 Basic Agreement and Class Maps for Addressee

The inter-annotator agreement for the AMI addressee annotations was given in Section 4.3. Recall that the value of Krippendorff's multi-annotator α was 0.44. This indicates a quite low level of agreement: it falls into the range usually reported on highly subjective annotation tasks. Before the contextual dependencies for the inter-annotator agreement are discussed in the next section, some more information about the basic agreement analysis is given here. Table 5.1 presents the pairwise agreement values expressed in Krippendorff's multi-annotator α [1980]. Table 5.2 shows an example of a confusion matrix for the addressee annotation, representative of the other confusion matrices. The values in the confusion matrix suggest that it is not so much problematic to decide *which* individual was addressed as it is to distinguish between I-addressed utterances versus utterances that are G-addressed or labeled UNKNOWN. In the remainder of this section, inter-annotator agreement is discussed for two derived versions of the label set, namely for the annotation without the label UNKNOWN and for the class map in which the annotation is reduced to the binary distinction I-addressed/G-addressed. Note that all results presented in this chapter concern only proper dialog acts and are based upon a pairwise comparison of agreed segments, as in the table below.

	MA	VKAR	DHA	S95
MA	.	0.57	0.32	0.46
VKAR		.	0.36	0.50
DHA			.	0.31
S95				.

Table 5.1: Pairwise agreement (Krippendorff's α) for addressee annotations by four annotators, on agreed segments annotated as proper dialog act.

	A	B	C	D	G	U	$\sum$
A	46				26	2	74
B	1	25			12	1	39
C			38	1	10	1	50
D				63	16	4	83
G	7	5	9	10	155	5	191
U	16	1	4	4	15	2	42
$\sum$	70	31	51	78	234	15	479

Table 5.2: Confusion matrix for annotators VKAR and MA for the addressee labels of agreed segments in meeting IS1003d. Krippendorff's α is 0.57 for this matrix.

5.1.1 Reliability for the Addressee Label UNKNOWN

The annotators indicated whether an utterance was addressed to a particular person or to the whole group. They could also use the label UNKNOWN, if they could not decide who was being addressed. All four annotators used this label at some point in their annotation of meeting IS1003d. Given the annotation guidelines, there might have been two reasons why an annotator would use the label UNKNOWN. Firstly, the utterance may have been ambiguously or unclearly addressed, making it impossible to choose a single label like the annotation task requires. The reason for assigning the label UNKNOWN then lies within the content. A certain amount of inter-annotator agreement for this label could be expected, and the applicability of the label could be learnable and worth learning. Secondly, the utterance may have been unambiguously addressed, but nevertheless the annotator may have been uncertain about his own judgement, for example because he was tired, or did not understand what was being said. In that case, the reason for assigning the label UNKNOWN lies completely with the annotator, rather than with the content. This second type of uncertainty would *not* cause the label UNKNOWN to exhibit a large inter-annotator agreement, and would by far be less interesting to learn to classify.

The question to be answered here is then: does the uncertainty in the addressee annotation, expressed by the annotator assigning the label UNKNOWN, reflect an attribute of the content, or rather an attribute of the specific annotator who assigned the label at a certain point? Inspection of the confusion matrices shows a clear answer to this question. The matrix displayed in Table 5.2 is certainly representative in this respect. Inter-annotator agreement on the applicability of the label is virtually non-existent for each and every pair of annotators. This means that the occurrence of the label UNKNOWN in the corpus does not seem to give any useful information about the annotated content at all.

For this reason, it was decided to remove all UNKNOWN labels from the corpus before proceeding with further analysis. That is, for all segments that an annotator labeled UNKNOWN, the label was removed, and the segment was taken as if the annotator had not labeled it with addressee at all — reducing the number of segments available for the analyses presented later in this chapter by one, for that annotator, but leaving the number of segments available from the other annotators unaffected.

The effect of this data set reduction on the inter-annotator agreement on the remaining segments is shown in Table 5.3. This table presents the α values for the addressee annotations computed on all proper dialog acts versus the α values calculated after removing all UNKNOWN labels from the corpus. The increase in level of inter-annotator agreement ranges from 0.10 to 0.16. This does not only hold for the overall data set reported in this table, but also for each and every contextual subset of the data set reported later in this chapter.

	Inc. UNKNOWN	Excl. UNKNOWN
MA vs VKAR	0.57	0.67
DHA vs S95	0.31	0.47
S95 vs VKAR	0.50	0.63
DHA vs VKAR	0.36	0.47
MA vs S95	0.46	0.59
DHA vs MA	0.32	0.43

Table 5.3: Inter-annotator agreement for all proper dialog acts versus only the dialog acts not annotated with the UNKNOWN addressee label.

5.1.2 Class Map: Group/A/B/C/D vs Group/Single

The second label that really contributed to the disagreement according to the confusion matrix of Table 5.2 is the GROUP label. However, in large contrast to the label UNKNOWN discussed above, the majority of its occurrences are actually agreed upon by at least some of the annotators. From the confusion matrices it can nevertheless be seen that annotators cannot make the global distinction between G-addressed and I-addressed utterances with a high level of agreement: there is a lot of confusion between the label G on the one hand and A, B, C and D on the other hand. If annotators see an utterance as I-addressed they subsequently do not have much trouble determining who of the single participants was addressed: there is much less confusion between the individual addressee labels A, B, C and D.

This observation is quantified using a class mapping of the addressee annotation in which the individual addressee labels A, B, C and D are all mapped onto the label S. Table 5.4 shows pairwise α agreement for this class mapping, next to the values obtained for the annotations after removing the label UNKNOWN from the data set (see also the previous section). Clearly, agreement on who of the participants was addressed individually is a major factor in the overall agreement.

	Normal label set (excl. UNKNOWN)	Class map $(A, B, C, D) => S$
MA vs VKAR	0.67	0.55
DHA vs S95	0.47	0.37
S95 vs VKAR	0.63	0.52
DHA vs VKAR	0.47	0.37
MA vs S95	0.59	0.46
DHA vs MA	0.43	0.32

Table 5.4: Pairwise α agreement for the unmapped label set (left) and for the class mapping $(A, B, C, D) => S$ (right), both after removing the label UNKNOWN from the data set.

5.2 The Multimodal Context of Utterances

The remainder of this chapter concerns multimodal contextual agreement. To a large extent multimodal behavior is a holistic phenomenon, in the sense that the contribution of a specific behavior to the meaning of an utterance needs to be decided upon in the context of other behaviors that coincide, precede or follow. A nod, for instance, may contribute to a conversation in different ways when it is performed by someone speaking or listening, when it is accompanied by a smile, or when it is a nod in a series of more than one. When we judge what is happening in conversational scenes, our judgements become more accurate when we know more about the context in which the actions have taken place. The occurrences of gaze, eye-contact, speech, facial expressions, gestures, and the setting determine our interpretation of events and help us to disambiguate otherwise ambiguous activities.

Annotators, who are requested to label certain communicative events, be it topic, focus of attention, addressing information or dialog acts, get cues from both the audio and the video stream. Some cues are more important than others: some may be crucial for correct interpretation whereas others may become important only in particular cases. The reliability of annotations may crucially depend on the presence or absence of certain features, even if these features are not mentioned in the annotator instructions. Using or not using the video and audio while annotating may therefore have a large impact on the agreement achieved for certain annotations. Also, one annotator may be more sensitive to one cue rather than to another. This means that the agreement between annotators may depend on particular variations in the multimodal input.

Within the AMI corpus, one of the more obvious annotations to which this bears relevance is the addressee annotation. The visual focus of attention (FOA) of speakers and listeners is an important cue in multimodal addressing behavior. The combination of these two layers will therefore be used in an attempt to determine a more reliable subset of the corpus.

5.3 Finding More Reliable Subsets

This section describes two 'more reliable subsets' within the AMI addressee annotations (with the UNKNOWN label removed as discussed in Section 5.1). The first is centered around the multimodal context of the utterance. The second uses the context determined by the type of dialog act for which the addressee was annotated. The aim of these contextual agreement analyses, as described in the introduction to this chapter, is to be able to pinpoint a more reliable subset in the data without having all training and test data be annotated by multiple annotators.

5.3.1 Context: Focus of Attention

Visual Focus of Attention (FOA) of speakers and listeners is an important cue in multimodal addressing behavior. In this section it is investigated to what extent this cue impacts the task of annotators who observe the conversational scene and have to judge who was addressing whom. FOA annotations are a manifest type of content, do not need extensive discourse interpretation, and can be annotated with a very high level of inter-annotator agreement. This makes them especially useful when they can serve as multimodal context for finding a more reliable subset of the addressing data, because it is more likely that this context can be retrieved for new, unseen data, too.[1]

Table 5.5 lists three different FOA contexts that each define a different subset of all addressee annotations. The contexts are defined with respect to the Focus of Attention of the speaker during the utterance. Context I concerns utterances during which the speaker's gaze is directed only to objects (laptop, whiteboard, or some other artefact) or nowhere in particular. One might expect that in this context the annotation task is harder and the inter-annotator agreement lower. Contexts II and III concern the utterances during which the speaker's gaze is directed at least some of the time to other persons (only one person, for context II, or any number of persons for context III). The expectation was that utterances in contexts II and III respectively would also exhibit a difference in inter-annotator agreement. When a speaker looks at only one participant, agreement may be higher than when the speaker looks at several (different) persons during an utterance.

Table 5.6 presents α values for the pairwise inter-annotator agreement for the three subsets defined by the three FOA contexts from Table 5.5, compared to the α values for the whole data set that were presented in Section 5.1.1. Inter-annotator

[1] Although it should be noted that state-of-the-art recognition rates are still too low for this, in the order of 60% frame recognition rate [Ba and Odobez, 2007; Voigt and Stiefelhagen, 2008].

Context	Description
I	Only those utterances during which the speaker does not look at another participant at all (he may look at objects, though)
II	Only those utterances during which the speaker does look at one other participant, but not more than one (he may additionally look at objects)
III	Only those utterances during which the speaker does look at one or more other participants (he may additionally look at objects)

Table 5.5: The three different contexts defined by different conditions on the FOA annotation that are used to find more reliable subsets of the addressee annotations.

	All (excl. UNKNOWN)	I	II	III
MA vs VKAR	0.67	0.60	0.78	0.77
DHA vs S95	0.47	0.41	0.57	0.57
S95 vs VKAR	0.63	0.59	0.69	0.66
DHA vs VKAR	0.47	0.42	0.48	0.51
MA vs S95	0.59	0.57	0.63	0.62
DHA vs MA	0.43	0.32	0.53	0.56

Table 5.6: Pairwise α agreement for the subsets defined by the three contextual FOA conditions, compared to α agreement for the full data set (without the label UNKNOWN).

agreement for the addressee annotation is consistently lowest for context I whereas contexts II and III consistently score highest. When a speaker looks at one or more participants, the agreement between annotators on addressing consistently becomes higher. Contrary to expectations there is no marked difference, however, between the contexts where, during a segment, a speaker only looks at one participant or at several of them (context II versus III).

In conclusion, it can be said that the subset of all utterances during which the speaker looks at some other participants at least some of the time, defined by context III, forms a more reliable subset of the addressee annotations as defined in the introduction to this chapter. This subset, containing two thirds of all utterances annotated with addressee, was used for the machine-learning experiments described in Section 5.5.

5.3.2 Context: Elicit Dialog Acts

The second contextual agreement analysis presented here concerns a certain specific group of dialog acts. Op den Akker and Theune [2008] discussed that forward looking dialog acts, and more specifically, 'Elicit' types of dialog act, are more often I-addressed, and tend to be addressed more explicitly. This might possible cause elicit dialog acts to exhibit a higher inter-annotator agreement. Table 5.7 presents the pairwise α inter-annotator agreement values for all proper dialog acts, the 'elicit' dialog acts only, and the proper acts without the 'elicit' acts. Clearly, the agreement for 'elicit' acts is a lot higher. Apparently the intended addressee of elicits is relatively easy to determine for an outsider (annotator); this may have to do with differences in how speakers express 'elicit' acts and other forward looking acts as suggested by Op den Akker and Theune [2008].

	All proper acts	Elicits only	No elicits
MA vs VKAR	0.67	0.87	0.64
DHA vs S95	0.47	0.84	0.38
S95 vs VKAR	0.63	0.80	0.61
DHA vs VKAR	0.47	0.58	0.41
MA vs S95	0.59	0.76	0.57
DHA vs MA	0.43	0.57	0.40

Table 5.7: Pairwise α agreement for all proper dialog acts and for the elicit dialog acts only.

5.4 Discussion and Summary of Addressing Agreement

Throughout this section pairwise α agreement scores have been presented for different class mappings and subsets of the addressee annotations in the AMI corpus. The different effects noted about these scores were consistent. That is, although only a few combinations of scores are reported, all different combinations of mappings and subsets consistently show the same patterns. For example, all relative differences between the FOA contexts hold for the 'all agreed proper dialog acts' condition, the 'excluding UNKNOWN' condition, and for the $(A, B, C, D) => S$ class mapping.

The following conclusions can be summarized for the inter-annotator agreement of addressee annotations: (1) the label UNKNOWN does not give any information about the annotated content; (2) there is a large confusion between dialog acts being G-addressed or I-addressed, but if the annotators agree on an utterance being I-addressed they typically also agree on the particular individual being addressed; (3) 'elicit' dialog acts are easier to annotate with addressee than other types of dialog act; and (4) utterances during which the speaker's focus of attention is directed to one or more other participants are consistently annotated with more agreement than those during which the speaker's FOA is not directed to any participant.

5.5 Contextual Performance of Classification

In Section 5.3.1, it was shown that FOA context III defines a more reliable subset (see introduction to this chapter) of addressee annotations. This section presents an exploration of machine-learning performance in relation to this more reliable subset. The expectation was, of course, that performance would be higher within the more reliable subset, for two reasons. Firstly, the high inter-annotator agreement achieved for this subset will result in the annotations containing more *consistent* information that a classifier can model, and, secondly, the higher agreement may have been caused by the fact that it is simply easier to determine the intended addressee of an utterance within the more reliable subset.

5.5.1 Approach

A Bayesian Network was trained to classify the addressee of utterances using a number of lexical and multimodal features from the AMI corpus. Compared to Jovanović [2007], a limited set of features was selected. Their lexical features, and features for focus of attention, were included. Their local context features, such as 'previous addressee' or 'previous dialog act type', were omitted. The full set of 14 meetings annotated with FOA and addressee was used for the experiment. As described in Section 5.3.1, this data set was transformed into three versions: (1) all data; (2) the subset defined by FOA context I (one third of the data); and (3) the subset defined by FOA context III (two thirds of the data). A training set was constructed from by taking 90% of all data, proportionally divided over the two subsets. Three test sets were defined, one each for the two FOA contexts containing the remaining 10% of that subset (TST_I and TST_{III}), and a test set TST_{ALL} that was the union of TST_I and TST_{III}. For training and testing, a cross-validation procedure was used on several of such 90%/10% divisions of the data in order to be able to graph mean and standard deviation values. The Bayesian Network, adapted from Jovanović [2007], was trained on all training data. The performance of the resulting network was evaluated using the three versions of the test data.

5.5.2 Results

Table 5.8 displays the results obtained using the approach from the previous section. Note that the standard deviations are very high, and often much larger than the differences between the performance on the different test sets. Table 5.9 therefore shows some *relative* performance values, that is, mean and standard deviation of the *difference in performance obtained on the different test sets in the same train/test run*. This figures show that on average, there is a slight but real increase in accuracy for TST_{III}, which was defined as the test set composed of data from the more reliable subset of the addressee annotations. This increase can be attributed to the increase in precision on the four I-addressed labels A, B, C, and D.

5.6 Summary and Discussion

This chapter showed that it is possible to take data which was annotated with a relatively low level of inter-annotator agreement, and use the results of an extended

Metric	TST_{ALL}	TST_I	TST_{III}
Accuracy	66.4 (8.4)	62.7 (10.2)	68.0 (8.1)
Precision on the four I-addressed labels	61.0 (14.3)	43.0 (12.9)	66.6 (17.5)
Recall on the four I-addressed labels	49.0 (9.3)	47.6 (9.1)	48.0 (7.9)

Table 5.8: Performance of the Bayesian classifiers on the three test sets using cross-validation. Numbers are given as *mean (standard deviation)*.

Metric	Value
$Acc(TST_{III}) - Acc(TST_{ALL})$	1.6 (2.0)
$Acc(TST_I) - Acc(TST_{ALL})$	-3.7 (5.0)
$PrecI(TST_{III}) - PrecI(TST_{ALL})$	5.6 (8.0)
$PrecI(TST_I) - PrecI(TST_{ALL})$	-18.0 (12.9)
$RecI(TST_{III}) - RecI(TST_{ALL})$	-1.0 (6.4)
$RecI(TST_I) - RecI(TST_{ALL})$	-1.4 (9.6)

Table 5.9: Relative performance differences obtained on the different test sets in the same train/test run. Numbers are given as *mean (standard deviation)*.

agreement analysis to determine certain ‘more reliable subsets’ in which the annotated data has a higher quality. FOA context III (‘at least some person in speaker’s FOA during utterance’) defines one such more reliable subset on the addressee annotations. Machine-learning performance restricted to that subset shows only a very slight increase of accuracy, which can mostly be attributed to increased precision obtained on the different individual addressee labels for I-addressed utterances. Nevertheless, it is argued here that it is worthwhile to have identified such a more reliable subset based on the multimodal context defined by the FOA annotations. In the first place, this particular more reliable subset can be identified without having to annotate all training and test data by several annotators. That allows one to collect and annotated more content, of which the ‘more reliable subset’ can serve as training material for a classifier. In the second place, FOA annotations are a manifest type of content, do not need extensive discourse interpretation, and can be annotated with a very high level of inter-annotator agreement. This makes it more likely that this context can be retrieved for new, unseen data, too, which allows one to determine for unseen data whether the classifier is qualified to render a judgement.

In theory at least, this could be exploited in a practical application, to build a classifier that limits its judgements to the more reliable parts of the data. Given the specific example presented in this chapter, an addressee detection module might be built which only assigns an addressee to an utterance in FOA context III, and in all other cases labels an utterance as ‘addressee cannot be determined’. Such a detection module would achieve a much higher precision than a module that tries to assign an addressee label regardless; this happens at the cost of overall recall because it will not even try to assign a label in the *less* reliable subset of the data.

Chapter 6

Explicitly Modeling (Inter)Subjectivity

This chapter — part of which was presented at the workshop on Human Judgements in Computational Linguistics [Reidsma and op den Akker, 2008] — continues the second part of this thesis, in which the relation between agreement, data quality and machine learning is explored.

The main questions for this chapter target the *subjective* aspect of annotator disagreement instead of the error aspect. Is it possible, given subjective annotations of different annotators, to separate the *overlap* and the *differences* of the mental conceptions of the annotators as reflected in the data? Can 'intersubjectivity', present in the way different annotators interpret the same observed interaction, be modeled explicitly? Rather than looking at inter-annotator agreement within one multiply annotated data fragment, such as AMI meeting IS1003d, the method followed in this chapter is designed to work with a number of much larger sets of data annotated by each of the annotators, without overlap between annotators.

Using these annotator-specific data sets, annotator-specific machine-learning models are trained. The resulting models are combined to find and model the overlap and disjunction in them, which ideally should reflect the overlap and disjunction in the mental conceptions of the annotators. As a side effect, this approach also offers a new way of looking at the 'more reliable subsets' that were the topic of the previous chapter. Instead of using the properties of a small multiply annotated data set to determine these subsets, the defining properties of the subset are directly inferred from the intersubjective overlap in the mental conceptions.

The chapter is structured as follows. Section 6.1 describes the setup and results of the first experiment towards modeling intersubjectivity in annotations, using 'Yeah-utterances', a subset of the AMI dialog act annotations. The experiment left a number of important questions unanswered. Section 6.2 addresses several of the most important ones. First, it is shown in Section 6.2.1 that the results can be generalized to at least one other data set, namely the addressee annotations that were also used in the previous chapter. Next, some characteristics of the machine-learning performance in relation to the modeling of intersubjectivity are discussed in Sections 6.2.2 and 6.2.3. Finally, the relation between the method presented in this chapter and the method of ensemble learning is discussed in some detail in Section 6.2.4. The chapter ends with a discussion of some of the many open questions and of future work in Section 6.3.

6.1 Modeling Subjectivity for 'Yeah' Utterances

The first illustration, and test case, for modeling the intersubjectivity in annotations concerns the human annotations and automatic classification of a particular type of utterances (dialog acts) in the AMI corpus: the *"Yeah-utterances"*, that is, utterances that start with the word "yeah". When one suspects that the inter-annotator disagreement in annotations originated from differences in the mental conceptions of the annotators, the first step is to test whether the differences in the annotations are systematic. If they are not, it makes no sense to try to model at subjective idiosyncrasies and the intersubjective overlap. The next step is to build a classifier that actually embodies this subjectivity and intersubjectivity. In this section, the approach and the results for both steps are discussed for yeah-utterances, a subset of the AMI dialog annotations.

6.1.1 'Yeah' Utterances

Heylen and Op den Akker [2007] discussed how response tokens such as "yeah", "okay", "right" and "no" have the interest of linguists because they may give a clue about the stance that the listener takes towards what is said by the speaker [Gardner, 2004]. Of these, this section concerns the yeah-utterances in the AMI corpus. Yeah-utterances are defined here as all utterances that either consist of the single world "yeah", or consist of a longer sentence *starting* with the word "yeah". They make up a substantial part of the dialog acts in the AMI meeting conversations (about eight percent). They are often ambiguous. In single word utterances, they are used, for example, as backchannel, or to express agreement with the opinion of the speaker. As part of a larger utterance, they may indicate agreement, they may serve as a signal to keep or take the floor, or may play yet other roles. In order to get information about the stance that participants take with respect towards the issue discussed in a meeting it is important to be able to distinguish such different occurrences of yeah-utterances.

	VKAR	DHA	S95
VKAR	(215)	0.36 (111)	0.36 (132)
DHA	.	(221)	0.45 (160)
S95	.	.	(224)

Table 6.1: The pairwise alpha values for meeting IS1003d, which was annotated by all three annotators. Between parenthesis the number of agreed DA segments that start with "Yeah". On the diagonal the total number of yeah-utterances identified in IS1003d by each single annotator are given.

The class variables for dialog act types of yeah-utterances that are considered in this chapter are: ASSESS (AS), BACKCHANNEL (BC), INFORM (IN), and OTHER (OT), as these are the most frequently occurring labels for these type of utterances. Yeah-utterances are particularly difficult to annotate reliably. The inter-annotator agreement, calculated on the yeah-utterances of meeting IS1003d, is low, as can be seen in Table 6.1. This makes them a suitable subject for the modeling of (inter)subjectivity in this section. Whether the disagreement between annotators that causes this low level of agreement is systematic will be investigated in Section 6.1.3.

6.1.2 Division of the Data into Training and Test Sets

In order to understand the experiments described later, it is necessary to have a clear image of how much data there is, how it was divided over the annotators and how it was split into training and test sets. The experiments were done on the full set of yeah-utterances that are found in then dialog act annotations. This is in contrast with the usual approach for analysis of similarities and differences between annotators, in which only the multiply annotated subset of reliability data is used. Figure 6.1 visualizes the composition of the relevant data. The complete set of dialog act annotations contained 13,017 yeah-utterances, spread over many meetings (top row of Figure 6.1). These annotations had been produced by three annotators (DHA, S95 and VKAR) who annotated non-overlapping parts of the corpus (see second row of Figure 6.1). The data of each single annotator was split into a training and a test set (third and lowest row of Figure 6.1). This resulted in the disjunct training and test sets called TRN_{DHA}, TRN_{S95} ... TST_{VKAR}. Additionally, all three training sets were combined into one training set called TRN_{ALL}, and similarly the test set TST_{ALL} was constructed from TST_{DHA}, TST_{S95}, and TST_{VKAR}. The exact sizes of all test and training sets as well as the distribution of the four relevant class labels are given in Table 6.2. Given these data sets, the next section will discuss how they have been used to model intersubjectivity using machine learning.

<table>
<tr><td colspan="6">13017 'Yeah' Utterances, spread over all meetings</td></tr>
<tr><td colspan="2">Annotated
by DHA</td><td colspan="2">Annotated
by S95</td><td colspan="2">Annotated
by VKAR</td></tr>
<tr><td>TRN_{DHA}</td><td>TST_{DHA}</td><td>TRN_{S95}</td><td>TST</td><td>TRN_{VKAR}</td><td>TST</td></tr>
</table>

Figure 6.1: The division of all yeah-utterances in the corpus into subsets according to which annotators labeled them. Sizes of the per-annotator training and test sets, as well as the label distributions, are given in Table 6.2.

6.1.3 Approach

Classifiers behave as they are trained. When two annotators differ in the way they annotate, that is, have different mental conceptions of the concepts being annotated, one can expect that a classifier trained on the data annotated by one annotator behaves differently from a classifier trained on the other annotator's data. As Rienks

Class	TRN/TST_{ALL}	TRN/TST_{DHA}	TRN/TST_{S95}	TRN/TST_{VKAR}
BC	3043/1347	1393/747	670/241	980/359
AS	3724/1859	1536/1104	689/189	1499/566
IN	782/377	340/229	207/60	235/88
OT	1289/596	316/209	187/38	786/349
Total	8838/4179	3585/2289	1753/528	3500/1362

Table 6.2: Sizes of training and test data sets used and the distribution of class labels over these data sets for the different annotators.

describes, this property allows us to use all data in the corpus for finding systematic differences between annotators, instead of just the multiply annotated part of it [Rienks, 2007, page 105], as long as there is at least enough data and there is not too large a variance between the characteristics of the data sets annotated by each annotator. One can expect that a classifier C_A trained on data TRN_A annotated by A will perform better when tested on data TST_A annotated by A, than when tested on data TST_B annotated by B. In other words, classifier C_A is geared towards modeling the mental conception of annotator A (assuming that the appropriate discriminating features have been used).

Using the data described in the previous section, three annotator-specific classifiers C_{DHA}, C_{S95}, and C_{VKAR}, were trained, each on a different annotator's data. These classifiers are called 'expert classifiers', for being an expert on a certain annotator's manner of assigning labels. In the next section, an analysis of their performance is carried out in order to find out whether the disagreement between the different annotators is systematic. To this end the performance of the classifiers C_{DHA}, C_{S95}, and C_{VKAR} will be presented for the four test sets defined in the previous section (TST_{DHA}, TST_{S95}, TST_{VKAR}, and TST_{ALL}).

The next step is to distinguish the overlap, or intersubjective parts of the models, and the parts of the learned models that mirror the idiosyncrasies of the annotators used to train the models. To achieve this, a Voting Classifier C_{vote} is built, based on the votes of the three classifiers C_{DHA}, C_{S95}, and C_{VKAR}. The classifier C_{vote} only makes a decision when all three annotator-specific agree on the class label. This means that of all instances in the test set TST_{ALL}, only a limited subset U will be assigned a label by the classifier C_{vote}. Ideally, this classifier C_{vote} should embody the overlap of the mental conceptions of the three annotators.

6.1.4 Results

Table 6.3 displays the classification accuracy for the expert classifiers C_{DHA}, C_{S95}, and C_{VKAR} on the four test sets (three annotator-specific TST_{DHA}, TST_{S95}, and TST_{VKAR}, and the combined test set TST_{ALL}). There is a clear performance drop between using the test data from the same annotator from which the training data was taken, and using the test data from other annotators or all test data. This suggests that at least some of the disagreement in the annotations is systematic and may have resulted from idiosyncrasies in the annotators' mental conceptions that were picked up by the annotator-specific classifiers in the training process.

Classifier	Test set			
	TST_{DHA}	TST_{S95}	TST_{VKAR}	TST_{ALL}
C_{DHA}	**69**	64	52	63
C_{S95}	59	**68**	48	57
C_{VKAR}	63	57	**66**	63

Table 6.3: Performance of the annotator-specific classifiers (in terms of accuracy values – i.e. percentage correct predictions) trained and tested on various data sets. Results were obtained with a decision tree classifier, J48 in the Weka toolkit.

Given the three classifiers C_{DHA}, C_{S95}, and C_{VKAR}, each trained on the training data taken from one single annotator, a Voting Classifier C_{vote} was built that only outputs a class label when all three annotator-specific classifiers return the same label. As was to be expected, the *overall accuracy* for C_{vote}, calculated on all test data, is much lower than the accuracy of each of the single voters and lower than the accuracy of a classifier trained on all training data (C_{ALL}) (see Table 6.4). This is due to the many times C_{vote} does not assign a label, because that counts as an erroneous classification. However, the per-class precision of C_{vote} (the amount of correct labels assigned to instances from TST_{ALL}) is higher than that of any of the other classifiers, for each of the single classes (see Table 6.5).

Classifier	Accuracy on TST_{ALL}
C_{ALL} (8838)	67
C_{VKAR} (3500)	63
C_{S95} (1753)	57
C_{DHA} (3585)	63
C_{vote} (8838)	43

Table 6.4: Performance of the MaxEnt classifiers (in terms of accuracy values, that is, percentage correct predictions) tested on the combined test set TST_{ALL} (4179 yeah-utterances). The first column repeats the size of the training sets between brackets.

Class	Classifier				
	C_{vote}	C_{DHA}	C_{S95}	C_{VKAR}	C_{ALL}
BC	71	65	63	71	69
AS	73	62	64	61	66
IN	60	58	34	52	50
OT	86	59	32	57	80

Table 6.5: Precision values per class label for all classifiers.

6.1.5 Discussion

The Voting Classifier C_{vote}, presented above, was built in an attempt to explicitly separate the intersubjective overlap and the annotators' idiosyncrasies as picked up in the training process by the annotator-specific classifiers C_{DHA}, C_{S95}, and C_{VKAR}. It assigns a class label only for the subset of the test data for which the annotator-specific classifiers agree on the class label. In this way, an increased precision was achieved. This is what one would expect when the overlap between the classifiers is related to overlap between the mental conceptions of the annotators, that is, when C_{vote} models the intersubjectivity between the annotators. Note that, although precision went up, accuracy went down, because the Voting Classifier is designed to restrict judgements to the cases where annotators would have agreed — and, presumably, therefore to the cases in which users of the data are able to agree to the judgements as well. All cases outside of that restriction are by definition not recalled. It is possible that a class label is even completely 'lost' for the Voting Classifier, which in the example above happened for the OTHER class which reached a high precision but a recall of less than five percent.

Assuming that the results presented in the previous section did actually express the intended effect of modeling intersubjectivity, a Voting Classifier built from annotator-specific expert classifiers forms a cautious kind of classifier that would generalize well to new annotators. The Voting Classifier picks up a kind of 'common sense' in the most literal meaning of the word. Building a Voting Classifier that models intersubjectivity makes it possible to reason about subjective constructs and the intersubjective elements to them as they are found in corpora. Another advantage is that when such a classifier is deployed in a practical application — assuming that the end users of the application have the same common sense — the classifier will render judgements with which the end users of the application would agree as well.

The results presented above do also raise a host of additional questions. The next section will address a few of the most important of those.

6.2 Exploring More Aspects of Subjective and Voting Classifiers

The initial results on developing a Voting Classifier for yeah-utterances presented in the previous section raise many questions. Do the results and conclusions generalize to other data sets? After all, yeah-utterances are a quite specific data set. Is it possible to find more evidence for the results being caused by the Voting Classifier modeling the intersubjectivity, and not by something else? What is the theoretical and practical relation between ensemble learning and the Voting Classifier discussed above? This question is very important, because the method of ensemble learning does not need to invoke concepts of subjectivity and intersubjectivity, and nevertheless can achieve increased performance by training separate classifiers on different subsets of data and combining the outcome. In this section, these questions are explored, supported by additional experiments.

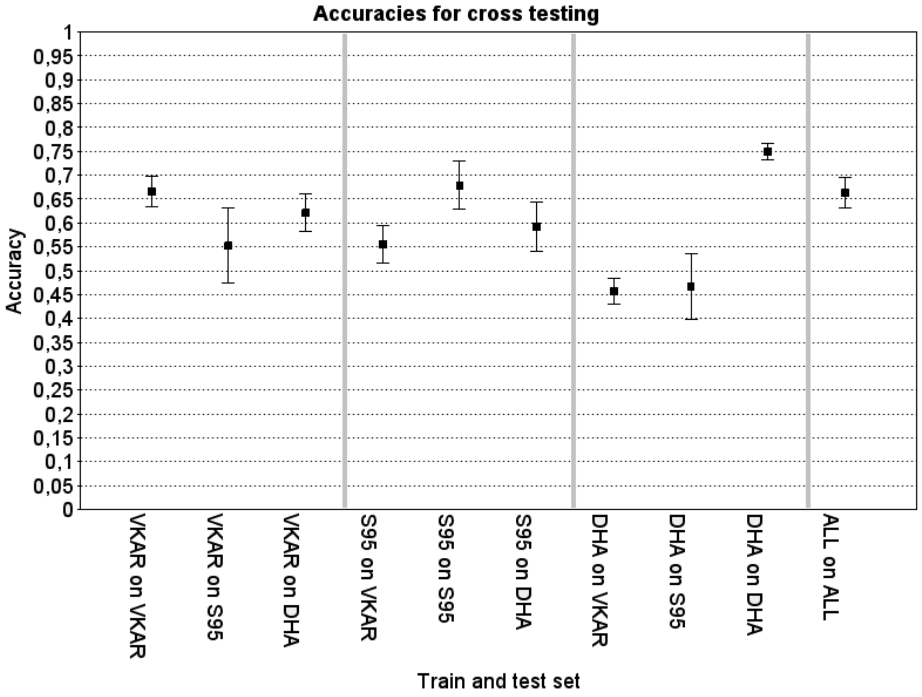


Figure 6.2: Cross annotator training/testing results for addressee annotations. “X on Y” shows the performance of the classifier C_X (trained on the training set from annotator X) that was achieved on the test set TST_Y from annotator Y.

6.2.1 Generalisation to Other Data

The first question that must be addressed is: can these results be repeated on different data? This question is explored using the addressing data that was introduced in Section 4.3. The data, 6,590 utterances annotated with addressee by the same three annotators (VKAR, S95, and DHA), was divided into training and test sets using exactly the same procedure explained in Section 6.1.2. For each annotator we split an equal amount of their data randomly in two parts: 90% of the data is designated as training data and 10% as test data. The three training sets are also combined in one large training set TRN_{ALL}; the three test sets into one test set TST_{ALL}. Five separate classifiers are constructed using the same Bayesian Networks and features described in Section 5.5. Classifier C_{ALL} is trained on all training data. On each of the three annotator-specific training sets an annotator-specific expert classifier is trained (resulting in C_{DHA}, C_{S95}, and C_{VKAR}). Finally, the three annotator-specific expert classifiers are combined into a Voting Classifier C_{vote}. Training and testing is carried out using 5-fold cross-validation in order to be able to graph accuracy values using mean and standard deviation.

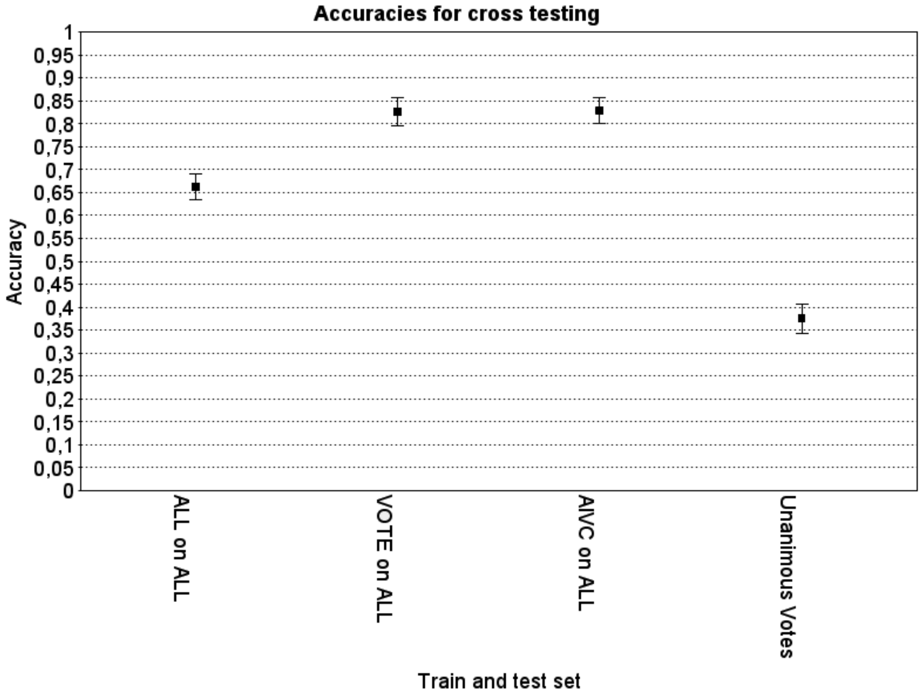


Figure 6.3: Results of the Voting Classifier for addressee annotations. "ALL on ALL" shows the overall accuracy of C_{ALL}. "VOTE on ALL" shows the accuracy of C_{vote} obtained on subset U. "AIVC on ALL" shows the accuracy of C_{ALL} obtained on subset U. "Unanimous votes" shows the number of unanimous votes from C_{vote}, that is, the size of subset U.

Figure 6.2 shows the performance of the annotator-specific classifiers on each of the three test sets. Clearly, each classifier C_X performs best on the test set TST_X. Figure 6.3 shows the performance of the classifier C_{vote} calculated on subset U in which the three annotator-specific classifiers reach a unanimous judgement, compared to the performance of the classifier C_{ALL}, tested on all data TST_{ALL}. Note that in this experiment, accuracy of the classifier C_{vote} is calculated on *only the subset U of test instances where the three expert classifiers give a unanimous judgement*. About 40 percent of the judgements of the experts were unanimous. In this subset U (40% of TST_{ALL}), C_{vote} obtains an accuracy of 83%, in contrast to the 66% obtained by C_{ALL} on the full set TST_{ALL}. This shows that the effects found for yeah-utterances, presented in Section 6.1, can be repeated for addressee data.

6.2.2 Performance Improvement or Context Selection?

The Voting Classifier returns a class label in only a subset $U \subset TST_{ALL}$ of 40% of the test instances. Within this subset it achieves an accuracy that is almost 20%

higher than the accuracy of the standard classifier C_{ALL} tested on TST_{ALL}. This does not necessarily mean that the accuracy of C_{vote} within subset U is higher than the accuracy of C_{ALL} within subset U of the test data. To put it differently, it is not yet clear from this difference whether C_{vote} achieves a higher performance within the limited subset U or whether C_{vote} serves as a *context selection mechanism* for the cases in which annotators would agree more as well — the latter also leading to a higher performance in subset U.

Figure 6.3 additionally shows the performance of C_{ALL} as calculated only on the subset $U \subset TST_{ALL}$ where C_{vote} returns a class label (the result marked "AIVC on ALL"). It turns out that *within subset* U both classifiers achieve comparable accuracy. For the addressee annotations at least, C_{vote} apparently serves as a context selection mechanism. Whether this is specific to the data set that is used or whether this is principally true for any such Voting Classifier is an open question that remains for future investigations. Note, though, that this does not affect the answer to the question whether the Voting Classifier C_{vote} models intersubjectivity — just the question *how well* it models it.

6.2.3 Precision and Recall

In this subsection, precision and recall within subset U are examined in more detail for the experiments described above. This is done in terms of precision and recall on G-addressed and I-addressed labels, because the difference between G-addressed and I-addressed is an important distinction in this data set with respect to the differences between annotators (see Section 5.1).

Figure 6.4 represents the results of the same experiments discussed in Section 6.2.1, but displayed differently. The graph should be read as follows. "PrecI" is an aggregate precision measure indicating the precision of a classifier on I-addressed utterances, that is, the degree to which the classifier correctly labeled an utterance with one of the labels A, B, C or D, relative to the total number of (any) individual labels assigned by the classifier. "RecI" indicates the recall on I-addressed utterances: the number of utterances correctly labeled A, B, C or D by the classifier, relative to the number of utterances that should have been labeled as such according to the test data. "PrecG" and "RecG" indicate the same, for the addressee label GROUP. The results marked "ALL on ALL" concern classifier C_{ALL} tested on TST_{ALL}. The results marked "VOTE on ALL" concern classifier C_{vote}, tested on subset $U \subset TST_{ALL}$. The precision and recall of C_{ALL} tested on subset U is not shown, as these values are similar to those for C_{vote} (see also Section 6.2.2). It should be noted that in the voting context defined by subset U, G-addressed utterances make up the same amount of data as in the whole test set (about 65%).

For the I-addressed utterances — the utterances with addressee label A, B, C or D — C_{vote} shows a strong increase of 20% in precision, but no increase in recall, compared to C_{ALL}. This means that C_{vote} labeled (relatively) less utterances with one of the labels A, B, C or D, but did so with higher precision. This fits very well with the observations on inter-annotator agreement for addressee, given in Section 5.1.2, which said that the annotators, *if* they agreed on an utterance being

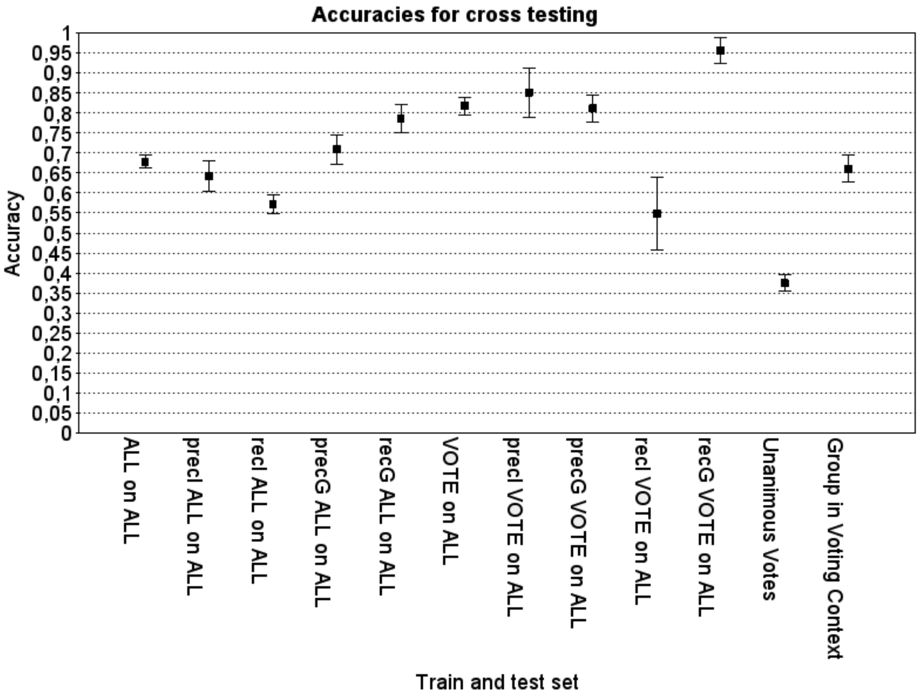


Figure 6.4: Precision and recall results of the voting classifier C_{vote} for addressee annotations.

I-addressed, could agree very well on *which* individual it was. This result, then at least does not contradict the idea that C_{vote} models the intersubjectivity between annotators.

6.2.4 Ensemble Learning

The Voting Classifier method presented in this chapter works by first training separate classifiers on subsets of the data and then combining their results to obtain a better classification. The same technique also underlies the *ensemble learning* approach. This subsection concerns the relation between the two methods. First, some detail is given on ensemble learning. Then, the question is phrased whether the results presented in this chapter can be explained without assuming that C_{vote} models intersubjectivity, using principles from ensemble learning. Finally, two experiments are presented that were carried out as attempts to do just this. Neither experiment could prove successfully that C_{vote} does *not* model intersubjectivity.

In ensemble learning, the goal is not to model (inter)subjectivity, but to have several classifiers learn different parts of the truth, hoping that the combination of the classifiers represents the overall truth of the classification problem better. As Dietterich [2000] discusses in his survey of ensemble learning, the separate classifiers must make different mistakes in order for the approach to work. When many classifiers make really different types of errors, in different situations, it becomes possible for the errors of one classifier to be compensated for by the judgements from all the other classifiers. There are several ways to build classifiers that make radically different errors, thus modeling different aspects of the truth, among which: (1) use the same training data but use completely different machine-learning methods for each classifier; (2) use 'unstable' classifiers or training methods that do not converge but rather end up in different local minima; (3) train the different classifiers on subsets of the data of which you suspect that they cover different aspects of the truth. The subjective Voting Classifiers described in this chapter are a clear example of the last: each subset of the data on which one of the annotator-specific expert classifiers was trained covers a different part of the common sense truth.

At this point the question arises whether the Voting Classifiers presented above have a higher performance in subset U because the training subsets mirror the mental conceptions of different annotators, or because *any* two subsets of the full training data would cover not completely overlapping parts of the overall truth. In order to answer this question two additional experiments are performed, in which Voting Classifiers are built on subsets of the training data divided on other properties than the annotator. If the mental conceptions of the annotators really make a difference, the annotator-specific Voting Classifier must show a stronger performance gain in subset U than the other Voting Classifiers.

Random Subsets

The first experiment is intended to find out whether it makes a difference that the three 'expert classifiers' that cast the votes for C_{vote} were trained on annotator-specific data. If the annotator-specificity does not introduce extra systematic differences between the training sets — so *any* three subsets of the full training data would cover 'not completely overlapping parts of the overall truth' to the same extent — then the performance gain observed in Section 6.2 would also be obtained by a Voting Classifier $C_{voteRnd}$ constructed of three expert classifiers each trained on *random* parts of the data.

Figure 6.5 shows the results obtained for $C_{voteRnd}$ constructed of three experts trained on the (not annotator-specific) subsets TRN_{rnd1}, TRN_{rnd2}, and TRN_{rnd3}. Each annotator is equally represented in these three training subsets, which are used to train the classifiers C_{rnd1}, C_{rnd2}, and C_{rnd3}. A comparison of the results in Figure 6.5 with the original results in Figure 6.3 shows that the annotator-specific classifier C_{vote} achieved twice as much performance gain as the classifier $C_{voteRnd}$. Concluding, the following can be said. Firstly, there is indeed an ensemble learning effect for $C_{voteRnd}$. Secondly, this effect is much smaller than for C_{vote}. Apparently there are systematic differences between the annotator-specific training sets that contribute a large portion of the performance gain.

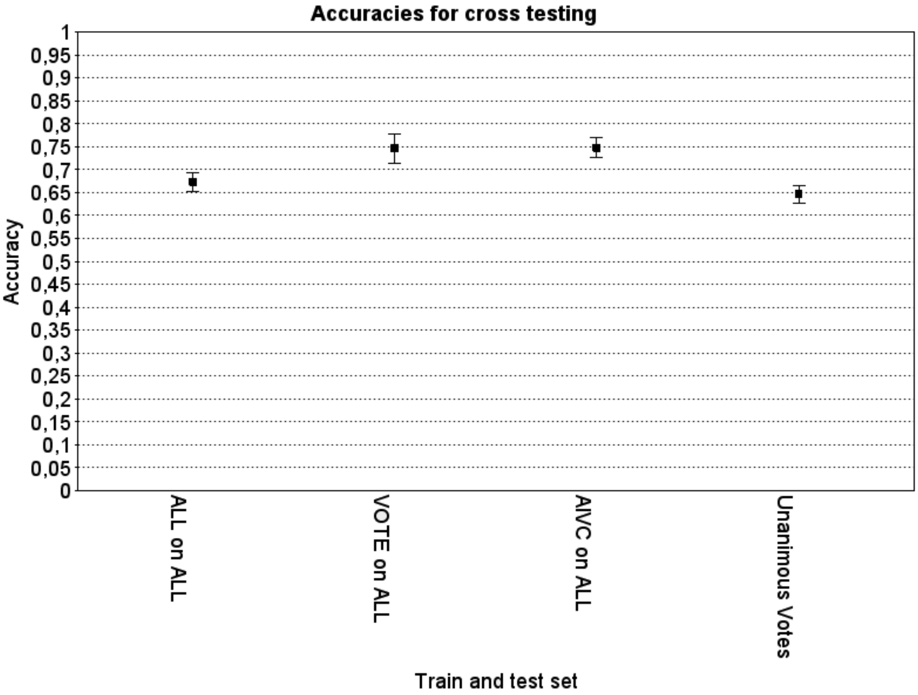


Figure 6.5: Results of the voting classifier $C_{voteRnd}$ for addressee, compared to the classifier C_{ALL} trained on all training data, for a voter constructed of three experts (C_{rnd1}, C_{rnd2}, and C_{rnd3}) trained on mixed data instead of annotator-specific data.

Different Meetings Instead of Different Annotators

The second experiment is intended to find out whether the systematic differences between the training sets that resulted in the performance gain were caused by the *annotator-specificity* of the training sets, or by a kind of meeting effect. The three annotators annotated non-overlapping parts of the corpus. This means that the 'annotator-specific' subsets are also 'meeting-specific' subsets. The different meetings are composed of different participants, and proceed therefore in different ways. Does this mean that the performance gain might as well be ascribed to the subsets of training data being meeting specific instead of their being annotator-specific? In order to answer this question, the experiments were repeated on different equally sized subdivisions of addressee data all annotated by the same annotator. The data from VKAR — who annotated many more meetings with addressee than the other annotators — was subdivided in two ways: randomly, leading to the voting classifier $C_{voteVkarRnd}$, and in *meeting-specific* subsets, leading to the voting classifier $C_{voteVkarMeeting}$. If the effect shown in Section 6.2.1 was caused by the meeting-specificity, rather than by the annotator-specificity, then two things would follow.

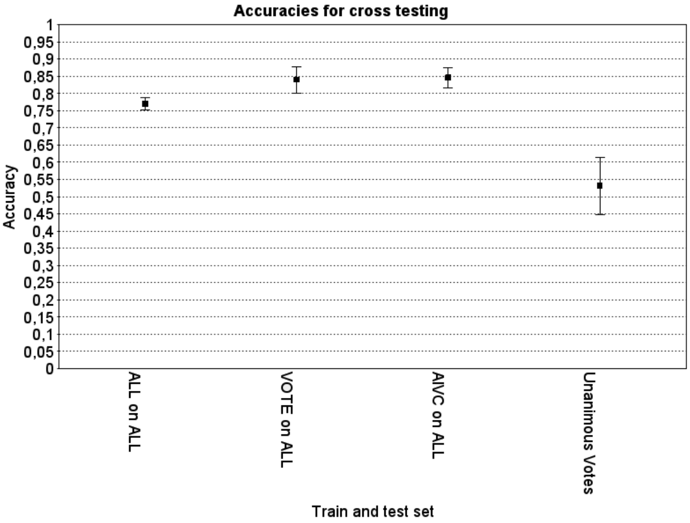
Firstly, $C_{voteVkarMeeting}$ would be expected to exhibit a performance gain equivalent to that achieved by the original annotator-specific classifier C_{vote}. Secondly, the performance gain of $C_{voteVkarRnd}$ and $C_{voteVkarMeeting}$ should exhibit the same differences as that of $C_{voteRnd}$ and C_{vote} shown in the previous section.

Figures 6.6(a) and 6.6(b) show the results for this second experiment. The gain achieved with a meeting-specific voting classifier is no better than the gain achieved with the voting classifier that is not meeting-specific. Apparently the performance gain achieved for the annotator-specific voting classifier can be ascribed to the differences between annotators instead of the differences between meetings, supporting the idea that the performance gain is in some way related to the overlap between the mental conceptions of the annotators.

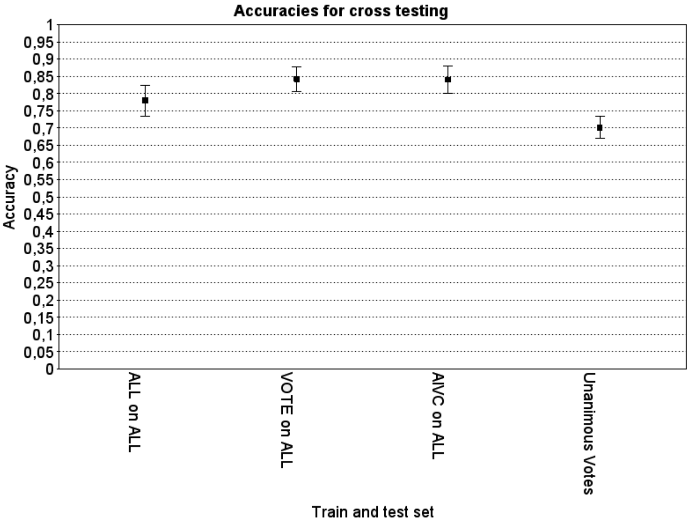
6.3 More Questions

The material presented in this chapter concerns a Voting Classifier based on the following theoretical principles: (1) annotations produced by annotators in a subjective task are a reflection of their personal “mental conceptions of a construct,” (2) those mental conceptions can be modeled using machine learning, and (3) agreement between the *learned classification models* may be indicative of situations where the annotators would have agreed, too. This way, the Voting Classifier can be seen as modeling the intersubjectivity, or ‘common sense’ shared between annotators. When such a classifier is deployed in a practical application, it will render judgements with which the end users of the application would agree more easily, too — assuming that the end users of the application have the same common sense.

The results presented above for Voting Classifiers are, if not conclusive, at least suggestive of supporting the theoretical ideas on which they are based. Section 6.2 explored some questions about how and why the Voting Classifier works. Although this resulted in no firm definitive conclusion regarding the hypothesis that the Voting Classifier is an embodiment of the intersubjectivity in the annotators’ mental conceptions, at least some objections to the hypothesis were countered. The results cannot be explained as ‘just’ ensemble learning, nor as a meeting-specific rather than annotator-specific effect, and the results generalise to at least one other data set. However, each attempt to answer a question leads to several new open questions. It is clear that there is still a lot of work to do in investigating Voting Classifiers and the relation between machine learning, subjectivity and inter-annotator agreement. In the remainder of this section a number of issues and open questions will be briefly touched upon. Some concern the underlying fundamentals of intersubjective voting classifiers and others concern the question as to how the results can be improved. Extensive experimentation with a really large data set of multiply annotated data would give the opportunity to address a number of these questions. This remains for future work.



(a) Results of the meeting-specific expert voting classifier $C_{voteVkarMeeting}$ for addressee annotations trained on meeting-specific data annotated by VKAR.



(b) Results of the voting classifier $C_{voteVkarRnd}$ for addressee annotations trained on *mixed* data from different meetings annotated by VKAR.

Figure 6.6: Results using meeting-specific data instead of annotator-specific data. The meeting-specific voting classifier $C_{voteVkarMeeting}$ performs similar to the non meeting-specific classifier $C_{voteVkarRnd}$, in contrast to the differences between C_{vote} and $C_{voteRnd}$ shown in Figures 6.3 and 6.5.

6.3.1 Questions About Improving the Results

Although the main goal of this chapter was to show that (inter)subjectivity in annotations could be modeled explicitly, and not to achieve the highest performance, the Voting Classifier allows for several obvious ways to improve classification results, related to its fundamental architecture, that have not been explored in this chapter.

Are there other (better) voting mechanisms for the Voting Classifier? The experiments in this chapter used a unanimous voting procedure for the voting classifiers. Other ways to do the voting might include majority voting or Bayesian averaging. A larger data set, annotated by more than three annotators, would be needed for systematic experimentation with such mechanisms.

How can the performance of the single annotator-specific classifiers be improved? The annotator-specific classifiers are intended to capture the personal view of a single annotator. This could be done better by tuning the classifiers more strongly to that single annotator's data, for example by doing separate feature selection for it. This in turn might conceivably lead to higher performance of the Voting Classifier within the subset of instances that it will venture an opinion about.

6.3.2 Questions About Underlying Fundamentals

Does the Voting Classifier actually model intersubjectivity? This remains the main open question. The results are suggestive, not conclusive. It might be possible to test this further: if true, the voting classifier would return a judgement on multiply annotated data in exactly those cases in which the annotators agreed, too. Note that not *all* data needs to be annotated by more annotators: just enough to test the hypothesis. Otherwise, it will suffice to have enough data for each single annotator, be it overlapping or not. This is especially advantageous when the corpus is really large, such as the 100h AMI corpus. Another way to test the hypothesis that the voting behavior relates to intersubjectivity is to look at the type and context of the points of agreement between annotators, found in the reliability analysis, and see if that relates to the type and context of the cases where the Voting Classifier renders a unanimous judgement. That would be strong circumstantial evidence in support of the hypothesis. The FOA context seems promising to use, given that the inter-annotator agreement for addressee data had such clear contextual dependencies. In this particular case, however, it is not such a good idea because the contextual feature (FOA) was also used as one of the most important training features for the classifier. Finding the most agreement between voters in a certain FOA context would point to FOA indeed being a strong feature, rather than unequivocally to the classifier embodying intersubjectivity.

What is the relation between the voting classifier and confidence metrics? The context selection behavior of the voting classifier can also be seen as a binary confidence metric. Is this 'analytical confidence metric' a kind of optimal estimation of a confidence metric that can also be approached by directly learning it? This might be something that can be figured out using another simulation approach, as was done in Chapter 3.

How do intersubjectivity for annotators and generalization to end users relate? Throughout this chapter it is mentioned several times that "the cases in which annotators would agree" more or less equal "the cases in which the judgement also generalizes well to the way end users of applications assess their own and others' (inter)actions." This position is based on the idea of "shared common sense," and the assumption that data has been annotated using "naive" annotators representative of the average end consumer of the data or end user of applications. It could be tested more rigorously using an experiment in which data is first annotated multiple times, and next offered to representative end users in an assessment procedure where they have to indicate whether they agree with the offered judgements or not.

What is the relation between the used features and classifier types, and whether the performance gain is observed? It is reasonable to assume that the performance gain that was achieved with the intersubjective voting classifier depends on the annotator-specific classifiers being able to pick up the idiosyncrasies of the individual annotators. This in turn will probably depend on things such as the selected features and modalities (see also the discussions in Chapter 5) and the used machine-learning methods.

How to distinguish disagreement that indicates subjectivity from the disagreement that indicates errors? In the introduction of this thesis it is mentioned that not all disagreement may be caused by errors: sometimes disagreement is simply an indication of differing mental conceptions. This does not mean, however, that errors are no longer an important problem in annotations as soon as one starts to look at subjective content! If an annotator misunderstands a label, the resulting disagreement may be systematic, but his annotations are still simply wrong, and not 'subjective'. The fact that a machine classifier is able to 'recognize something' is not enough. "[Such a position] could be dangerous if taken as recommendation, since it would legitimize the detection of patterns that have little to do with any reality that can be observed by humans" [Artstein 2008, personal communication]. However, currently there is no clear way to distinguish the two types of disagreement. For this, the field needs to develop new methods.

6.3.3 Final Thoughts

The final conclusions that can be drawn from the results and questions discussed in this chapter are the following.

(1) Explicitly modeling subjectivity and intersubjectivity of the annotations produced by different annotators is an interesting direction that certainly deserves more future attention.

(2) There are numerous open questions regarding this approach, of which the most important one may be how to measure reliability for subjective annotations.

(3) For all of the issues and directions discussed in this chapter, provenance data is very important. Projects that perform annotation on a large scale need to consider storing (anonymized) information about the annotators who produced the separate sections of the annotated data. This requires a major change in working procedures in the field.

Part III

Reflection

Chapter 7

Designing for Interaction

In this chapter a somewhat broader view on corpus based research is discussed. The previous parts of the thesis concerned the annotation process, inter-annotator agreement, the quality of corpora and the training of automatic classifiers for the annotated content. This chapter focuses on the relation between classifiers for projective latent content and the interactive systems in which they are to be used. First, a few different applications of automatic recognition modules are discussed. After that, an analysis of the role of classifiers for projective latent content in the different types of application leads to a final reflection on subjective machines.

7.1 Application Contexts for Recognition Modules

Within projects such as AMI many recognition technologies are developed. Some of these concern manifest content and others projective latent content. There are many ways to apply these technologies in HCI, both inside the meeting domain and elsewhere. An important distinguishing dimension for such applications is the relation between the person(s) whose behavior is being recognized, and the person(s) who are the end consumers of the recognition results.

For some of these applications the person whose (communicative) behavior is recognized is also the end user of the system. Recognition of manifest content is used to control a computer through speech and gestures. Recognition of projective latent content such as emotion and attitude can be used in tutoring situations to determine the best student feedback or in entertainment systems to ascertain that the user is still enjoying himself.

It is also possible that the end user is not necessarily one of the people whose behavior was recognized. For example, an automatic meeting summarizer may also present the (interpretation of) the recognized behavior to people other than the original participants. The summaries are useful as minutes for the participants, but managers, colleagues who were not there and other people may have access to them, too. The summaries can contain recognized speech (manifest content), but also recognized decision points, or information about who did or did not agree with a decision (projective latent content).

Finally, in some applications the intended end consumer of the recognition result is clearly and unequivocally *not* the person whose behavior is being recognized. For automatic call centers, there is a lot of work on automatic recognition of certain emotions of the customer based on telephone speech, to allow intervention from a human operator when a user gets frustrated and angry at the automatic response system [Petrushin, 2000; Devillers et al., 2002; Morrison et al., 2007; Gupta and Rajput, 2007]. In a distributed communication environment, one can present not only the audio and/or video to the communication partners, but also additional information such as mood. In some games and chat environments people can display their emotions through an avatar; usually this is done with explicit commands, but it could be an interesting extension to do this based on automatic recognition. As a last example, in the AMI project a demonstration setup is developed in which communication in a distributed meeting is enhanced with a remote participant in an otherwise co-located meeting is represented by an avatar. The expressions of the avatar are to be controlled through a mixture of user commands and automatic recognition of the behavior of the remote participant. The use scenario is a situation in which the remote participant cannot use a full teleconferencing setup because he is otherwise occupied, for example driving a car. The goal is to enhance the presence of the remote participant in the meeting and to increase the level of participation by the remote participant.

7.2 Requirements for Subjectively Annotated Data

The different application contexts described in the previous section place different demands on the type of subjectivity that needs to be modeled in classifiers, and therefore on the type of data that the classifiers are trained on. Below, three possible requirements — that do not necessarily go together very well — are described: a requirement for *agreeable judgements*, a requirement for *consistent and cohesive judgements*, and a requirement for *user adapted judgements*.

For some systems it is very important that the recognition modules will only render judgements that all end users would agree to. To build such a system, one needs annotated data from many different people, to make sure that the full range of human variability in interpretation is covered. These annotations can then be used as in Chapter 6 to build classifiers that focus on the overlap between the mental conceptions of all these annotators.

For other systems, it may be more important that the recognition judgements made by the system simulate a believably complete ‘mental conception’, and are as consistent and cohesive as if they come from one real human. For a Virtual Human in a tutoring system, for example, it is more important that the working of the recognition modules suggests a consistent personality than that it is never controversial in its judgements. Real human teachers are not expected to only make judgements of the behavior of the pupils to which everybody would agree, either. But if the different interpretations made by the system do not form a coherent whole, the system will appear to be whimsical in its perceptions and interpretations of the interaction partner. To obtain consistent and cohesive judgements from a classifier, it probably needs to be trained on annotated data from only one annotator, to make sure that the corpus contains a balanced reflection of the mental conceptions of one person.

Yet other applications may need classification modules that are perfectly adapted to the behavior of the specific user whose behavior is being recognized. For example, the remote meeting participation demo, in which the remote participant is represented through an avatar, needs to be adapted to interpreting each particular remote participant correctly. Otherwise the other participants in the meeting will get entirely the wrong image of the opinions and contributions of the remote participant.

7.3 Two New Types of Classifiers

Many annotation tasks require judgements with a large measure of subjectivity. When this is simply taken as a given, and the systematic disagreement resulting from the different mental conceptions of the annotators is not taken into account while training a machine classifier on the resulting data, there is no simple reason to assume that the resulting classifier is any less idiosyncratic in the judgements it makes. Without additional analyses one cannot suppose the classifier did not pick up those idiosyncrasies from the annotators. Indeed, in the previous section we saw that this may be a desirable feature of a classifier. Models may be trained with the goal of having them mirror the mental conceptions of one person. A judgement made by such a classifier should be approached in a similar manner as a judgement made by another person. Such classifiers can therefore be called *'subjective entity'* classifiers.

A careful analysis of the inter-annotator (dis)agreement makes it possible to build classifiers that (partly) embody the intersubjective overlap between the mental conceptions of the annotators. Because the classifier is designed only to give a judgement in situations where one can expect annotators or users to agree, one can, despite the subjective quality of the annotation task, approach the judgements made by the classifier as a kind of common sense truth ('this is more or less what people can agree on'). Such classifiers can be called *'consensus objective'* classifiers.

7.4 Subjective Machines?

The two types of classifiers introduced in the previous section are fundamentally not the same as manifest content classifiers that detect persons in video streams, fighter planes in radar images, or characters in hand written text. The process of collecting data, annotating the data, and learning from the annotations to automatically classify new data is superficially not so much different. Nevertheless, subjective entity classifiers and consensus objective classifiers, which both operate on projective latent content, yield classifications that have a quite different meaning than those of manifest content classifiers.

The manifest content classifiers are like a calculator. When a computer performs a calculation for you, you expect it to be fast and precise. Some calculations may be complex and yield approximations rather than an exact answer, but in general the answers can be trusted. You know how to use the answers as facts to form an opinion or make a decision.

The projective latent content classifiers are not like a calculator. The results of their classifications should be approached in a fundamentally different way. To see this, think about the following. A person who attended a meeting that you missed tells you what happened during the meeting. He tells you what the general mood in the meeting was, which participants had a conflict, and he tells about the decisions that were most hotly disputed. When you form your own ideas about that meeting, based on his story, you will take into account that the information was given to you by another person, with his own preconceptions and views. You will approach part of his story not as fact, but as opinions and interpretations. Subjective entity classifiers are not persons. Nevertheless, their classifications should be approached not as facts, but rather as opinions and interpretations, due to the way they were trained. And even consensus objective classifiers, the classifications of which are supposed to only concern 'that what people can agree on' yield interpretations rather than facts — although, philosophically speaking, one can wonder what the distinction is between a fact and an interpretation on which everybody agrees.

The question rises who is responsible for the interpretations that are given by the classifiers. Is it the annotator who produced the training data? Is it the designer who build a program around the classifier? Or should the machine be redefined as a 'subjective actor' to replace the 'machine as calculator'? Looking for answers to these questions is well outside the scope of this thesis. This chapter will merely end with pointing out that systems that use projective latent content classifiers to interpret the behavior of humans should be very clear towards the user with respect to the possible subjective status of the information being presented. Maybe information drawn from such classifiers should be presented to a user as the opinions of a Virtual Human, as recently suggested by Reidsma et al. [2007].

Chapter 8

Conclusions

In this chapter the results and conclusions from this thesis are summarized, in the light of the research questions that were formulated in Section 1.2. The main research question was as follows.

Main RQ — What are the relations between inter-annotator agreement, subjective judgements in annotation, and whether a corpus is fit for the purpose for which it was constructed?

This question was addressed in detail using three more concrete research questions that are discussed below. First, it is necessary to understand how people have in the past been looking at these respective relationships. The first part of the thesis started in Chapter 2 with a discussion of literature concerning reliability analysis and corpus based research, with a focus on the themes that were considered most relevant to the thesis. The use of inter-annotator agreement as a metric for determining the quality and reliability of an annotated corpus was explained. Next, related work was summarized on two additional steps in the analysis of inter-annotator: how and why disagreement occurs in certain annotation tasks and what the consequences of such disagreement is for the use that can be made of the resulting corpus. Although the last two topics are somewhat under represented in the literature, several practical examples were presented. The literature overview furthermore contained a discussion of the role of subjective judgements in corpus annotation. The concept of *projective latent content* from the field of Content Analysis was presented, arguing that there is much work in computational linguistics and corpus based computer science to which this concept applies. It was remarked that for projective latent content it is more likely that inter-annotator disagreement is systematic than for other types of content.

After this introductory chapter, the first of the three concrete research questions was taken up.

RQ 1 — What is the relevance of placing a threshold on the level of inter-annotator agreement for assessing the reliability of a corpus, especially if the errors that caused reduced agreement may not have been homogeneous?

To answer this question, a series of simulation experiments has been performed. The results of the experiments challenge one of the common practices in corpus based research, namely that of equating "reliability assessment" with "calculating an agreement metric". In this practice researchers consider the task of reliability assessment to be finished, and the quality of their data to be tolerable, as soon as they have been able to show that κ or α exceeds a certain threshold value. The experimental results presented in Chapter 3 showed that not just the *amount* but also the *type* of annotators' errors that caused a reduced inter-annotator agreement can have large consequences for whether the data is fit for purpose, especially if this purpose is machine learning. If the errors appear to be more or less random then the amount of information in the annotations goes down, limiting the performance of the machine classifiers. But since machine-learning algorithms are designed specifically to look for, and predict, patterns in noisy data, more data will yield more signal and the learner will ignore the noise. In theory, this makes random disagreement relatively unimportant. However, if the errors are systematic in some way, they introduce patterns in the data. Machine classifiers are designed to pick up patterns, so they will learn to emulate these errors. Because the same errors may also be present in the test data it is possible that a performance analysis of the resulting machine classifiers will not reveal this weakness in the learned models. Simulating the annotation process, the experiments demonstrated that machine learning can tolerate data with a low reliability measurement as long as the disagreement looks like random noise, and that when it does not, data can have a reliability measure commonly held to be acceptable but produce misleading results. This means that the value of reliability metrics, compared to the usual threshold of 0.8, is not readily interpretable. It was concluded that although inter-annotator agreement analysis can be a powerful tool for assessing the quality of an annotated corpus, this is only so if it includes an analysis of the type of errors made by annotators as well as of the potential impact of those particular errors on the use that is made of the data.

The second part of the thesis contains several explorations, on data from the AMI Corpus, of the relation between inter-annotator agreement, data quality, machine learning and subjectivity. Chapter 5 addressed research question two.

RQ 2 — Given annotation with a low inter-annotator agreement, how can one pinpoint more reliable subsets of the annotated data for which a higher agreement was achieved?

It was shown that it is possible to take data which was annotated with a relatively low level of inter-annotator agreement, and use the results of an extended agreement analysis to determine certain 'more reliable subsets' in which the annotated data has a higher quality. FOA context III ('at least some person in speaker's FOA during utterance') defines one such more reliable subset on the addressee annotations. One can then train and test a classifier on only the more reliable subset of the data, which probably leads to a better performance. If the context defining the more reliable subset can be derived for new, unseen data, too, this allows one to have the trained classifier only return a class label for those instances for which it is qualified to make a judgement.

The third and last concrete research question addressed the subjective aspect of annotator disagreement.

RQ 3 — Is it possible to find out how subjective the annotations are, and to model the subjectivity explicitly as it relates to the overlap and disjunctions between the personal points of view of the annotators, using machine-learning methods?

A series of experiments has been performed on data that was annotated with a certain amount of subjectivity. In Chapter 6 it was shown how it is possible to take this subjectivity in annotation into account during the machine-learning process. Ensemble classifiers can be constructed that are designed to make judgements in a certain subset of instances in such a way that they reflect the overlap in the mental conceptions of the annotators who produced the annotations. This allows one to separate the intersubjectivity in the classifiers' judgement from the idiosyncrasies that were picked up from the annotations of one particular annotator.

Based on these results, two new concepts are defined in Chapter 7. Subjective annotations can be used to build a *subjective entity classifier*. Such a classifier should classify in a way with which users would not necessarily always agree but which can be seen as reflecting a consistent subjective judgement. A subjective entity classifier can either be conceived of as implementing the mental conception of one specific annotator, or as embodying a generic 'average user' mental conception. Subjective annotations can also be used to build a *consensus objective classifier*. Such a classifier ideally only renders a judgement for instances where all annotators as well as end users would agree with the judgement.

The most important contributions of this thesis to the field of corpus based research consist of four elements. Firstly, it presents an approach to the use of annotated data that takes the interaction between inter-annotator agreement analysis and the possible use of the annotated data into account, more so than is the usual practice in the field. Secondly, it also contains clear arguments — based on simulations — for *why* the usual practice is insufficient. Thirdly, the thesis presents new ways of looking at subjective judgements from annotators in the context of machine classification tasks, to show how the field can approach the increased presence of projective latent content in corpus based research, namely (1) deriving mode reliable subsets from data annotated with a low inter-annotator agreement and (2) explicitly modeling the (inter)subjectivity present in annotations. This particular contribution has resulted in many open issues and questions that can be a starting point for future work. Fourthly, there are several places in the thesis where a need for a change of practices in the field is pointed out. They can be summarized as follows.

- An analysis of how and why annotators disagree should be included as a necessary part of reliability analysis.

- An analysis of the impact of disagreement on the intended use of the corpus should be included as a necessary part of reliability analysis.
- The field needs to develop new methods for these analyses.
- Methods need to be developed to distinguish disagreement that stems from errors from disagreement that follows from the subjective nature of an annotation task.
- Information about which annotator produced which data in a corpus should be stored as an essential part of the corpus data.
- Systems that use projective latent content classifiers to interpret the behavior of humans should be very clear to the user with respect to the possible subjective status of the information being presented.

The final conclusion that can be drawn from this thesis is based on the observation that for every tentative answer to a question that was found more new questions were raised: *“This thesis is only a starting point.”*

Bibliography

op den Akker, H. and C. Schulz. Exploring features and classifiers for dialogue act segmentation. In van Leeuwen, D., A. Nijholt, A. Popescu-Belis, and R. Stiefelhagen, editors, *Proceedings of the MLMI*, September 2008.

op den Akker, H. J. A. and M. Theune. How do I address you? modelling addressing behavior based on an analysis of multi-modal corpora of conversational discourse. In *Proceedings of the AISB symposium on Multi-modal Output Generation, MOG'08*, pages 10–17, April 2008.

Al-Hames, M., T. Hain, J. Cernocky, S. Schreiber, M. Poel, R. Müller, S. Marcel, D. van Leeuwen, J.-M. Odobez, S. O. Ba, H. Bourlard, F. Cardinaux, D. Gatica-Perez, A. Janin, P. Motlicek, S. Reiter, S. Renals, J. van Rest, R. J. Rienks, G. Rigoll, K. Smith, A. Thean, and P. Zemcik. Audio-visual processing in meetings: Seven questions and current AMI answers. In *Proceedings of the 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, volume 4299 of *Lecture Notes in Computer Science*, pages 24–35, London, 2007. Springer.

Allwood, J., L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio. A coding scheme for the annotation of feedback, turn management and sequencing phenomena. In Martin, J.-C., P. Kühnlein, P. Paggio, R. Stiefelhagen, and F. Pianesi, editors, *Proc. of the LREC Workshop "Multimodal Corpora From Multimodal Behaviour Theories to Usable Models"*, pages 44–48. ELRA, ELRA, May 2006.

AMI Consortium. Deliverable D5.2, implementation and evaluation results, 2005a.

AMI Consortium. Guidelines for dialogue act and addressee. Technical report, 2005b.

Aron, A. and E. N. Aron. *Statistics for Psychology*. Prentice Hall, USA, 2003.

Artstein, R., G. Boleda, F. Keller, and S. Schulte im Walde, editors. *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*. Coling 2008 Organizing Committee, Manchester, UK, August 2008.

Artstein, R. and M. Poesio. Bias decreases in proportion to the number of annotators. In *Proceedings of FG-MoL 2005*, pages 141–150, August 2005.

Artstein, R. and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, ISSN 0891-2017, to appear.

Ba, S. O. and J.-M. Odobez. A study on visual focus of attention recognition from head pose in a meeting room. In Renals, S. and S. Bengio, editors, *Machine Learning for Multimodal Interaction Third International Workshop, MLMI 2006, Bethesda, MD, USA, May 1-4, 2006, Revised Selected Papers*, volume 4299 of *Lecture Notes in Computer Science*, pages 75–87. Springer, 2006.

Ba, S. O. and J.-M. Odobez. Head pose tracking and focus of attention recognition algorithms in meeting rooms. In *Multimodal Technologies for Perception of Humans, Proceedings of MLMI07*, volume 4122 of *LNCS*, pages 345–357, Berlin / Heidelberg, 2007. Springer.

Bakeman, R. and J. M. Gottman. *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press, 1986.

Bales, R. F. *Interaction Process Analysis; A Method for the Study of Small Groups*. Addison Wesley, Reading, Mass, 1950.

Batliner, A., C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russel, and M. Wong. "you stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proceedings of the 4th International Conference of Language Resources and Evaluation (LREC 2004)*, pages 171–174, 2004.

Bayerl, P. S. and K. I. Paul. Identifying sources of disagreement: Generalizability theory in manual annotation studies. *Computational Linguistics*, 33(1):3–8, ISSN 0891-2017, 2007.

Beigman Klebanov, B., E. Beigman, and D. Diermeier. Analyzing disagreements. In Artstein et al. [2008], pages 2–7.

Beigman Klebanov, B. and E. Shamir. Reader-based exploration of lexical cohesion. *Language Resources and Evaluation*, 40 (2):109–126, ISSN 1574-020X, May 2006.

Berelson, B. *Content analysis in communication research*. Free Press, Glencoe, Illinois, 1952.

Bhowmick, P. K., A. Basu, and P. Mitra. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In Artstein et al. [2008], pages 58–65.

Bos, P., D. Reidsma, Z. M. Ruttkay, and A. Nijholt. Interacting with a virtual conductor. In Harper, R., M. Rauterberg, and M. Combetto, editors, *Proc. of 5th International Conference on Entertainment Computing, Cambridge, UK*, number 4161 in Lecture Notes in Computer Science, pages 25–30. Springer Verlag, September 2006.

Carletta, J. C. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254, ISSN 0891-2017, 1996.

Carletta, J. C. Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, ISSN 1574-020X, May 2007.

Carletta, J. C., S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. M. Post, D. Reidsma, and P. Wellner. The AMI meeting corpus: A pre-announcement. In Renals, S. and S. Bengio, editors, *Machine Learning for Multimodal Interaction, Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers*, volume 3869 of *Lecture Notes in Computer Science*, pages 28–39. Springer, 2006.

Carletta, J. C., S. Evert, U. Heid, J. Kilgour, J. Robertson, and H. Voormann. The NITE XML toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments and Computers*, 35(3):353–363, ISSN 0743-3808, 2003.

Carletta, J. C., S. Isard, G. Doherty-Sneddon, A. Isard, J. C. Kowtko, and A. H. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31, ISSN 0891-2017, 1997.

Carletta, J. C., D. McKelvie, A. Isard, A. Mengel, M. Klein, and M. B. Møller. A generic approach to software support for linguistic annotation using xml. In Sampson, G. and D. McCarthy, editors, *Corpus Linguistics: Readings in a Widening Discipline*. Continuum International, London, 2005.

Carlson, L., D. Marcu, and M. E. Okurowski. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proc. of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

Ciceri, M. R., S. Balzarotti, F. Beverina, F. Manzoni, and L. Piccini. Meed: the challenge towards a multimodal ecological emotion database. In Devillers, L., J.-C. Martin, R. Cowie, and A. Batliner, editors, *Proc. of the LREC2006 Workshop on Corpora for Research on Emotion and Affect*, pages 29–33, May 2006.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, ISSN 0013-1644, 1960.

Cohen, J. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum, Hillsdale, NJ, 2 edition, 1988.

Craggs, R. and M. McGee Wood. A categorical annotation scheme for emotion in the linguistic content of dialogue. In André, E., L. Dybkjær, W. Minker, and P. Heisterkamp, editors, *Affective Dialogue Systems*, volume 3068 of *LNCS*, pages 89–100. Springer, 2004.

Craggs, R. and M. McGee Wood. Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 31(3): 289–296, ISSN 0891-2017, September 2005.

Devillers, L., I. Vasilescu, and L. Lamel. Annotation and detection of emotion in a task-oriented humanhuman dialog corpus. In *Proceedings of ISLE Workshop on Dialogue Tagging*, 2002.

Dhillon, R., S. Bhagat, H. Carvey, and E. Shriberg. Meeting recorder project: Dialog act labeling guide. Technical Report TR-04-002, International Computer Science Institute, Berkeley, USA, 2004.

Di Eugenio, B. and M. Glass. The kappa statistic: A second look. *Computational Linguistics*, 20(1):95–101, ISSN 0891-2017, March 2004.

Dielmann, A. and S. Renals. DBN based joint dialogue act recognition of multiparty meetings. In *Proc. IEEE ICASSP*, volume 4, pages 133–136, April 2007.

Dietterich, T. G. Ensemble methods in machine learning. In Kittler, J. and F. Roli, editors, *MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems*, volume 1857 of *LNCS*, pages 1–15, London, UK, 2000. Springer-Verlag.

Falcon, V., C. Leonardi, F. Pianesi, and M. Zancaro. Annotation of group behaviour: a proposal for a coding scheme. In *Proc. of Workshop on Multimodal Multiparty Multimodal Processing at ICMI 2005*, pages 39–46, 2005.

Fleiss, J. L. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 31(3):651–659, ISSN 0006-341X, 1975.

Galley, M., K. McKeown, J. Hirschberg, and E. Shriberg. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Meeting of the ACL*, pages 669–676, Morristown, NJ, USA, July 2004. Association for Computational Linguistics.

Gardner, R. Acknowledging strong ties between utterances in talk: Connections through right as a response token. In *Proceedings of the 2004 Conference of the Australian Linguistic Society*, pages 1–12, 2004.

Goffman, E. Footing. In *Forms of Talk*, pages 124–159. Philadelphia: University of Pennsylvania Press, 1981.

Greene, J. O. and J. N. Cappella. Cognition and talk: The relationship of semantic units of temporal patterns of fluency in spontaneous speech. *Language and Speech*, 29(2):141–157, ISSN 0023-8309, 1986.

Gupta, P. and N. Rajput. Two-stream emotion recognition for call center monitoring. In *Proceedings of INTERSPEECH-2007*, pages 2241–2244, 2007.

Heylen, D. and H. J. A. op den Akker. Computing backchannel distributions in multi-party conversations. In Cassell, J. and D. Heylen, editors, *Proceedings of the ACL Workshop on Embodied Language Processing, Prague*, volume W07-19, pages 17–24, Prague, Czech Republic, June 2007. Association of Computational Linguistics.

Hsueh, P.-Y. and J. D. Moore. Automatic decision detection in meeting speech. In Popescu-Belis, A., S. Renals, and H. Bourlard, editors, *Machine Learning for Multimodal Interaction IV*, volume 4892 of *Lecture Notes in Computer Science*. Springer, 2007.

Hung, H. and D. Gatica-Perez. Identifying dominant people in meetings from audio-visual sensors. In *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG), Special Session on Multi-Sensor HCI for Smart Environments*, September 2008.

Jovanović, N. *To Whom It May Concern - Addressee Identification in Face-to-Face Meetings*. Phd thesis, University of Twente, 2007.

Jovanović, N., H. J. A. op den Akker, and A. Nijholt. A corpus for studying addressing behavior in multi-party dialogues. In *Proceedings 6th SIGdial Workshop on Discourse and Dialogue*, pages 107–116, Lisbon, Portugal, 2005.

Jovanović, N., H. J. A. op den Akker, and A. Nijholt. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation*, 40(1):5–23, ISSN 1574-020X, February 2006.

Kita, S., I. v. Gijn, and H. vd. Hulst. Movement phases in signs and co-speech gestures, and their transcription by human coders. In Wachsmuth, I. and M. Frühlich, editors, *Gesture and Sign Language in Human-Computer Interaction, Proceedings of the International Gesture Workshop Bielefeld*, volume 1371 of *LNCS*, pages 23–35, Berlin, 1998. Springer Verlag.

Kleinbauer, T., S. Becker, and T. Becker. Indicative abstractive summaries of meetings. In *Proceedings of MLMI 2007*, 2007.

Krippendorff, K. *Content Analysis: An Introduction to its Methodology*, volume 5 of *The Sage CommText Series*. Sage Publications, Beverly Hills, London, 1980.

Krippendorff, K. On the reliability of unitizing continuous data. *Sociological Methodology*, 25:47–76, ISSN 0081-1750, 1995.

Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Inc, 2 edition, 2004a.

Krippendorff, K. Reliability in content analysis. some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433, ISSN 0360-3989, 2004b.

Laskowski, K. and S. Burger. Development of an annotations scheme for emotionally relevant behavior in multiparty meeting speech. In *Interspeech (to appear)*, 2005.

Lazarsfeld, P. Latent structure analysis. In Stouffer, S. A., L. Guttmann, E. Suchman, P. Lazarsfeld, S. Star, and J. Claussen, editors, *Measurement and Prediction*. Wiley, New York, 1966.

TER MAAT, M., R. M. EBBERS, D. REIDSMA, and A. NIJHOLT. Beyond the beat: Modelling intentions in a virtual conductor. In *INTETAIN '08: Proceedings of the 2nd international conference on INtelligent TEchnologies for interactive enterTAINment*, pages 1–10. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), January 2008.

MARCU, D., E. AMORRORTU, and M. ROMERA. Experiments in constructing a corpus of discourse trees. In WALKER, M, editor, *Towards Standards and Tools for Discourse Tagging: Proceedings of the Workshop*, pages 48–57. Association for Computational Linguistics, Somerset, New Jersey, 1999.

MARTELL, C. and J. KROLL. Using FORM gesture data to predict phase labels. In MARTIN, J.-C., P. KÜHNLEIN, P. PAGGIO, R. STIEFELHAGEN, and F. PIANESI, editors, *Proc. of the LREC Workshop "Multimodal Corpora From Multimodal Behaviour Theories to Usable Models"*, pages 29–32. ELRA, ELRA, May 2006.

MORAN, T. P., L. PALEN, S. HARRISON, P. CHIU, D. KIMBER, S. MINNEMAN, W. VAN MELLE, and P. ZELLWEGER. I'll get that off the audio: a case study of salvaging multimedia meeting records. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 202–209. ACM Press, 1997.

MORRISON, D., R. WANG, and L. C. DE SILVA. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112, ISSN 0167-6393, February 2007.

MÜLLER, R., B. SCHULLER, and G. RIGOLL. Enhanced robustness in speech emotion recognition combining acoustic and semantic analyses. In *From Signals To Signs of Emotion and Vice Versa*, 2004.

OTSUKA, K., J. YAMATO, Y. TAKEMAE, and H. MURASE. Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns. In *CHI '06 extended abstracts on Human factors in computing systems*, pages 1175–1180, New York, NY, USA, 2006. ACM.

PAIVA, A., R. PRADA, and R. W. PICARD, editors. *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, volume 4738 of *Lecture Notes in Computer Science*. Springer, 2007.

PASSONNEAU, R. J. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth Internation al Conference on Language Resources and Evaluation (LREC)*, 2006.

PASSONNEAU, R. J., N. HABASH, and O. RAMBOW. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, 2006.

PASSONNEAU, R. J., T. YANO, T. LIPPINCOTT, and J. KLAVANS. Relation between agreement measures on human labeling and machine learning performance: Results from an art history image indexing domain. In *Proceedings of the LREC 2008*, 2008.

PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

PELACHAUD, C., J.-C. MARTIN, E. ANDRÉ, G. CHOLLET, K. KARPOUZIS, and D. PELÉ, editors. *Intelligent Virtual Agents, 7th International Conference, IVA 2007, Paris, France, September 17-19, 2007, Proceedings*, volume 4722 of *Lecture Notes in Computer Science*. Springer, 2007.

PETRUSHIN, V. A. Emotion recognition in speech signal: experimental study, development, and application. In *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 222–225, 2000.

POESIO, M. and R. ARTSTEIN. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

POOLE, M. S. and J. P. FOLGER. Modes of observation and the validation of interaction analysis schemes. *Small Group Behavior*, 12(4):477–493, ISSN 0090-5526, 1981.

POST, W. M., A. H. M. CREMERS, and O. A. BLANSON HENKEMANS. A research environment for meeting behavior. In *Proceedings of the 3rd Workshop on Social Intelligence Design*, pages 159–165, 2004.

POTTER, J. W. and D. LEVINE-DONNERSTEIN. Rethinking validity and reliability in content analysis. *Journal of applied communication research*, 27(3):258–284, ISSN 0090-9882, 1999.

QUEK, F., F. T. ROSE, and D. MCNEILL. Multimodal meeting analysis. In *Proceedings of the international conference on intelligence analysis*, May 2005.

REIDSMA, D. and H. J. A. OP DEN AKKER. Exploiting 'subjective' annotations. In Artstein et al. [2008], pages 8–16.

REIDSMA, D. and J. C. CARLETTA. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326, ISSN 0891-2017, September 2008.

REIDSMA, D., D. HEYLEN, and H. J. A. OP DEN AKKER. On the contextual analysis of agreement scores. In MARTIN, J.-C., P. PAGGIO, M. KIPP, and D. HEYLEN, editors, *Proceedings of the LREC Workshop on Multimodal Corpora*, pages 52–55. ELRA, ELRA, May 2008a.

REIDSMA, D., D. HEYLEN, and R. ORDELMAN. Annotating emotions in meetings. In *Proc. of the fifth international conference on Language Resources and Evaluation, LREC 2006*, pages 1117–1122. European Language Resources Association, May 2006.

REIDSMA, D., D. H. W. HOFS, and N. JOVANOVIĆ. Designing focused and efficient annotation tools. In NOLDUS, L. P. J. J., F. GRIECO, L. W. S. LOIJENS, and P. H. ZIMMERMAN, editors, *Measuring Behaviour*, pages 149–152, Wageningen, NL, September 2005a.

REIDSMA, D., D. H. W. HOFS, and N. JOVANOVIĆ. A presentation of a set of new annotation tools based on the NXT API. Poster at Measuring Behaviour 2005, 2005b.

REIDSMA, D., A. NIJHOLT, and P. BOS. Temporal interaction between an artificial orchestra conductor and human musicians. *ACM Computers in Entertainment*, in press, ISSN 1544-3574, 2008b.

REIDSMA, D., R. J. RIENKS, and N. JOVANOVIĆ. Meeting modelling in the context of multimodal research. In BENGIO, S. and H. BOURLARD, editors, *Machine Learning for Multimodal Interaction, First International Workshop 2004, Revised Selected Papers*, volume 3361 of *Lecture Notes in Computer Science*, pages 22–35, Martigny, 2005c. Springer Verlag.

REIDSMA, D., Z. M. RUTTKAY, and A. NIJHOLT. *Challenges for Virtual Humans in Human Computing*, chapter 16, pages 316–338. Number 4451 in LNAI: State of the Art Surveys. Springer Verlag, Berlin/Heidelberg, June 2007.

REITHINGER, N. and M. KIPP. Large scale dialogue annotation in Verbmobil. In *Workshop Proceedings of ESSLLI 98*, 1998.

RIENKS, R. J. *Meetings in Smart Environments: Implications of progressing technology*. Phd thesis, SIKS Graduate School / University of Twente, Enschede, NL, June 2007.

RIENKS, R. J., A. NIJHOLT, and P. BARTHELMESS. Pro-active meeting assistants: attention please! *AI & Society, The Journal of Human-Centred Systems*, ISSN 0951-5666, 2007.

RIENKS, R. J., A. NIJHOLT, and D. REIDSMA. *Meetings and Meeting Support in Ambient Intelligence*, chapter 17, pages 359–378. Mobile communication series. Artech House, Norwood, MA, USA, 2006.

SHRIBERG, E., R. BATES, P. TAYLOR, A. STOLCKE, D. JURAFSKY, K. RIES, N. COCCARO, R. MARTIN, M. METEER, and C. VAN ESS-DYKEMA. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41 (3-4):443–492, ISSN 0023-8309, 1998.

SHRIBERG, E., R. DHILLON, S. BHAGAT, J. ANG, and H. CARVEY. The ICSI meeting recorder dialog act (MRDA) corpus. In STRUBE, M. and C. L. SIDNER, editors, *Proc. 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, April 2004.

SHRIBERG, E. and G. LOF. Reliability studies in broad and narrow phonetic transcription. *Clinical Linguistics and Phonetics*, 5:225–279, ISSN 0269-9206, 1991.

STEGMANN, J. and A. LÜCKING. Assessing reliability on annotations (1): Theoretical considerations. Technical Report, SFB360, Project B3 05-02, University of Bielefeld, 2005.

STEIDL, S., M. LEVIT, A. BATLINER, E. NÖTH, and H. NIEMANN. “of all things the measure is man” automatic classification of emotion and intra labeler consistency. In *ICASSP 2005, International Conference on Acoustics, Speech, and Signal Processing*, 2005.

VERBREE, D., R. J. RIENKS, and D. HEYLEN. Dialogue-act tagging using smart feature selection; results on multiple corpora. In *Proceedings of the IEEE Spoken Language Technology Workshop*, pages 70–73, December 2006.

VIEIRA, R. How to evaluate systems against human judgment on the presense of disagreement? In *Proc. workshop on joint evaluation of computational processing of Portugese at PorTAL 2002*, June 2002.

VOIGT, M. and R. STIEFELHAGEN. Visual focus of attention in dynamic meeting scenarios. In *Proceedings of Machine Learning for Multimodal Interaction ’07*, volume 5237 of *LNCS*, pages 1–13, Berlin / Heidelberg, 2008. Springer Verlag.

VOORHEES, E. M. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, ISSN 0306-4573, September 2000.

VOORHEES, E. M. and D. HARMAN. Overview of the fifth Text REtrieval Conference (TREC-5). In *Proceedings of the Fifth Text REtrieval Conference (TREC-5)*, pages 1–28. NIST, October 1997.

VYAS, D. and A. BAJART. Artefact ecologies: Supporting embodied meeting practices with distance access. In MULLER, H. and T. STRANG, editors, *Proceedings of UbiComp (Ubiquitous Computing) 2007 Workshops*, pages 117–122, Innsbruck, 2007. University of Innsbruck, Ubicomp.

WEBER, R. P. Measurement models for content analysis. *Quality and Quantity*, 17(2):127–149, ISSN 0033-5177, 1983.

WHITTAKER, S., S. TUCKER, K. SWAMPILLAI, and R. LABAN. Design and evaluation of systems to support interaction capture and retrieval. *Personal and Ubiquitous Computing*, 12(3):197–221, ISSN 1617-4909, March 2008.

WIEBE, J. M., R. F. BRUCE, and T. P. O'HARA. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 246–253, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

WILSON, T. Annotating subjective content in meetings. In *Proceedings of the Language Resources and Evaluation Conference (LREC-2008)*, 2008.

WITTEN, I. H. and E. FRANK. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

WITTGENSTEIN, L. *Philosophical Investigations*. 1953.

WRIGLEY, S. N., S. TUCKER, G. J. BROWN, and S. WHITTAKER. Effect of sound spatialisation on multitasking in remote meetings. In *Proceedings of Acoustics'08*, 2008.

Abstract

Researchers who make use of multimodal annotated corpora are always presented with something of a dilemma. On the one hand, one would prefer to have research results that are reproducible and independent of the particular annotators who produced the corpus that was used to obtain the results. A low level of inter-annotator agreement achieved on an annotation task implies a risk that this requirement is not met, especially if any disagreement between annotators was caused by them making errors in their task. On the other hand, many very interesting research issues concern phenomena for which annotation is an inherently subjective task. The judgements required of the annotators are then heavily dependent on the personal way in which the annotator views and interprets certain communicative behavior. In that case, the research results may become less easily reproducible, and certainly are no longer independent of the particular annotators who produced the corpus.

The usual practice in assessing whether a corpus is fit for the purpose for which it was constructed is to calculate the level of inter-annotator agreement, and when it exceeds a certain fixed threshold the data is considered to be of tolerable quality. There are two problems with this approach. Firstly, it depends on the assumption that any disagreement in the data is *not systematic*, but looks like noise. This assumption may not always be warranted. Secondly, the approach is not well suited for annotations that are subjective to a certain degree, as in that case annotator disagreement is (partly) an inherent property of the annotation, expressing something about the *level of intersubjectivity* between annotators in how they interpret certain communicative behavior versus the *amount of idiosyncrasy* in their judgements with respect to this behavior.

This thesis addresses both problems. In the theoretical part, it is shown that when disagreement is systematic, obtaining a certain level of inter-annotator agreement may indeed not be enough of a guarantee for the data being fit for its purpose. Simulations are used to investigate the effect of systematic disagreement on the relation between the level of inter-annotator agreement and the validity of machine-learning results obtained on the data. In the practical part, two new methods are explored for working with data that has been annotated with a low level of inter-annotator agreement. One method is aimed at finding a *subset* of the annotations that has been annotated more reliably, in a way that makes it possible to determine for new, unseen data whether it should belong to this subset — and therefore, whether a classifier trained on this more reliable subset is qualified to make a judgement for the new data. The other method is designed to use machine learning for explicitly modeling the overlap and disjunctions in the judgements of different annotators. Both methods put together should make it possible to build classifiers that, when deployed in a practical application, yield decisions that make sense for the human end user of the application, who indeed also may have his or her own way of interpreting the communicative behavior that is subjected to the classifier.

Samenvatting

Onderzoekers die werken met multimodale annotaties gemaakt voor audio- en video-opnames van conversaties tussen mensen worden vaak geconfronteerd met een dilemma. Onderzoeksresultaten moeten bij voorkeur reproduceerbaar zijn. Bovendien zouden ze niet afhankelijk moeten zijn van de specifieke individuen (waarnemers) die de annotaties gemaakt hebben. Als verschillende waarnemers voor dezelfde data, met dezelfde instructies, tot verschillende annotaties komen, is er sprake van "data met een lage overeenstemming tussen waarnemers". Zeker als dat gebrek aan overeenstemming veroorzaakt wordt doordat één of meer van de waarnemers fouten maken in het uitvoeren van hun taak wordt er niet aan bovengenoemde vereisten voldaan. Aan de andere kant is er veel interessant onderzoek waarbij de annotatietaak een subjectief oordeel van de kant van de waarnemer vereist. Het oordeel is dan afhankelijk van de persoonlijke manier waarop de waarnemer naar bepaalde vormen van communicatief gedrag kijkt en dergelijk gedrag interpreteert. In dat geval worden de onderzoeksresultaten minder reproduceerbaar. Bovendien zijn ze zeker niet meer onafhankelijk van de specifieke waarnemers die de annotaties gemaakt hebben.

De gebruikelijke aanpak om te bepalen of annotaties geschikt zijn voor het doel waarvoor ze gemaakt zijn is om te berekenen hoeveel overeenstemming er is tussen waarnemers; als dit meer is dan een bepaalde drempelwaarde wordt er van uitgegaan dat de annotaties goed genoeg zijn. Er zijn twee problemen met deze aanpak. Ten eerste gaat die uit van de veronderstelling dat verschillen tussen de waarnemers niet systematisch zijn, maar op ruis lijken. Deze veronderstelling is niet altijd terecht. Ten tweede is deze aanpak niet erg geschikt voor annotaties die met een bepaalde mate van subjectiviteit uitgevoerd moeten worden, omdat in dat geval de verschillen als inherente eigenschap van de annotaties iets zeggen over de *mate van intersubjectiviteit* tussen waarnemers in hoe ze bepaald communicatief gedrag interpreteren ten opzichte van de *persoonsspecifieke* manieren van interpreteren.

Dit proefschrift heeft betrekking op beide problemen. In het eerste, theoretische, deel wordt aangetoond dat, als de verschillen tussen waarnemers systematisch zijn, het bereiken van een bepaalde drempelwaarde in de overeenstemming tussen waarnemers niet noodzakelijkerwijs genoeg is om de bruikbaarheid van de annotaties te garanderen. Met behulp van simulaties wordt onderzocht wat het effect is van systematische verschillen in hoe waarnemers hun taak uitvoeren op de relatie tussen de mate van overeenkomst in hun annotaties en de validiteit van de resultaten van automatische herkenningsalgoritmes behaald op die annotaties. In het praktijkgerichte deel worden twee methodes onderzocht waarmee gewerkt kan worden met annotaties waarbij een lage mate van overeenstemming tussen waarnemers optreedt. E'e methode richt zich op het vinden van een *subset* van de annotaties die met meer overeenstemming geannoteerd is, op een manier die het mogelijk maakt om voor nieuwe data te bepalen of het tot die subset behoort — en daarmee of de automatische herkenningsmodule die op de data getraind was gekwalificeerd is een oordeel te vellen over deze nieuwe data. De andere methode is ontworpen om met behulp van automatische classificatietechnieken expliciet de overlap en het verschil te modelleren van de oordelen zoals de verschillende waarnemers die vellen. Samen make de twee methodes het mogelijk om classificatiemodules te maken die, ingezet in een praktsiche toepassing, oordelen zou moeten vellen op een manier waar de eindgebruiker van de toepassing iets mee kan, in het licht van het feit dat deze eindgebruiker ook zijn of haar eigen manier heeft om communicatief gedrag te beoordelen.

SIKS Dissertation Series

Since 1998, all dissertations written by Ph.D. students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series. This thesis is the 193th in the series.

2008-29 Dennis Reidsma (UT), *Annotations and Subjective Machines – Of Annotators, Embodied Agents, Users, and Other Humans*

2008-28 Ildiko Flesch (RUN), *On the Use of Independence Relations in Bayesian Networks*

2008-27 Hubert Vogten (OU), *Design and Implementation Strategies for IMS Learning Design*

2008-26 Marijn Huijbregts (UT), *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*

2008-25 Geert Jonker (UU), *Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency*

2008-24 Zharko Aleksovski (VU), *Using background knowledge in ontology matching*

2008-23 Stefan Visscher (UU), *Bayesian network models for the management of ventilator-associated pneumonia*

2008-22 Henk Koning (UU), *Communication of IT-Architecture*

2008-21 Krisztian Balog (UVA), *People Search in the Enterprise*

2008-20 Rex Arendsen (UVA), *Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven.*

2008-19 Henning Rode (UT), *From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search*

2008-18 Guido de Croon (UM), *Adaptive Active Vision*

2008-17 Martin Op 't Land (TUD), *Applying Architecture and Ontology to the Splitting and Allying of Enterprises*

2008-16 Henriëtte van Vugt (VU), *Embodied agents from a user's perspective*

2008-15 Martijn van Otterlo (UT), *The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains.*

2008-14 Arthur van Bunningen (UT), *Context-Aware Querying; Better Answers with Less Effort*

2008-13 Caterina Carraciolo (UVA), *Topic Driven Access to Scientific Handbooks*

2008-12 József Farkas (RUN), *A Semiotically Oriented Cognitive Model of Knowledge Representation*

2008-11 Vera Kartseva (VU), *Designing Controls for Network Organizations: A Value-Based Approach*

2008-10 Wauter Bosma (UT), *Discourse oriented summarization*

2008-09 Christof van Nimwegen (UU), *The paradox of the guided user: assistance can be counter-effective*

2008-08 Janneke Bolt (UU), *Bayesian Networks: Aspects of Approximate Inference*

2008-07 Peter van Rosmalen (OU), *Supporting the tutor in the design and support of adaptive e-learning*

2008-06 Arjen Hommersom (RUN), *On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective*

2008-05 Bela Mutschler (UT), *Modeling and simulating causal dependencies on process-aware information systems from a cost perspective*

2008-04 Ander de Keijzer (UT), *Management of Uncertain Data – towards unattended integration*

2008-03 Vera Hollink (UVA), *Optimizing hierarchical menus: a usage-based approach*

2008-02 Alexei Sharpanskykh (VU), *On Computer-Aided Methods for Modeling and Analysis of Organizations*

2008-01 Katalin Boer-Sorbán (EUR), *Agent-Based Simulation of Financial Markets: A modular, continuous-time approach*

2007-25 Joost Schalken (VU), *Empirical Investigations in Software Process Improvement*

2007-24 Georgina Ramírez Camps (CWI), *Structural Features in XML Retrieval*

2007-23 Peter Barna (TUE), *Specification of Application Logic in Web Information Systems*

2007-22 Zlatko Zlatev (UT), *Goal-oriented design of value and process models from patterns*

2007-21 Karianne Vermaas (UU), *Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005*

2007-20 Slinger Jansen (UU), *Customer Configuration Updating in a Software Supply Network*

2007-19 David Levy (UM), *Intimate relationships with artificial partners*

2007-18 Bart Orriëns (UvT), *On the development an management of adaptive business collaborations*

2007-17 Theodore Charitos (UU), *Reasoning with Dynamic Networks in Practice*

2007-16 Davide Grossi (UU), *Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems*

2007-15 Joyca Lacroix (UM), *NIM: a Situated Computational Memory Model*

2007-14 Niek Bergboer (UM), *Context-Based Image Analysis*

2007-13 Rutger Rienks (UT), *Meetings in Smart Environments; Implications of Progressing Technology*

2007-12 Marcel van Gerven (RUN), *Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic*

Decision-Making under Uncertainty

2007-11 Natalia Stash (TUE), *Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System*

2007-10 Huib Aldewereld (UU), *Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols*

2007-09 David Mobach (VU), *Agent-Based Mediated Service Negotiation*

2007-08 Mark Hoogendoorn (VU), *Modeling of Change in Multi-Agent Organizations*

2007-07 Nataša Jovanović (UT), *To Whom It May Concern – Addressee Identification in Face-to-Face Meetings*

2007-06 Gilad Mishne (UVA), *Applied Text Analytics for Blogs*

2007-05 Bart Schermer (UL), *Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance*

2007-04 Jurriaan van Diggelen (UU), *Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach*

2007-03 Peter Mika (VU), *Social Networks and the Semantic Web*

2007-02 Wouter Teepe (RUG), *Reconciling Information Exchange and Confidentiality: A Formal Approach*

2007-01 Kees Leune (UvT), *Access Control and Service-Oriented Architectures*

2006-28 Börkur Sigurbjörnsson (UVA), *Focused Information Access using XML Element Retrieval*

2006-27 Stefano Bocconi (CWI), *Vox Populi: generating video documentaries from semantically annotated media repositories*

2006-26 Vojkan Mihajlović (UT), *Score Region Algebra: A Flexible Framework for Structured Information Retrieval*

2006-25 Madalina Drugan (UU), *Conditional log-likelihood MDL and Evolutionary MCMC*

2006-24 Laura Hollink (VU), *Semantic Annotation for Retrieval of Visual Resources*

2006-23 Ion Juvina (UU), *Development of Cognitive Model for Navigating on the Web*

2006-22 Paul de Vrieze (RUN), *Fundaments of Adaptive Personalisation*

2006-21 Bas van Gils (RUN), *Aptness on the Web*

2006-20 Marina Velikova (UvT), *Monotone models for prediction in data mining*

2006-19 Birna van Riemsdijk (UU), *Cognitive Agent Programming: A Semantic Approach*

2006-18 Valentin Zhizhkun (UVA), *Graph transformation for Natural Language Processing*

2006-17 Stacey Nagata (UU), *User Assistance for Multitasking with Interruptions on a Mobile Device*

2006-16 Carsten Riggelsen (UU), *Approximation Methods for Efficient Learning of Bayesian Networks*

2006-15 Rainer Malik (UU), *CONAN: Text Mining in the Biomedical Domain*

2006-14 Johan Hoorn (VU), *Software Requirements: Update, Upgrade, Redesign – towards a Theory of Requirements Change*

2006-13 Henk-Jan Lebbink (UU), *Dialogue and Decision Games for Information Exchanging Agents*

2006-12 Bert Bongers (VU), *Interactivation – Towards an e-cology of people, our technological environment, and the arts*

2006-11 Joeri van Ruth (UT), *Flattening Queries over Nested Data Types*

2006-10 Ronny Siebes (VU), *Semantic Routing in Peer-to-Peer Systems*

2006-09 Mohamed Wahdan (UM), *Automatic Formulation of the Auditor's Opinion*

2006-08 Eelco Herder (UT), *Forward, Back and Home Again – Analyzing User Behavior on the Web*

2006-07 Marko Smiljanic (UT), *XML schema matching – balancing efficiency and effectiveness by means of clustering*

2006-06 Ziv Baida (VU), *Software-aided Service Bundling – Intelligent Methods & Tools for Graphical Service Modeling*

2006-05 Cees Pierik (UU), *Validation Techniques for Object-Oriented Proof Outlines*

2006-04 Marta Sabou (VU), *Building Web Service Ontologies*

2006-03 Noor Christoph (UVA), *The role of metacognitive skills in learning to solve problems*

2006-02 Cristina Chisalita (VU), *Contextual issues in the design and use of information technology in organizations*

2006-01 Samuil Angelov (TUE), *Foundations of B2B Electronic Contracting*

2005-21 Wijnand Derks (UT), *Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics*

2005-20 Cristina Coteanu (UL), *Cyber Consumer Law, State of the Art and Perspectives*

2005-19 Michel van Dartel (UM), *Situated Representation*

2005-18 Danielle Sent (UU), *Test-selection strategies for probabilistic networks*

2005-17 Boris Shishkov (TUD), *Software Specification Based on Re-usable Business Components*

2005-16 Joris Graaumans (UU), *Usability of XML Query Languages*

2005-15 Tibor Bosse (VU), *Analysis of the Dynamics of Cognitive Processes*

2005-14 Borys Omelayenko (VU), *Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics*

2005-13 Fred Hamburg (UL), *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen*

2005-12 Csaba Boer (EUR), *Distributed Simulation in Industry*

2005-11 Elth Ogston (VU), *Agent Based Matchmaking and Clustering – A Decentralized Approach to Search*

2005-10 Anders Bouwer (UVA), *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments*

2005-09 Jeen Broekstra (VU), *Storage, Querying and Inferencing for Semantic Web Languages*

2005-08 Richard Vdovjak (TUE), *A Model-driven Approach for Building Distributed Ontology-based Web Applications*

2005-07 Flavius Frasincar (TUE), *Hypermedia Presentation Generation for Semantic Web Information Systems*

2005-06 Pieter Spronck (UM), *Adaptive Game AI*

2005-05 Gabriel Infante-Lopez (UVA), *Two-Level Probabilistic Grammars for Natural Language Parsing*

2005-04 Nirvana Meratnia (UT), *Towards Database Support for Moving Object data*

2005-03 Franc Grootjen (RUN), *A Pragmatic Approach to the Conceptualisation of Language*

2005-02 Erik van der Werf (UM)), *AI techniques for the game of Go*

2005-01 Floor Verdenius (UVA), *Methodological Aspects of Designing Induction-Based Applications*

2004-20 Madelon Evers (Nyenrode), *Learning from Design: facilitating multidisciplinary design teams*

2004-19 Thijs Westerveld (UT), *Using generative probabilistic models for multimedia retrieval*

2004-18 Vania Bessa Machado (UvA), *Supporting the Construction of Qualitative Knowledge Models*

2004-17 Mark Winands (UM), *Informed Search in Complex Games*

2004-16 Federico Divina (VU), *Hybrid Genetic Relational Search for Inductive Learning*

2004-15 Arno Knobbe (UU), *Multi-Relational Data Mining*

2004-14 Paul Harrenstein (UU), *Logic in Conflict. Logical Explorations in Strategic Equilibrium*

2004-13 Wojciech Jamroga (UT), *Using Multiple Models of Reality: On Agents who Know how to Play*

2004-12 The Duy Bui (UT), *Creating emotions and facial expressions for embodied agents*

2004-11 Michel Klein (VU), *Change Management for Distributed Ontologies*

2004-10 Suzanne Kabel (UVA), *Knowledge-rich indexing of learning-objects*

2004-09 Martin Caminada (VU), *For the Sake of the Argument; explorations into argument-based reasoning*

2004-08 Joop Verbeek (UM), *Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politile gegevensuitwisseling en digitale expertise*

2004-07 Elise Boltjes (UM), *Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes*

2004-06 Bart-Jan Hommes (TUD), *The Evaluation of Business Process Modeling Techniques*

2004-05 Viara Popova (EUR), *Knowledge discovery and monotonicity*

2004-04 Chris van Aart (UVA), *Organizational Principles for Multi-Agent Architectures*

2004-03 Perry Groot (VU), *A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving*

2004-02 Lai Xu (UvT), *Monitoring Multi-party Contracts for E-business*

2004-01 Virginia Dignum (UU), *A Model for Organizational Interaction: Based on Agents, Founded in Logic*

2003-18 Levente Kocsis (UM), *Learning Search Decisions*

2003-17 David Jansen (UT), *Extensions of Statecharts with Probability, Time, and Stochastic Timing*

2003-16 Menzo Windhouwer (CWI), *Feature Grammar Systems – Incremental Maintenance of Indexes to Digital Media Warehouses*

2003-15 Mathijs de Weerdt (TUD), *Plan Merging in Multi-Agent Systems*

2003-14 Stijn Hoppenbrouwers (KUN), *Freezing Language: Conceptualisation Processes across ICT-Supported Organisations*

2003-13 Jeroen Donkers (UM), *Nosce Hostem – Searching with Opponent Models*

2003-12 Roeland Ordelman (UT), *Dutch speech recognition in multimedia information retrieval*

2003-11 Simon Keizer (UT), *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks*

2003-10 Andreas Lincke (UvT), *Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture*

2003-09 Rens Kortmann (UM), *The resolution of visually guided behaviour*

2003-08 Yongping Ran (UM), *Repair Based Scheduling*

2003-07 Machiel Jansen (UvA), *Formal Explorations of Knowledge Intensive Tasks*

2003-06 Boris van Schooten (UT), *Development and specification of virtual environments*

2003-05 Jos Lehmann (UVA), *Causation in Artificial Intelligence and Law – A modelling approach*

2003-04 Milan Petković (UT), *Content-Based Video Retrieval Supported by Database Technology*

2003-03 Martijn Schuemie (TUD), *Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy*

2003-02 Jan Broersen (VU), *Modal Action Logics for Reasoning About Reactive Systems*

2003-01 Heiner Stuckenschmidt (VU), *Ontology-Based Information Sharing in Weakly Structured Environments*

2002-17 Stefan Manegold (UVA), *Understanding, Modeling, and Improving Main-Memory Database Performance*

2002-16 Pieter van Langen (VU), *The Anatomy of Design: Foundations, Models and Applications*

2002-15 Rik Eshuis (UT), *Semantics and Verification of UML Activity Diagrams for Workflow Modelling*

2002-14 Wieke de Vries (UU), *Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems*

2002-13 Hongjing Wu (TUE), *A Reference Architecture for Adaptive Hypermedia Applications*

2002-12 Albrecht Schmidt (Uva), *Processing XML in Database Systems*

2002-11 Wouter C.A. Wijngaards (VU), *Agent Based Modelling of Dynamics: Biological and Organisational Applications*

2002-10 Brian Sheppard (UM), *Towards Perfect Play of Scrabble*

2002-09 Willem-Jan van den Heuvel (KUB), *Integrating Modern Business Applications with Objectified Legacy Systems*

2002-08 Jaap Gordijn (VU), *Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas*

2002-07 Peter Boncz (CWI), *Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications*

2002-06 Laurens Mommers (UL), *Applied legal epistemology; Building a knowledge-based ontology of the legal domain*

2002-05 Radu Serban (VU), *The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents*

2002-04 Juan Roberto Castelo Valdueza (UU), *The Discrete Acyclic Digraph Markov Model in Data Mining*

2002-03 Henk Ernst Blok (UT), *Database Optimization Aspects for Information Retrieval*

2002-02 Roelof van Zwol (UT), *Modelling and searching web-based document collections*

2002-01 Nico Lassing (VU), *Architecture-Level Modifiability Analysis*

2001-11 Tom M. van Engers (VUA), *Knowledge Management: The Role of Mental Models in Business Systems Design*

2001-10 Maarten Sierhuis (UvA), *Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design*

2001-09 Pieter Jan 't Hoen (RUL), *Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes*

2001-08 Pascal van Eck (VU), *A Compositional Semantic Structure for Multi-Agent Systems Dynamics.*

2001-07 Bastiaan Schonhage (VU), *Diva: Architectural Perspectives on Information Visualization*

2001-06 Martijn van Welie (VU), *Task-based User Interface Design*

2001-05 Jacco van Ossenbruggen (VU), *Processing Structured Hypermedia: A Matter of Style*

2001-04 Evgueni Smirnov (UM), *Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets*

2001-03 Maarten van Someren (UvA), *Learning as problem solving*

2001-02 Koen Hindriks (UU), *Agent Programming Languages: Programming with Mental Models*

2001-01 Silja Renooij (UU), *Qualitative Approaches to Quantifying Probabilistic Networks*

2000-11 Jonas Karlsson (CWI), *Scalable Distributed Data Structures for Database Management*

2000-10 Niels Nes (CWI), *Image Database Management System Design Considerations, Algorithms and Architecture*

2000-09 Florian Waas (CWI), *Principles of Probabilistic Query Optimization*

2000-08 Veerle Coupé (EUR), *Sensitivity Analyis of Decision-Theoretic Networks*

2000-07 Niels Peek (UU), *Decision-theoretic Planning of Clinical Patient Management*
2000-06 Rogier van Eijk (UU), *Programming Languages for Agent Communication*
2000-05 Ruud van der Pol (UM), *Knowledge-based Query Formulation in Information Retrieval.*
2000-04 Geert de Haan (VU), *ETAG, A Formal Model of Competence Knowledge for User Interface Design*
2000-03 Carolien M.T. Metselaar (UVA), *Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief.*
2000-02 Koen Holtman (TUE), *Prototyping of CMS Storage Management*
2000-01 Frank Niessink (VU), *Perspectives on Improving Software Maintenance*
1999-08 Jacques H.J. Lenting (UM), *Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.*
1999-07 David Spelt (UT), *Verification support for object database design*
1999-06 Niek J.E. Wijngaards (VU), *Re-design of compositional systems*
1999-05 Aldo de Moor (KUB), *Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems*
1999-04 Jacques Penders (UM), *The practical Art of Moving Physical Objects*
1999-03 Don Beal (UM), *The Nature of Minimax Search*
1999-02 Rob Potharst (EUR), *Classification using decision trees and neural nets*
1999-01 Mark Sloof (VU), *Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products*
1998-05 E.W. Oskamp (RUL), *Computerondersteuning bij Straftoemeting*
1998-04 Dennis Breuker (UM), *Memory versus Search in Games*
1998-03 Ans Steuten (TUD), *A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective*
1998-02 Floris Wiesman (UM), *Information Retrieval by Graphically Browsing Meta-Information*
1998-01 Johan van den Akker (CWI), *DEGAS – An Active, Temporal Database of Autonomous Objects*

Propositions, supplemental to the PhD thesis

Annotations and Subjective Machines

Of Annotators, Embodied Agents, Users, and Other Humans

by

DENNIS REIDSMA

1. The practice of applying a fixed threshold to Krippendorff's alpha in order to assess whether the quality of an annotation is good enough implies the assumption that inter-annotator agreement values can be compared between annotation tasks.

2. The fact that a person does not agree with the majority opinion does not imply that he is wrong.[1]

3. But not every disagreement is a matter of subjectivity, either.[2]

4. Getting a careful and conscientious review for a paper, submitted to a conference or journal, can have a fundamental impact on the course of a research project. The review does not need to be nice. It does not have to lead to acceptance of the paper. The best reviews are those that inspire the author to ask new questions that he or she wants to answer.

5. Even if annotated data has an adequate inter-annotator agreement, there is no clear reason why machine learning applied to that data should learn its model mostly from the agreed annotations rather than mostly from the disagreed annotations, especially if disagreement in the annotations is systematic. It is therefore important to spend enough time investigating the generalizability of machine-learning results in an application context.[3]

6. Writing a paper, and designing a software system, are very similar activities. For both, one can think about the end goal, the functions of the product, and its requirements. On the other hand, both activities can also be carried out as an exploration of the space of possibilities, without knowing the intended outcome beforehand.

[1] Chapter 1
[2] Chapter 6
[3] Chapter 3

7. Information about which person annotated a certain part of a corpus is important, and needs to be stored along with the corpus.[4]

8. The comics of Jorge Cham[5] are sometimes painfully accurate.

9. Distractions and detours tend to turn out to have been part of the principal research topic after all.

10. Given that nonverbal synchrony is such a powerful and positive element in cooperative communication behavior between humans, it is a worthwhile endeavour to teach computers to adapt their interactive performance to the natural working rhythms of their users.[6]

Dennis Reidsma
Enschede, October 2008

[4] Chapter 8

[5] `http://www.phdcomics.com/`

[6] A. Nijholt, D. Reidsma, H. van Welbergen, H. J. A. op den Akker, and Z. M. Ruttkay (2008) "Mutually Coordinated Anticipatory Multimodal Interaction", in: A. Esposito, N. Bourbakis, N. Avouris, and I. Hatzilygeroudis (eds.) Nonverbal Features of Human-Human and Human-Machine Interaction, pages 73–93, Springer Verlag, Berlin